Dynamically Forecasting Airline Departure Delay Probability Distributions for Individual Flights using Supervised Learning

MSc Thesis

Maarten Beltman

TU Delft Delft University of Technology

Dynamically Forecasting Airline Departure Delay Probability Distributions for Individual Flights using Supervised Learning

by

Maarten Beltman

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday December 12th, 2023.

Student number:	4772350	
Project duration:	February 2023 – December 2023	
Thesis committee:	Dr. B.F. Lopes Dos Santos	Delft University of Technology
	Dr. M.J. Ribeiro	Delft University of Technology
	Dr. J. Sun	Delft University of Technology
	Dr. N.A.K. Doan	Delft University of Technology
	Ir. J. de Wilde	KLM Royal Dutch Airlines
	Ir. C. Evertse	KLM Royal Dutch Airlines

Acknowledgements

While writing this report, I was reflecting on my time as a Bachelor and Master student at Delft University of Technology. In 2018, I left the small village to move to the *'big city'*. Soon enough, I found out that the *'big city'* was not that *'big'* after all. For me, Delft was the perfect place to spend my previous five years.

I am very thankful to have been given the opportunity to perform my internship and MSc thesis projects at KLM. During my time at KLM, I have been able to catch a glimpse of the aviation industry from up close. It was true when they told me there is never a dull moment in the aviation industry. I am very thankful to Jasper de Wilde for his daily supervision and constructive feedback. I also want to thank Christiaan Evertse and Frans Huisman for their involvement in the project. Furthermore, a special thanks goes out to my fellow graduate intern Frans Vossen, whom I've had a great time with at the company.

The thesis project would not have been possible without the support of Marta Ribeiro and Junzi Sun from TU Delft. Our densely packed update meetings steered me towards the right direction. I would like to especially thank Marta Ribeiro for providing her constructive feedback, for both the draft literature study report and the draft thesis paper. Furthermore, I would like to express my gratitude towards Bruno Santos for his support in the progress sessions.

A final word of *'thank-you!'* goes out to my family, friends and girlfriend for their support and interest in the project. Now, it is time for me to relax!

Maarten Beltman
Delft, December 2023

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

ADS-B	Automatic Dependent Surveillance-Broadcast
AIBT	Actual In-Block Time
ALDT	Actual Landing Time
AOBT	Actual Off-Block Time
ARIMA	Autoregression Integrated Moving Average
ARMA	AutoRegression Moving Average
ATC	Air Traffic Control
ATFM	Air Traffic Flow Management
ATM	Air Traffic Management
ATOT	Actual Take-Off Time
AUC	Area Under Curve
BTS	Bureau of Transportation Statistics
CART	Classification and Regression Trees
CDM	Collaborative Decision Making
CG	Center of Gravity
COVID-19	Coronavirus Disease 2019
CPCLS	Cascade Principal Component Least Squares
CRPS	Continuous Ranked Probability Score
DBN	Deep Belief Network
DST-GAT	Dynamic Spatial-Temporal Graph Attention
DT	Decision Tree
EOBT	Estimated Off-Block Time
ETFMS	Enhanced Tactical Flow Management System
ETOT	Estimated Take-Off Time
FAA	Federal Aviation Administration
FF	Feed-Forward
FN	False Negative
FP	False Positive
FPR	False Positive Rate

GBDT	Gradient Boosted Decision Tree
GNB	Gaussian Naive Bayes
GPS	Global Positioning System
GPU	Ground Power Unit
IATA	International Air Transport Association
ICAO	International Civil Aviation Organization
ID	Identification
IQR	Inter-Quartile Range
KLM	Koninklijke Luchtvaart Maatschappij
kNN	k-Nearest Neighbor
LM	Levenberg-Marquardt
LR	Linear Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MDN	Mixed Density Network
MIDT	Marketing Information Data Tapes
MILP	Mixed Integer Linear Programming
MJLS	Markov Jump Linear System
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MVE	Mean Variance Estimation
NAS	National Aviation System
NN	Neural Network
OD	Origin-Destination
PAX	Passengers
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
Q1	25th percentile
Q3	75th percentile
RF	Random Forest
RF	Random Forests
RFE	Recursive Feature Elimination
RM	Revenue Management
RMSE	Root Mean Square Error

RNN	Recurrent Neural Network
ROC	Receiving Operator Characteristic
SIBT	Scheduled In-Block Time
SLDT	Scheduled Landing Time
SMOTE	Synthetic Minority Oversampling Technique
SOBT	Scheduled Off-Block Time
STOT	Scheduled Take-Off Time
SVM	Support Vector Machine
SVR	Support Vector Regression
SVR	Support Vector Regression
TN	True Negative
TOBT	Target Off-Block Time
TP	True Positive
TPR	True Positive Rate
TSAT	Target Start-up Approval Time
UAC	Upper Area Control
ULD	Unit Load Devices
US	United States
WP	Work Package

Symbols

β	Scaling Factor for F_β computation	-
$\delta_{\sigma_r(f)}$	Delay of Previous Flight	min
ϵ	Residual	-
$\hat{p}$	Probability	-
$\hat{y}$	Estimated Output	-
ϕ	Neural Network Activation Function	-
ρ_{rf}	Minimum Required Turnaround Time	min
σ	Sigmoid Function	-
τ_{rf}	Slack Time	min
θ	Model Parameter Vector	-
cr_{rf}	Scheduled Turnaround Time	min
$d(x,y)$	k-Nearest Neighbor Distance	-
D_{arr}	Arrival Delay	min
D_{dep}	Departure Delay	min

f	Flight	-
k	Number of Nearest Neighbors	-
$P(X)$	Probability of X	-
q	Minkowski Power	-
r	Rotation	-
R^2	Coefficient of Determination	-
SS_{RES}	Sum of Squares of Residuals	-
SS_{TOT}	Total Sum of Squares	-
x	Input Vector	-

Introduction

The airline industry is highly competitive, with small profit margins and complex flight schedules. Potential flight delays are not only frustrating for passengers but also costly for airlines. European law EU261 forces airlines to compensate passengers for delays, both monetarily and in the form of hotel accommodation and transport for stranded passengers. Additionally, due to experienced discomfort, passengers may be reluctant to fly the airline for trips yet to be booked, leading to a loss of future value.

There are additional strong internal incentives for airlines to minimize departure delays. To ensure on-time arrivals, flights that were delayed at departure have to compensate for the lost time whilst airborne, thereby increasing fuel consumption. This is particularly true for hub-and-spoke airlines, to maintain passenger connections at the arrival airport.

The goal of this MSc Thesis project is to develop explainable supervised learning models that forecast the probability distributions of departure delays for individual flights at multiple moments relative to their scheduled departure times. Doing so allows flight dispatchers to better determine the planned fuel amounts. For larger departure delays, generally more fuel is required to compensate for the delay. Given the costs and emissions associated with the decisions made following the departure delay forecasting model, the models should explain the reasoning behind the model's predictions and provide context to the flight dispatcher.

This study is unique because of the research collaboration with KLM Royal Dutch Airlines, who provide access to their extensive databases. This allows to explore a multitude of novel flight characteristics which are potentially useful for forecasting departure delays. Furthermore, through the collaboration with KLM, expertise is directly available because of their daily supervision. Combining this with academic supervision from Delft University of Technology allows to perform scientific research on a subject that is interesting from both an academic and business point of view.

This thesis report is split up in two parts. Although both parts may overlap, they were performed as separate pieces of research. Part I presents the scientific paper on dynamic probabilistic departure delay forecasting. Part II contains the relevant literature study that was conducted during the first phase of the MSc Thesis.

I

Scientific Paper

Delft University of Technology, Aerospace Engineering (2023), 1–33
MSc Thesis, Control & Simulation (C&S)

ARTICLE

Dynamically Forecasting Airline Departure Delay Probability Distributions for Individual Flights using Supervised Learning

Maarten Beltman[*,1,2]

[1]Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands
[2]KLM Royal Dutch Airlines, Amstelveen, The Netherlands
[*]Corresponding author

Daily Supervision by Dr. Marta Ribeiro[1], Dr. Junzi Sun[1] and Ir. Jasper de Wilde[2]
Supporting Supervision by Dr. Bruno Lopes dos Santos[1] and Ir. Christiaan Evertse[2]

Abstract
Punctuality is a key performance indicator for any airline. Hub-and-spoke airlines are particularly committed to on-time arrivals to guarantee passenger connections. Flights that are delayed at departure need to compensate for the lost time whilst airborne. Because fueling takes place well before scheduled departure, predicted departure delays determine the planned fuel amounts for en-route speed optimization. To prevent unnecessary fuel burn, airlines benefit from highly accurate departure delay predictions. This study aims to extend previous work on airline departure delay forecasting to a dynamic and probabilistic domain, whilst incorporating novel day-of-operations airline information to further minimize prediction errors. Random Forest, CatBoost, and Deep Neural Network models were proposed for a case study on KLM departures from Amsterdam Airport Schiphol between 1 January 2020 and 1 August 2023. The Random Forest model was selected for its superior probabilistic performance and high accuracy in predicting delays between 5 and 25 minutes, for which en-route speed optimization has the largest effect. The departure delay probability distribution forecasts are made at six distinct prediction moments: 90, 75, 60, 45, 30, and 15 minutes before scheduled departure time. At the 90-minute prediction horizon, the model reaches a Mean Absolute Error (MAE) of 8.46 minutes and a Root Mean Square Error (RMSE) of 11.91 minutes. Simultaneously, for 76% of flights, the actual delay is within the predicted probability distribution range. Considering the costs and emissions associated with the decision-making following the departure delay prediction model, this study puts strong emphasis on explainability. Flight dispatchers are therefore provided not only the predicted departure delay but also the main factors impacting the prediction, explaining the context of the flight. The versatility of the model was demonstrated in two shadow runs, where delays caused by familiar and unfamiliar factors were successfully predicted.

Keywords: Airline Operations; Dynamic Probabilistic Forecasting; Explainability; Flight Departure Delay; Flight Planning; Hub-and-Spoke Airline; Regression Problem; Supervised Learning

Abbreviations: ATFM: Air Traffic Flow Management; ETA: Estimated Time of Arrival; GBDT: Gradient Boosted Decision Trees; IQR: Inter-Quartile Range; MAE: Mean Absolute Error; METAR: Meteorological Aerodrome Report; R2: Coefficient of Determination; RMSE: Root Mean Square Error; TAF: Terminal Aerodrome Forecast; TAT: Turn-around Time; TSAT: Target Start-up Approval Time

1. Introduction

Punctuality is a key performance indicator for airlines, especially those utilizing hub-and-spoke business models. Such airlines offer short passenger connections that yield competitive advantages but require high punctuality. A significant factor that affects punctuality is the departure delay of the flight, the time difference between scheduled and actual departure times. Departure delays not only affect a current flight but possibly also that of subsequent flights in the airline's network, potentially affecting operations for the entire day.

Delays are not only frustrating for passengers, but also costly for airlines. European law EU261 forces airlines to compensate passengers for delays, both monetarily and in the form of hotel accommodation and transport for stranded passengers. Additionally, due to experienced discomfort, passengers may be reluctant to fly the airline for trips yet to be booked, leading to a loss of future value. There are additional strong internal incentives for

Thesis paper written by Maarten Beltman to obtain the Master of Science at Delft University of Technology, to be defended publicly on the 12th of December 2023. An electronic version of this thesis is available at https://repository.tudelft.nl/.

airlines to minimize departure delays. To ensure on-time arrivals, flights that were delayed at departure have to compensate for the lost time whilst airborne, thereby increasing fuel consumption. This is particularly true for hub-and-spoke airlines, to maintain passenger connections at the arrival airport.

The final flight plan, which indicates the required fuel amounts, is usually issued by the airline's flight dispatcher around 90 minutes before scheduled departure time. Accurate departure delay predictions greatly benefit the fuel amount calculations. In case of underpredicted departure delays, insufficient fuel is carried for compensating time, possibly leading to the loss of passenger connections. Alternatively, in the case of overpredicted departure delays, excessive fuel is carried, leading to unnecessary fuel burn as a result of increased aircraft mass.

This study aims to extend current departure delay prediction models towards a dynamic and probabilistic domain, thereby forecasting departure delay probability distributions for individual flights at different moments relative to their scheduled departure times. Simultaneously, this study aims to further minimize prediction errors by exploring novel flight characteristics, available from the airline's detailed flight dataset. Given the importance of the decisions that follow from the model predictions, this study places strong emphasis on explainability by outlining the reasoning behind the predictions of the model.

The remainder of the paper is structured as follows. A brief literature overview on the topic of airline departure delay forecasting is presented in Section 2, where the research gaps are also outlined. Thereafter, the methodology adhered to in this study is elaborated upon in Section 3. The results of the proposed models are then presented in Section 4, where one of the models is selected. The model validation is performed in Section 5. Subsequently, the results are discussed in Section 6. Finally, the conclusion and recommendations are presented in Section 7.

2. Literature Review

The topic of airline departure delay forecasting has been studied since the 1960s, where early literature provided qualitative explanations for the existence of such delays. Since the new millennium, a shift was noticed towards quantitative research, mainly based on statistical methods. Mueller and Chatterji [1] found that departure delay probability distributions were best modelled using Poisson distributions and that arrival delay probability distributions better fitted Normal distributions. In early quantitative research, departure delays were generally considered to be a sum of temporal components. Abdel-Aty et al. [2] proposed a model building on daily, monthly, and seasonal patterns and Tu et al. [3] summed daily and seasonal patterns with a residual.

Historical data shows that departure delays are not evenly distributed: the majority of flights experience minor delays whereas only a few flights are delayed more significantly. This unevenness, referred to as *positive* skew, was the motivation for Pérez-Rodríguez et al. [4] to compare symmetric and asymmetric Bayesian logistic models for predicting flight delays. The skewed nature of the dataset favoured the performance of the latter model.

From the late 2010s onwards, the use of supervised learning approaches has become increasingly popular for the prediction of departure delays. Such approaches make use of large-dimensional datasets, consisting of numerous features that describe flight characteristics. This allows to either predict the class that the delay of a given flight belongs to (*classification*) or the magnitude of the delay for the respective flight (*regression*). Although a large share of papers studied classification approaches, the following paragraphs will focus on regression approaches only, since this is the approach that is taken in this study.

2.1 Departure Delay Forecasting using Regression Approaches

In previous work on departure delay prediction using supervised learning, the applicability of both tree-based models and neural network models has been studied. For tree-based models, Kalliguddi and Leboulluec [5] and Khan [6], concluded that simple decision trees were outperformed by random forests. At the same time, Manna et al. [7] showed that random forests, in their turn, were outperformed by boosting models. Ye et al. [8] further investigated boosting models, specifically Extra-Trees and LightGBM, by introducing features that describe accumulated airport characteristics. Examples of such features are the number of aircraft on-ground and their respective averaged delays in some given time interval. Incorporating these features resulted in an improvement of the model performance compared to using only features that describe individual flights.

With the aim of evaluating the performance of EUROCONTROL's Enhanced Tactical Flow Management System (ETFMS) against a supervised learning model, Dalmau et al. [9] proposed a different boosting model: Gradient-

Boosted Decision Trees (GBDT). Using a large number of features (over 30), it was found that the existing system is outperformed by the GBDT model, especially for prediction horizons larger than 60 minutes.

Vorage [10] extended departure delay prediction towards the probabilistic domain. Random Forests and Mixture Density Networks were developed to generate probability density functions for individual flights from Amsterdam Airport Schiphol. From these distributions, the probability of a forecasted delay to be accurate within some time-error interval could be computed. Later, this approach was used by Zoutendijk and Mitici [11], constructing similar models to predict departure delays using Rotterdam Airport flight data. The models reached a Mean Absolute Error (MAE) of around 12.5 minutes.

The work of Sun et al. [12] aimed to predict airline delays from a network perspective, testing the applicability of several neural networks including a Dynamic Spatial-Temporal Graph Attention (DST-GAT) network and a Long Short-Term Memory (LSTM) network. Whilst the network architectures differed significantly, the outcomes for both models were comparable, with Root Mean Square Error (RMSE) values between 5-10 minutes, differing per airport in the network.

Finally, Birolini and Jacquillat [13] collaborated with European airline Vueling, comparing the performance of linear regression, random forest, and Extreme Gradient boosting models for predicting the airline's flight delays. For each flight, only the primary flight delay[1] was considered, eliminating the effect of precedent flights. Taking into account airline-specific information, such as crew rosters and aircraft availability, the Extreme Gradient boosting model reached an MAE of approximately 7 minutes, outperforming the other considered models.

2.2 Research Gaps and Contribution of this Paper

Most reviewed literature considered temporal features, flight schedule features, and weather features. Among others, Sternberg et al. [14] demonstrated that including weather features benefits the model performance. Other novel features such as flight de-icing status [9] and take-off runway [6, 15] were proposed. Moreover, Yu et al. [16] especially focused on short-term features, including the boarding option (jet bridge or bus), closing time of cargo doors and passenger doors and the time between check-in, boarding, and gate closure.

Only very few papers considered passenger data. As previously explained, connecting passengers are of high importance for hub-and-spoke airlines. Only two papers considered passenger connection information. Ciruelos et al. [17] assumed monthly connecting passenger percentages, thereby not specifying the numbers per individual flight. Sismanidou et al. [18] had access to real passenger itineraries from a Marketing Information Data Tapes (MIDT) dataset, but used this data to determine "*a proportion of connecting passengers for a specific itinerary by a specific air carrier*", therefore also averaging the connecting passenger numbers, missing out on the opportunity to use the flight-specific passenger connection data for predicting the departure delays. The research gap of flight-specific passenger connection information thus remains, and can be described as follows:

> *To the best of the author's knowledge, the planned number of connection passengers for every unique inbound to outbound flight combination has not been considered in any previous research on airline departure delay forecasting.*

Furthermore, only few studies had access to detailed airline data, through research partnerships with Peach and Vueling respectively [13, 19]. Despite these partnerships, the studies refrain from proposing detailed day-of-operations features, but instead hold on to mostly booking and schedule information. Thus, there remains a research gap for the effect of day-of-operation features, such as last-minute airframe assignments, Target Start-up Approval Time (TSAT) changes, and airport delay levels. This second research gap is described as follows:

> *To the best of the author's knowledge, there are novel airline day-of-operations features, likely to improve the model performance, that have not been considered in previous research on airline departure delay forecasting.*

The third and final research gap that was identified covers the time aspect of the forecasting process. Whereas for the vast majority of papers (around 85% of reviewed literature) only a single forecasting moment was considered, several papers compared the performance of the model at multiple moments relative to scheduled departure times [9, 12, 20, 21, 22, 23]. While these six papers compared the forecasts over time, all of them did so in a deterministic manner. Although probabilistic departure delay forecasts for individual flights were proposed [10, 11], it has not

[1] The part of the departure delay originating only from the considered outbound flight, therefore eliminating effects of previous flights.

yet been investigated how such probabilistic forecasts change over time. The concept of dynamic probabilistic forecasts was previously used by Felder et al. [24] in the domain of dynamic wind power forecasting. No such research has yet been performed for the prediction of departure delays of individual flights, however. For that reason, the third research gap is described as follows:

> *To the best of the author's knowledge, it has not yet been investigated how probabilistic departure delay forecasts for individual flights change over time relative to scheduled departure given the availability of new data.*

3. Methodology

This section aims to describe the methodology adopted for this study. First, the case study is introduced in Section 3.1. Thereafter, the data preprocessing is discussed in Section 3.2. The model development process is then elaborated upon in Section 3.3. Finally, the feature engineering processes are explained in Section 3.4 and the model training and result processing are discussed in Section 3.5.

3.1 Description of the Case Study

Several departure delay prediction methods are proposed and tested through a case study on data provided by KLM Royal Dutch Airlines (*in the remainder referred to as KLM*), for departing flights from its hub airport, Amsterdam Airport Schiphol (*in the remainder referred to as Schiphol*). KLM is the largest airline in the Netherlands and operated flights to over 170 destinations in the summer of 2023 [25]. The case study considers flights operated both by KLM and its subsidiary KLM Cityhopper, which operates the regional flights.

Given the data availability, the case study is performed for the period between 1 January 2020 and 1 August 2023. An overview of the KLM traffic and delays at Schiphol for this period is presented in Figure 1. This period includes various notable moments, all of which influence the model outcomes differently. In the period between March 2020 and January 2022, as a result of the COVID pandemic, the number of departing flights lowered drastically, leading to significantly smaller departure delays. In the period after, the demand for air travel grew faster than anticipated, leading to logistical problems that resulted in above-average departure delays [26].

Most of the logistical problems were solved from November 2022 onwards, which translated into traffic and delay levels almost back at pre-COVID standards, still impacted by seasonality. Including the data of these significantly different dynamics turns out to be beneficial as it allows the model to be trained on a wider variety of historical data. The presence of more flights towards the extremes of the departure delay spectrum simultaneously improves the sampling practices, which is discussed in Section 3.2.2.

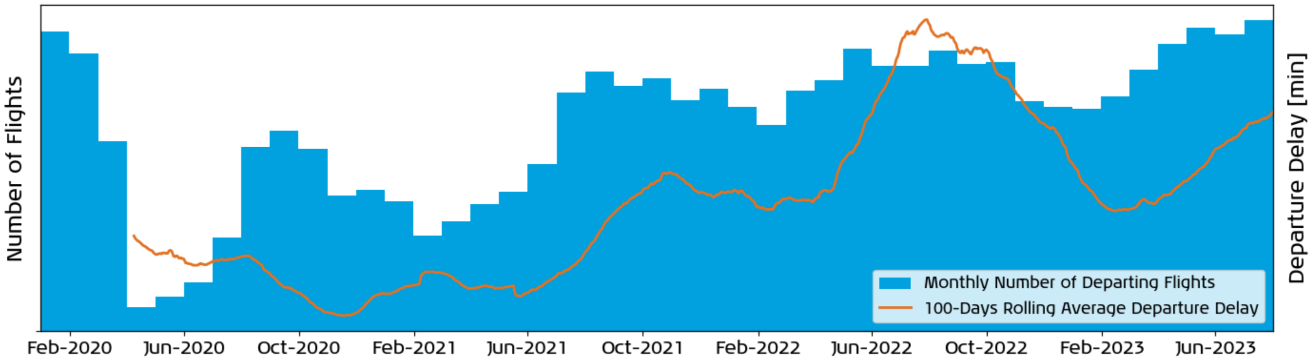


Figure 1. Flight Counts and Departure Delays for KLM Departures from Schiphol (January 2020 to August 2023)

The importance of accurate departure delay predictions is presented in Appendix 1. The case study involves the use of supervised learning algorithms to improve on an existing statistical model, currently in use at KLM. For predicting departure delays, the statistical model solely uses historical data of other flights within the same flight series. the supervised learning algorithms, coded using Python, are trained to draw patterns based on all historical flights. Next to further minimizing the prediction error, the case study also aims to better explain the predicted delays by predicting departure delays as probability distributions over time and by explaining what inputs contribute to the predicted output. The study is split up into multiple stages, illustrated in Figure 2. These stages are elaborated upon in more detail in Section 3.2 to Section 3.5.

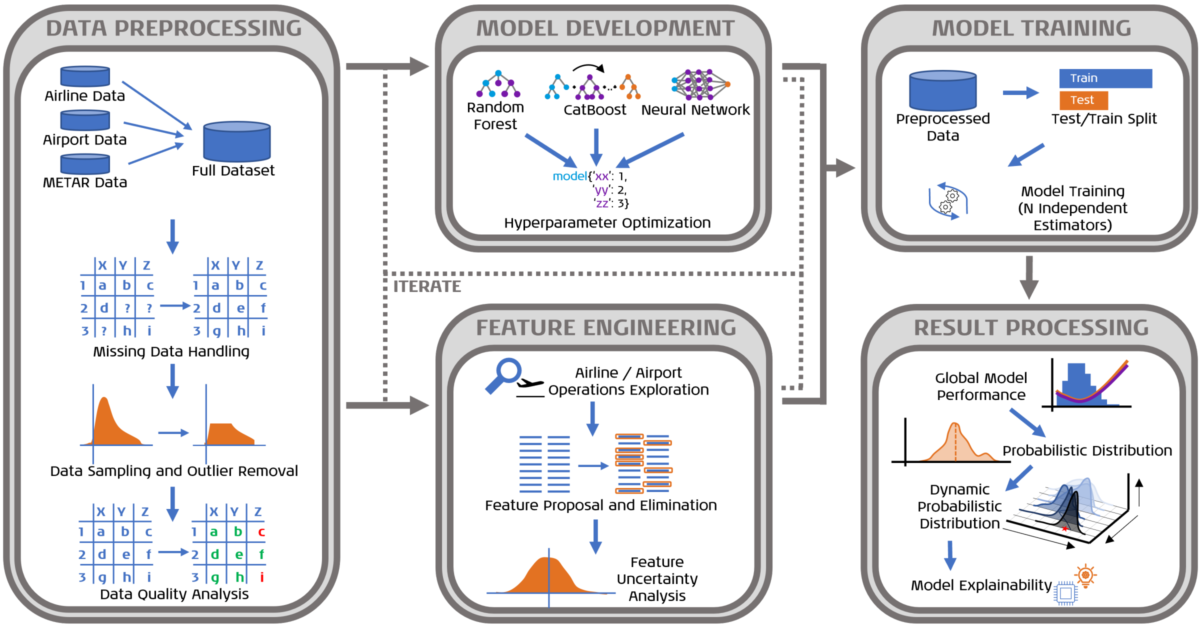


Figure 2. Methodology for Dynamically Forecasting Airline Departure Delay Probability Distributions

3.2 Data Preprocessing

Through the collaboration with KLM, detailed airline data and basic Schiphol traffic data is available. Historical weather data is available from the Iowa State University Environmental Mesonet [27]. Following these three datasets, a full list of proposed features is presented in Appendix 2.

- **KLM Data:** After eliminating cancelled flights, flights with unpredictable delays[2], ferry flights, test flights and flights without recorded departure delays, the KLM dataset consists of 286,000 unique flights. This dataset also includes (anonymized) passenger information and flight events such as TSAT updates and Estimated Time of Arrival (ETA) predictions of previous flights. Additionally, connection times and passenger counts from any flight arriving at Schiphol are available for over 90% of the outbound flights.
- **Schiphol Data:** After similar initial data cleaning as the KLM data, the raw Schiphol dataset consists of 562,000 flights. From the available information on origin-destination pairs, scheduled and actual departure times, and flight numbers, the number of flights and average delays at Schiphol can be determined for any time interval within the case study period.
- **METeorological Aerodrome Report (METAR) Data:** Available from the Iowa State University Environmental Mesonet, weather reports with an update frequency of 30 minutes (*25 and 55 minutes past the hour*) are obtained for Schiphol. These reports include air temperature, dew point, humidity, wind direction, wind speed and gust, precipitation, visibility, cloud coverage, and cloud height information [27].

The remainder of this section explains the process of transforming the above data sources into a usable dataset. For that, the missing data handling practices are presented in Section 3.2.1 and the sampling and outlier removal processes are elaborated upon in Section 3.2.2.

3.2.1 Missing Data Handling

To guarantee the quality of the input data for the models, several other data handling techniques are applied on top of the previously mentioned filter. The range of feature values is inspected for *incorrect* values, requiring additional preprocessing for two features. Flights with aircraft tail swaps after actual departure are corrected for by replacing these clearly wrong values with the median feature value. Similarly, for three flights, wind speeds of over 100 knots[3] were recorded and are thus also replaced by the median feature value.

To adjust for *incomplete* data, missing data for the number of passengers with reduced mobility is set to zero because empty cells in the database represent flights without such passengers. Missing data for all other features

[2] Delays are registered using IATA Delay Codes [28], a classified selection of which is deemed unpredictable, e.g. technical issues.
[3] Well above wind speeds for hurricanes.

is replaced with the median feature value, to avoid introducing biases into the model. Given the importance of the effective delay of previous flights feature, which is discussed in Section 4.3.1, flights with missing entries for this feature are removed completely, as this was found to improve the performance of the model.

3.2.2 Data Sampling and Outlier Removal

The blue bins in the left figure of Figure 3 confirm that the aforementioned positive skew is also present in the departure delay data for this study. To mitigate the problem, logarithmic and square-root transformations were tested but did not yield the desired results. Therefore, the data has to be balanced through undersampling. Selecting the correct bin to undersample the excess data introduces a trade-off between two metrics: R2 and RMSE, illustrated in the right figure of Figure 3. No data sampling yields relatively low RMSE values because the model is overfitted to the part of the departure delay spectrum with most data points. As a result, the errors for flights with uncommon departure delays are relatively high, causing the fit of the model to remain low.

To guarantee a balance of global fit and optimal model performance, it is chosen to select a sampling strategy whereby the number of flights is limited to the number of flights in the 15-minute bin. This yields the purple distribution in the left figure of Figure 3. For this sampling strategy, the R2 value is above its trend and the RMSE value is below its trend, where additional importance is given to the RMSE value as it increases relatively faster than the R2 value. The sampling process is performed at random, because other methods such as cherry-picking may introduce biases to the model.

To improve the accuracy and reliability of the model, outliers are removed from the dataset. Because of the positive skew in the departure delay distribution, visible in the left figure of Figure 3, only values towards the extreme positive end are removed, since these values lie much farther from the median value than those towards the negative end. When comparing three different outlier removal percentages (1%, 2%, and 3%), the model fit in terms of Coefficient of Determination (R2) remains constant, with only the MAE and RMSE error metrics changing. Logically, removing fewer outliers leads to higher error values. Because the model fit remains unaffected, only 1% of outlier data is removed, thereby preventing the potential loss of valuable data patterns. After outlier removal, the remaining departure delay spectrum ranges from -20 to +97 minutes.

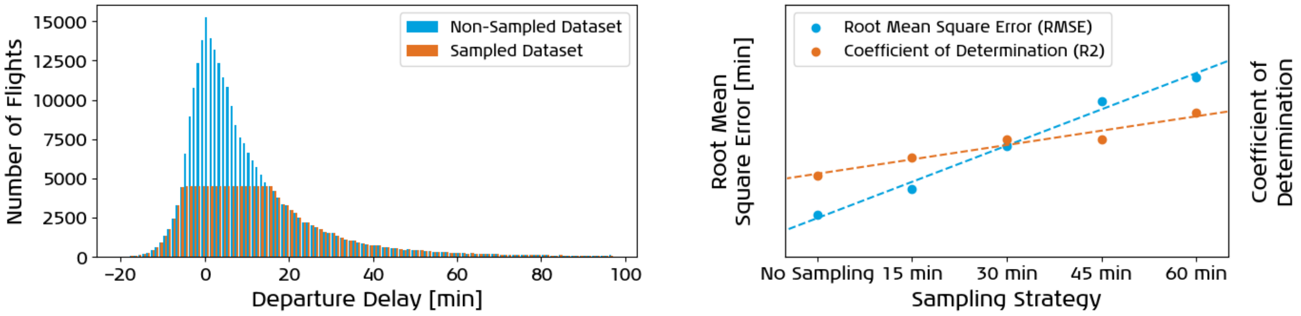


Figure 3. Non-Sampled and Sampled Departure Delay Distributions (left) and Sampling Strategy Trade-Off (right)

3.3 Model Development

When comparing the results of different studies, one has to take into account the differences in underlying datasets. Fortunately, multiple studies compared the performance of different models while using the same dataset. From these studies, it became apparent that simple decision trees would nearly always be outperformed by random forests and boosting models [5, 6, 7]. Additionally, more complex neural network structures would outperform simpler structures as long as the available training dataset is of sufficient size [8, 12, 13]. When insufficient training data is available or the predicted processes are too random, simpler neural networks or tree-based models may outperform more complex models.

Since it is unknown whether the data set is of sufficient size beforehand, it is chosen to propose both tree-based and neural network models in this study. Although in early design stages, Gradient Boosted Decision Tree (GBDT), Graph Neural Network (GNN), and Mixture Density Network (MDN) models were tested, initial results were in favour of Random Forest, CatBoost and Deep Neural Network models. The latter three models are further developed and are elaborated upon in Section 3.3.1, Section 3.3.2 and Section 3.3.3, respectively. The pseudo-code for training these models is presented in Appendix 3. Whilst the focus is on the 90-minute prediction horizon, the delays are also predicted at five other prediction moments: 75, 60, 45, 30, and 15 minutes before scheduled departure time. To prevent the models from falsely propagating errors and uncertainty from one prediction moment to another, separate models are trained for each prediction moment. For all models, hyperparameter optimization is performed empirically, as the effect of all parameters was understood through numerous model iterations in the test phase. Bayesian search hyperparameter optimization schemes were tested but not applied because they resulted in very marginal improvements only, especially considering the extremely long run times.

3.3.1 Random Forest

Given its computational efficiency, robustness to outliers, and interpretability, random forests have been a popular choice for modelling stochastic processes [5, 12, 13, 20]. The random forest is an ensemble method, a forest of decision trees that serve as independent predictors. Furthermore, the method relies on the concept of bagging (bootstrap-aggregating), meaning that sub-samples of the dataset are used to construct unique decision trees and that, for regression problems, the final prediction is the mean of all individual results. For a probabilistic approach, a probability distribution can be created using all individual tree results [10, 11]. This is illustrated in Figure 4. Further details about the random forest model are provided in the original paper by Breiman [29].

To allow for generating detailed probability distributions, the number of decision trees (*nr_estimators*) is set to 1000. The maximum number of features for the model (*max_features*) is set to 4, following from the binary logarithm of the number of features[4]. The maximum model depth (*max_depth*) is set to 10 to prevent overfitting as a result of extremely large trees. The minimum number of samples for splits (*min_samples_split*) and leaves (*min_samples_leaf*) is set to 4 and 2 respectively, because larger values may yield too simple decision trees. The hyperparameters are summarized in Table 1.

Table 1. Random Forest Hyperparameters

Hyperparameter	Assigned Value
nr_estimators	1000
max_features	4
max_depth	10
min_samples_split	4
min_samples_leaf	2

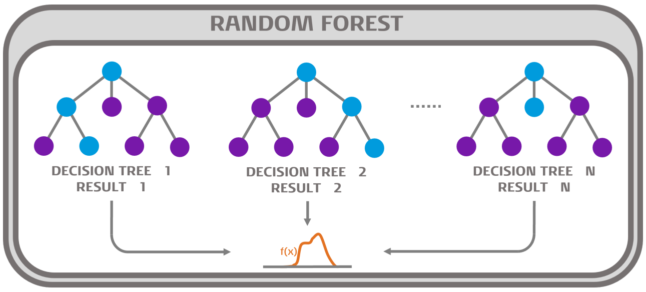


Figure 4. Probabilistic Approach for Random Forest Method

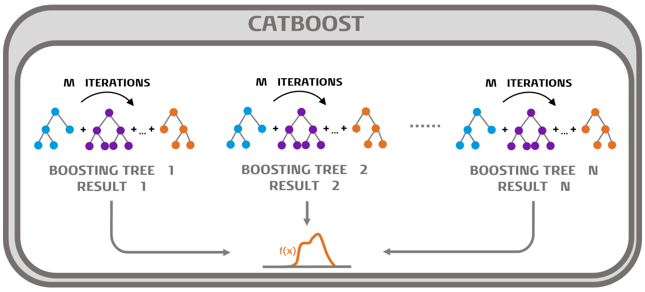


Figure 5. Probabilistic Approach for CatBoost Method

3.3.2 CatBoost

CatBoost is an open-source gradient boosting library that allows for efficient and fast predictions. The model treats data sequentially to prevent data leakage. The use of symmetric trees makes weaker learners for the boosting process, resulting in faster computation times. Moreover, the underlying boosting scheme of CatBoost prevents overfitting and eases hyperparameter tuning [30]. CatBoost models require M iterations to reach the final prediction. To allow for probabilistic modelling, a total of N independent CatBoost models are created, see Figure 5, where N is set to 1000 to match the number of estimators of the Random Forest model. Further details about the CatBoost model are provided in the original paper by Prokhorenkova [31].

Two different CatBoost models are proposed, one with a smooth iteration scheme (*iterations* = 400 and *learning_rate* = 0.02), in the remainder referred to as *CatBoostTH* for its theoretical application, and one with a rougher iteration scheme (*iterations* = 10 and *learning_rate* = 0.5), in the remainder referred to as *CatBoostPR* for its probabilistic application. This distinction is made for probabilistic prediction purposes, as the smooth iteration scheme yields extremely confident models. For both models,

Table 2. CatBoost Hyperparameters

Hyperparameter	Assigned Value (CatBoostTH)	Assigned Value (CatBoostPR)
iterations	400	10
learning_rate	0.02	0.5
depth	10	10
posterior_sampling	*True*	*True*
random_state	*Random*	*Random*

[4] Using all features may lead to overfitting the model.

the model depth (*depth*) is set to 10, complying with that of the Random Forest model. Finally, posterior sampling (*posterior_sampling*) is enabled to *"obtain uncertainty predictions with good theoretical properties"* [31]. To ensure that all N independent models are unique, the unique state (*random_state*) is set randomly for every estimator of the model. The hyperparameters are summarized in Table 2.

3.3.3 Deep Neural Network

Inspired by the human brain, neural networks are machine learning models consisting of interconnected nodes that are activated by activation functions. Neural networks are able to handle nonlinear feature relationships and are trained using error back-propagation, a feedback loop that tunes the internal model parameters to achieve the optimal performance [32]. For adequate training, a substantial amount (10^5 to 10^6) of historical flights is required to achieve meaningful results [33]. Compared to the other models, neural networks present more challenges in terms of model explainability.

For the neural network structure, the number of input neurons (*nr_input_neurons*) is the number of input features. Since the departure delay is the only output, the number of output neurons (*nr_output_neurons*) is equal to 1. There is more flexibility in determining the number of hidden neurons (*nr_hidden_neurons*), which is set to 16 for the most optimal result. Controlling the depth of the model, the number of hidden layers (*nr_hidden_layers*), is set to 4, to prevent an overly complex model from forming. Since preliminary results yielded comparable results for in-sample and out-of-sample data, the model was not overfit. Therefore, dropout (*dropout*) is not required and is set to 0. Finally, for the Leaky ReLU (Rectified Linear Unit) activation functions in each of the layers of the model, the negative slope (*leakyrelu_negative_slope*) is set to 0.1, to allow for negative inputs. The hyperparameters are summarized in Table 3.

Table 3. Deep Neural Network Hyperparameters

Hyperparameter	**Assigned Value**
nr_input_neurons	15
nr_output_neurons	1
nr_hidden_neurons	16
nr_hidden_layers	4
dropout	0
leakyrelu_negative_slope	0.1
n_epochs	1000
batch_size	2048
lr_initial	1e-4
lr_increase	1.2
lr_decrease	1.2
lr_stop	1e-10

For deep neural network training purposes, several training parameters are set. The number of epochs (*n_epochs*) is set to 1000, to facilitate enough learning iterations. For the regression algorithm, the MSE loss function is used. Batch sizes (*batch_size*) of 2048 datapoints are used, resulting in sufficient batches considering that the full dataset is two orders of magnitude larger. For the training process, an adaptive learning rate is used to ease searching for global optima. The initial learning rate (*lr_initial*) is set to 1e-4. After every epoch, the learning rate is increased by an increase factor (*lr_increase*) of 1.2 if the current epoch prediction is better than the current best prediction and decreased by a decrease factor (*lr_decrease*) of 1.2 otherwise. If the learning rate becomes smaller than the stop criteria learning rate (*lr_stop*) of 1e-10, the model is deemed to have converged.

3.4 Feature Engineering

The feature elimination process is explained in Section 3.4.1. The selected features are then presented in Section 3.4.2 and their correlation is discussed in Section 3.4.3. Finally, potential uncertainty related to some of the features is evaluated in Section 3.4.4.

3.4.1 Feature Elimination

In a Recursive Feature Elimination (RFE) process, the least valuable feature of a model is eliminated, one at a time. This approach was taken for all proposed models, considering all features listed in Appendix 2. Due to the randomness in the models, the elimination process was repeated several times, averaging the rankings. As such, a preliminary list of features was obtained. A number of features were manually eliminated to avoid overfitting as a result of strong correlations between comparable features, known as multicollinearity. For example, the transfer passenger percentage and the local passenger percentage are related as the one is 100% minus the other.

In the end, 15 features are selected for the Random Forest and CatBoost models, listed in Table 4. Some of these features are dynamic, meaning that their values are updated at every prediction moment. For the Deep Neural Network, trigonometric variations[5] to the month of the year and hour of the day are used, resulting in a total

[5] Creating sine and cosine components.

of 17 features. Doing so allows the model to learn that December and January are subsequent months and 23:00 and 0:00 are subsequent hours. Tree-based models do not require such trigonometric functions because of the splitting nature of these models that differentiates certain moments in time already.

3.4.2 Selected Features

The selected features from Table 4 are described in more detail below. Note that for the Deep Neural Network, there are trigonometric variations to the month of the year and hour of the day features.

Table 4. Selected Features after Feature Elimination

Feature Name	**Unit**	**Numeric**	**Dynamic**	**Example**
Month of Year	[-]	✓		*4*
Hour of Day	[-]	✓		*13*
Passenger Load Factor	[-]	✓		*0.73*
Baggage Load Factor	[-]	✓		*1.14*
Transfer Passenger Percentage	[-]	✓		*67*
Number of Passengers Reduced Mobility	[-]	✓		*2*
Total Passengers Day Flying Blue Schiphol	[-]	✓		*35000*
Median Delay of Flight Number	[min]	✓		*7*
Effective Delay Previous Flight	[min]	✓		*12*
Current Number of Flights Schiphol [a]	[-]	✓	✓	*20*
Current Average Delays Schiphol [a]	[min]	✓	✓	*32*
Current TSAT Delay [b]	[min]	✓	✓	*4*
Last Aircraft Tail Swap [b]	[min]	✓	✓	*1500*
Wind Speed Longitudinal Direction	[kts]	✓		*8.32*
Wind Speed Latitudinal Direction	[kts]	✓		*-4.25*

[a] In the 30-minute interval before prediction moment.
[b] At prediction moment.

- **Month of Year:** The numeric[6] month of the flight, following from the Scheduled Off-Blocks Time (SOBT)[7].
- **Hour of Day:** The departure hour of the flight, following from the SOBT[7].
- **Passenger Load Factor:** The number of booked passengers relative to the number of seats available on the aircraft. For a fully booked flight, the passenger load factor is 1. Taking a ratio rather than an absolute number of passengers allows to consider flights operated on aircraft with different seating capacities in the same model.
- **Baggage Load Factor:** The number of booked pieces of baggage relative to the number of booked passengers. Similarly to the passenger load factor, taking a ratio rather than an absolute number of baggage pieces allows to consider flights operated on aircraft with different seating capacities in the same model.
- **Transfer Passenger Percentage:** The percentage of booked passengers for an outbound flight that connect from any inbound flight at Schiphol.
- **Number of Passengers Reduced Mobility:** The number of booked passengers with wheelchair assistance.
- **Total Passengers Day Flying Blue Schiphol:** The total number of daily passengers departing from Schiphol on flights operated by Flying Blue member airlines. Unfortunately, for airlines outside the Flying Blue alliance, the passenger numbers were unavailable to this study.
- **Median Delay of Flight Number:** The median departure delay of all flights for a given flight series. For flight series with fewer than 25 recordings, the median delay calculation is considered to be too random because of the small number of data points. For these uncommon flights, a zero median delay is registered.
- **Effective Delay Previous Flight:** The effective arrival delay of the previous flight that propagates to the outbound flight for the same aircraft. Airlines incorporate slack times into their schedules to mitigate potential delays [13, 34]. For inbound flight i and outbound flight j, operated by airline k using aircraft type l at airport m, the Scheduled Turn-Around Time ($STAT_{i,j}$) is the time difference between the Scheduled Off-Blocks Time of the outbound flight ($SOBT_j$) and the Scheduled In-Blocks Time of the inbound flight ($SIBT_i$), see Equation 1. The slack time ($\rho_{i,j,k,l,m}$) is the time difference between the Minimum Turn-Around Time[8] ($MTAT_{k,l,m}$) and the Scheduled Turn-Around Time, see Equation 2. MTAT values may differ per airline, aircraft type, and airport. The arrival delay of the inbound flight (δ_{arr_i}) is the time difference between the Actual In-Blocks Time ($AIBT_i$) and Scheduled In-Blocks Time of the inbound flight, see Equation 3. Finally, the effective arrival delay of the inbound flight is the arrival delay minus the slack time, see Equation 4. Since negative slack times do not exist, the effective arrival delay can never exceed the arrival delay itself. All negative effective arrival delays are set to 0 as there is enough available time to turn around the aircraft, regardless of how early the previous flight arrived. Figure 6 illustrates the effect of slack times on effective arrival delays.

$$STAT_{i,j} = SOBT_j - SIBT_i \tag{1}$$

$$\delta_{arr_i} = AIBT_i - SIBT_i \tag{3}$$

$$\rho_{i,j,k,l,m} = STAT_{i,j} - MTAT_{k,l,m} \tag{2}$$

$$\delta_{arr,eff_{i,j,k,l,m}} = \delta_{arr_i} - \rho_{i,j,k,l,m} \tag{4}$$

[6] e.g. January → 1, February → 2, etc.
[7] In Coordinated Universal Time (UTC).
[8] Airline-issued times that indicate the minimum number of minutes required between arrival and departure of two consecutive flights.

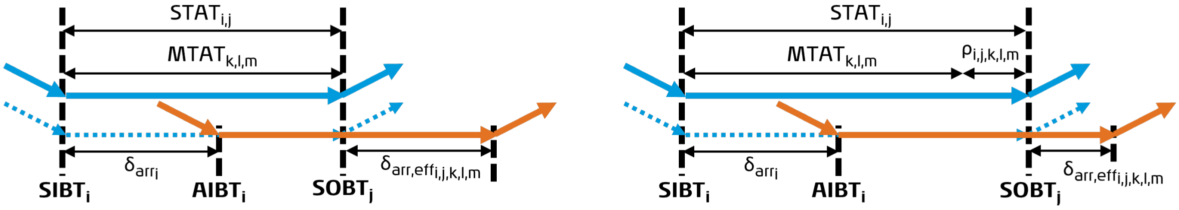


Figure 6. Aircraft Turn-Around without Slack Times (left) and with Slack Times (right)

- **Current Number of Flights Schiphol:** The total number of flights departing from Schiphol in a 30-minute time interval before the prediction moment. This feature is dynamic because for each prediction moment, the 30-minute time interval is different, possibly resulting in a different total number of flights.
- **Current Average Delays Schiphol:** The average delay of all flights departing from Schiphol in a 30-minute time interval before the prediction moment. This feature is dynamic because for each prediction moment, the 30-minute time interval is different, possibly resulting in a different average delay. Mathematically, for every prediction moment t, the average departure delay (δ_{dep,avg_t}) is the sum of individual departure delays (δ_{dep_j}), divided by the total number of flights in the 30-minute time interval (N_t), see Equation 5.

$$\delta_{dep,avg_t} = \frac{\sum_{j=1}^{N_t} \delta_{dep_j}}{N_t} \tag{5}$$

- **Current TSAT Delay:** The latest available Target Start-up Approval Time delay update. This delay, imposed by Air Traffic Control[9], is dynamic because it may update between prediction moments.
- **Last Aircraft Tail Swap** The time difference between the last aircraft tail swap (new airframe allocation) and scheduled departure time. This feature is dynamic because tail swaps may occur between prediction moments.
- **Wind Speed Longitudinal / Latitudinal Direction:** The longitudinal (East-West) and latitudinal (North-South) components of the wind speed at the departure airport. Including the wind direction and wind speed by themselves may train the model to believe some wind direction would favour delays, even if the wind speed is almost zero. This can be avoided by combining wind speed (V_w) and wind direction (Γ_w) into longitudinal and latitudinal wind components, as presented in Equation 6 and Equation 7.

$$V_{w_{lon}} = -V_w \cdot \cos\left(\Gamma_w - 90^\circ\right) \tag{6}$$

$$V_{w_{lat}} = V_w \cdot \sin\left(\Gamma_w - 90^\circ\right) \tag{7}$$

3.4.3 Feature Correlation

The statistical correlation between all features at the 90-minute prediction horizon is presented in Figure 7. The feature describing the daily total number of Flying Blue passengers at Schiphol shows relatively strong correlations with three other features: the passenger load factor (0.59), the current number of flights at Schiphol (0.49), and the current average delays at Schiphol (0.44). The latter two correlations follow from the coupling of traffic levels and associated delays, previously identified in Figure 1. The former correlation follows from the fact that during busy periods, the number of daily passengers increases faster than the number of flights, thus resulting in higher passenger load factors. This also explains the negative correlation with the baggage load factor (-0.40), as it appears that for busy periods, the number of passengers increases faster than the pieces of baggage that are carried along.

Furthermore, the median delay of flight number shows a relatively strong correlation with the number of passengers with reduced mobility. Data reveals that the 500 flights with the highest number of passengers with reduced mobility, were operated under 11 unique flight numbers only. The correlation is evident because the number of passengers with reduced mobility heavily impacts the turnaround process and therefore potential departure delays.

The feature correlations between the non-dynamic features remain the same at the 15-minute prediction moment. The correlations between the dynamic features (e.g. current average delays Schiphol and current TSAT delay) become significantly larger. This can be explained by the fact that the updated dynamic features are closer to the actual values. Finally, for the trigonometric features used for the Deep Neural Network, the cosine of the hour

[9] In the case of Schiphol, air traffic is controlled by Luchtverkeersleiding Nederland (LVNL).

of day has a relatively strong negative correlation (-0.49) with the current number of flights at Schiphol. This makes sense as the cosine value of the hour of day is high for the early morning and late evening, but gradually decreases for the middle of the day. The number of scheduled flights develops in the exact opposite manner.

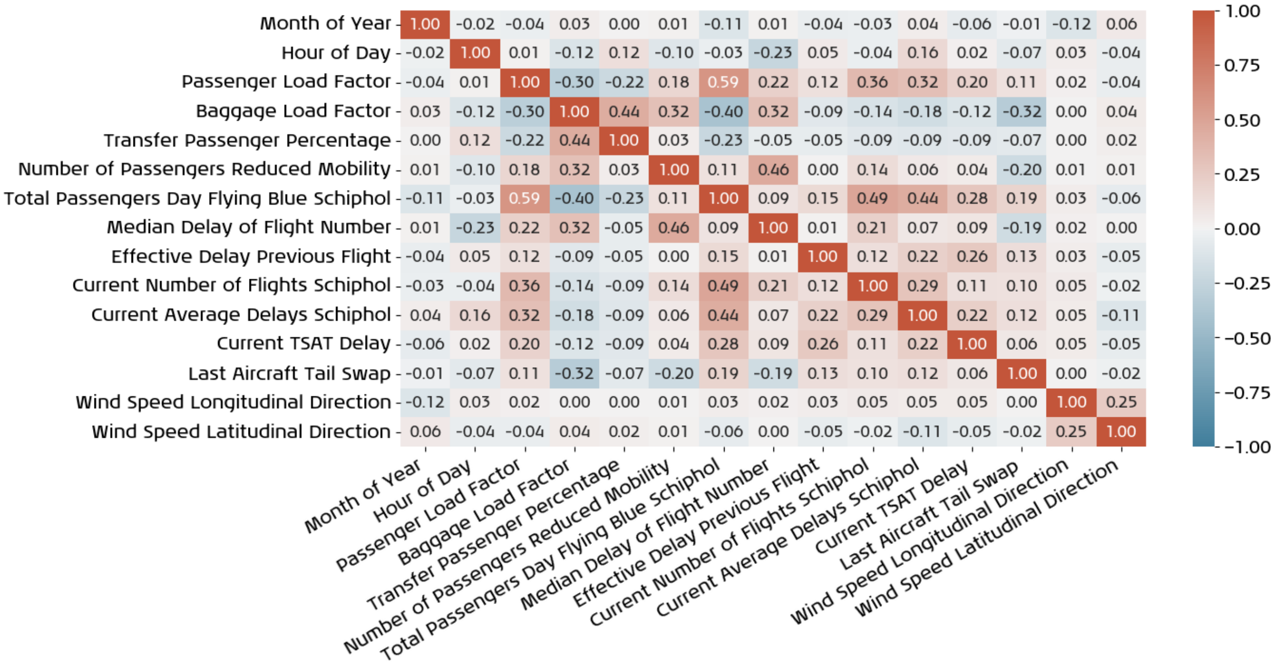


Figure 7. Feature Correlation Matrix for the 90-Minute Prediction Horizon

3.4.4 Feature Uncertainty

For flights yet to be predicted, not all input features may be exactly known at each of the different prediction moments. Since the historical training dataset consists of *actual* values, the use of predicted input values may introduce noise and biases. To guarantee the accuracy of the model, such potential uncertainty needs to be evaluated. Most of the features mentioned in Table 4 are either constant throughout the prediction horizon of a flight (e.g. month of year, hour of day, median delay of flight number), can be exactly computed for all prediction moments (e.g. current number of flights, delays and TSAT delays) or undergo only very minor changes over the prediction horizon of a flight (e.g. passenger and baggage numbers). For a check on a subset of the flight data, large differences in actual and booking passenger and baggage numbers were observed only very rarely, leading to the assumption of using constant passenger and baggage numbers during the prediction horizon. For the effective delay of previous flights and the wind speed features, however, the uncertainty is higher.

Effective Delay Previous Flight Uncertainty: KLM flights operated on single-aisle aircraft have minimum turn-around times of 35-50 minutes and thus may still be airborne 90 minutes prior departure of the outbound flight. Flights operated on twin-aisle aircraft have minimum turn-around times of over 120 minutes. However, these flights may experience delays on the inbound flight, impeding the aircraft from being at the gate 90 minutes prior to departure of the outbound flight. Throughout the case study period, 42% of KLM flights arrived more than 90 minutes before departure of the outbound flight; for these flights, there is no feature uncertainty. For the remaining flights, the effective delay of the previous flight can be computed from the ETA of the previous flight.

For the flights that were not yet at Schiphol 90 minutes prior to the outbound flight, the error distributions between actual and predicted arrival times are presented in Figure 8, differentiating the total set of flights and that of the three shortest routes KLM operates. Inbound flights on these three routes are most likely to still be on-ground at the origin airport 90 minutes before departure of the outbound flight from Schiphol, affecting the uncertainty of ETA predictions more than whilst airborne. As a result of using the predicted ETA, noise is added to the model. The Inter-Quartile Range (IQR) in Figure 8 is at most 6 minutes. Although the ETA prediction model seems slightly conservative, predicting a large number of flights to arrive later than they did in reality[10], there are no clear biases.

[10] Most likely as a result of en-route shortcuts.

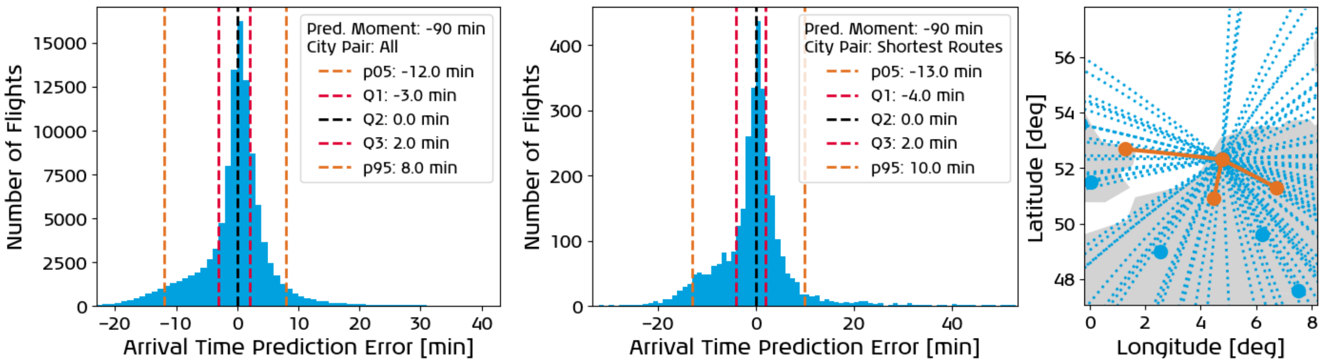


Figure 8. Previous Flight Arrival Time Prediction Error Distributions for All Flights (left) and the 3 Shortest Flights (centre). The three Shortest Routes are visualized on the right.

Wind Speed Longitudinal / Lateral Direction Uncertainty: The wind speed features contain uncertainty because only METAR data is available to train the model. For flights yet to be predicted, the METAR at scheduled departure time is still unknown at the prediction moment. Therefore, Terminal Aerodrome Forecasts (TAF) are used for predicting new flights instead. The unavailability of open-source TAF reports restricts training the model on TAF data. This problem was previously addressed by Schoesser and Schoenberger [23], who wrote that their "*intention is to use forecasted weather data*" but that "*historical true weather data is used due to a lack of weather forecast data*".

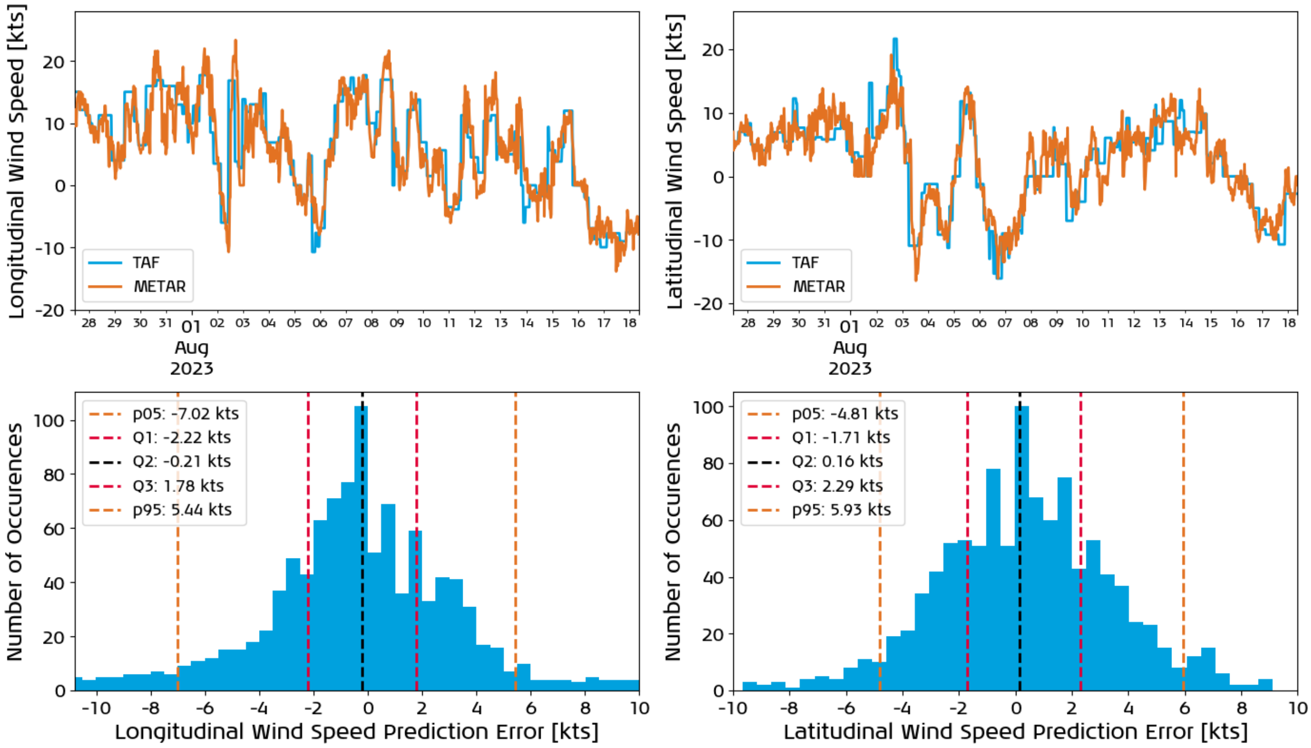


Figure 9. Longitudinal and Latitudinal Wind Speed Developments (upper left, upper right) and Prediction Error Distributions (lower left, lower right)

To validate the use of METAR data for Schiphol, for 22 days (1056 recordings) with various wind conditions[11], the predicted wind (TAF) is compared to the actual wind (METAR). The developments over time for the longitudinal and lateral wind components are presented in the upper figures of Figure 9. In general, the TAF is capable of adequately predicting long-term weather developments. Some of the prediction errors can be attributed to the TAF reports describing how the weather is expected to change over a longer period of time. This is represented

[11] Wind speeds ranging from 0 to 27 kts, from every direction (rounded to 10 degrees) at least once.

by the horizontal sections on the blue lines in the graphs. METARs have higher update frequencies and vary more heavily, leading to other small prediction errors. The bottom figures in Figure 9 present the wind speed prediction error distributions for longitudinal and lateral directions. It can be concluded that the IQR for both directions never exceeds 4 knots. Similar to the feature describing the effective delay of the previous flight, the use of predicted weather data introduces noise to the model, but no bias as both error distributions are symmetric and the medians are near-zero.

3.5 Model Training and Result Processing

To train the models, the full dataset is randomly split into a training dataset (80% of data) and a test dataset (20% of data). The departure delays in the training and test datasets are similarly distributed as in the full dataset. To minimize data leakage, the split was made per day instead of per flight. This ensures that when testing the model, it has no prior knowledge about the dynamics on the day of the flight.

To facilitate probabilistic departure delay forecasting, the models use all independent predictions to create probability distributions. This is preferred over using majority voting or computing the mean of all independent predictions since probability distributions indicate the likelihood of a delay value being predicted. To ensure high granularity, 1000 unique independent estimators are considered. To create dynamic probabilistic departure delay predictions, the predictions for all prediction moments are combined. The dynamic probabilistic departure delay predictions not only show how the predicted delay value changes over time but also the evolution of the associated probability density and certainty.

The costs and emissions associated with the decisions made using the departure delay prediction model make explainability an important aspect of this study. For that reason, the probabilistic model performance is one of the considerations for the model selection. The Shapley Additive Explanations (SHAP) library was tested but yielded run times of several days because of the size of the dataset. Therefore, a method that explains the predictions based on relative feature scaling is introduced, illustrated in Figure 10.

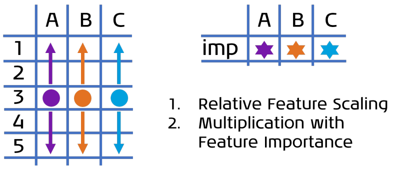

A
B
C
1
2
3
4
5
A
B
C
imp
1. Relative Feature Scaling
2. Multiplication with Feature Importance

Figure 10. Explainability Through Feature Scaling Method

First, the feature values of all flights in the test dataset are scaled by fitting a StandardScaler()[12] on the dataset. The size of the dataset should be at least order of magnitude 1000 to obtain meaningful scaled feature values[13]. The scaling is performed for each feature individually and returns the number of standard deviations a feature value differs from the mean of all flights. Large scaled feature values indicate that a flight stands out from others in the respective feature. Simply scaling the feature values does not explain the model prediction; it only considers the model inputs, not what the model is doing with this input data. For that reason, the scaled values are weighted by the feature importances, thereby including the importance assessment of the model. Although this method does not explain the exact decision-making of the model, it indicates how the model has treated the underlying data and how this affects the prediction for a certain flight. Finally, for usability purposes, thresholds are determined for classifying the scaled weighted feature values towards large, moderate, and small impacts[14]. Following these impacts, an explainability message is constructed and provided to the flight dispatchers.

4. Results

This section aims to present and analyse the results from the four proposed models. The global model performance is presented in Section 4.1, where the most suitable model is selected. The dynamic prediction behaviour of this model is then presented in Section 4.2. The model explainability results are discussed in Section 4.3.

4.1 Global Model Performance and Selection

The performance metrics used for the model selection are outlined in Section 4.1.1. The global and probabilistic model performances are presented in Section 4.1.2 and Section 4.1.3, respectively. Finally, the model performance per departure delay bin is discussed in Section 4.1.4.

[12] From the scikit-learn module.

[13] To have enough data points to generate feature distributions of sufficient granularity.

[14] $X \geq 0.5$ → large impact; $0.3 \geq X > 0.5$ → moderate impact; $0.1 \geq X > 0.3$ → small impact, for scaled weighted feature value X.

4.1.1 Performance Metrics

Five performance metrics are proposed for evaluating the four proposed models, covering the model error, model fit, and probabilistic performance. Each metric is briefly elaborated upon in Table 5. Whereas the first three metrics are commonly used, the final two metrics are introduced in this study, especially to quantify the probabilistic performance of the models. The *ActInDistr* metric represents the percentage of flights for which the actual departure delay is within the predicted departure probability distribution. The *AvgIQR* metric represents the average inter-quartile range of predicted flights, a measure of the model confidence. Ideally, models score high on *ActInDistr* and low on *AvgIQR*, resulting in confident and correct predictions. Overconfident models predict narrow probability distributions and thus score low for both metrics, while underconfident models predict wide probability distributions and thus score high for both metrics.

Table 5. Performance Metrics

Performance Metric	Explanation
Mean Absolute Error (*MAE*)	Absolute error between predicted and actual values
Root Mean Square Error (*RMSE*)	Standard deviation of errors between predicted and actual values
Coefficient of Determination (*R2*)	Proportion of variation in dependent variable predictable from independent variable
Actual in Distribution Percentage (*ActInDistr*)	Percentage of flights for which actual value is in predicted probability distribution
Average Inter-Quartile Range (*AvgIQR*)	Average inter-quantile range for predicted flights

4.1.2 Global Model Performance

Table 6. Global Model Performance for Out-of-Sample Data

		Prediction Moment [min]					
		-90	-75	-60	-45	-30	-15
Random Forest	MAE [min]	8.46	8.31	8.11	7.90	7.67	7.37
	RMSE [min]	11.91	11.69	11.42	11.13	10.81	10.44
	R2 [-]	0.55	0.57	0.59	0.61	0.63	0.65
CatBoostTH	MAE [min]	8.20	8.06	7.88	7.69	7.47	7.15
	RMSE [min]	11.67	11.46	11.20	10.93	10.62	10.23
	R2 [-]	0.57	0.58	0.60	0.62	0.64	0.67
CatBoostPR	MAE [min]	8.26	8.11	7.93	7.74	7.52	7.20
	RMSE [min]	11.72	11.51	11.25	10.98	10.67	10.28
	R2 [-]	0.56	0.58	0.60	0.62	0.64	0.67
Deep Neural Network	MAE [min]	8.35	8.26	8.07	7.91	7.82	7.31
	RMSE [min]	11.81	11.64	11.39	11.16	10.92	10.42
	R2 [-]	0.56	0.57	0.59	0.61	0.62	0.66

For each model, distinct sub-models area developed for every prediction moment. For each respective model, the MAE, RMSE, and R2 are tabulated in Table 6 and graphically presented in Figure 11. Given that all models have comparable performance for in-sample and out-of-sample datasets, it can be concluded that no models are overfit.

For all models, the errors decrease for shorter prediction horizons, whilst the R2-values increase. This follows logically from the perceived updates on the dynamic features. For these features, the change in correlation with the departure delay is illustrated in Figure 12. As expected, the current average delays and current TSAT delays yield much higher correlations over time. Figure 11 shows that between 30 and 15 minutes before scheduled departure, the model improves most significantly. This can be explained by the fact that a large share of delays occur just before departure. Including more short-term features would further strengthen this effect, but adding such parameters is invaluable at larger prediction horizons, which is the main focus of this study.

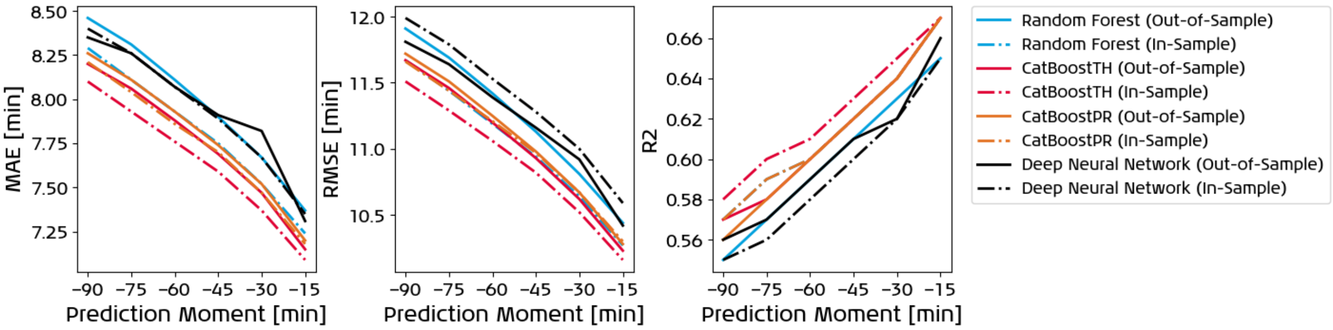


Figure 11. Global Model Performance over Time in terms of MAE, RMSE, and R2

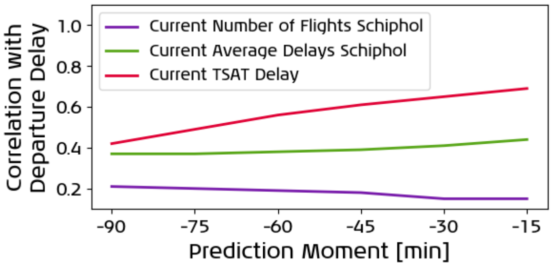


Figure 12. Dynamic Feature Correlation over Time

Several other conclusions can be drawn from Figure 11. First, in terms of MAE and RMSE errors, the CatBoostTH model slightly outperforms the CatBoostPR model. The absolute differences in MAE, RMSE, and R2 between the four proposed models are small and decrease over time. When expressed as percentages relative to the best model, the differences remain constant over time (3.1% for the MAE and 2.1% for the RMSE). The CatBoostTH model has a quasi-constant advantage over the other models because only for the Deep Neural Network at the 30-minute prediction horizon, the relative differences are larger. This kink may be explained by the model reaching the maximum number of epochs, whereas for other prediction horizons, it converged earlier.

Following Figure 11, the Deep Neural Network does not outperform the other models. Next to that, the model imposes difficulties for providing probabilistic and explainable results. If the model had yielded smaller errors, these challenges would have been investigated further. For the sake of project planning, it was decided to eliminate the model from further selection[15].

The prediction accuracy of the other models is illustrated in Figure 13, where the actual and predicted delays for a 90-minute prediction horizon are plotted as a heatmap. In the ideal situation, the data points follow the diagonal dotted line, where the predicted delay equals the actual delay. For all models, the distributions follow this line to some extent, albeit with notable noise. All models tend to overpredict for flights with small delays and underpredict for flights with larger delays. This behaviour is partially caused by the splitting nature of the structure of the models as well as the absence of possibly valuable information and the uncertainty associated with large prediction horizons. The overpredicting behaviour is further discussed in Section 6.3.

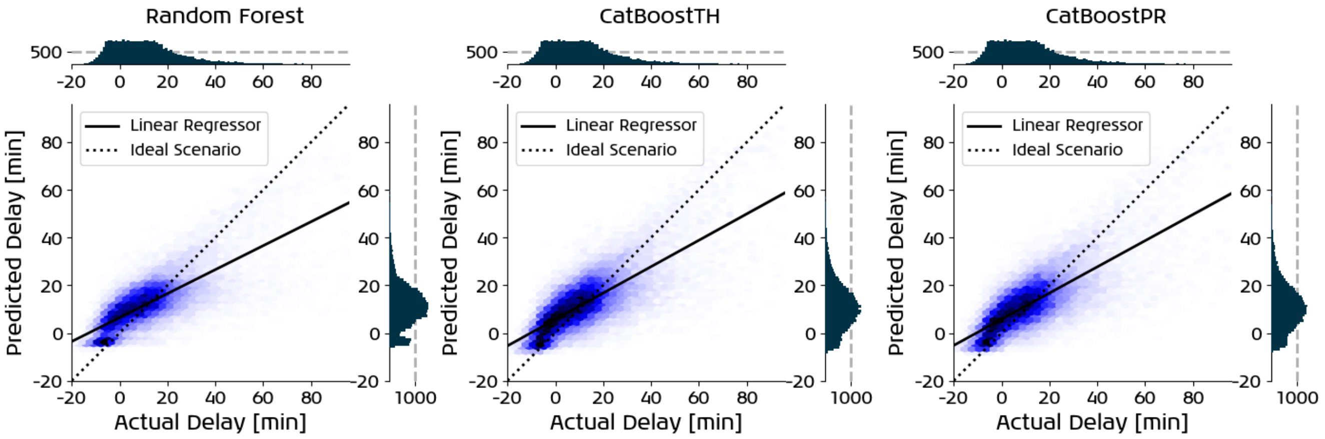


Figure 13. Prediction Accuracies for Random Forest, CatBoostTH and CatBoostPR models for the 90-Minute Prediction Horizon

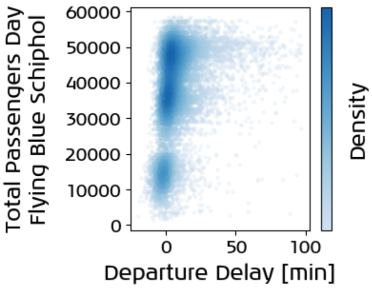


Figure 14. Correlation Total Passengers Day Flying Blue Schiphol with Departure Delay

For the Random Forest model in Figure 13, the distribution of predicted delays shows a valley around 0 minutes and a peak around -5 minutes. This behaviour is caused by the data from the COVID-19 period; excluding this data removes the peak. The total daily Flying Blue passengers feature is especially impacted during this period. Whereas normally these passenger counts range between 30000 and 55000, during the COVID-19 period the numbers were much lower creating the almost separated cluster at the bottom of Figure 14. The flights in this cluster are the same flights that form the peak in the distribution in Figure 13. The valley in the same distribution can be attributed to the low number of available data points in the range of 20000 to 27000 passengers. Following their respective distributions, both CatBoost models were better at correcting for this data anomaly.

[15] Time was instead reserved for accurately implementing the other models.

4.1.3 Probabilistic Model Performance

Based solely on the three global performance metrics, it may seem straightforward to select the CatBoostTH model for further use. Given the emphasis on explainability in this study, however, the probabilistic performance also needs to be evaluated. For the three remaining models, the ActInDistr and AvgIQR are tabulated in Table 7 and graphically presented in Figure 15.

Table 7. Probabilistic Model Performance for Out-of-Sample Data

		Prediction Moment [min]					
		-90	**-75**	**-60**	**-45**	**-30**	**-15**
Random Forest	ActInDistr [%]	76.09	76.82	77.51	78.60	79.17	80.00
	AvgIQR [min]	5.29	5.28	5.24	5.20	5.10	4.99
CatBoostTH	ActInDistr [%]	7.43	7.67	7.61	7.93	7.87	7.89
	AvgIQR [min]	0.39	0.39	0.39	0.38	0.38	0.36
CatBoostPR	ActInDistr [%]	42.23	42.71	43.30	43.44	43.99	45.20
	AvgIQR [min]	2.35	2.35	2.33	2.29	2.26	2.21

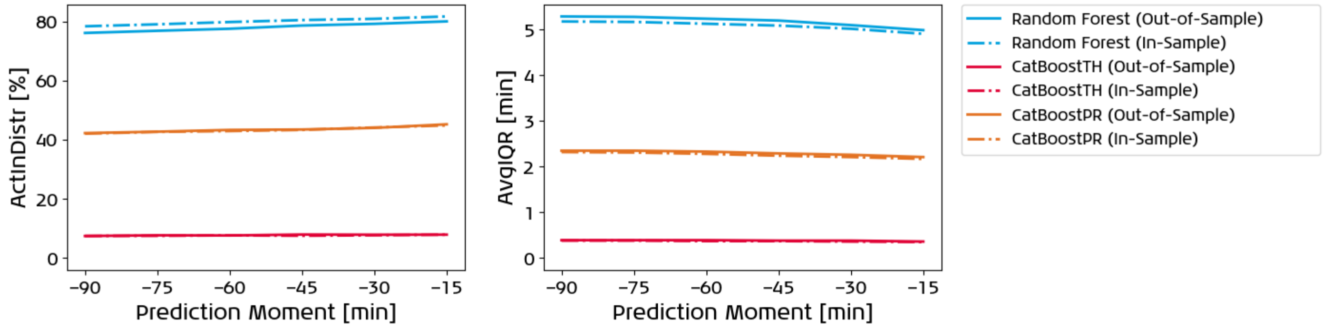


Figure 15. Probabilistic Model Performance over Time in terms of ActInDistr and AvgIQR

Evaluating the probabilistic performance is a trade-off between the two metrics. Ideally, a model has high correctness (i.e. high ActInDistr) and high confidence (i.e. low AvgIQR). Models with high ActInDistr and above-average AvgIQR are preferred over models with low ActInDistr and below-average AvgIQR, as the latter are confident yet incorrect models. Figure 15 shows that for around 78% of flights, the actual delays are within the predicted probability distributions for the Random Forest model, greatly outperforming the other two models (8% and 44%). Although the AvgIQR is higher for the Random Forest, the model refrains from predicting wide delay probability distributions; the model is still able to distinguish high and low likelihoods for different delay values.

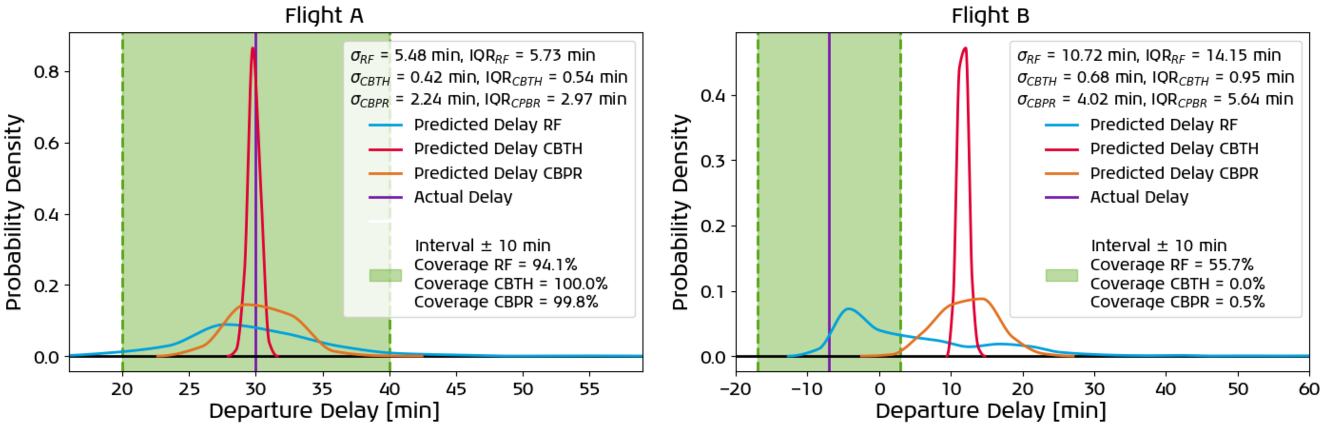


Figure 16. Probabilistic Departure Delay Predictions for Flight A (left) and Flight B (right) by Random Forest, CatBoostTH and CatBoostPR models (abbreviated RF, CBTH and CBPR respectively) for the 90-Minute Prediction Horizon

Figure 16 illustrates the probabilistic advantage of the Random Forest method over the two CatBoost models. To quantify the accuracy of the probabilistic prediction, a coverage is calculated. For the example flights, the percentage of predicted probability density within an interval of ±10 minutes around the actual delay is determined. For flights with small prediction errors (such as Flight A), all models achieve high coverages, especially the confident models. For flights with larger prediction errors (such as Flight B), the Random Forest reaches much higher coverages than both CatBoost models. The Random Forest model, despite being less confident, thus provides a better probabilistic prediction. From a user perspective, models are desired to indicate uncertainty associated to possibly incorrect predictions. If a model fails to do so, the user is likely to lose faith in the model over time.

Finally, the in ActInDistr and AvgIQR in Figure 15 only marginally improve over time. Although the AvgIQR is expected to decrease over time as a result of the updated dynamic features, this is only very slightly the case. Since flight delays commonly arise close to departure, the average predicted delays are higher for shorter prediction horizons. It is empirically observed that some relation exists between the magnitude of the predicted delay and the certainty of the model, as illustrated in Figure 17. As the Random Forest and CatBoostPR models predict higher delays, their predictions become less certain. This can be explained by the sampled dataset presented in Figure 3, which contains relatively more flights with small delays than flights with large delays. Statistically, the probability of the model being trained on comparable flights is higher for flights with small delays, which results in more confident predictions for such flights. For the extremely confident CatBoostTH model, this pattern is hardly visible, as the IQR remains small for almost all predictions.

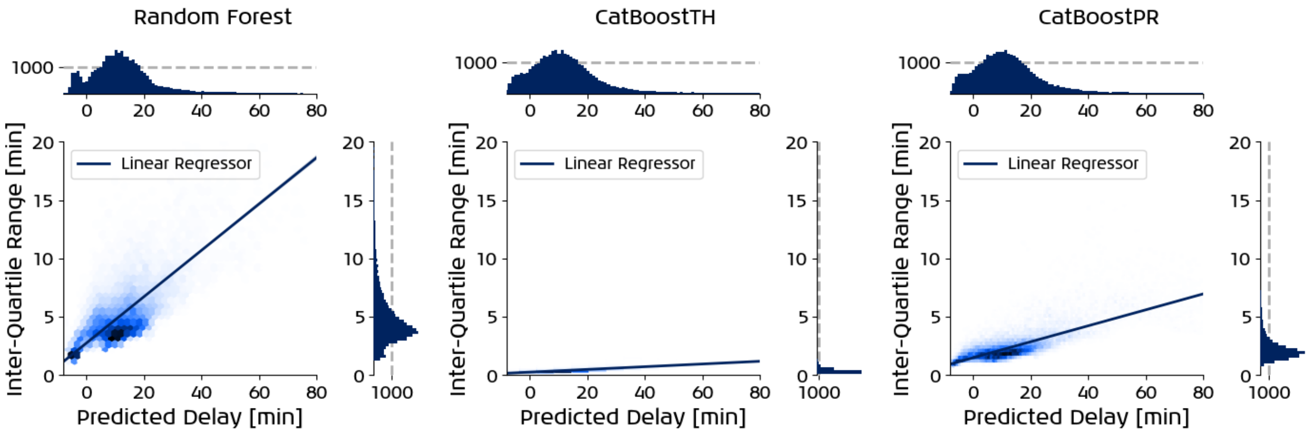


Figure 17. Relation between Uncertainties and Predicted Delays for Random Forest, CatBoostTH and CatBoostPR models for the 90-Minute Prediction Horizon

4.1.4 Binned Model Performance

Finally, the model performance per departure delay bin should be considered for the model selection. There is a physical limit[16] to the number of minutes a flight dispatcher can speed up a flight to compensate for departure delays; delays of over 30 minutes can hardly be compensated for, not even in the case of long-distance flights. Flights with departure delays smaller than 25 minutes are particularly interesting for flight dispatchers to slow down or speed up. For that reason, the sampling strategy outlined in Section 3.2.2 was adopted to optimize the model performance in this part of the departure delay spectrum. Figure 18 presents the model errors for departure delay bins of 5 minutes. The vertical bars represent the percentage of flights belonging to each bin.

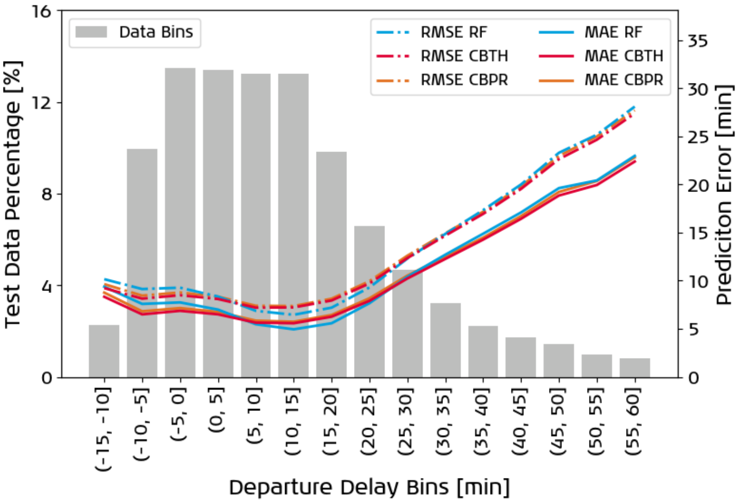


Figure 18. Binned Performance for Random Forest, CatBoostTH and CatBoostPR models (abbreviated RF, CBTH, and CBPR respectively) for the 90-Minute Prediction Horizon

The binned error distributions in Figure 18 are comparable, particularly those of the CatBoostTH and CatBoostPR models and for delays larger than 25 minutes. Although the errors for the departure delay range of -15 to 5 minutes are larger for the Random Forest than the CatBoost models, the errors for the range

[16] Dependent on many factors, e.g. flight distance, weather, aircraft weight, and ATFM.

of 5 to 25 minutes are smaller for the Random Forest model. In summary, despite having slightly larger global MAE and RMSE errors, the Random Forest model outperforms the two CatBoost models for the flights most suitable for en-route speed optimization. Combining this with the model's superior probabilistic performance leads to the decision to select the Random Forest for the remainder of the study.

4.2 Dynamic Model Prediction Behaviour

To obtain dynamic probabilistic departure delay predictions, the predicted probability distributions for all prediction horizons are combined. There exist millions of different ways in which the prediction error and certainty (IQR) may vary over the full prediction horizon. This makes it impossible to classify all unique flights into a reasonable number of categories. For that reason, the changes in prediction error and IQR between every two consecutive prediction moments are evaluated instead and are presented in Table 8.

Table 8. Percentage of Test Flights for which Prediction Error and IQR Increase or Decrease Between Prediction Moments

	Prediction Interval [min]				
	-90 to -75	**-75 to -60**	**-60 to -45**	**-45 to -30**	**-30 to -15**
Decreasing Error Decreasing IQR	29.06%	30.34%	30.44%	31.42%	33.96%
Decreasing Error Increasing IQR	24.47%	24.03%	24.20%	23.23%	20.37%
Increasing Error Decreasing IQR	22.14%	22.64%	22.52%	23.34%	26.26%
Increasing Error Increasing IQR	24.33%	22.99%	22.84%	22.01%	19.41%

Between the 90-minute and 75-minute prediction moments already, the prediction error and IQR decrease for more flights than for which they increase. These ratios only improve for shorter prediction horizons. Between the 30-minute and 15-minute prediction horizons, for over 60% of flights, the prediction confidence still improves. Although at shorter prediction horizons, a significant number of flights is predicted more accurately and with greater certainty, it should be pointed out that this is not the case for all flights. As previously mentioned, there are countless patterns for prediction error and IQR propagation over the prediction horizon. Even if one were to bin the flights into five prediction error bins and five IQR bins at the six prediction moments, as is done in Appendix 4, a total of $(5{\cdot}5)^6$ = 244,140,625 combinations are possible. A few (common) examples are discussed in the next paragraphs.

Flight C, illustrated in Figure 19, is an example for which the prediction error decreases as the certainty increases. Although the initial prediction error is relatively large, the model is able to correct for it as the dynamic features are updated. Alternatively, Flight D, illustrated in Figure 20, is an example for which the opposite is true. The model diverges from its initial prediction because, for smaller prediction horizons, it considers the TSAT delay (which is 11 minutes for all prediction moments) to be more important, see Section 4.3.1. Therefore, the predicted delay slightly increases over time. Just one minute after the final prediction moment, the TSAT delay was updated to -5 minutes, which explains that the actual delay is much lower than the predicted delays.

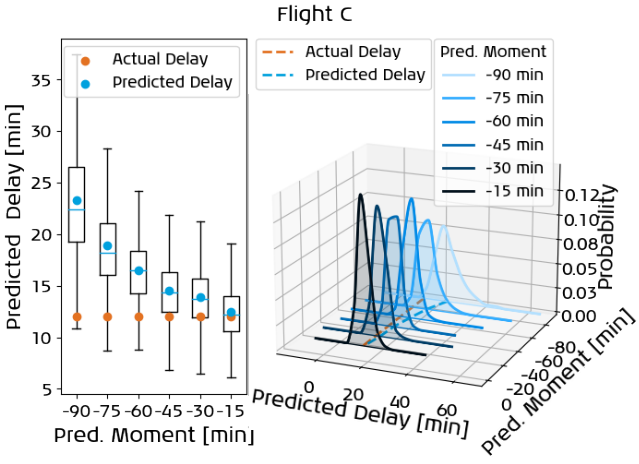


Figure 19. Dynamic Probabilistic Departure Delay Prediction for Flight C

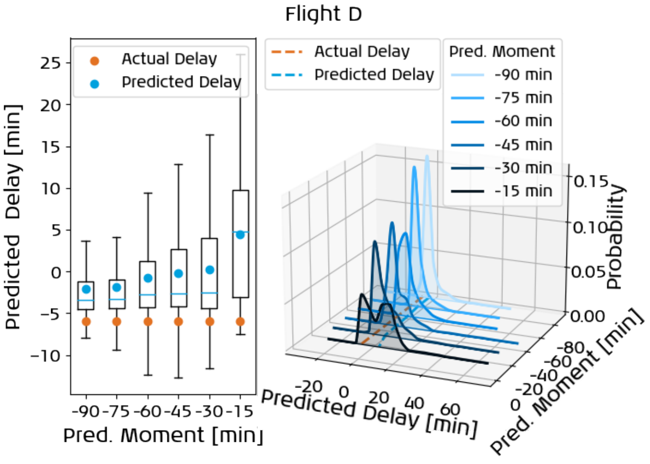


Figure 20. Dynamic Probabilistic Departure Delay Prediction for Flight D

For a large share of flights, the prediction error and IQR hardly vary over time. For 54% of flights, the prediction error changes for all prediction intervals combined is smaller than 5 minutes. Similarly, for 61% of flights, the

IQR changes for all prediction intervals combined is smaller than 3 minutes. For these flights, the model is either able to accurately predict the delays at the first prediction moment already or it is unable to improve its initial prediction. Flight E, illustrated in Figure 21, is an example for which the prediction error and IQR are small for all prediction moments. If instead, the error would be constantly large, this would result in a translation of the distribution with respect to the *Predicted Delay*-axis. Alternatively, changes in certainty would flatten or steepen the distribution curves.

Next to evaluating dynamic model errors, it is worth evaluating the dynamic model certainty. For a number of flights, the prediction certainty decreases as the prediction horizon becomes smaller. This is mostly caused by the model predicting higher delays, which results in additional uncertainty, previously explained in Section 4.1.3. Moreover, increasing uncertainty can be caused by contradicting features, for example flights with large effective delays of previous flights where the TSAT delay has not yet been updated. Fortunately, in many cases, the certainty increases over the prediction horizon, in line with Table 8. Flight F, illustrated in Figure 22, is an example for which the TSAT delay was large (45 minutes) for the first five prediction horizons and changed to just 12 minutes at the final prediction horizon, allowing the model to make a final prediction with much higher confidence.

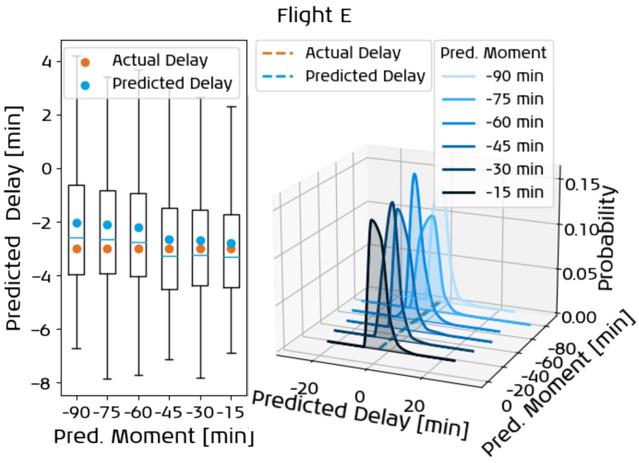


Figure 21. Dynamic Probabilistic Departure Delay Prediction for Flight E

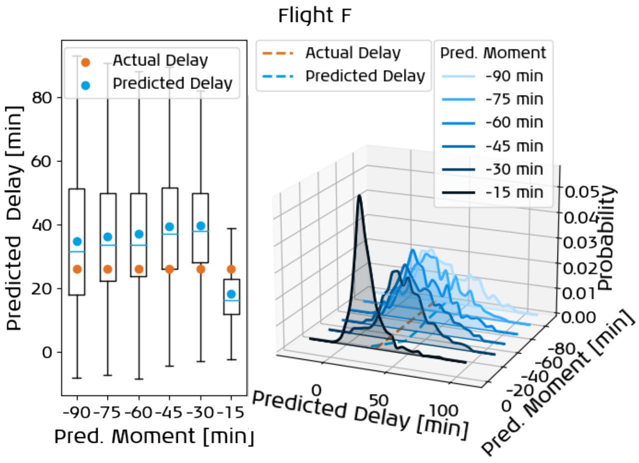


Figure 22. Dynamic Probabilistic Departure Delay Prediction for Flight F

For some flights, the prediction error and IQR temporarily increase for some parts of the prediction horizon. Temporary changes in prediction error are almost always the cause of temporary changes in prediction certainty. Flight G, illustrated in Figure 23, is an example for which the prediction error is temporarily higher, caused by a temporarily larger TSAT delay value. This causes the probability distribution to temporarily translate with respect to the *Predicted Delay*-axis. Flight H, illustrated in Figure 24, is an example for which temporary high average delays at Schiphol cause the prediction error to increase, thereby temporarily decreasing the prediction certainty. The coupling between the prediction error and uncertainty causes the probability distribution to be translated and flattened simultaneously.

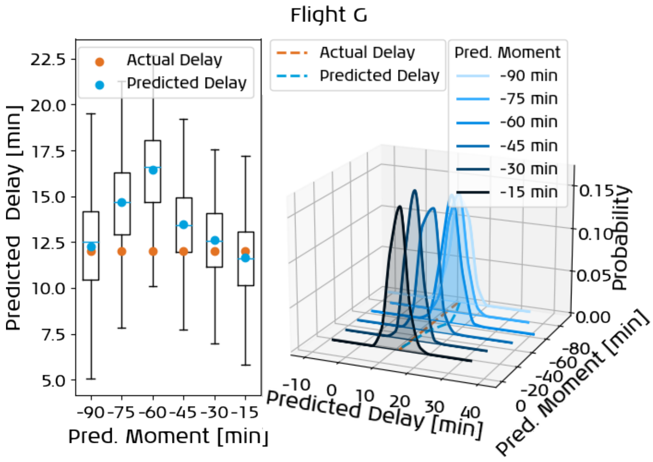


Figure 23. Dynamic Probabilistic Departure Delay Prediction for Flight G

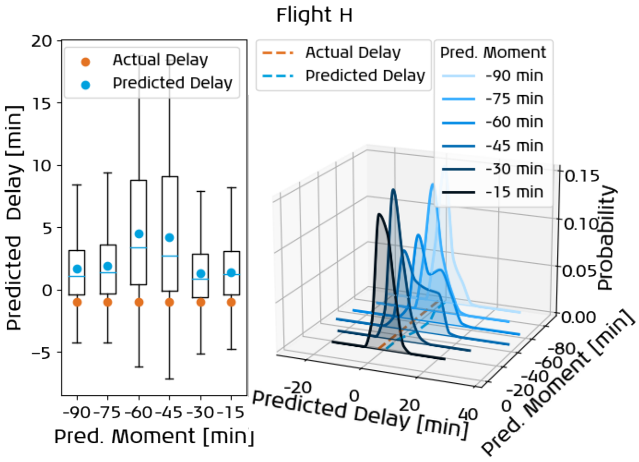


Figure 24. Dynamic Probabilistic Departure Delay Prediction for Flight H

4.3 Model Explainability Results

The feature importances associated with the Random Forest model are presented in Section 4.3.1 and the results of the relative feature scaling method are presented in Section 4.3.2, with two examples of explainability messages outputted to the flight dispatcher.

4.3.1 Feature Importance

Given the dynamic feature updates and their improving correlations with departure delays, presented in Figure 12, the feature importances are expected to change over time as well. The feature importances for the 90-minute prediction horizon as well as their propagation over time are presented in Figure 25. The ten least important features do not change significantly over the full prediction horizon. The shifts noticed in the five most important features are more interesting to evaluate. Whereas at longer prediction horizons, the model considers passenger information more important (ranked 3rd and 4th), for the shorter prediction horizons, these importances drop to 4th and 5th place respectively, almost halving in magnitude. Dynamic features describing the current TSAT delay and average Schiphol delays become more important for shorter prediction horizons (1st and 3rd compared to 2nd and 5th place). This is in line with the increasing dynamic feature correlations for shorter prediction horizons.

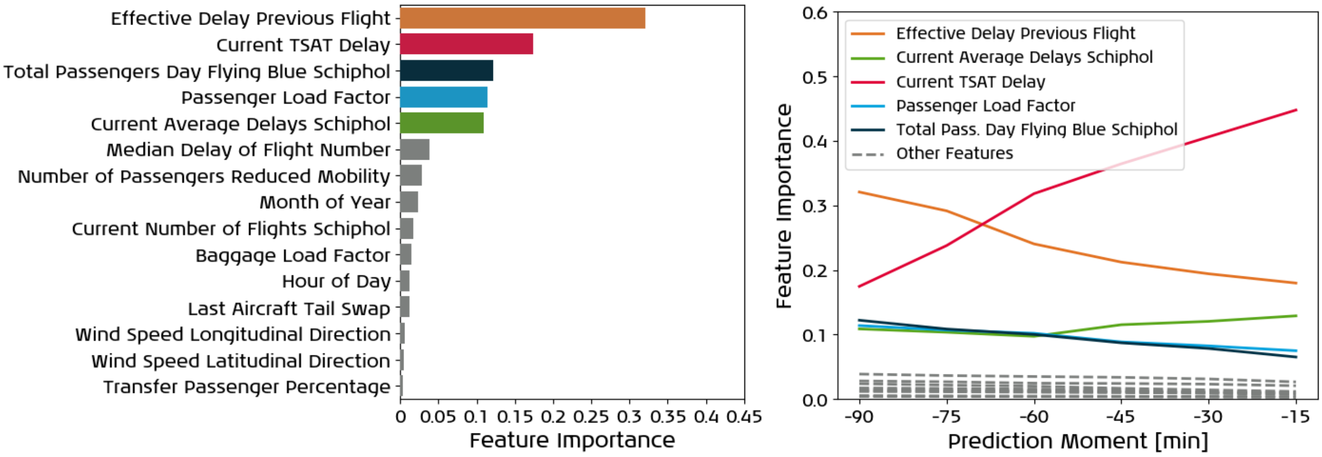


Figure 25. Feature Importance for the Random Forest Model for the 90-Minute Prediction Horizon (left) and the Feature Importance Developments over the Full Prediction Horizon (right)

4.3.2 Explainability Through Feature Scaling

Following the relative feature scaling method presented in Section 3.5, explainability messages are provided to the flight dispatchers, indicating the significance of a given feature for this flight relative to all others. Including this information in the explainability message was one of the wishes of the flight dispatchers. The explainability message for Flight E, previously discussed in Section 4.2, is presented below for both the 90-minute and 15-minute prediction horizon. From the messages, the user is informed that the lower delay prediction at the 15-minute prediction horizon is caused by the decrease in current TSAT delay and current average delays at Schiphol. Finally, the impact of delays is excluded for flights with predicted delays of less than 15 minutes because these delays typically occur due to *quasi-random* operational factors close to departure, not due to the major delay causes the model was trained for.

```
Flight Date:          2022-01-12
Flight:               KL****
Departure Airport:    AMS
Prediction Moment:    -90 min
Predicted Delay:      34.65 min
Large impact from:    Current TSAT Delay (90 min. before Scheduled  Departure) [min]: 45.0 (average = 6.7)
Moderate impact from: Current Average Delays Schiphol (120 to 90 min. before Scheduled Departure) [min]: 72.44 (average = 14.1)

Flight Date:          2022-01-12
Flight:               KL****
Departure Airport:    AMS
Prediction Moment:    -15 min
Predicted Delay:      18.25 min
Small impact from:    Current Average Delays Schiphol (45 to 15 min. before Scheduled Departure) [min]: 43.31 (average = 14.12)
Small impact from:    Current TSAT Delay (15 min. before Scheduled Departure) [min]: 12.0 (average = 8.42)
```

5. Model Validation

This section aims to present the validation practices that were performed to guarantee the applicability of the model. First, the sensitivity study is discussed in Section 5.1. Thereafter, the conclusions drawn from two shadow runs are presented in Section 5.2. Finally, an error analysis is performed in Section 5.3, evaluating the flights that were most difficult for the model to predict.

5.1 Sensitivity Analysis

To guarantee the robustness of the model, a sensitivity analysis was performed. The effect of removing features completely is presented in Appendix 5. This section focuses on the effect of altering a single input feature, one at a time. For this, a baseline flight was set up, the values and respective changes are tabulated in Table 9. The baseline values and changes were chosen such that hypothetically, outputs are most likely to lead to meaningful changes[17]. The sensitivity of the mean predicted departure delay is shown in Figure 26, for each of the six prediction moments separately. Furthermore, the probabilistic sensitivity of the model is illustrated in Figure 27, for the prediction moment 90 minutes before scheduled departure.

Table 9. Sensitivity Flight Baseline Parameters

Feature	Baseline	Change
Month of Year	4	±3
Hour of Day	14	±4
Passenger Load Factor	0.75	±0.1
Baggage Load Factor	0.85	±0.1
Transfer Passenger Percentage	60	±10
Number of Passengers Reduced Mobility	5	±4
Total Passengers Day Flying Blue Schiphol	42000	±5000
Median Delay of Flight Number [min]	10	±10
Effective Delay Previous Flight [min]	15	±10
Current Number of Flights Schiphol	15	±5
Current Average Delays Schiphol [min]	10	±10
Current TSAT Delay [min]	20	±10
Last Aircraft Tail Swap [min]	-900	±300
Wind Speed Longitudinal Direction [kts]	10	±5
Wind Speed Latitudinal Direction [kts]	10	±5

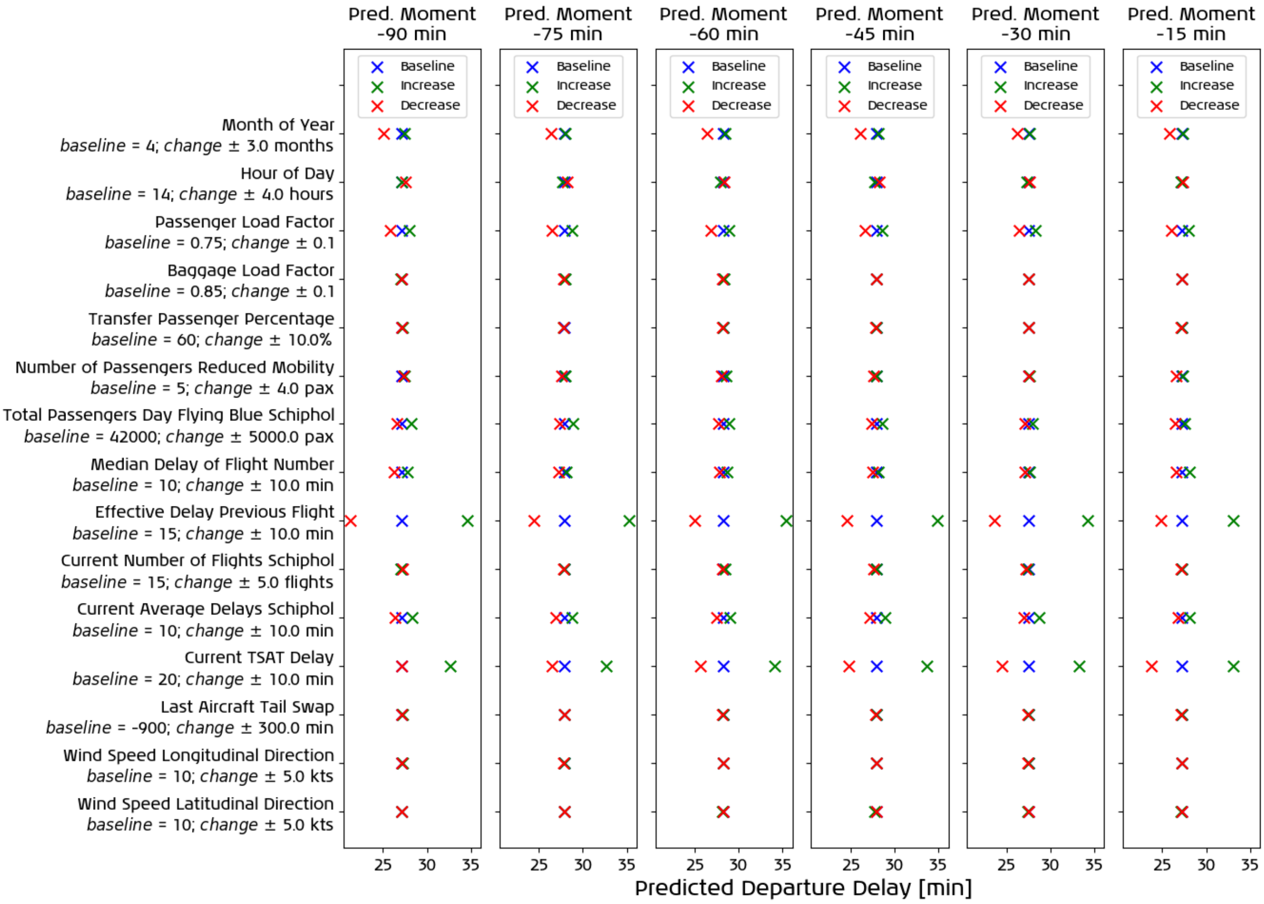


Figure 26. Sensitivity Analysis for Random Forest Model using Baseline and Change Values from Table 9

[17] By choosing baseline values and changes that are not towards the extremes of the feature value range, e.g. if the feature ranges from 0.4 to 1, changes are expected to be more meaningful when comparing 0.65, 0.75 and 0.85 instead of 0.4, 0.41 and 0.42 or 0.98, 0.99 and 1.

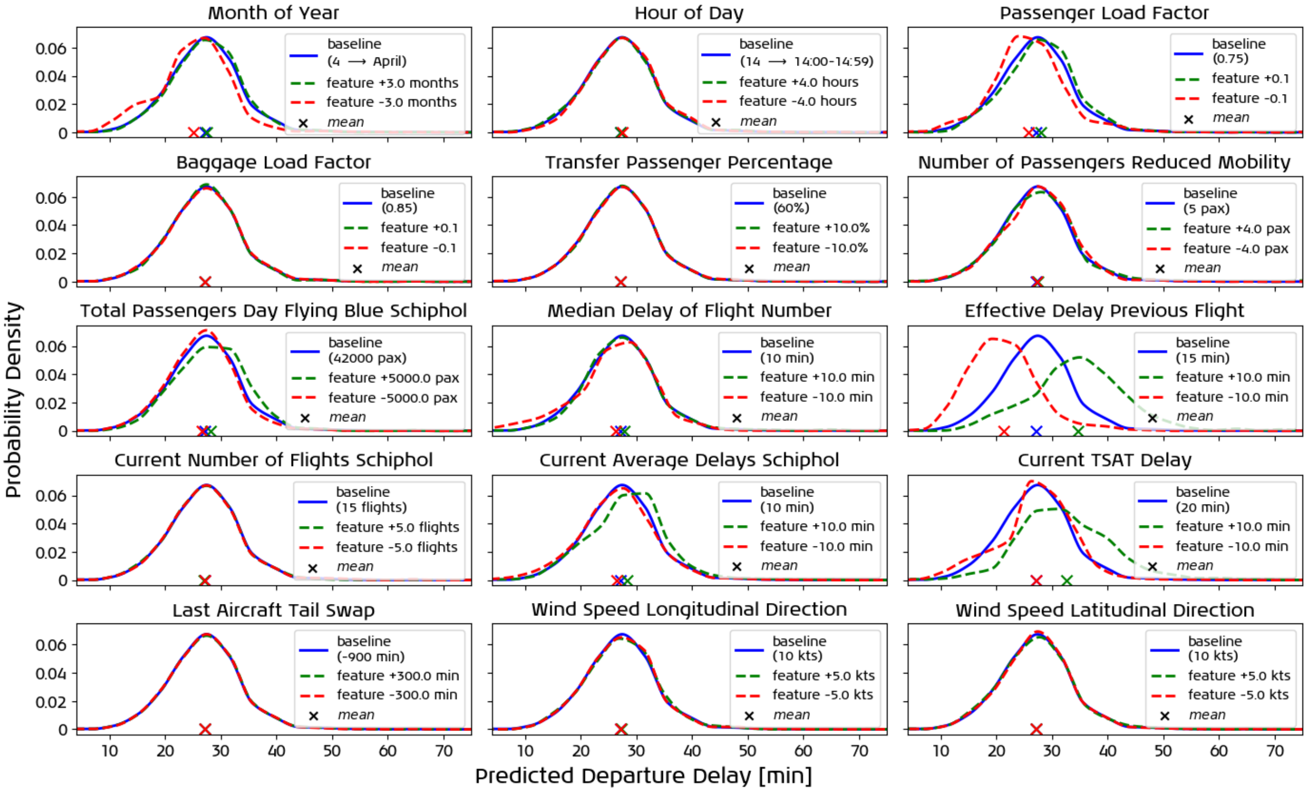


Figure 27. Probabilistic Sensitivity Analysis for Random Forest Model at 90-Minute Prediction Horizon using Baseline and Change Values from Table 9

In Figure 26 and Figure 27, the predicted departure delay hardly varies for changes on the hour of day, baggage load factor, transfer passenger percentage, passengers with reduced mobility, current number of flights at Schiphol, last aircraft tail swap and wind speeds features. This is explained by their relatively low importances in Section 4.3.1. Although among the six most important features in Section 4.3.1, the model is hardly sensitive to changes in the median delay associated with the flight number. The effect of changes in other parameters is better visible. Compared to the baseline month of April, departure delays are predicted to be slightly lower if the flight was to be scheduled in January. Similar relations are obtained for the passenger load factor, current average delays at Schiphol, and the total daily Flying Blue passengers. These small differences can be explained by seasonality, as historically the delays are smaller in months with fewer passengers (such as January).

The effective delay of previous flight is the most sensitive feature. At the 90-minute prediction horizon, an input change of 10 minutes results in an output change of almost the same size. At closer prediction moments, complying with the decreasing feature importance from Section 4.3.1, the effective delay of previous flight becomes slightly less sensitive. From the probabilistic sensitivity in Figure 27, it is visible that the model is less confident in predicting larger departure delays. The same behaviour was previously observed in Section 4.1.3.

The model becomes more sensitive to the TSAT delay feature for smaller prediction horizons. It was previously found that this feature becomes more important for smaller prediction horizons. Interestingly, the model is more sensitive to TSAT delay increases than TSAT delay decreases. From historical data, it is observed that TSAT delays increase more frequently than they decrease. Once a new TSAT delay is issued, operations are often centred around accommodating this new time. As such, it rarely happens that a TSAT delay decreases after it has previously increased. Even in the case of decreases, it is often observed that a new increase is issued later in the process. The model has thus successfully learned that a TSAT delay increase is a stronger indicator for higher delays than a TSAT delay decrease is for smaller delays. Concluding, the robustness of the model has been demonstrated in this sensitivity analysis, which contributes to the overall model validation.

5.2 Shadow Run

It is insightful to analyse the responses of the model to real-life scenarios it has not been trained for. For this reason, two shadow runs were conducted: one for European flights and one for intercontinental flights.

Table 10 lists the flights considered during both shadow runs. Because the input data was not available in real-time, the model performance could only be evaluated as soon as the data became available. Some of the flights were delayed for reasons the model was not trained for, these cases are elaborated upon in the next paragraphs. Although the model can predict delays caused by untrained factors to some extent, it cannot predict all such delays. The model strongly benefits from the current TSAT delay feature[18], since it covers a wide range of operational delay causes. Due to time constraints and the added workload for flight dispatchers, it was infeasible to conduct more than two shadow runs. Before implementation of the model, however, additional shadow runs are necessary to assess the model performance for a larger variety of untrained delay causes.

Table 10. Shadow Run Flights

Date	Flight Number	Aircraft Type	Actual Delay [min]	Predicted[a] Delay [min]
2023-10-10	Flight I	A330	14.00	29.36
	Flight J	B787	48.00	59.91
	Flight K	B787	32.00	32.41
	Flight L	B777	27.00	27.22
	Flight M	B777	29.00	28.32
	Flight N	B787	5.00	15.07
2023-10-27	Flight O	B737	9.00	8.41
	Flight P	B737	32.00	32.96
	Flight Q	B737	7.00	12.87
	Flight R	B737	12.00	15.09
	Flight S	B737	-4.00	15.31
	Flight T	B737	15.00	18.59
	Flight U	B737	16.00	15.34
	Flight V	B737	15.00	17.25
	Flight W	B737	5.00	16.04

[a] Predicted delay at the 90-minute prediction horizon.

- **ATFM Slot:** Flight M, illustrated in Figure 28, received an ATFM slot of 27 minutes just after the first prediction moment. This caused the TSAT delay of the flight to get extended by the same amount. To avoid the slot, the flight dispatcher searched for an alternative route to avoid the overcrowded sector. After finding a suitable route, the slot time and TSAT delay returned to their previous values, as is visible in the departure delay prediction, where temporarily higher delays with larger uncertainty are predicted. The flight dispatcher thus is part of the loop, as his/her actions influence new predicted delays.
- **High-priority Flight:** Flight I, illustrated in Figure 29, was a high-priority flight because of an important part delivery. Because of a delayed previous flight, the model severely overpredicts for this flight. Given the high priority, every effort was made to keep the turn-around time as small as possible. The model is unable to capture the high priority, as it is an exceptional circumstance.

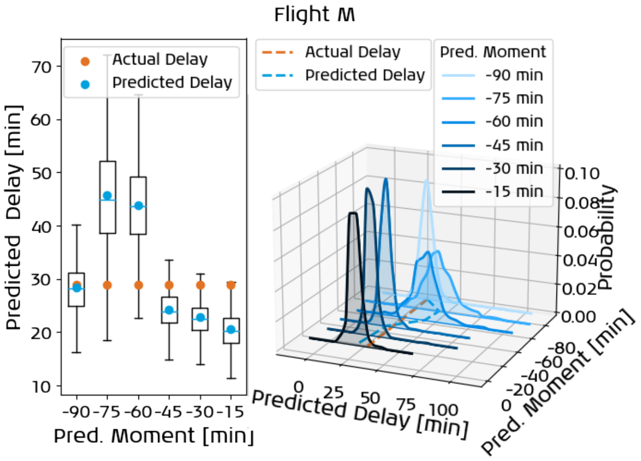


Figure 28. Dynamic Probabilistic Departure Delay Prediction for Flight M

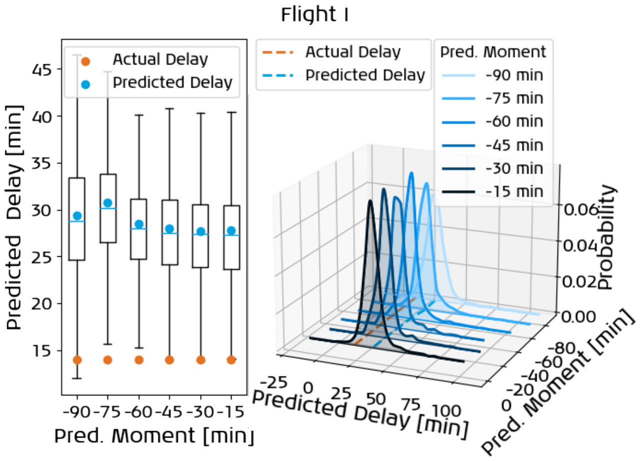


Figure 29. Dynamic Probabilistic Departure Delay Prediction for Flight I

- **Late Fueling:** Flight T, illustrated in Figure 30, was delayed because the fueling team arrived later than planned. The model captures this effect, as the predicted delay increases at smaller prediction horizons, as a result of increasing TSAT delays. The TSAT delay feature thus covers the fueling delay.
- **Late Arrival Baggage:** Flight R, illustrated in Figure 31, was delayed because the baggage carts arrived later than planned. The model is unable to capture the late arrival of baggage carts, despite small increases in TSAT delays. Instead, the model most likely overpredicts because of a fully booked flight.
- **Delayed Maintenance:** Flight K was delayed for 32 minutes because of a late return from maintenance. Since the model is not trained with maintenance data, the effective delay of the previous flight is thought to be 0 minutes, as the aircraft had arrived the day before already. Although the model does not understand

[18] TSAT delays often follow from Target Off-Blocks Time (TOBT) delays, which are issued by the airline itself.

the maintenance delay directly, it correctly predicts the departure delay because the maintenance delay was already known 125 minutes before scheduled departure. For that reason, the TSAT was already updated before the first prediction moment. The TSAT delay feature thus covers the maintenance delay.

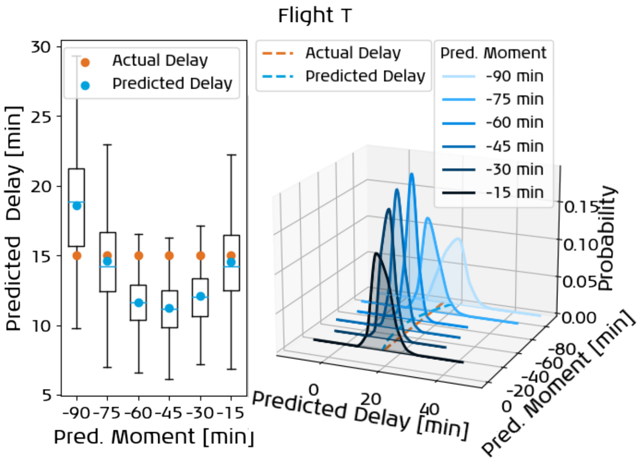


Figure 30. Dynamic Probabilistic Departure Delay Prediction for Flight T

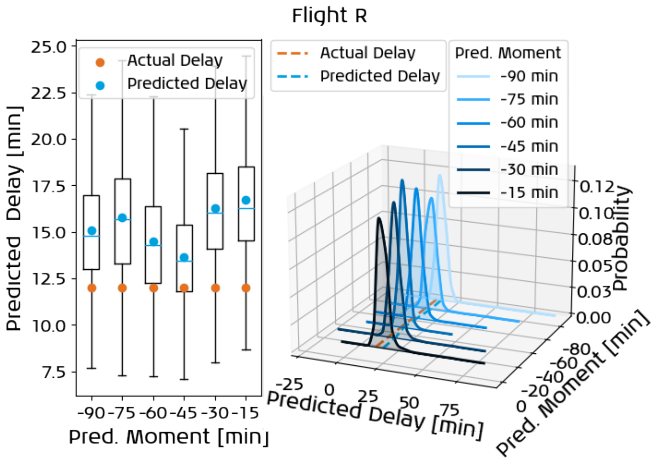


Figure 31. Dynamic Probabilistic Departure Delay Prediction for Flight R

5.3 Error Analysis

For 179 out of the 33532 flights in the test dataset, the model predicts with an error of over 45 minutes. It is worth investigating the reason for such errors. Following IATA guidelines [28], primary (and secondary) delay codes are issued for delayed flights. These delay codes explain the cause of the flight delay and are thus useful for explaining high prediction errors. It should be noted that ambiguity may exist for the issued delay cause, as different stakeholders have different interests for the delay code issuing. For the 179 flights, an overview of issued delay codes[19] is presented in Figure 32. Primary and secondary delays are represented by Delay Code 1 and Delay Code 2, respectively.

Firstly, it can be observed that over three-fourths (75.4%) of flights with large prediction errors are caused by just 17 delay codes. For the majority of flights with large prediction errors, the model is not trained for the underlying delay causes because the data is simply unavailable (e.g. loading, fuelling, ATFM delay, and crew rotations) or unpredictable (e.g. security/immigration, missing passengers, and flight deck crew request). This may be one of the reasons that only these delay causes appear more often in this analysis.

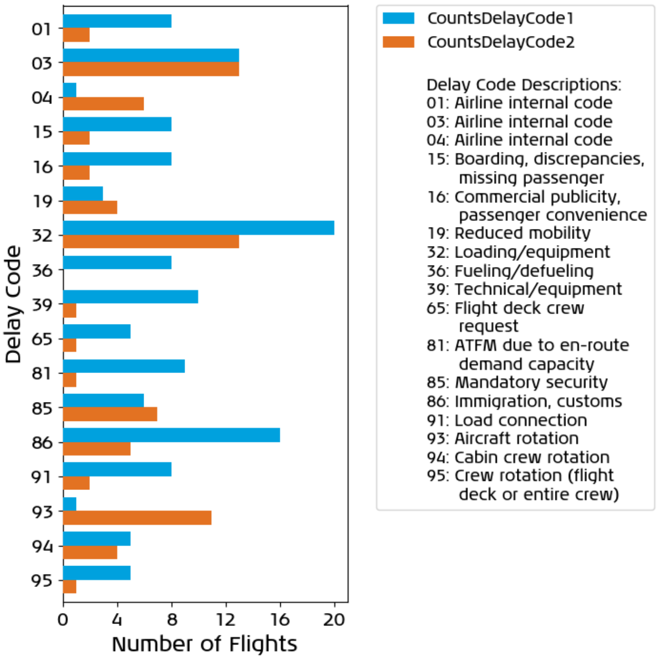


Figure 32. Delay Code Issued for Flights with Highest Model Prediction Errors

Given that primary delay codes contribute the most to flight delays and that the model accounts for the number of passengers with reduced mobility and the effective delay of previous flights, it is worth exploring why these factors were responsible for the main delay of some flights. For the three flights with delays caused by passengers with reduced mobility, the number of such passengers was 2, 7, and *unknown* respectively. Seeing that most other flights with these numbers of passengers with reduced mobility are only little delayed, it is no surprise that the model underpredicted the delay for these three flights. For the one flight with a delay due to aircraft rotations, the inbound flight effectively arrived 26 minutes late, however, the departure delay was much larger (91 minutes).

[19] That were issued at least five times.

Another unknown reason must have caused the remainder of the delay. It thus makes sense that the model is not able to correctly predict the delay for this flight just based on the effective delay of the previous flight.

In conclusion, the model is considered to be robust to changes in input features and has demonstrated its capability of predicting delays caused by untrained factors. Due to the limited time frame of this study and the fact that before a shadow run, it is unknown what will cause the delays for the upcoming flights, only a number of these untrained factors could be tested. Further shadow runs are necessary to guarantee the applicability of the model in a broader sense. Finally, the error analysis demonstrated the limitations of the model in situations it has not been trained for or that are simply unpredictable.

6. Discussions

This section aims to further discuss some of the results. The feature elimination process is reflected upon in Section 6.1. The effect of using data from the COVID period is discussed in Section 6.2. The overpredicting behaviour for small delays is evaluated in Section 6.3. The model performance is compared to that of the model currently in use at KLM in Section 6.4. Finally, the retraining of the model is discussed in Section 6.5.

6.1 Selected Features

The list of selected features was presented in Table 4. Some interesting features from the full feature list, see Appendix 2, were not selected. Although visibility was an important feature in previous work by Sternberg et al. [14], it was either covered by one of the other features, reducing the need for a separate feature, or the number of recordings with extreme fog was too low to make the models understand its effect. It should be noted that Schiphol possesses cutting-edge technology to support reduced visual operations.

The first research gap presented in Section 2.2 considers the use of flight-specific passenger connection data for departure delay forecasting. Features 18-22 and 35 from the list in Appendix 2 were developed for this purpose. Furthermore, features 28 and 29 were developed to mimic the decision-making by airlines on whether or not to delay a flight because of connection passengers. None of these features were selected, however. Instead, feature 34, describing the *percentage* of transfer passengers, was selected. Although this feature does not consider every inbound-outbound flight combination, it still stands out from other research as the feature is unique per outbound flight. This does imply, however, that the flight-specific passenger connection data was not valuable enough to predict departure delays. As a side note, KLM strives for timely arrivals to guarantee connections. The model refraining from selecting the flight-specific passenger connection data could suggest that KLM's passenger connection operations are smooth. If more historic flights had been delayed because of connection passengers, the model might have placed more value on the flight-specific passenger connection features.

6.2 Including Data from COVID period

Including the data from the COVID period allowed the model to be trained on a wider variety of historical data. Additionally, in Section 3.2.2, the sampling practices were improved because of the presence of more flights with negative and near-zero delays. Using only post-COVID flight data would mean a large reduction in available training data and would result in higher prediction errors and worse model fit.

On the other hand, the data from the COVID period was responsible for the peak in predicted delays around -5 minutes in Figure 13 and the separated cluster in Figure 14. Including the COVID period data therefore trained the model to believe that the range of total daily passengers is larger than it is for regular operations (i.e. without the separated cluster). Over time, it could make the model predict higher delays because the total daily passengers feature value appears to be relatively high, although in reality it is not so large but it was unusually low in the COVID period.

6.3 Overpredicting For Small Delays

Following Figure 13, at the 90-minute prediction horizon, flights with actual delays up to 15 minutes are overpredicted by 5.17 minutes on average. Alternatively, flights with actual delays of over 15 minutes are underpredicted by 9.68 minutes on average. The overall overprediction distribution for the Random Forest model is presented in Figure 33, for each of the six distinct prediction horizons.

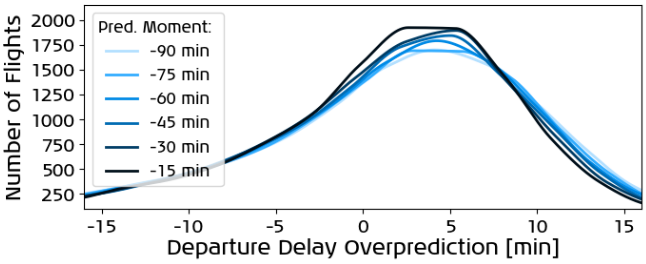


Figure 33. Random Forest Departure Delay Overprediction

As explained in Appendix 1, overprediction leads to unnecessary fuel burn. For that reason, every effort was made to find the root cause of the overpredicting behaviour. All of the following options were tested: splitting the model into European and intercontinental flights, sampling at different strategies (sampling to other bins, normal and bimodal distributions), changing the outlier removal percentages (2% and 3% instead of 1%), changing the criterion/loss function (from MSE loss to MAE loss[20] and Friedman MSE loss), changing the hyperparameters (number of estimators, maximum number of features, maximum depth, minimum samples for splits and leaves), removing features with high inter-correlation (such as the Total Passengers Day Flying Blue Schiphol feature), removing the COVID period data and removing outliers on each of the feature value ranges.

None of these actions led to a reduction of the overpredicting behaviour. Although increasing the maximum depth of the model slightly lowered the overpredictions, it resulted in models that were overfitted and therefore not usable. Changing the loss function to MAE loss is promising as it would assign less value to larger errors for larger predicted delays. This would increase RMSE and decrease global fit, however. Given the extremely long run times[20], it was infeasible to test this loss function for the 1000 estimators and around 135,000 training flights.

Finally, the overpredicting behaviour for smaller delays is attributed to two things. Firstly, almost all feature-target relations are positive[21], see Figure 34. As a result, when some feature values are above average, the model may already be inclined to predict higher delays since there are no features that impact the predicted delays negatively. Secondly, given the randomness involved with the turn-around process, the model is not always able to accurately predict the delay (R2 is *only* 0.55 at the 90-minute prediction horizon). The overpredicting behaviour may also partially be caused by the fact that not all required information is already known at early prediction moments. This is visible in Figure 33, where the overpredicting behaviour reduces for shorter prediction horizons, but does not disappear completely. From a business operations point-of-view, mitigating the overpredicting problem can be done by implementing a rule that corrects for the overprediction tendency, for example using a correction factor. Further scientific research is required to investigate how such correction factors can adequately mitigate the overprediction problem.

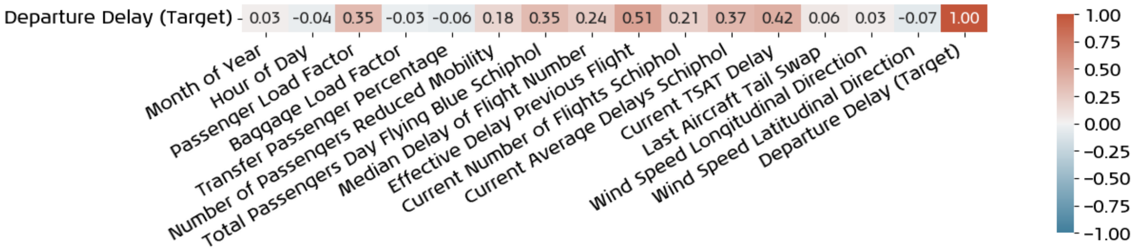


Figure 34. Target Correlation Matrix for the 90-Minute Prediction Horizon

6.4 Improvement Compared to Existing Model

As outlined in Section 3.1, one of the aims of the study is to improve on an existing statistical model. When considering the same case study period, the existing model reaches an MAE of 9.51 minutes, an RMSE of 18.62 minutes, and an R2 of 0.13. Since this model does not consider multiple prediction moments, these performance metrics values are the same for all prediction horizons. Following Table 6, the proposed Random Forest model outperforms the existing model at all prediction metrics, for all prediction moments.

The prediction performance of both models at the 90-minute prediction moment is presented in Figure 35, considering just a single flight series (upper figure) and a single day at Schiphol (lower figure). Although the existing model captures the global dynamics, judging from the upper plot of Figure 35, it is unable to predict large delays. The proposed Random Forest model is much better capable of predicting such large delays. Whereas the existing

[20] The MAE loss function is several orders of magnitude slower than the MSE loss function because it is not differentiable at zero.

[21] The feature-target relations that are not positive are near zero.

model overpredicts less for smaller delays, it is useless for days with disrupted operations, such as 2023-03-13, see the lower figure in Figure 35. The proposed Random Forest model is particularly of added value for the prediction of severely delayed flights.

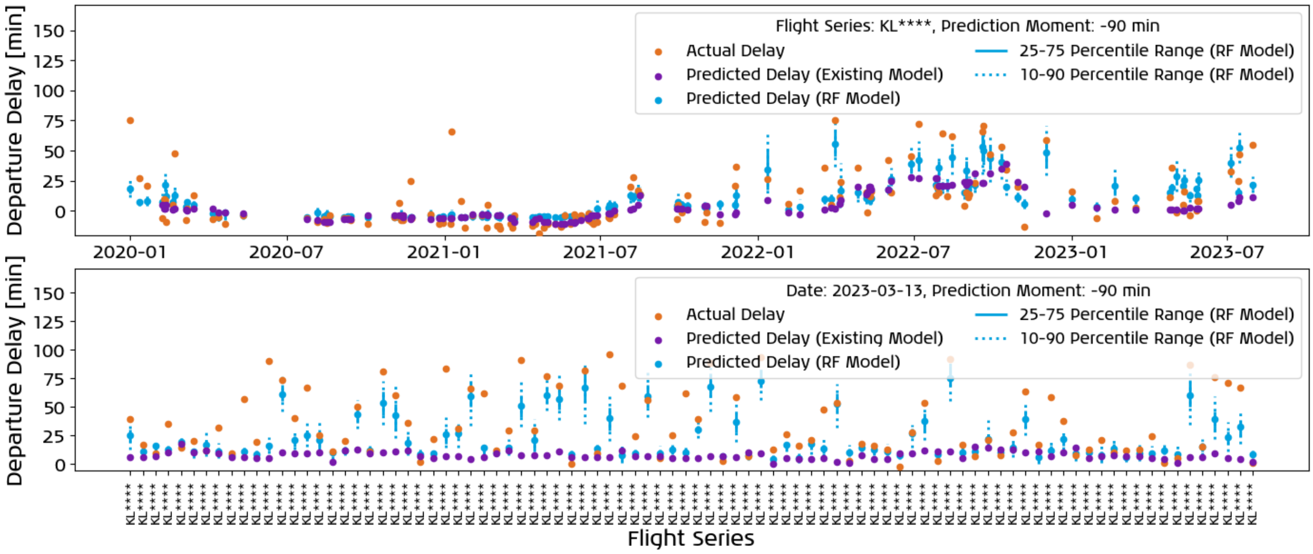


Figure 35. Existing and Proposed Random Forest (RF) Model Prediction Performance for Flight Number KL**** (upper figure) and Date 2023-03-13 (lower figure) at 90-Minute Prediction Horizon

6.5 Model Retraining

The prediction errors of the Random Forest model over time are presented in Figure 36. To prevent data leakage, the training and testing data was split by date, which explains that part of the time-axis seems empty. A 100-flight rolling average of the prediction error was computed to reduce the impact of outliers. The longer the training dataset period, the less frequent updates are required, as the large data evens out some temporary effects. Upon retraining, the feature describing the median delay of the flight number automatically updates to more representative values. Taking this into account, together with the 3.5 years of training data, it is advised to retrain the model every 5-7 months[22]. Should the prediction error systematically increase for too long, it is advised to retrain the model sooner.

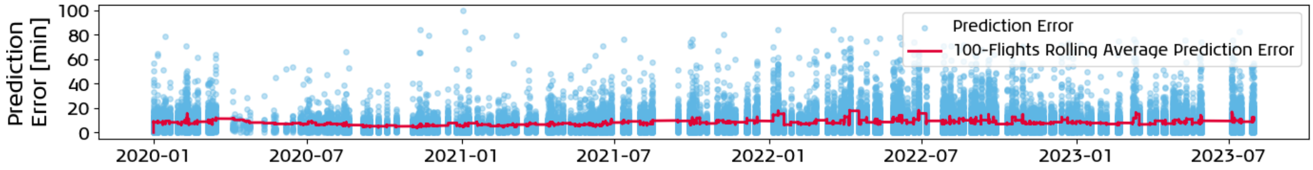


Figure 36. Random Forest Prediction Error over Time

7. Conclusions and Recommendations

Hub-and-spoke airlines generally adjust their operations to guarantee passenger connections. For that reason, punctuality is one of the key performance indicators of such airlines. To ensure on-time arrivals, flights that were delayed upon departure need to compensate for the lost time whilst airborne. For adequate fueling, flight dispatchers use departure delay predictions. The goal of this study was to propose an explainable supervised learning model that improves on an existing departure delay prediction model, as there was room for improvement.

From the large flight dataset available through the research collaboration with KLM, a substantial number of features were proposed. Simultaneously, Random Forest, CatBoost, and Deep Neural Network models were developed. The feature elimination process addressed the first research gap, revealing that including flight-specific passenger connection data would not improve the models. Several day-of-operations features, such as current

[22] Preferably at the beginning of the winter and summer flight schedules.

average delays, TSAT delays, and aircraft tail swaps, were selected in the feature elimination process and thereby contributed to the second research gap. Finally, through the probabilistic approach for the Random Forest and CatBoost models by using the independent predictions of all estimators, it was possible to obtain dynamic probabilistic departure delay forecasts for individual flights, contributing to the third research gap.

Despite reaching slightly larger overall MAE and RMSE errors, the Random Forest was selected as it outperformed the other models for the flights most suitable for en-route speed optimization and demonstrated superior probabilistic performance. The dynamic probabilistic model performance analysis then indicated that for shorter prediction horizons, the model was able to improve on initial predictions for a large number of flights, both in terms of correctness and certainty. At the default 90-minute prediction horizon, the model reaches an MAE of 8.46 minutes, an RMSE of 11.91 minutes, and an R2 of 0.55. At the 15-minute prediction horizon, these values improve to 7.37 minutes, 10.44 minutes, and 0.65, respectively. At all prediction moments, for around 78% of flights, the actual delay was within the predicted departure delay probability distribution. For the flights that are most suitable for en-route speed optimization, the Random Forest model reached MAE values of around 5 minutes.

The model was validated through a sensitivity analysis in which features were altered and removed, along with two shadow runs performed in the KLM Operational Control Centre (OCC) and an analysis of the largest prediction errors of the model. Thereafter, the proposed Random Forest model was critically reflected upon, leading to the conclusion that the proposed model outperforms the existing model in terms of error, explainability and probabilistic performance. Additionally, the proposed model is particularly suitable for detecting and predicting large departure delays. Compared to the existing model, the improvement for predicting smaller delays is relatively small, as a result of the overpredicting behaviour due to positive feature correlations and the absence of valuable information at large prediction horizons. Despite thorough evaluation of all model parameters, training settings and the underlying dataset, the root cause of the overprediction behaviour was not found and thus requires further research.

Several other logical follow-up steps can be taken to further improve dynamic probabilistic airline departure delay forecasting. Firstly, one can test the impact of incorporating other novel features. It is recommended to explore the effect of push-back truck availability data and crew rotations as both are critical in the turnaround process. Secondly, for future work it is recommended to consider actual departures rather than planned departures for the current average delays at the airport, in order to always have complete data. Thirdly, the granularity of the prediction horizons can be increased to 5 or 10 minutes to allow for quicker incorporation of dynamic feature updates. Moreover, the performance of the Deep Neural Network can be further investigated, also taking into account how to better explain its predictions. Finally, for implementation of the proposed Random Forest model, it is recommended to perform further validation shadow runs to guarantee the applicability of the model to a wider range of untrained delay causes.

Appendix 1. Cost Index Planning Trade-Off

Figure 37 shows the influence of flight time decreases on costs[23] (left) and trip fuel (right). For the six intercontinental flights considered in the shadow run, the flight dispatcher created several flight plans at different cost indexes. The same pattern was observed for all flights. Small percentual decreases in flight time can be achieved at little additional costs and trip fuel. For larger flight time decreases, however, the costs and trip fuel increase relatively faster. These graphs thus stress the importance of accurate departure delay predictions, as unnecessary time compensation may result in more than proportional increases in costs and trip fuel.

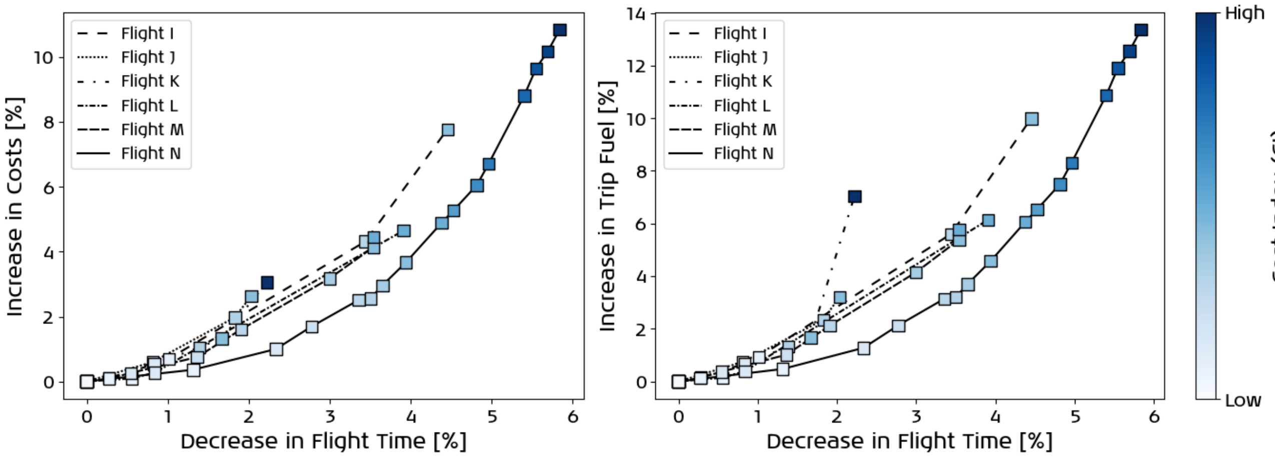


Figure 37. Influence of Flight Time Decrease on Costs (left) and Trip Fuel (right) for Shadow Run Flights

Appendix 2. Full List of Proposed Features

The full list of features considered in the Recursive Feature Elimination process is alphabetically presented below. The selected features are highlighted using bold letters. Feature 2 is one-hot encoded, with jet bridge and bus being the boarding options. Similarly, feature 31 is one-hot encoded, with 09, 18L, 18C, 27, 24, 36C, and 36L being the take-off runway options at Schiphol.

To include the flight-specific passenger connections, the number of connecting passengers with critical actual turn-around times was computed for every flight, resulting in features 18-22. Feature 35 weighs the number of passengers belonging to the bins with the number of minutes these passengers fall short. For the respective bins, the passengers on average fall short of 80, 60, 40, 20, and 0 minutes.

Features 28 and 29 attempt to capture the decision-making by airlines for delaying flights based on connection passengers. Since a large share of destinations is served multiple times a day, passengers with missed connections are often re-booked to a later flight. Towards the end of the day, when there are no more later flights, delays are more frequently initiated to accommodate passenger connections. For that reason, the ratio of remaining KLM flights was determined, with and without including codeshare flights. Taking a ratio rather than an absolute number of remaining flights allows to consider routes with different daily flight frequencies in the same model.

1. **Baggage Load Factor**
2. Boarding Option
3. **Current Average Delays Schiphol**
4. **Current Number of Flights Schiphol**
5. **Current TSAT Delay**
6. Day of Week
7. **Effective Delay Previous Flight**
8. Great Circle Distance
9. **Hour of Day**
10. **Last Aircraft Tail Swap**
11. Last Deicing Cancel Request
12. Last Deicing Request
13. Local Baggage Percentage
14. Local Passenger Percentage
15. **Median Delay of Flight Number**
16. Median Delay of Origin-Destination Pair
17. **Month of Year**
18. Nr. of Connection Passengers with -40 < Actual TAT < -20 min.
19. Nr. of Connection Passengers with -20 < Actual TAT < 0 min.
20. Nr. of Connection Passengers with 0 < Actual TAT < 20 min.
21. Nr. of Connection Passengers with 20 < Actual TAT < 40 min.
22. Nr. of Connection Passengers with 40 < Actual TAT < 60 min.
23. Number of Hand Luggage Pieces Collected at Gate
24. **Number of Passengers Reduced Mobility**

[23] This number does not include costs for potential missed connections.

25. Number of Transfer Passengers Arrival Airport
26. Number of Transfer Passengers Departure Airport
27. **Passenger Load Factor**
28. Remaining Flights Origin-Destination Pair KLM
29. Remaining Flights Origin-Destination Pair KLM + Codeshare
30. Scheduled Block Time
31. Take-Off Runway
32. **Total Passengers Day Flying Blue Schiphol**
33. Transfer Baggage Percentage
34. **Transfer Passenger Percentage**
35. Transfer Passenger Severity
36. Visibility
37. Wind Gust Speed
38. **Wind Speed Latitudinal Direction**
39. **Wind Speed Longitudinal Direction**

Appendix 3. Pseudo-Code of Random Forest, CatBoost, and Deep Neural Network Training

Training: Random Forest Model

```
split flight data into training and testing datasets
set hyperparameters
construct Random Forest model
fit model to training data
calculate feature importances
generate predictions
store model, predictions and feature importances
```

Training: CatBoost Model

```
split flight data into training and testing datasets
set hyperparameters
for $estimator \leftarrow 1$ to $nr\_estimators$ do
    construct CatBoost model
    fit model to training data
    calculate feature importances
    generate predictions
end
merge predictions
merge feature importances
store models, predictions and feature importances
```

Training: Deep Neural Network

```
split flight data into training and testing datasets
set hyperparameters
prepare $nr\_batches$ from $batch\_size$
$MSE\_best \leftarrow \infty$
$lr \leftarrow lr\_initial$
for $epoch \leftarrow 1$ to $nr\_epochs$ do
    train model
    for $batch \leftarrow 1$ to $nr\_batches$ do
        generate predictions
        compute model loss
        perform back-propagation
    end
    compute MSE
    if $MSE < MSE\_best$ then
        $MSE\_best \leftarrow MSE$
        $lr \leftarrow lr \cdot lr\_increase$
    else
        $lr \leftarrow lr \cdot lr\_decrease$
    end
    store model state
    if $lr < lr\_stop$ then
        break epochs loop
    end
end
load best model state
store model and predictions
```

Appendix 4. Dynamic Model Prediction Error and IQR as Multi-Level Sankey Diagrams

The evolution of the prediction error and prediction certainty (measured in terms of IQR) is illustrated in Figure 38 and Figure 39. As there exist millions of unique ways in which the prediction error and IQR may vary over the full prediction horizon, the errors and IQR values were binned to 4 minutes and 2 minutes respectively. It is worth pointing out that within the bins, the prediction error and IQR can also improve.

The main conclusion that can be drawn from Figure 38 and Figure 39 is that for a large share of flights, the prediction error decreases over time, whereas the prediction certainty increases over time. Between 90 and 15 minutes before scheduled departure, the number of flights with prediction errors between 0 and 8 minutes increases from 19994 to 22439, whereas the number of flights with prediction errors larger than 12 minutes decreases from 7063 to 5313. For the same prediction moments, the number of flights with IQRs between 0 and 4 minutes increases from 13801 to 15467, whereas the number of flights with IQRs larger than 6 minutes decreases from 8549 to 7571. Although a significant number of flights is predicted more accurately and with greater certainty, it should be pointed out that this is not the case for all flights.

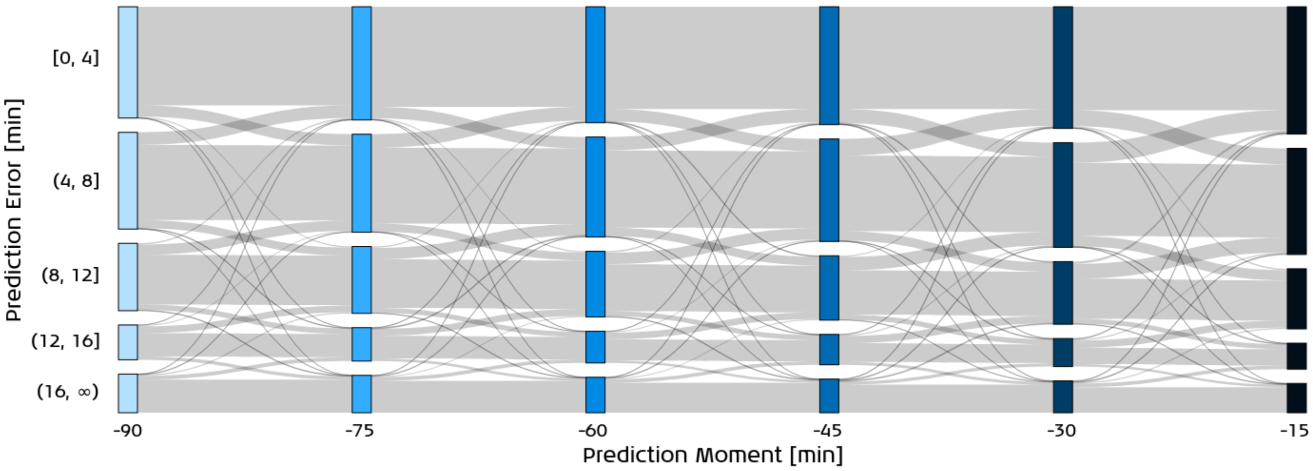


Figure 38. Prediction Error Propagation over Time for Random Forest model

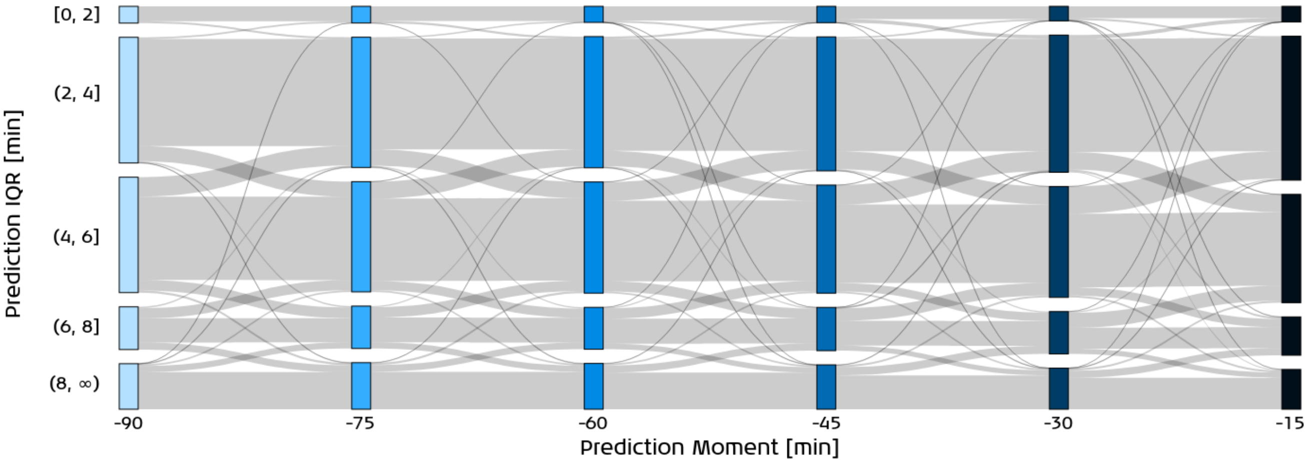


Figure 39. Prediction Inter-Quartile Range Propagation over Time for Random Forest model

Appendix 5. Model Performance when Removing Features

Table 11. Model Performance Changes for Feature Removal

Removed Feature	Δ*MAE*	Δ*RMSE*	Δ*R2*	Δ*ActInDistr*	Δ*AvgIQR*
Month of Year	+ 1.30 %	+ 2.27 %	- 3.64 %	- 0.59 %	- 4.16 %
Hour of Day	+ 0.95 %	+ 1.93 %	- 3.64 %	- 1.45 %	- 3.21 %
Passenger Load Factor	+ 2.72 %	+ 3.02 %	- 5.45 %	- 1.03 %	- 4.73 %
Baggage Load Factor	+ 0.83 %	+ 1.93 %	- 3.64 %	- 0.43 %	- 3.59 %
Transfer Passenger Percentage	+ 0.47 %	+ 1.60 %	- 1.82 %	- 0.07 %	- 2.27 %
Number of Passengers Reduced Mobility	+ 0.59 %	+ 1.76 %	- 1.82 %	- 0.21 %	- 1.89 %
Total Passengers Day Flying Blue Schiphol	+ 2.01 %	+ 2.52 %	- 3.64 %	+ 0.12 %	- 1.51 %
Median Delay of Flight Number	+ 1.42 %	+ 2.27 %	- 3.64 %	- 0.55 %	- 2.84 %
Effective Delay Previous Flight	+ 13.48 %	+ 13.85 %	- 23.64 %	+ 0.57 %	- 1.70 %
Current Number of Flights Schiphol	+ 0.47 %	+ 1.60 %	- 1.82 %	- 0.12 %	- 2.46 %
Current Average Delays Schiphol	+ 2.01 %	+ 3.53 %	- 5.45 %	- 2.00 %	- 8.13 %
Current TSAT Delay	+ 6.38 %	+ 6.97 %	- 10.91 %	- 0.68 %	- 7.18 %
Last Aircraft Tail Swap	+ 0.47 %	+ 1.51 %	- 1.82 %	+ 0.32 %	- 1.89 %
Wind Speed Longitudinal Direction	+ 0.47 %	+ 1.60 %	- 1.82 %	- 0.32 %	- 2.46 %
Wind Speed Latitudinal Direction	+ 0.47 %	+ 1.60 %	- 1.82 %	- 0.01 %	- 2.08 %

Apart from testing the model sensitivity by changing the baseline feature values, it was investigated how the model performance changes when removing the features altogether, one at a time. Compared to the Random Forest model with all selected features, the observed percentual performance differences at the 90-minute prediction horizon are tabulated in Table 11. For all removed features, the MAE and RMSE errors increase and the R2 decreases. As expected, the more important a feature was considered in Section 4.3.1, the higher the percentual changes in MAE, RMSE, and R2. The ActInDistr metric changes are relatively small. The removal of Hour of Day, Passenger Load Factor, and Current Average Delays Schiphol leads to the largest decrease in ActInDistr. Finally, for all removed features the confidence increases.

This makes sense as a higher number of features complicates the model and thereby decreases confidence. The removal of the current average delays Schiphol and current TSAT delay features particularly increases the confidence. This can be explained by the large fluctuations in these dynamic features that are observed for some flights.

Acknowledgement

I would like to express my gratitude to KLM Royal Dutch Airlines for offering me the opportunity to perform my internship and graduation thesis with them. A special thanks goes to Jasper de Wilde, Christiaan Evertse and Frans Huisman for their supervision, support and feedback. Additionally, I would like to thank Marta Ribeiro, Junzi Sun and Bruno Lopes dos Santos from TU Delft for their assistance, critical questions and feedback.

References

[1] Eric Mueller and Gano Chatterji. “Analysis of aircraft arrival and departure delay characteristics”. In: *AIAA’s Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*. 2002, p. 5866.

[2] Mohamed Abdel-Aty, Chris Lee, Yuqiong Bai, Xin Li, and Martin Michalak. “Detecting periodic patterns of arrival delay”. In: *Journal of Air Transport Management* 13.6 (2007), pp. 355–361.

[3] Yufeng Tu, Michael O Ball, and Wolfgang S Jank. “Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern”. In: *Journal of the American Statistical Association* 103.481 (2008), pp. 112–125.

[4] Jorge Vicente Pérez–Rodríguez, José María Pérez–Sánchez, and Emilio Gómez–Déniz. “Modelling the asymmetric probabilistic delay of aircraft arrival”. In: *Journal of Air Transport Management* 62 (2017), pp. 90–98.

[5] Anish M Kalliguddi and Aera K Leboulluec. “Predictive modeling of aircraft flight delay”. In: *Universal Journal of Management* 5.10 (2017), pp. 485–491.

[6] Waqar Ahmed Khan, Hoi-Lam Ma, Sai-Ho Chung, and Xin Wen. “Hierarchical integrated machine learning model for predicting flight departure delays and duration in series”. In: *Transportation Research Part C: Emerging Technologies* 129 (2021), p. 103225.

[7] Suvojit Manna, Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, and Subhas Barman. “A statistical approach to predict flight delay using gradient boosted decision tree”. In: *2017 International conference on computational intelligence in data science (ICCIDS)*. IEEE. 2017, pp. 1–5.

[8] Bojia Ye, Bo Liu, Yong Tian, and Lili Wan. “A methodology for predicting aggregate flight departure delays in airports based on supervised learning”. In: *Sustainability* 12.7 (2020), p. 2749.

[9] Ramon Dalmau, Franck Ballerini, Herbert Naessens, Seddik Belkoura, and Sebastian Wangnick. “An explainable machine learning approach to improve take-off time predictions”. In: *Journal of Air Transport Management* 95 (2021), p. 102090.

[10] Laurence Vorage. *Predicting Probabilistic Flight Delay for Individual Flights using Machine Learning Models.* MSc Thesis. 2021.

[11] Micha Zoutendijk and Mihaela Mitici. “Probabilistic flight delay predictions using machine learning and applications to the flight-to-gate assignment problem”. In: *Aerospace* 8.6 (2021), p. 152.

[12] Junzi Sun, Tristan Dijkstra, Constantinos Aristodemou, Vlad Buzetelu, Theo Falat, Tim Hogenelst, Niels Prins, and Benjamin Slijper. “Designing Recurrent and Graph Neural Networks to Predict Airport and Air Traffic Network Delays”. In: *10th International Conference for Research in Air Transportation*. FAA & Eurocontrol. 2022.

[13] Sebastian Birolini and Alexandre Jacquillat. “Day-ahead Aircraft Routing with Data-driven Primary Delay Predictions”. In: *European Journal of Operational Research* (2023).

[14] Alice Sternberg, Diego Carvalho, Leonardo Murta, Jorge Soares, and Eduardo Ogasawara. “An analysis of Brazilian flight delays based on frequent patterns”. In: *Transportation Research Part E: Logistics and Transportation Review* 95 (2016), pp. 282–298.

[15] Hugo Alonso and António Loureiro. “Predicting flight departure delay at Porto Airport: A preliminary study”. In: *2015 7th International Joint Conference on Computational Intelligence (IJCCI)*. Vol. 3. IEEE. 2015, pp. 93–98.

[16] Bin Yu, Zhen Guo, Sobhan Asian, Huaizhu Wang, and Gang Chen. "Flight delay prediction for commercial air transport: A deep learning approach". In: *Transportation Research Part E: Logistics and Transportation Review* 125 (2019), pp. 203–221.

[17] C Ciruelos, A Arranz, I Etxebarria, S Peces, B Campanelli, P Fleurquin, VM Eguiluz, and JJ Ramasco. "Modelling delay propagation trees for scheduled flights". In: *Proceedings of the 11th USA/EUROPE Air Traffic Management R&D Seminar, Lisbon, Portugal.* 2015, pp. 23–26.

[18] Athina Sismanidou, Joan Tarradellas, and Pere Suau-Sanchez. "The uneven geography of US air traffic delays: Quantifying the impact of connecting passengers on delay propagation". In: *Journal of Transport Geography* 98 (2022), p. 103260.

[19] Yuji Horiguchi, Yukino Baba, Hisashi Kashima, Masahito Suzuki, Hiroki Kayahara, and Jun Maeno. "Predicting fuel consumption and flight delays for low-cost airlines". In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 31. 2. 2017, pp. 4686–4693.

[20] Juan Jose Rebollo and Hamsa Balakrishnan. "Characterization and prediction of air traffic delays". In: *Transportation research part C: Emerging technologies* 44 (2014), pp. 231–241.

[21] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris. "Prediction of weather-induced airline delays based on machine learning algorithms". In: *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC).* IEEE. 2016, pp. 1–6.

[22] Karthik Gopalakrishnan and Hamsa Balakrishnan. "A comparative analysis of models for predicting delays in air traffic networks". In: *Twelfth USA/Europe air traffic management research and development seminar.* ATM Seminar. 2017.

[23] Delia Schösser and Jörn Schönberger. "On the Performance of Machine Learning Based Flight Delay Prediction–Investigating the Impact of Short-Term Features". In: *Promet-Traffic&Transportation* 34.6 (2022), pp. 825–838.

[24] Martin Felder, Anton Kaifel, and Alex Graves. "Wind power prediction using mixture density recurrent neural networks". In: *Poster presentation Gehalten auf der European wind energy conference.* 2010.

[25] KLM Royal Dutch Airlines. *KLM Royal Dutch Airlines Summer Schedule 2023.* https://news.klm.com/klm-royal-dutch-airlines-summer-schedule-2023/ (accessed on 04/10/2023). 2023.

[26] Reuters. *Amsterdam's Schiphol airport limits number of summer passengers.* https://www.reuters.com/business/aerospace-defense/amsterdams-schiphol-airport-limits-number-summer-passengers-2022-06-16/ (accessed on 04/10/2023). 2022.

[27] Iowa State University. *Iowa Environmental Mesonet.* Data retrieved from Iowa State University Environmental Mesonet, https://mesonet.agron.iastate.edu/request/download.phtml (accessed on 22/03/2023). 2023.

[28] EUROCONTROL. "All-Causes Delays to Air Transport in Europe Annual 2022". In: *CODA Digest* (2023).

[29] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[30] CatBoost. *CatBoost.* https://catboost.ai/ (accessed on 10/10/2023). 2023.

[31] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. "CatBoost: unbiased boosting with categorical features". In: *Advances in neural information processing systems* 31 (2018).

[32] Daniel Svozil, Vladimir Kvasnicka, and Jiri Pospichal. "Introduction to multi-layer feed-forward neural networks". In: *Chemometrics and intelligent laboratory systems* 39.1 (1997), pp. 43–62.

[33] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, and Vineeth Vijayaraghavan. "A machine learning approach for prediction of on-time performance of flights". In: *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC).* IEEE. 2017, pp. 1–6.

[34] Shan Lan, John-Paul Clarke, and Cynthia Barnhart. "Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions". In: *Transportation science* 40.1 (2006), pp. 15–28.

II

Literature Study
previously graded under AE4020

1 Introduction

After a few years in which the COVID-19 pandemic hit the aviation sector like no other, air traffic is growing rapidly again [22]. Although not back yet at 2019-levels, European airports and airlines have had great difficulty managing this rapid growth of demand in the previous year (2022), resulting in five-year-high flight delays. In the months of June, July and August, the percentage of delayed departures was around 65-70%, which is around 27 percentage points higher than pre-COVID levels [27].

Departure delays are also known as the delay at which the aircraft is pushed away from the gate. Due to the large costs associated with such delays [23], there are great incentives for airlines to better forecast and reduce these delays. For this reason, a thesis project will be performed in collaboration with KLM Royal Dutch Airlines, a Dutch airline operating from Amsterdam Airport Schiphol. Generating more accurate departure delay forecasts is desirable for airlines as it allows flight dispatchers to more effectively plan the aircraft speed at different phases of flight. As such, compared to current forecasting models, improved models may reduce emissions if flights can be planned at relatively lower speeds and more passenger connections can be made if flights have to be planned at relatively higher speeds.

The main goal of the upcoming thesis project is to improve academic state-of-the-art departure delay forecasting models using airline operational data, thereby enhancing the current departure delay forecasting model used by KLM. Due to the labelled nature of the data describing the airline departure delay, supervised learning solutions are especially applicable. For that reason, it was chosen to focus the attention on such computational models only, giving rise to the following main research question:

> *What combination of supervised learning models and features allows for the best dynamic departure delay probability distribution forecasts for individual flights and how does airline operational data affect the model performance?*

This *literature* study will contribute to the upcoming thesis work as it will allow to identify the state-of-the-art models referred to in the main research question. Furthermore, through conducting this literature study, the understanding of the departure delay forecasting field of research will improve. This literature study report aims to review existing work on departure delay forecasting, whilst providing background information about departure delay causes and costs, airport operations and forecasting methods.

This report is further structured as follows. First, Chapter 2 will provide background information on airline departure delays, discussing the departure delay causes, costs and the impact on flight planning. Then, Chapter 3 will elaborate on the influence that airports have on airline departure delays. Afterwards, Chapter 4 will provide an overview of supervised learning methods, applicable for solving the previously described problem. Thereafter, in Chapter 5, existing research will be reviewed in detail. Following from the reviewed existing research, the research gaps will be introduced in Chapter 6 and the research proposal will be outlined in Chapter 7. Finally, the report will be concluded in Chapter 8.

2

Airline Departure Delays

Flight delay can refer to multiple types of delay: departure delay, take-off delay, landing delay and arrival delay. These four types of delay are illustrated in Figure 2.1, where the delay is said to be the difference between the scheduled and actual timestamp. While most existing research considers departure and/or arrival delays, there are some research projects that focus on take-off and/or landing delay [23, 78].

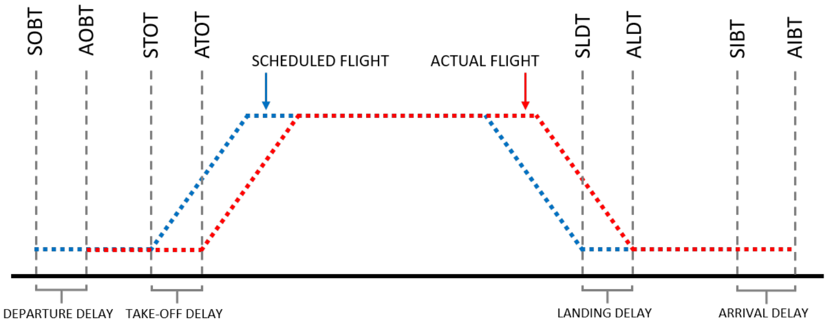


Figure 2.1: Delay Definitions given Scheduled and Actual Flight Times

Different approaches to defining departure delay and arrival delay exist. One could argue that the moment of closing the passenger door or retracting the passenger boarding bridge is the moment of departure. This thesis project, however, will use the start of towing as the moment of departure. This approach was also adopted in most previous research, for example in the work of Tu et al. [76], Sternberg et al. [72], Gui et al. [36] and Ye et al. [84]. As such, the departure delay (D_{dep}) is defined as the difference between actual off-blocks time (AOBT) and scheduled off-blocks time (SOBT) and the arrival delay (D_{arr}) is defined as the difference between actual in-blocks time (AIBT) and scheduled in-blocks time (SIBT), as visible in Figure 2.1 and Equation 2.1 and Equation 2.2. Following the work of Wesonga et al. [80], flights are more likely to be delayed at departure than at arrival. With cruise speed control, departure delays can be mitigated whilst en-route [4].

$$D_{dep} = AOBT - SOBT \tag{2.1}$$

$$D_{arr} = AIBT - SIBT \tag{2.2}$$

This chapter is further structured as follows. An overview of aircraft turnaround operations is presented in Section 2.1, followed by a historical overview of airline departure delays, presented in Section 2.2. Thereafter, in Section 2.3, the different causes for flight delays are elaborated upon and the associated costs are discussed in Section 2.4. Finally, the influence of departure delays on airline flight planning is discussed in Section 2.5.

2.1. Aircraft Turnaround Operations

The time between on-block and off-block of two consecutive flights is known as the turnaround time. During this period of time, a large number of processes are executed, which will be elaborated upon in Subsection 2.1.1. Depending on weather conditions, aircraft are required to be de-iced. Such procedures are described in Subsection 2.1.2. Finally, an increasing number of airports, including KLM's hub Amsterdam Airport Schiphol, is adopting an Airport Collaborative Decision Making (A-CDM)-approach, which will be further explained in Subsection 2.1.3.

2.1.1. Step-by-Step Turnaround Processes

The aircraft turnaround process can be split into landside and airside ground operations [48]. Landside operations entail everything that happens up until security/passport control and thus includes the passenger and baggage check-in, security/passport control and the cargo handling. Whilst these three steps could delay a flight, there are more steps to be taken in the airside operations, thus more possible delay causes arise from the airside operations.

The airside operations can be split into five parallel processes, as indicated in Figure 2.2, elaborated upon one by one with the numbers between brackets referring to the numbers in the figure. Stream 1 depicts the cabin-related processes: disembarking passengers (1), cleaning the cabin (2) and embarking new passengers (3). Stream 2 depicts the baggage-related processes: unloading baggage (1) and loading new baggage (2). Stream 3 depicts the crew processes: crew change (1) and flight planning (2). Stream 4 depicts the catering-related processes: replacing used water with fresh water (1) and supplying new food and drinks (2). Finally, stream 5 depicts the refuelling processes: connecting the Ground Power Unit (GPU) (1) and refuelling the aircraft (2).

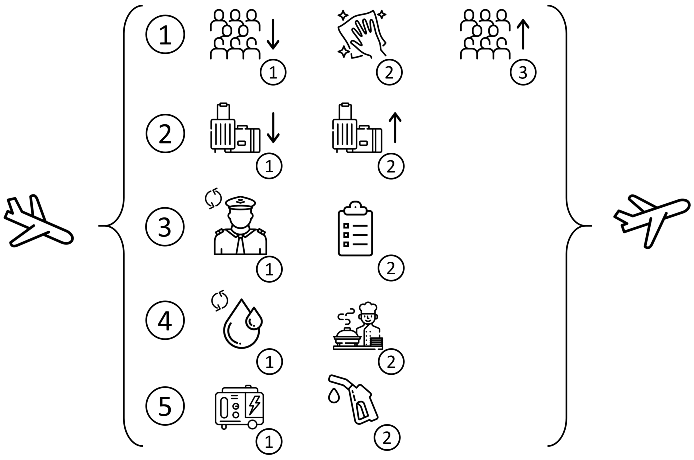


Figure 2.2: Schematic Representation of Aircraft Turnaround Processes

Additionally, it should be noted that maintenance is also part of the airside operations, although it does not have to be performed between every two flight legs. Finally, Khan et al. [48] mentioned that the turnaround processes differ significantly between domestic and international flights, with the international flights ending up taking more time than domestic flights. This time difference is also true for European and intercontinental flights at KLM, where more turnaround time is allocated between intercontinental flights [32].

2.1.2. De-Icing Processes

When temperatures are close to or below zero degrees Celcius, thin layers of ice can form that degrade the lift performance of aircraft and/or moisturize the engine intakes [69]. During de-icing, a de-icing fluid removes these layers of frost. For many airports, including Amsterdam Airport Schiphol, de-icing is only required a few days every year. The de-icing process itself and possible queuing at the de-icing station both contribute to increases in turnaround time, leading to higher departure delays [59].

2.1.3. Airport Collaborative Decision Making (A-CDM)

EUROCONTROL, the European organisation for safety of air navigation, launched the concept of Airport Collaboartive Decision Making (A-CDM, *or CDM for short*) in the 2010s, with the aim to "improve the efficiency and resilience of airport operations by optimising the use of resources and improving the predictability of air traffic" [28]. Therefore, aircraft operators, airport operators, ground handling parties and air traffic control are stimulated to share relevant information to enhance the turnaround process or aircraft. KLM is an active user and contributor to the A-CDM system at Amsterdam Airport Schiphol. Through this system, airlines like KLM are able to plan their operations given the information provided by other parties. The data, such as Air Traffic Control (ATC) approval times, is gathered and may be used in the upcoming thesis work.

2.2. Historical Overview of Flight Delays

Both in early and recent papers, it is indicated that around 20% of all flights arrive beyond schedule [57, 73]. Furthermore, from data between 2013 and 2022 from Bureau of Transportation Statistics [11], similar results were obtained and are visualized in Figure 2.3. Clearly, the average arrival delay percentage fluctuates around 20% for these 10 years, with an exception for the 2020 and 2021 years. Most likely, the lower delays in these two years can be attributed to the lower traffic demands with the COVID-19 pandemic hitting hardest. This gives rise to the hypothesis that in the other years, the aviation industry could not adequately keep up with high demand, leading to overcrowded airports and high departure and arrival delays. Moreover, the total number of delay minutes in 2022 for US domestic flights was 100 million minutes.

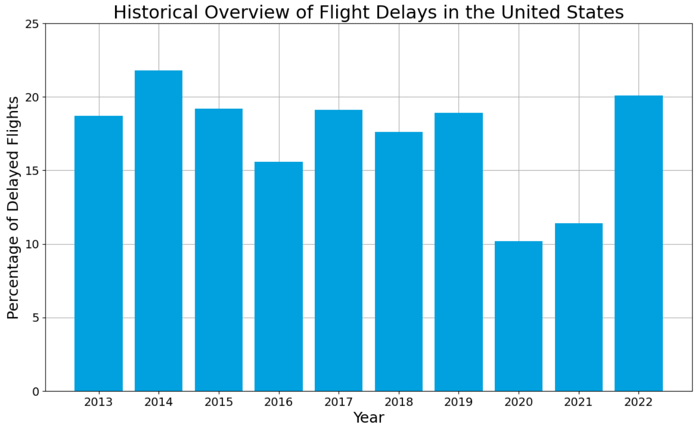


Figure 2.3: Percentage of Flights with Arrival Delays in the United States, Data obtained through Bureau of Transportation Statistics [11]

Furthermore, the majority of papers has indicated that the distribution of flight delays is highly positively skewed [48, 62, 64, 76]. This was confirmed in Figure 2.4, where using 2022 EUROCONTROL annual results, the departure and arrival delay distributions were plotted for 2022. The Bureau of Transportation Statistics considers a flight to be delayed if it arrives or departs 15 minutes or more behind schedule [12]. This threshold has been indicated with red dotted line in Figure 2.4. Since the vast majority of flights has departure delays between -5 and 15 minutes, only a minority (31.4% of flights in Figure 2.4) would actually be considered delayed. 31.4% may seem high compared to Figure 2.3, but in 2022, departure delays were much higher in Europe than in the United States [27]. Nonetheless, the skewed nature of the data remains and might cause difficulties for training supervised learning algorithms. This will be covered in more detail in Chapter 4 and Chapter 5.

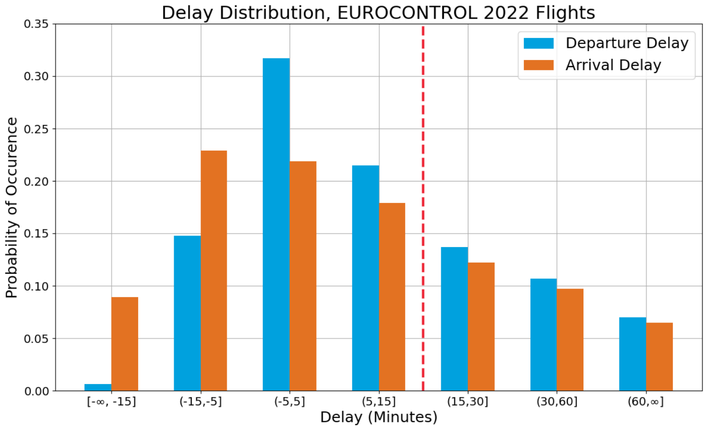


Figure 2.4: Departure and Arrival Delay Distributions for Flights through EUROCONTROL Area in 2022, Data obtained through EUROCONTROL [27]

2.3. Departure Delay Causes

There is a large number of causes that could possibly cause flights to be delayed. Fortunately, the International Air Transport Association (IATA) has established universal delay codes, that allocates the delays to eleven categories: one internal category (codes 00-09) and ten external categories (codes 10-99) [25, 48]. The ten external categories of delay codes are listed below. EUROCONTROL reported that the most commonly used delay codes are those referring to reactionary delays (codes 90-96) [24].

- 10-19: Passenger and Baggage
- 20-29: Cargo and Mail
- 30-39: Aircraft and Ramp Handling
- 40-49: Technical and Aircraft Equipment
- 50-59: Damage to Aircraft & Automated Equipment Failure
- 60-69: Flight Operations and Crewing
- 70-79: Weather
- 80-89: Air Traffic Flow Management and Governmental Authorities
- 90-96: Reactionary
- 97-99: Miscellaneous

According to Abdel-Aty et al. [1], the fundamental reason for flight delay is the lack of supply being able to match demand of air travel. Mueller and Chatterji [57] listed a large number of delay causes that can be summarized with the list above. He finds that 50% of all delays originate before the aircraft is pushed back from the gate and another 26% of delays originate between push-back and take-off. Furthermore, the author argues that hub operations attract delays due to the higher concentrations of flights around peak hours. For the sake of clarity, a few of the listed delay categories are explained in more detail: Subsection 2.3.1 will elaborate on technical delays (codes 40-49), Subsection 2.3.2 will elaborate on weather delays (codes 70-79), Subsection 2.3.3 will elaborate on air traffic control delays (codes 80-89) and finally Subsection 2.3.4 will elaborate on reactionary delays (codes 90-96).

2.3.1. Technical Delays

Technical delays are caused by malfunctioning or damaged aircraft. Whilst aircraft are scheduled for A-, B-, C- and D-checks after a fixed number of flights or flight hours, unplanned maintenance might be required from time to time. The number of flights that are delayed due to technical delays is very small, however. Khan et al. [48] found that only 0.01% of the flights in his data set were delayed for technical reasons.

Because of the low number of technical delays and their unpredictability, the current departure delay forecast model that KLM uses regards flights with maintenance delays as outliers. Despite the small chance of such delays to happen, the associated delay minutes are often high, therefore flights with technical delays would have quite large influences on the forecasting result had they not been excluded from the dataset.

2.3.2. Weather Delays

The influence of weather on departure delays has been studied thoroughly in existing research. Two potential reasons for this observation could be the large number of meteorological datasets that are publicly available [44] and the fact that weather is statistically highly correlated with departure delays [46, 57]. Choi et al. [16] noted that adverse weather conditions are a dominant cause for flights to be delayed, but that they are not always reported to be the cause of the delay because they are very closely related to other delays such as air traffic control delays and reactionary delays.

Two weather components that largely influence departure delay and flight delay in general are visibility and wind speed [19, 72]. Sternberg et al. [72] found that fog (visibility-related) is found to increase delay chances by 148%. Furthermore, thunderstorms and rain are found to increase delay chances by 67%. Intuitively, it makes sense that for such adverse weather conditions, delays tend to be higher.

2.3.3. Air Traffic Control Delays

Air travel demand has grown rapidly over the past decades. Airports have difficulties keeping up with this growth rate, as airport infrastructure upgrades require time, money and political considerations [2]. The number of ATC delays has increased as a result [47]. According to the annual delay report by EUROCONTROL [27], in 2022 the en-route Air Traffic Flow Management (ATFM) delay was 1.8 minutes per flight, significantly higher than previous years. To airlines, air traffic control delays are important as potentially, slots can be missed if these delays are large.

Managing the traffic flow becomes more complex, both on ground (due to limiting airport infrastructure) and en-route (due to sector capacity constraints). A well-known European bottleneck is Karlsruhe Upper Area Control (UAC), as it is located in the centre of the continent. The closure of Ukrainian and Russian airspace further increases traffic in neighbouring airspaces, but also in Western European airspaces [27].

2.3.4. Reactionary Delays

Reactionary delays are delays for a current flight that originate from earlier flight(s). Already in early research, reactionary delays were studied [76, 82]. According to several researchers, reactionary delays are one of the largest delay factors [1, 17, 62, 79, 85]. This observation is supported by Cook et al. [18] who found that reactionary delays due to late aircraft accounted for 42% of all delayed flights in Europe in 2013. Furthermore, Cook et al. [18] differentiated rotational reactionary delays (late arrival of aircraft) and non-rotational reactionary delays (late arrival of airline crew and/or connecting passengers). These different reactionary delays will be elaborated upon in the following sections.

Aircraft Rotations

Aircraft are often scheduled to fly multiple flight legs per day. As for any other mode of (public) transport, if an incoming flight arrives late, it is likely that the outgoing flight will depart late, depending on the time that is scheduled between the flights. This can be explained following the approach of Birolini and Jacquillat [7] in Figure 2.5, where flight delay is seen as the sum of primary delays (delay associated to current flight) and propagated delays (delay associated with previous flight(s)), see Equation 2.3. For flight f in rotation r, a slack time (τ_{rf}) can be introduced to a flight schedule to mitigate delays, and is defined as the difference between the scheduled turnaround time (cr_{rf}) and the minimum required turnaround time (ρ_{rf}), following Equation 2.4. Since the delay of the previous flight ($\delta_{\sigma r(f)}$) is known, it is possible to determine the part of the total delay that is associated with previous flight(s) using Equation 2.5. In case the slack time is larger than the previous flight delay, logically there is no propagated flight delay. Using these equations, it is possible to quantify the primary and propagated delay per flight. Stand-by aircraft and more slack time are two solutions to mitigate departure delays due to aircraft rotations [7, 18]. Alternatively, Ciruelos et al. [17] posed the option of swapping departure slots within a single alliance.

$$\delta_{rf} = \delta_{rf}^{primary} + \delta_{rf}^{propagated} \tag{2.3}$$

$$\tau_{rf} = ct_{rf} - \rho_{rf} \tag{2.4}$$

$$\delta_{rf}^{propagated} = max\left(0, \quad \delta_{\sigma_r(f)} - \tau_{rf}\right) \tag{2.5}$$

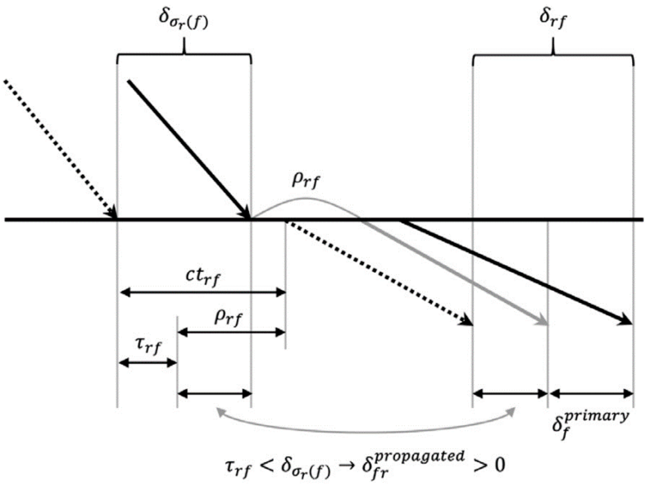


Figure 2.5: Primary and Propagated Delay Distinction Technique (Visualization by Birolini and Jacquillat [7])

Crew Connectivity

Similar to late arrival of aircraft, late arrival of crew may impose flight delays. Especially with airlines that operate many short-haul flights, it is common that cockpit and cabin crew fly more than one flight leg per day. For airlines that operate from multiple hubs, the crew scheduling becomes an even larger and more complex task because the different hubs add a new variable to the problem. Stand-by crews and larger buffer times in crew itineraries are proposed to mitigate departure delay due to crew connectivity [17, 18].

At KLM Royal Dutch Airlines, to facilitate connecting passengers through Amsterdam Airport Schiphol in the morning peak hours, many European flights stay for the night at destination airports (*further referred to as outstations*) to allow for early flights to Amsterdam. The crew onboard of these flights would only work for a maximum of 3 hours on that European flight, which therefore allows them to work on at least one more European flight. As such, there is again crew available at an outstation for the early flight towards Amsterdam the next day. This does mean, however, that there are many crews that have connecting flights at Amsterdam every day.

Passenger Connectivity

Little research has yet been performed on the influence of passenger connectivity on departure delays, most likely because for most researchers this data is not available. It is known that sometimes airlines wait for connecting passengers, accepting the associated departure delay as a result. Ciruelos et al. [17] proposed a simulation model where connecting passengers numbers are assumed. For a simulation with around 20000 Lufthansa flights, if the airline were not to wait for connecting passengers, the total passenger delay minutes would almost double for that day.

Bratu and Barnhart [9] analyzed passenger delays using passenger booking data and finds that delay minutes of passengers are higher than that of airlines, as delay minutes due to missed connections do count towards passenger delay minutes but not towards airline delay minutes. Furthermore, the paper identifies that connecting passengers are 2.8 times more likely to be disrupted than local passengers. Even though this study is relatively old, it does demonstrate the relation between connecting passengers and the delays they experience. To mitigate departure delays due to passenger connectivity, larger buffer times in passenger itineraries are proposed [17, 18].

Fortunately, for the thesis project, this data will be available due to the partnership with KLM Royal Dutch Airlines. Minimum transfer times are determined that indicate whether tickets can be sold for a certain inbound-outbound connection. For Amsterdam Airport Schiphol, the minimum connection time is 40 minutes for Schengen flights and 50 minutes for non-Schengen flights [50]. However, these thresholds are based on scheduled flight times, not on actual flight times. In practice, it thus happens that passengers have shorter connection times and might not make their connection.

2.4. Departure Delay Costs

Departure delay costs have been extensively researched in existing literature, from scientific, governmental and industrial points of view. In the United States alone, the total cost of delay in 2019 was 33 billion US Dollars, 55% of which is classified as passenger costs (resulting from delays, cancellations and missed connections) [30]. The second largest cost (8.3 billion US Dollars, *or 25% of total costs*) is related to increased expenditure on fuel, crew, rescheduling and maintenance [30]. It should be noted that not only unforeseen delays but also schedule buffers are included in this cost calculation. Thus, time that has been scheduled to mitigate potential delays is accounted for in these delay cost figures.

Although there are initiatives to improve the current Air Traffic Management (ATM) processes to reduce delays, such as the planning of capacity and demand [45], delay costs have increased significantly in recent years and are predicted to do so by another 25% in 2039 [23, 30]. In Europe, since 2004 a law named EC261 is in place that requires airlines to compensate passengers if flight delays exceed 3 hours. The compensations are between 250 and 600, depending on the flight distance [29]. For disruption management, these compensation costs are often taken into account as they significantly contribute to total disruption costs [38].

Finally, to some extent, flight delays worsen airline reputations. Customers will be dissatisfied and might choose another carrier for their next journeys, which is referred to as loss of future value [75]. This will not only decrease the future income of the airline itself, but will also increase the future income of competitors.

2.5. Departure Delay Influence on Flight Planning

Apart from the associated costs discussed in Section 2.4, departure delays also affect flight planning. The departure delay effects on the long-term schedule planning are discussed in Subsection 2.5.1 and the departure delay effects on the short-term flight planning are discussed in Subsection 2.5.2.

2.5.1. Departure Delay Influence on Schedule Planning

Flight schedules are set up with tight margins, thereby maximizing aircraft utilization. This statement is backed by Sternberg et al. [72], who found that 89% of flights that are delayed at departure are still delayed at arrival. In the research of Dalmau et al. [20], the aim was to early detect whether an aircraft would be able to reach its final destination of the day before a night curfew (nightly airport closure to minimize noise nuisance), taking into account the uncertainties for all flight legs yet to be completed. 15 hours before the curfew, the model was already able to predict with 94% accuracy if the curfew would be infringed. To some extent, this research assesses the robustness of flight schedules and how much encountered delays can be made up for. Lambelho et al. [52] assessed strategic flight schedules with respect to flight delays and suggests to integrate delays and cancellations into the flight schedule.

For the development of flight schedules, block times (time between SOBT and SIBT) are critical since they determine the allocated time per flight. While airlines adopt different strategies for determining these block times, it is common to have some reserve to anticipate potential delays [67].

Although block times are corrected for departure delay, the minimum turnaround time of an aircraft is not. This minimum turnaround time of the aircraft is often fixed per airport and per aircraft type. One method to mitigate delays is to make use of slack times between flights [7]. While these slack times allow for a longer turnaround time, and thus the possibility to overcome (parts of) the delay, they also reduce aircraft utilization and revenue.

2.5.2. Departure Delay Influence on Flight Plan Creation

Next to the departure delay effects on the long-term schedule planning, also the day-of-operation flight planning is affected by departure delays. In many airlines, the day-of-operation flight planning is the responsibility of flight dispatchers. With information about weather, traffic flow, aircraft availability, ground operations and more, these dispatchers determine the flight path and the cost index at different phases of flight.

Flight departure delays significantly influence the normal operations and require the flight dispatcher to make changes to the flight plan to allow for on-time arrival. Khan et al. [48] stated that forecasting the departure delay at the moment of creating the flight plan allows the airline to establish well-grounded measures to overcome the delays, such as flying a different route or flying at higher cost index. Better departure delay forecasts will allow for more accurate cost index planning, potentially reducing fuel emissions if the improved departure delay forecast allows to plan flights at lower cost indexes. Alternatively, if a flight has to be planned at higher cost indexes, more passenger connections will be made.

3 Airport Operations

Since departure delay is defined to be the delay at which the aircraft is pushed away from the airport gate, airport characteristics directly influence airline departure delay. This chapter aims to explain how airports and airline departure delays are related to each other. First, in Section 3.1, some of the most important aspects covering the airport infrastructure are elaborated upon. Then, Section 3.2 aims to introduce the reader to the airports in the KLM network, stressing the dissimilarities between airports.

3.1. Airport Infrastructure

Countless different airport designs exist over all parts of the world. However, there are some aspects that are similar for all airports, as they were universally agreed upon. Some of the governing aspects of airport infrastructure are the security and passport control, further elaborated upon in Subsection 3.1.1, as well as the gate architecture and runway and taxiway architecture, presented in more detail in Subsection 3.1.2 and Subsection 3.1.3, respectively.

3.1.1. Security and Passport Control

For local passengers (passengers who start their journey from the respective airport), apart from baggage check-in, often the security and passport control are most time intensive. The queue sizes depend not only on the hour of day (as there are different departure peak hours, also known as departure banks), but also on the number of airport staff available. In 2022, Amsterdam Airport Schiphol, amongst other European airports, encountered great difficulties recruiting enough staff to keep up with increasing passenger flows, leading to the implementation of passenger caps [56].

Security control is required for every flight, whether it is a domestic flight in Chile or a intercontinental flight between Europe and Asia. Passengers are also often required to pass through security control when transferring at a connection airport. Passport control is only required for international flights. Within Europe, the Schengen Zone was created that allows passengers to freely move from one Schengen country to another [37].

3.1.2. Gate Architecture

A number of different airport gate architectures exist and can be found all over the world. Figure 3.1 shows five common airport layouts that were presented by The Geography of Transport Systems [74]. The different layouts will be elaborated upon one by one:

- *Terminal a:* Standard Layout: This is the most common design, especially suitable for small-size airports, and can be easily extended to pier layouts or concourses. The main drawback is that for larger airports with a standard layout, the walking time between gates can be large. Examples of airports that have adopted a standard layout are: Lisbon (LIS), Nice (NCE), Buenos Aires (EZE).
- *Terminal b:* Pier Layout: This design is an extension of a standard layout, where piers are often used for different categories of flights (e.g. international and domestic piers or piers for a single airline). Examples of airports that have adopted a pier layout are: Amsterdam (AMS), Glasgow (GLA), New York (JFK).

- *Terminal c:* Satellite Layout: This layout with underground links to other satellites or the main terminal can be an efficient solution to airports with little available space. Examples of airports that have adopted a satellite layout are: Seattle (SEA), Orlando (MCO), Kuala Lumpur (KUL).
- *Terminal d:* Concourses Layout: This design is an extension of a standard layout, where different concourses are linked underground. Examples of airports that have adopted a concourses layout are: Madrid (MAD), London (LHR), Atlanta (ATL).
- *Terminal e:* Shuttles Layout: This layout, where passengers walk to their aircraft or where shuttle busses are used, allows airports to reduce the terminal size, at the cost of longer boarding times. Examples of airports that have adopted a shuttles layout are: Eindhoven (EIN), Tirana (TIA), Malta (MLA).

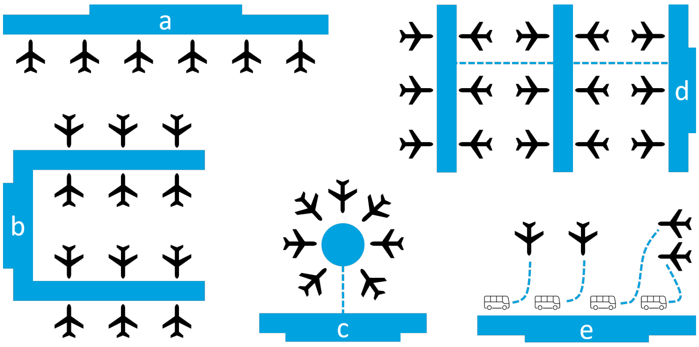


Figure 3.1: Different Gate Layouts: *a:* Standard, *b:* Pier, *c:* Satellite, *d:* Concourses, *e:* Shuttles [74]

3.1.3. Runway and Taxiway Architecture

Wind is known to have great influence on the performance of flights and storms may cause large disruptions on airports. For that reason, for airport design, the direction of the runways is often determined based on the average wind direction, thereby avoiding cross-winds. Therefore, in the Netherlands, almost all airports have runways running in the South-Western to North-Eastern direction.

Another aspect that is frequently taken into account for airport design is noise abatement. Governmental instances often require (new) airports to be built such the population living below the arriving and departing routes is minimized. Finally, for both runway and taxiway design, redundancy is important as maintenance needs to be performed once in a while. In such cases, it is desired that the total capacity is only affected to some extent.

3.1.4. Capacity Levels

For airports, the maximum number of aircraft or passengers that can be served in some period of time is referred to as the airport capacity. For many airports, there is more demand for flights than the airport can offer. For such congested airports, Ye et al. [84] found that "once some delay happened in airports, the delays can hardly be released in short term; delays in airports have some time correlation". This holds especially for airports that run at their maximum capacity, resulting in increasing delays over the span of the day.

For congested airports, the capacity is managed using landing slots, which allow airlines to land at some airport in a certain time-frame (often 20 minutes). Airlines can plan their schedules but have to be rewarded landing slots at congested airports. Such slots are of high importance for the marketing position of an airline with respect to its competitors. These slots are distributed by airport slot coordinators who have to follow worldwide slot guidelines, guaranteeing fair competition. The process of slot allocation is performed twice a year, for the summer and winter seasons. The congestion is regulated by applying a maximum to the number of slots. In order to increase the number of available slots at a time of day, additional infrastructure such as terminals and runways are the only possible solution. Finally, Tu et al. [76] was able to observe several steep spikes in departure delays at peak hours. During these moments of day, aircraft formed queues before lining up at the runway, showcasing that the airport in question was operating close to its maximum capacity.

3.2. Airports in KLM Network

Being the oldest airline still operating under its original name, KLM is one of Europe's largest flag carriers. At the moment of writing, KLM has 170 aircraft in fleet, 62 of which are used by KLM Cityhopper (a daughter company of KLM) and 3 of which are used by KLM Cargo [32]. Using a hub-and-spoke network, KLM focuses on transferring passengers through its hub airport in Amsterdam. For that reason, punctuality is of utmost importance and there is an incentive for considering connection passengers in departure delay forecasts.

KLM operates from over 150 airports worldwide, each of which has its own characteristics. Sternberg et al. [72] analyzed the punctuality of four airlines (TAM, Gol, Avianca and Azul) at different Brazilian airports and concluded that "airlines performances differ from airport to airport" and that "all airlines are punctual at some airports and unpunctial at other ones". The remainder of this section first focuses on KLM's hub airport Amsterdam Airport Schiphol in Subsection 3.2.1, followed by the KLM destination airports in Subsection 3.2.2.

3.2.1. Amsterdam Airport Schiphol

Amsterdam Airport Schiphol is the largest airport in the Netherlands and welcomed 70.7 million passengers in 2019 [65]. KLM has been operating from Schiphol since its foundation and departed from Amsterdam Airport Schiphol over 100,000 times in 2019. In total, 496,826 flight movements were registered in 2019. Figure 3.2 shows a detailed map of Amsterdam Airport Schiphol, including its 6 runways and the pier-like gate infrastructure in the centre of the map.

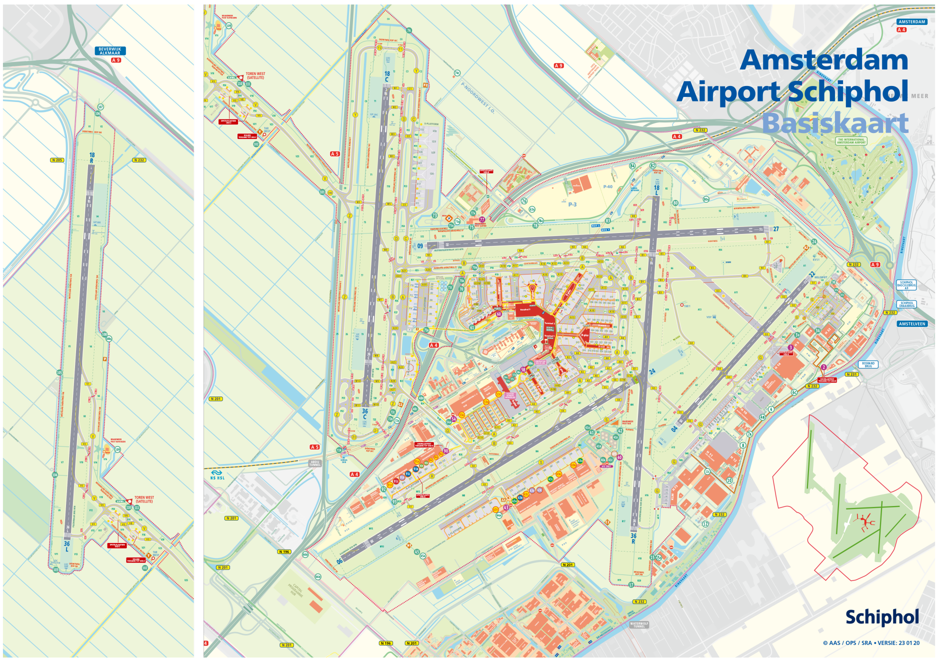


Figure 3.2: Detailed Map of Amsterdam Airport Schiphol. Image Courtesy of Royal Schiphol Group

Regular operations at Amsterdam Airport Schiphol require a 2-1 combination of runways: during departure banks, 2 runways are used for departures and 1 runway is used for arrivals and during arrival banks, 2 runways are used for arrivals and 1 runway is used for departures.

3.2.2. KLM Destination Airports

KLM connects Amsterdam to over 150 destinations around the world, which are shown in Figure 3.3, for intercontinental and European destinations separately. It can be seen that not all flights depart or arrive in Amsterdam. Whereas within the European network, only return flight operations exist at the moment of writing, outside of the European network, there also exist circle flight operations (such as Amsterdam (AMS) - San Jose (SJC) - Liberia (LIR) - Amsterdam (AMS)) and tail-end operations (such as Amsterdam (AMS) - Singapore (SIN) - Denpasar (DPS) - Singapore (SIN) - Amsterdam (AMS)).

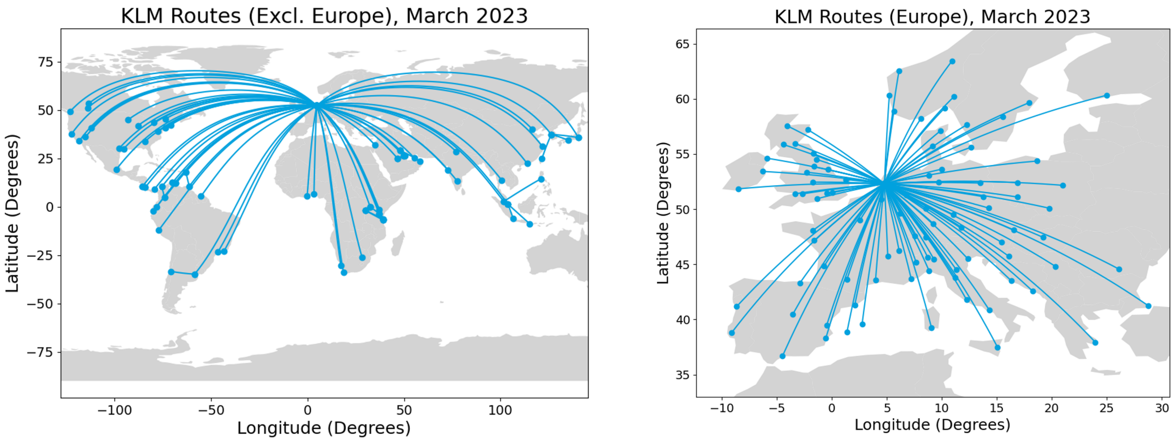


Figure 3.3: KLM Destination Airports: Intercontinental (left) and Europe (right) [32]

The European flights are partially executed by KLM and partially executed by KLM Cityhopper. The distinction lies in the type of aircraft planned in the schedule, as KLM Cityhopper only operates the flights with Embraer aircraft. Intercontinental flights are always operated by KLM and are mainly concentrated towards the Americas, Middle East and Far East. The network of KLM is complementary to it's partner airline Air France, allowing to collectively generate the largest economically feasible network.

As was earlier found by Sternberg et al. [72], the on-time performance of airlines differs per airport. This holds for KLM and makes sense because logically the airport operations differ from Jakarta to Paris and from New York to Entebbe. Furthermore, Wu [81] found that the planning of airline operations for international flights is different than that of domestic flights. KLM does not serve any national flights due to the small size of their hub country, but on average, intercontinental flights have larger departure delays than European flights, confirming the analysis of Wu [81] that longer flights accumulate larger departure delays.

4

Supervised Learning for Forecasting Applications

Given the large amounts of data that are generated in the aviation industry, supervised learning is a suitable method for forecasting applications [66]. This chapter aims to present the main steps required for applying supervised learning methods. First, in Section 4.1, the data pre-processing process will be elaborated upon. Then, Section 4.2 will introduce the reader to a number of supervised learning models, categorized into classification models and regression models. Then, Section 4.3 will explain the process of feature selection for the models to be trained on, followed by Section 4.4 where the process of hyper-parameter tuning will be explained. Finally, common performance metrics will be discussed in Section 4.5 and the effect of the forecasting horizon will be explained in Section 4.6.

4.1. Data Pre-Processing

Data processing covers the handling of outliers and missing values [8]. Even though data pre-processing is the first step, it is an important step as the performance of the model may vary significantly with different datasets. It may not always be necessary to include all data points as unusual values due to special circumstances can be disadvantageous for any type of computational method that is based on historical patterns [8]. The removal of such unusual values is referred to as outlier elimination.

Outlier elimination is commonly used in existing research. In the work of Sun et al. [73], all flights delays over 90 minutes or smaller than -30 minutes were removed from the dataset. These flights make up approximately 0.7% of all flights. Manna et al. [55] applied a more complex procedure for outlier elimination. Given the Inter-Quartile Range (IQR) between the 25th percentile (Q1) and the 75th percentile (Q3), all datapoints that are outside of the [Q1-1.5*IQR, Q3+1.5*IQR] interval were removed from the dataset.

Furthermore, the distribution of departure delays in the empirical data is often highly skewed, as was shown in Figure 2.4. For that reason, some researchers perform sampling-practices to generate a balanced input dataset. As such, it is aimed to have approximately an equal distribution between on-time flights and delayed flights and the model would not be biased towards either of the two. As shown in Figure 4.1, two methods exist to balance the dataset: undersampling (limiting the majority dataset size to the minority dataset size by taking only a selection of samples) and oversampling (generating samples to increase size of minority dataset to the majority dataset size).

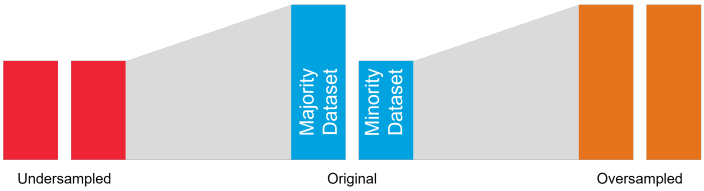


Figure 4.1: Graphical Representation of Undersampling and Oversampling Techniques

It was found that especially in classification studies, sampling was performed. In the studies by Thiagarajan et al. [75] and Khan et al. [48], a combination of oversampling (Synthetic Minority Oversampling Technique (SMOTE)) and undersampling (Tomek Links) was applied to generate the sampled dataset, as SMOTE alone was unable to capture the skewed nature of the airline delay distribution. Finally, with machine learning it is common to divide the dataset into a train dataset and a test dataset, such that the model can be tested with unseen data. Frequently used ratios are 75:25 and 80:20, where the majority of data is devoted to training [66, 73].

4.2. Model Selection

Forecasting departure delays is a typical supervised learning problem, as for the training data, all data points (flights) have an associated label (departure delay). Following from Géron [33], supervised learning problems can be subdivided into classification problems (predicting classes) and regression problems (predicting values). An overview of the supervised learning problems is presented in Table 4.1 and the different models will be elaborated upon in Subsection 4.2.1 until Subsection 4.2.9, after which in Subsection 4.2.10, the relation between model performance and explainability will be discussed.

Table 4.1: Supervised Learning Models

Supervised Learning Model	Applicable for Classification	Applicable for Regression
Linear Regression		✓
Logistic Regression	✓	✓
Support Vector Machine	✓	✓
k-Nearest Neighbor	✓	✓
Naive Bayes	✓	
Decision Tree	✓	✓
Random Forest	✓	✓
Boosting Algorithm	✓	✓
Neural Network	✓	✓

4.2.1. Linear Regression

Linear regression models are simple linear functions that predict output $\hat{y}$ by computing the weighted sum of model parameters (vector $\boldsymbol{\theta}$) and input features (vector $\boldsymbol{x}$) plus a constant bias term (scalar ϵ), which is the prediction when all input features are zero [33]. The governing equation of this method is presented in Equation 4.1.

$$\hat{y} = \boldsymbol{\theta} \cdot \boldsymbol{x} + \epsilon \tag{4.1}$$

4.2.2. Logistic Regression

Logistic regression differs slightly from linear regression in the sense that a probability is predicted using Equation 4.2, using a sigmoid function with the model parameters (vector $\boldsymbol{\theta}$) and input features (vector $\boldsymbol{x}$) as inputs [33]. Given probability $\hat{p}$, output $\hat{y}$ can be predicted according to Equation 4.3.

$$\hat{p} = \sigma(\boldsymbol{x}^T\boldsymbol{\theta}) = \frac{1}{1+e^{-\boldsymbol{x}^T\boldsymbol{\theta}}} \tag{4.2}$$

$$\hat{y} = \begin{cases} 0 & \hat{p} < 0.5 \\ 1 & \hat{p} \geq 0.5 \end{cases} \tag{4.3}$$

4.2.3. Support Vector Machine

Suitable for both classification and regression problems, Support Vector Machine (SVM) aims to find optimal hyper-planes that separate datapoints of different categories [33], as visualized in Figure 4.2. The SVM method is easy to understand and implement, but might not be applicable for high-dimensional problems.

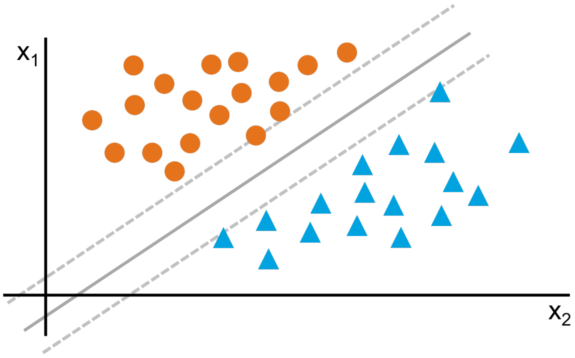


Figure 4.2: Graphical Representation of Underlying Principle of Support Vector Machine Models

4.2.4. k-Nearest Neighbor

For the k-Nearest Neighbor (kNN) model, the distances between a datapoint of interest and a number k neighbors are calculated, taking either the Euclidean distance (Equation 4.4), the Manhattan distance (Equation 4.5) or the Minkowski distance (Equation 4.6). For these equations, k is the number of nearest neighbors and q is the Minkowski power. For a classification problem, the most frequent category of the k nearest neighbors is selected and for a regression problem, the average of these categories is selected.

$$d(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{4.4}$$

$$d(x,y) = \sum_{i=1}^{k}\left|x_i - y_i\right| \tag{4.5}$$

$$d(x,y) = \left(\sum_{i=1}^{k}\left(\left|x_i - y_i\right|\right)^q\right)^{1/q} \tag{4.6}$$

4.2.5. Naive Bayes

As shown in Equation 4.7, using Naive Bayes, one can predict the likelihood of y given X by multiplying the likelihood of y with the likelihood of X given y, then dividing by the likelihood of X [43]. For departure delay, if y would be the label "delayed" the and X would be a possible delay cause, e.g. "fog", then the likelihood of the flight being delayed due to fog ($P(y|X)$) would be the likelihood of fog for historic delays ($P(X|y)$) multiplied by the likelihood that there is a delay ($P(y)$), divided by the likelihood that there is fog after all ($P(x)$). Pérez-Rodríguez et al. [62] showed that asymmetric Bayesian models outperform their symmetric counterparts for forecasting departure delay probabilities.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{4.7}$$

4.2.6. Decision Trees

Decision trees allow to perform both classification and regression tasks with high dimensional sets of data, whilst remaining somewhat easy to understand [33]. Starting at the root node (the top box in Figure 4.3), some decision has to be made, resulting in a move towards the right or the left. Depending on the number of nodes, at some point a decision box no longer has lower-level nodes (this is the leaf node), but rather a prediction is made for the outcome: in this case red or green. The number of nodes and leaves and the depth of the tree are some of the hyper-parameters that determine the performance of this method. Whereas Figure 4.3 shows two classes as output (green/red), it is also possible to use decision trees for regression applications. In that case, at each node a value is predicted instead of a class [33].

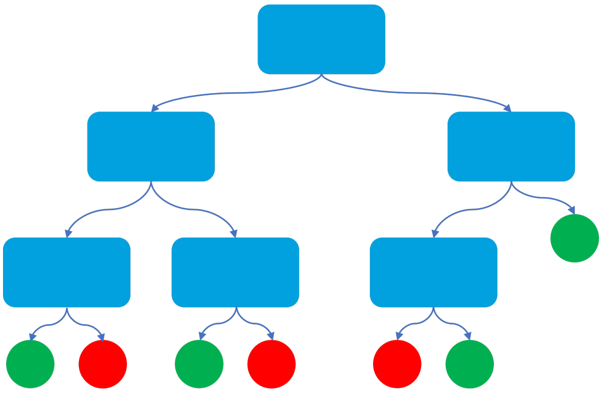


Figure 4.3: Graphical Representation of Underlying Principle of Decision Tree Models

4.2.7. Random Forests

Knowing the principle of decision trees, the principle of random forests follows logically. Rather than taking the outcome of just one decision tree, random forests use an ensemble of decision trees that all train on different parts of the dataset [33]. This concept is referred to as "bagging" and reduces the variance of the estimations [58]. In the end, a majority vote is performed, leading to the final output of the random forest model. This is visualized in Figure 4.4. A variation to regular random forests is Extra-Trees, where the different trees are trained all with the entire dataset, rather than with parts of it [10]. Like decision trees, random forests can be used for both classification and regression problems. Random forest models are among the most powerful supervised learning models [33, 64].

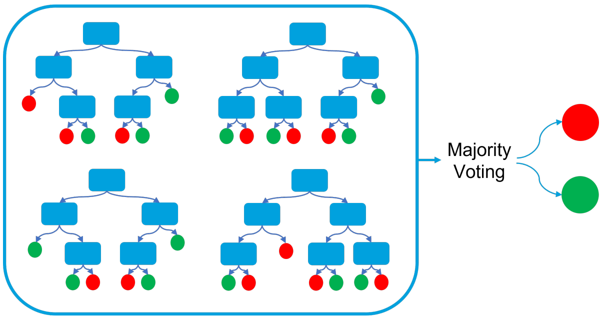


Figure 4.4: Graphical Representation of Underlying Principle of Random Forest Models

4.2.8. Boosting Algorithms

Whereas random forest is a bagging algorithm, where all trees are trained in parallel, boosting algorithms train the different elements in series. As such, the next tree to be trained can learn from the previous tree and can prevent making the same errors. This is illustrated in Figure 4.5. State-of-the-art research shows the potential of boosting algorithms, where Gradient Boosting [55], Extreme Gradient Boosting [7] and LightGBM [84] are being used for both classification and regression applications. More traditional methods such as linear regression and kNN are outperformed by these novel methods.

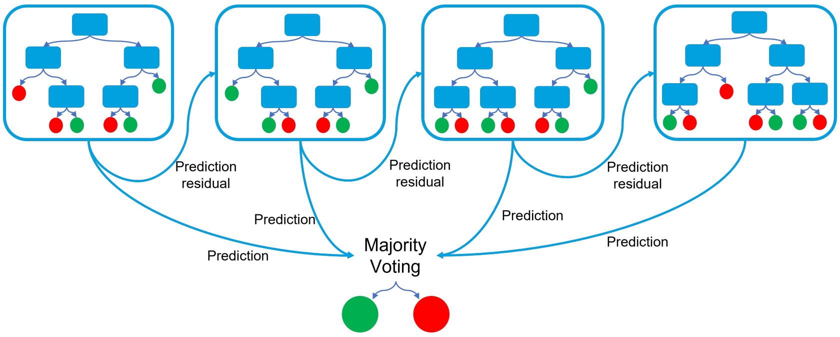


Figure 4.5: Graphical Representation of Underlying Principle of Boosting Algorithms

4.2.9. Neural Networks

Neural networks, inspired by biological neurons, have seen an increase in popularity over the past decade. The smallest sub-unit of a neural network is a perceptron, which finds the weighted sum of inputs (all links in Figure 4.6 have some weight), and feeds this input to an activation function (ϕ in Figure 4.6), which results in the output of the perceptron [33]. This activation function can be linear, sigmoidal, tangential or in the form of a radial basis function [21]. Depending on the use case, neural networks of different kinds can be created, varying the number of hidden layers, number of neurons per layer, type of activation functions and the loss function that is used for error back-propagation [21, 33]. These parameters are referred to as the hyper-parameters of the neural network.

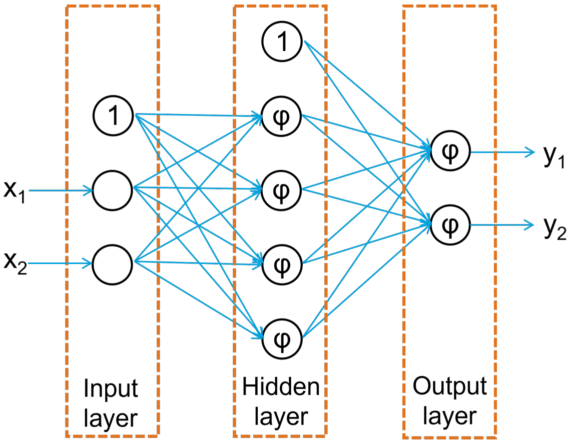


Figure 4.6: Graphical Representation of Underlying Principle of Neural Networks

Existing research on departure delay has seen different approaches. Pamplona et al. [60] used artificial neural networks such as the one explained in the previous section. Sun et al. [73] used recurrent and graph neural networks. When processing sequential data, standard recurrent neural networks have a recurrent unit in the hidden layers that saves the state from a previous timestamp and uses this information for the current timestamp prediction [61]. Long Short-Term Memory (LSTM), which was used by Kim et al. [49], Gui et al. [36] and Sun et al. [73], is more complex as it determines what previous timestamp data is relevant and only uses that data for the current timestamp prediction [39].

4.2.10. Relation Performance and Explainability

Machine learning methods can be categorized into black-box models (inner workings unknowable) and clear-box models (inner workings knowable). With model selection, a trade-off has to be performed in terms of performance and explainability, because explainable models generally lack in performance and high-performing models lack in explainability [63]. This is illustrated in Figure 4.7, where each of the models discussed in Subsection 4.2.1 to Subsection 4.2.9 is located on the graph. Note that the exact location of each of the listed methods in Figure 4.7 may vary slightly, depending on design choices made given the application of the model. Dalmau et al. [19] aimed to find the best fit between performance and explainability and ends up using random forests as a result.

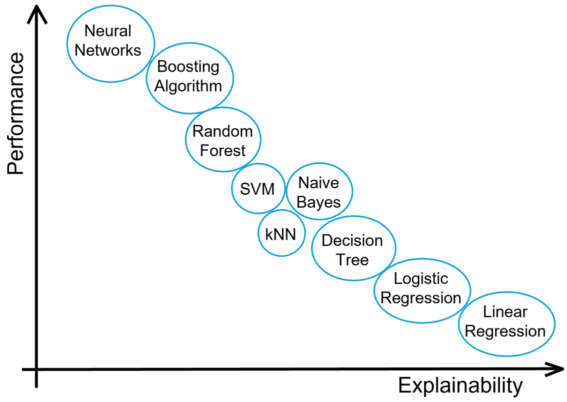


Figure 4.7: Supervised Learning Methods in terms of Performance and Explainability, following Prince [63]

4.3. Feature Engineering

Feature engineering entails the selection of input data into the model. From existing research, it has been shown that feature selection greatly influences the algorithm performances [34]. Bojer [8] explained that "the goal of feature engineering is to come up with a set of transformations that increases the predictive performance of the model". For feature engineering, one should carefully decide what features to select, but also what features not to select. Including irrelevant features may lead to overfitting, which is undesired.

Horiguchi et al. [40] introduced sinusoidal functions for periodical features. As such, it is possible to model December (index 12) and January (index 1) as consecutive months. The same would apply for consecutive days at the beginning of a new month and consecutive hours at the beginning of a new day. For this, Equation 4.8 is consulted where N could be the number of months (12), number of days per month (28/29/30/31) or number of hours per day (24).

$$y = \sin\left(\frac{2\pi}{N}x\right) \tag{4.8}$$

Two different approaches exist for feature selection: filter methods and wrapper methods. For filter methods, features are eliminated based on their correlation with the output. For wrapper methods, the effect of features is studied by adding and subtracting these features to models, seeing the relative differences. Thiagarajan et al. [75] implemented different techniques to perform feature elimination and found that the Recursive Feature Elimination (RFE) algorithm (a wrapper method) resulted in the highest accuracy levels. This algorithm removes the weakest feature until a steep accuracy drop is observed and takes the last stable feature set.

4.4. Hyper-Parameter Tuning

Many of the discussed supervised learning methods from Section 4.2 require a number of user inputs to run. The architecture - and therefore performance - of decision trees, random forests and neural networks is greatly influenced by hyper-parameters, such as the tree depth, number of hidden layers or learning rate.

Due to the endless possible combinations of hyper-parameter values, tuning will most probably result in a local optimum Thiagarajan et al. [75]. Figure 4.8 shows the differences between two commonly used methods, namely Grid Search (which evaluates all parameter combinations) and Random Search (which only evaluates a random sample of the parameter combinations). Whereas Grid Search is more likely to find hyper-parameter combinations with higher accuracy, Random Search (used by Schösser and Schönberger [66]) may reach acceptable hyper-parameter combinations in a much smaller amount of time, especially if the number of hyper-parameters is high [60].

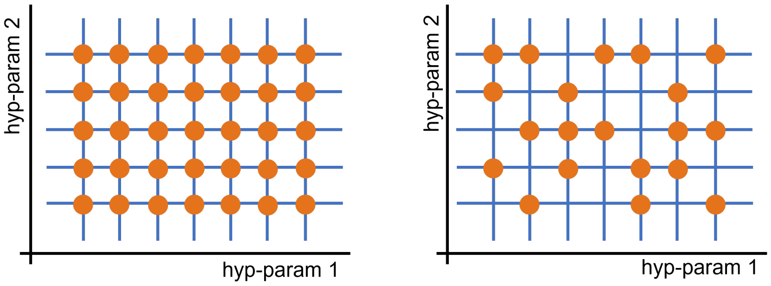
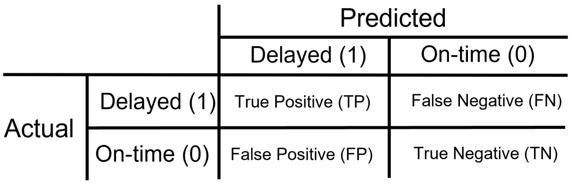


Figure 4.8: Hyper-Parameter Tuning Methods: Grid Search (left) and Random Search (right)

4.5. Performance Metrics

Different performance metrics exist for classification and regression problems. For binary classification problems, there are four possible observations that are summarized in the confusion matrix presented in Figure 4.9. Given the number of observations for True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN), it is possible to determine the accuracy (overall performance), recall (rate of actual delayed flights predicted to be delayed), precision (rate of predicted delayed flights that are actually delayed) and F_β (weighted combination of recall and precision) [41, 66]. The formulas for calculating the accuracy, recall, precision and F$_\beta$-score are presented by Equation 4.9, Equation 4.10, Equation 4.11 and Equation 4.12, respectively. Additionally, one can create a Receiving Operator Characteristic (ROC) curve of the False Positive Rate (FPR) versus the True Positive Rate (TPR). The area under this curve is a measure of performance, often referred to as Area Under Curve (AUC).



		Predicted	
		Delayed (1)	On-time (0)
Actual	Delayed (1)	True Positive (TP)	False Negative (FN)
	On-time (0)	False Positive (FP)	True Negative (TN)

Figure 4.9: Confusion Matrix for Binary Classification of Departure Delays

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.9}$$

$$Precision = \frac{TP}{TP+FP} \tag{4.11}$$

$$Recall = \frac{TP}{TP+FN} \tag{4.10}$$

$$F_\beta = (1+\beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{4.12}$$

For regression problems, some simple performance metrics such as the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE), which better captures the effect of outliers, are shown in Equation 4.13 and Equation 4.14. Furthermore, the coefficient of determination (R^2) is calculated following Equation 4.15. These metrics have been used frequently in existing research [7, 19, 34, 55].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) \tag{4.13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{4.14}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{4.15}$$

Furthermore, some existing research on regression problems used quartiles [19]. When using quartiles, the 25 percentile is referred to as Q1, the 50 percentile is referred to as Q2 or Median and the 75 percentile is referred to as Q3. The inter-quartile range (IQR) is known as the difference between Q3 and Q1. Finally, the mean of a distribution is the average value and the mode of a distribution is the value for which the probability is highest.

4.6. Forecasting Horizon

For forecasting problems such as departure delay, one important choice is the moment at which the forecast is being created. For many applications, such as weather, traffic and stock, forecasts are updated as time passes by. The reason for this may be that new indicators can improve the accuracy of the forecast. For departure delay forecasts, it was found that the forecast accuracy increased towards the moment of departure, partially due to better available weather forecasts.

In his research for predicting arrival delays, Schösser and Schönberger [66] created two feature sets: a long-term feature set including general flight information and a short-term feature set including the arrival delay of the previous flight and the departure delay of the current flight. It was demonstrated that the accuracy of the classification model increased from 69.5% to 89.9% when, next to the long-term features, also the short-term features were taken into account.

Rebollo and Balakrishnan [64] and Gopalakrishnan and Balakrishnan [34] also evaluated the model performance over time. Rebollo and Balakrishnan [64] tested the average median test error at 2 hours, 4 hours, 6 hours and 24 hours before flight and found that this error was 7.4 minutes smaller for the 2 hours forecast relative to the 24 hours forecast. Gopalakrishnan and Balakrishnan [34] tested the accuracy of his models at the same time intervals and found that the average prediction error decreased from around 12 minutes at 24 hours before flight to around 7.5 minutes at 2 hours before flight.

Finally, Dalmau et al. [19] evaluated the model performance at 8 intervals before the estimated departure time of the flight. It was shown that the MAE decreased from 15.1 minutes 4 to 6 hours before flight to 8.8 minutes at 15 minutes before estimated departure. This value may still seem very high, and could be dominated by some extreme delays that the model cannot accurately predict. The median of the error, however, decreased from 8.7 minutes 4 to 6 hours before flight to 4.6 minutes at 15 minutes before flight, which is a more acceptable value.

5 Literature Review

This chapter aims to provide an overview of the research that has been performed on the topic of departure delay forecasting in the past. First, Section 5.1 will elaborate on early departure delay forecasting research, covering mainly the papers that proposed statistical methods. With the increase in popularity of machine learning, it became clear that especially supervised learning methods were applicable to this field of research. For that reason, Section 5.2 covers the research that has been performed on departure delay forecasting using supervised learning methods. Finally, in Section 5.3, general trends following the content of the different papers will be discussed.

5.1. Early Departure Delay Forecasting

Airline departure delay has received a lot of attention in the academic world over the past decade, however this has not always been the case. Only little research was performed on this topic before 2010. Most early papers on departure delay forecasting took a statistical approach [1, 57, 76], and tried to detect recurring temporal patterns in available datasets.

Analyzing departure and arrival flight data of the ten largest US airports between October 14th, 2001 and November 3rd, 2001, Mueller and Chatterji [57] aimed to generate probability distributions that characterize the delay observed in the data. The research demonstrated that even though the traffic numbers differ per day of week, the day-to-day variation of delay departures is rather small, fluctuating between 12% and 18% of all flights. It should be noted that this paper only took into account 3 weeks of data, therefore the effect of the day of week is only based on 3 days for each day of week. On average, at these 10 US airports, the departure delay was found to be 31.7 minutes. The most significant finding by Mueller and Chatterji [57] was that the departure delay probability distribution is best modelled using a Poisson distribution whereas the arrival delay probability distribution is best modelled using a Normal distribution.

Similarly to Mueller and Chatterji [57], the aim of Abdel-Aty et al. [1] was to identify periodic patterns in available data from Orlando Airport (MCO) in 2006. In this research, a two-stage approach was proposed, where first a frequency analysis was performed to detect periodic patterns. Thereafter, in the second stage, the correlation of delay factors was tested based on the observed frequencies. Abdel-Aty et al. [1] found that there exist daily, monthly (30.7 days) and seasonal (173 days) patterns, that match the traffic demands. For example, higher demands were visible for peak hours and the summer holiday season. Furthermore, it was found that season, day of week, time of day, precipitation, flight distance and available turnaround time are factors that were significantly correlated with single flight delay.

Taking a different approach, Tu et al. [76] summed three components that together make up a prediction of the departure delay: a seasonal trend (covering weather conditions and holidays), a daily propagation pattern (covering reactionary delays and peak hours) and a random residual (covering remaining influences). Given the delay prediction method, based on 2000/2001 data from United Airlines at Denver Airport (DEN), a decision model was developed that aims to support airlines in operating from overcrowded airports. Similarly to Mueller and Chatterji [57], the dataset used by Tu et al. [76] was highly positively skewed, meaning that there are many flights with little to no delay and only few flights with large delays. Because of this asymmetric nature of the data, Pérez-Rodríguez et al. [62] compared symmetric and asymmetric Bayesian logistic models and found that the asymmetric model had better in-sample fit than the symmetric model.

5.2. Departure Delay Forecasting using Supervised Learning

For high-dimensional forecasting problems like departure delays, the relations between the causing factors and the resulting flight delay is very complex and highly non-linear [36]. For that reason, statistical prediction models may not capture the full dynamics of the problem. From the literature review, a shift towards the use of machine learning methods was observed around 2015, which also stimulated more researchers to explore the topic as the total number of publications grew rapidly after 2015.

Within the field of machine learning, the departure delay problem can be categorized as a typical supervised learning problem, as for the training data, all data points (flights) have an associated label (departure delay). Following from Géron [33], supervised learning problems can be subdivided into classification problems (predicting classes) and regression problems (predicting values).

In the departure delay forecasting research of Rebollo and Balakrishnan [64], classification errors and regression errors were compared and although a strong positive correlation of 0.78 was found, it was concluded that classification and regression problems differ from each other. The mean reason for this is that different input information is required for predicting the actual delay rather than predicting the range of the delay. According to numerous researchers including Khan et al. [48], due to the uneven distribution of departure delays, the classification methods are more suitable for the problem than regression methods are, since it is difficult to find regressors that can solve the problem with high enough precision. Nonetheless, some interesting papers took a regression approach, albeit with different results.

All papers considered in this literature review are categorized to take a binary classification approach, a multiclass classification approach, a deterministic regression approach or a probabilistic regression approach, and will be discussed in Subsection 5.2.1, Subsection 5.2.2, Subsection 5.2.3 or Subsection 5.2.4, respectively. It was observed that the majority of papers regard the departure delay forecasting problem as a classification problem, with only little papers trying to predict the actual departure delays. Thiagarajan et al. [75] observed the same phenomenon, stating that "in existing literature, researchers have worked extensively on the problem of classification of flight delays and comparatively little research exists on prediction of departure as well as arrival flight delays using machine learning methods".

Since for the upcoming thesis work, a regression problem is to be solved, those papers covering regression problems were deemed more interesting and will be elaborated upon in more detail. Nonetheless, due to the large number of papers considering classification problems, this part could not be overlooked in the literature review.

5.2.1. Binary Classification

In total, 15 papers were reviewed that tackle the departure (or arrival) delay forecasting problem using a binary classification approach. With binary classification, an output is predicted to be 0 or 1. In essence, this means that if a departure delay value is above some threshold, it is marked as 1 (delayed), else it is marked as 0 (on-time). A frequently used threshold is 15 minutes, as the US Bureau of Transportation Statistics (BTS) considers flights to be delayed only if the delay is above 15 minutes [12]. Even though all papers in the upcoming paragraphs use a binary classification approach and similar performance metrics (*accuracy*, *recall*, *precision*, F_β - explained in more detail in Section 4.5), the differences in underlying datasets and the data pre-processing make it difficult to directly compare outcomes of different papers. The 15 papers are listed in chronological order in Table 5.1.

Table 5.1: [**SEE TABLE ON NEXT PAGE:**] Overview of Existing Literature using Binary Classification Approach. Abbreviations used in the table: A: Arrival; D: Departure; NN: Neural Network; DT: Decision Tree; RF: Random Forest; kNN: k-Nearest Neighbor; SVM: Support Vector Machine; OD: Origin-Destination; SM: SMOTE; AUC: Area Under Curve.

Author(s)	Year		Method						Features																							Case Study			Result	Forecast Moment
									Current Flight													Previous Flight				Calendar						Location, Period	Data			
									Numerical							Categorical						Numerical	Categorical													
		Delay Type	Neural Network (NN)	Decision Tree (DT)	Random Forest (RF)	k-Nearest Neighbor (kNN)	Support Vector Machine (SVM)	Other	Scheduled Departure Time	Scheduled Arrival Time	Actual Departure Time	Actual Arrival Time	Flight Time	Flight Distance	Other	Aircraft Type	Aircraft Operator	Origin	Destination	Flight Type / Number	Other	Other	Other	Weather	Passenger	Day of Week	Day of Month	Day of Year	Month of Year	Quarter of Year	Other		Real-Life Data	Generated Data		
Rebollo and Balakrishnan	2014	D			x										NAS Delay State, Type of Delay Day, Type of Delay Day of Previous Day, Influential Airport / OD Pair Delays											x					x	United States, 2007-2008	x		2 hour horizon: Threshold 45 min: Average Test Error = 21%, Threshold 60 min: Average Test Error = 19%, Threshold 90 min: Average Test Error = 16%.	2 hours, 4 hours, 6 hours and 24 hours before flight
Kim et al.	2016	A, D	x						x	x								x	x		Delay Status Origin Airport, Delay Status Destination Airport			x		x		x	x		x	United States, 2010-2015	x		Airport: Accuracy = 87%-91% (71% for low data airport). Flight: Accuracy = 85%-87%	Not mentioned
Belcastro et al.	2016	A			x				x	x	x	x						x	x					x								United States, 2009-2013	x		Threshold 60 min, Accuracy = 85.8% and Recall = 86.9%	Several days before flight
Choi et al.	2016	A		x	x	x		AdaBoost	x	x								x	x					x		x	x		x	x		United States, 2005-2015	x	SM	Day of Flight: RF: Accuracy = 81%, AdaBoost: Accuracy = 78%, kNN: Accuracy = 62%, DT: Accuracy = 77%. 1 day before Flight: RF: Accuracy = 30% and 5 days before Flight: RF: Accuracy = 27%	5, 1 and 0 days before flight
Horiguchi et al.	2017	D	x		x			XGBoost	x	x			x		Fuel On Board			x	x		Domestic / International Flight, Airframe ID, Standby Position ID, Passenger Gender				x	x	x	x	x			Far East, 2012-2015	x		AUC: differs per airport, roughly around 0.60 - 0.80. Slightly higher for Deep Learning, XGBoost and RF	1 day before flight
Thiagarajan et al.	2017	A, D			x			Extra-Trees, AdaBoost, Gradient Boosting	x	x							x	x	x	x				x		x	x		x	x		United States, 2012, 2016	x	SM	Random forest: Arrival Delay: Accuracy = 94%, Departure Delay: Accuracy = 86%	Not mentioned
Gopalakrishnan and Balakrishnan	2017	D	x	x											OD-Pair Delay, OD-Pair Delay Adjacent Pairs						Delay Mode, Type-of-Hour					x					x	United States, 2011-2012	x		Unbalanced Data, 2 hour horizon: Neural Network: Accuracy = 92-93%, Decision Tree: Accuracy = 91%; Balanced Data, 2 hour horizon: Neural Network: Accuracy = 60-66%, Decision Tree: Accuracy = 69%	2hr, 4hr, 6hr and 24 hrs before flight
Pamplona et al.	2018	A, D	x						x	x	x	x					x	x	x	x	Authorization Code, Line Type Code, Justificaiton Delay Code, Flight Situation							x				Brazil, 2017	x		Accuracy = 81-95%, with a mean of 87%	Not mentioned
Chakrabarty et al.	2019	A			x	x	x	Gradient Boosting	x	x								x	x	x						x	x		x	x		United States, 2015-2016	x		Random Forest: Recall = 79%, SVM: Recall = 78%, kNN: Recall = 78%, Gradient Boosting: Recall = 80%	Not mentioned
Yazdi et al.	2020	A, D	x						x	x								x	x	x	Carrier Delay, Weather Delay, NAS Delay, Security Delay	Late Aircraft Delay				x	x	x	x			United States, 2019	x		3 Neural Networks proposed. Balanced dataset: SDA-LM: Accuracy = 92%, SAE-LM: Accuracy = 83%, SDA: Accuracy = 80%; Unbalanced dataset: SDA-LM: Accuracy = 96%, SAE-LM: Accuracy = 86%, SDA: Accuracy = 89%	Not mentioned
Lambelho et al.	2020	A, D	x		x			LightGBM	x	x				x	Number of Seats	x		x	x		Terminal					x	x	x	x			London Heahtrow, 2013-2018	x		Departure Delay: LightGBM: Accuracy = 79%; Neural Network: Accuracy = 77%, Random Forest: Accuracy = 77%. Arrival Delay: LightGBM: Accuracy = 79%; Neural Network: Accuracy = 77%, Random Forest: Accuracy = 76%.	6 months before flight
Huo et al.	2020	A		x	x	x		Logistics Regression, Naïve Bayes	x	x				x	Average Taxi Time, Block Time, Aircraft Group Size		x	x	x	x	Airline Group					x			x			Hong Kong, 2018	x		Random Forest: Accuracy = 73%, Logistic Regression: Accuracy = 63%, kNN: Accuracy = 64%, Decision Tree: Accuracy = 62%, Naïve Bayes: Accuracy = 65%	Not mentioned
Yi et al.	2021	A, D						Stacking (Gaussian Naïve Bayes, Logistic Regression, kNN, DT, RF)	x	x			x	x							Diverted					x	x		x	x		United States, 2019	x	SM	Accuracy of all proposed Stacking Algorithm configurations around 80%. Best algorithm: Accuracy = 82% with AUC = 0.82	Not mentioned
Khan et al.	2021	D	x	x	x		x	Gradient Boosting, XGBoost, hyp-free CPCLS	x	x			x	x	Altitude, Ramp Weight, Speed, Engine Performance	x		x	x		Alternative Destination, Airframe ID, Runway Direction, Runway Surface			x		x	x	x				Hong Kong, 2015-2017	x	SM	Balanced data: hyp-free CPCLS: Accuracy = 83%	4 hours before flight
Schoesser and Schoenberger	2022	A	x		x			XGBoost	x	x			x	x	Departure Delay		x	x	x			Arrival Delay		x		x		x	x		x	Unites States, 2015	x		Week before flight: Random Forest: Accuracy = 69%, XGBoost: Accuracy = 70%, Neural Network: Accuracy = 66%. Just after departure: Random Forest: Accuracy = 90%, XGBoost: Accuracy = 90%, Neural Network: Accuracy = 87%	One week before flight until just after departure

From Table 5.1, it can be concluded that most studies compared the performance of neural networks and decision trees or random forests to that of simpler methods, such as kNN or SVM. All studies, except two, used scheduled departure and arrival times and airports. 6 out of 15 papers used weather information to forecast the flight delays. 4 out of 15 papers forecasted the departure delay at multiple timestamps before scheduled departure. The accuracies of the various proposed models varied between 60% and 95% and most case studies focused on the United States. *A more detailed description of each of the papers can be found in the following paragraphs.*

Rebollo and Balakrishnan [64] considered flight delays both from a classification and a regression point of view. In both problems, data for the 100 most-delayed Origin-Destination pairs (*from now on referred to as OD-pairs*) in the United States in 2007/2008 was used. The used random forest model takes into account features provided by the Bureau of Transportation Statistics such as National Airspace System (NAS) delay state, type of delay day and the delay on adjacent OD-pairs, but no weather data is included in the dataset. For the classification problem, the accuracy of the model is determined for different thresholds (45 min, 60 min and 90 min) and for different forecast horizons (2-h, 4-h, 6-h and 24-h before flight). Intuitively, it was found that for higher classification thresholds (90 min. vs 60 min.) and shorter forecasting horizons, the test error would decrease. Similarly, Choi et al. [16] classified arrival delays using four different methods, where a random forest outperformed a decision tree, an AdaBoost algorithm and a kNN model. Similar to Rebollo and Balakrishnan [64], Choi et al. [16] also found that the forecast accuracy is superior on the day of flight (80%) than one to five days before flight (27-30%).

Kim et al. [49] was the first to solve the classification problem with a deep learning approach. He selected Recurrent Neural Networks (RNN) because of their ability to capture sequential and temporal relations. The paper claims that "the accuracy of RNN improves with deeper architectures", based on previous work of Graves et al. [35] and Pascanu et al. [61]. With the increase in accuracy and complexity in the model of Kim et al. [49], it was found that overfitting is more likely to happen. One way to overcome this problem is by applying the dropout technique, as was demonstrated by Srivastava et al. [70]. This regularization technique randomly drops parts of the neural networks, therefore "thinning" the network structure. As a result, the model is less likely to overfit to one particular node. For 9 out of 10 major US airports, accuracies in the range of 85% to 92% were obtained, with Phoenix Airport (PHX) lacking behind at 71% accuracy, presumably because there was too little training data for that airport. This gives rise to the hypothesis that for deep learning to be effective, the size of the training dataset should be above a certain threshold, something which was later also concluded by Thiagarajan et al. [75]. For individual flights, on average an accuracy of 87% was obtained, slightly outperforming the results of Rebollo and Balakrishnan [64].

Belcastro et al. [6] constructed a random forest model due to its high potential. Similarly to what Rebollo and Balakrishnan [64] found, this paper suggests that increasing the delay threshold leads to higher accuracies. This makes sense because it becomes gradually easier to predict that a flight delay will be below that threshold as statistically more flights are below that threshold. Furthermore, this paper suggests that including weather information does only improve the prediction accuracy marginally. This conclusion is counter-intuitive and was not shared in other existing research.

In the work of Horiguchi et al. [40], the departure delay was predicted using the data of Peach Aviation. Three methods were compared: random forests, XGBoost and deep neural networks. The deep neural network was found to outperform the XGBoost and random forest methods, as the Area Under Curve (AUC) was found to be largest. It should be noted, however, that the AUC varied heavily per airport, which could be caused by the fact that the dataset was unbalanced, or that the selected feature set performed better for some airports than for others.

Thiagarajan et al. [75] used 5 years of domestic US flights to train four classification models: gradient boosting, random forest, Extra-Trees and AdaBoost. This paper extensively elaborated on data pre-processing, covering feature engineering and sampling. Both Choi et al. [16] and Thiagarajan et al. [75] combined oversampling and undersampling for balancing the dataset. Combinations of SMOTE [14] and Tomek Links were used as oversampling alone did not capture the skewed distribution well. For the classification problem, the random forest and gradient boosting methods by Thiagarajan et al. [75] managed to reach accuracies of 86% for departure delay and 94% for arrival delay.

From the research by Gopalakrishnan and Balakrishnan [34] with US commercial flight data from 2011/2012, neural networks slightly outperformed decision trees. Using different feature sets, the accuracies fluctuated between 90 and 94% for a 2 hour forecast horizon and a 60-minute delay threshold. For a lower thresholds of 30 minutes, the accuracy dropped to 85%. In contrast to Rebollo and Balakrishnan [64] and Kim et al. [49], the decrease in accuracy for longer forecasting horizons was only very small (around 1%). This gives rise to the hypothesis that the model may be overfit during training.

The paper by Pamplona et al. [60] only considered flight data on one Brazilian route (Rio de Janeiro (GIG) - São Paulo (GRU)), for which a neural network was found to reach 90% accuracies, without even taking into account meteorological data. A possible explanation for such high accuracy without weather information could be that the weather in these two nearby cities is similar and thus would not influence flights on an individual basis. From other research, it was found that generally higher accuracies are obtained when weather data is included in the feature set [6, 49, 53], whereas research without weather data, such as Lambelho et al. [52], Chakrabarty et al. [13] and Huo et al. [42], reached lower accuracies.

Lambelho et al. [52] compared LightGBM with a neural network and random forests, where all methods reached similar accuracies (77-79%). Chakrabarty et al. [13] found that gradient boosting outperforms random forest, SVM and kNN methods. Huo et al. [42] compared five methods (random forest, logistic regression, kNN, decision tree and naive bayes) with relatively basic features, then found the improved accuracy after adding fewer common features such as average taxi out time, airline group, block time, distance and aircraft group size. The best performing method was random forest, with an accuracy of 73%. Since all of these papers lacked weather data, it could be concluded that the absence of weather data in the feature set leads to lower performance of the models. On the other hand, hour of day was found to be an important feature, both in the three papers named above and in the work of Ye et al. [84].

Yazdi et al. [83] and Yi et al. [85] took a completely different approach. In the work by Yazdi et al. [83], a deep learning model with a de-noising auto-encoder and Levenberg-Marquardt (LM) algorithm was found to reach accuracies of 92-96% outperforming random forests and recurrent neural networks. Yi et al. [85] used a stacking approach, combining different algorithms (Gaussian Naive Bayes (GNB), decision tree, logistic regression, random forest and kNN). Similarly to Choi et al. [16] and Thiagarajan et al. [75], a SMOTE algorithm was applied for oversampling. Using a Boruta algorithm [51], ineffective features were eliminated to reach a forecasting accuracy of around 80%.

Khan et al. [48] aimed to solve the classification problem using a hyper-parameter-free method, reaching an accuracy of 65.5%. The benefit of not having to tune hyper-parameters does not weigh up to the perceived loss in accuracy. Finally, Schösser and Schönberger [66] compared a neural network with XGBoost and random forest for forecasting arrival delays. Feature sets containing long-term and short-term parameters were set up, showing significant improvements for all methods when short-term parameters are used. Accuracies improved from 66-70% for including only long-term parameters to 86-90% when also including short-term parameters. Not completely unexpectedly, it was found that the departure delay of some flight (short-term parameter) is highly correlated to the arrival delay of that flight.

5.2.2. Multi-Class Classification

In total, 6 papers were reviewed that tackle the departure (or arrival) delay forecasting problem using a multi-class classification approach. With multi-class classification, an output is predicted to within one of (three or more) pre-defined bins. The more bins considered for multi-class classification, the closer the approach is to a regression approach. Similar performance metrics (*accuracy*, *recall*, *precision*, F_β, previously explained in Section 4.5), are used for multi-class classification, but their computation is different due to the higher number of bins. The 6 papers are listed in chronological order in Table 5.2.

From Table 5.2, it can be concluded that, similar to the binary classification studies, most multi-class classification studies compared the performance of neural networks and decision trees or random forests to that of simpler methods, such as SVM. All studies used scheduled departure and arrival times and airports. Furthermore, all studies used weather information to forecast the flight delays. No papers forecasted the departure delay at multiple timestamps before scheduled departure, however. Different to binary classification, for multi-class classification, the number of classes is larger than two. Whereas most papers considered 3 up to 6 classes, the paper by Chen and Li [15] considered 15 classes, therefore getting closest to a regression approach. The accuracies of the various proposed models varied between 40% (neural network trained with too little data) and 99% (most probably overfitted neural network). The case studies applied in these papers focused on China, Europe and the United States. *A more detailed description of each of the papers can be found in the following paragraphs.*

Alonso and Loureiro [3] used a multi-class classification approach in which departure delays were classified into one of five classes: (-∞,0], (0, 15], (15, 30], (30, 60] and (60,∞). Fed with data from Porto Airport (OPO) from 2012, a neural network and decision tree were trained. In the end, Alonso and Loureiro [3] found a correlation between predicted and true departure delay of 0.7 and 0.66 for the neural network and decision tree, respectively.

Chen and Li [15] presented an Air Traffic Management (ATM) prediction model that combined random forest classification and an approximated delay propagation model. This chained model has the ability to predict delays of one aircraft for multiple flights of its itinerary. Flight delays were classified into classes of 15 minutes. Furthermore, the SMOTE technique was used for oversampling and all ineffective features were eliminated using Recursive Feature Elimination (RFE). In the end, an average accuracy of 86% was achieved using the model.

Esmaeilzadeh and Mokhtarimousavi [23] proposed a SVM method to classify flight delays into three classes: 0-15 minutes, 15-45 minutes and 45+ minutes. To overcome the black-box nature of the SVM method, Esmaeilzadeh and Mokhtarimousavi [23] performed a sensitivity analysis on the feature inputs, where the impact of minor, moderate and severe changes to the feature values was discussed. A positive correlation was found between weather condition impacts and departure delay, with increased delays due to bad visibility, high wind speed and thunderstorms. Furthermore, Esmaeilzadeh and Mokhtarimousavi [23] found that departure delays are higher when demand exceeds the available airport capacity. Finally, as expected from Figure 2.1, a strong positive correlation between departure delay and take-off delay was noticed, as the two are logically closely related.

The research of Gui et al. [36] and Liu et al. [54] is closely related as both papers used the same Automatic Dependent Surveillance-Broadcast (ADS-B) generated dataset as input. For that reason, the outcome of the results can easily be compared. The dataset was imbalanced due to the large share of on-time flights; undersampling was performed to overcome this issue. Gui et al. [36] proposed a Long-Short Term Memory (LSTM) neural network and a random forest method. The LSTM neural network was most likely overfitted as it reached up to 99% accuracies with training data, but only reached 42% accuracies with test data. The random forest was less prone to overfitting and reached accuracies of 90% for binary classification and 81% and 70% for three-class and four-class classification respectively. Liu et al. [54] proposed a gradient boosting model, which was found to reach accuracies of 88% for binary classification and 79% and 67% for three-class and four-class classification. Therefore, it can be concluded that for this dataset, the random forest model by Gui et al. [36] slightly outperformed the gradient boosting model by Liu et al. [54], leaving the overfitted LSTM model far behind.

Table 5.2: [**SEE TABLE ON NEXT PAGE:**] Overview of Existing Literature using Multi-Class Classification Approach. Abbreviations used in the table: A: Arrival; D: Departure; NN: Neural Network; DT: Decision Tree; RF: Random Forest; kNN: k-Nearest Neighbor; SVM: Support Vector Machine; OD: Origin-Destination; SM: SMOTE.

Author(s)	Year		Method						Features																							Case Study			Result	Forecast Moment
									Current Flight													Previous Flight				Calendar						Location, Period	Data			
									Numerical							Categorical						Numerical	Categorical													
		Delay Type	Neural Network (NN)	Decision Tree (DT)	Random Forest (RF)	k-Nearest Neighbor (kNN)	Support Vector Machine (SVM)	Other	Scheduled Departure Time	Scheduled Arrival Time	Actual Departure Time	Actual Arrival Time	Flight Time	Flight Distance	Other	Aircraft Type	Aircraft Operator	Origin	Destination	Flight Type / Number	Other	Other	Other	Weather	Passenger	Day of Week	Day of Month	Day of Year	Month of Year	Quarter of Year	Other		Real-Life Data	Generated Data		
Alonso and Loureiro	2015	D	x	x					x	x						x	x	x	x		Take-Off Runway; Aircraft Parking Stand	Arrival Delay		x			x	x				Porto Airport, 2012	x		Correltion predicted and true departure delay: Neural Network: r = 0.70; Decision Tree: r = 0.66	Not mentioned
Chen and Li	2019	D			x				x	x			x		Number of Scheduled Departures and Arrivals			x	x					x		x			x			Chicago, 2016-2017	x	SM	Accuracy = 87% (unclear what accuracy is referred to)	Not mentioned
Esmaeilzadeh and Mokhtarimousavi	2020	D					x		x	x					Traffic counts, Number of Scheduled Departures and Arrivals			x	x		National Strategy			x				x				New York Airports, 2018	x		Training Accuracy: 87%, Testing Accuracy: 86%	Not mentioned
Gui et al.	2020	A	x		x				x	x					Traffic Flow			x	x	x	ICAO ID			x		x	x		x		x	China, 2018-2019	x		LSTM Neural Network: Training Accuracy = 99%, Testing Accuracy up to 40%. Random Forest: Binary Classification: Testing Accuracy = 90%, Three-Class Classification: Testing Accuracy = 81%, Four-Class Classification: Testing Accuracy= 70%	Not mentioned
Liu et al.	2020	A						Gradient Boosting	x	x					Traffic Flow			x	x	x	ICAO ID			x		x	x		x		x	China, 2018-2019	x		Binary Classification: Testing Accuracy = 88%, Three-Class Classification: Testing Accuracy = 79%, Four-Class Classification: Testing Accuracy= 67%	Not mentioned
Stefanovic et al.	2020	A, D	x	x	x		x	Tree Ensemble, Gradient Boosting	x	x							x	x	x	x				x				x				Lithuania, 2019-2020	x	SM	Arrival Delay: Probabilistic NN: Accuracy = 34%, MLP NN: Accuracy = 32%, DT: Accuracy = 58%, RF: Accuracy = 59%, Tree Ensemble: Accuracy = 66%, Gradient Boosted Tree: Accuracy = 81%, SVM: Accuracy = 33%; Departure Delay: Probabilistic NN: Accuracy = 82%, MLP NN: Accuracy = 82%, DT: Accuracy = 88%, RF: Accuracy = 85%, Tree Ensemble: Accuracy = 89%, Gradient Boosted Tree: Accuracy = 93%, SVM: Accuracy = 83%	Not mentioned

Whereas a random forest model outperformed a gradient boosting model in the research by Gui et al. [36] and Liu et al. [54], the research of Stefanovič et al. [71] concluded the opposite. Using Lithuanian airport data, oversampled using SMOTE technique, 7 classification techniques (Probabilistic Neural Network (PNN), Multi-Layer Perceptron (MLP), decision tree, random forest, tree ensemble, gradient boosted trees and SVM) were considered to predict arrival and departure delays in 6 classes. For arrival delay, only gradient boosted trees reached an acceptable accuracy (80.9%) with unseen data. For departure delays, however, all listed methods reached accuracies above 80% with unseen data, with gradient boosted trees reaching 92.63%.

5.2.3. Deterministic Regression

In total, 13 papers were reviewed that tackle the departure (or arrival) delay forecasting problem using a deterministic regression approach. Deterministic regression models predict a value rather than a class. Although deterministic regression models converge towards some numerical estimate, no likelihood for that estimate to be true is given. The performance metrics that are mainly used for regression problems include MAE, RMSE, R^2, previously explained in Section 4.5. The 13 papers are listed in chronological order in Table 5.3.

From Table 5.3, it can be concluded that most studies compared the performance of neural networks to that of boosting or bagging methods. All studies, except four, used scheduled departure and arrival times and airports. 8 out of 13 papers used weather information to forecast the flight delays. 4 out of 13 papers forecasted the departure delay at multiple timestamps before scheduled departure. The mean absolute errors of the various proposed models varied between 2 and 37 minutes, very much depending on the methods used. Whereas early case studies mainly focused on the United States, the more recent case studies focused on Europe and the Far East instead. *A more detailed description of each of the papers can be found in the following paragraphs.*

Rebollo and Balakrishnan [64] considered flight delays both from a classification and a regression point of view. In both problems, data for the 100 most-delayed OD-pairs in the United States in 2007/2008 was used. The used random forest model took into account features provided by the Bureau of Transportation Statistics such as NAS delay state, type of delay day and the delay on adjacent OD-pairs, but no weather data was included in the dataset. Similarly to what was observed with their classification problem, Rebollo and Balakrishnan [64] found that for the regression problem, the average median test error increased for larger forecast horizons: the average median test error was 27.4 minutes for a 24 hour forecast horizon and 20 minutes for a 2 hour forecast horizon. This test error improvement is not extremely large, which may be caused by the fact that no weather information is included in the feature set. Most of the features in the dataset do not change much in the period of 2 to 24 hours before flight because they are either fixed already (e.g. calendar features) or because they are still subject to change. Since Rebollo and Balakrishnan [64] forecasted departure delays using classification and regression approaches, the two can be directly compared. Both performed average relative to the other papers in the categories, but adequate given that this study was already performed in 2014.

Thiagarajan et al. [75] used 5 years of domestic US flights to train four regression models: gradient boosting, random forest, Extra-Trees and a MLP neural network. The dataset was sampled using a combination of SMOTE [14] and Tomek Links and scaled using StandardScalar and RobustScalar methods, resulting in the best MSE for the RobustScalar method. It was found that selective training (training on a single origin airport) greatly improved the MSE values, resulting in the best MSE value of 26.36 min^2 for the Extra-Trees method with selective training. Finally, a deep neural network was proposed, but its results were significantly worse than those of the other methods considered (MSE = 892 min^2 for departure delay). Thiagarajan et al. [75] explained this poor performance by the small size of the training data. However, with 5 years of data available for training, this paper had access to one of the largest databases of all reviewed papers in terms of years.

Table 5.3: [**SEE TABLE ON NEXT PAGE:**] Overview of Existing Literature using Deterministic Regression Approach. Abbreviations used in the table: A: Arrival; D: Departure; P: Primary; NN: Neural Network; DT: Decision Tree; RF: Random Forest; kNN: k-Nearest Neighbor; SVM: Support Vector Machine; OD: Origin-Destination; SM: SMOTE; MAE: Mean Absolute Error; MSE: Mean Square Error; RMSE: Root Mean Square Error; R^2: Coefficient of Determination.

Author(s)	Year		Method						Features																							Case Study			Result	Forecast Moment
									Current Flight													Previous Flight				Calendar						Location, Period	Data			
									Numerical							Categorical						Numerical	Categorical													
		Delay Type	Neural Network (NN)	Decision Tree (DT)	Random Forest (RF)	k-Nearest Neighbor (kNN)	Support Vector Machine (SVM)	Other	Scheduled Departure Time	Scheduled Arrival Time	Actual Departure Time	Actual Arrival Time	Flight Time	Flight Distance	Other	Aircraft Type	Aircraft Operator	Origin	Destination	Flight Type / Number	Other	Other	Other	Weather	Passenger	Day of Week	Day of Month	Day of Year	Month of Year	Quarter of Year	Other		Real-Life Data	Generated Data		
Rebollo and Balakrishnan	2014	D			x										NAS Delay State, Type of Delay Day, Type of Delay Day of Previous Day, Influential Airport / OD Pair Delays											x					x	United States, 2007-2008	x		2 hour horizon: Median test error = 19.1 min. 24 hour horizon: Median test error = 27.4 min.	2 hours, 4 hours, 6 hours and 24 hours before flight
Thiagarajan et al.	2017	A, D	x		x			Extra-Trees, Gradient Boosting	x	x							x	x	x	x				x		x	x		x	x		United States, 2012, 2016	x	SM	Gradient Boosting with Extra-Trees regressors and selective training: Arrival Delay: MSE = 26.4 min^2, Departure Delay: MSE = 70.2 min^2	Not mentioned
Kalliguddi and Leboulluec	2017	D		x	x			Multiple Linear Regression						x	Taxi In Time, Taxi Out Time, Carrier Delay, NAS Delay, Security Delay							Late Aircraft Delay		x								United States, 2016	x		Multiple Linear Regression: RMSE = 21.2 min, Decision Tree: RMSE = 26.5 min, Random Forest: RMSE = 12.5 min	Not mentioned
Manna et al.	2017	A, D						Gradient Boosting	x	x							x	x	x							x						United States, 2013	x		Arrival Delay: MAE = 7.6 min, RMSE = 10.7 min, R^2 = 0.92. Departure Delay: MAE = 4.7 min, RMSE = 8.2 min, R^2 = 0.95	Not mentioned
Gopalakrishnan and Balakrishnan	2017	D	x	x				Linear Regression, Markov Jump Linear System (MJLS)							OD-Pair Delay, OD-Pair Delay Adjacent Pairs						Delay Mode, Type-of-Hour					x					x	United States, 2011-2012	x		Unbalanced Data: 2 hour horizon: NN: MAE = 8.5-12 min, RT: MAE = 9 min, LR: MAE = 11 min, MJLS: MAE = 4.5 min; Balanced Data: 2 hour horizon: NN: MAE = 18.5-36 min, RT: MAE = 29.5 min, LR: MAE = 33 min, MJLS: MAE = 44 min	2 hours, 4 hours, 6 hours and 24 hours before flight
Yu et al.	2019	D				x	x	Linear Regression, DBN-SVR							Number of Scheduled Flights, Crowdedness (PAX, Scheduled Flights, Actual Flights)						Base Airport, Airline Size, Boarding Option, Origin vs Pass-by Flight, Air Route Situtation, Market Concentration, ATP, Gap Boarding - Gate	Delay, Interval, Number of Previous Flights Legs										China, 2017-2018	x		DBN-SVR: MAE = 8.4 min, kNN: MAE = 11.0 min, SVM: MAE = 12.0 min, LR: MAE = 15.6 min	2 hours before flight
Shao et al.	2019	D	x				x	Linear Regression, Gradient Boosting	x	x					GPS Flight Trajectories						Cause of Delay, Tail Number			x								Los Angeles, 2016	x		Full Dataset (schedule + weather + ATC): Gradient Boosting: RMSE = 37 min, LR: RMSE = 44.3 min, MLP Neural Network: RMSE = 44.1 min, SVR: RMSE = 43.9 min	Not mentioned
Ye et al.	2020	D					x	Multiple Linear Regression, ExtraTrees, Gradient Boosting	x	x	x	x				x		x	x	x	Number of planned and actual flights, Accumulated delay of planned and actual flights			x		x			x			Nanjijng, 2017-2018	x		Gradient Boosting: MAE = 6.6 min, Extra-Trees: MAE = 6.8 min, SVM: MAE = 6.7 min, Multiple Linear Regression: MAE = 6.8 min	1 hour before flight
Khan et al.	2021	D	x	x	x		x	Gradient Boosting, XGBoost, hyp-free CPCLS	x	x			x	x	Altitude, Ramp Weight, Speed, Engine Performance	x		x	x		Alternative Destination, Airframe ID, Runway Direction, Runway Surface			x		x	x	x				Hong Kong, 2015-2017	x	SM	MAE = 36-46 min across 8 methods, with RF, GBDT, XGBoost having MAE of around 36 min	4 hours before flight
Sun et al.	2022	A, D	x		x				x	x					Departure delay, Airport Capacity		x	x								x			x			Europe, 2017-2019	x		Flight arrival delay: Random Forest: MAE = 3.8-7.7 minutes. Airport arrival delay EGGL: LSTM: MAE = 4.6-5.4 min (lookahead 30 to 180 min), DST-GAT: MAE = 4.3-6.0 min (lookahead 30 to 180 min). Airport departure delay EGGL: LSTM: MAE = 2.2-2.3 min (lookahead 30 to 180 min), DST-GAT: MAE = 2.2-2.6 min (lookahead 30 to 180 min).	30 to 180 minutes before flight
Wang et al.	2022	A, D	x						x	x					Number of planned and actual departures and arrivals, Accumulated delay of planned and actual departures and arrivals			x	x		Traffic Management Initiative (TMI) active			x		x			x			United States, 2019	x		Departure delays: MSE = 20.46-27.40 minutes, MAE = 2.48-3.60 minutes. Arrival delays: MSE = 13.32-18.97 minutes, MAE = 2.12-2.81 minutes	Not mentioned
Dalmau et al.	2023	D						Gradient Boosting	x	x			x		Available Turnaround Time, EOBT shift, ETOT shift, Time to EOBT, ETOT, TOBT and TSAT, ATFM Delay	x	x	x	x	x	Flight state, CDM stateus, De-icing status, last letter of SID, Ready-status, Causes of ATFM Delay	Flight Time, EOBT shift, ETOT shift	Flight State, Aircaft operator, Origin, Flight type	x		x			x			Maastricht Upper Area Control (MUAC), 2019	x		(0, 15] minutes to EOBT: MAE = 8.8 min, σ = 19.6 min, Median = 4.6 min; (240, 360] minutes to EOBT: MAE = 15.1 min, σ = 24.8 min, Median = 8.7 min	8 intervals between 6 hours and 0 minutes before flight
Birolini and Jacquillat	2023	P			x			Linear Regression, XGBoost	x	x	x	x	x		Seating Capacity, Aircraft Age, Airport traffic levels, En-route traffic levels	x		x	x		Seating Configuration, Aircraft Airport Base			x	x	x			x		x	Vueling Network (Europe), 2017-2019	x		Linear Regression: MAE = 12.6 min, RF: MAE = 9.8 min, XGBoost: MAE = 5.4 min	1 day before flight

Both Kalliguddi and Leboulluec [46] and Manna et al. [55] performed an initial statistical analysis before moving on to a supervised learning approach. Kalliguddi and Leboulluec [46] created a predictor correlation plot comparing the correlations between all considered predictors. In the remainder of his paper, Kalliguddi and Leboulluec [46] proposed a multiple regression model, a decision tree structure and a random forest algorithm to predict the departure delay, taking into account the arrival delay of the previous flight leg. For the multiple linear regression model, a RMSE of 21.2 minutes was found, outperforming a simple decision tree that obtained a RMSE of 26.5 minutes. Random forests, on the other hand, performed significantly better than both methods, with a RMSE of 12.5 minutes. From this, it can be concluded that a simple decision tree will not reach the required performance level, and one is better off using a random forest algorithm.

Manna et al. [55] determined the correlation between departure and arrival delay (which was found to equal 0.94) and presented box-plots for the departure delay distribution per day of week and per airline carrier. Then, Manna et al. [55] proposed a gradient boosted decision tree reaching a RMSE of 8.19 minutes and a MAE of 4.70 minutes for departure delays and a RMSE of 10.72 minutes and a MAE of 7.56 minutes for arrival delays. The model was trained with 2.18 million flight instances and tested with 544 thousand flight instances. It should be noted, however, that quite some outliers were removed in order to obtain these results: only flights with delays in the (Q1-1.5·IQR, Q3+1.5·IQR) range were considered for the model training. Furthermore, normalization was executed to a uniform scale of 0 to 1.

For his regression problem, Gopalakrishnan and Balakrishnan [34] used 2011/2012 US commercial flight data and proposed a Feed-Forward (FF) neural network, a generalized regression neural network, a decision tree, a Linear Regression (LR) model and a Markov Jump Linear System (MJLS) model. The latter model is "a useful model class for capturing abrupt changes in system behaviour that are temporally random" [5]. Testing different feature combinations, Gopalakrishnan and Balakrishnan [34] found that the prediction errors varied to some extent by using different features. The general pattern, however, remained the same: the generalized regression neural network and the decision tree outperformed the FF neural network and the LR model. For an unbalanced dataset, the prediction error was found to increase with larger forecast horizons for all methods except MJLS, where the MJLS outperformed the other methods. For the balanced dataset however, the MJLS model had large prediction errors of up to 44 minutes, hinting that it might be prone to overfitting.

Different from the rest, Yu et al. [86] proposed an unsupervised deep belief network (DBN) which, combined with Support Vector Regression (SVR) as supervised component, was found to be competent for forecasting departure delays. According to Yu et al. [86], DBN is "an effective method to reduce the dimension of the input data and eliminate redundant information". The results were average, with a MAE-value of 8.41 minutes and a RMSE-value of 12.65 minutes. It should be noted that for the applied case study, only departing domestic flights at Beijing (PEK) were considered, therefore not capturing the full dynamics of an airport or airline. The research by Yu et al. [86] is one of the few where the introduction of micro-level features such as cargo door closure, gate bridge retraction and type of boarding was evaluated. In the end, however, these features were not included into the model as their correlation with departure delay was deemed to be too low.

Shao et al. [67] also took a unique approach, including Global Positioning System (GPS) coordinates in take-off delay forecasts. The research uses Air Traffic Control (ATC) and GPS data and tested 4 models: Linear Regression (LR), Support Vector Regression (SVR), Multi-Layer Perceptron (MLP) neural network and LightGBM (a gradient boosting method). The noise was removed from the data and using Principal Component Analysis (PCA) and the high dimensional correlated attributes were transformed to a set of variables that are linearly uncorrelated. Using flight schedule, weather and airport GPS data, the methods reached RMSE values of 37 minutes and above. Thus, the performance of the models in this paper does not come close to that of other papers discussed before.

Another interesting paper was written by Ye et al. [84], where 4 different regression methods were compared and reached hopeful results. Due to standardization, the data was less affected by outliers. The paper used a one-hour forecasting horizon and found a MAE of 6.80 minutes for Multiple Linear Regression (MLR), 6.69 minutes for SVM, 6.82 minutes for Extra-Trees and 6.64 minutes for LightGBM. Furthermore, the paper found that once some delay occurs in airports, these delays can hardly be resolved on the short term. Thus, it was found that airport delays have some time correlation. The reason for such low MAE values could be that the methods took into account aggregate characteristics for airport states, looking at the number of aircraft on ground and their respective delay, rather than just that of a single flight.

Like for the classification problem, in the regression problem, Khan et al. [48] tested a hyper-parameter free method. Similarly to the classification problem, however, this method did not reach performance standards that methods from other papers reached. The benefit of not having to tune hyper-parameters thus did not weigh up to the perceived loss in performance.

Similar to Kalliguddi and Leboulluec [46] and Manna et al. [55], the research by Sun et al. [73] presented a correlation plot between the considered features. Although the paper only considered a small number of features (8 for flight arrival delays and 12 for airport arrival and departure delays), still above-average results were obtained. All flights with delays over 90 minutes or under -30 minutes were removed from the dataset (accounting for 0.7% of the data). For flight arrival delays, random forests were found to reach the best results with MAE of around 3.8 to 7.7 minutes, varying per airport. For the airport arrival and departure delays, two neural networks were compared at look-ahead times of 30 to 180 minutes: a Long Short-Term Memory (LSTM) network and a Dynamic Spatial-Temporal Graph Attention (DST-GAT) network. Whilst the network architectures differed significantly, the results presented for London Heathrow Airport (EGLL) were somewhat similar. The MAE for the LSTM network was 4.6-5.4 minutes for arrival delay and 2.2-2.3 minutes for departure delay. The MAE for the DST-GAT network was 4.3-6.0 minutes for arrival delay and 2.2-2.6 minutes for departure delay. For both networks, the prediction errors increased with increasing look-ahead times.

Using a deep learning approach, Wang et al. [78] was able to determine departure delays for the North Eastern corridor of the United States airspace. Given weather information on a 1-by-1 kilometer grid, and making use of transformers that have the ability to capture long-term dependencies, the obtained results on average had a MSE of around 15 min^2 and a MAE of around 2.7 minutes. Additionally, Wang et al. [78] found that the training process of the transformer network was faster than that of regular recurrent neural networks.

The aim of the study of Dalmau et al. [19] was to compare the current EUROCONTROL Enhanced Tactical Flow Management System (ETFMS) prediction error to that of a Gradient Boosted Decision Tree (GBDT) model. It is concluded that the GBDT model only had marginal improvements compared to the current ETFMS model for forecast horizons of up to one hour. For longer forecast horizons, it was found that the GBDT model outperformed the current model. From a feature importance analysis, it was found that the available turn-around time and the Air Traffic Flow Management (ATFM) delay have the largest influence on the result. Finally, the MAE for the GBDT model was found to be 8.8 minutes, which was higher than Ye et al. [84] and Sun et al. [73]. This might be caused by the fact that the research of Dalmau et al. [19] considered over 30 features which may be too much for a tree-based model.

Finally, the research of Birolini and Jacquillat [7] was performed in collaboration with Vueling. The paper proposed linear regression, random forest and extreme gradient boosting models to predict the delay of Vueling flights. It was aimed to only predict primary delays, thus excluding the propagated delay parts, as was previously explained in Subsection 2.3.4. Using a quantile regression model, these primary delays were retrieved from the total delay, turn-around time and propagated delays. Segmentation was used to train a number of separate models per airport in the Vueling network such that the airport-specific features can be taken into account. Flights with primary delays over 60 minutes were filtered out to reduce the effect of outliers. The linear regression model, which was chosen to be the baseline model, scored an out-of-sample MAE of 18.35 minutes. The best performing segmented sparse outlier-removed XGBoost model performed much better and was able to reach an out-of-sample MAE of 7.21 minutes.

5.2.4. Probabilistic Regression

Already in 2005, Xu et al. [82] figured that traditional statistical methods are inadequate to analyze micro-level causes to flight delay. Instead, the paper proposed the use of Bayesian networks. Furthermore, Xu et al. [82] argued that "because the NAS is a stochastic control system, it must be characterized by probability density functions". Interestingly, only recent papers that forecasted departure delays using supervised learning methods include this probabilistic aspect. In total, 3 papers were reviewed that tackle the departure (or arrival) delay forecasting problem using a probabilistic regression approach, which not only predict a value but also its probability. The used performance metrics differ per paper and include the MAE, the Continuous

Ranked Probability Score (CRPS) that generalizes the MAE to probabilistic forecasts and the area under the probability curve for some time interval. The 3 papers are listed in chronological order in Table 5.4.

The thesis work of Vorage [77] was the first to take a probabilistic approach in flight delay forecasting, yielding probability density functions for individual flights. Amsterdam Airport Schiphol data from 2012-2019 was used to train four models: a Mean Variance Estimation (MVE) neural network, a Mixed Density Network (MDN), a dropout neural network and a random forest model. The outcome of the models is a probability curve that shows the probability of occurrence as a function of the departure delay time. From this distribution, one can determine the performance from the distribution MAE, distribution MSE, or the area under this curve for the given departure delay interval. Vorage [77] demonstrated that MDN slightly outperformed the other methods. Although the MAE of both MDN, Dropout and random forest methods was around 10 minutes, the MDN model resulted in the lowest MSE for departure delay forecasts of 15.98 min^2.

Using a similar approach as Vorage [77], Zoutendijk and Mitici [87] extended the concept of probabilistic departure delay forecasting. Similar to Vorage [77], a Mixture Density Network and Random Forest model was proposed, now trained with 2017-2020 data from Rotterdam Airport. For the random forest model, the RMSE and MAE were determined from the distribution of point estimates from each decision tree. The MDN model was found to reach a MAE of 13.23 minutes for forecasting departure delays, whereas the random forest model reached a slightly lower MAE of 12.51. The reason for the MDN model by Zoutendijk and Mitici [87] performing slightly worse than the MDN model by Vorage [77] could be that the training data of the model by Zoutendijk and Mitici [87] was of significant smaller size, which is known to negatively impact the performance of neural networks.

Wang et al. [79] modeled departure and arrival delays as normal distributions, even though Mueller and Chatterji [57], already in early research, concluded that departure delays can be modelled better with Poisson distributions. Therefore, large parts of the forecast distribution of Wang et al. [79] were relatively far away from the median of the distribution, resulting in relatively large errors. Three models were proposed by Wang et al. [79]: a MLP neural network, a LightGBM algorithm and random forest. It was found that 50% of arrival and 65% of departure flights were predicted correctly taking a 35 minute departure delay interval. For the same 35-minute interval, Vorage [77] would reach around 85%-88% of departure flights, indicating (again) that normal distributions are better avoided in departure delay forecasts. This also confirms the potential of mixed density neural networks and random forest for the goal of the upcoming thesis project: forecasting the departure delay probability distribution for individual flights.

Table 5.4: [**SEE TABLE ON NEXT PAGE:**] Overview of Existing Literature using Probabilistic Regression Approach. Abbreviations used in the table: A: Arrival; D: Departure; NN: Neural Network; DT: Decision Tree; RF: Random Forest; kNN: k-Nearest Neighbor; SVM: Support Vector Machine; CRPS: Continuous Ranked Probability Score; MAE: Mean Absolute Error.

Author(s)	Year	Delay Type	Method						Features																							Case Study			Result	Forecast Moment
									Current Flight													Previous Flight				Calendar						Location, Period	Data			
									Numerical							Categorical						Numerical	Categorical													
			Neural Network (NN)	Decision Tree (DT)	Random Forest (RF)	k-Nearest Neighbor (kNN)	Support Vector Machine (SVM)	Other	Scheduled Departure Time	Scheduled Arrival Time	Actual Departure Time	Actual Arrival Time	Flight Time	Flight Distance	Other	Aircraft Type	Aircraft Operator	Origin	Destination	Flight Type / Number	Other	Other	Other	Weather	Passenger	Day of Week	Day of Month	Day of Year	Month of Year	Quarter of Year	Other		Real-Life Data	Generated Data		
Vorage	2021	A, D	x		x									x	Seats, Flights last hour, Flights coming hour		x	x	x		Terminal, Schengen			x		x	x	x				Amsterdam Airport Schiphol, 2012-2019	x		Departures: Statistical: MAE = 11.79 min, MVE: MAE = 10.52 min, MDN: MAE = 9.97 min, Dropout: MAE = 9.59, Random Forest: MAE = 9.99; Arrivals: Statistical: MAE = 14.59 min, MVE: MAE = 12.85 min, MDN: MAE = 12.52 min, Dropout: MAE = 12.23, Random Forest: MAE = 12.94	Not mentioned
Zoutendijk and Mitici	2021	A, D	x		x									x	Seats, Scheduled flights day, Scheduled flights 2h	x	x	x	x					x		x	x	x	x		x	Rotterdam Airport, 2017-2020	x		Departures: MDN: CRPS Mean = 9.12 min, MAE_M = 13.23 min. RFR: CRPS Mean = 8.86 min, MAE_M = 12.51 min. Arrivals: MDN: CRPS Mean = 10.95 min, MAE_M = 15.62 min. RFR: CRPS Mean = 10.85 min, MAE_M = 14.99 min.	Several days before flight
Wang et al.	2022	A, D	x		x			Gradient Boosting	x	x			x		Number of Scheduled Hourly Flights	x	x	x	x		Days of week this Flight Number is Operated, Hub at Guangzhou or not					x						Guangzhou Airport, 2017-2020	x		35 minute interval: Departures: MLP: 72% of area, LightGBM: 62% of area, Random Forest: 62% of area. Arrivals: MLP: 52% of area, LightGBM: 50% of area, Random Forest: 50% of area.	Around 6 months before flight

5.3. General Literature Patterns

Based on all papers that were reviewed in Table 5.1, Table 5.2, Table 5.3 and Table 5.4, some patterns and general conclusions can be drawn. Until 2017, the majority of research focused on the United States. Ciruelos et al. [17] and Cook et al. [18] share this view, claiming that also European research institutes depended on US data. From 2017 onwards, however, more European data (such as EUROCONTROL [26]) became available, giving rise to case study applications in Europe. Additionally, flight delays in Brazil and the Far East have been studied more in recent research. 57% of the reviewed papers took classification approaches, the remaining 43% took regression approaches. Nonetheless, in recent years more studies took a regression approach. Only in 2021, Vorage [77] introduced a probabilistic regression approach, possibly encouraging more research to be conducted to fill remaining research gaps.

In general, it was observed that the majority of papers considered some complex supervised learning models (e.g. neural networks, bagging/boosting methods) and compared their performance to that of other complex models or to simpler models such as kNN and SVM. Esmaeilzadeh and Mokhtarimousavi [23] was the only to solely consider a SVM method. Furthermore, it can be observed that neural networks and deep learning models have gained more attention in recent years. From the research of Kim et al. [49] and Thiagarajan et al. [75], however, it was observed that deep learning models require more training data than was available in the respective researches. As such, the performance of deep learning models was lower than that of simpler neural networks or tree-based models.

Various papers have identified the skewed distribution of flight delays and showed that as a result, there are large performance differences between balanced and unbalanced training datasets. Various sampling, normalization and standardization methods were applied to generate such balanced datasets. Additionally, it was found that almost all papers considered calendar and flight schedule features. Such features (e.g. scheduled departure/arrival time, origin/destination airport) are publicly available. Furthermore, some papers included publicly available weather data. Including weather-related data significantly improved the predictions of almost all supervised learning models. Especially visibility and thunderstorms have large influences on flight delay [72], and thus including these features improves the model accuracies. On the other hand, based on the research of Dalmau et al. [19], using a feature set of too large size (over 50 features) may result in the model being of too high dimension to generate results with low prediction errors.

For classification papers, it was found that the accuracies fluctuated between 60% and 95%, largely depending on the used methods and the data preparation and outlier removal procedures. For regression approaches, the best performing models reached mean absolute errors of 2-3 minutes. The results are difficult to compare to one another since different studies were conducted using different datasets and different outlier removal techniques.

A number of papers investigated the model performance at different forecasting horizons [16, 34, 64, 66, 73]. In every one of these five papers, as expected, it was found that the performance of the model is worse for longer forecasting horizons (i.e. timestamps longer before scheduled departure). Additionally, the model performance was found to improve as accumulated airport data is added to the input feature set. Ye et al. [84] proposed a model with averaged traffic numbers at airports and flight delays of other flights than just the considered flight. This model was found to be amongst the best in terms of mean absolute prediction errors.

The previous paragraphs generalized the work that has already been performed. Still, some research gaps exist, especially for the probabilistic regression approach (Table 5.4). These research gaps will be further elaborated upon in Chapter 6, explaining how the upcoming thesis project will innovate and fill these research gaps.

6 Research Gap

This chapter aims to present the research gaps that were identified based on the literature review that was presented in Chapter 5. Filling the research gaps that will be described in this section is relevant for the academic community as the influence of novel features on departure delay will be examined and is relevant for industry because it aims to further improve flight delay forecasts, most probably leading to cost and emission reductions. For KLM, with whom this thesis project is conducted, improving the current departure delay forecasting model is desirable, since the current model leaves room for improvement, as the day-of-operations conditions are not fully included.

Although the classification of airline departure delays has been thoroughly studied, fewer studies took a regression approach. Recently, a probabilistic regression approach for forecasting airline departure delays was proposed by Vorage [77] and was later used by Zoutendijk and Mitici [87]. Since this probabilistic approach, which also inspired Wang et al. [79], is the state-of-the-art and has only been applied in three studies at the moment of writing, a large number of research gaps exist that have yet to be filled. The upcoming thesis project aims to fill three of these research gaps related to probabilistic departure delay forecasting. These research gaps will be elaborated upon one by one in Section 6.1, Section 6.2 and Section 6.3.

6.1. Detailed Passenger Connection Booking Information

Due to large costs related with passengers missing their connections, airlines occasionally decide to delay an outbound flight when an inbound flight with many connecting passengers arrives behind schedule. Whereas shorter flights may be scheduled multiple times a day and connecting passengers can be re-booked onto the next flight, for intercontinental flights, this is often not the case. The airline is more likely to delay these outbound intercontinental flights. Additionally, the size of the group of connecting passengers from one specific flight to another is of importance. Airlines are more likely to delay their outbound flights if the group of connecting passengers is large. As such, the number of passengers connecting between any two flights in an airline schedule can be a good predictor for flight delays. The planned number of connection passengers, specific to every inbound and outbound flight combination, can be determined from passenger booking information.

In previous work, only two papers considered connecting passenger information. Ciruelos et al. [17] assumed monthly connecting passenger percentages. This does not necessarily add valuable information because it does not specify the actual number of connecting passengers from one flight to another, therefore it does not allow to forecast the delay of specific flights. Sismanidou et al. [68] did use real passenger itineraries from a Marketing Information Data Tapes (MIDT) dataset, but uses this data to determine "a proportion of connecting passengers for a specific itinerary by a specific air carrier". Thus, this study again averages the number of connecting passengers, losing the opportunity to use this information to predict delays for individual flights. As such, the research gap of flight-specific passenger connection information remains, and can be described as follows.

To the best of the author's knowledge, flight-specific passenger connections data (the planned number of

> *connection passengers for every inbound and outbound flight combination) has not been considered in any previous research on airline departure delay forecasting for individual flights.*

6.2. Detailed Airline Day-of-Operations Information

As previously described in Chapter 2, there are many parameters that in some way influence departure delays of individual flights. From Chapter 5, it was concluded that most papers took into account temporal features (e.g. day of week, month), weather features and flight schedule features (e.g. departure and arrival times, origin, destinations, flight time). Some papers proposed some novel features. Khan et al. [48] and Alonso and Loureiro [3] took into consideration the take-off runway and Dalmau et al. [20] integrated the de-icing status of flights. Finally, Yu et al. [86] took into consideration the boarding option (jet bridge or bus), the closing-time of cargo and passenger doors, the time gap between check-in and boarding and the gate closing time.

Through the collaboration with KLM, it is possible to access a large database with flight information, covering the features named above, but also extending it with novel day-of-operations features. Such novel features include first and foremost the flight-specific passenger connections described in Section 6.1. Moreover, novel features include the gate and parking stand information, the total number of (transfer) baggage and hand luggage pieces. Furthermore, International Air Transport Association (IATA) delay codes, previously explained in Section 2.4, are known per delayed flight. Finally, there are many timestamps available for every flight, including the time of load-sheet acceptance, start and end of cleaning, aircraft tail changes, Target Off-Block Time (TOBT) changes and Target Start-up Approval Time (TSAT) changes. Finally, it is also possible to derive previously unexplored features from existing ones, as for example demonstrated by Ye et al. [84], who aggregated flight amounts and delays. The availability of the novel day-of-operations features explained in this section covers the research gap on detailed airline day-of-operations information, which is described as follows.

> *To the best of the author's knowledge, there are day-of-operations features available from KLM data that have not been considered in previous research on airline departure delay forecasting, including gate allocation information, delay codes and operational timestamps. These day-of-operations features directly influence departure delays and when included in the feature set, the forecasting model performance is expected to improve.*

6.3. Dynamical Updates

The third research gap that was identified covers the time aspect of the forecasting process. Whereas the vast majority of papers (around 85%) only considered one forecasting moment before flight, a number of papers compared the accuracy of the results at multiple moments in time [16, 19, 34, 64, 66, 73]. While these 6 papers compared the forecasts over time, they all forecast departure delays in a deterministic manner. No research has yet been performed on how the probabilistic departure delay forecasts for individual flights change over time. Felder et al. [31] researched how probabilistic wind power forecasts change over time, but no such research has yet been performed on individual flight departure delays. For that reason, the third research gap is described as follows.

> *To the best of the author's knowledge, for probabilistic airline departure delay forecasting, it has not yet been investigated how the departure delay probability density forecasts for individual flights change as the scheduled time of departure approaches.*

7 Research Proposal

This chapter aims to present the research proposal for the thesis project. First, in Section 7.1, the main research question and the supporting side questions are outlined. Then, Section 7.2 discusses the objective of the research and finally, in Section 7.3, a work breakdown is presented, illustrating the work to be performed in each of the phases of the thesis project.

7.1. Research Questions

Following from the research gaps that were identified and described in Chapter 6, the main research question is defined as follows:

> **Main Research Question**
>
> What combination of supervised learning models and features allows for the best dynamic departure delay probability distribution forecasts for individual flights and how does airline operational data affect the model performance?

This main research question is then supported by four questions, each consisting out of a number of sub-questions:

1. **What supervised learning models are applicable for dynamic forecasting problems?**
   (a) How do different models compare in terms of performance?
   (b) How do different models compare in terms of explainability?
   (c) How do different models compare in terms of run times?
2. **What features should be included for forecasting departure delay probability distributions?**
   (a) What features may influence airline departure delays?
   (b) What is the relative importance of the features for each of the considered models?
   (c) How should the models handle incomplete and/or incorrect data?
   (d) What is the effect of including flight-specific passenger connection data into the models?
   (e) What is the effect of including detailed airport operational data into the models?
3. **How can the considered models forecast departure delay probability distributions over a dynamic forecasting horizon?**
   (a) What data should be included for generating non-biased and representative training and testing datasets?
   (b) When is a departure delay probability distribution forecast deemed trustworthy?
   (c) How should new data be used to retrain the supervised learning model over time?
4. **How do the models perform for a case study with KLM operational data?**

(a) How do the models perform for a case study with KLM operational data at Amsterdam Airport Schiphol?

(b) How do the models perform for a case study with KLM operational data outside of Amsterdam Airport Schiphol?

(c) How do the models perform during a live trial with real-time data at KLM Flight Dispatch?

(d) How does the performance of the supervised learning models compare to the performance of the method currently adopted by KLM?

7.2. Research Objective

The main goal of the upcoming thesis project is to improve academic state-of-the-art departure delay forecasting models using airline operational data, thereby enhancing the current departure delay forecasting model used by KLM. The model should be able to demonstrate how the probability density forecast for a flight changes as the scheduled departure time approaches. Furthermore, the model should be widely applicable, meaning that it should work for flights operated by different airlines, departing from different airports. Concluding, the objective of the research can be defined as follows:

> **Research Objective**
>
> The research objective is to support airline flight planning by developing supervised learning models that dynamically forecast the departure delay probability distribution for individual flights at different airports.

7.3. Work Breakdown

This section aims to present the planning of this thesis project. The work that is to be performed has been divided into multiple phases, each of which has its own sub-goal. The thesis project has been divided into 5 phases: the orientation phase, literature study phase, initial phase, final phase and defense phase. These five phases are also color-coded in the work breakdown, presented in Figure 7.1. Here, the work has been divided into 7 work-packages (WP) and possible iterations required in the work breakdown are outlined. WP0 and WP1 have been completed at the moment of delivering the literature study report. WP2 will cover the data pre-processing and WP3 will cover the model development. Then, WP4 will cover the verification and validation purposes and WP5 will cover the conclusions and thesis report writing. Finally, WP6 will cover the activities required for the final defense of the thesis.

It is aimed to defend the thesis work in early December 2023, thereby completing the research project. Along the way, 4 weeks of holidays have been planned, two weeks during the initial phase and two weeks during the final phase. The number of weeks allocated to each phase and the sub-goal of each of the project phases is presented in Table 7.1. This table also includes the deliverable per project phase.

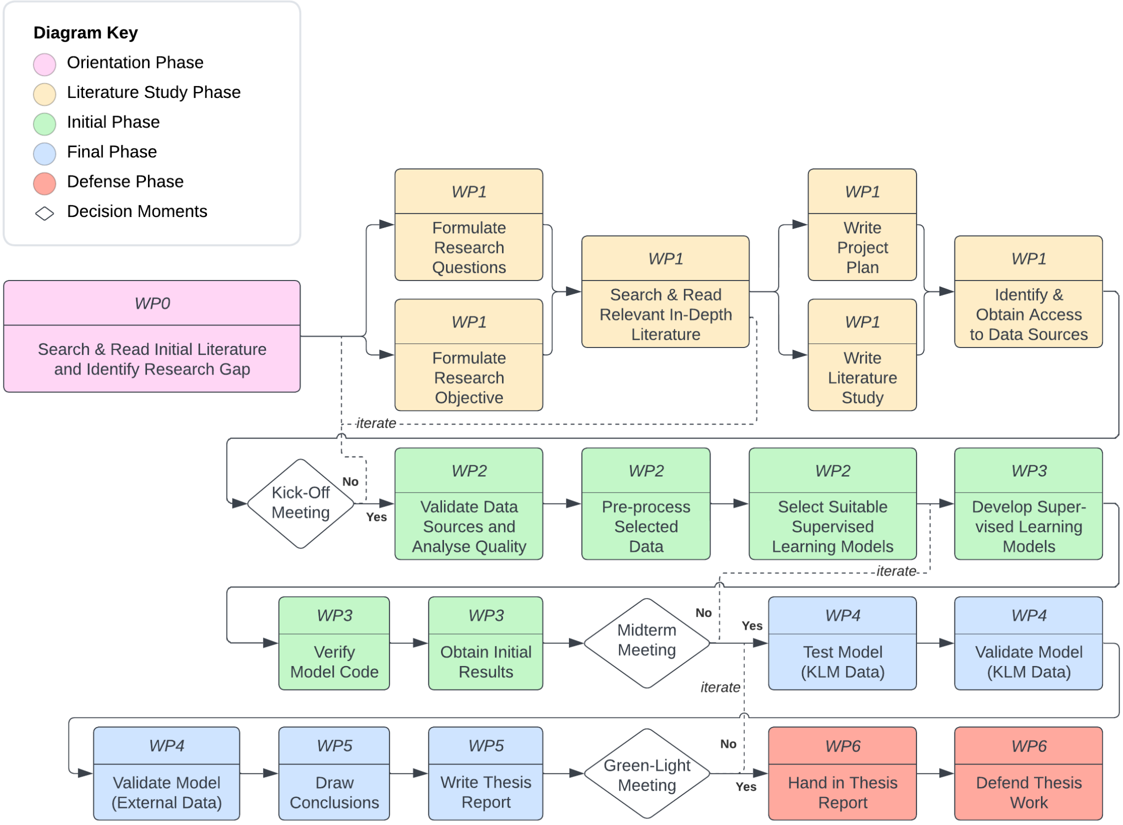


Figure 7.1: MSc Thesis Project Work Breakdown Structure

Table 7.1: MSc Thesis Project Planning

Phase	Nr. of Weeks	Sub-goal	Deliverable (end of Phase)
Orientation	2.5	Find research gap	Research Proposal
Literature Study	8.5	Familiarize with field of research	Project Plan, Literature Study Report
Initial	13 (+2 holiday)	Develop supervised learning model	Midterm Presentation
Final	12 (+2 holiday)	Verify and validate developed model	Green Light Presentation, Final Thesis
Defense	5	Defend performed work and graduate	Defense Presentation

8 Conclusion

This chapter aims to conclude this literature study on airline departure delay forecasting. The aim of the performed literature study was to identify the state-of-the-art in the departure delay forecasting field of research, whilst also gaining a better understanding of related concepts such as departure delay causes and costs, airport operations and forecasting methods.

Departure delay forecasting was found to be a field of research for which supervised learning solutions are especially applicable, because of the labelled nature of the available data. Also from existing work, it was noticed that supervised learning models were frequently used to solve departure delay forecasting problems. As a result, for this literature study and the upcoming thesis project, it was chosen to focus the attention on supervised learning models only.

In total, 56 papers were reviewed, 37 of which were elaborated upon in more detail in the literature review chapter. These papers were divided into four categories, based on the supervised learning approach that was taken. 15 papers took a *binary classification* approach, thereby predicting whether the departure of an individual flight would be delayed or on-time. 6 papers took a *multi-class classification* approach, for which departure delays were forecasted to be within pre-defined delay classes. 13 papers took a *deterministic regression* approach, where the departure delay of individual flights was forecasted as a point estimate. Finally, 3 relatively recent papers took a *probabilistic regression* approach, where the forecasted departure delay would be in the form of a probability density function.

A large share of the reviewed papers performed case studies where multiple supervised learning models were tested and compared to one another, varying from simple linear regression models to complex tree structures and deep learning solutions. Most frequently, bagging methods (such as random forests), boosting methods and neural networks were proposed, albeit at different performance levels. For classification approaches, the forecasting model accuracies varied between 60% and 95%, whereas for the regression approaches, the performance in terms of mean absolute error varied between 3 and 35 minutes. These large differences can be attributed partially to the dataset preparation, as it was found that pre-processing practices such as outlier removal, heavily influenced the performance of the models.

Almost all studies took into account features describing flight schedules (e.g. scheduled departure and arrival time) and calendar features (e.g. day of week and month of year). Additionally, 59% of the papers took into account some form of forecasted weather data. From papers that directly compared models with and without weather data, it was found that including weather data enhances the forecasting performance. Moreover, it was found that the model performance would improve when averaged traffic numbers and averaged departure delays of other flights would be considered in the model.

Although departure delay forecasting has been researched quite extensively, there are still many open ends that leave room for future work. In previous work, mostly general flight features were tested. The influence of detailed flight features, such as flight-specific connection passenger data and airline day-of-operations data has not yet been researched, most probably because such airline data is confidential and unavailable to most researchers. Furthermore, departure delay forecasting using a probabilistic regression approach has been researched in only three studies. It has yet to be determined how such probabilistic departure delay forecasts change as the scheduled departure time approaches.

Filling the previously identified research gaps may result in improved departure delay forecasting models. Such improved models are desirable for airlines as it allows flight dispatchers to more effectively plan the aircraft speed at different phases of flight. As such, compared to current forecasting models, improved models may reduce emissions if flights can be planned at relatively lower speeds and more passenger connections can be made if flights have to be planned at relatively higher speeds. The academic community will also benefit from the proposed research as the influence of novel, previously unavailable, features on airline departure delay will be investigated.

In order to fill the research gaps, a research proposal was established, including several research questions. The main research question was formulated as follows.

> *What combination of supervised learning models and features allows for the best dynamic departure delay probability distribution forecasts for individual flights and how does airline operational data affect the model performance?*

From the performed literature review, it is possible to form a hypothesis to this research question. For case studies with little data, bagging and boosting tree structures were found to generate the best results. It is worth investigating neural network structures when larger training datasets are available, however. From the literature review, it became apparent that weather information generally leads to higher model performances. Additionally, it is expected that novel features describing passenger connections and airline day-of-operation processes will improve the model performance.

Through the collaboration with KLM Royal Dutch Airlines, it is be possible to access data covering flights and their corresponding flight plans, passenger itineraries (anonymized), baggage, security queues at Amsterdam Schiphol Airport and aircraft maintenance. This data, combined with publicly available meteorological data, will be used to train different supervised learning models to obtain probabilistic departure delay forecasts. These forecasts should allow for dynamical updates and should be applicable to individual flights from different airports worldwide. Verification and validation will not only be performed with KLM data, but also using external, multi-airline data samples from EUROCONTROL.

Bibliography

[1] Mohamed Abdel-Aty, Chris Lee, Yuqiong Bai, Xin Li, and Martin Michalak. Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, 13(6):355–361, 2007.

[2] Alaska Airlines. Whats an atc delay? heres what can cause them and what alaska airlines is doing behind the scenes, 2018. https://news.alaskaair.com/travel-tips/what-is-an-atc-delay/ (visited on 30/03/2023).

[3] Hugo Alonso and António Loureiro. Predicting flight departure delay at porto airport: A preliminary study. In *2015 7th International Joint Conference on Computational Intelligence (IJCCI)*, volume 3, pages 93–98. IEEE, 2015.

[4] Uğur Arıkan, Sinan Gürel, and M Selim Aktürk. Flight network-based approach for integrated airline recovery with cruise speed control. *Transportation Science*, 51(4):1259–1287, 2017.

[5] Mark P Balenzuela, Adrian G Wills, Christopher Renton, and Brett Ninness. Parameter estimation for jump markov linear systems. *Automatica*, 135:109949, 2022.

[6] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Using scalable data mining for predicting flight delays. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1–20, 2016.

[7] Sebastian Birolini and Alexandre Jacquillat. Day-ahead aircraft routing with data-driven primary delay predictions. *European Journal of Operational Research*, 2023.

[8] Casper Solheim Bojer. Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, 38(4):1555–1561, 2022.

[9] Stephane Bratu and Cynthia Barnhart. An analysis of passenger delays using flight operations and passenger booking data. *Air Traffic Control Quarterly*, 13(1):1–27, 2005.

[10] Jason Brownlee. How to develop an extra trees ensemble with python, 2020. https://machinelearningmastery.com/extra-trees-ensemble-with-python/ (visited on 31/03/2023).

[11] Bureau of Transportation Statistics. Airline on-time statistics and delay causes, 2022. https://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp (visited on 28/03/2023).

[12] Bureau of Transportation Statistics. On-time performance, 2023. https://www.transtats.bts.gov/HomeDrillChart.asp (visited on 30/03/2023).

[13] Navoneel Chakrabarty, Tuhin Kundu, Sudipta Dandapat, Apurba Sarkar, and Dipak Kumar Kole. Flight arrival delay prediction using gradient boosting classifier. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*, pages 651–659. Springer, 2019.

[14] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[15] Jun Chen and Meng Li. Chained predictions of flight delay using machine learning. In *AIAA Scitech 2019 forum*, page 1661, 2019.

[16] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris. Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE, 2016.

[17] C Ciruelos, A Arranz, I Etxebarria, S Peces, B Campanelli, P Fleurquin, VM Eguiluz, and JJ Ramasco. Modelling delay propagation trees for scheduled flights. In *Proceedings of the 11th USA/EUROPE Air Traffic Management R&D Seminar, Lisbon, Portugal*, pages 23–26, 2015.

[18] AJ Cook, G Tanner, S Cristobal, and M Zanin. Delay propagation–new metrics, new insights. In *Eleventh USA/Europe air traffic management research and development seminar*, 2015.

[19] Ramon Dalmau, Franck Ballerini, Herbert Naessens, Seddik Belkoura, and Sebastian Wangnick. An explainable machine learning approach to improve take-off time predictions. *Journal of Air Transport Management*, 95:102090, 2021.

[20] Ramon Dalmau, Giuseppe Murgese, Yves De Wandeler, Ricardo Correira, Alan Marsden, and EUROCONTROL Innovation Hub EIH. Early detection of night curfew infringements by delay propagation with neural networks. 2021.

[21] Coen de Visser. Ae4320 system identification of aerospace vehicles: Introduction to neural networks, 2022. [PowerPoint slides]. Delft University of Technology, Department of Control & Simulation (consulted on 31/03/2023).

[22] Kaitano Dube, Godwell Nhamo, and David Chikodzi. Covid-19 pandemic and prospects for recovery of the global aviation industry. *Journal of Air Transport Management*, 92:102022, 2021.

[23] Ehsan Esmaeilzadeh and Seyedmirsajad Mokhtarimousavi. Machine learning approach for flight departure delay prediction and analysis. *Transportation Research Record*, 2674(8):145–159, 2020.

[24] EUROCONTROL. All-causes delays to air transport in europe quarter 3 2019. *CODA Digest*, 2019.

[25] EUROCONTROL. All-causes delays to air transport in europe quarter 3 2022. *CODA Digest*, 2022.

[26] EUROCONTROL. Aviation data for research, 2023. Data retrieved from EUROCONTROL https://www.eurocontrol.int/dashboard/rnd-data-archive (visited on 20/03/2023).

[27] EUROCONTROL. All-causes delays to air transport in europe annual 2022. *CODA Digest*, 2023.

[28] EUROCONTROL. Airport collaborative decision making, n.d. https://www.eurocontrol.int/concept/airport-collaborative-decision-making (visited on 30/03/2023).

[29] European Union. Regulation (ec) no 261/2004 of the european parliament and of the council. *Official Journal of the European Union*, 2004.

[30] FAA. Cost of delay estimates 2019, 2020. ttps://www.faa.gov/data_research/aviation_data_statistics/media/cost_delay_estimates.pdf (visited on 11/04/2023).

[31] Martin Felder, Anton Kaifel, and Alex Graves. Wind power prediction using mixture density recurrent neural networks. In *Poster presentation Gehalten auf der European wind energy conference*, 2010.

[32] Flightradar24. Klm routes and destinations, 2023. https://www.flightradar24.com/data/airlines/kl-klm/routes (visited on 30/03/2023).

[33] Aurélien Géron. Hands-on machine learning with scikit-learn, keras, and tensorflow. 2019.

[34] Karthik Gopalakrishnan and Hamsa Balakrishnan. A comparative analysis of models for predicting delays in air traffic networks. In *Twelfth USA/Europe air traffic management research and development seminar*. ATM Seminar, 2017.

[35] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[36] Guan Gui, Fan Liu, Jinlong Sun, Jie Yang, Ziqi Zhou, and Dongxu Zhao. Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, 69(1):140–150, 2019.

[37] Marta Guimarães, Cláudia Soares, and Rodrigo Ventura. Decision support models for predicting and explaining airport passenger connectivity from data. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):16005–16015, 2022.

[38] LK Hassan, Bruno Filipe Santos, and Jeroen Vink. Airline disruption management: A literature review and practical challenges. *Computers & Operations Research*, 127:105137, 2021.

[39] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[40] Yuji Horiguchi, Yukino Baba, Hisashi Kashima, Masahito Suzuki, Hiroki Kayahara, and Jun Maeno. Predicting fuel consumption and flight delays for low-cost airlines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4686–4693, 2017.

[41] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[42] Jiage Huo, Kin Lok Keung, Carman Ka Man Lee, Kam KH Ng, and KC Li. The prediction of flight delay: Big data-driven machine learning approach. In *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 190–194. IEEE, 2020.

[43] IBM. What are naive bayes classifiers?, n.d. https://www.ibm.com/topics/naive-bayes (visited on 31/03/2023).

[44] Iowa State University. Iowa environmental mesonet, 2023. Data retrieved from Iowa State University Environmental Mesonet, https://mesonet.agron.iastate.edu/request/download.phtml (visited on 22/03/2023).

[45] Alexandre Jacquillat and Amedeo R Odoni. A roadmap toward airport demand and capacity management. *Transportation Research Part A: Policy and Practice*, 114:168–185, 2018.

[46] Anish M Kalliguddi and Aera K Leboulluec. Predictive modeling of aircraft flight delay. *Universal Journal of Management*, 5(10):485–491, 2017.

[47] Lisa Kestens. Root cause analysis of atc delays: A case study on klm flights at schiphol airport. [Master Thesis, Delft University of Technology]. 2021.

[48] Waqar Ahmed Khan, Hoi-Lam Ma, Sai-Ho Chung, and Xin Wen. Hierarchical integrated machine learning model for predicting flight departure delays and duration in series. *Transportation Research Part C: Emerging Technologies*, 129:103225, 2021.

[49] Young Jin Kim, Sun Choi, Simon Briceno, and Dimitri Mavris. A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE, 2016.

[50] KLM Royal Dutch Airlines. Transfers, n.d. https://www.klm.co.uk/information/airport/transfers (visited on 30/03/2023).

[51] Miron B Kursa and Witold R Rudnicki. Feature selection with the boruta package. *Journal of statistical software*, 36:1–13, 2010.

[52] Miguel Lambelho, Mihaela Mitici, Simon Pickup, and Alan Marsden. Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management*, 82:101737, 2020.

[53] Qiang Li, Cinga Guan, and Jinpeng Liu. A cnn-lstm framework for flight delay prediction. *Available at SSRN 4283680*, 2022.

[54] Fan Liu, Jinlong Sun, Miao Liu, Jie Yang, and Guan Gui. Generalized flight delay prediction method using gradient boosting decision tree. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5. IEEE, 2020.

[55] Suvojit Manna, Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, and Subhas Barman. A statistical approach to predict flight delay using gradient boosted decision tree. In *2017 International conference on computational intelligence in data science (ICCIDS)*, pages 1–5. IEEE, 2017.

[56] Bart Meijer. Amsterdam's schiphol airport to limit passenger flow until march, 2022. https://www.reuters.com/world/europe/amsterdams-schiphol-airport-limit-passenger-flow-until-march-2022-09-29/ (visited on 18/04/2023).

[57] Eric Mueller and Gano Chatterji. Analysis of aircraft arrival and departure delay characteristics. In *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, page 5866, 2002.

[58] Anuja Nagpal. Decision tree ensembles - baggin and boosting, 2017. https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9 (visited on 31/03/2023).

[59] Anna Norin, Di Yuan, Tobias Andersson Granberg, and Peter Värbrand. Scheduling de-icing vehicles within airport logistics: a heuristic algorithm and performance evaluation. *Journal of the Operational Research Society*, 63:1116–1125, 2012.

[60] Daniel Alberto Pamplona, Li Weigang, Alexandre Gomes de Barros, Elcio Hideiti Shiguemori, and Claudio Jorge Pinto Alves. Supervised neural network with multilevel input layers for predicting of air traffic delays. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[61] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*, 2013.

[62] Jorge Vicente Pérez-Rodríguez, José María Pérez-Sánchez, and Emilio Gómez-Déniz. Modelling the asymmetric probabilistic delay of aircraft arrival. *Journal of Air Transport Management*, 62:90–98, 2017.

[63] S. Prince. Explainability i: local post-hoc explanations, 2022. https://www.borealisai.com/research-blogs/explainability-i-local-post-hoc-explanations/ (visited on 31/03/2023).

[64] Juan Jose Rebollo and Hamsa Balakrishnan. Characterization and prediction of air traffic delays. *Transportation research part C: Emerging technologies*, 44:231–241, 2014.

[65] Royal Schiphol Group. Jaarcijfers 2019, 2020. https://nieuws.schiphol.nl/jaarcijfers-2019/ (visited on 18/04/2023).

[66] Delia Schösser and Jörn Schönberger. On the performance of machine learning based flight delay prediction–investigating the impact of short-term features. *Promet-Traffic&Transportation*, 34(6):825–838, 2022.

[67] Wei Shao, Arian Prabowo, Sichen Zhao, Siyu Tan, Piotr Koniusz, Jeffrey Chan, Xinhong Hei, Bradley Feest, and Flora D Salim. Flight delay prediction using airport situational awareness map. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 432–435, 2019.

[68] Athina Sismanidou, Joan Tarradellas, and Pere Suau-Sanchez. The uneven geography of us air traffic delays: Quantifying the impact of connecting passengers on delay propagation. *Journal of Transport Geography*, 98:103260, 2022.

[69] Skybrary. Aircraft ground de/anti-icing, n.d. https://www.skybrary.aero/articles/aircraft-ground-deanti-icing (visited on 30/03/2023).

[70] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15 (1):1929–1958, 2014.

[71] Pavel Stefanovič, Rokas Štrimaitis, Olga Kurasova, et al. Prediction of flight time deviation for lithuanian airports using supervised machine learning model. *Computational intelligence and neuroscience*, 2020, 2020.

[72] Alice Sternberg, Diego Carvalho, Leonardo Murta, Jorge Soares, and Eduardo Ogasawara. An analysis of brazilian flight delays based on frequent patterns. *Transportation Research Part E: Logistics and Transportation Review*, 95:282–298, 2016.

[73] Junzi Sun, Tristan Dijkstra, Constantinos Aristodemou, Vlad Buzetelu, Theo Falat, Tim Hogenelst, Niels Prins, and Benjamin Slijper. Designing recurrent and graph neural networks to predict airport and air traffic network delays. In *10th International Conference for Research in Air Transportation*. FAA & Eurocontrol, 2022.

[74] The Geography of Transport Systems. Airport components and terminal configurations, n.d. https://transportgeography.org/contents/chapter6/airport-terminals/airport-terminals-configuration/ (visited on 18/04/2023).

[75] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, and Vineeth Vijayaraghavan. A machine learning approach for prediction of on-time performance of flights. In *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE, 2017.

[76] Yufeng Tu, Michael O Ball, and Wolfgang S Jank. Estimating flight departure delay distributionsa statistical approach with long-term trend and short-term pattern. *Journal of the American Statistical Association*, 103(481):112–125, 2008.

[77] Laurence Vorage. Predicting probabilistic flight delay for individual flights using machine learning models. [Master Thesis, Delft University of Technology], 2021.

[78] Liya Wang, Alex Tien, and Jason Chou. Multi-airport delay prediction with transformers. In *AIAA AVIATION 2022 Forum*, page 3707, 2022.

[79] Ziming Wang, Chaohao Liao, Xu Hang, Lishuai Li, Daniel Delahaye, and Mark Hansen. Distribution prediction of strategic flight delays via machine learning methods. *Sustainability*, 14(22):15180, 2022.

[80] Ronald Wesonga, Fabian Nabugoomu, and Peter Jehopio. Parameterized framework for the analysis of probabilities of aircraft delay at an airport. *Journal of Air Transport Management*, 23:1–4, 2012.

[81] Qian Wu. A stochastic characterization based data mining implementation for airport arrival and departure delay data. In *Applied Mechanics and Materials*, volume 668, pages 1037–1040. Trans Tech Publ, 2014.

[82] Ning Xu, George Donohue, Kathryn Blackmond Laskey, and Chun-Hung Chen. Estimation of delay propagation in the national aviation system using bayesian networks. In *6th USA/Europe Air Traffic Management Research and Development Seminar*. FAA and Eurocontrol Baltimore, 2005.

[83] Maryam Farshchian Yazdi, Seyed Reza Kamel, Seyyed Javad Mahdavi Chabok, and Maryam Kheirabadi. Flight delay prediction based on deep learning and levenberg-marquart algorithm. *Journal of Big Data*, 7:1–28, 2020.

[84] Bojia Ye, Bo Liu, Yong Tian, and Lili Wan. A methodology for predicting aggregate flight departure delays in airports based on supervised learning. *Sustainability*, 12(7):2749, 2020.

[85] Jia Yi, Honghai Zhang, Hao Liu, Gang Zhong, and Guiyi Li. Flight delay classification prediction based on stacking algorithm. *Journal of Advanced Transportation*, 2021:1–10, 2021.

[86] Bin Yu, Zhen Guo, Sobhan Asian, Huaizhu Wang, and Gang Chen. Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, 125:203–221, 2019.

[87] Micha Zoutendijk and Mihaela Mitici. Probabilistic flight delay predictions using machine learning and applications to the flight-to-gate assignment problem. *Aerospace*, 8(6):152, 2021.