Considering Airport Planners' Preferences and Imbalanced Datasets when Predicting Flight Delays and Cancellations

Hendrickx, Rik ; Zoutendijk, Mike; Mitici, Mihaela; Schäfer, Jeffrey

DOI
10.1109/DASC52595.2021.9594367

Publication date
2021

Document Version
Final published version

Published in
40th Digital Avionics Systems Conference, DASC 2021 - Proceedings

Considering Airport Planners' Preferences and Imbalanced Datasets when Predicting Flight Delays and Cancellations

Rik Hendrickx
Department of Control & Operations
Delft University of Technology
Delft, The Netherlands
rikhendrickx@live.be

Mike Zoutendijk
Department of Control & Operations
Delft University of Technology
Delft, The Netherlands
m.zoutendijk@tudelft.nl

Mihaela Mitici
Department of Control & Operations
Delft University of Technology
Delft, The Netherlands
m.a.mitici@tudelft.nl

Jeffrey Schäfer
Royal Schiphol Group
Schiphol, The Netherlands
jeffrey.schafer@schiphol.nl

2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC) | 978-1-6654-3420-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/DASC52595.2021.9594367

***Abstract*—A key part of efficient airport operational planning is to have insight into potential flight delays and cancellations. For airport planners, it is important to obtain flight delay or cancellation predictions with a high degree of certainty, i.e. a high precision. This allows planners to make sound decisions based on these predictions. To obtain such predictions, machine learning classification techniques are often applied. An important issue for classification problems is that of imbalanced class distributions: the number of actually cancelled/delayed flights is low. In general, the imbalance is addressed by resampling the data using one or more sampling techniques. However, resampling does not necessarily correspond to an imbalance ratio that leads to the best classification results. In this paper a systematic approach is presented to deal with imbalanced data for classification problems, while taking into account the preferences of airport planners. A range of feasible imbalance ratios, together with several classification algorithms and sampling techniques, are considered. An optimal imbalance ratio is identified with respect to relevant performance metrics. The approach is illustrated by performing binary classification of flight cancellations and delays at a large European airport. The results show that the highest prediction precision is obtained using a base imbalance ratio, whereas a higher imbalance ratio is needed to obtain the highest F1-score. Specifically, the cancellation prediction performance is increased by up to 243%, while its optimal imbalance ratio does not correspond to resampling. In general, the results underline the need to investigate the influence of varying data imbalance ratios on the performance of classification algorithms.**

***Index Terms*—flight delay, machine learning, imbalance, classification**

I. Introduction

Flight on-time performance is an important measure for airport and airline service quality. Before the COVID-19 crisis, the continuous growth of air traffic led to challenging scheduling situations and an increase in flight delays and cancellations: In 2018, more than 11 million flights were operated in Europe, with an average delay of 14.7 minutes, an increase of 3.8% and 17% from 2017, respectively [1], [2]. After the crisis, the air traffic volume is expected to restore to its pre-crisis level within 5 years [3]. An increase in the number of flight delays and cancellations has detrimental effects on an airline's and airport's quality of service and revenue [4]. As such, having the ability to anticipate which flights may be cancelled or delayed is of great value for airports and airlines, as it allows for pro-active decision making to mitigate the effects of cancellations/delays. In order to anticipate flight delays and cancellations it is necessary to predict these events ahead of time, preferably with a high certainty, in order to allow efficient managing of the airports resources.

One class of techniques that can be used to predict flight delays and cancellations is that of machine learning classification techniques. In the past years, several studies have developed machine learning algorithms to predict flight delays and cancellations [5], emphasizing the importance of flight on-time performance. One of the challenges of classification problems is the fact that the used datasets can have an imbalanced class distribution, i.e., the amount of samples in the class of interest is only a fraction of the amount of samples in the majority class. This imbalance leads to a low performance of the classification algorithms [6], which usually work best when having a balanced class distribution. Binary flight cancellation and delay prediction is one example of a classification problem where the issue of imbalanced class distribution needs to be addressed. When considering regular operations, a large majority of the flights are not delayed or cancelled, causing the problem to be imbalanced.

In order to address the limitations caused by data imbalance, many studies use oversampling and under-sampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) [7], and Random Undersampling (RUS) [8]. However, when using these techniques, a $50\% - 50\%$ sampling ratio is most often used, which is not necessarily the ratio that leads to the best performance of the prediction algorithms with

respect to the performance metrics considered. Moreover, the performance metric of choice is usually accuracy, while this may not be the most relevant performance metric for the problem considered.

In this paper a systematic approach is proposed to analyse and deal with the effects of highly imbalanced datasets when predicting flight delays and cancellations. First, the most relevant performance metric for the prediction problem is selected. Then an adaptive sampling methodology is used to determine which sampling technique and which imbalance ratio yield the best classification performance with regard to this metric. This approach is demonstrated using several sampling techniques and classification algorithms, which are applied to data on flights arriving/departing to and from a large, European hub-airport, in the period 2015 - 2019. In addition to flight operational data, weather data from METAR weather reports [9] are considered. To the best of our knowledge, this paper is the first to propose a systematic approach to deal with the inherent imbalance of the prediction of flight on-time performance, when formulated as a binary classification problem.

This paper contributes to the current body of knowledge concerning highly imbalanced datasets and flight on-time performance as follows. From a practical point of view, this proposed approach provides support for air transport stakeholders such as airport coordinators who can use the predictions and the proposed approach to assess flight schedules in advance of the flight execution and take action in order to mitigate the effects of flight delays and cancellations. Second, both flight delays and cancellations are addressed, while existing studies mainly focus on flight delay predictions and not flight cancellations. Predictions for flight cancellations in particular make use of highly imbalanced datasets, which are the focus of this paper. Third, the approach presented in this paper can be used to deal with imbalanced data in other fields of research, when considering binary classification problems.

The remainder of this paper is structured as follows. Section II presents the systematic approach to deal with the inherent data imbalance, including the binary classification algorithms, feature selection and relevant performance metrics, and addresses the classification results. Section III concludes the research by discussing the approach, summarizing the most important observations and providing suggestions for future research directions.

Related work

In recent years, many studies have addressed the flight delay prediction problem using machine learning techniques. Usually, the authors express the problem as a classification task: in [10], the authors predict airline delay on prediction horizons of 5 days, 1 day and 0 days, using Decision Trees, Random Forests, AdaBoost and k-Nearest-Neighbors classifiers. The data is sampled using a combination of SMOTE [7] and RUS [8]. In general, the Random Forest classifier is found to have the best performance, with an accuracy of 0.80. In [11] flight delays are predicted on prediction horizons of 5 months, 1 week and 1 day using Random Forests, XGBoost and Deep Neural Networks. These algorithms make use of airline data, originating from a low cost carrier. The classifiers attain an average Average Under Curve (AUC) score of 0.65 for a horizon of 1 day, with a maximum of 0.75 for certain airports. In [12] flight delay and cancellations predictions are used to rank IATA strategic flight schedules at London Heathrow Airport. The predictions are made using three different classification algorithms, of which LightGBM performs best, attaining a maximum F1-score of 0.60 for the cancellation prediction problem. In [13] deep learning algorithms are used to predict flight delays for airports in the US, several hours before the operation. Weather data is also considered in this study. It is found that the Recurrent Neural Networks architecture results in the most reliable delay prediction: an accuracy of 0.87 is obtained. In [14] an air traffic delay prediction model is proposed that combines multi-class Random Forests and an approximated delay propagation model, which results in an accuracy of 0.87. Additionally, it is found that departure delay and late arriving aircraft delay are the most important features for the prediction. The authors use SMOTE to resample the dataset. Finally, [15] perform multi-class predictions for departing flight delay at Porto Airport, several hours before the flight.

Other studies express the flight delay prediction problem as a regression task. The authors of [16] investigate the prediction of flight delays several months before the operation for US airports. Using Gradient Boosted Decision Trees, the authors find that the model predicts flight delay patterns with a root mean square error (RMSE) of 8.2 and 10.7 minutes for departure and arrival delay, respectively. Next, [17] estimate flight delay several hours ahead of operation using several algorithms, of which Random Forests performs best, with an RMSE of 12.5 minutes. It is concluded that late aircraft delay, carrier delay, weather delay and national airspace delay have the largest effect on on-time performance. Furthermore, [18] perform both classification and regression on the flight delay prediction problem. Classification using the Gradient Boosting Classifier with a combination of SMOTE and Tomek Links [19] yields an accuracy of 0.94 and a recall of 0.91. Regression with Random Forests produced an RMSE of 8.7 minutes. Lastly, [20] combine individual predictions made using Random Forests regression to obtain delay probability density functions for individual aircraft.

The topic of flight cancellation has been approached in varying ways in the literature: both [21] and [22] are studies utilising on-time performance data to propose an accurate decision-support tool, integrating flight delays and cancellations. They apply network models with minimum cost and maximum profit objectives, respectively. The tool returns an optimal set of flights to either delay or cancel. Furthermore, [23] investigate flight cancellation behaviour by using an econometric discrete choice model. The purpose of the research is to identify factors that influence flight cancellations and to predict cancellation probabilities. The results are incorporated in a queuing model, which visualises the effects flight cancellations have on flight

delays. Lastly, [4] analyze the effect of an airline being part of a global alliance on cancellations. It is concluded that airlines belonging to an alliance are likely to have more flight cancellations compared to non-alliance airlines. Complementary to these studies, in this paper the cancellation problem is posed as a binary classification problem.

On-time performance datasets are generally imbalanced, and so are flight delays and cancellation datasets. Regarding imbalance, multiple studies have been carried out on different topics. First, [6] establish an approach to handle imbalanced healthcare data by incorporating multiple different rebalancing techniques. The proposed framework successfully improves the detection of rare healthcare events due to look-alike sound-alike mix-ups. A 45% increase in recall is observed when combining a logistic regression algorithm with SMOTE. Another study on the effects of data imbalance is [24]. Four different rebalancing strategies are presented, combined with a binary classification framework for scientific artifacts in the evidence-based medicine domain. An increase of up to a factor of three in the F1-score of the minority class was found for some of the strategies. Within the field of aircraft on-time performance the most popular approach is to reduce imbalance by sampling with over- or undersampling techniques, such as random oversampling [25], random undersampling [10], [18], [26], SMOTE [10], [14], [18], [27], [28] and Tomek Links [18]. Most studies choose to resample the delayed and undelayed classes, without using a systematic approach to choose the sampling ratio.

This paper aims to elaborate on previous work regarding handling of imbalanced datasets and the prediction of flight delay and cancellation using machine learning, by developing a general approach to handle imbalance in on-time performance datasets.

II. Dealing with imbalance: a systematic approach

In this section, a systematic approach is presented to select an optimal imbalance ratio for an imbalanced dataset in the context of binary classification for flight cancellation and delay. The approach is demonstrated by predicting cancellation and delays with two different classification algorithms and two different sampling techniques, on a one-day prediction horizon.

A. Data description and definitions

In this study, Amsterdam Airport Schiphol (AAS) is considered as the reference airport where flights are scheduled to depart from/arrive at. Two datasets are considered for the proposed prediction algorithms: i) cancelled arrival/departure flights and, ii) delayed arrival/departure flights.

i) Cancelled flights - Highly imbalanced dataset

A total of 1,956,418 arriving and departing flights to and from AAS in the period 2015-2018 are considered. The dataset is based on the strategic flight schedules [12] available in 2015-2018 and contains information such as scheduled date and time of the flight arrival/departure, origin/destination airport of the scheduled flight and the airline that operates the flight. These flights are operated by 256 airlines that fly to/from 649 airports. Furthermore, 54% of the flights have both the destination and origin airport in the Schengen area. Out of all considered flights 1.6% (30,695) are cancelled. Therefore this dataset is considered to be highly imbalanced.

An arriving/departing flight is considered to be *cancelled* if this flight is scheduled to arrive/depart at the reference airport, but it is not operated on the day of the scheduled arrival/departure.

ii) Delayed flights - Moderately imbalanced dataset

The flight delay dataset contains a total of 479,400 arriving and departing flights to and from AAS during 2019. Similar to the cancelled flights dataset, this dataset is based on the strategic flight schedules available in 2019 and contains information such as date and time of arriving/departing flights, origin/destination airport and the airlines that operate the flights. Specifically, the flights are operated by 99 different airlines, flying from 336 unique origin airports and to 323 unique destination airports. This delay dataset is considered to be moderately imbalanced with 34% (82,350) of all departing flights being delayed, and 24% (57,253) of all arriving flights being delayed.

An arriving/departing flight is considered to be *delayed* if during operation, this flight arrives/departs 16 min or more after the scheduled time of arrival/departure.

With regard to imbalance in datasets, the following definitions are introduced. The *imbalance ratio* of a dataset of flights is defined as the ratio of delayed (cancelled) flights to non-delayed (non-cancelled) flights. The *base imbalance ratio* of a flight dataset is defined as the imbalance ratio the considered dataset initially has. Lastly, the *sampling ratio* applied to a flight dataset is defined as the ratio between the amount of delayed (cancelled) flight samples after sampling and the amount of delayed (cancelled) flight samples before sampling.

As an example, a dataset of 100 flights, of which 20 are delayed, has an imbalance ratio of $20/80$, i.e. 0.25. If the minority class is oversampled to a size of 40, the imbalance ratio increases to $40/100$, i.e. 0.40. An imbalance ratio of 100% corresponds with perfect resampling, where the number of delayed (cancelled) and non-delayed (non-cancelled) flights are equal.

Fig. 1 shows the delay distribution of the arriving/departing flights in 2019 at and from AAS. These histograms show that both the distributions of the arrival and departure flight delays are unimodal with positive skew, i.e., the flights are more likely to arrive/depart later than scheduled compared to earlier than scheduled. Also, as expected, the histograms show that the arriving flights generally experience less delay than the departing flights.

Apart from the flight schedule specific datasets, the weather conditions at the origin/destination airports such as the air temperature, wind speed, visibility and pressure at sea level are considered. These data are obtained from METAR [9].

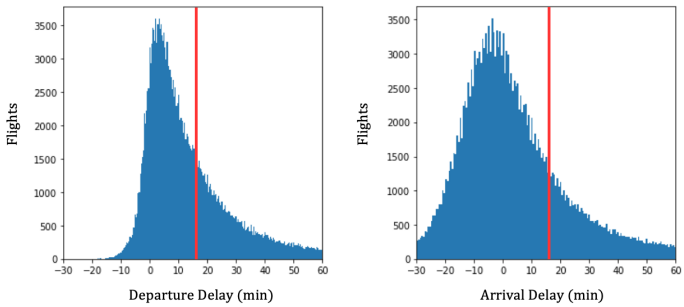


Fig. 1: Departure and arrival delay distribution of flights arriving and departing at/from AAS in 2019. The vertical red line shows the delay threshold of 16 min.

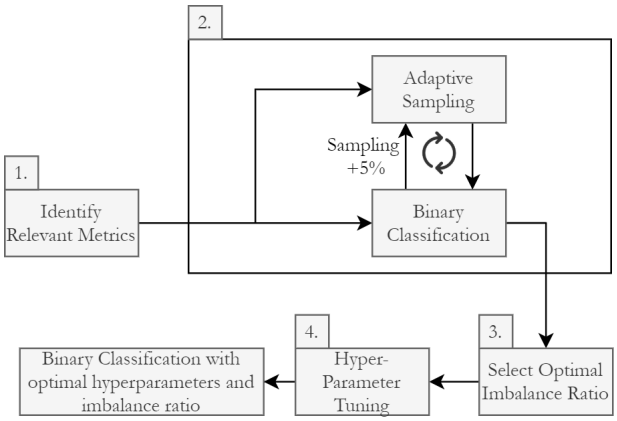


Fig. 2: A flow diagram of the systematic approach to deal with imbalanced data.

B. A systematic approach to deal with imbalanced data for flight delay and cancellation predictions

Given the fact that the flight cancellation and delay datasets are highly and moderately imbalanced, respectively, a systematic approach is proposed to deal with these imbalances when predicting flight delays and cancellations. Fig. 2 shows a schematic overview of the proposed approach. First, the relevant performance metrics for flight delay and cancellation prediction algorithms are identified. Next, an adaptive sampling procedure is iteratively applied to the flight delay and cancellation prediction algorithms. Finally, an optimal imbalance ratio is determined. The available data is sampled such that this imbalance ratio is attained and several binary classification algorithms are run to predict whether flights are delayed or cancelled.

Step 1: Identifying relevant performance metrics

First, performance metrics relevant for the prediction problem are identified. Common metrics for binary classification algorithms are accuracy, precision, recall and F1-score. However, given that the datasets are highly and moderately imbalanced, accuracy is not considered as a relevant performance metric.

Given the specific problem of flight delay/cancellation prediction, in practice it is preferred by airport planners to be able to predict whether flights are delayed/cancelled with a high certainty, even at the cost of mis-classifying many delayed/cancelled flights as not delayed/not cancelled. Otherwise, a low certainty in the flight delay/cancellation prediction may lead to less-informed decisions from an airport planner, which may negatively affect stakeholders such as airlines, passengers, etc. As such, in this study, precision is considered to be the main performance metric (high certainty of predictions), and F1-score as the second most important metric (overall performance of the prediction algorithm).

Step 2: Prediction algorithms and adaptive sampling

In this step, several binary classification algorithms are employed to predict flight delays and cancellations. Below the feature selection and an adaptive sampling approach for these classification algorithms are discussed.

1) Feature encoding and selection: Table I indicates whether each feature is categorical, numerical or time-related. The categorical features are target-encoded. Here, the target-encoded value of a categorical feature is the probability of the flight being delayed/cancelled, based on all samples that fall into the same category [29]. For example, if 20 out of all 50 flights from an airline X are delayed, then airline X is encoded with value 0.4. The time features such as hour, day of week and month are encoded using trigonometric functions that preserve periodicity [11]. Lastly, all feature values are scaled to the interval $[0, 1]$ to eliminate feature domination or ranking [10], [11].

Table I also shows which features have been selected for predicting the departure delay, arrival delay and cancellations using binary classification algorithms. The selection is performed based on Pearson's correlation coefficients. These features are the flight number, the airline operating the flight, the apron handler assigned to a flight at the airport, the aircraft type used for the flight, the aircraft registration number, the airport and country of origin/destination, the number of times an origin-destination airport route is operated per day by all aircraft arriving/departing at/from AAS, the service type of the flight (passenger or freight), the month of the year, the time of day, and, for both the destination and origin airport: the wind speed, gust speed, air temperature, air pressure, visibility and snow presence. Table I shows that the delay classifiers make more use of time features, since busy periods in the flight schedules are causes for flight delay. The cancellation classifiers, however, make more use of weather features such as visibility and snow presence, as they often cause flight cancellations.

2) Binary classification algorithms: The flights are classified as delayed or cancelled using two binary classification algorithms: Random Forests (RF) and Multilayer Perceptron (MLP). Random Forests [30] is a collection of many classification trees which are each constructed using a different

TABLE I: Selected features for the delay and cancellation prediction problems.

Classifier	Features
Departure delay	Flight number[c], Airline[c], Handler[c], Aircraft type[c], Aircraft registration[c], Destination airport[c], Route frequency[n], Month[t], Time[t], Gust speed (origin)[n], Temperature (origin)[n], Temperature (destination)[n]
Arrival delay	Flight number[c], Handler[c], Aircraft type[c], Aircraft registration[c], Origin airport[c], Month[t], Time[t], Gust speed (destination)[n]
Cancellations	Flight number[c], Airline[c], Handler[c], Aircraft registration[c], Origin/destination airport[c], Origin/destination country[c], Service type[c], Wind speed[n], Pressure[n], Visibility[n], Snow[n]

[c] Categorical feature, target encoding
[n] Numerical feature
[t] Time feature, trigonometric encoding

subset of the training set, and using a different selection of features. Each tree carries out a class vote, after which the RF classifies using the majority vote. This approach reduces overfitting and sensitivity to outliers, and enhances the predictive accuracy. The Multilayer Perceptron [31] is a feed-forward neural network with backpropagation, non-linear activation functions and hidden layers. The MLP has the advantage that it can learn non-linear relations. Both the MLP and RF algorithms are well-established and often used in the field of machine learning classification and are therefore fitting to be used in the demonstration of our adaptive sampling approach.

For both algorithms the datasets are split into train and test data, with an 80%-20% ratio. Thus, a 5-fold Cross Validation is used for these classifiers.

3) Adaptive sampling: In this part of the procedure, adaptive sampling is used to investigate the relation between the imbalance ratio of the dataset used for the prediction problem at hand (flight cancellation or delay prediction), and the performance metrics considered relevant for the problem (see Step 1). Adaptive sampling is performed as follows: starting at the base imbalance ratio, the imbalance ratio is iteratively increased by 5%, until it reaches 100%. For each such imbalance ratio, the classification is performed using the two classification algorithms introduced previously and the sampling is performed using two sampling techniques. The resulting values of the performance metrics selected in Step 1, i.e., precision and F1-score, are thus obtained for each imbalance ratio. Lastly, for every combination of algorithm and sampling technique, an optimal imbalance ratio is selected such that precision and F1-score are highest.

The two sampling techniques, used to sample the considered dataset for every imbalance ratio, remain to be introduced. The first is an oversampling technique and the second is an undersampling technique: Synthetic Minority Oversampling Technique (SMOTE) [7] over-samples the minority class, i.e. the cancelled/delayed flights, by creating synthetic samples between samples and their nearest neighbours. When using SMOTE, the samples are not duplicated. Random Undersampling (RUS) [8] undersamples the majority class by leaving out random samples from this class. Both techniques are well-known in literature, and the approach presented in this paper can be extended to different sampling techniques. In summary, for every value of the imbalance ratio, the classification is performed with four different settings: RF sampled with SMOTE, RF sampled with RUS, MLP sampled with SMOTE and MLP sampled with RUS.

Figs. 3 to 5 show the precision, recall and F1-score as functions of the imbalance ratio for the cancellations, departure delays and arrival delays, respectively, obtained using the RF and MLP algorithms and the features as described in Section II-A. The sampling techniques SMOTE and RUS are indicated by S and R, respectively. The models are run with the default hyper-parameter settings as hyper-parameter tuning is performed at a later stage.

i) Cancellations

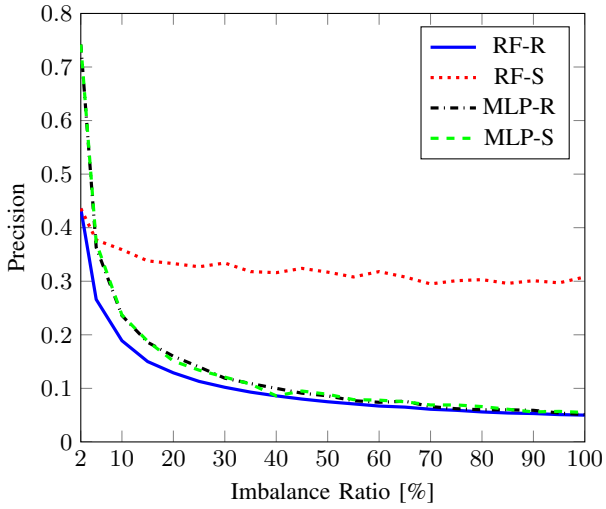
Fig. 3 shows that the precision score is highest at the base imbalance ratio, 1.6%, for all combinations of algorithms and sampling techniques. The precision rapidly decreases with increasing imbalance ratio, until it levels at 0.05. The opposite can be seen for the recall, which starts at a minimum and increases with increasing imbalance ratio. There is a clearly visible trade-off between recall and precision. Finally, the peak of the F1-score is observed near a ratio of 10%. Since the F1-score constitutes the harmonic mean between precision and recall, the peak is observed at an imbalance ratio where neither of the precision and recall attain extreme values. The results also show that RF with SMOTE is insensitive to the imbalance ratios for all metrics.

ii) Departure delays

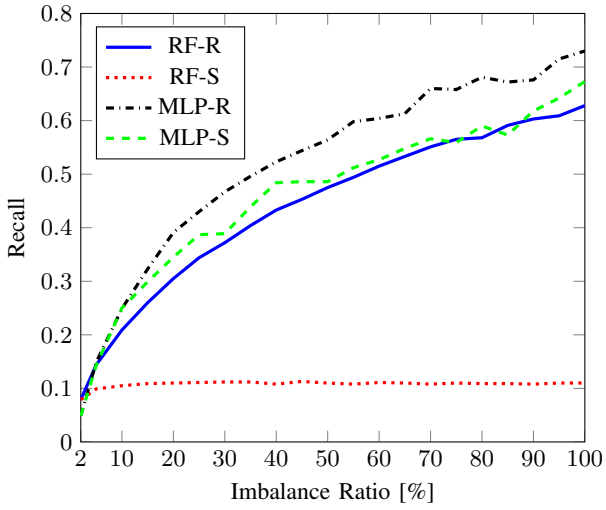
For the departure delays, the imbalance ratio ranges between 55%, the base imbalance ratio, and 100%. The graphs for precision, recall and F1-score are shown in Fig. 4. The general trends are the same as for the performance of the cancellation classifiers, but the performance differences are smaller. Precision decreases with increasing imbalance ratio, while recall increases with increasing imbalance ratio, for both algorithms and sampling techniques. The F1-score also gradually increases with the imbalance ratio.

iii) Arrival delays

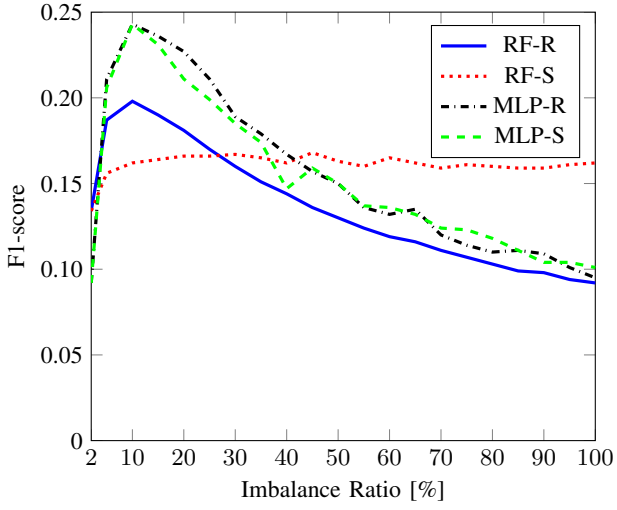
Finally, for the arrival delays, the precision, recall and F1-score graphs are shown in Fig. 5. The base imbalance ratio for arrival delay lies at 33%. Again, there is a clear decreasing trend for precision and an increasing trend for recall, with the F1-score graph corresponding to their harmonic mean.



(a)

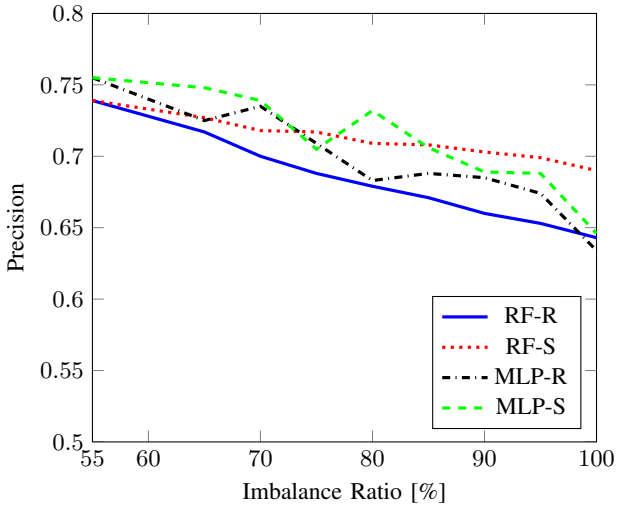


(b)

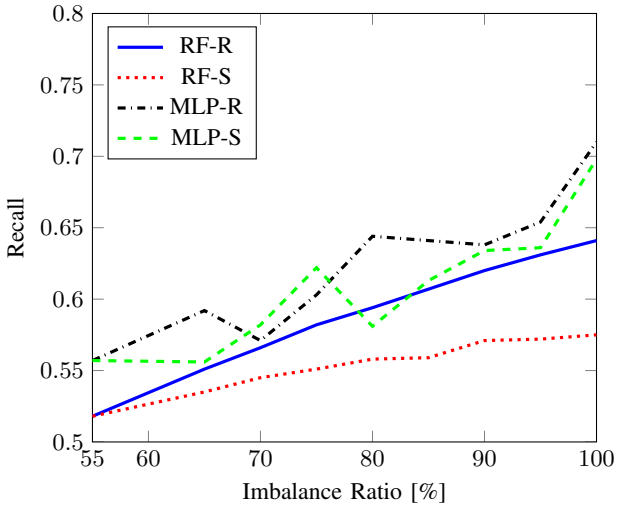


(c)

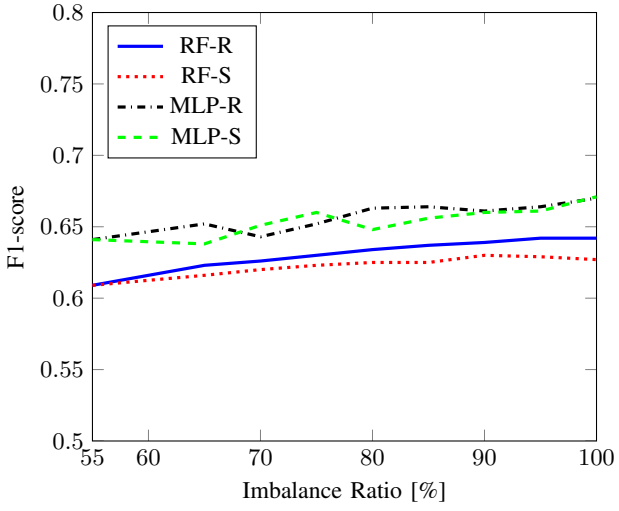
Fig. 3: Precision (a), recall (b) and F1-score (c) as function of the imbalance ratio, for cancellation prediction (RF = Random Forest, MLP = Multilayer Perceptron, R = RUS, S = SMOTE.)



(a)

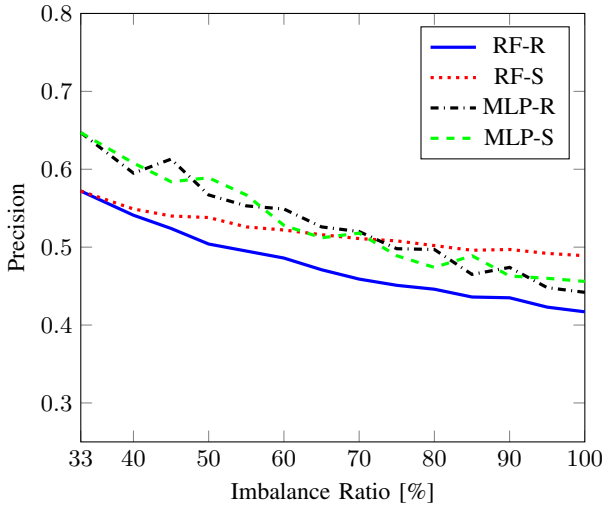


(b)

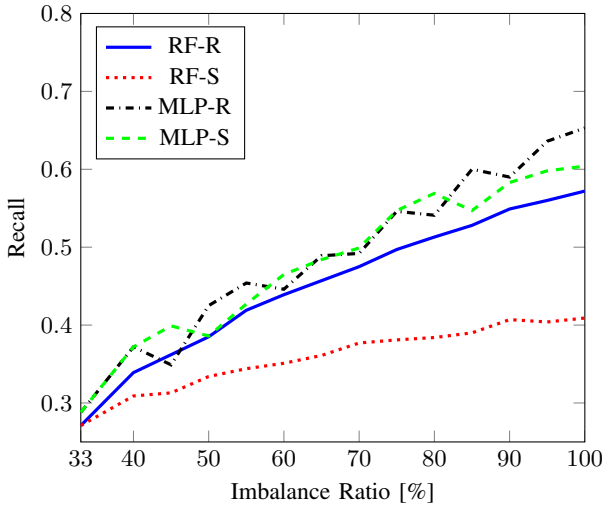


(c)

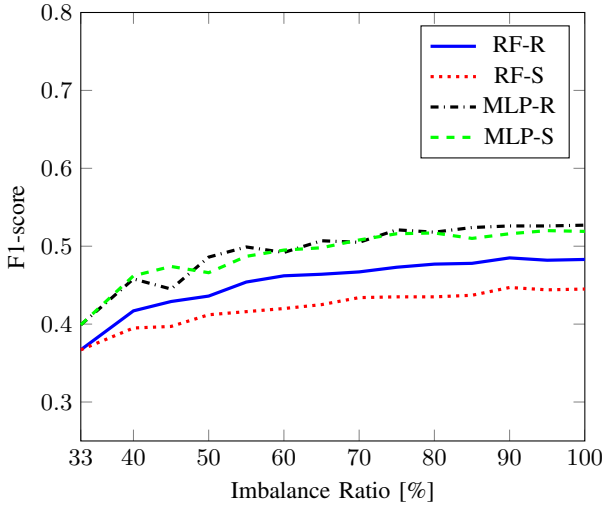
Fig. 4: Precision (a), recall (b) and F1-score (c) as function of imbalance ratio, for departure delay prediction (RF = Random Forest, MLP = Multilayer Perceptron, R = RUS, S = SMOTE).



(a)



(b)



(c)

Fig. 5: Precision (a), recall (b) and F1-score (c) as function of the imbalance ratio, for arrival delay prediction (RF = Random Forest, MLP = Multilayer Perceptron, R = RUS, S = SMOTE).

Step 3: Selecting an optimal imbalance ratio

In this step an optimal imbalance ratio is selected based on the performance achieved in Step 2. As mentioned above, an optimal imbalance ratio is the ratio for which the relevant performance metric value (see Step 1) is highest.

Figs. 3a, 4a and 5a show that the highest precision is attained at the base imbalance ratio, i.e. without using sampling, for both classification algorithms. This shows that at the base imbalance ratio the algorithms only classify those samples as positive that have high certainty of being positive. This leads to a small amount of false positives, and consequently to a higher precision than for greater imbalance ratios. As expected, the large amount of positive samples that cannot be classified as such with high certainty by the algorithm lead to a large amount of false negatives, and consequently to a lower recall.

For the F1-score, the highest performance is obtained as follows. For the cancellation results, an optimal F1-score for MLP is obtained when using a 10% imbalance ratio sampled with SMOTE (see Fig. 3c). An optimal F1-score for RF is located at the 10% RUS imbalance ratio. Considering departure delay results, the MLP achieves the best performance at 100% SMOTE and the RF at 100% RUS, as shown in Fig. 4c. Finally, for the arrival delay results, the highest F1-score is obtained at an imbalance ratio of 100% RUS for MLP and 90% RUS for RF (see Fig. 5c). Due to the greater imbalance in the cancellation dataset a larger range of imbalance ratios is considered for the cancellation prediction during the adaptive sampling procedure. This leads to a larger range of precision and recall values for cancellations (Figs. 3a and 3b), as opposed to the values for flight delay (Figs. 4a, 4b, 5a and 5b). This explains why a clear optimum imbalance ratio appears for the cancellation F1-score near 10% (Fig. 3c), while for the delay F1-score the values are similar for all considered imbalance ratios, and the optimum is less pronounced compared to that of the cancellation prediction (Figs. 4c and 5c).

A summary of these optimal selected imbalance ratios for each classifier is shown in Table II.

Step 4: Performing hyper-parameter tuning

Following the selection of an optimal imbalance ratio, hyperparameter tuning is performed for the flight cancellation, departure flight delay and arrival flight delay classifiers. For the RF classifier, the number of trees, selection criterion, maximum tree depth and maximum features per tree are considered for tuning. For the MLP classifier, the hidden layer size, the batch size, activation function, solver and the learning rate are considered. In all cases, a random grid search is performed. Table III and Table IV show the best hyperparameters for the considered classifiers.

TABLE II: Optimal imbalance ratios corresponding to the maxima in the performance metric plots, for all classification problems and both the Multilayer Perceptron (MLP) and Random Forest (RF) classifiers.

	Cancellations		Departure Delay		Arrival Delay	
	MLP	RF	MLP	RF	MLP	RF
Highest precision	no sampling	no sampling	no sampling	no sampling	no sampling	no sampling
Highest F1-score	10% SMOTE	10% RUS	100% SMOTE	100% RUS	100% RUS	90% RUS

TABLE III: Final hyper-parameters for Multilayer Perceptron (MLP).

		Sampling	Hidden layer size	Batch size	Activation	Solver	Learning rate
Cancellations	Highest precision	no sampling	100 (1 layer)	1000	ReLu	sgd	constant
	Highest F1-score	10% SMOTE	100 (1 layer)	1000	ReLu	adam	constant
Departure Delay	Highest precision	no sampling	100 (1 layer)	auto	ReLu	adam	constant
	Highest F1-score	100% SMOTE	100 (1 layer)	auto	ReLu	adam	constant
Arrival Delay	Highest precision	no sampling	100 (1 layer)	1000	logistic	sgd	adaptive
	Highest F1-score	100% RUS	100 (1 layer)	auto	ReLu	adam	constant

TABLE IV: Final hyper-parameters for Random Forest (RF).

		Sampling	Number of trees	Criterion	Max depth	Max features
Cancellations	Highest precision	no sampling	100	Entropy	10	0.2
	Highest F1-score	10% RUS	300	Entropy	6	1.0
Departure Delay	Highest precision	no sampling	500	Gini	8	0.1
	Highest F1-score	100% RUS	500	Entropy	6	1.0
Arrival Delay	Highest precision	no sampling	100	Gini	6	0.1
	Highest F1-score	90% RUS	300	Entropy	6	0.7

TABLE V: Final performance metric results for cancellation, departure delay, and arrival delay prediction.

		Cancellations		Departure delays		Arrival delays	
	Indicator	MLP	RF	MLP	RF	MLP	RF
Highest precision	Accuracy	0.986	0.986	0.682	0.681	0.768	0.765
	Precision	**0.809**	**0.853**	**0.614**	**0.660**	**0.692**	**0.713**
	Recall	0.041	0.035	0.303	0.203	0.054	0.028
	F1-score	0.079	0.068	0.406	0.311	0.101	0.054
	AUC	0.772	0.850	0.691	0.691	0.680	0.693
Highest F1-score	Accuracy	0.978	0.981	0.666	0.645	0.710	0.640
	Precision	0.263	0.284	0.524	0.493	0.406	0.362
	Recall	0.237	0.198	0.491	0.601	0.528	0.624
	F1-score	**0.249**	**0.233**	**0.507**	**0.542**	**0.459**	**0.458**
	AUC	0.854	0.839	0.679	0.685	0.712	0.700

C. Results - Binary classification for flight delays and cancellations with optimal imbalance ratios and hyper-parameter tuning

Using the obtained optimal imbalance ratios and sampling techniques for each prediction problem and selected metric of interest, the classification algorithms are applied once more to perform the final flight delay and cancellation predictions. The results are summarized in Table V. All results are the mean of a 5-Fold Cross Validation. In this table, "highest precision" and "highest F1-score" indicate that the imbalance ratios have been used that produce optimal results for the respective metric (see Table II). For example, the highest F1-score of 0.507 for departure delays with MLP is obtained using 100% SMOTE.

Table V can be used to compare the performance of the two used classification algorithms, RF and MLP. For the cancellation problem, the table shows that the precision performance of RF is higher than that of MLP when optimizing for precision (no sampling). The opposite is observed for the value of the F1-score when optimizing for F1-score (10% sampling). For the departure delay problem RF outperforms MLP for both metrics of interest. For the arrival delay problem the difference between the classifier performances is smaller and in the case of F1-score the performance is similar, although the MLP does attain a greater accuracy.

Table V shows that the general performance, as illustrated by the F1-score, is better when the base imbalance ratio is larger. When aiming for a high precision, the results show that the departure delay results have the smallest difference between recall and precision, followed by the arrival delay and cancellation results. The trade-off between precision and recall is therefore stronger for smaller base imbalance ratios, as expected.

As shown in Step 3, sampling does not improve the precision in any of the cases. However, for F1-score a clear improvement is observed when choosing an optimal imbalance ratio. For example, when using the MLP classifier, the increase

is 243% for cancellation predictions, 74% for the departure delays, and 354% for the arrival delays, compared to the base imbalance ratio.

In general, the fact that large differences in the classification performance are observed when comparing the precision, recall and F1-score between the different imbalance ratios, confirms the need for a systematic approach to deal with imbalanced datasets regarding the flight cancellation and delay classification problem.

III. Conclusion

In this paper, a systematic approach to deal with highly imbalanced data for binary classification problems is developed, in order to enhance the performance of machine learning algorithms predicting flight delays and cancellations, while taking into account the preferences of airport planners regarding this performance. The presented approach emphasises the need to identify the performance metrics relevant for the considered problem. In the case of predicting flight delays and cancellations, correct predictions are valuable to airport coordinators. The predictions can be used to propose changes to strategic flight schedules. However, the airlines, which are subject to these change proposals, are expected to accept such change proposals only if the predictions have a high certainty. Hence, in this paper the performance metric considered to be most relevant has been the precision, as a high precision implies a high certainty in predictions. Additionally, the F1-score has been considered.

The algorithms Random Forests and Multilayer Perceptron are trained and tested with flight operational data from a large European hub airport and weather data. The imbalance of the data is mitigated by applying an adaptive sampling procedure to the prediction problem using the sampling techniques Random Undersampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE), and investigating its effects on the classifier performance.

The imbalance analysis and its results show that optimal performance with respect to the metrics can be obtained by varying the data imbalance ratios. Optimal precision is shown to be found at base imbalance ratio (data without sampling), for all algorithm and sampling technique combinations. In order to find the optimal F1-score, sampling is shown to be necessary. Increasing the imbalance ratio to the optimal amount improves the F1-score by a significant factor for each prediction problem. In the case of cancellation prediction, the optimal imbalance ratio greatly differs from the ratio corresponding to the conventional resampling (100%).

The proposed approach provides support for major hub-airports to perform on-time performance prediction. Furthermore, the approach can be applied within other research areas when considering imbalanced classification problems. Moreover, the presented approach is not dependent on the type of machine learning algorithm, the features considered, nor on the type of data. Therefore, it is generic and can be applied to any imbalanced binary classification problem.

As future work we plan to develop a systematic approach to deal with imbalanced datasets on which multiclass classification or regression is performed, which use different performance metrics than are used for binary classification. Lastly, we plan to apply our approach in an on-time performance analysis of regional airports.

Acknowledgment

The authors wish to thank the European Fund of Regional Development (EFRD) for partly funding this work under grant number KVW-00235.

References

[1] Eurocontrol. Network manager annual report. 2018, accessed 02-2020.
[2] Eurocontrol. Network operations report 2018. 2018, accessed 02-2020.
[3] Eurocontrol five-year forecast 2020-2024. Accessed on 24-02-2021.
[4] Marco Alderighi and Alberto A Gaggero. Flight cancellations and airline alliances: Empirical evidence from europe. *Transportation Research Part E: Logistics and Transportation Review*, 116:90–101, 2018.
[5] Alice Sternberg, Jorge Soares, Diego Carvalho, and Eduardo Ogasawara. A review on flight delay prediction. *arXiv preprint arXiv:1703.06118*, 2017.
[6] Yang Zhao, Zoie Shui-Yee Wong, and Kwok Leung Tsui. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering*, 2018.
[7] 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.
[8] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: one-sided selection. *Proc. of the 14th Int. Conf. on Machine Learning*, pages 179–186, 1997.
[9] IowaStateUniversity. Asos-awos-metar data download. 2020, accessed 05-2020.
[10] 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.
[11] Yuji Horiguchi, Yukino Baba, Hisashi Kashima, Masahito Suzuki, Hiroki Kayahara, and Jun Maeno. Predicting fuel consumption and flight delays for low-cost airlines. In *Twenty-Ninth IAAI Conference*, pages 4686–4693, 2017.
[12] 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.
[13] 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.
[14] Jun Chen and Meng Li. Chained predictions of flight delay using machine learning. In *AIAA Scitech 2019 Forum*, page 1661, 2019.
[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)*, volume 3, pages 93–98. IEEE, 2015.
[16] 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.
[17] Anish M Kalliguddi and Aera K Leboulluec. Predictive modeling of aircraft flight delay. *Universal Journal of Management*, 5(10):485–491, 2017.
[18] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, and Vineeth Vijayaraghavan. A machine learning approach for prediction of on-time performance of flights. *Proc. of the 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, 2017.
[19] Ivan Tomek. An experiment with the edited nearest-neighbour rule. *IEEE Transactions on Systems. Man, and Cybernetics*, 6:448–452, 1976.

[20] 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.

[21] Jia-Ming Cao and Adib Kanafani. Real-time decision support for integration of airline flight cancellations and delays part i: mathematical formulation. *Transportation Planning and Technology*, 20:3:183–199, 1997.

[22] Ahmad I Z Jarrah, Gang Yu, Nirup Krishnamurthy, and Ananda Rakshit. A decision support framework for airline flight cancellations and delays. *Transportation Science*, 27(3):266–280, 1993.

[23] Michael Seelhorst and Mark Hansen. Flight cancellation behavior and delay savings. In *5th International Conference on Research in Air Transportation*, 2012.

[24] Hamed Hassanzadeh, Tudor Groza, Anthony Nguyen, and Jane Hunter. Load balancing for imbalanced data sets: classifying scientific artefacts for evidence based medicine. *Pricai 2014: Trends in Artificial Intelligence*, 8862:972–984, 2014.

[25] Keshav Ram Chandramouleeswaran, David Krzemien, Kevin Burns, and Huy T. Tran. Machine learning prediction of airport delays in the us air transportation network. *Proc. of the 2018 AIAA Aviation Technology, Integration, and Operations Conference*, 2018.

[26] 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):5, 2016.

[27] Nicholas Bambos and Michael Bloem. Ground delay program analytics with behavioral cloning and inverse reinforcement learning. *Journal of Aerospace Information Systems*, pages 299–313, 2015.

[28] Shon Grabbe, Banavar Sridhar, and Avijit Mukherjee. Clustering days and hours with similar airport traffic and weather conditions. *Journal of Aerospace Information Systems*, 11(11):751–763, 2014.

[29] Daniele Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3, 2001.

[30] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[31] Geoffrey E. Hinton. Connectionist learning procedures. *Machine Learning*, III:555–610, 1990.