Dynamic Airline Booking Demand Forecasting

by

Telmo Luís Eleutério Marquês

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on Friday January 13, 2017 at 2:00 PM

Student Number:	4402189	
Thesis Committee	Prof. Dr. Richard Curran	TU Delft, Chair
	Dr. ir. Bruno Santos	TU Delft, Supervisor
	Dr.ir. Eelco Doornbos	TU Delft

Executive Summary

Nowadays one of the most important asset of an airline is its revenue management system. This system will provide the airline with the right price to sell the right seat at the right time providing a competitive advantage and optimizing profits. No matter how good the optimization algorithm of a revenue management system is, the key for a good revenue management system is a good forecaster. If the wrong information is being fed to the optimizer then wrong information will come out of it.

When reviewing the literature it was identified that there is a lack of new research in terms of new forecasting algorithms applied to the airline industry demand problem while at the same time new forecasting algorithms are having good results in other areas. This lead to a further investigation of the algorithms that could potentially improve the current forecasts as well as the forecasting process itself. Two models were identified: Ridge regression, as an improvement to the commonly used linear regression and Random Forests, a binary decision tree based algorithm that is capable of including categorical features in its modeling.

This work has then 2 main objectives that came from the review and identification of the problem. The first one is to prove that Ridge regression can be an improvement to the commonly used linear regression. The second objective is to prove that Random Forests can effectively capture the influence of categorical features, more specifically the month of departure, day of the week of departure, route and flight number, to automate the data clustering (or pooling) stage of the forecasting while maintaining or increasing the accuracy.

The research was done in partnership with Brussels Airlines. So the methodology was constrained by the context of Brussels Airlines and the data they had available. The context was analyzed and one of the main limitations to come out of it was the absence of data on constraining of bookings. This means that there is no possibility of unconstraining the demand data and as such net bookings was used as the objective of the forecast.

After defining which models were going to be used and which data was available the metrics to evaluate our model were defined. The models were evaluated based on three different metrics. The first two are root mean squared error and mean absolute error which are standard metrics for this type of problems, with the main difference being that root mean squared error heavily penalizes predictions with a larger absolute error. The third metric used was defined in this project and called DINT. The idea behind it is use the standard deviation of the increment in bookings between two points in time to evaluate our error for different time periods. At the end of the project a discussion on the computational times is also given in terms of feasibility.

To test the hypothesis proposed at this work and evaluate the objectives 6 different tests were designed. The first three tests had the objective of optimizing the models to our data. This included parameter tuning for each mode, do feature (or variable) selection and treating the outliers. The last three tests were designed to test our hypothesis.

After running all the tests we concluded that Ridge regression is not an im-

provement to linear regression in this type of problem. The other conclusion is that the proposed solution of Random Forests indeed can improve the forecasting accuracy, specially for forecasts when early into the booking curve, while also automating the process of clustering the data to do those forecasts.

The main conclusions from this study are then that for short-term predictions linear regression is the better performer as it has results that do not differ too much from Random Forests while being much faster to compute. Nevertheless considerations on the automation of the data clustering process should also be taken into consideration here, this saving in time might compensate for the larger computational time. We also conclude that on short-term predictions the most important features are the most recent demand data, with categorical features not having a big effect on the prediction and being ignored by the model. On long-term forecasting Random Forest clearly outperforms linear regression in terms of the defined metrics and we see that here the demand data on current number of bookings is less important but the categorical features given to the model get a much higher importance. An estimation on computational times also shows that despite taking longer to compute, Random Forests is still feasible under the given time constrains.

Acknowledgments

This thesis represents the culmination of my studies at TU Delft and it would not be possible without all the people that helped shaping that path and to whom I am truly indebted.

First of all I would like to thank my thesis supervisor, Professor Bruno Santos, for all the encouragement, guidance and support given during these 9 months where the master thesis was developed. He not only helped me a lot moving forward with the project but also showed complete availability during the time of the work developed to provide the necessary support to solve the all the issues that were arising. His knowledge of the airline industry and mathematics/statistics was essential for the development of this project.

I want to acknowledge my project supervisors, Yorick Buys and Michiel Gilles, from Brussels Airlines, for the insightfulness of proposing this topic. I would also like to thank them for the experience shared on project management and all the "real-world" business knowledge that they provided during the course of these months. I would also like to thank all the people at Brussels Airlines for the way I was received there and for other smaller co-operations to the project.

I would like to thank my brother and all my closest friends that shared with me all these years that will finally lead to my graduation, as they played an essential part in who I am today and so, indirectly, contributed to this work.

Finally, I would like to thank my parents for all the support at all levels that they provided during the course of my studies, for all the guidance and for being there at every important moment of my life.

Contents

List of Figures

List of Tables

Acronyms

BUD Route that connects Budapest to Brussels.

DINT Distribution Interval, developed evaluation metric.

FCO Route that connects Rome to Brussels.

GVA Route that connects Geneva to Brussels.

IAD Route that connects Washington to Brussels.

JFK Route that connects New York to Brussels.

LR Linear Regression mathematical model.

MAE Mean Absolute Error evaluation metric.

RF Random Forests mathematical model.

RMSE Root Mean Squared Error evaluation metric.

RR Ridge Regression mathematical model.

SN Brussels Airlines code.

1 Introduction

1.1 Motivation

With the deregulation of the market with the Airline Deregulation Act of 1978, airlines started having the responsibility to manage by themselves variables such as: number of seats to make available at each stage of the booking process, prices at each stage of the booking process, network capacity, routes to operate and capacity to offer in each route, competition, among others. As such, airlines felt the need for a methodical process to evaluate that data and get a competitive advantage out of it.

The revenue management concept arise out of that need. The main objective of revenue management is to sell the right product to the right client at the right price so as to maximize the revenues. In order to save space for clients willing to pay more, that usually book on a later stage, airlines decline requests of lower paying costumers under the risk of not selling the seat and ending up losing money. To predict then if a certain type of costumer is going to make a request for a ticket is to have a competitive advantage. Lee [36] shows that "10% improvement in forecast accuracy on high demand flights can potentially result in a $10 to $60 million increase in total annual revenue for a major U.S. airline.".

Cleaz-Savoyen [18] and William L. Cooper [68] defend that in the current price-driven market common revenue management system assumptions such as the demand being independent between classes no longer hold true. The authors proceed to show that wrong assumptions can lead to a spiral-down effect on the revenues.

Hence pursuing more accurate forecast tools prove to be quite beneficial.

During the literature review stage of this research it was identified that there is a lack of recent research on forecasting in the airline industry whereas in other industries new algorithms are being widely used with big improvements into the forecasting capabilities. The structure of some of these algorithms permits it to capture the influence of categorical variables (or features) contrary to the current ones which has potential to automate the data clustering stage of the forecasting process. As such, it was proposed for this work to test two different algorithms identified as having good performance in different articles that compare the performance of different algorithms [9], [15], [16]. The proposed algorithms are Ridge regression, a variation of the commonly used linear regression, and Random Forests, a binary decision tree based algorithm.

This work has then 2 main objectives that came from the review and identification of the problem. The first one is to prove that Ridge regression can be an improvement to the commonly used linear regression. The second objective is to prove that Random Forests can effectively capture the influence of categorical features, more specifically the month of departure, day of the week of departure, route and flight number, to automate the data clustering (or pooling) stage of the forecasting while maintaining or increasing the accuracy.

1.2 Organization of the Thesis

The rest of this thesis will be organized as follows.

Chapter 2 Literature Review will show and discuss the research done on revenue management. The chapter starts with a description of the airline booking process and reinstates the need for good forecasts in the airline industry. Then a broad overview of the research on the 4 main topics of revenue management: Overbooking, Inventory Control, Pricing and Forecasting will be given. The research will shift its focus to a more in-depth analysis of the forecasting research by analyzing each of its key-areas: demand arrival, demand behavior, demand volume and demand detruncation. The main issues and areas of interest found are finally discussed at the end of the chapter.

Chapter 3 Research Plan will provide all the structuring of this research. It starts with a description of the research scope and goals. Then the research questions that will lead this research are presented. After that it describes the main hypothesis to be tested during this work and the conceptual model used to model our problem and a proposed framework on how to approach the problem. Finally the chapter closes with a discussion on the innovation elements of this research.

Chapter 4 Brussels Airlines Context describes the current business situation at Brussels Airlines and which implications come from it. The chapter starts by describing the context of Brussels Airlines in the Airline market. We then proceed to give an overview of the current forecasting system implemented at Brussels Airlines and to present the problem from their perspective. To finish the available data and the main implications of that are discussed.

Chapter 5, entitled Methodology, explains the methodology used during the course of this project. The chapter starts by describing the theoretical framework used during this research. Next an outline of the two main mathematical models used during the course of this work is given. This includes an explanation of how each algorithm works, their advantages and disadvantages. Further, the chosen model evaluation methods are depicted. We then finish with a discussion on the main assumptions and limitations of the methodology.

Chapter 6 Case Study describes how the methodology together with the Brussels Airlines context is applied to a specific case study. There the data sets used during the course of this project delineated. This includes all the data that will be used, including which part of the network and for which reason. The chapter will proceed with an exploration of the obtained data where the main found patterns will be discussed. The chapter closes with a description of the tests designed to prove or refute the research hypothesis.

Chapter 7 Results and Discussion will present all the results obtained during this project as well as the discussion associated with those results. We start the chapter by providing a proof of concept for both models used and exploring the parameters that will influence the model performance. After that an automated way for selecting the best features while solving the problem of dimensionality for Ridge regression is presented. Then the chapter proceeds to analyze the outliers and process them. Later an analysis of the performance over time of both models is given. The chapter will then present a comparison between the developed models and the currently implemented models at Brussels Airlines. The chapter will finalize with an analysis on pooling vs non-pooling of the obtained data and how does that influence the performance of both models.

Chapter 8 Conclusion contains the conclusions of the work developed and proposals for future development of this work. The main contributions and results of this master thesis are going to be presented in the chapter and the main limitations of the work developed will be discussed.

2 Literature Study

A revenue management system can be split into four modules: (1) Overbooking, (2) Inventory Control, (3) Pricing and (4) Forecasting McGill and van Ryzin [42]. This chapter will first introduce a description of the airline booking process and the need for forecasting and then proceed to provide an overview of the research done in those four areas with a special focus on the forecasting component.

2.1 The Airline Booking Process and the Need for Forecasts

Airlines have as their main source of revenues the selling of seats in one or multiple aircraft that will connect origin point A to destination point B. Their inventory, seats sold, is perishable. This means that if the airline does not sell a specific seat until the departure date of a flight they cannot sell that product anymore and will be losing money. On top of that the sold seats will increase the ancillary revenues by providing other services to the costumers such as in-flight sales, entertainment, etc..

The inventory capacity of the product provided is limited to the size of the aircraft for a certain flight or to the capacity of the fleet for the whole network. Therefore airlines will be compelled to maximize revenues by taking the highest amount of people at the highest possible price while minimizing the risk of not selling certain seats. Since different people have a different willingness to pay or have different service level requirements, airlines will price the same product differently across the selling period or include different service levels for the same type of seat Belobaba [4]. To balance these variables to obtain the maximum return airlines have in place revenue management systems. Since the performance of the optimization have a big dependency on how good the forecasting was, the forecasting module is central to the overall performance of the revenue management system. The general process of a revenue management system and the forecaster place in that structure can be found on figure 2.1.

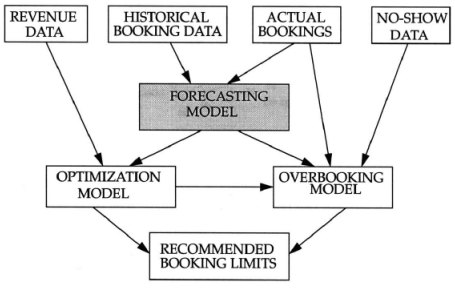

Figure 2.1: Revenue management system process. Belobaba [5]

2.1.1 The Airline Booking Process

Airlines forecast the demand for the seats in each flight and, ultimately, that demand will turn into a certain number of bookings. A potential costumer will go through the airline booking process before having a confirmation of the requested service. That process can be split into three different stages: the reservation phase, the cancellation phase and the boarding phase. A schematic of that process can be found in figure 2.2:

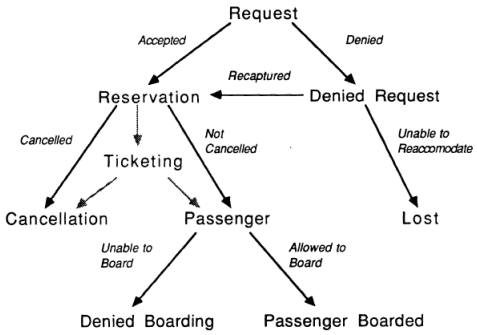

Figure 2.2: Airline booking process. Lee [36]

Reservation phase

The reservation phase starts with a request sent by a costumer to the air transport service provider. The request enters the reservation system and will be accepted or denied depending whether or not the airline still has space for the fare class and time-frame requested. A fare class is considered as a certain level of service offered at a certain price. It is also important to establish the difference between a space and a seat. To correct for no-shows, defined as passengers that miss their flight, airlines have widely implemented the practice of overbooking. Overbooking is the process of selling more seats than the ones available. As a consequence of that, an airline might not have any more seats available for sale but still have selling space when overbooking is taken consideration.

If the request of the costumer is denied there is still the possibility of the provider recapturing that costumer on a different fare class (vertical recapture) or in a flight that occurs in a different time-frame (horizontal recapture), otherwise it is considered as lost Belobaba [5]. If the request is accepted then a reservation happens. This means that the costumer successfully requested a space in a certain flight departing at a certain time for a certain price.

When the costumer pays for the service he requested a ticketing occurs and a ticket confirming that information is issued.

Cancellation phase

The cancellation phase comes next on the booking process. This stage goes from the day a reservation is made to the day of departure. At any point between those two days the costumer or the airline, under the previously agreed conditions, can cancel the reservation. Changes of fare class such as re-bookings

are also considered as cancellations.

Boarding phase

The last phase is the boarding phase happening on the day of departure of the flight. A costumer arrives at the airport and he becomes a passenger if his reservation/ticket was not canceled at any prior point in time. If everything goes smoothly the passenger will be allowed to board the aircraft. On the situation where too many passengers show-up and there are not enough seats some passengers will be denied boarding and will receive the correspondent legal compensation.

2.1.2 The Need for Forecasts

The need for a good forecaster to the airline is two-folded. On a high level and long-term, forecasts are very important tools for the company to take strategic decisions. These decisions can include routes to operate, fleet required to accommodate the expected demand, work-force necessary to support that fleet, etc.. On short-term/medium-term, and the focus of this study, the forecaster is very important as the source of information to the optimizer that will try to maximize the revenues. As mentioned before in this report and shown in figure 2.1 the forecaster is the central part in the revenue management system and, in line with the objectives of this research, the main focus of the literature review.

2.2 Revenue Management

The revenue management problem has been one of the main drivers of airlines ability to thrive in a very competitive market. Different authors have different definitions for what a revenue management system is. In Belobaba et al. [6] they establish that "(...) "Pricing" refers to the process of determining the fare levels, combined with various service amenities and restrictions, for a set of fare products in an origin–destination market. "Revenue management" is the subsequent process of determining how many seats to make available at each fare level.(...)". Despite believing that to be a proper definition, due to pricing being included in revenue management research papers McGill and van Ryzin [42], we will make the following assumption: every time revenue management is mentioned in this report, it refers to the whole revenue management system, which includes the following four blocks: (1) Overbooking, (2) Inventory Control, (3) Pricing and (4) Forecasting. Although pricing will be included as it is fundamentally correlated to the revenue management system, its approach in research is derived from a more macro-economic or strategic point of view and as such will be discussed in less detail.

Revenue management practices gained preponderance with the Airline Deregulation Act of 1978. Airlines were now free to change the prices, schedules, routes, etc. without the approval of the U.S. Civil Aviation Board (CAB). This introduced in the market a lot of flexibility while at the same time increasing the risks and unpredictability of the said market. This introduced the need for more robust and complex revenue management systems in the 80s. American Airlines started the development of the Dynamic Inventory Allocation and Maintenance

Optimizer system (DINAMO) which might be considered the first large scale revenue management system. The deployment of the software was done during the year of 1985. The success of DINAMO was immense with PeopleExpress's CEO, one of American Airlines main competitors attributing their failure to the introduction of DINAMO by American Airlines as stated in Talluri and Ryzin [58].

Revenue management systems are nowadays a fundamental tool for every major airline and most of the smaller ones. Boyd [10] reports that, depending on the carrier the increment in revenue for implementing a full revenue management system is of 2-8%. On another example, Smith et al. [54] presents a case study of American Airlines which were one of the earlier adopters of the revenue management system. This system consisted of three composing blocks: overbooking, discount allocation and traffic management. In his paper he shows a total revenue of 1.4$ billion for American Airlines during a period of three years directly correlated to the usage of revenue management compared to a profit of roughly 800$ million for the same period. This data already let us evaluate the preponderance of a revenue management system in the viability of a carrier financial year. Although revenue management is intimately related to the airline industry, it has recently found its way to other areas of business. Two other industries where the impact of revenue management is of great importance are the hotel industry and the car rental business. The fact that these are the two industries where revenue management is most extensively used besides the air transportation industry is of no coincidence. The three share three important characteristics: the perishable value of the products they offer, the relative bigger size of the fixed costs when compared to the variable costs and the variable nature of demand. Later in this report some research that is applied industries other than the airline industry will be discussed but the main focus will be research directly correlated to the airline industry.

The success of revenue management practice in the airline industry is directly correlated with its needs. A major carrier might deal with thousands of flights per day each with more than 10 different fare structures. Those flights will process between hundreds or thousands of OD pairs. The dimension of the problem together with the unstable nature of the air transport industry market led to the development of automated systematic approaches to solve the issue. In Talluri and Ryzin [58] the authors clarify that the process and the idea behind revenue management is not new, what is new is how the decisions on seat-control, overbooking, etc., are made. The size of the problem related above made it impossible for a human to make those decisions with consistent good results so the decisions are now not taken by a human being but by a computer software that takes into account numerous mathematical models and data inputs to provide what it thinks to be the best possible decisions in an automated way. Human input is still relevant but not the main component anymore. It can be inferred by the above that there were two main key drivers in the success of revenue management systems: scientific advances and technological advances. Research developed new mathematical models to better model demand or economic factors while also providing better performing optimization techniques. Side-by-side with the scientific advancement the technological advancement brought better storage capacity which led to carriers

saving increasingly more data on bookings, costumers, routes, etc., while bigger computational power enables the utilization of more complex algorithms and the possibility to update the calculations made at an higher frequency.

Revenue Management is one area of research where the pressure faced by the industry leads to major carriers handling their own research and development. While it leads to new developments that do not need to be motivated by academic backgrounds, there are some cons associated with this such as the proprietary nature of that research. This issue is discussed in Talluri and Ryzin [58] where the authors mention "()...there are few published reports that document the performance of various forecasting methods in RM applications... often suffer from the proprietary nature of the material, with key details either omitted or disguised.".

Most recent developments and their findings and challenges will be discussed more in detail in the next sections.

2.2.1 Overbooking

Overbooking is the area of research in revenue management system that dates further back in time. The overbooking problem has not only been researched for the longest period but is also the area that as been most extensively covered by researchers. Here in this report the research covered will be grouped by research trend, instead of being presented chronologically. That way it becomes easier to explain how research in the same topic correlates to each other.

The overbooking concept is based on the assumption that some costumers will not show or cancel their flight. Some of the cancellations (that come from premium-fares which implies a return of the fare by the airline to the costumer) are made too close to the departing time leading to a lost of revenue by the airline. Carriers make use of that to overbook, i.e., sell more tickets than the current capacity of the aircraft. This comes with the risk of the carrier having to pay a pre-established monetary compensation to the costumer that gets his service denied. Usually besides the established monetary compensation there will also be a "goodwill" compensation such as a paid hotel night, etc., depending on the situation. A good overbooking policy has then the objective of maximizing revenue by maximizing the sales of seats while trying to minimize the probability of denied boarding.

Following the previous idea, most of the early research in overbooking was directed to control the denied boardings within limits set by external regulating bodies or carriers itself. The first nondynamic optimization model for overbooking was set by Beckmann and Bobkowski [3]. The author tries to fit statistical distributions such as poisson, compound poisson and gamma distribution to histograms of demand to try to estimate the values for no-show, late cancellations and miss-connections. During his research he found that it was difficult to deal with two issues, the first being the unstable nature of the airline industry, i.e. most of the factors that would cause no-shows would hardly maintain stable for a large period of time. Besides that the models did not take into account that the air transportation market was facing a large growth. The second issue that

he faced was that, at a disaggregate level, the flights for which demand was higher than capacity did not have the data on the total demand as the data would be constrained. These are two challenges that are still very present as of today and motivate a lot of research. Beckmann's research was then one of the first extensive research on the overbooking process while also incorporating the utilization of forecasting techniques.

Following that line of research Thompson [59] comes up with a similar research to the one of Beckmann and Bobkowski [3]. The two researches present three major differences: (1) Thompson [59] finds no-shows of little importance due to the rarity of those events and decides to ignore it; (2) Thompson [59] concludes that a binomial distribution would better fit the characteristics of the data, opposing to the notion shown in Beckmann and Bobkowski [3] where he finds that a Gaussian distribution would better fit the data; (3) Thompson [59] uses for his research cancellation data patterns instead of demand patterns as in Beckmann and Bobkowski [3]. It can be concluded that the usage of different data in both studies led to different conclusion and to the differences explained above. Being one of earliest research in the area both studies have some serious limitations. The two limitations that can be deemed more relevant are the fact that it only considers the full capacity of the aircraft and ignores the different classes available, and the fact that they are static optimization processes. While other articles from the following years use similar approaches to the previous papers, with the major change being the statistical models used, Shlifer and Vardi [53] extends that research to new levels. In their research three different criteria are used to try to limit the expected number rejections and expected probability of rejecting a costumer would not exceed certain limits and to maximize the revenue on a given flight. This is already covered in previous research papers and it is not an innovation, although the criteria are slightly different from the one in the previous research. The true innovation about this paper relative to its previous counterparts is the inclusion of two-fare decision process, as well as a two-leg, two-fare decision process. One of the draw-backs of this research is that the data-pool is very small. They only make use of two flight legs for their data: Tel-Aviv - London and London - New York.

Taking down a different road from the articles mentioned before, following his PhD. thesis, Rothstein [49] comes with the first overbooking model that uses the concept of dynamic optimization by using a dynamic programming algorithm. He tries his model using two different optimization characteristics: (1) the probability of denied boardings and (2) the proportion of reserved passengers that had their boarding denied. One of the major drawbacks of his research is the fact of only considering the aircraft capacity has a whole and so ignoring different booking fares. Alstrup et al. [2] solves partially this problem by presenting a model that is capable of solving for two different classes using stochastic dynamic programming. His research assumes that the booking process is a Markovian nonhomogeneus sequential decision process which is the same assumption made in Rothstein [49]. One of the main difficulties faced in their research was the computing time required to obtain results, which is understandable considering the computing power at the time. Those constraints led to simplifications of the model, such as reducing the variable included in the problem formulation, so that the computing time would match the practical industry requirements. Karaesmen and van Ryzin [32] builds up on the two previous papers mentioned by including substitution in his model. He considers

different classes as possible substitutes for one-another. The reason for this only being considered so late in time is explained in Rothstein [49] when he says mentions that Civil Aviation Board at that time expressly prohibited airlines from basing reservation policies on the possibility of reassigning passengers to a different cabin. IATA also made enforced similar constraints for some time. Another issue that this paper comes to solve is the computing times. Karaesmen makes use of a stochastic gradient descent algorithm to significantly reduce the computing time without having the need to simplify the model by decreasing the number of variables used.

2.2.2 Inventory Control

Inventory control is intrinsically correlated to the overbooking problem and some people consider overbooking to be part of the inventory control problem. There is a clear evolution through time from Littlewood [37] work on the rule for two fare classes, to expected marginal seat revenue (EMSR) control for multiple classes proposed in Belobaba [4]. On more recent years the research shifted from single-leg control to segment control and to Origin-Destination control. Following the two main areas of research on inventory control mentioned before, this section will be organized in two main subsections: (1) Single-Leg Seat Inventory Control and (2) Network capacity control.

Single-Leg Seat Inventory Control

Littlewood was the first to introduce a solution to the inventory control problem with the proposed Littlewood's rule Littlewood [37], that discount fare bookings should be accepted as long as their revenue value exceeded the expected revenue of future full fare bookings. His model had 6 main assumptions:

1. The demand of different booking classes is considered to be independent;
2. No cancellation effects considered;
3. The model does not consider batch bookings (groups, etc.);
4. No network effect considerations;
5. The booking arrival pattern will consider low-before-high fares arrival;
6. Sequential booking classes.

Most research followed Littlewood's work with the intent of modeling or softening the effects of the previous assumptions. Belobaba [4] is the first to bring a true innovation and proposes what he calls the Expected Marginal Seat Revenue (EMSR), a decision framework for seat inventory control that considers multiple fare classes instead of just two classes like the previous model. One of the major limitations of the work is that it only considers individual flight legs (ignores network effects). The EMSR model also assumes independent demand between different fare classes and that demand will arrive on a low-before-high order. The principle behind EMSR is that the expected revenue value for the last seat (marginal seat) protected for a certain class is bigger than the fare from the previous class.

The author also tries to model the first assumption, which states that the demand of different booking classes is considered to be independent. The author tries to model the first item by using probability distributions to include it in the full model.

EMSR method has the problem of performing poorly when classes have very similar fares. To solve that Belobaba proposed EMSRb as a variation of EMSR. The main difference between EMSRa and EMSRb is that EMSRb aggregates the demand between classes instead of aggregating protection levels. Both EMSRa and EMSRb are widely used on the the airline industry. A description of the EMSRb model is provided in van Ryzin and McGill [63]. The authors also mention EMSRb to be the most commonly used model in the industry.

In van Ryzin and McGill [63] the authors also try to solve the problem of unsatisfactory demand forecasting processes by trying to eliminate that part of the revenue management process. Instead of having multiple forecasting and optimization cycles, the authors analyze booking policy parameters performance for historical departures and use a adaptive approach to correct those parameters instead of resourcing to forecasting to derive the new parameters.

Curry [20] tries to couple the EMSR model with mathematical formulations that consider an Origin - Destination system. His work shows that is possible to reduce the influence of the assumption that stated no network effect is considered. Nevertheless, the proposed solution is not a full Origin - Destination system as it only considers the inventory to be non-shared between OD pairs, i.e., the allocation of seats to each Origin - Destination pair is made first through a mathematical formulation and only then it proceeds to the seat protection algorithm. S. L. Brumelle [52] proposes the use of joint demand probability distributions for higher fare classes. The author obtains results that would increase revenues by 0.5% comparing to EMSR. Despite that, the test set with which these results were obtained was quite limited to infer if these results would be consistent under various situations.

To try to model factors such as concurrent demand arrivals of customers from different classes, i.e., to drop the assumption that the booking arrival pattern will consider low-before-high fares arrival, GOSAVI et al. [29] proposes a Semi-Markov Decision Process (SMDP) and they solve it using a stochastic optimization technique called reinforcement learning which is capable of working with very high dimension problems. The study also compares the model with the most common EMSR model and the results show the model outperforming EMSR. As the main drawback, it has the fact that the reinforcement learning approach for solving such an high dimension problem will lead to longer computational times when comparing to the EMSR.

Lan [34] follows van Ryzin and McGill [63] by also proposing a model that does not rely on accurate demand forecasting models. The method is based on online methods to define the inventory control policies and depends only on upper and lower bounds of the expected demand. The authors show that it obtains results similar to EMSRa and EMSRb if there is a good estimation of the demand boundaries.

The main drawback of all the papers discussed above is the fact that they do not consider network effects or consider them in a poor way.

Network capacity control

With the deregulation of the airspace and the increasingly importance of Hub-and-spoke networks, a necessity for models that could include network effects in the inventory control problem appeared. In a Hub-and-spoke network passengers with very different origins and destinations compete for the same seats on the aircraft. These different costumers have a different value for the airline as they take different itineraries with different demand levels and fares.

Glover et al. [28] was the first to mention this "new" aspect of the airline business. The author uses Frontier Airlines as a case study where he develops a theoretical model for an airline network where he considers various itineraries and fares and formulated the problem as a minimum cost network flow. The author considers the demand to be deterministic in this problem which is one of the major drawbacks of his work. Another limitation of the early OD models is the fact that they generate non-nested allocations. The research done in Curry [20] already brought some innovations such as the nesting of classes. This paper was already covered in the previous section so it will not be described here.

One approach that was introduced more recently was the bid-price. One of the earliest examples of that approach can be found in Williamson [69]. There they define bid-price as "(...) the marginal value to the network of the last seat available on a given flight leg.". The concept behind the bid-price is to have a clear guiding value to which the airline can relate when deciding to accept or refuse an OD request for a certain itinerary. The bid-price concept has as the main advantage the fact that every request is dealt by a very simple decision rule.

In Feng and Xiao [24] a case with multiple origins and one destination, passing through the hub, is presented. The authors develop a stochastic control model and develops optimal control rules to provide an optimal seat allocation for competing ODs. Future research could extend of their work by formulating the problem for multiple fare classes, time dependent demands and various departure times.

Bertsimas and Popescu [7] propose a alternative to the widely used bid-price model. The author proposed a dynamic programming optimization for the OD problem. To simplify the problem and decrease the computation times the author uses a linear programming relaxation. In the results the author reports that the model is more robust and out-performs the bid-price model without compromising on very big computational times. The model being more difficult to understand by flight-annalists when compared to the bid-price model might be considered as its main drawback.

In Moller et al. [44] they propose a stochastic programming solution with tree scenarios generation. The main advantage of this method is that it does not need to make assumptions on demand distributions. In terms of results the report is scarce as it only tests the model for one specific leg and not on a full OD. The performance of the model in terms of increased revenue is also not tested.

Adelman [1] presents an alternative to the usual static bid-prices policy. The author reports that his approach of using dynamic bid-prices can improve the revenues, specially under less usual conditions. Nevertheless the model is much more computational expensive than the regular bid-price model, which

can be seen as a limitation.

2.2.3 Pricing

This section will be studied in less depth than the others as most of the pricing studies are done on a macro-economical level from a strategic point of view, which is not covered in the scope of this literature study as it is focused on revenue management at an operational level.

Dana [21] describes the market where the airlines implement advance-purchase discounts as having the same pattern described by monopoly second-degree price discrimination. Various airlines provide costumers with a price structure where buying in advance is usually associated with discounts as those lower valuation passengers represent more certain demand. On a later stage the bookings will be done by higher valuation passengers that represent demand with larger uncertainty. The model does not include some other airline pricing strategies such as the discounts associated with the Saturday overnight stay or the round-trip discount.

While the last paper focused on how to model the market considering its pricing strategy, Gallego and Van Ryzin [27] tries to build an optimal pricing model and concludes that under an optimal pricing policy, revenue management does not have a big effect on increasing the revenue. The authors defend revenue management systems are more useful when the pricing policy is not optimal. You [70] proposes a similar approach to the one in Gallego and Van Ryzin [27], he concludes that the main constraint is the extensive computational time. He proposes to use the method only in critical decision stages of the booking process as means to reduce the amount of data used, and so, the computational time associated.

Another concept that is used in pricing, although seldom used in the airline industry, is the concept of auctions. That concept is discussed in Eso [23] where the author discusses an auction process for the extra-capacity flights of the airlines. He proposes a solution using integer programming. The author presents some computational results but they should be expanded more extensively on future research.

Next section will deal with the forecasting component of a revenue forecasting system.

2.3 Forecasting

As explained in section 2.1, the forecasting block of a revenue management system will be the focus of this literature review. The reasons for that choice are the following:

1. Airlines rely on good forecasts to implement their booking policies. Bad forecasts can lead to booking policies that will completely deteriorate revenues on the long run by having values of yield too high and little bookings or a yield that is too low considering the number of bookings. Other areas of business such as catering services, etc., also rely on the aggregate value of demand forecast;

2. It was a topic of interest for the author, and so, a topic the author had interest in researching during his final thesis;

3. The final thesis will be done in partnership with an European Airline, Brussels airlines. The aforementioned airline had also an interest in developing the forecasting module of their revenue management system.

This section will be organized as follows: the section will start by giving an overview of the importance of the forecasting system component inside the complete revenue management system. To follow a similar structure to the one in section 2.2, a brief summary of the initial forecasting research in air transportation market context will be given. The research in forecasting can be subdivided in smaller areas of research, as such, sections 2.3.1 Demand Arrival, 2.3.2 Demand Behavior, 2.3.3 Demand Volume, 2.3.4 Demand Detruncation, of this report will expand on the research on the respective research topics. In section 2.3.5 Findings, the main findings will be explained, i.e., after analyzing the historical evolution and the state-of-art research in forecasting for airlines, some areas that need some improvement or further research will be presented and discussed, setting the stage for the last section 2.3.6 Machine Learning Models, and for what will be the research work that will come from this thesis research.

Both overbooking techniques and inventory control rely on forecasts to produce their results, as such it is of crucial importance that those forecasts produce the minimum amount of error possible. van Ryzin [62] describes forecasting as the following: "(...) it is really the entire system for estimating demand and market response – the data sources, the information technology for collecting and storing data, the various statistical estimations models and algorithms used to process and analyse these data and the infrastructure for deploying model outputs – in short, everything that is required to turn raw data into actionable market information. This is normally called the 'forecasting system' in traditional RM parlance, though forecasting is merely one of its many functions.".

The importance of the forecasting system is matched only by its challenges. The dimension of the problem can be daunting. According to Weatherford [64] bigger carriers have to produce forecasts in the order of the tenths of millions at a time. Those numbers together with the increasing amount of data airlines possess makes it a challenge for airlines to deal with, as they need to find accurate forecasting methods that are feasible in terms of computing time.

The other main challenge in forecasting for the airline market is its variable nature. Listed below are some of the factors that should be taken into account when trying to produce forecasts in the airline environment (based on McGill and van Ryzin [42]):

1. Demand volatility - Demand varies a lot even for flights with the exact same conditions;

2. Seasonality - Including time of the day, day of week and week of the year variation;

3. Special events - Holidays, major sports events, musical events, etc.;

4. Sensitivity to pricing - More and more costumers are driven by price;

5. Demand dependencies between booking classes - costumers can divert to different classes from that they initially intended;
6. Return itineraries - Most costumers will return after a short period meaning an opportunity to sell two tickets;
7. Batch bookings - Bookings of groups;
8. Cancellations - Costumers can decide to not take the their trip at any point before departure;
9. Censorship of historical demand data - demand data will be censored mostly due to capacity constraints of the aircraft or booking policies;
10. Defections from delayed flights - Costumers can ask for refunds from delayed flights;
11. Diversions - Flight needed to change is initial predicted route;
12. Go-Shows - Some costumers try to take a flight without a prior reservation for that specific flight and stay on standby;
13. No-shows - Costumers not showing for the flight without any prior warning;
14. Recapture - Recapture of costumers that the airline failed to sell a ticket on a first instance;
15. Upgrades - Some costumers after buying the ticket decide to upgrade their current service;
16. External factors - Such as wars, major economic changes, fuel prices, etc.

Beckmann and Bobkowski [3] is the earliest paper that includes research on forecasting directly related to the airline industry date to 58 years ago. As its content and contributions were already discusses in section 2.2.1 it will not be covered again here. Most of the important research produced in forecasting during the 60s and the 70s was presented in the annual symposiums of the Airline Group of the International Federation of Operational Research Societies (AGIFORS) which did not enable the access for review. All posterior research will be covered in the following sections.

2.3.1 Demand Arrival

Knowing the arrival process of the demand through time can give carriers important operational information on which to make decisions, and its the basis behind inventory control decisions. The resulting product of a demand arrival model is the booking curve.

Two of the earliest works that can be related to demand arrival were already covered above in this report. Beckmann and Bobkowski [3] used a poison and gaussian models to fit the demand while Shlifer and Vardi [53] tried to fit the data with normal distributions.

The most common way to model demand arrival is to use Poisson processes as a special case of Markovian processes. Poisson processes have a memory-less property, i.e., it only takes into account the latest event. Talluri and Van Ryzin [57] presents an example on how to incorporate a Poisson arrival process together with modeling customer choice behavior. The authors defend that their choice based approach could pave a new path in revenue management although a lot of work is still to be done. Among the future paths this research could take, one would be to expand the model from a single flight model to a network model.

Subramanian et al. [56] proposes a Markov decision process to model demand together with dynamic optimization to solve the revenue management problem. In the paper he considers the problem of a single-flight with multiple fares.

In Stefanescu et al. [55] the authors try to come up with a demand forecasting model where the innovation was the inclusion of the correlation between the demand for different products at the same point of time and the correlation of demand for different points of time. The authors report that the model is capable of capturing some of the proposed correlations efficiently. As future research the model can be extended to include overbooking and cancellations as proposed by the authors.

This section summarized some of the research paths taken in demand arrival research. Most of it makes use of the Poisson processes as a special case of Markovian processes to try to model that arrival of demand. Despite having results that would improve current forecasting system results, these models are not widely implemented in the industry for demand forecasting mostly due to their complexity.

2.3.2 Demand Behavior

With the markets being increasingly more liberalized and at the same time with the exponential growth of access to information through the development of technology, costumer behavior is changing and with it new lines of research opened. The last century saw a shift of focus of forecasting research to demand behavior. Costumers are now better informed and most costumers already take care of their own reservations online instead of going to a travel agency. This allows costumers to compare all their available options within a very short time and without any major difficulties.

Belobaba [4] already approaches a part of this problem in terms of his inventory control system when he takes into consideration the option that lost costumers might be recaptured by the same airline to different classes in the same flight or different flights from the same airline.

One of the main challenges that arises is the growth and change of different business models, particularly the proliferation of low-cost carriers offering unrestricted fare structures, i.e., fares that do not restrict to having a Saturday night stay, specific number of nights before return, etc.. Their business model created products that do not have any differences besides the price. This creates turns the main focus to which price to offer in which period of time.

A curious effect is described by William L. Cooper [68]. The authors describe their spiral-down effect as the process when "incorrect assumptions about customer behavior cause high-fare ticket sales, protection levels, and revenues to systematically decrease over time." This occurs when the assumptions be-

hind the revenue management and forecasting system do not hold leading to a dynamic performance increasingly worse. The authors propose a mathematical framework to analyze the process of forecasting and seat protection optimization by the airlines and observe that some of the models stand on wrong assumptions which lead to deterioration of the revenues. The authors propose that the dynamic behavior of the system should be tested when facing situations where its assumptions do not hold.

The challenge in a market driven by prices is for airlines to avoid a consistent buy-down by costumers with an higher willingness to pay. The same spiral-down effect that can occur in an unrestricted fare environment was studied in a bigger extension by Cleaz-Savoyen [18] in his PhD. thesis. He proposes to different models called Q-forecasting and Fare Adjustment method to deal with this effect. His conclusions are rather positive. Being the proposed solution to include a combination of both of the approaches above. Q-forecasting deals with the fact that the assumption most revenue management systems have of the demand being independent between classes no longer holds. Fare Adjustment method tries to work together with the booking limits optimizer to take into account the sell-up rates. By implementing this the authors claim to improve the revenue 0.63%, taking into account the data used for this simulation.

Reyes [47] tries to expand on the aforementioned research by proposing a hybrid model containing unrestrictred and restricted fare structures. The author assumes that the market contains two main segments, one where costumers are considered to be product oriented and another where costumers are considered to be price oriented.

Neuling et al. [45] takes a different approach from the rest of the research done up to that point as he tries to use the passenger name records (PNR) to make more accurate O-D demand predictions. The second part of their work consisted on using that same PNR to forecast no-shows using a machine learning inductive tree. This approach not only yields good results here but presents a area that can be fruitful in the future with the increasing availability of data.

2.3.3 Demand Volume

Predicting demand volume is important not only to individual flights but also, at an aggregate level, provides valuable insights to managers so they can support their strategical decisions. Besides covering demand volume, this section will include the problem of aggregation vs dis-aggregation of the demand. The research carried on this topic showed a lack of available research on this topic. McGill and van Ryzin [42] states that "Airlines are understandably reluctant to share information about their forecasting methodologies because their revenue management activities are so heavily dependent on accurate forecasting. As far as we know at this time, most disaggregate forecasting systems depend on relatively simple moving average and smoothing techniques augmented with careful analysis of recent booking profiles, as mentioned above.".

As mentioned before, most recent research was focused on demand behavior, leaving the demand volume research focused on the macroeconomic aspect of demand. An example of that can be found in Tobias Grosche [61] where he proposes a two gravity models for estimating demand between city-pairs. The input variables in the model are: air-service level, economic characteristics and geographic characteristics. This model can be used in to make strategic deci-

sions about city-pairs where there is currently no historical data and has the advantage of only considering geo-economic effects for most of it, i.e., does not need to take into account the ever changing competition in the airline industry.

One of the major decisions that need to be taken when considering to build a forecasting system is to decide the aggregation level of the forecasting. Different approaches are required for different objectives. Gravity models are models that consider macroscopic relations between two places, such has GDP, population, among others. Gravity models like Tobias Grosche [61] used are suitable for fully aggregated demand but do not perform as good in a more disaggregated environment. One of the main challenges in forecasting for fully disaggreggated fares is the scarcity of data. Some fare-classes have such a small amount of demand that it makes the forecasting too volatility. This also presents the same issue for low demand routes where demand can be really scattered. Williamson [69] proposes the aggregation of itineraries with similar revenue values. His proposal is done from an inventory control perspective and not from a demand forecasting perspective. It is difficult to apply the same kind of model to forecasting as there would be the challenge to disaggregate the demand back into the many itineraries.

Wickham [67] in his PhD thesis covers a comparison for forecasting methods for short-term forecasting, i.e., he limits the forecasting horizon to 8 weeks. He also includes a historical data-set of only 10 weeks. This data-set was chosen to not be wider so the author could avoid seasonality effects on the data-set. The author thinks that the volatility on the demand seems purely stochastic and as such the data set chosen complies to his needs. There was notwithstanding a demand peak near a special event, an holiday. The demand data that was censored for that period or time was detruncanted. The author in the results discusses that the unconstraining of the booking data helped improving the overall performance of the forecasting models. This conclusions go in accordance to what is discussed in section 2.3.4.

In his comparison he uses the following models:

1. A simple mean of the final bookings
2. A Simple Exponential Smoothing of Final Bookings
3. A regression
4. Classical pick-up model
5. Advanced pick-up model

Both the classical pick-up model and the advanced pick-up model first appeared in the AGIFORS symposiums brought by Ducanson and L'Heureux respectively and as explained in section 2.3 they could not be accessed for review. A brief explanation on both methods is given below.

Classical Pickup Models use booking data from departed flights only. In the table below we can see in gray the data values that are used. Being day 0 the day of departure and week 0 the current week. This means in the current week the flight departed with 39 people and 2 weeks before the flight it had 30 reservations.

Week	Day0	Day7	Day14	Day21	Day28	Day35	Day42	Day49	Day56
-5	25	22	10	5	3	3	2	0	0
-4	30	21	15	17	12	7	3	1	0
-3	23	25	14	9	8	5	5	2	1
-2	40	34	30	16	11	6	3	0	0
-1	35	29	20	12	13	8	3	1	0
0	39	33	30	21	14	6	4	2	1
1	—	28	22	18	10	5	3	0	0
2	—	—	18	11	10	7	4	2	1
3	—	—	—	15	9	8	6	6	2
4	—	—	—	—	11	7	3	2	0
5	—	—	—	—	—	9	8	5	2

Figure 2.3: Matrix of data used for the classical pick-up model. Wickham [67]

To calculate the final number of reservations with the pick-up method let's give the following example: 2 weeks prior to departure we have 18 reservations for the flight as seen in the table. The final number of reservations will be the number of reservations on that week, summed with the difference between the averages of the reservations for closed flights between the day of departure and the day two-weeks before the departure. In this case the final number of reservations would be 18+ average of gray values on column 1 − average of gray values on column 3. Different classical pick-up models differ in the way this average is calculated. For one the average is the simple mean, while other models use exponential smoothing instead of the simple mean.

Advanced Pick-up Models try to make use of all the data, i.e., not only closed flights but also flights that still have their booking process ongoing. The structure behind the table is similar to the one explained for the classical pick-up method.

Week	Day0	Day7	Day14	Day21	Day28	Day35	Day42	Day49	Day56
-5	25	22	10	5	3	3	2	0	0
-4	30	21	15	17	12	7	3	1	0
-3	23	25	14	9	8	5	5	2	1
-2	40	34	30	16	11	6	3	0	0
-1	35	29	20	12	13	8	3	1	0
0	39	33	30	21	14	6	4	2	1
1	—	28	22	18	10	5	3	0	0
2	—	—	18	11	10	7	4	2	1
3	—	—	—	15	9	8	6	6	2
4	—	—	—	—	11	7	3	2	0
5	—	—	—	—	—	9	8	5	2

Figure 2.4: Matrix of data used for the advanced pick-up model. Wickham [67]

Let's consider a 3 week before the flight example in this case. For the

advanced booking pick-up model it will be needed the disaggregate intervals of demand. This values are shown in the figure 2.5. In gray we can see as an example if we decided to use only 4 flight points to make our final reservation prediction. In the advanced pick-up model the pick-up for the 3 weeks before the departure would be:

$$3weekpickup = pickupweek3 + pickupweek2 + pickupweek1 \quad (1)$$

The pick up for each week can again be given by a simple mean or by exponential smoothing depending on the model used.

Week	Pickup in Wk. 1 before departure	Pickup in Wk. 2 before departure	Pickup in Wk3 before departure
-5	3	12	5
-4	9	6	-2
-3	-2	11	5
-2	6	4	14
-1	6	9	8
0	6	3	9
1	—	6	4
2	—	—	7
3	—	—	—
4	—	—	—
5	—	—	—

Figure 2.5: Matrix of incremental reservations. [67]

Wickham shows in his results that the pick-up models consistently beat the other models used. However, the advanced pick-up model lacked robustness and overestimated when the demand turned volatile. His recommendations are that pick-up models are the best option for short-term forecasting. His opinion is supported on the results obtained in his study, as well as the characteristics of the models: they are easy to implement, do not require a great amount of data and are computationally fast to run.

With the lack of research papers that cover the many demand volume algorithms used in airlines and its performances, the focus turned to other similar industries. Weatherford and Kimesb [65] has one of the major studies in comparing forecasting algorithms in the context of revenue management. His focus lays on the hotel industry for this comparison. The hotel industry has similar characteristics to the ones of the airline industry which makes it relevant for this study. Weatherford & Kimesb decide to group the forecasting methods in 3 different groups in the same way as Lee [36]. Forecasting methods are then divided as follows:

Historical Makes use only of historical booking data, i.e., only uses data of flights that already departed

Advanced Booking Uses data from flights that are still open to predict future demand. Can be divided in additive models and multiplicative models.

Additive models assume that the number of bookings in a certain day before departure is independent of the final number of seats sold while multiplicative models assume that the number of future bookings is dependent on the number of current bookings

Combined models Uses data from historical booking that and flights that have not yet departed and combines them.

The following 7 forecasting algorithms were used in Weatherford's analysis:

1. "Simple exponential smoothing, using the smoothing factor α values between 0.05 and 0.95.
2. Moving average methods with the number of periods in the average varying between 2 and 8.
3. Linear regression methods which assumed that there was a correlation between the number of reservations on hand currently (day n) and final number of reservations (day 0) (e.g., Forecast@Day 0)= $a+b*(Bookings@Dayn)$, where Forecast@Day 0 is the forecast at the day of departure while Bookings@Day n is the number of bookings n days before departure.
4. Logarithmic linear regression methods (e.g log(Forecast@Day 0)= $a + b * log$(Bookings at Day n).
5. Additive, or 'pickup', method which adds the current bookings to the average historical pickup in bookings from the current reading day to the actual stay night (e.g., Forecast@Day 0)= $Bookings@Dayn+AveragePickup(DayntoDay0)$
6. Multiplicative method, which multiplies the current bookings by the average historical pickup ratio in bookings from the current reading day to the actual stay night. (e.g., Forecast@Day 0) = $Bookings@Dayn * AveragePickupRatio(DayntoDay0)$
7. Holt's Double Exponential smoothing, using smoothing factors α values between 0.05 and 0.95, β values between 0.05 and 0.95."

Weatherford finds the most robust and better performing algorithms across his set of data to be the exponential smoothing and pick-up methods. This is confirmed by Wickham [67] research where pick-up methods also out-perform the historical methods weighted averages and simple averages. One constraint of this work is the absence of combined forecasting methods. The author mentions this type of methods in the beginning of the report and even states that "Unpublished, proprietary hotel industry methods advocate a weighted average of the historical forecast and the advanced booking forecast. When the day of arrival is far in the future, more weight is put on the historical forecast, whereas when the day of arrival is imminent, more weight is put on the advanced booking forecast." but does not take this kind of methods into consideration in his comparison.

On Lawrence R. Weatherford and Wilamowski [35] we find the first approach to neural networks in an airline context. The main contribution of the paper is the comparison of this forecasting method with the traditional methods used in the industry. The methods used for comparison in is research were the following:

1. Simple moving averages
2. Weighted moving averages
3. Simple exponential smoothing
4. Exponential smoothing with trend
5. Linear regression
6. Quadratic regression
7. Cubic regression
8. Single layer neural network
9. Functional link network

The data used in this paper is a 85 week period with the final number of reservations for a given flight. The author kept weeks 31-85 as a test sample while using weeks 1-30 to train the models. Later he compared the results obtained on the training data and on the test data.

As most of these algorithms are known or explained in other parts of this report, the focus will be to explain the Neural Network used by the author. More details into neural networks will be expanded in section 4. The author uses the most simple Neural Network with just one input layer (with 8 different inputs for 8 different historical points) and one output layer. The Neural network used by the author is shown on figure 2.6

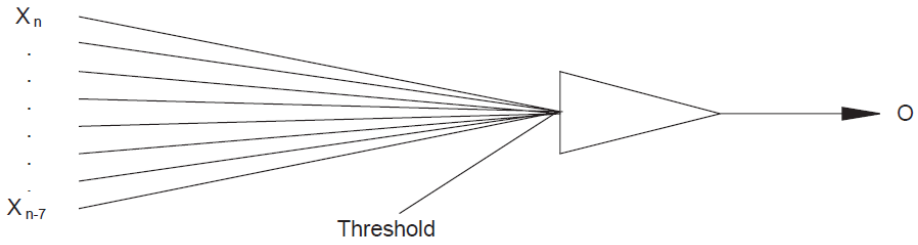

Figure 2.6: Single-layer feedforward network. Lawrence R. Weatherford and Wilamowski [35]

Having no hidden layers, i.e., a third layer between the input and output layer, the Neural Networks used by the author can only produce linear combinations of the inputs. To solve for this problem without growing the network one layer the author tries to use functional links as shown in figure 2.7. This functional links produce non-linear combinations of the inputs.

Weatherford & Wilamowski uses the Mean absolute percent error (MAPE) as an accuracy metric. The results found in his paper can be seen in figure 2.8.

As a first point to be discussed, Neural networks using functional links did produce good results in the training set but failed to generalize for the holdout sample. This can be explained by the model over-fitting the data. The same can be seen to be happening to cubic regression. As it can be seen just a simplest

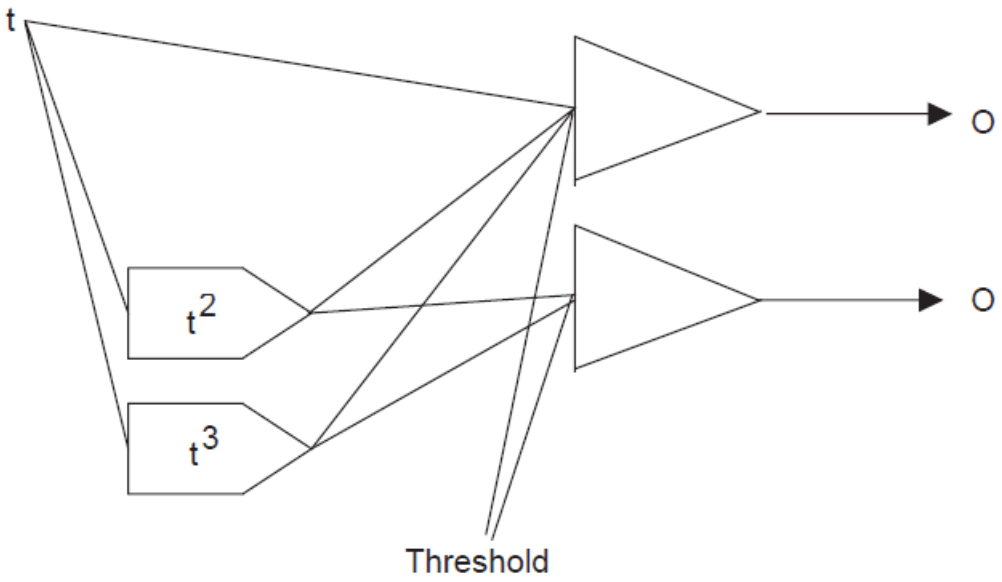

Figure 2.7: Functional link network. [35]

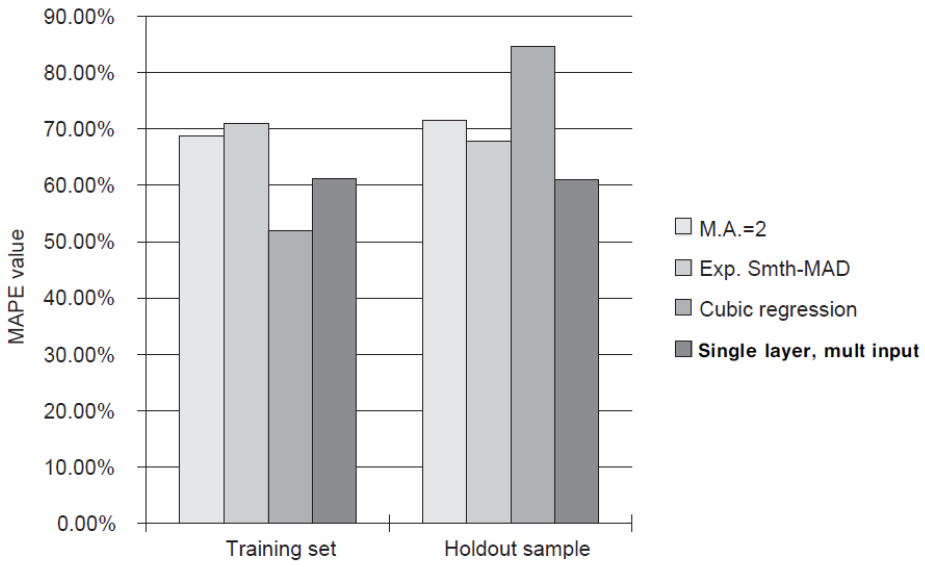

Figure 2.8: Predicted vs actual forecasting error for best of all methods. [35]

neural-network without much work behind it managed to outperform the more traditional models. The author defends that it might be a promising area of research, but no further research is available later on the application of neural networks to airline forecasting.

Talluri and Ryzin [58] mentions the utilization of Neural Networks in the airline industry but does not expand on the topic.

2.3.4 Demand Detruncation

One of the main issues when using forecasting methods that take into account historic data is the fact that the data carriers possess often has periods for which it is constrained, i.e., inventory control techniques lead to some fares being closed for certain periods of time. As airlines do not save any data on denied requests the historical data becomes biased. From the four areas of research related to demand forecast mentioned in this report demand detruncation or demand unconstraining is the least explored by researchers.

Lee [36] was the first to address this issue in a formal way in his PhD thesis. Lee uses what he calls a "Censored Poisson Process" being the idea behind to use a poisson process not only to fit the arrival data but also to model the potential censored demand on that. The model is computational expensive as it needs to be solve the differential equations with numerical methods and is not further developed in the literature nor is it used in the industry according to the information gathered. There is no more important research on the topic until 10 years later. That is confirmed in Bobb and Veral [8] when he describes the following: "As Zeni (2001) illustrates in his dissertation, very little research has been done on un-constraining censored data for revenue management systems. Thus, valuable information is lost which may affect the forecasting model used.". The full dissertation of Zeni could not be accessed so there was no way to review it. Weatherford and Polt [66] confirms the information that we provided above by stating "(...) not one scholarly paper had been written on the topic as it dealt with the airlines(...)", so this is actually the first explicit available paper on the topic of unconstraining. His research is supposed to follow Zeni's research by comparing 6 different unconstraining methods. Weatherford uses simulated data to run the comparison. That simulated data is truncated by the authors of the paper in order to create virtual censored data. This is a good approach as the "virtual true demand" is known, being the data created artificially, whether if real data from an airline would be used it would be impossible to know, and as such, to evaluate the results. Below the methods compared will be listed and an explanation of the mathematical process behind it will be given:

1. Naive 1 or N1: uses all the data for its estimation whether it is closed or open booking data points. It replaces the closed booking data points with the mean of all observations

2. Naive 2 or N2: uses only the open data points to estimate the closed data points. The closed booking data points are then replaced by the mean of the open ones

3. Naive 3 or N3: replaces the closed observation with the largest value comparing the closed observation value and the mean of the open data points

4. Booking profile or BP: will use the mean of the data from other flights to fill the closed points

5. Expectation-Maximization or EM: will be explained in more detail below

6. Projection Detruncation or PD: will be explained more in detail below

The Expectation-Maximization algorithm is a two-step iteration process. First let's suppose we have M + N observations of bookings for a given product, z_1, ... z_{M+N}, of which M observations are constrained because the product was closed. We ignore the time-series aspect of the observations and treat z_1, ... z_{M+N} as an unordered set of observations generated by an i.i.d. process. Assuming the underlying distribution is a normal distribution with mean μ and standard deviation σ. For the M constrained observations we have the booking limits where $z_1 = b_1$, ... , $z_M = b_M$.

If the data was not constrained the complete-data log-likelihood function would be given by:

$$lnL(\mu, \sigma, M+N) = -\frac{M+N}{2} ln2\pi - (M+N)ln\sigma - \frac{\sum_{i=1}^{M=N}(z_i - \mu)^2}{2\sigma^2} \quad (2)$$

And the μ and σ that maximize L() have a close form solution as follows:

$$\hat{\mu} = \frac{1}{M+N} \sum_{i=1}^{M+N} z_i \quad (3)$$

$$\hat{\sigma}^2 = \frac{1}{M+N} \sum_{i=1}^{M+N} (z_i - \hat{\mu})^2 \quad (4)$$

However since M observations are constrained we have to take a two step iterative approach.

Step 0 (Initialization): Let $\mu^{(0)}$ and $\sigma^{(0)}$ be the mean and standard deviation of the unsconstrained observations, these form good candidates to initialize the algorithm:

$$\mu^{(0)} = \frac{\sum_{i=M+1}^{M+N} z_i}{N} \quad (5)$$

$$\sigma^{(0)} = \sqrt[2]{\frac{\sum_{i=M+1}^{M+N}(z_i - \mu^{(0)})^2}{N}} \quad (6)$$

Step 1 (Estimation step or E-step): Calculate the expected value of the censored data in the log-likelihood function, where L() is the likelihood function, assuming that they come from a normal distribution X with parameters $(\mu^{(k-1)}, \sigma^{(k-1)})$. That is, for $i = 1, ..., M$ it should be calculated the following:

$$\hat{Z_i}^{k-1} = E[X | X \geq b_i, X\ N(\mu^{(k-1)}, \sigma^{(k-1)}] \quad (7)$$

$$\hat{Z_i^2}^{k-1} = E[X^2 | X \geq b_i, X\ N(\mu^{(k-1)}, \sigma^{(k-1)}] \quad (8)$$

Next, for each censored observation $i = 1, ..., M$, replace z_i by $\hat{Z_i}^{k-1}$ and z_i^2 by $\hat{Z_i^2}^{k-1}$ to form the complete-data log-likelihood function $Q(\mu, \sigma) = lnL()$. What was done here was to replace the constrained values with their expected

value, given the current estimates of mean and standard deviation.

Step 2 (Maximization step or M-step): Maximize $Q(\mu, \sigma)$ with respect to μ and σ to obtain $\mu^{(k)}$ and $\sigma^{(k)}$:

$$\mu^{(k)} = \frac{1}{M+N}[\sum_{i=1}^{M} \hat{Z}_i{}^{(k-1)} + \sum_{i=M+1}^{M+N} Z_i] \tag{9}$$

$$\sigma^{(k)} = \frac{1}{M+N}[\sum_{i=M+1}^{M+N} (\hat{Z_i^2}{}^{(k-1)} - 2\hat{Z}_i\mu^{(k-1)} + (\mu^{(k-1)})^2) + \sum_{i=M+1}^{M+N} (z_i - \mu^{(k-1)})^2] \tag{10}$$

The decision for leaving the iteration can be done by setting a fixed δ and stop the iterative process when $||\mu^{(k)} - \mu^{(k-1)}|| < \delta$ and $||\sigma^{(k)} - \sigma^{(k-1)}|| < \delta$

Projection Detruncation is in all similar to the Expectation-Maximization algorithm, except in the E step instead of replacing the constrained values by an estimate of the conditional mean, it replaces the values with solution of the equation:

$$\int_{\hat{Z}_i{}^{k-1}}^{\infty} f(x|(\mu^{(k-1)}, \sigma^{(k-1)}))dx = \tau \int_{b_i}^{\infty} f(x|(\mu^{(k-1)}, \sigma^{(k-1)}))dx \tag{11}$$

The Expectation-Maximization method obtains consistently best results, according to Zeni's research. The drawback of this method is the fact of being computationally expensive compared to naive methods as Expectation-Maximization must go through an iterative process before reaching a result. Some of the other issues that this kind of approach presents are, as identified in Weatherford and Polt [66]:

1. The uncostrained estimations still have the risk of "exploding" even with the more robust methods and as such they should be capped at a maximum value established by the designer of the system
2. The research considers independent demand between classes, while this is not true in the real worlds. A closed fare might lead a potential costumer to change to another fare.
3. This methods are considered for a leg-basis but are more difficult to implement in an O-D network, as in that situation it can occur that a particular fare is only partially closed.

Ferguson et al. [25]proposes the use of a double exponential smoothing model that is not covered in Weatherford and Polt [66]. The proposed method obtains similar, although slightly worse, results when compared to the Expectation-Maximization method. This paper have similar limitations to the one of Weatherford. It also only considers a single-leg situation, it is also computationally expensive as it requires an iterative process and it also considers demand between different fare-classes to be independent.

To summarize the little research that exists on this topic, we can conclude that the Expectation-Maximization method looks as the most promising algorithm for unconstraining despite its inherent limitations, with focus on being computationally expensive. Caution is nevertheless required as the algorithm and its effectiveness have never been tested on an industry setting.

2.3.5 Findings

After a careful review of the historical literature on the topic, as well as the most recent and state-of-art literature some of the following conclusions were derived:

1. Weatherford and Kimesb [65] indicate that there has been a dearth of research on new forecasting methods and ideas. There is a shortage of significant research addressing the suitability of various forecasting tools for revenue management applications specially in the airline industry. Despite being from 2003 this remains true until most recent times as research focused on costumer behavior models.

2. William L. Cooper [68], as discussed above proposes that the dynamic behavior of the system should be tested when facing situations where its assumptions do not hold. This is one important issue and one of the main reasons why airlines usually decide to go with more simplified models. Traditional models are robust in most situations and rarely have the problem with having their results exploding to inconceivable prediction.

3. Being a research area that is highly driven by industry needs with carriers facing an ever-changing ultra-competitive environment, a lot of the researched is developed in-house by the airlines. This leads to a lot of the research having proprietary value and not being available.

4. In terms of forecasting for demand volume it is concluded by Lee [36] that combined models work better than advanced models or historical models alone while Weatherford and Kimesb [65] reports the same conclusions from the hotel industry, although he does not include those types of methods in his analytic comparison.

5. The growth of low cost carriers with the unrestricted fare structure associated with their business model brought a change to the market. The traditional revenue management system assumes that demand is independent between classes but this assumption may not hold in this kind of market.

6. Detruncation of demand does not have an extensive research on the topic. Despite that Weatherford and Polt [66] produces a solid comparison between some of the most used statistical models.

Since there is very little research on the utilization of more advanced predictive algorithms to forecast demand, that can be a good area where to expand the research. A big advantage of those types of algorithms is their ability to capture hidden relationships in the data. This kind of characteristic might provide a tool capable of integrating the relationships between the demand of the various classes for example.

Provided this the next section some of these types of models will be discussed such as their advantages or disadvantages.

2.3.6 Machine Learning Models

Some of the most powerful mathematical algorithms used today in the many areas for predictive analysis are machine learning algorithms. Machine Learning is intrinsically related to technology and computer science. The algorithms make use of the ever increasing computational power to process large amounts of data and, through the process of many iterations, return a prediction.

Machine learning algorithms had success in a wide variety of industries, from research engines to computer vision to spam filtering. Machine learning algorithms can fall into three categories:

Supervised Learning The task of generating an algorithm that will predict results with the use of labeled training data. The algorithm will use the labeled training data provided to design a model that can be later used to predict the label of unseen examples. A good supervised learning algorithm will generalize well and avoid over-fitting to the given data.

Unsupervised Learning The task of generating an algorithm that will derive an hidden structure from unlabeled data. Examples of unsupervised learning algorithms are clustering or the Expectation - Maximization algorithm mentioned above.

Semi-Supervised Learning As the name describes, it is in the middle between supervised and unsupervised learning algorithms. This type of algorithms will make use of both labeled and unlabeled to try to obtain a better accuracy.

All this three types provide important insights and different applications. In terms of forecasting there can be various approaches. One important thing to take into consideration is that airlines store a lot of data, specially when we consider past bookings. This stored data can be very important as training data. In this research the focus will be on supervised learning algorithms as those kind of algorithms have a wider range of implementation and success.

Machine learning algorithms can perform really different tasks. The two main tasks are classification and regression. The two different tasks are described below:

Classification These tasks generally try to split training examples into different classes. An example of that can be a travel agency that wants to predict the country to where a new costumer will travel. As training data the travel agency has some characteristics gathered on past clients and the country they visited. That data will be used to train the algorithm and build a model. Next, the built model can be given the characteristics of a new coming costumer and will analyze those characteristics to predict the country that that costumer wants to visit

Regression These tasks take the input features to predict a continuous value. An example of that can be predicting the housing prices. A real-estate agency can have various characteristics or features from a place such as the area, number of rooms, etc., and the correspondent price. That will be used to train a model which will be later able to predict houses prices depending on its characteristics.

As mentioned there are a lot of different possible machine learning algorithms and it is impossible a priori to find the best model for the problem. For the problem of demand forecasting with airline data the focus will be on supervised regression algorithms. The approach taken was to narrow down the research to the algorithms which are studied to have the best results.

In Bowles [9] penalized linear regression methods and ensemble methods are covered.

Penalized linear regression methods differ from the regular linear regression as they introduce a regularization parameter to avoid issues such as possible collinearity between variables, overfitting to the training data and even to perform variable selection. One of the main models of this family is Ridge regression that uses an L2 type of regularization. The general concept behind this is that it will be introduced a parameter to the cost function (function we want to optimize) that will control the weights of the model and the model complexity. All these concepts will be explained more in detail in the methodology chapter of this report.

An ensemble model is the combination of multiple models to obtain a stronger predictor. These combinations can be done with a vast set of different models such as penalized linear regression, neural networks, etc.. In the case of a regression problem the results given by each model are averaged whether on classification problems a voting between all the ensembled models is done. Nevertheless, this is not very simple as different models have to be optimized using different parameters and have very different configurations. For this reason binary decision threes usually provide a good base predictor as you can build a large number of slightly different models easily. Random Forests produces different decision trees by random sampling smaller portions of the data pool. These different decision trees should capture different patterns in the data. As with Ridge regression, here just a general idea of the model was given and all these concepts and the model itself will be explained more in detail in the correspondent methodology chapter.

The author of Bowles [9] considers the two families of algorithms mentioned to be the ones which can achieve the most consistent results and presents some evidence by testing them on different data-sets. Caruana and Niculescu-Mizil [15] arrived to similar results, having as the best performing algorithms methods from the ensemble family. Caruana et al. [16] tried to make a similar analysis but considering the high dimensional problem (high number of features where feature importance is usually sparse). Both for high dimension problems and low dimension problems ensemble methods were the best performing methods, while neural networks variations were right behind it for high dimensions. Taking into account the conclusions from the aforementioned papers this report will focus on penalized linear regression methods, ensemble methods and neural networks. Although Neural Networks did not provide the best results in the studies, the existence of prior research in the area of airline demand forecasting weights in its favor, some of that research can be found in Lawrence R. Weatherford and Wilamowski [35].

After analyzing the literature on machine learning, three main families of algorithms were identified. These families are the following:

1. Penalized Linear Regression

2. Ensemble Methods

3. Neural Networks

Each one has its own advantages and disadvantages. While penalized linear regression models can be trained very fast when in comparison with other models, they struggle to model more complex systems. Ensemble methods are used mainly with decision trees as the base predictor as decision trees make it easier to create hundreds of just slightly different models that are later going to be combined Bowles [9]. These models have decently fast training times and can get a good accuracy for a wide set of problems but have a tendency to over-fit the training data. Gradient boosting and random forests try to minimize the problem of over-fitting. Neural-Networks are usually the model that can get the best accuracy, even/specially in highly complex systems, but have associated training times that do not make it feasible for a real-world application in some situations.

To summarize this literature review: there are 3 different learning processes and our model falls into the supervised learning category; Inside that category the problems can be of regression or classification type depending on our objective, ours is to predict demand and as such falls into the regression type of problem. Later we identified penalized linear regression, ensemble methods and neural networks as the families of algorithms with the better results in some comparison studies and identified the above mentioned models, Ridge regression model from the penalized linear regression family and random forests regression from the ensemble methods, as the most promising ones. Although Neural Networks could also provide good results and has prior research on the area, it will be discarded due to the higher processing times and higher complexity in implementing that turn it difficult to evaluate within the time-frame of the project. The selected models after reviewing the literature will be explained more in detail in chapter 5.

3 Research Plan

In this chapter we will define all the research framework used during this project. First it the chapter will start by defining the scope of this research and the main goals to be achieved during the time of the project in section 3.1. Section 3.2 presents the research questions that guided this project. On section 3.3 the main hypothesis proposed for this research are stated. After the conceptual model and the framework used during the project is presented in section 3.1 and finally section 3.5 describe the main innovation concepts of this research.

3.1 Research Scope And Goals

The main goal of this thesis is to develop the required framework so that the models proposed in this work produce an accurate forecasts of the number of net bookings using real airline data. Net bookings are the difference between reservations made by costumers and cancellations. To reach this goal the following objectives were defined. Net bookings is what actually the revenue management wants to evaluate, i.e., how many people will actually book in each class, with a priority given to higher classes. Demand could also be used but we are constrained by the available data.

The first objective was to identify the models with the larger potential to improve the accuracy of forecasts when evaluated with the selected metrics RMSE, MAE and DINT by analyzing the the current literature not only on airline demand forecasting but also the forecasting literature on other industries. Ridge regression was identified as it is a fast algorithm and is considered an improvement to linear regression, the most used model in practice. Random Forests was chosen due to its potential in using categorical data. This can shift the current approach on the market of running forecasting algorithms in data that was previously clustered together by analysts and a model was fit, for example, per flight number and month, with a lot of manual work, to a model that can identify that segmentation by itself. These models were chosen due to their characteristics mentioned above and for their good results in literature on a wide set of problems [9] [15] [16].

The second objective was to evaluate the performance of the selected forecasting models when compared to the models currently used in practice. This performance is not only measured in terms of accuracy and computation times, but also on a more qualitative approach in terms of complexity, ease of implementation and robustness.

With the results obtained, recommendations as when and under which circumstances the model should be used are going to be provided.

3.2 Research Questions

The research questions and sub-questions for the Master thesis as follows:

1. Which forecasting methods are better established in the airline industry and which forecasting methods could be an improvement to those methods?

2. How do the chosen forecasting methods compare between themselves and between the implemented system in terms of the following aspects?
   (a) Modeling time
   (b) Computational time
   (c) Accuracy of the results according to the selected metrics MSE, MAE and DINT
   (d) Robustness
   (e) Transparency
3. How do different periods of the booking curve influence the forecasting results?
4. Can the models capture the effects of the following characteristics of each flight without the data having to be manually selected?
   (a) Month of departure
   (b) Day of the week of departure
   (c) Flight number associated with the flight
   (d) Route flown

In the first research question the main goal is to identify the state-of-art and practical applications of forecasting methods in the airline industry.

In the second research question, the general characteristics of the forecasting algorithms are addressed. The main goal here is to design a forecasting system using the following algorithms: (1) Ridge regression and (2) Random forests regression. The characteristics of these algorithms will be compared and finally discussed their applicability and limitations of each model. The following characteristics are going to be analyzed: (a) modeling time, (b) computational time, (c) accuracy of results, (d) robustness and (e) transparency.

In the third research question, the forecasting for different time-frames is addressed. The main goal here is to verify at which points of the booking curve does the model perform better or worse and the amount of demand data that would be important to maintain a good forecast.

In the fourth and last research question, the identification of input features is addressed. The main goal here is to feed the models with different inputs, that include as the main one historical booking data but are not limited to it. Other inputs such as route flown, flight number or month of departure are fed to the model and their influence is analyzed.

3.3 Hypothesis to be Tested

Within this research there are two main hypothesis to be tested that are derived from the objectives of this research.

The hypothesis to be tested are:

1. *Ridge Regression is an improvement to the conventionally used linear regression in the problem of airline demand forecasting.*

2. *Random Forests is able to capture categorical features such as month and day of departure, route and flight number to provide automated clustering of data without losses in performance*

These hypotheses will be evaluated during the course of this research are confirmed or dismissed by the end of this report.

3.4 Conceptual Model

The conceptual model is shown below on figure 3.1. This shows our three main influencing variables such as scheduling, route flown and historic of demand which is constrained by the constrain of demand done manually by flight analysts, by a revenue management system or simply because an aircraft is full. Those variables are then treated by a forecasting mathematical model that is subject to pre-defined parameters to give our final demand.

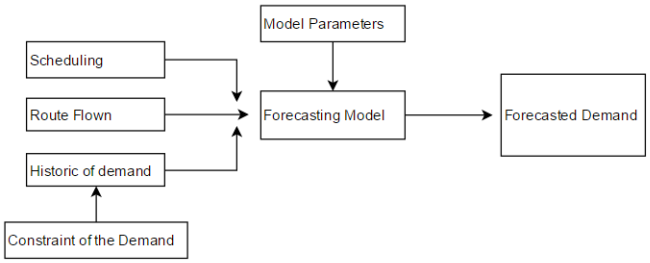

Figure 3.1: Conceptual Model

Below on figure 3.2 we can see the approach developed to obtain the results and that complements the conceptual model shown before and the theoretical framework shown on section 5.1.

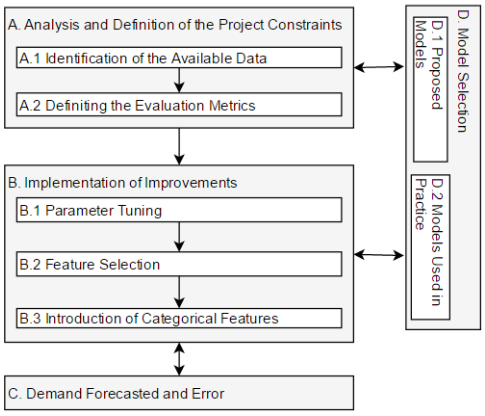

Figure 3.2: Multi-stage approach

This two diagrams give us the necessary info to understand the work developed during this research and the approach taken to obtain the results. This will be nevertheless expanded on subsequent sections.

3.5 Innovation Elements

In this section the innovation elements on this work will be shortly discussed.

This work have two minor innovations and one that is the main thing to come up from this research. The first innovation done is that the variation of the linear regression model (that is commonly used in practice), ridge regression, is used to infer if the regularization parameter included into the model affects the performance of the said model. The second innovation is the use of Random Forests model to try to model demand in the airline industry. While most research focus on the use of the same mathematical models such as linear regression, logarithmic regression and moving averages, only one study uses machine learning models, in this case neural networks [35]. This is the first time that the Random Forests is used for forecasting in this type of situations.

The final innovation of this work and the one with the most practical applications is to automate a part of the forecasting process. Currently most practical systems and research papers consider as input data to forecast the demand on a certain flight, the demand for that same flight in similar periods in the previous years. This usually requires the manual processing of considering which flight in the past are "the most similar" to the flight we want to forecast. Random Forests will make use of categorical features such as month of departure, day of the week of departure, route and flight number to "decide" which flights better model the flight we want to predict, automating the process of pooling flights together, i.e., of manually picking our historical flights.

This chapter gave an overview of the research plan, its goals and scope, the research questions, which hypothesis were to be tested and the concept model that would be used during this project. This section was finalized by summarizing the innovation we pretended to achieve during the course of this work.

4 Brussels Airlines Context

This chapter was designed in order to identify and describe the specific case of Brussels Airlines and its constraints. The chapter starts with a description of the context at Brussels Airlines including a description of their current business operations in section 4.1. Then we proceed to section 4.2 where the two main forecasting models currently implemented are presented. Further, on section 4.3 we identify the problem from the company side and which main objectives are their trying to achieve. Section 4.4 will describe the available data at Brussels airlines and finally section 4.5 will discuss how the available data constrains the subsequent methodology.

4.1 SN Context

The analysis started with an overview of the overall Brussels Airlines network.

As most airlines based on a hub structure, i.e. airlines that have a main airport that operates all their flights, two main types of flights can be found in their network:

1. Point-to-point flights. Routes where the passenger's trip has the hub as origin or destination of their flight;

2. Connecting flights. Routes where the passenger will pass through the airline's hub but the hub is neither the origin nor the destination of their trip.

Exceptionally there are also cases in their African network where they will provide flights between two different African airports but those are not going to be considered due to their exceptional status.

Currently Brussels Airlines is present on the following geographical regions:

1. Short-Haul - European market;

2. Long-Haul - African and North American market.

In the European market Brussels Airlines have routes that are clearly more business oriented while others are leisure destinations. These routes have not only different frequencies and competition but also operate in different times of the year. The different characteristics of the different routes and the sheer size of their complete network lead to picking some routes that fulfill certain defined criteria. This will be discussed more in detail in section 6.1.

Another important point to be considered is that Brussels Airlines currently have implemented an Origin-Destination system. This means tickets are sold to costumers according to their origin and destination airports and not by individual flight legs. This is very typical on airlines such as Brussels Airlines that use a central hub airport for their operations. As an example, on a flight from Brussels to Rome there can be passengers that had their origin in places such as New York, Dakar or London.

Despite the current system implemented selling tickets on a Origin-Destination logic, the data kept on the various databases is only available on a leg basis. This means that we have information on the number of bookings on a certain flight from Brussels to Rome but we have no information on the specific origin

of the various passengers.

4.2 Current Forecasting Method

One of the important steps of the research is to later compare the developed models not only among themselves but with the current implemented system. Brussels airlines forecaster uses as a priority the linear regression model and the logarithmic regression models, this is described in the PROS Revenue Management System user guide which is the manual given by provider of the revenue management software. These two models are going to be described in detail on this section.

PROS Revenue Management System only evaluates the Linear model if a future departure segment class has at least one booking at the last actual DCP. The Linear model excludes historical records with zero bookings. For PROS Revenue Management System to select the Linear model, the model must pass the F-test, which expresses the statistical confidence that the model is better than the Mean model for forecasting demand. Logistic regression is only used if linear regression fails the F-test.

Their current system forecasts on 23 different periods of the booking curve called data collection points.

4.2.1 Linear Regression Model

The linear regression model uses a standard linear equation to fit a line to the historical data to forecast the demand. The equation that describes the final model is the following:

$$y = a + bx \tag{12}$$

Considering that we are forecasting for example between data collection point 3 and data collection point 23 (day of departure), a and b are calculated as follows:

$$b = \frac{\sum_{t=1}^{n} y_t x_t - \frac{1}{n}\sum_{t=1}^{n} x_t \sum_{t=1}^{n} y_t}{\sum_{t=1}^{n} {x_t}^2 - \frac{1}{n}(\sum_{t=1}^{n} x_t)^2} \tag{13}$$

$$a = \frac{\sum_{t=1}^{n} y_t - b\sum_{t=1}^{n} x_t}{n} \tag{14}$$

Where y_t would be the historical net booking for flights at the data collection point 23 and x_t would be the historical net booking for flights at the data collection point 3.

4.2.2 Logarithmic Regression Model

The logarithmic regression model uses the natural logarithm of the last actual class bookings in the linear equation instead of the bookings themselves. This is given by the following equation:

$$y = a + b\log x \tag{15}$$

Following the same path of the liner regression model a and b are defined as:

$$b = \frac{\sum_{t=1}^{n} y_t \log x_t - \frac{1}{n}\sum_{t=1}^{n} \log x_t \sum_{t=1}^{n} y_t}{\sum_{t=1}^{n} \log {x_t}^2 - \frac{1}{n}(\sum_{t=1}^{n} \log x_t)^2} \tag{16}$$

$$a = \frac{\sum_{t=1}^{n} y_t - b\sum_{t=1}^{n} \log x_t}{n} \tag{17}$$

Where, if we take the same example as in the linear regression model, y_t would be the historical net booking for flights at the data collection point 23 and x_t would be the historical net booking for flights at the data collection point 3.

4.3 Problem Definition - SN Perspective

Brussels Airlines is the flagship carrier of Belgium. As any privately owned company their main objective is to generate profit and this is tied directly to the their capacity of selling the right seats for the right price.

From the Brussels Airlines perspective the current forecasting system is too simple to model and adapt to the changing situation in the Belgian market. This change of situation is lead by the proliferation in the Belgian market of low-cost carriers with a special focus on Ryanair. Because of the model not being able to catch up with all these changes, flight analysts are overloaded with a lot of manual work trying to fix over-predictions by the model that take into account only past data and as such do not consider the poor evolution of the current booking curve.

SN objective is to find a solution that will provide better forecasts and reduce the load of the manual work of flight analysts.

4.4 Data Available

The airline industry, alongside with the stock market industry, is one of the industries with highest amount and better quality of data Bobb and Veral [8]. Nevertheless there is still a lot of relevant data that is not captured that could improve the forecasting performance and the fact that is not available limits the obtained results.

Most airlines have different fare classes adapted to their business situation. Each fare class will represent a certain level of service provided to the costumer that will be sold at a certain price as previously stated in this report. Brussels Airlines have currently implemented 25 different fare classes, each one described by a letter of the alphabet. The implemented fare classes are split as follows:

1. Business Fare Classes: J, C, D, Z, P, I, R, O;
2. Economy Fare Classes: Y, B, M, U, H, Q, V, W, S, T, E, L, K, N, G, X, A.

A common situation is that, according to business necessities, those classes can change during the course of time. Some issues that had to be dealt with when treating with the data was not only the changing of the classes available,

but also the changing of the allocation of passengers between classes. As an example, groups would be treated as class N in the beginning of the considered period of data, but mid-way through that period, due to business decision, they started being allocated on class G.

The airline has available the net bookings for each day of the booking period for each of the 25 fare classes, except for weekends where data is not kept track of. This booking period goes from the opening of bookings, one year before the departure date, to the day of departure. These data is composed by flights which have already departed and as such have a complete booking curve, and flights that are yet to depart and so have an incomplete booking curve. Something important to notice is that only net bookings are recorded in that data. Net bookings are the balance between reservations and cancellations. According to Lee [36] only keeping net bookings seemed to be standard practice across most of the airline industry. This means most airlines do not keep an actual track of requests, reservations and cancellations individually.

4.5 Methodology Implications

As described before the only data that is available then is the net bookings for each flight and the departure date of each of those flights. This leaves some space for improvement on additional data that could provide valuable in the future. Besides the separating net bookings into requests, reservations and cancellations other sources of data could be important. To start with special events could be tracked. Special events can mean not only events that can generate a lot of extra traffic such as important music or sport events, but also exogenous events that lead to a disruption of operations such as strikes or weather problems.

Another issue is the nature of the data. The booking data is constrained due to the constrained capacity that an airline has to offer. Frequently fare classes or routes are closed, i.e., they are denying all the requests they receive. This means that the demand data is actually constrained demand. This effect could be minimized by taking into account the periods where the fare classes were closed and use some of the unconstraining algorithms mentioned in section 2.3.4. This is nevertheless not possible as with the current Origin-Destination system there is no data on when a fare class was or not closed during the booking period.

Business choices can also have a big influence in the demand. Major fare changes or frequency chances can lead to a change in demand. Fare changes, by changing the price or service level for a certain number of fares, can lead to relevant variations on demand. Changes in the frequency with which a route is flown can lead to an increased demand per flight if the market is to react as expected, or to a distribution of the same demand to between a higher number of flights if the increased frequency does not lead to the airline capturing a higher market share.

Last but not least is competition data. Proliferation of low-cost carriers in the European Markets is driving down the demand for legacy carriers. An external change in the competitors price or frequency in a certain route can lead to changes in the demand. Although important, this data is very difficult to obtain. Below we can find a summary of all the sources of data not available that could probably prove to be valuable for the forecasting of the demand:

1. Splitting net bookings into requests, reservations and cancellations;
2. Include in the flight data the existence of special events such as holidays, sports events, music events, etc. as well as operation disruptive events including strikes, weather issues, etc.;
3. Closing of fare classes to support unconstraining of the demand data;
4. Changes in the business. This can include for instance big chances in the fare class structure or in the flight frequency;
5. Competition data. The actions of the competitors can bring a new stability to the market. This data is nevertheless the most difficult of all the ones mentioned to get.

These data constraints will constrain the methodology that will be presented next.

5 Methodology

In this chapter we will discuss the methodology used during this research in order to obtain the desired goals. The chapter starts with a description of the theoretical framework used in section 5.1, this is complemented by the section 3.1 where there is also presented a multi-stage approach to the problem. Sections 5.2 and 5.3 describe the models proposed during this research work. Section 5.4 will delve into the metrics that were proposed to evaluate the forecasting models and it finishes by describing some of the main assumptions and limitations of this work in section 5.5.

5.1 Framework

On the figure bellow we can find the research framework of this project:

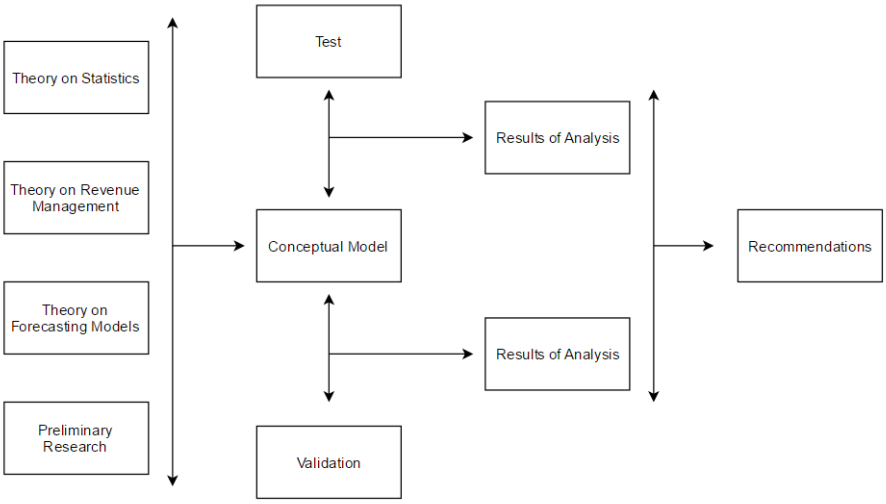

Figure 5.1: Research Framework

A study of the theories of statistics, revenue management and forecasting algorithms, and preliminary research results in a conceptual model, to be used to forecast and evaluate that same forecasting in 2 different data-sets, a confrontation of these evaluations results in recommendations for the improvement of the current demand forecasting techniques.

5.2 Ridge Regression

When two independent variables on the data are highly correlated the least squares solution provides in general bad results. To solve that problem without changing to a different model Hoerl and Kennard [30] proposed Ridge Regression. Ridge Regression builds up into the linear regression model by adding what is called L2 regularization.

Regularization is the process of introducing an extra-parameter to the cost function not only as a means to cancel the effects of highly correlated variables, but also to work as a mechanism to prevent over-fitting. The regularization

parameter, identified in this report as λ, controls the complexity of our function. Explaining some of the mentioned concepts, the cost function is the function that we want to optimize in order to do parameter estimation, i.e., to obtain the parameters of our model. Over-fitting is when our cost function increases a lot when we try to generalize our model to a new set of data drawn independently from the same population, i.e., the model is capturing the noise in the training data as patterns in the data and will fail to fulfill its main function of generalizing to previously unseen data.

There are two main types of regularization for linear regression: L2, Ridge regression and L1, Lasso.

L2 is of the form [31]:

$$R_{L2} = \lambda \sum_{j=1}^{p} \theta_j^2 \tag{18}$$

Where λ is the regularization control parameter, p the number of weights (or parameters, depending on nomenclature) in the model and θ the weights of the model.

L1 is of the form [60]:

$$R_{L1} = \lambda \sum_{j=1}^{p} |\theta_j| \tag{19}$$

Where the various symbols are the same as in the previous equation.

L2 regularization introduces the square of the weights to the cost function, inducing a shrinkage of bigger weights while L1 regularization uses the absolute value. The use of the absolute value of the weights in the cost function when the L1 regularization is introduced will not only do the shrinkage of the weights with the control parameter λ but also perform feature selection by leading some weights to be zero. This is the reason L1 leads very well with very sparse features.

A graphical example with just two weights (or parameters) of the difference between L2 and L1 is given in figure 5.2[1].

In the Ridge Regression algorithm the objective is to minimize the following equation [14]:

$$minimize J(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta x_i)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \tag{20}$$

Where $J(\theta)$ is the cost function, n the number of examples in our training data, y_i is the target variable for each training example (in our case the net bookings), x_i is the input feature vector for each training example, θ is the vector of the weights, λ is our control regularization constant and p is the number of weights in our weight vector. The last term as explained above is the regularization term with the function of shrinking the weights and subsequently reduce over-fitting and function complexity.

To solve the cost function, which in this case the objective is to minimize it, we use an iterative optimization algorithm called gradient descent. In the

[1] *https : //en.wikipedia.org/wiki/Regularization_(mathematics)#/media/File : Sparsityl1.png*

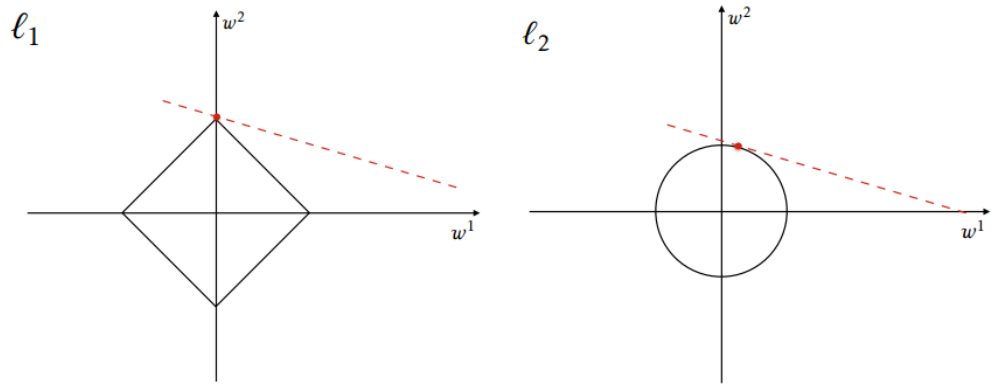

Figure 5.2: L1 regularization vs L2 regularization with just 2 weights in the model.

gradient descent solution each θ is updated in iterative way according to the following equation, with j being the number of features in our x vector and α an optimization constant:

$$\theta_j := \theta_j - \alpha \frac{\delta}{\delta \theta j} J(\theta) \tag{21}$$

This becomes:

$$\theta_j := \theta_j - \alpha \frac{2}{n} \sum_{i=1}^{n} (y_i - \theta x_i) x_i^j + \frac{\lambda}{n} \theta j \tag{22}$$

The iterative process stops after a pre-defined number of iterations or when the cost function is varying less than a pre-defined threshold. The system has a closed form solution given by the normal equation [14]:

$$\theta = (x^T x + \lambda I)^{-1} x^T y \tag{23}$$

Marquardt and Snee [40] studied the how would ridge regression would perform in practice and concluded as well that ridge regression is superior to the least square method. This last equation shows an easy implementation of the Ridge regression model where if we have our feature matrix X and our target variable vector Y we can calculate the solution for the weights that optimize the cost function using algebra. Having these weights θ, i.e., having the parameters of our model, we then can use them to make predictions on new data by having:

$$Y = X\theta \tag{24}$$

This will give us our prediction final prediction.

5.3 Random Forests

The Random Forests algorithm was introduced in Breiman [12]. As with the introduction of regularization in linear regression models, Random Forests was

introduced with the main objective of avoiding the over-fitting inherent to decision tree models.

Random Forests is an ensemble model. An ensemble is in reality not a single model but a combination of multiple models. Combining the results given by multiple less complex models can lead to better results than doing just one very complex model to fit our data. As an example let's consider a classification task where we have 3 classes which we call red, blue and green. Consider we have 3 different models which all obtain an accuracy of 66%, the models are built differently in two ways: (1) by fitting the same model to different samples of data drawn from on the same population or (2) by using different models to fit the same data. While probably all the 3 models in our example classify most examples as the same class, for instance, blue, there will be classes where different models give different classifications. In this case a voting will be used. This usually increases the accuracy of the result by combining these results from multiple sources. Decision trees allows to build a big number of models just slightly different by using the first strategy mentioned above. One of the most used base predictors in decision tree ensembles is the binary decision trees [9].

Decision Trees

Induction of decision trees concept was explained in terms of the classification problem in Quinlan [46]. The models use a top down approach where for each value a feature can assume the model will produce a split point for each feature taking into account the distribution of features. It is considered a top-down approach as we first consider higher level splits that will maximize the reduction of the loss. On a binary decision tree all internal nodes have exactly two children. To explain better all this process we have as an example of a binary decision tree figure 5.3.

While this image is taken from a wine evaluation model where the evaluation is given as a grade from 0 to 10. What the model first does is to run all the different features (in this case let's consider we have 10 features) and all the possible splits considering the interval of values for each feature. The mean squared error is considered for each of those possible splits and the split that minimizes the cost will be the one chosen. Here its feature 9 from the feature vector X that is first split. With the same rationality we see that its feature 1 on the left branch and feature 1 and 10 on the right branch that provide the better split of data. On the last level we have the predicted value if we follow each branch of the tree.

The complexity of the three would be much bigger if we varied the stopping condition of the algorithm. As we see most leafs on the last level have a high number of samples, generally the stopping condition will be to stop only when a branch as a minimum of 2 samples for instance. This will provide more complex trees but that also can capture a wider range of patterns in the data. While a very complex decision tree can lead to over-fitting the data, models such as Random Forests discussed below will solve that problem

In the case that there are multiple options to model the data-set equally good, the most simple decision tree should be prioritized. This not only improves computation times but also makes the model more interpretable.

Here we provide a more formal approach to decision tree modeling. To model a decision tree the following general steps are required:

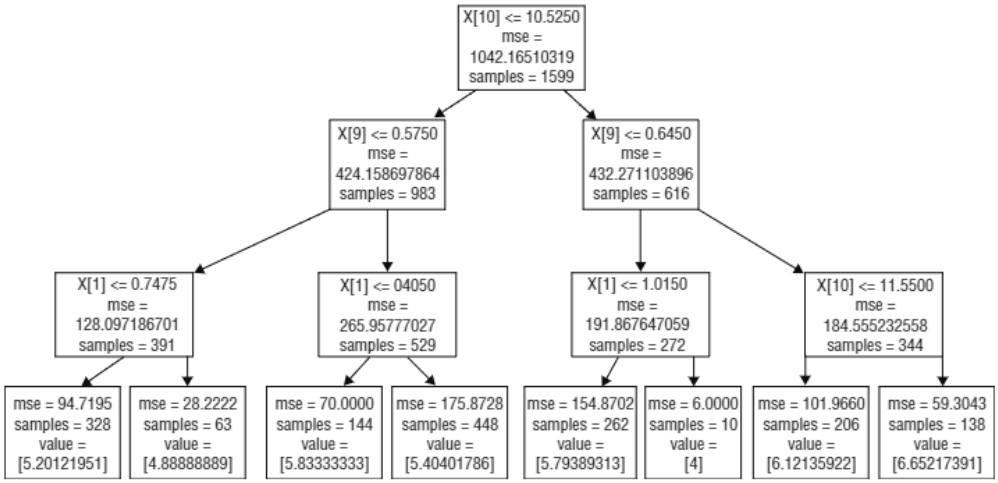

Figure 5.3: Decision tree example. [9]

1. Iterate through the variables and find the best splitting point for each variable;
2. Find the variable that after splitting minimizes the objective function;
3. Repeat for each "children", i.e., the new nodes of the decision tree, until the stopping condition is met.

The first step is then to define what the best split is. Decision tress can be used in classification or regression problems. In the regression problem the best split is going to be defined by the mean squared error (MSE). The model is going to take into account all the possible splits for each feature and then choose the feature and the split that minimize the MSE. As an example, if feature x_1 has the following set of values [1,4,6,7]. The model is going to do 3 different splits and test each one of those splits:

1. x<=1 and x>1
2. x<=4 and x>4
3. x<=5 and x>5

The stopping condition is one of the pre-defined parameters of the tree. It can be either the minimum number of samples per split or the minimum of samples per node.

Ensemble Methods

As explained before random forests is an ensemble method. Random forests model will generate multiple independent decision trees that will later be combined to provide a stronger forecaster [12]. While there are different techniques to build different decision trees, Random Forests does this by training decision

trees on different substets of data. This is a random sampling of a portion of the total data with replacement. In the regression type of problems all these decision trees will be combined by averaging the final result.

The way that was found to create some variability between decision trees and subsequently diminish over-fitting and variance was with a technique called bootstrap aggregation or "bagging". This method was proposed in [11].
Taking a set L with n training examples, the usual approach is to use those n training examples to train a model that will then later be used to obtain predictions when facing new sets of data. In bagging, m sub-sets $L^{(B)}$ will be taken from the main set L by random sampling with replacement. Those m sub-sets will then be used independently to build m different predictive models

$$y_B = \phi_B(L^{(B)}) \tag{25}$$

For the regression problem the final predictive model will be given by;

$$y = \phi(L) = \frac{1}{m}\sum_{i=1}^{m}(\phi_B(L_i^{(B)})) \tag{26}$$

Where ϕ is our predictors, i.e., our different decision tree models trained on different sub-sets of data. The author shows in his paper for 5 different data-sets that the method brings an error reduction between 20% and 50% for the subsets considered Breiman [12]. This is also confirmed by other independent authors. According to [15] and [16] random forests is one of the best performing methods in a big range of problems.

5.4 Error Estimation

To evaluate the performance of the algorithms under different situations and parameters and to compare how different models perform, evaluation metrics have to be chosen. This section will provide a small summary on the evaluation metrics that are going to be used during this study and the reasoning behind that.

The main objective of this work is to forecast the net bookings. We consider net bookings to be a continuous variable. Although in reality there are no 0.5 net bookings we will consider the target variable of our forecasts to be continuous and will round the result to the closest integer after the forecasting process. Being a continuous variable, as defined in 2.3.6, we are in presence of a regression problem.

Regression problems adopt certain specific types of metrics that adapt specifically to this problem. Two of the main ones that are used across different forecasting works such as Lee [36], Weatherford and Kimesb [65], Lawrence R. Weatherford and Wilamowski [35] are the root-mean-square error (RMSE) and the mean absolute error (MAE).

Both measures have different characteristics. RMSE punishes more bigger errors putting a bigger weight on those errors while MAE does not and considers all errors similar.

RMSE is given by:

$$RMSE = \sqrt[2]{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}} \tag{27}$$

Where n is the number of elements in our sample, $\hat{y}_t$ the forecasted number of net bookings and y_t the real value of net bookings.

MAE is given by:

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |\hat{y}_t - y_t| \tag{28}$$

Where n, $\hat{y}_t$ and y_t are the same as in the previous example.

A third metric that is used sometimes in regression problems is the mean absolute percentage error (MAPE). This metric was not used due to its associated problems such as situations where a division by zero occurs.

As such we used a third metric which we will call DINT as it is related to the distribution interval. When forecasting subsequent values in the booking curve the acceptable values of MAE can be very different if we are in the beginning of the booking period where the standard deviation is almost null, to when we are near departure date. For that reason this third metric is going to be used. The objective of this third model evaluation method is to compare the forecast for different times in the booking curve.

Assuming the increment in the number of net bookings for two different periods in the booking curve follows a normal distribution, we will calculate the standard deviation from that increment and see how does the MAE compare with the standard deviation for that certain period. This will give us the z value that we will use to calculate our DINT. The DINT will be the percentage of cases that would have a larger error than our currently obtained error. The z value is calculated has:

$$z = \frac{MAE - \mu}{\sigma} \tag{29}$$

Where MAE is our mean average error, μ is the mean which in this case will be considered 0 and σ is the standard deviation.

To give some examples, imagine the case where the standard deviation is 1. If the MAE is 0 our z value is 0 which will return that 100% of the cases should have a similar or bigger error than what we obtained. If the MAE is 1 our z is equal to 1 we obtain that 31.7% of the cases should have a similar or bigger error than what was obtained, and so on.

5.5 Main Assumptions and Limitations

To finalize this chapter and before passing to the presentation of the results we summarize here the main assumptions taken during this project and some of the limitations:

1. No yearly trends were considered and there is no possibility to test how the models react to them as only one full year of data is considered due to time consumption of processing data from the server;
2. Due to data constraints only flight legs are considered, as such, we assume the model is able to capture the network effects;
3. The model is predicting net bookings as there is no data available on the closure of fare classes to unconstrain the demand;
4. Individual classes and sub-classes were not modeled, the forecasts are done on a flight level of aggregation;
5. It is assumed that the extrapolation done to fill the missing data will not lead to bias of the results;
6. We assume that group or batch bookings will not constrain the performance of our model.

The implications of these assumptions and limitations is going to be discussed after the results are analyzed. In this chapter the methodology for the work developed was presented and discussed. The data available and its main constraints were identified. Next chapter will present the results achieved using the described methodology.

6 Case Study

In this chapter we describe the particular case study used and how the methodology applies to it. The chapter starts by giving a small description of the routes that are going to be used for this research and the fundamentals behind that decision in section 6.1. Then sections 6.2, 6.3, 6.4 will deal with the data used in this project, from describing the data sets used and how was the data processed to a small exploration of the said data. This chapter ends in section 6.5 with a description of the tests designed in order to attain the research goals.

6.1 Routes Analyzed

Brussels Airlines has a wide set of routes in the markets referenced in 4.1. To test the models, as the whole set of routes was too big to investigate, a set of routes that would be representative of the full Brussels Airlines network was chosen. The routes chosen would have certain characteristics that were identified as relevant. Below we give the list of the five routes analyzed during this work as well as the criteria those routes fulfilled.

1. Route gva ; BRU-GVA ; Very high frequency short-haul route. Business oriented route;
2. Route fco ; BRU-FCO ; Short-haul route with the highest level of competition: Leisure oriented;
3. Route bud ; BRU-BUD ; Short-haul route without relevant competition;
4. Route jfk ; BRU-JFK ; Long-Haul with competition;

5. Route iad ; BRU-IAD ; Recent Long-haul route with scarcity of data.

The airport acronyms above are the following: BRU is the Brussels Airport in Brussels; FCO is the Leonardo da Vinci–Fiumicino Airport in Rome; BUD is the Budapest Ferenc Liszt International Airport in Budapest; JFK is the John F. Kennedy International Airport in New York and IAD is the Washington Dulles International Airport in Washington.

From here onward the routes will be described in this report with the identifiers associated with the routes: route gva, route fco, route bud, route jfk and route iad. The proof of concept and first analysis steps will consider only route gva. The reason to choose that route was because of the consistency and quantity of its data. The analysis will then be extended to the other routes.

6.2 Data Sets

After deciding which routes to analyze we still have to discuss which part of the data is going to be used, as they have stored data on net bookings for more than 5 years

Period of Data Considered

The data was processed from the Brussels Airlines server. The booking data was spread across daily files from where the net booking data for each of the 25 fare classes on each flight was processed and compiled together. Due to the big time consumption accessing and processing that data only one year of data was covered. This means we will not cover yearly trends in demand between years.

The Brussels airport was target of a terrorist attack on the 22nd of March of 2016. This caused the closure of the airport for few days as well as a major disruption in operations and demand to/through Brussels during the subsequent period. For that reason the period after the attacks is not going to be considered for the analysis and, consequently, the year of data processed is from the 21st of March 2015 to 21nd of March 2016.

Besides those periods that are not considered, special flight numbers that represent special flights created, for instance, after a disruption in operations will not be considered either. Codeshare flights are flights that advertised and sold by more than one airline as being its own. Since these flights might have an extra set of demand that is generated only due to the partner airline, they are not processed as relevant data either.

Aggregation Levels

As mentioned before the Brussels Airlines fare structure include 25 different fare classes between business and economy as described in 4.4. During this study the different classes will be aggregated together and the forecasting is going to be done on a flight level. The reason behind this decision is two-folded. The first reason is that the fare classes content have changed throughout time. For instance groups were attributed to fare class M in earlier dates of the data gathered while on later dates they were attributed to fare class G. The second reason for aggregating on a flight level is that a lot of those classes have a very sparse demand, specially for business fare classes in Europe, this is more difficult to evaluate and usually needs special optimization techniques for the model to

not predict always zero.

After all the previous steps described in this section of the report we got to the final matrix with the process data. The number of rows of the data matrix is the the number of flights in our data. The columns are organized as follows: first all the days of the booking curve, between the opening of the bookings to the departure day are going to be included, with t_0 being the day of departure. t_340 being the day 340 days before departure day and other days following the same type of logic; the next column identify the route; then we have column identifying the flight number associated with that flight; month of departure is the following column; finally day of the week of the departure is identified.

The year was not included as a column in the data because the data only comprises a full year of data so there are no overlaps. To finalize this section we can find on table a example of the data structure used:

t_0	**...**	**t_340**	**flight_number**	**route**	**month**	**DoW**
140	...	0	2720	gva	8	1
...	...	...	..	...	...	...
132	...	1	3182	fco	2	5

Table 1: Data matrix structure example

This section gave an overview of the forecasting environment and some of the decisions required during the data gathering process that were important for the definition of the project.

6.3 Data Treatment

Being defined which part of the network to use as data, in which period of time and which aggregation level of the said data, the next step was to identify gaps in the data. The following problem was identified: net bookings daily files were not recorded on holidays and weekends which created gaps in the data.

This data had to be filled. There are two proposed methods to fill these gaps in data depending on the situation.

The first situation encountered is holidays and weekends that do not happen in the day of departure but during earlier stages of the booking period. Assuming that the bookings come sequentially and there will be no big volume of bookings and cancellations during those days when comparing to the surrounding days, the following equation was used:

$$b_t = \frac{b_{t+1} - b_{t-1}}{2} \tag{30}$$

With b_t being the net bookings on the day where the data is missing and b_{t+1} and b_{t-1} the net bookings on the following and previous day, respectively. When there is the case of a weekend where the next day also does not have data on the number of net bookings, b_{t+1} is replaced by b_{t+2} in the equation.

The other situation is when the gap in the data happens for the day of departure. This means that there will be no data on days after to compare with. Any pattern filling here would always bias the data. When analyzing all

the routes in the data we conclude that the average increment from the day before departure to the day of departure for the flights that have a complete booking curve was of 0.28 net bookings. The final decision was then to consider that the variation is small enough to use the following imputation method:

$$b_t = b_{t-1} \tag{31}$$

As in the previous equation, b_t being the net bookings on the day where the data is missing and b_{t-1} the net bookings on the previous day.

The other main issue was the unconstraining of the data. As explained in 2.3.4, the booking data is not the actual demand. During the booking period airlines do not sell certain fare classes which leads to the loss of potential demand if the costumers are not willing to join a different fare class or different flight. Some methods of unconstraining were reviewed. The main requirement is that there should be information of the period for which a certain fare class was not taking requests for tickets. This information is currently not available in the system. The current origin-destination system does not have the information on when a class is closed for a certain flight. This means that the unconstraining could not be done. For this reason in this work we will use net bookings as the forecaster input as well as forecasting for net bookings and not unconstrained demand.

Another topic that is important in any project that makes use of real that is the identification and processing of outliers. This will be discussed more in detail in 6.5

6.4 Data Exploration

After the data processing stage of the work an analysis of the data was conducted. This analysis had as the major objective verifying the data distribution and booking patterns when grouping the data by different parameters such as route, flight number, month of departure and day of the week of departure.

The analysis started with an analysis of the overall distribution of flights per number of net bookings at the day of departure. This can be seen in figure 6.1. We can clearly see two different areas in the overall distribution chart. The first one is related to short haul flights with peak frequency between 50 and 150 net bookings. The second clear region is related to long-haul flights, with peak at 280 net bookings.

To investigate further the distribution of the demand we plotted the same type of distribution per route and separating it by long-haul and short haul. This can be found in figures 6.2 and 6.3.

The clear conclusions that can be taken from this is that the different routes have very different distribution patterns. On long-haul while iad route has a very uniform distribution, jfk flights have most of their flights departing with around 280 passengers. Jfk is a well established route that provides most of their flights with a high load-factor, on the other side iad is a new route where the main costumers are more business oriented, these type of passengers in general

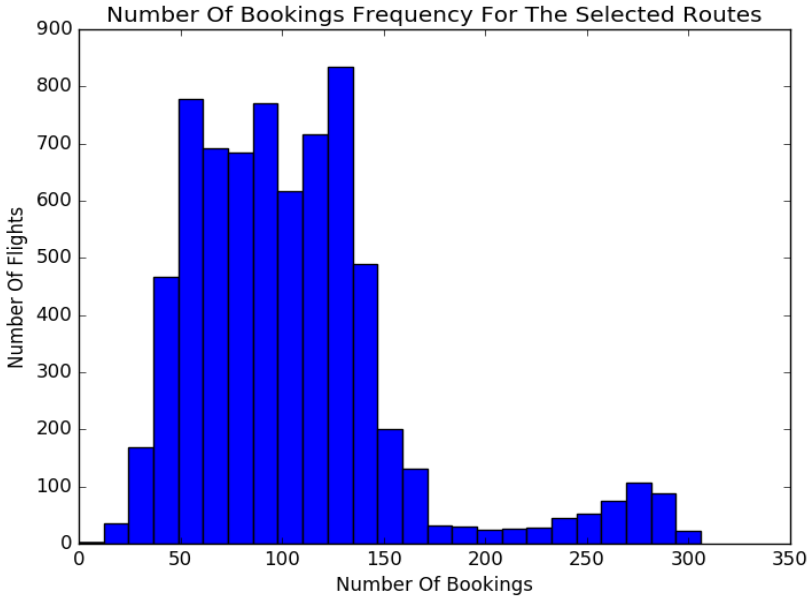

Figure 6.1: Flight distribution per number of bookings

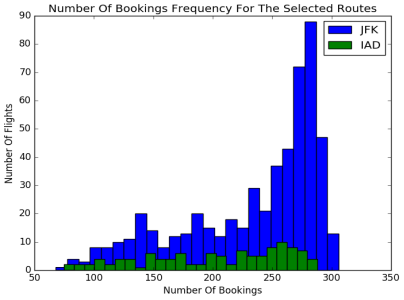

Figure 6.2: Long-Haul Flight Distribution

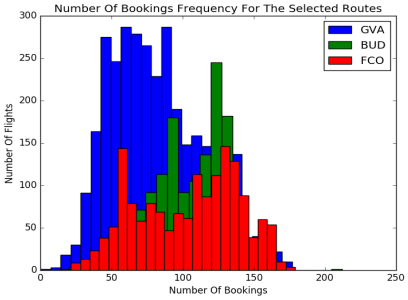

Figure 6.3: Short-Haul Flight Distribution

prefer to stay within the same carrier and as such are harder to capture which explains why there is a strugle to capture higher load-factors as of now. On short-haul some patterns can also be identified. Not only the different routes have different peak regions, but also fco route has peaks in two very different regions of the distribution chart. This lead to a further investigation on other factors that can justify that type of behavior such as day of the week of the departure, month of departure, flight number, so that we could make a better segmentation of the demand.

The next step of the analysis was to identify if the booking pattern varied depending on the departure month. Figures 6.4, 6.5 and 6.6 show the average net bookings per day per month for each respective route.

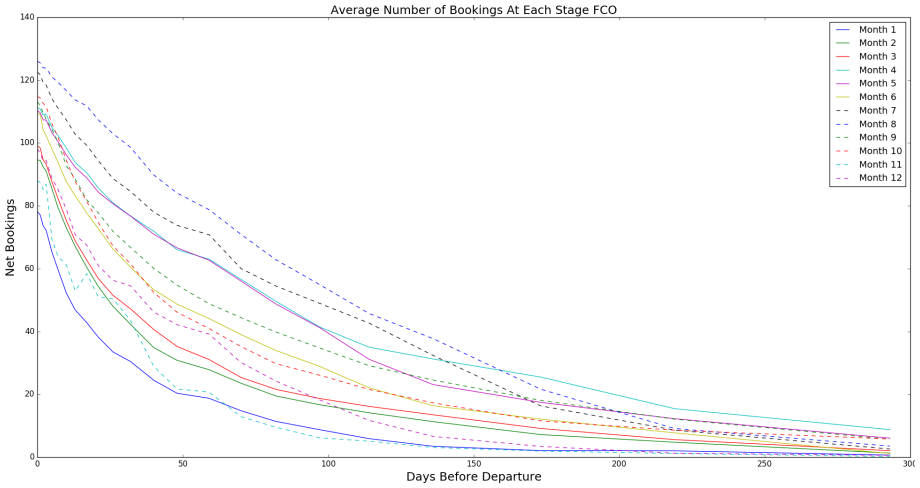

Figure 6.4: FCO average number of net bookings per month

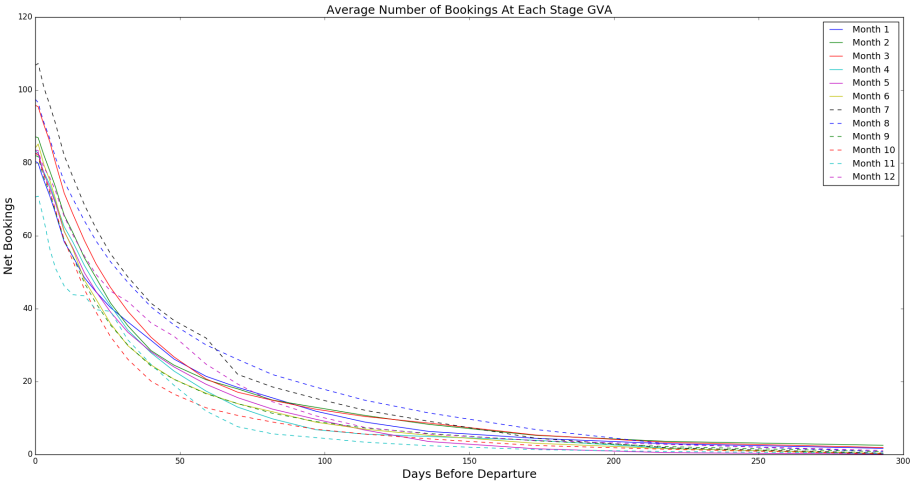

Figure 6.5: GVA average number of net bookings per month

Comparing these two short-haul routes, we can clearly see two evident patterns. The first conclusion is related to the type of route. GVA being a more business oriented route when compared to fco has its bookings coming closer to departure date. Logically, the opposite is found in fco that, being a leisure oriented type of route starts getting bookings earlier in the booking period. In figure 6.6 we can also verify that jfk, being a long-haul route, starts having a good percentage of its bookings earlier in time when compared to both short-haul routes.

Something that can also be clearly seen in this analysis is that different departing months lead to different average number of net bookings and different evolution patterns. For instance the two strongest months in fco route are August and July which is logical considering those are the peak periods for tourism in Southern European destinations.

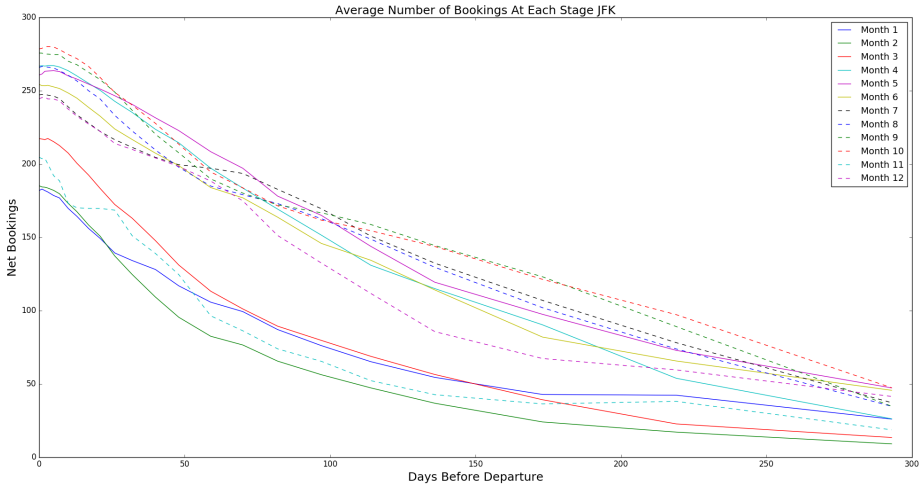

Figure 6.6: JFK average number of net bookings per month

To finalize this section we show below the average booking curve per route depending on the day of the week of the flight departure. The most important insights can be found on figures 6.7,6.8 and 6.9. The days of the week in the figures are identified from 1 to 7 with 1 being Monday and 7 being Sunday.

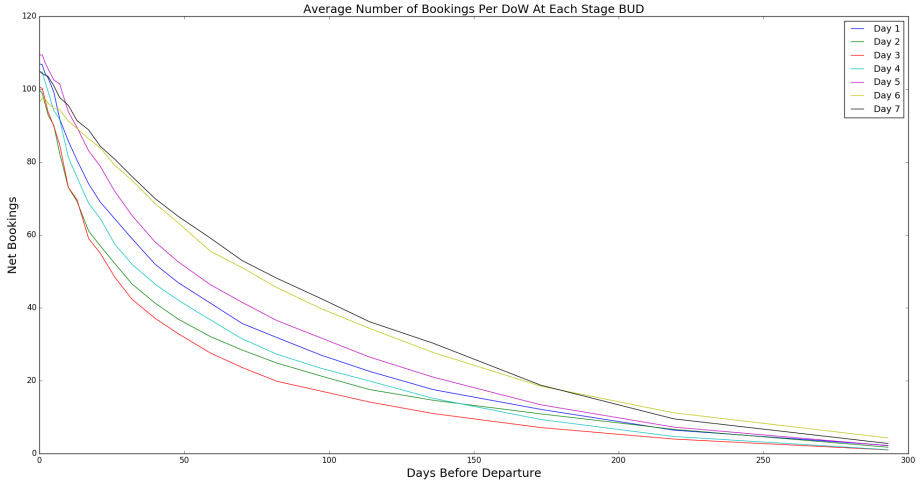

Figure 6.7: BUD average number of net bookings per day of week

We can conclude from the short-haul routes that the weekend flights not only start to get a big batch of the bookings earlier but also have higher average number of net bookings. Specially on the bud route we can identify that flights that depart during weekdays have bookings on a later stage of the booking period. This is probably connected with the type of passengers traveling during the week being business passengers.

On jfk route the distinction between days is not as noticeable as in the

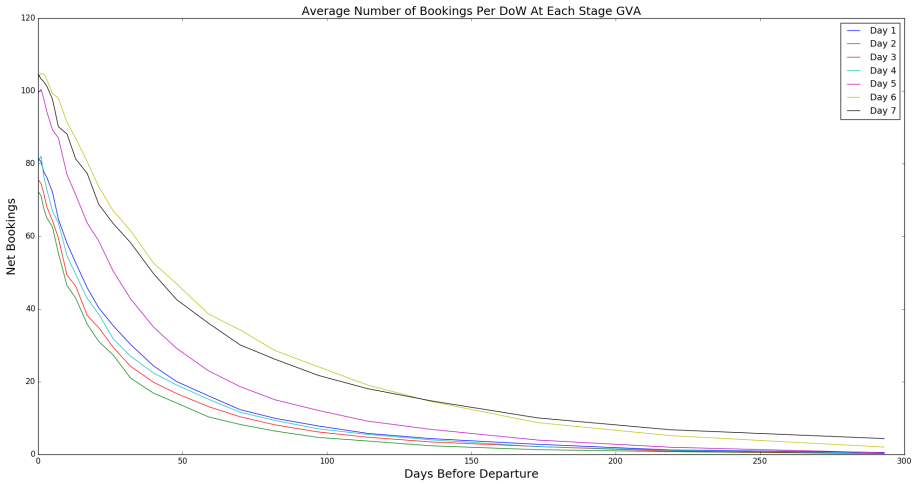

Figure 6.8: GVA average number of net bookings per day of week

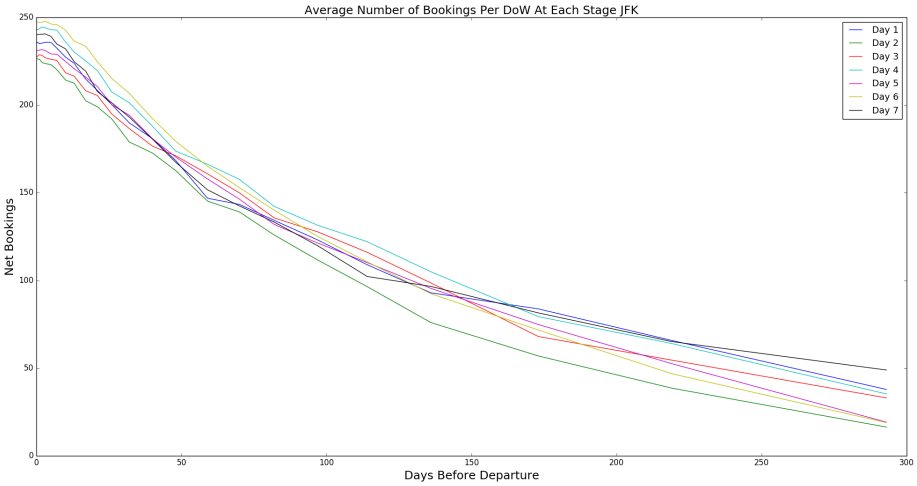

Figure 6.9: JFK average number of net bookings per day of week

short-haul routes.

The analysis done on this section shows that many factors take a part into the volume and arrival of demand. In this study we work with the data available, and as such, these factors will be considered during the forecasting procedure and influential factors. For sure other factors play a role in driving the demand up or down but due to their non-availability as data or simply for the model to not be over-complex they will not be considered in this study.

6.5 Tests

To optimize our models to the data used in this study and to study the hypothesis proposed at the beginning of the developed work six different tests were designed. These tests had different purposes. Tests I and II are made to improve the models considering the type and amount of data we have. Tests III and IV are made to evaluate the performance of the model under different situations. Tests V and IV were designed to confirm or deny the hypothesis stated in 3.3.

Three different times were chosen for analysis:

1. 5 days before departure;
2. 28 days before departure;
3. 98 days before departure.

The objective was to chose a period of the booking curve close to departure, hence the 5 days prior to departure, a medium-term period of forecasting, 28 days before departure, and a long-term period, hence the 98 days before departure. These specific days were chosen because they matched the data collection points of Brussels Airlines system closer to the wanted dates.

Most of these tests will be tested for the following 6 different time frames:

1. With data until 5 days before departure forecast for the day after - situation 05-04
2. With data until 28 days before departure forecast for the day after - situation 28-27
3. With data until 98 days before departure forecast for the day after - situation 98-97
4. With data until 5 days before departure forecast for the departure date - situation 05-00
5. With data until 28 days before departure forecast for the departure date - situation 28-00
6. With data until 98 days before departure forecast for the departure date - situation 98-00

This will give us very-short term forecasting in different regions of the booking curve. One long-term forecast, one medium-term forecast and one short-term forecast. During the rest of this report the different forecast time-frames will be identified as 05-04, 28-27, 98-97, 05-00, 28-00 and 98-00.

Something important to note is the machine where the computations were done. The machine used had 12Gb of ram, a 64 bits architecture, Intel Core i7-4720HQ @ 2.60GHz. The machine having 8 cores is relevant for these computation times obtained. While Ridge regression is done sequentially and the number of cores does not affect the computation time, the decision trees are being built in parallel in each processor core of the computer. This means, for

example, that on a machine where everything else is the same but with only 4 cores the computational times for Random Forests will be doubled.

Test I: Proof of concept and Parameter Tuning

The first step of the analysis was to do a proof of concept. For that both Ridge regression and Random Forests were tested for their feasibility and it was identified if the results obtained were reasonable. After, the main parameters of both models were tuned using the data.

The models, as explained previously in this report, will be first tested for the gva route. The data input to the model was 66% of all the historical flights for that route, the other 33% are going to be left out for later for the validation stage of the project. The features input to both mathematical models are the historical net bookings of each flight from the day bookings open to the current day, which depending on the situation being studied can be 5 days before departure, 28 days before departure or 98 days before departure. The metrics used as explained in chapter 5 Methodology are going to be the RMSE, MAE and DINT.

The second part of this test was to do the parameter tuning of the models. The parameters to tune differ between both models. Ridge regression has as the only parameter to tune the regularization parameter, λ. Random Forests has 3 different parameters to tune: (1) number of trees that are going to be created, the general idea is that more trees are better for the forecasting, but at a certain points the increased number of decision trees (and respective increase in computation time) do not compensate the increase in accuracy; (2) minimum number of samples in a node to be able to split an internal node, this is a stopping criteria; (3) minimum number of samples per leaf so that a new node will be created, this is also a stopping criteria.

These parameters were submitted to a search using the range of values described bellow.

Ridge regression:

1. Regularization parameter: 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000

Random Forests:

1. number of trees: 10, 30, 100, 300, 1000;
2. minimum number of samples in a node: 2, 3, 4, 5, 6, 7, 8, 9, 10;
3. minimum number of samples per leaf: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

For ridge regression a bigger regularization parameter means we achieve the optimum faster but if the regularization is too big then the model might not achieve the optimum. As such, it should be chosen the biggest regularization parameter before the solution starts degrading.

For random forests the solution is a bit more tricky since there are more parameters to analyze. We want to have the least number of nodes and leafs (to reduce model complexity) that do not degrade the solution while the number of trees is a trade off between computational times and accuracy.

Test II: Feature Selection

This test will cover the feature selection of the model. In the previous test we test the models by giving them as features the net bookings for all the days from the opening of the bookings to the current day, i.e., if we were at 5 days before departure trying to make a forecast, we would train the model with the data of all the days between the opening of the bookings and 5 days before departure.

In this case the model will have as an input 340 features for a n number of flights in our historical data. In the case of a short-haul route such as gva it will happen that for the first days the number of bookings will vary very little and, as such, not give much information to our model. This can be seen in figure 6.10 where we have the average build-up of the net bookings for gva route.

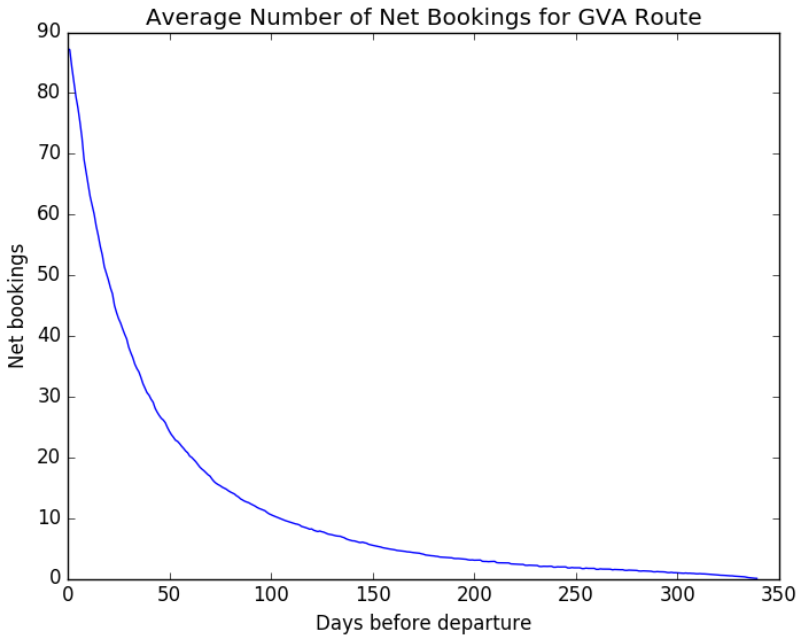

Figure 6.10: Average number of net bookings per day in GVA route

It can be seen in the figure that for the first 250 days of bookings there will be no more than 10 bookings. With the very small variation presented, most probably those variables will be ignored by the models.

Machine Learning models have the advantage to be able to pick the features that are the most important for our prediction and ignore the ones that do not have an added-value. Nevertheless, more features as input means higher computation times. In this test we try to decrease the number of features to the essential ones and see if the trade-off between accuracy reduction and computation time reduction is advantageous.

Test III: Outliers

The next step of analysis was to screen the data for severe outliers. Outliers can come from two sources: (1) errors in the measuring phase and (2) real values that for some reason vary too much from the normal value distribution. Errors of the type (1) can occur for example due to bugs or errors in the software that captures the values. Human error might also be a reason if a person commits a mistake when entering a value into the system.

Errors of the type (2) can happen due to one-off events that will create an abnormal variation in the target variable. In our case where the main data is historical net bookings we can find many examples. For instance, a big number of bookings when comparing to the historical data can arise due to the cancellation of a flight by another airline or a special promotion with very low prices.

Outliers can have a big influence on the estimation of the parameters of the model (or on the training of the model) which will lead ultimately to worse forecasts. According to Lee [36], in practice, outliers are treated by the mean and standard deviation method. Here we will approach the problem by removing flights that deviate more than 3 standard deviations from the average. To see how the performance varies we are going back to the gva route. The testings will be done already with the feature selection implementation.

Test IV: Variation of Performance for Different Forecasting Periods

In this test it will be analyzed how both models perform in terms of accuracy across time. For that we will forecast every day from current day until day of departure for the 3 periods considered throughout this report:

1. 5 days prior to departure;
2. 28 days prior to departure;
3. 98 days prior to departure.

The main objective here is to identify how does the models forecasting capabilities vary for different periods of the booking curve.

Test V: Benchmark Comparison

As previously covered in section 4, the two most used models at Brussels airlines are the linear regression and the logistic regression. In both these models it will take the data from our current data collection point (pre-defined days in the booking curve where data is collected and the model re-forecasted) and the historical data collection point that we want to forecast. A regression will be fit to that historical data and used on the new points. Making the parallel to our models, only the current day and target day are used to train the model.

We will test both these models for the 6 situations in gva route under the constraints described and compare with the results obtained on the previous test. The main objective here is to test the first hypothesis stated on section

3.3.

Test VI: Pooling vs Non-Pooling

After all the previous steps on this chapter we arrive to the conclusion that the two best models, depending on the booking horizon, are linear regression (or Ridge regression, since they yield the same results) with just the current day as input feature or Random Forests with all the days as input features. While linear regression is a lot faster in terms of computing time, Random Forests can get a lower error on long-term forecasts.

In this final test we want to test the pooling vs non-pooling issue. The linear regression model is not capable of using categorical features such as the day of the week, month, flight number, etc., to train the model while Random Forests can. As such we tested including the following features: (1) flight number; (2) day of the week of the departure; and (3) month of the departure. The data was first tested in the gva route and later expanded to include all the routes. We tried including each of the new features individually and coupled together.

The objective of this final test is to test the second hypothesis stated in section 3.3.

This chapter presented some data analysis and exploration while also defining how this same data would be treated before starting to fit it to the model. In this last part of the chapter the tests run during this project were presented. The most important results will be discussed in the next chapter.

7 Results and Discussion

This chapter presents the results obtained after running the tests described in the previous chapter and will provide a discussion on those results. The first section 7.1 will show the results on model optimization while section 7.2 presents the results on the hypothesis testing and validation of the results. The chapter will close with a summary and discussion of the results in section 7.3.

Only the main contributing results of each test will be presented on this chapter.

7.1 Model Feature and Parameter Optimization: Tests I, II and III

This section will summarize the results for the first three tests. This tests were, as explained before, designed to prove the concept and optimize our models in terms of its own pre-defined parameters as well as the data that is fed to the said models.

Starting with test I the first results for the 6 situations described in the beginning of this chapter can be found on tables 2 and 3.

Metric/Situation	**05-04**	**28-27**	**98-97**	**05-00**	**28-00**	**98-00**
RMSE	2.41	1.61	1.23	6.25	12.82	21.51
MAE	1.22	0.84	0.44	3.98	9.46	16.64
DINT	0.68	0.67	0.69	0.67	0.66	0.58

Table 2: Random Forests performance for GVA route - proof of concept

Metric/Situation	**05-04**	**28-27**	**98-97**	**05-00**	**28-00**	**98-00**
RMSE	2.76	1.84	1.02	8.45	19.24	27.49
MAE	1.62	1.07	0.44	5.94	14.92	22.01
DINT	0.59	0.60	0.69	0.54	0.49	0.45

Table 3: Ridge Regression performance for GVA route - proof of concept

The first thing we can conclude from these results is that the models have results that can be considered reasonable. This consideration is done first by looking into the DINT metric. For Random Forests depending on the case, it will have an error smaller than 58% to 69% of the cases. The fact that MAE and RMSE are consistent indicates that there are no huge errors in terms of predictions, as RMSE penalizes big errors a lot. The last metric shows a small degradation of the models performance when we get further from the prediction date.

Another conclusion we can take from tables 2 and 3 is that Random Forests is outperforming Ridge regression, as Ridge regression only outperforms Random Forests in 1 of 18 cases that can be seen in the table. This conclusions should be taken with care as Random Forests is able to model much more complex interactions and as such is still prone to overfitting the data-set we are using. The possible overfitting will be investigated more in detail in section 7.2.

After having the proof of concept and confirming that the results are as theoretically were expected the next stage was to do the parameter tuning of both models. These parameters are described in the previous chapter when we describe the various tests.

Below on table 4 we will show only the results for the best combination of parameters and how these results could be compared with the ones obtained before. For Ridge regression the standard parameters with the regularization parameter equal to 1 as in the proof of concept tests was the best of the grid search and as such will be kept and the results not be provided below.

The parameters chosen to provide the best trade-off between computational time and results were the following:

1. number of trees: 100;
2. minimum number of samples in a node: 2;
3. minimum number of samples per leaf: 1.

Metric/Situation	**A**	**B**	**C**	**D**	**E**	**F**
RMSE	1.26	0.92	0.71	3.53	7.35	14.95
MAE	0.73	0.54	0.32	2.40	5.64	10.70
DINT	0.80	0.79	0.78	0.80	0.79	0.72

Table 4: Random Forests error results after parameter tuning

We see comparing both Random Forest tables that the parameter tuning largely improved the metrics given. This, again, can be a result of overfitting the training data so for now should be interpreted carefully. Nevertheless the concept is proved to work and that the models are obtaining promising results.

The tests II and III are not very relevant to the conclusions of this study and as such the analysis is not included here.

Feature selection was motivated by figures 7.1 and 7.2:

Here we can find that for short term-predictions as well as long term prediction the most recent features take all the weight. While the weights are more distributed in situation 98-00 depicted in figure 7.2, it is still seen that the first 15 features, i.e., the 15 most recent days in the booking curve, will take 80% of the weight in the model.

Trying to make use of this on both models we tried to include thresholds that define the minimum weight for a feature to be considered but the conclusions were that this would not improve the performance in terms of accuracy metrics neither would be noticeable in terms of computation time and as such there is no use in implementing.

The only situation that can come up where this feature selection process is important is for Ridge Regression when the number of features is bigger than the number of flights in the training data. This will make the problem to not have a closed form solution and as such for the results to explode.

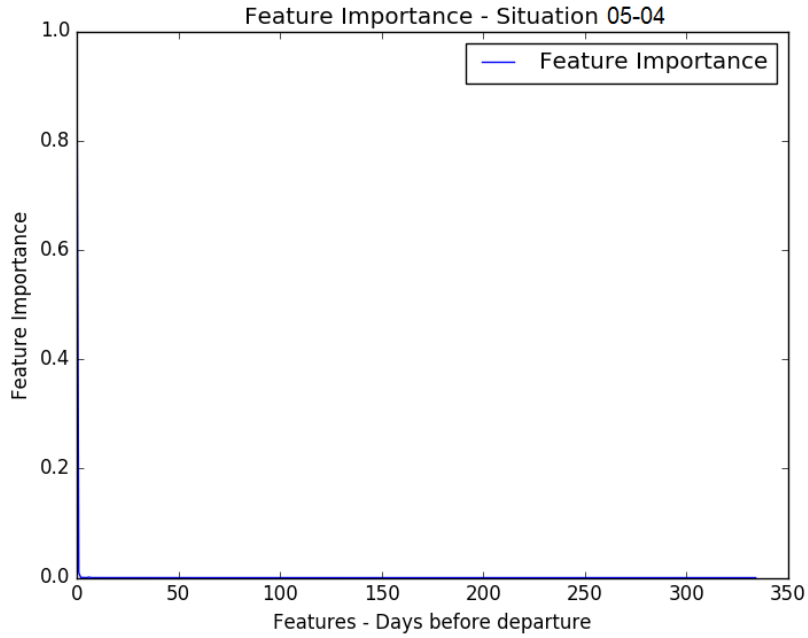

Figure 7.1: Feature importance GVA route - Situation 05-04

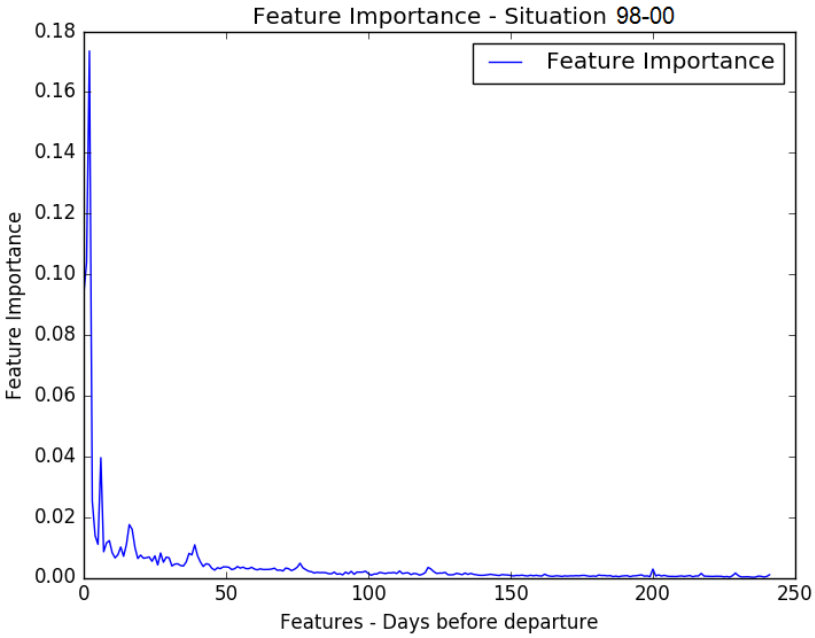

Figure 7.2: Feature importance GVA route - Situation 98-00

Considering test III, the data was analyzed for outliers of the type (1), which were clear errors such as errors in entering the data. After searching for impossibly high values or negative net booking values, no errors of that type were found in the data. There were type (2) outliers, i.e., data that is realistically feasible but deviates too much from the standard set of values. Nevertheless, after running the model and eliminating examples that varied more than 3 standard deviations, it was found that those values did not have a big influence on the forecaster as seen in the tables presented in this section and, as such, they will not be removed from the data pool.

To summarize, feature selection did not prove to improve the performance nor in computation times neither in the accuracy of the forecast, the only recommendation is in the case where there are more features than flights on the training data, the number of features should be reduced, and this should be done by only using the most recent days, so that there is a closed form solution again. Outlier removal did not make any noticeable changes in the forecasting results either and as such is also not included in the method.

7.2 Hypothesis Testing and Results Validation: Test IV, V, VI

In this section a summary of the results of the last three tests is going to be presented as well as the validation of those results to be included. The main objective of this section is to test the two hypothesis proposed earlier on this work. To recap, the two hypothesis are:

1. *Ridge Regression is an improvement to the conventionally used linear regression in the problem of airline demand forecasting.*

2. *Random Forests is able to capture categorical features such as month and day of departure, route and flight number to provide automated clustering of data without losses in performance*

When we compared from Ridge regression to linear regression we see that the model built is roughly the same given the exact same results when using 2 decimal cases. This means two things, the first is that Ridge regression in this case is not an improvement over linear regression and as such the simpler model, linear regression should be chosen. The second is that this probably indicates no overfitting from the model and that the generalization when presented with the validation data-set will provide similar results to the training set.
This then refutes the first proposed hypothesis.

As a result of previously refuting the hypothesis that Ridge regression would be an improvement over linear regression, we then compare Linear Regression with Random Forests. The first part of the results was to identify how both models perform when forecasting the full booking curve for the periods of 5, 28 and 98 days before. The graphs for 5 days and 98 days of training are presented in figures 7.3, 7.4, 7.5 and 7.6.

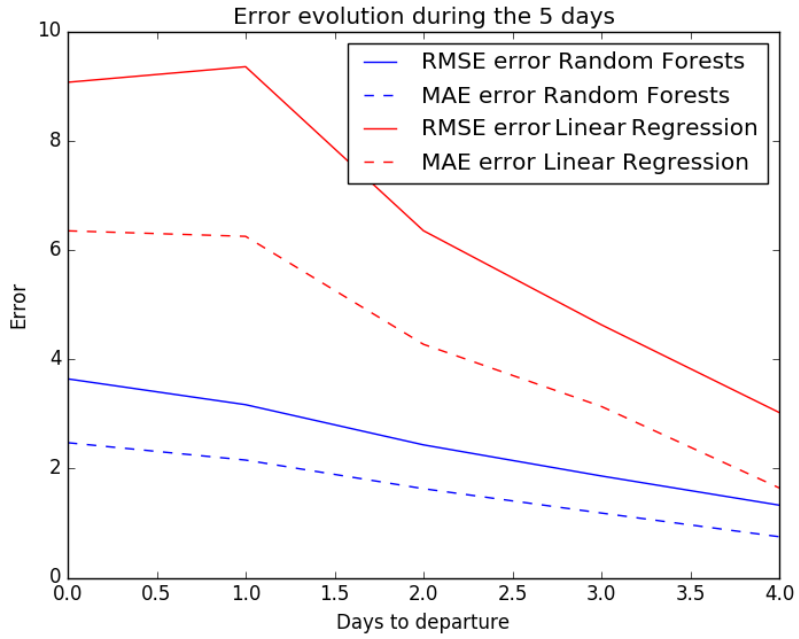

Figure 7.3: Error metrics RMSE and MAE, GVA route - Situation 05-00 - Training data

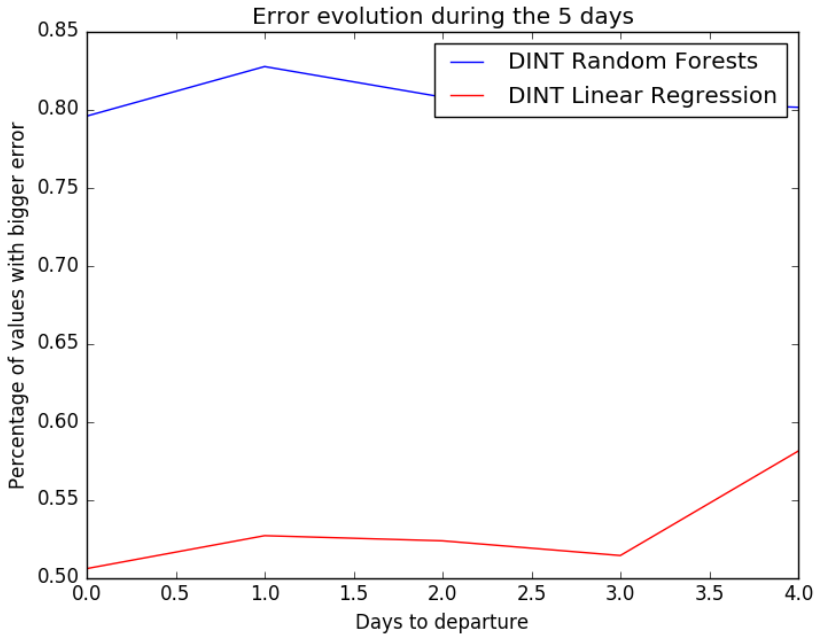

Figure 7.4: Error metric DINT, GVA route - Situation 05-00 - Training data

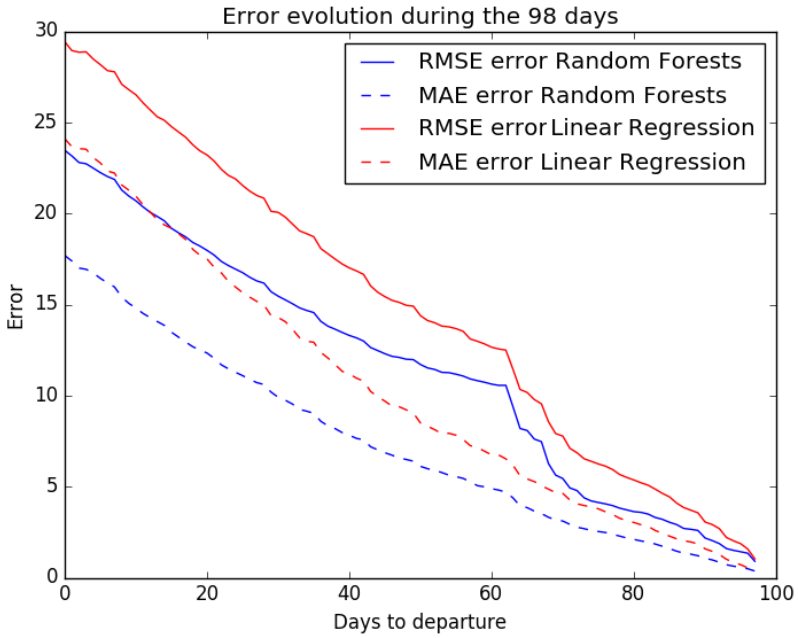

Figure 7.5: Error metrics RMSE and MAE, GVA route - Situation 98-00 - Training data

These results show us Random Forests outperforming linear regression in every situation. We see that we have a DINT of around 50% for linear regression which means that 50% of the values will have a similar or bigger error while Random Forests is putting that value at 80%. The models do not seem to degrade much either. By looking into the RMSE we conclude that there shouldn't many predictions with a big error or the RMSE would explode more and looking at the MAE we see that both on short term, 5 day prediction, and long term, 98 day prediction the difference is considerable, with the MAE for Random Forests being roughly $\frac{1}{3}$ for the 5 day prediction.

This results nevertheless might mean an overfitting from the Random Forests model as it is a much more complex algorithm which can better model the training data. As such we must validate these results using the 33% of data that was randomly left aside at the beginning of the project. This means the models are then trained on the training data and later used on this new set of data to make prediction.

On figures 7.7 and 7.8 we can see some of the results when using the models to make forecasts on the validation data. We see confirmed that Random Forests was indeed overfitting the training data as its performance in terms of error metrics largely declines.

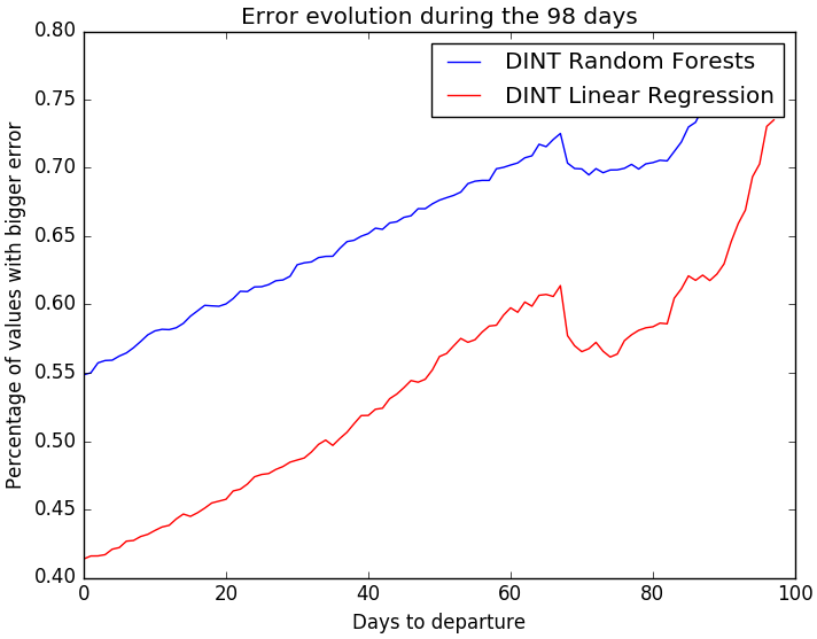

Figure 7.6: Error metric DINT, GVA route - Situation 98-00 - Training data

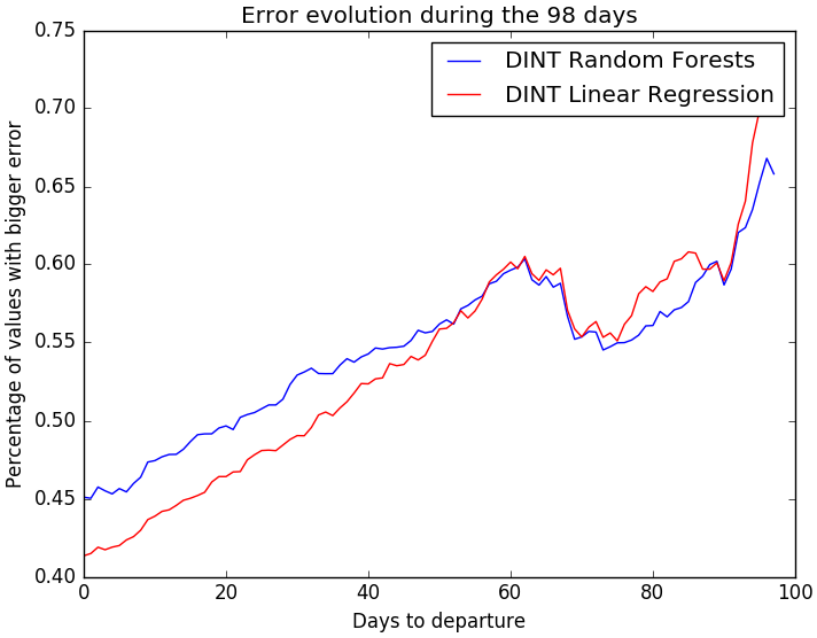

Figure 7.8: Error metric DINT, GVA route - Situation 98-00 - Test data

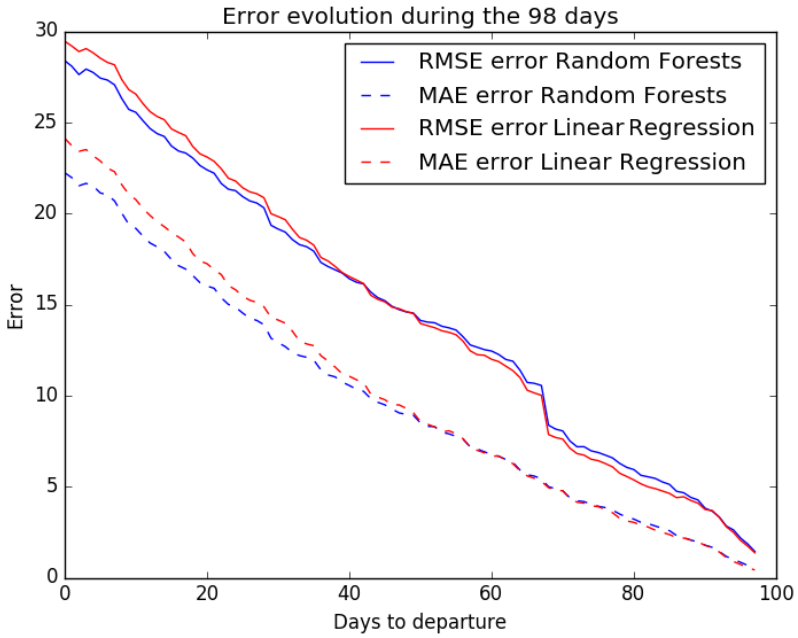

Figure 7.7: Error metrics RMSE and MAE, GVA route - Situation 98-00 - Validation data

We see while Random Forests on the long term still outperforms linear regression, the difference is now much smaller with for instance the MAE at the day of departure having a difference of 10%, i.e., around 2 people. A 10% difference in this case is still noticeable as it is 2 extra persons that are booking tickets.

Something about the results that is also important to mention is that while these results are for predictions in GVA, the results are the same for the other routes but they were not included here.

Acknowledging this there are two things remaining to test. The first is to confirm the ideas used in practice that pooling improves the results of forecasts for linear regression. The second is to confirm or refute the hypothesis that when introducing categorical features to the Random Forests model, this process can be automated while at the same time obtaining a better accuracy.

After some tests were conducted we confirm that indeed linear regression performs better when the data is pooled. Figure 7.9 shows the evolution of the metrics RMSE and MAE when the data is just from flight number 2711 from GVA.

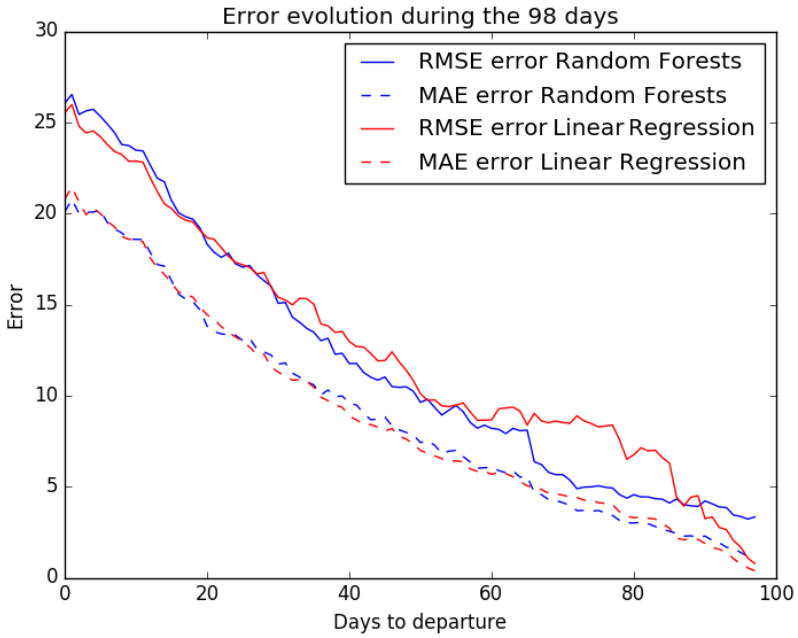

Figure 7.9: Error metrics RMSE and MAE, GVA route - Situation 98-00 - Test data - Pooled

We can see here that actually not only linear regression but also Random Forests improve their performance when the data is being pooled. For this type of solution, Random Forests having a very similar results while being more computationally expensive makes linear regression a better solution even on long term predictions. Short-term predictions are not shown here because linear regression always performed better on those cases.

As such the final tests were to introduce new categorical features into to the Random Forests model such as flight number, route, day of the week of departure and month of departure and to introduce data from flights from all the routes into the data that would be used to train the model. The best results were obtained when all the data was fed together to the model and the model was let to "decide" how to do the pooling. These results are shown on figure 7.10.

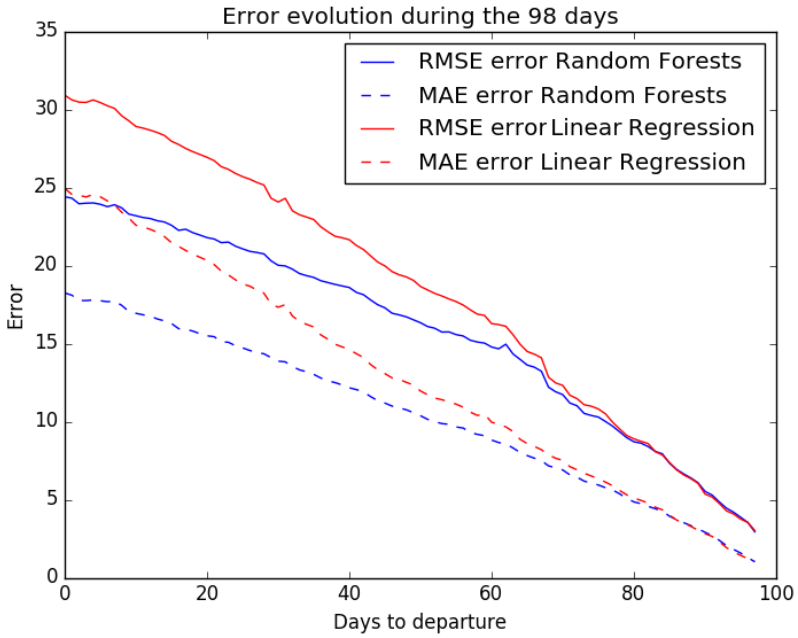

Figure 7.10: Error metrics RMSE and MAE, GVA route - Situation 98-00 - Test data - Not Pooled

The first thing we see in this graph is the confirmation that, as discussed before, the linear regression model performs worse when the data is not pooled. The second, and most important conclusion, is that we confirm our hypothesis that the Random Forests model performs better in terms of accuracy when fed categorical features that let it do its own clustering and data from all the routes.

Looking at MAE and RMSE, and comparing figure 7.10 and 7.9, we see that it even performs better than linear regression when pooled. This are very important results because it shows that an approach that goes against the practical applications where its considered that pooling each time more can provide similar to better results. Another important consideration is also that in the figure 7.10 it is taking into account also 2 long-haul routes that have inherently bigger errors due to bigger standard deviation in the distribution. The DINT metric was not included in the graphs here as it is not good to compare between models that use different data-sources. Since including long-haul flights will cause a much bigger standard deviation, this will also lead to a better DINT.

Finally bellow we have the tables 5, 6 and 7 that show all these results summarized for the three metrics.

Situation	**05-00**	**28-00**	**98-00**
Train LR	9.07	20.97	29.43
Train RF	3.62	8.89	23.44
Validation LR	9.51	21.10	29.45
Validation RF	9.87	20.67	28.44
LR Pooled	9.11	18.04	25.54
RF Pooled	9.65	19.09	25.85
LR All data	9.86	21.05	30.90
RF All data	9.53	17.43	24.37

Table 5: RMSE summary for all the tests conducted during this study

Situation	**05-00**	**28-00**	**98-00**
Train LR	6.34	16.50	24.10
Train RF	2.47	6.51	17.66
Validation LR	6.33	16.80	24.12
Validation RF	6.59	15.74	22.28
LR Pooled	6.74	14.10	20.79
RF Pooled	7.16	14.16	19.95
LR All data	6.57	16.62	24.93
RF All data	6.18	12.87	18.22

Table 6: MAE summary for all the tests conducted during this study

Situation	**05-00**	**28-00**	**98-00**
Train LR	0.51	0.44	0.41
Train RF	0.80	0.76	0.55
Validation LR	0.50	0.44	0.41
Validation RF	0.49	0.47	0.45
LR Pooled	0.43	0.43	0.40
RF Pooled	0.40	0.43	0.41
LR All data	0.52	0.47	0.48
RF All data	0.55	0.58	0.61

Table 7: DINT summary for all the tests conducted during this study

The conclusions of this study will be discussed on the last section of this chapter.

7.3 Discussion of the Results

This section will make a summary of the insights obtained by analyzing the results as well as the recommendations associated with it.

The first conclusion we get from both the tables and the figures is that the Random Forests model was overfitting the training data. This is seen when we use the models on the validation data set. There we see that the average

performance of both models is very similar when only using demand data.

It was also tested and concluded that Ridge regression is not an improvement to linear regression in this case with both models having virtually the same results when considering the level of accuracy we are considering here. As such our first hypothesis for this work was refuted.

The tests conducted led to conclude that linear regression performs better when the data is pooled which confirms what is done in practical applications in the airline industry.

We confirm the the second hypothesis of this work that the Random Forests model is able to capture the influence of categorical features such as flight number, day of week of the departure, month of the departure and route. We can see by the figures and tables presented that this is specially noticeable on the long term predictions while on short-term, i.e., for the 05-00 case, the difference is small enough to be worth using linear regression. This also indicates there is a probability that adding even more routes and different features could improve the results already obtained here.

The main conclusions from this results are then that for short-term predictions linear regression is the better performer as it has results that do not differ to much from Random Forests while being much faster to compute. We also conclude that on short-term predictions the most important features are the most recent demand data, with categorical features not having such a big effect. On long-term forecasting Random Forest clearly outperforms linear regression in terms of the defined metrics and we see that here the demand data is less important but the categorical features given to the model get a higher importance. This gives then a new approach to select the data for forecasting in an automated way without losing forecasting accuracy.

To finish, a small discussion on the computational times and feasibility is required. Currently Brussels Airlines has roughly 82,000 flights per year. The forecasting is done every night and takes around one hour to do. For linear regression the computation was fast and is the currently implemented model so the focus is to estimate how fast Random Forests could perform. For this two different situations were tried: using GVA route which has 4559 flights on our data and all the routes which have a total of 9648 flights in our data. We see that the computation jumps from nearly 40 seconds to 62 seconds. First we see the increase is not linear, second, even if it was, it would take around 10 minutes to run the forecasting on all the data. This is a very good time but some other factors should be taken into consideration. The first is that the Random Forests model can run in parallel computing and was running in 8 cores at the same time. The second is that currently their data is not in the format that is required to be run by this model and as such that processing time might be a lot larger. The last one but not the least is the fact that this is done on a flight aggregation level. The number of forecasts when predicting for all the different classes is prone to increase a lot.

This chapter presented all the main results obtained during the course of this research, the main findings and conclusions.

8 Conclusion

This work has delve into the problem of forecasting for a revenue management system in the airline industry. During the literate review stage of this project the main problem of lack of recent research on forecasting algorithms applied to the airline industry was identified and possible approaches to it were explored culminating in the identification of two main algorithms that could improve not only the forecast but also the process behind it. After identifying Ridge regression and Random Forests as potencial solutions, the context of Brussels Airlines was explored and it's implications on the methodology analyzed. Therefore all this prior information was used to define the methodology and identify it's main limitations.

After analyzing the available data, to test the main hypothesis proposed on this work a set of different tests were designed. All this work led to the main conclusions of this work discussed on the next section.

8.1 Research Findings and Contributions

The first hypothesis proposed by this work was refuted as Ridge regression was proved to not be an improvement to linear regression in the levels of accuracy pretended for this work. The regularization that is introduced with Ridge regression changed very little the weights of each variable, corresponding to a variation on the third decimal case in the forecasts. Since we were discussing bookings here, this kind of variation can be considered to be irrelevant.

The tests conducted led to conclude that linear regression performs better when the data is pooled which confirms what is done in practical applications in the airline industry.

The second hypothesis proposed by this work, that the Random Forests model is able to capture the influence of categorical features such as flight number, day of week of the departure, month of the departure and route to automate the data pooling process and improve the forecasting accuracy, is confirmed by the results obtained. When introducing all the aforementioned categorical variables and all the flights from the chosen routes into the data pool to train the Random Forests model the error of the model decreases even when comparing with the linear regression model after pooling. This indicates that there is potential in the future for other features to be tested and introduced into the model as well as the other routes on the network.

This improvement obtained with Random Forests is more evident on long-term prediction but less noticeable on short-term.

We confirm the the second hypothesis of this work, that the Random Forests model is able to capture the influence of categorical features such as flight number, day of week of the departure, month of the departure and route. We can see by the figures and tables presented that this is specially noticeable on the long term predictions while on short-term, i.e., for the 05-00 case, the difference is small enough to be worth using linear regression. This also indicates there is a probability that adding even more routes and different features could improve the results already obtained here.

Some of the big errors obtained during the testing phase are correlated with the batch booking, i.e., when groups book a large amount of tickets at once. The model is not capable of dealing with this kind of group forecasting the way it is designed.

The main conclusions from this study are then that for short-term predictions linear regression is the better performer as it has results that do not differ too much from Random Forests while being much faster to compute. Nevertheless considerations on the automation of the data pooling process should also be taken into consideration here, this saving in time might compensate the for the larger computational time. We also conclude that on short-term predictions the most important features are the most recent demand data, with categorical features not having a big effect on the prediction and being ignored by the model. On long-term forecasting Random Forest clearly outperforms linear regression in terms of the defined metrics and we see that here the demand data on current number of bookings is less important but the categorical features given to the model get a much higher importance.

We conclude then that Random Forests gives a new approach to forecast for the airline industry with the advantage of being able to select the data for forecasting in an automated way without losing forecasting accuracy.

8.2 Future Work

After all that was discussed before and all the findings that were achieved the following future contributions can be valuable.

It was shown in this research work that the Random Forests model is capable of capturing even categorical features to enhance the performance of the model. It would be important to investigate which extra features could be added to the model to improve even further the performance. These features might range from the signaling of an holiday to a big variation in price to a major event happening.

Another important notion is that the Random Forests model cannot predict values outside of the range of values in the training set. A future step would be to investigate the effect of yearly trends in the performance of the model.

The research done studied forecasts on a flight aggregation level. As a further step into improving the research, dis-aggregating the forecasting into different fare classes would be important because current revenue management systems use fare classes forecasting into their optimization algorithm.

The last future work recommendation that can be given is that with the proliferation of the origin-destination system into most main carriers the forecasting should be done on a origin-destination basis instead of on a leg basis.

References

[1] D Adelman. Dynamic bid prices in revenue management. *Operations Research*, 55(4):647–661, 2007.

[2] J. Alstrup, S. Boas, O. B. G. Madsen, R. Vidal, and V. Victor. Booking policy for flights with two types of passengers. *Eur. J. Oper. Res.*, 27(1): 274–288, 1986.

[3] M. J. Beckmann and F. Bobkowski. Airline demand: an analysis of some frequency distributions. *Naval Research Logistics Quarterly*, 5(1):43–51, 1958.

[4] P. P. Belobaba. Air travel demand and airline seat inventory management. *Ph.D. thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA*, 1987a.

[5] Peter Belobaba. The revenue enhancement potential of airline yield management systems. *Proceedings of the Advanced Software Technology for Air Transport International Conference & Exhibition*, 1992.

[6] Peter Belobaba, Amadeo Odoni, and Cynthia Barnhart. *The Global Airline Industry*. John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, 2009.

[7] D. Bertsimas and I. Popescu. Revenue management in a dynamic network environment. *Transportation science*, 37(3):257–277, 2003.

[8] Leslie M. Bobb and Emre Veral. Open issues and future directions in revenue management. *Journal of Revenue and Pricing Management*, 7(3): 291–301, 2008. doi: http://dx.doi.org/10.1057/rpm.2008.25.

[9] Michael Bowles. *Machine Learning in Python: Essential Techniques for Predictive Analysis*. John Wiley & Sons, Inc., 10475 Crosspoint Boulevard, Indianapolis, IN 46256, 2015.

[10] E.A. Boyd. Airline alliance revenue management. *OR/MS Today*, 25(1): 28–31, 1998.

[11] Leo Breiman. Bagging predictors. *Technical Report No. 421, Department of Statistics, University of California, Berkeley, California 94720*, 1994.

[12] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[13] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984.

[14] Sarah Beth Carney. Optimal and sequential design for bridge regression with application in organic chemistry. *University of Southampton*, 2010.

[15] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms using different performance metrics. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[16] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[17] W-C. Chiang, J.C.H. Chen, and X. Xu. An overview of research on revenue management: current issues and future research. *Int. J. Revenue Management*, 1(1):97–128, 2007.

[18] R. Cleaz-Savoyen. Airline revenue management methods for less restricted fare structures. *Ph.D. thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA*, 2005.

[19] C. Cleophas, M. Frank, and N. Kliewer. Recent developments in demand forecasting for airline revenue management. *Int. J. Revenue Management*, 3(3):52–269, 2009.

[20] Renwick E. Curry. Optimal airline seat allocation with fare classes nested by origins and destinations. *Transportation Science*, 24(3):193–204, 1990.

[21] J. D. Dana, Jr. Advancepurchase discounts and price discrimination in competitive markets. *Journal of Political Economy*, 106(2):395–422, 1998.

[22] H. Deng and E. Runger, G.and Tuv. Bias of importance measures for multi-valued attributes and solutions. *International Conference on Artificial Neural Networks, Business and Economics Statistics Section*, pages 293–300, 2011.

[23] M. A. R. T. A. Eso. An iterative online auction for airline seats. *IMA Volumes In Mathematics And Its Applications*, 127:45–58, 2001.

[24] Y. Feng and B. Xiao. A dynamic airline seat inventory control model and its optimal policy. *Operations Research*, 49(6):938–949, 2001.

[25] M. Ferguson, C. Crystal, J. Higbie, and R. Kapoor. A comparison of unconstraining methods to improve revenue management systems. *3rd edition, Working Paper, Georgia Institute of Technology*, 2007.

[26] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *AT&T Bell Laboratories*, 1995.

[27] G. Gallego and G. Van Ryzin. A multiproduct dynamic pricing problem and its applications to network yield management. *Operations Research*, 45(1):24–41, 1997.

[28] F. Glover, R. Glover, J. Lorenzo, and C. McMillan. The passenger-mix problem in the scheduled airlines. *Interfaces*, 12(3):73–80, 1982.

[29] ABHUIT GOSAVI, Naveen Bandla, and Tapas K. Das. A reinforcement learning approach to a single leg airline revenue management problem with multiple fare classes and overbooking. *IIE transactions*, 34(9):729–742, 2002.

[30] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

[31] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[32] Itir Karaesmen and Garrett van Ryzin. An airline overbooking model. *Operations Research*, 52(1):83–104, 2004. doi: http://www.jstor.org/stable/30036562.

[33] Michael Kearns. Thoughts on hypothesis boosting. *Machine Learning class project*, 1988.

[34] et al. Lan, Yingjie. Revenue management with limited demand information. *Management Science*, 54(9):1594–1609, 2008.

[35] Travis W. Gentry Lawrence R. Weatherford and Bogdan Wilamowski. Neural network forecasting for airlines: A comparative analysis. *Journal of Revenue and Pricing Management*, 1(4):319–331, 2003.

[36] Anthony O. Lee. Airline reservations forecasting: Probabilistic and statistical models of the booking process. *Ph.D. thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA*, 1990.

[37] K Littlewood. Forecasting and control of passenger bookings. *Proc. 12th AGIFORS Symposium*, 1972.

[38] Philip M. Long and Rocco A. Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010.

[39] J.C. MacKay, David. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[40] Donald W. Marquardt and Ronald D. Snee. Ridge regression in practice. *The American Statistician*, pages 3–20, 1975.

[41] Warren; Walter Pitts McCulloch. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[42] Jeffrey I. McGill and Garrett J. van Ryzin. Revenue management: Research overview and prospects. *Transportation Science*, 33(2):233–256, 1999. doi: http://dx.doi.org/10.1287/trsc.33.2.233.

[43] Tom Mitchell. *Machine Learning*. McGraw Hill, 1997.

[44] A. Moller, W. Romisch, and K. Weber. A new approach to o&d revenue management based on scenario trees. *Journal of Revenue and Pricing Management*, 3(3):265–276, 2004.

[45] R. Neuling, S. Riedel, and K-U. Kalka. New approaches to origin and destination and no-show forecasting: excavating the passenger name records treasure. *Journal of Revenue and Pricing Management*, 3:62–72, 2004.

[46] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[47] M. Reyes. Hybrid forecasting for airline revenue management in semi-restricted fare structures. *Ph.D. thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA*, 2006.

[48] Raul Rojas. Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Computer Science Department, Freie Universitat Berlin*, 2009.

[49] Marvin Rothstein. An airline overbooking model. *Transportation Science*, 5(2):180–192, 1971. doi: http://dx.doi.org/10.1287/trsc.5.2.180.

[50] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3), 1988.

[51] P. Rusmevichientong, Z. J. M. Shen, and D. B. Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58(6):1666–1680, 2010.

[52] J. I. McGill S. L. Brumelle. Airline seat allocation with multiple nested fare classes. *Operations Research*, 41(1):127–137, 1993.

[53] R. Shlifer and Y. Vardi. An airline overbooking policy. *Transp. Sci.*, 9(2): 101–114, 1975. doi: http://dx.doi.org/10.1287/trsc.9.2.101.

[54] B. C. Smith, J. F. Leimkuhler, and R. M. Darrow. Yield management at american airlines. *Interfaces*, 22(1):8–31, 1992.

[55] C. Stefanescu, V. de Miguel, K. Fridgeirsdottir, and S. Zenios. Revenue management with correlated demand forecasting. *Proceedings of the American Statistical Association, Business and Economics Statistics Section*, 2004.

[56] J. Subramanian, S. Stidham Jr, and C. J. Lautenbacher. Airline yield management with overbooking, cancellations, and no-shows. *Transportation Science*, 33(2):147–167, 1999.

[57] K. Talluri and G. Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

[58] Kalyan T. Talluri and Garret J. Van Ryzin. *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers, Boston, 2004.

[59] H. R. Thompson. Statistical problems in airline reservations control. *Oper. Res. Q.*, 12:167–185, 1961.

[60] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

[61] Armin Heinzl Tobias Grosche, Franz Rothlauf. Gravity models for airline passenger volume estimation. *Journal of Air Transport Management*, 13: 175–183, 2007.

[62] G. van Ryzin. Future of revenue management: models of demand. *Journal of Revenue and Pricing Management*, 4(2):204–209, 2005.

[63] Garrett van Ryzin and Jeff McGill. Revenue management without forecasting or optimization: An adaptive algorithm for determining airline seat protection levels. *Management Science*, 46(6):760–775, 2000.

[64] Larry Weatherford. The history of forecasting models in revenue management. *Journal of Revenue and Pricing Management*, page 1–10, 2016. doi: http://dx.doi.org/10.1057/rpm.2016.18.

[65] Larry R. Weatherford and Sheryl E. Kimesb. A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, 19:401–415, 2003.

[66] Larry R Weatherford and Stefan Polt. Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. *Journal of Revenue and Pricing Management*, 1(3): 234–254, 2002.

[67] Richard Robert Wickham. Evaluation of forecasting techniques for short-term demand of air transportation. *Ph.D. thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA*, 1995.

[68] Anton J. Kleywegt William L. Cooper, Tito Homem-de-Mello. Models of the spiral-down effect in revenue management. *Operations Research*, 54(5): 968–987, 2006.

[69] E. L. Williamson. Airline network seat inventory control: Methodologies and revenue impacts. *Ph.D. thesis, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA*, 1992.

[70] P. S. You. Dynamic pricing in airline seat management for flights with multiple flight legs. *Transportation Science*, 33(2):192–206, 1999.