ESTIMATING RATE OF NO-SHOW PASSENGERS:

A RESEARCH ON AN AIRLINE COMPANY IN TURKEY

ENES BAŞOĞLU

BOĞAZİÇİ UNIVERSITY

2024

ESTIMATING RATE OF NO-SHOW PASSENGERS:

A RESEARCH ON AN AIRLINE COMPANY IN TURKEY

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirement for the degree of

Master of Arts

in

Business Information Systems

by

Enes Başoğlu

Boğaziçi University

2024

Estimating Rate of No-Show Passengers:

A Research on an Airline Company in Turkey

The thesis of Enes Başoğlu

has been approved by:

Prof. Bertan Badur _____
(Thesis Advisor)

Prof. Sona Mardikyan _____

Assist. Prof. Tuğrul Hakyemez _____
(External Member)

April 2024

DECLARATION OF ORIGINALITY

I, Enes Başoğlu, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature _____

Date _____

ABSTRACT

Estimating Rate of No-Show Passengers:

A Research on an Airline Company in Turkey

This research aims to investigate the model that best predicts no-show rates for a flight in the civil aviation industry. Predicting no-show numbers on flights is crucial for effective operational planning and revenue management for airlines. It enables airlines to manage flight capacity more efficiently by minimizing empty seats and maximizing revenue. Accurate predictions aid in optimizing pricing strategies, allowing airlines to capitalize on empty seats effectively. Features were generated by aggregating from the ticket data, including region, scheduled day part, season, day of the week (Monday, Tuesday, etc.), first half of the year (first six months) or second half (last six months), month, proximity of official holidays based on origin and destination, flight time, and average no-show rates on previous flights. The study has employed seven predictive models: Random Forest, Bagging, Ada Boost, Gradient Boosting, Extreme Gradient Boost, Neural Networks, and Decision Trees. For comparison, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2) and Max Error values have been calculated and compared. Among the models, Decision Tree received the lowest performance score, while Gradient Boosting received the highest. Unlike previous studies, it has been observed that the previous one-month no-show averages are the most powerful variable in predicting the no-show rates of future flights. These findings were discussed in the light of previous research.

ÖZET

Rezervasyona Gelmeyen Yolcu Oranının Tahminlenmesi: Türkiye'deki Bir Havayolu Şirketi Üzerine Araştırma

Bu araştırma, sivil havacılık sektöründe bir uçuş için uçuşa gelmeme oranlarını en iyi tahmin eden modeli araştırmayı amaçlamaktadır. Uçuşlardaki rezervasyona gelmeme sayılarını tahmin etmek, havayollarının etkili operasyonel planlaması ve gelir yönetimi açısından çok önemlidir. Boş koltukları en aza indirip geliri en üst düzeye çıkararak havayollarının uçuş kapasitelerini daha verimli yönetmesine olanak tanır. Uçulan bölge, planlanan saat, sezon, haftanın günü (Pazartesi, Salı vb.), yılın ilk yarısı (ilk altı ay) veya ikinci yarısı (son altı ay), ay, kalkış ve varış noktasına göre resmî tatillerin yakınlığı, uçuş süresi ve önceki uçuşlardaki ortalama rezervasyonun kullanılmama oranları dahil olmak üzere bilet verilerinden özellikler oluşturulmuştur. Çalışmada Sinir Ağları, Karar Ağaçları, Rastgele Orman, Bagging, Ada Boost, Gradient Boosting, Extreme Gradient Boost, olmak üzere yedi sınıflandırma modeli kullanılmıştır. Karşılaştırma için Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2) and Max Error değerleri hesaplanıp kullanılmıştır. Modeller arasında en düşük performans puanını Karar Ağacı alırken, en yüksek performansı Gradient Boosting almıştır. Önceki çalışmalardan farklı olarak geçmiş bir aylık no-show ortalamalarının, gelecekteki uçuşların no-show oranlarını tahmin etmede en güçlü değişken olduğu gözlemlenmiştir. Bu bulgular daha önce yapılan araştırmalar ışığında tartışılmıştır.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: LITERATURE REVIEW 5

CHAPTER 3: MODEL BUILDING METHODOLOGY 11

3.1 Problem definition 11

3.2 Data preparation 11

3.3 Predictive models 21

3.4 Model evaluation criteria 22

CHAPTER 4: RESULTS 23

4.1 Model results 23

4.2 Model evaluation 29

4.3 Discussion 33

CHAPTER 5: CONCLUSION 36

REFERENCES 41

LIST OF TABLES

Table 1. Summary Statistics of No-show Percentages According to the Year …….12

Table 2. Summary Statistics of No-Show Percentages According to the Flights in the Year Part ……..…….…...…..…13

Table 3. Summary Statistics of No-Show Percentages According to the Season……..…13

Table 4. Summary Statistics of No-Show Percentages According to the Months….14

Table 5. Summary Statistics of No-Show Percentages According to the Days of Week.....15

Table 6. Summary Statistics of No-Show Percentages According to the Day Part…15

Table 7. Summary Statistics of No-Show Percentages According to the Flights on a Holiday……16

Table 8. Summary Statistics of No-Show Percentages According to the Region….18

Table 9. Summary Statistics of No-Show Percentages According to Flight Duration and Previous No-Show Percentages…….....……...20

Table 10. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2), and Max Error Values of the Models…...….24

Table 11. Correlation Coefficients Between Features and No-Show Percentage……...26

Table 12. Performance Scoring According to MAE…...……..….30

Table 13. Performance Scoring According to MSE…….……....30

Table 14. Performance Scoring According to RMSE……..……..31

Table 15. Performance Scoring According to R^2……..…...32

Table 16. Performance Scoring According to Max Error……..……32

Table 17. Total Performance Scores……...……...33

CHAPTER 1

INTRODUCTION

The no-show term is generally used to mean that the person who has agreed to purchase a service or product for a certain period of time is not present at the time of receipt of the service without notice or cancellation. (AlMuhaideb, Alswailem, Alsubaie, Ferwana, & Alnajem, 2019; Chin, Ting, & Cham, 2023; Topuz, Uner, Oztekin, & Yildirim, 2018; Zenkert, 2017). The absence of a person in a hospital at the time of the appointment can be seen as an example of a no-show in the health sector and in the airline sector, where a person receiving a flight ticket is not there at the departure time. This is an important concept that companies encounter in their revenue-management and resource-planning processes. (Berg et al., 2013).

As is well known, the aviation sector is growing every day. According to a report by the International Civil Aviation Organization (ICAO), demand for air transport is expected to increase by an average of 4.3 percent per year by 2036. The situation in Turkey is not different. According to an aviation report by the T.C. Ministry of Transport and Infrastructure published in 2023, by the end of 2021, the airport will be ranked 2nd in Europe and 6th in the world in terms of passenger traffic. It was ranked 7th among European countries in 2003, 5th in 2016, and 2nd in 2020. It was ranked 18th in the world in 2003, eleventh in 2016, and seventh in 2020. This growth in civil aviation is thus becoming an opportunity for airlines and an incentive for companies to identify new strategies in a competitive market to increase their revenue.

No-shows have a significant impact in the aviation sector, as in other sectors. Efficiency becomes crucial because the costs of the services provided by air transport

are too high and they are not easy to reduce. In order to increase profitability, airlines need to minimize vacant seat flight activity and fly aircraft at maximum capacity. Despite getting a ticket at this point, seats for passengers who do not board the flight are an important opportunity (Gorin, Brunger, & White, 2006). Therefore, many airlines sell overcapacity tickets on certain assumptions to increase their revenue.

Typically, overbooking choices are made by operation experts. To minimize bias and the procedure's reliance on the individual, choices must be supported by a separate automated process. Forecasts that are clearly accurate can increase airline profitability. Usage prediction is therefore regarded as a significant subject by academics and industry professionals (Somboon, & Amaruchkul, 2017).

Airlines may suffer a major loss of reputation as a result of passenger complaints spreading to broad customers, especially in light of the increased usage of social media in recent times. Airlines are at significant risk from this since excessive overbooking could result in issues including dissatisfied customers, harm to their reputation, and revenue losses (Perez, 2021).

If someone fails to show up at their scheduled service time, it is not only a waste of time; it may also result in an appointment or reservation reserved for the same person not being filled. The cost of cancellation and non-use of the reservation is real pain and suffering, and it can happen often in many areas of competence in many locations. This can result in lower revenues, less human resource utilization and loss of commissions, and lower staff morale. According to a healthcare center study, a daily net gain of $4433.32 per day without a no-show was to be a net loss of $725,42 per day, with an 18% no-show rate. This represents a net loss of 16 percent per day. Again, the same study found that from 36% to 76.9% of losses caused by no-shows could be compensated with overbooking (Berg et al., 2013).

It is difficult to find factors that influence the prediction of no-shows. Passenger purchasing habits can be affected by many sub-factors consisting of social, political, and economic changes. We know that civil aviation, in particular, is vulnerable to these developments. One of the sectors most affected during the pandemic period was civil aviation transportation. While passenger transportation has come to a dramatic halt, there has been a sharp increase in cargo flights due to the influence of online shopping and cargo transportation (Adediran, 2020; Lim, Yuan, Khaw, & Chew, 2023). In fact, there is a very wide set of variables that can affect customer behavior in the transportation industry. Large data sets contain thousands of variables, making them difficult to handle and manage efficiently with traditional techniques. As a result, variable selection has been the subject of many studies in various fields by experts without using modern forecasting techniques in the aviation industry (Chowdhury, & Turin, 2020). Finally, building accurate predictive models is important for business, finance, healthcare, etc. It is a big task in many industries, including creating predictive models that can estimate future no-show rates using historical data and relevant factors has gained popularity recently. Well-designed forecasting models can help businesses and organizations make informed decisions, save money, and improve their financial results.

The aim of this study is to predict future flights by developing an analytical data model that explains the predictive variables that lead to passenger no-shows. Once the data set was created, predictive models were developed to determine which model performed best by predicting no-show rates. In the study, prediction models such as Neural Networks, Decision Trees and Random Forest, Bagging, Ada Boost, Gradient Boosting and Extreme Gradient Boost algorithms were used, and their results were compared.

This study consists of five chapters in total. Chapter 2 provides an overview of the existing literature on determining no-show rates in civil aviation. Chapter 3 focuses on the development methodology, including problem definition, data preparation, model building, and model evaluation. In chapter 4, you will find the details of the findings. Chapter 5 concludes the study, evaluates the findings of this study, mentions the limitations of the study, and offers various suggestions for further studies.

CHAPTER 2

LITERATURE REVIEW

To develop a better prediction model by taking advantage of previous studies, the literature was examined. In this context, it is aimed at contributing to the literature by conducting an in-depth literature review and trying new approaches. For this purpose, in this chapter, previous studies conducted in the aviation sector are mentioned and their results are evaluated.

When a comprehensive literature review was conducted, it was seen that the pioneers of no-show research in the civil aviation sector were generally based on statistical analysis of cancellations and no-show rates based on historical observations, as part of overbooking research. Recently, it has been observed that ticket data coming with reservations is mostly included in these studies.

An airline has to know who cancels far in advance of the arrival date or does not show up, since this information is essential for figuring out how much overbooking is suitable. Faraway and Chatfield (1998) applied neural networks to anticipate passenger numbers in airline data with considerable seasonality variance during a time when neural networks were extensively pushed in the literature. This made it possible to estimate the number of people who failed to arrive. According to some past research, show rates for closed flights can reach as high as 15%, which might result in a significant portion of empty seats (Smith, Leimkuhler, & Darrow, 1992).

In a study conducted at Lufthansa German Airlines, Neuling, Riedel, and Kalka (2004) estimated the number of no-shows using data mining techniques on reservation-based data. In the research on overbooking in aviation conducted by

Suzuki (2002), it was seen that extra sales made positive contributions to companies in the long term.

Several studies highlight how crucial it is to predict the number of no-shows. Lawrence, Hong, and Cherrier's (2003) study was among the first to highlight the importance of this problem. According to their research, a carrier's profits could be increased by reducing the number of denied passengers loading up and the number of vacant seats that could have been sold by correctly predicting no-show passengers for each flight correctly. Garrow and Koppelman (2004) carried out a further comparative study one year later.

A study by Garrow and Koppelman (2004a) found that traveling in high-cost classes, traveling in groups, and having a loyalty membership reduced the number of no-shows. Additional services and ticket prices are therefore important indicators among the variables that determine whether passengers will come or not on their flights. On the other hand, in this study, it was observed that the days of the week were insufficient to determine the no-show behavior, while the hours of the flight were the significant factor. As a result of the study, more no-shows were observed on early morning flights and late-night flights. In another study done by the same researchers to estimate the number of no-shows, it was observed that passengers with e-tickets were significantly less likely to have a no-show (Garrow and Koppelman, 2004b). In these two studies, it was observed that inbound flights tended to show more no-shows than outbound flights. Expensive ticket prices and the high cost that may arise due to postponement of bookings made abroad may be persuasive in bringing passengers to the flight.

Researchers raised that predicting the number of passengers that will not show up on a flight could help with many business decisions. Various examinations

in the past have proposed and supported the utilization of Case Based Reasoning (CBR) in aviation. Chang, Hsieh, Yeh, and Liu (2006) managed to increase the performance of the seat allocation system by using dynamic probability with CBR, giving better results compared to the old methods. As revealed in this study, the proposed system presented with the Case Based Reasoning method had many advantages. The framework had the option to learn and adjust by gaining new cases, and the proposed arrangement was exceptionally clear and straightforward.

Fildes, Nikolopoulos, Crone, and Syntetos (2008) have demonstrated via their extensive analysis of these models' applications that forecasting models are a necessary tool for airlines. They also need to choose a forecasting system that is suitable and capable of analyzing many large data sets at once, such as the ones used for no-show forecasting. Unnumkal, Talluri, and Topaloglu (2012) used a randomized linear programming method to address the network revenue management issue while taking cancellations and/or no-shows into consideration.

After that, a different researcher from Malaysia estimated the number of travelers over time using the Geometric Brownian Motion (GBM) method (Asrah, Nor, Rahim, & Leng, 2018). A rather straightforward mathematical model called GBM is used to forecast future share prices, frequently over a brief time frame. The distributional behavior data of passengers on two nearby airlines is what the researcher is comparing. Additionally, a Swedish researcher used data from passenger bookings to estimate the percentage of passengers that did not travel (Zenkert, 2017). To find out if decision trees, gradient boosting, or neural networks could outperform the basic model, the researcher employed a range of methods for the prediction. Therefore, on some key performance indicators, slope boosting

yielded results that were similar to decision trees but slightly worse. The neural network performed the worst out of all the models.

Numerous airlines base their decisions about how many seats to permit overbooking on estimations of the number of passengers who are unlikely to change their minds about a particular flight. In order to address this difficulty, a study suggested a decision support system that extracts suggestions from both experts and algorithms and forecasts the number of people who would not show up (Petrović & Milošević, 2021). This system integrates the CBR method with Interpolative Boolean Algebra (IBA). In addition, one study predicts the number of passengers flying in Malaysia using the Box-Jenkins model and an artificial neural network model with delay variables as input variables (Idrus, & Mohamed, 2020).

In the airline business, demand for seats is erratic, capacity is constrained and difficult to increase, and costs are prohibitively high. Therefore, demand forecasting is of utmost importance to the sector. Based on previous data, a researcher developed an ideal model with a multilayer feedforward neural network to forecast passenger demand at flight origination and arrival levels (Tissera, Waduge, Perera, Nawinna, & Kasthurirathna, 2020). In order to guarantee the greatest outcomes in capacity utilization decisions, researchers conducted a study with data such as area, departure location, date, and number of passengers to predict customer demand.

In the study conducted by Chin, Ting, and Cham (2023), meaningful findings were obtained using reservation data. The models compared include Random Forest, Decision Tree, Gradient Boosting, and Neural Networks. With Random Forest leading the pack at an accuracy rate of 90.4%, followed closely by Decision Tree at 90.2%, the research sheds light on the effectiveness of these models in managing and potentially decreasing the issue of no-shows in the aviation sector. The relatively

lower performance of Neural Networks at 67.6% indicates that not all advanced algorithms perform equally and that choosing the right model is pivotal.

In the medical field, patients who skip outpatient appointments for diagnostic or clinical testing are referred to as "no-shows." In order to optimize resource utilization and enhance healthcare efficiency, physicians and healthcare facilities need to recognize these patients. For instance, a researcher developed a predictive model to forecast patients' nonattendance at scheduled appointments using decision trees and AdaBoost (Alshammari, Almalki, & Alshammari, 2021). Model analysis reveals that patients who miss appointments are typically younger, male, scheduled in the morning, and do not get phone calls or reminders. This research points out that the tendency to miss appointments is associated with certain patient characteristics and behaviors. Key findings suggest that younger patients, particularly those between 25 and 34 years old, have higher no-show rates compared to older patients, with rates decreasing linearly as age increases. It's also observed that patients without a reminder, such as an SMS, are more likely to not show up.

The study uses important attributes like diabetes, hypertension, alcoholism, lead time (the duration between the booking and the appointment date), scheduling time, handicap status, and whether an SMS reminder was received to predict no-show probabilities. Machine learning models, such as AdaBoost and Decision Tree, were trained on these attributes to produce predictions. Although both models show promise, the DT model was highlighted for its higher performance, despite some limitations like the complexity and robustness of the decision tree algorithm.

It's essential to monitor the performance of such predictive models and to regularly update them with new data to ensure accuracy. Furthermore, the consideration of broader factors, including environmental, educational, and cultural

influences, may help in better understanding and mitigating no-show behavior. This holistic approach toward predicting and managing no-show rates can lead to significant improvements in healthcare delivery and operational effectiveness. In order to assist hospitals in making defensible decisions and lessen the probability of patient no-show behavior, Chinese researchers also created a prediction model for patient no-shows (Fan, Deng, Ye, & Wang, 2021).

CHAPTER 3

MODEL BUILDING METHODOLOGY

In this chapter, the purpose of the study and the problem it focuses on are mentioned in Section 3.1, and then how the data was prepared, the variables, and their descriptive statistics are included in Section 3.2. The algorithms used in the models are mentioned in Section 3.3, and the model evaluation criteria are included in Section 3.4.

3.1 Problem definition

No-show rates pose a significant challenge in various sectors, such as civil aviation, healthcare, hospitality, and service industries, impacting operational efficiency and revenue. Factors influencing no-show rates may include demographic information, appointment scheduling patterns, historical attendance records, external variables like weather conditions, and any specific characteristics of the service being provided. By understanding and quantifying these factors, the aim is to devise strategies to minimize no-shows, optimize resource allocation, and enhance overall customer satisfaction and organizational performance. The main focus of this study is how best to estimate no-show rates. For this purpose, the variables determining the no-show rates were identified, and the model with the highest performance was tried to be determined by using different algorithms.

3.2 Data preparation

All flight data between the years 2022 and 2023 was used to estimate no-show rates. 160 million ticket-level data were aggregated and converted into flight data, and

characteristic variables of the flight data were created from the ticket information. While calculating no-show rates for flights, all tickets that have been booked and missed the flight in some way are included, while reservation cancellations are not considered. A total of 778,879 flight data were obtained after eliminating outliers noticed in the descriptive analysis due to disruptions based on data that was corrupted or that occurred due to reasons that were difficult to predict. The summary statistics of no-show percentages by year are shown in Table 1.

Table 1. Summary Statistics of No-Show Percentages According to the Year

Year	*N*	*f*	Mean	*SD*	Min	Max
2023	778879	414492	4.38	2.96	0	19.89
2022	778879	364387	4.28	2.9	0	19.8

A total of 74 features were generated by aggregating from the ticket data, including region, scheduled day part, season, day of the week (Monday, Tuesday, etc.), first half of the year (first six months) or second half (last six months), month, proximity of official holidays based on origin and destination, flight time, and average no-show rates on previous flights.

The year part variable was created by marking the flights that took place in the first six months and the last six months of the year according to the local schedule departure date over two years of flight data. Summary statistics of no-show percentages according to the year part are shown in Table 2.

Table 2. Summary Statistics of No-Show Percentages According to the Flights in the Year Part

Year Part	*f*	Mean	SD	Min	Max
First Six Months	363429	4.41	2.92	0	19.89
Last Six Months	754639	4.35	2.94	0	19.89

The season variable was created by marking the flights according to the seasons of the year according to the local schedule departure date over two years of flight data. Summary statistics of no-show percentages by season are shown in Table 3.

Table 3. Summary Statistics of No-Show Percentages According to the Season

Season	*f*	Mean	*SD*	Min	Max
Summer	187087	4.33	2.85	0	19.89
Spring	233055	4.11	2.95	0	19.87
Autumn	207576	4.31	2.92	0	19.79
Winter	151161	4.72	3	0	19.8

The month variable is created by marking flights based on the month of the local schedule departure date over two years of flight data. Summary statistics of no-show percentages by season are shown in Table 4.

Table 4. Summary Statistics of No-Show Percentages According to the Months

Month	*f*	Mean	SD	Min	Max
January	54359	4.7	3.03	0	19.8
February	47846	4.9	3.01	0	19.46
March	56302	4.64	2.93	0	19.71
April	59826	4.21	2.8	0	19.53
May	70959	4.18	2.8	0	19.89
June	74137	4.11	2.88	0	19.87
July	78618	4.1	2.95	0	19.34
August	80300	4.13	3.01	0	19.69
September	76316	4.25	2.96	0	19.76
October	69631	4.33	2.92	0	19.79
November	61629	4.37	2.86	0	19.77
December	48956	4.56	2.95	0	19.8

The days of week variable is created by marking flights based on days of the local schedule departure date over two years of flight data. Summary statistics of no-show percentages by season are shown in Table 5.

Table 5. Summary Statistics of No-Show Percentages According to the Days of Week

Days of week	*f*	Mean	SD	Min	Max
Monday	111518	4.37	2.91	0	19.77
Tuesday	107275	4.34	2.92	0	19.76
Wednesday	108006	4.34	2.93	0	19.87
Thursday	108494	4.36	2.95	0	19.37
Friday	114980	4.37	2.96	0	19.89
Weekend	228606	4.29	2.93	0	19.86

The day part variable is created by marking the flights taking place according to the local schedule departure date time over two years of flight data. 'Morning' flight from 6:00 to 11:59, 'Afternoon' from 12:00 to 17:59, 'Evening' flight from 18:00 to 20:59, and 'Night' flight from 21:00 to 5:59 in the morning are marked as. Summary statistics of no-show percentages by day part are shown in Table 6.

Table 6. Summary Statistics of No-Show Percentages According to The Day Part

Day Part	*f*	Mean	SD	Min	Max
Night	175950	4.29	3.02	0	19.8
Morning	251359	4.19	2.68	0	18.1
Evening	128098	4.73	3.2	0	19.71
Afternoon	223472	4.31	2.96	0	19.89

Public holidays close to the departure and arrival days of each flight point were found and calculated according to their proximity. If at least one of the departure or arrival points is close to a public holiday, that flight is marked as close to a public holiday. A total of six independent variables were created: on a public holiday, within one day, within two days, within three days, within four days, and within a week. Summary statistics of no-show percentages by day part are shown in Table 7.

Table 7. Summary Statistics of No-Show Percentages According to Flights on a Holiday

	f	Mean	SD	Min	Max
On a holiday					
Yes	24240	4.03	2.74	0	17.95
No	754639	4.35	2.94	0	19.89
One Day Before a Holiday					
Yes	28548	4.02	2.7	0	19.57
No	750331	4.35	2.94	0	19.89
Two Days Before a Holiday					
Yes	24853	4.01	2.68	0	17.58
No	754026	4.35	2.94	0	19.89
Three Days Before a Holiday					
Yes	23660	4.04	2.71	0	18.6
No	755219	4.34	2.94	0	19.89
Four Days Before a Holiday					
Yes	21720	4.03	2.74	0	18.6
No	757159	4.34	2.94	0	19.89
One Week Before a Holiday					
Yes	158945	4.05	2.75	0	19.8
No	619934	4.41	2.98	0	19.89

Each flight is divided into 11 regions in total, according to the departure and arrival countries: Domestic, Sub-Saharan Africa, Africa, Central Europe, Southern Europe, Middle East-Cyprus, North Africa, Northern Europe, Eastern Europe and Balkans, Asia-Fareast and America. Summary statistics of no-show percentages by region are shown in Table 8.

Table 8. Summary Statistics of No-Show Percentages According to the Region

Origin Region	Destination Region	*f*	*Mean*	*SD*	*Min*	*Max*
Domestic	Domestic	308354	4.91	3.24	0	19.89
Central Europe	Domestic	52654	2.85	1.83	0	15.41
Domestic	Central Europe	46678	4.62	2.57	0	15.61
Middle East-Cyprus	Domestic	46328	3.45	2.39	0	18.24
Domestic	Middle East-Cyprus	39274	5.16	2.97	0	18.25
Northern Europe	Domestic	38909	2.68	2.07	0	14.89
Domestic	Northern Europe	35474	3.96	2.87	0	16.1
Southern Europe	Domestic	33020	3.63	2.23	0	18.78
Domestic	Southern Europe	28791	4.97	2.69	0	18.29
Asia-Fareast	Domestic	22878	2.59	1.86	0	16.6
Domestic	Eastern Europe and Balkans	21260	5.3	2.83	0	18.1
Eastern Europe and Balkans	Domestic	20791	3.27	2.38	0	17.22
Domestic	Asia-Fareast	20587	4.07	2.6	0	18.32
America	Domestic	14470	4.21	2.48	0	17.99
Domestic	America	14276	4.79	2.37	0	16.36
North Africa	Domestic	10193	3.96	2.25	0	13.38
Domestic	North Africa	8236	5.89	3.06	0	16.23
Domestic	Sub-Saharan Africa	7482	5.75	3.36	0	19.38
Sub-Saharan Africa	Domestic	6345	4.3	2.87	0	17.65
Africa	Domestic	1619	4.27	2.09	0	12.75
Domestic	Africa	1260	7.21	3.39	0	19.46

In the research, flight duration and no-show percentages of the previous days were used to estimate the no-show percentages. To calculate the flight duration, the time zones of the countries were included, and the difference between arrival and departure times was used. To calculate the no-show percentages of previous days, the "No-Show Percentage One Day Before" variable was created by using the no-show rate of a flight on the same route and time of the previous day. Using the same method, no-show percentages until seven days before the flight were calculated and included in the study as a variable. Finally, the "No-Show Percentage Average for Last One Week" variable was created by taking the no-show averages of the flights one week before the flight, and the "No-Show Percentage Average for Last One Month" variable was created by taking the no-show averages of the flights one month ago. Summary statistics of these variables are shown in Table 9.

Table 9. Summary Statistics of No-Show Percentages According to Flight Duration and Previous No-Show Percentages

Variable	*N*	Min	Max	Mean	SD	Median	Mode
Flight Duration	778879	50	883	189.89	155.97	138	72
No-Show Percentage One Day Before	778879	0	19.94	4.64	3.53	3.82	0
No-Show Percentage Two Days Before	778879	0	19.96	4.64	3.53	3.82	0
No-Show Percentage Three Days Before	778879	0	19.96	4.65	3.53	3.82	0
No-Show Percentage Four Days Before	778879	0	19.96	4.65	3.53	3.82	0
No-Show Percentage Five Days Before	778879	0	19.96	4.65	3.53	3.82	0
No-Show Percentage Six Days Before	778879	0	19.96	4.65	3.54	3.82	0
No-Show Percentage Seven Days Before	778879	0	19.96	4.65	3.54	3.82	0
No-Show Percentage Average for Last One Week	778879	0	18.29	4.65	2.5	4.18	0
No-Show Percentage Average for Last One Month	778879	0	17.62	4.68	2.31	4.29	0
No-Show Percentage	778879	0	19.89	4.34	3.82	4.34	0

3.3 Predictive models

In this study, flight data from an airline company in Turkey was used to estimate no-show rates. A total of seven predictive models were used, including Neural Networks and Decision Trees, which are singular learning techniques, and Random Forest, Bagging, Adaboost, Gradient Boosting, and Extreme Gradient Boost, which are ensemble learning techniques, and their results are compared.

3.3.1 Single learning techniques

The neural network model has been applied using Keras. First, variables have been scaled using Sklearn's Min Max Scaler for neural networks. The correlation-based feature selection method (CFS) was applied to detect the most correlated variables among the 74 candidate features. Among all variables, 30 variables with the highest correlation with a no-show percentage have been selected. The model then has begun with the input shape set to 30, indicating that just the top 30 features have been supplied into it. Using the Relu activation function, the architecture consists of two hidden layers with 32 and 20 units, respectively. There has been just one unit with sigmoid activation in the output layer. Mean Absolute Error (MAE) loss and MAE metrics have been selected, together with the Stochastic Gradient Descent (SGD) optimizer, in order to assemble the model.

The decision tree model has been imported from the Scikit-Learn library using the Decision Tree Regressor algorithm, which recursively partitions the dataset into subsets based on the features that minimize the mean squared error (MSE) of the target variable. To choose the split at each node, the “best” splitter has been used. The model has been applied with unrestricted maximum depth, two minimum

samples per split, and one minimum sample per leaf with a random state, and has been tested with 25 percent of the test data.

3.3.2 Ensemble learning techniques

Sklearn has been used to implement the Random Forest Regressor. After trying many different numbers of estimators, a Random Forest model has been created with a random state consisting of 100 estimators (decision trees) and a random state of 1, which is fast in terms of performance and allows for an accurate performance evaluation when compared with other algorithms. The random state has allowed the results to be compared between runs.

The bagging regression for the decision tree has been imported from the Sklearn library. After the model has been fitted with 100 decision trees with 3 max depts, it has been tested with 25 percent test data. Later, Sklearn has been used in the application of Adaboost, Gradient Boosting, and Extreme Gradient Boost regressors. While training the model, a number of 100 decision trees, three max depts, and least square regression have been used as the loss function and have been tested with 25 percent of the test data.

3.4 Model evaluation criteria

The no-show percentage is calculated by dividing the number of reservations by the number of passengers who do not show up. As the dependent variable is a continuous variable in predictions made according to seven different models, regression is used. For comparison, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2) and Max Error values are calculated and compared.

CHAPTER 4

RESULTS

Seven different predictive models are applied to estimate the no-show percentages, and their results on test data are compared. Model results are given in Section 4.1, and comparisons of model results are given in Section 4.2. In Section 4.3, the results are discussed within the framework of the literature.

4.1 Model results

No-show values are obtained by calculating the percentage of passengers who do not show up for the flight based on the number of reservations. In the predictions made according to seven different models, regression is used, and Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2) and Max Error values of the models are shown in Table 10.

Table 10. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2) and Max Error Values of the Models

Model	MAE	MSE	RMSE	R^2	Max Error
AdaBoost	1.65	3.99	1.99	0.53	5.77
Bagging	1.65	4.25	2.06	0.50	7.86
Gradient Boosting	1.56	3.77	1.94	0.55	5.67
Decision Tree	2.23	7.95	2.82	0.08	10.15
Random Forest	1.61	4.00	2.00	0.53	7.22
Extreme Gradient Boost (XGBoost)	1.56	3.77	1.94	0.56	6.16
Neural Network	1.58	3.88	1.95	0.55	6.33

The importance of features is compared in the predictive model where the Decision Tree algorithm is used. Decision Tree importance levels reflect the contributions of the features used in the construction of the decision tree in predicting the target variable. These values are normalized between 0 and 1, set to sum to 1. Therefore, a feature with a higher value has more importance. No-Show Percentage Average for Last One Month (0.48) has the highest value in the order of importance of the variables of the model created with the Decision Tree algorithm. The importance of the variable No-Show Percentage Average for Last One Week (0.15) is in second place. The variables No-Show Percentage One Day Before, No-Show Percentage Two Days Before, No-Show Percentage Three Days Before, No-Show Percentage Four Days Before, No-Show Percentage Five Days Before, No-Show Percentage Six Days Before, No-Show Percentage Seven Days Before and Flight Duration (0.04) each have the same importance. The significance of the remaining

variables is 0.0. The impact of these variables on the model is insignificant. Figure (1) shows the decision tree plot of the model for predicting no-show percentages.

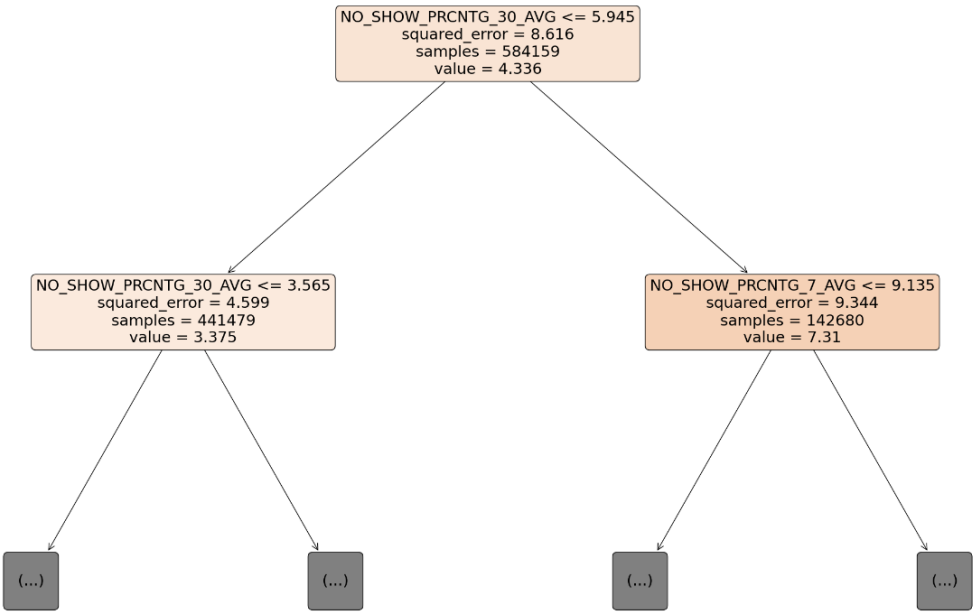

Figure 1. Decision tree plot of model for predicting no-show percentages

In the study, the correlation-based feature selection method (CFS) is applied to detect the most correlated variables among the 74 candidate features for the neural network. Among all variables, 30 variables with the highest correlation with no-show percentage are selected, and their correlation coefficients are shown in Table 11.

Table 11. Correlation Coefficients Between Features and No-Show Percentage

	Correlation Coefficients
	No-show Percentage
No-Show Percentage Average for Last One Month	0.73
No-Show Percentage Average for Last One Week	0.72
No-Show Percentage One Day Before	0.53
No-Show Percentage Two Days Before	0.52
No-Show Percentage Three Days Before	0.51
No-Show Percentage Six Days Before	0.51
No-Show Percentage Seven Days Before	0.51
No-Show Percentage Four Days Before	0.51
No-Show Percentage Five Days Before	0.51
Origin Domestic First	0.35
Origin Domestic	-0.17
Origin Central Europe	-0.14
Origin Northern Europe	-0.13
Origin Asia-Fareast	-0.10
Flight Duration	-0.09
Origin Middle East Cyprus	-0.08
Destination Domestic First	-0.07
Destination Middle East Cyprus	0.06
Winter Flight	0.06
Origin Eastern Europe and Balkans	-0.06
Evening Flight	0.06
Origin Eastern Europe and Balkans	0.06
Destination North Africa	0.05
Origin Southern Europe	-0.05
Summer Flight	-0.05
February Flight	0.05
Destination Sub-Saharan Africa	0.05
Destination Southern Europe	0.04
Destination Africa	0.04

The Bagging Regressor from Sklearn is used for the other model. To check feature importance, the averages of each ensemble tree are used. These importance values indicate the relative contribution of each feature to predicting the target variable. No-Show Percentage Average for the Last One Month (0.51) holds the highest importance. No-Show Percentage Average for Last One Week (0.12) is following behind. Other variables such as No-Show Percentage One Day Before, No-Show Percentage Two Days Before, No-Show Percentage Three Days Before, No-Show Percentage Four Days Before, No-Show Percentage Five Days Before, No-Show Percentage Six Days Before, No-Show Percentage Seven Days Before, and Flight Duration (0.04) each have the same importance. The remaining variables, including seasonal ticket purchases and regional details, show no discernible importance in the model. Their impact on predicting 'no-show' behavior is minimal, as indicated by their importance scores of 0.0.

When calculating the importance of a particular variable, XGBoost combines the attribute importance of the decision trees at the nodes of all trees in which that variable is used. Consequently, the importance of a variable is the sum of the contributions of this variable across all trees and all nodes. The importance order and importance percentages of the variables of the prediction model made with the XGBoost algorithm are calculated, and the No-Show Percentage Average for Last One Month (0.58) variable is identified as the most important variable of the model. The No-Show Percentage Average for Last One Week (0.30) variable is the second most important feature of the model. The significance of other variables is 0.01: No-Show Percentage One Day Before, Origin Domestic, Origin Central Europe, Flight Duration, Destination Domestic, Destination Cyprus, Origin Southern Europe, and

Summer Ticket. The remaining variables have a significance of 0% or less and do not affect model performance.

Gradient Boosting from the Sklearn library reports the importance of each feature used in the decision tree model. These importance values indicate the relative contribution of each feature to predicting the target variable. The values are normalized and range between 0 and 1, with a higher value indicating greater importance. The importance ranking of variables in a prediction model was created using the Gradient Boosting algorithm. According to the model results, the No-Show Percentage Average for the Last One Month (0.51) has the highest importance. No-Show Percentage Average for Last One Week (0.48) is following closely behind. The remaining variables have negligible importance in the model. Their impact on predicting 'no-show' behavior is minimal, as indicated by their importance scores of 0.0.

The Random Forest Regressor from the Sklearn library reports the importance of each feature used in the Random Forest model. These importance values indicate the relative contribution of each feature to predicting the target variable. The importance ranking of variables in a prediction model is created using the Random Forest algorithm. No-Show Percentage Average for the Last One Month (0.51) holds the highest importance. No-Show Percentage Average for Last One Week (0.12) is following behind. Other variables such as No-Show Percentage One Day Before, No-Show Percentage Two Days Before, No-Show Percentage Three Days Before, No-Show Percentage Four Days Before, No-Show Percentage Five Days Before, No-Show Percentage Six Days Before, No-Show Percentage Seven Days Before, and Flight Duration (0.04) each have the same importance. The remaining variables, including seasonal ticket purchases and regional details, show

no discernible importance in the model. Their impact on predicting 'no-show' behavior is minimal, as indicated by their importance scores of 0.0.

The feature importance function is used with the Ada Boost Regressor from the Sklearn library. These importance values indicate the relative contribution of each feature to predicting the target variable. However, it's important to note that AdaBoost does not inherently provide feature importance estimates like some other algorithms. Instead, it indirectly calculates feature importance based on how often features are used in the ensemble of weak learners. Features used more often are considered more important. According to the model results, the No-Show Percentage Average for the Last One Month (0.51) has the highest importance. No-Show Percentage Average for Last One Week (0.47) is following closely. The remaining variables, including day-specific 'no-show' percentages, regional details, and seasonal ticket purchases, showed negligible importance in the model. These variables have minimal impact on predicting 'no-show' behavior within the AdaBoost framework.

4.2 Model evaluation

After calculating the MAE values of each of the seven models, the model that gives the best results in the comparison is the Gradient Boosting and Extreme Gradient Boost models. Since both models have equal MAE values, they are both given six points. Subsequently, the neural network becomes the model that gives the second-best result, and therefore five points are given. According to the MAE score, it is followed by Random Forest, AdaBoost, Bagging, and Decision Tree, respectively. Details of MAE scores are shown in Table 12.

Table 12. Performance Scoring According to MAE

Model	MAE	Given Performance Score
Gradient Boosting	1.56	6
Extreme Gradient Boost	1.56	6
Neural Network	1.58	5
Random Forest	1.61	4
AdaBoost	1.65	3
Bagging	1.65	3
Decision Tree	2.23	2

The MSE values of seven models are compared and scored. According to MSE values, the Gadient Boosting and Extreme Gradient Boost models achieved the best results. They received equal points because they have the same values. The Neural Network model that followed received five points. Afterwards, AdaBoost, Random Forest, Bagging, and Decision Tree models were followed and scored as in Table 13.

Table 13. Performance Scoring According to MSE

Model	MSE	Given Performance Score
Gradient Boosting	3.77	6
Extreme Gradient Boost	3.77	6
Neural Network	3.88	5
AdaBoost	3.99	4
Random Forest	4.00	3
Bagging	4.25	2
Decision Tree	7.95	1

After calculating the RMSE values of each of the seven models, the model that gave the best results in the comparison was the Gradient Boosting and Extreme Gradient Boost models. Since both models had equal RMSE values, they were both given six points. Subsequently, Neural Network became the model that gave the second-best result, and therefore five points were given. According to the RMSE score, it was followed by AdaBoost, Random Forest, Bagging, and Decision Tree, respectively. Details of RMSE scores are shown in Table 14.

Table 14. Performance Scoring According to RMSE

Model	RMSE	Given Performance Score
Gradient Boosting	1.94	6
Extreme Gradient Boost	1.94	6
Neural Network	1.95	5
AdaBoost	1.99	4
Random Forest	2.00	3
Bagging	2.06	2
Decision Tree	2.82	1

The R^2 values of seven models were compared and scored. According to R^2 values, the Extreme Gradient Boost model achieved the best results. Gradient Boosting and Neural Network received equal points because they have the same values. The Random Forest model that followed received five points. Afterwards, AdaBoost, Bagging, and Decision Tree models were followed and scored as in Table 15.

Table 15. Performance Scoring According to R^2

Model	R^2	Given Performance Score
Extreme Gradient Boost	0.56	6
Gradient Boosting	0.55	5
Neural Network	0.55	5
Random Forest	0.53	4
AdaBoost	0.53	4
Bagging	0.50	3
Decision Tree	0.08	2

After calculating the Max Errors values of each of the seven models, the model that gave the best results in the comparison was Gradient Boosting. It's followed by AdaBoost, which scored five. According to the Maximum Errors, it was followed by Extreme Gradient Boost, Neural Network, Random Forest Bagging, and Decision Tree, respectively. Details of Max Error scores are shown in Table 16.

Table 16. Performance Scoring According to Max Error

Model	Max Error	Given Performance Score
Gradient Boosting	5.67	6
AdaBoost	5.77	5
Extreme Gradient Boost	6.16	4
Neural Network	6.33	3
Random Forest	7.22	2
Bagging	7.86	1
Decision Tree	10.15	0

The scores given according to each criterion were calculated for seven models, and a total value was obtained. These obtained values are shown in Table 17, from highest to lowest.

Table 17. Total Performance Scores

Model	MAE Scores	MSE Scores	RMSE Scores	R^2 Scores	Max Error Scores	Total Performance Score
Gradient Boosting	6	6	6	5	6	29
Extreme Gradient Boost	6	6	6	6	4	28
Neural Network	5	5	5	5	3	23
Ada Boost	3	4	4	4	5	20
Random Forest	4	3	3	4	2	16
Bagging	3	2	2	3	1	11
Decision Tree	2	1	1	2	0	6

As seen in Table 17, among the models, Decision Tree received the lowest performance score, while Gradient Boosting received the highest. It has been observed that the previous one-month no-show averages are the most powerful variable in predicting the no-show rates of future flights.

4.3 Discussion

The quest to accurately predict no-show rates, whether in the aviation or healthcare sectors, is a vital endeavor with far-reaching implications for operational efficiency and resource allocation. The study conducted aimed to identify the most effective model for predicting no-show rates in the civil aviation industry, utilizing various predictive algorithms. Among the seven models examined, Decision Tree exhibited the lowest performance, while Gradient Boosting emerged as the top performer. This

disparity underscores the significance of choosing the appropriate model for accurate predictions in flight management.

The study under review adopts a comprehensive approach to forecasting no-show rates for flights. Utilizing machine learning algorithms, including Random Forest, Bagging, Adaboost, Gradient Boosting, Extreme Gradient Boost, Neural Networks, and Decision Tree, the study evaluates the performance of these models against various metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2), and Max Error values. Notably, Decision Tree emerges with the lowest performance score, while Gradient Boosting exhibits the highest efficacy.

It's noteworthy that the margin of error, considering the distribution of no-show rates between 0 and 20%, is deemed acceptable at 1.56%. This suggests that despite the inherent complexities in predicting human behavior, the models achieved satisfactory results within a reasonable range of error.

The research distinguishes itself by aggregating 74 features from ticket data, encompassing factors like region, scheduled day part, season, day of the week, month, and flight time, among others. Particularly noteworthy is the inclusion of historical no-show rates from previous flights, which significantly contribute to predictive accuracy across all models. Notably, the analysis revealed that variables like the Last One Month and Last One Week averages of previous flight no-show rates significantly contributed to predictive accuracy across all models.

The broader literature on no-show prediction encompasses studies spanning both the aviation and healthcare domains, offering valuable insights into predictive modeling techniques and their applications. Noteworthy studies include Lawrence, Hong, and Cherrier’s seminal work, which underscores the economic benefits of

accurately forecasting no-shows for airlines. Similarly, Garrow and Koppelman's comparative analysis highlights the influence of variables such as ticket prices, travel class, and flight timing on no-show behavior. The study builds upon previous research efforts that have underscored the economic implications of accurately forecasting no-show rates, highlighting the potential for increased profitability through optimized seat utilization and revenue management strategies. Notably, advancements in predictive analytics, including methodologies like Case-Based Reasoning (CBR) and dynamic probability modeling, have shown promise in enhancing forecast accuracy and informing strategic decision-making in the aviation industry.

Additionally, researchers have employed randomized linear programming and Geometric Brownian Motion (GBM) methods to address network revenue management and passenger forecasting challenges.

The findings from this study contribute valuable insights to the growing body of research aimed at enhancing decision-making processes in industries reliant on accurate demand forecasting. By leveraging advanced analytical techniques and integrating diverse data sources, organizations can better navigate the challenges posed by unpredictable human behavior and optimize resource allocation for improved operational efficiency and customer satisfaction.

Notably, the literature and this study highlight the complexity of no-show prediction, influenced by multifaceted factors ranging from economic incentives and travel preferences to demographic characteristics and reminder systems. The findings underscore the need for tailored predictive models that account for domain-specific nuances and evolving consumer behaviors.

CHAPTER 5

CONCLUSION

This research aims to investigate the model that best predicts no-show rates for a flight in the civil aviation industry. Although there are several similar studies in the literature, this study differs both methodologically and in that it has not been conducted before in our country. Seven predictive models; Random Forest, Bagging, Adaboost, Gradient Boosting, Extreme Gradient Boost, Neural Networks, and Decision Tree have been applied in this study, utilizing several machine learning algorithms.

A total of 74 features were generated by aggregating from the ticket data, including region, scheduled day part, season, day of the week (Monday, Tuesday, etc.), first half of the year (first six months) or second half (last six months), month, proximity of official holidays based on origin and destination, flight time, and average no-show rates on previous flights.

Seven models were used to estimate no-show percentages in the civil aviation industry. The models were compared using regression and Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), R-square (R^2), and Max Error values. Decision Tree received the lowest performance score, while Gradient Boosting received the highest. It should be noted here that the no-show variable is the percentage of the total number of reservations. For instance, on a flight with a reservation for 300 people, a MAE value of 1.56 corresponds to approximately 4 passengers. Considering the 2-year flight data, knowing that no-show rates are distributed between zero and 20 percent, it is possible to say that 1.56

% MAE is actually a good result. Although the R^2 value is around 0.5, it could be considered a good fit when estimating the no-show rate in the civil aviation industry.

Unlike other research, this study used ticket data such as season, date, flight time, and the no-show rates seen on previous flights together. The strongest variables that predicted the no-show percentage for all models were the last one month and last one-week averages of the previous flight. This reveals that the characteristics of the flight, as well as the historic no-show rates of the flight, play an important role in predicting.

No-show percentages of a flight offer practical advantages for the aviation industry that go beyond operational efficiency. They provide insights that can be used to optimize revenue, improve customer satisfaction, and inform strategic decisions.

By accurately predicting no-show rates, airlines can optimize seat and revenue management strategies. This allows them to minimize revenue losses from empty seats while increasing profitability. They can also adjust ticket prices and available seats based on expected no-shows.

Moreover, predicting no-show percentages, airlines can improve customer satisfaction by reducing overbooking and denied boarding incidents. They can also ensure that seat availability is in line with actual demand, reducing passenger inconvenience and improving customer experience, leading to loyalty and repeat business for the airline.

Strategically, estimating no-show rates allows airlines to gain valuable insights into customer behavior and market dynamics. This information can be used to inform route planning and scheduling decisions, as well as resource allocation.

understanding no-show behavior patterns across different demographics, different travel segments, and different seasons, airlines can customize their services and marketing campaigns to reach high-value customers.

Beyond immediate operational benefits, accurate predictions of flight no-show rates contribute to broader industry objectives, such as environmental sustainability and resource efficiency. By reducing unnecessary fuel consumption and emissions associated with empty seats, airlines can minimize their environmental footprint and contribute to global efforts to mitigate climate change.

Estimating flight no-show percentages offers tangible and multifaceted benefits for airlines, ranging from revenue optimization and customer satisfaction to strategic decision-making and environmental stewardship. By leveraging advanced predictive analytics techniques and comprehensive data analysis, airlines can unlock new opportunities for innovation and competitive advantage in an increasingly dynamic and competitive market landscape.

One limitation of the study is that the nature of the flight lines is distributed all over the world, which causes extreme values depending on many variables in the world, and therefore it is impossible to predict these extreme predictions, causing the correct prediction rate to remain only at a certain level.

No-show rates can vary significantly based on factors such as holidays, weather conditions, or special events, making it challenging to accurately predict them year-round. Changes in booking patterns, last-minute cancellations, or seat upgrades can affect the accuracy of predictions. Different passenger demographics may exhibit varying levels of reliability in showing up for their flights, making it difficult to generalize predictions across all passengers. Unexpected events like medical emergencies, family emergencies, or travel restrictions can lead to higher-

than-expected no-show rates. Errors in reservation systems or communication failures can lead to corruption of the data and inaccuracies in predicting no-show rates. Competitive pricing strategies or last-minute deals offered by other airlines can influence passenger behavior and affect no-show predictions. Limited or incomplete historical data, especially for new routes or seasonal flights, can hinder the accuracy of predictions. Developing accurate predictive models for no-show rates requires sophisticated algorithms and may still be subject to errors or biases. Changes in regulations or security measures may impact passenger behavior and affect the reliability of no-show predictions. Ultimately, predicting human behavior is inherently uncertain, and no-show rates may be influenced by a wide range of personal factors and preferences beyond the scope of predictive models.

For future work, several promising avenues emerge from the current research findings. There's a need to delve deeper into the integration of advanced machine learning algorithms to enhance both accuracy and robustness in predicting no-show percentages. Exploring alternative data sources, including customer-based features such as age, gender, and other demographic variables, and refining the estimation method to account for dynamic factors such as weather or traffic volume present exciting avenues for future investigation, enabling more comprehensive and accurate predictions of no-show rates on flights. Further exploration is warranted in the incorporation of big data analytics and real-time data streams, enabling more dynamic predictions that can adapt to evolving passenger behaviors and external influences. Additionally, future research could focus on the development of hybrid models that blend statistical and machine learning approaches, leveraging the strengths of each methodology. Moreover, there's a potential for significant advancements by adopting explainable AI techniques, which can improve the

interpretability and trustworthiness of predictive models, crucial for decision-making in the airline industry. Lastly, collaboration with academia, industry partners, and regulatory bodies remains essential to establish standardized methodologies and benchmarks, ensuring consistency and comparability across different estimation methods.

REFERENCES

Adediran, A. O. (2020). Addressing the effects of pandemics on civil aviation: The case of Covid-19. *Issues Aviation Law & Policy, 20*, 219.

AlMuhaideb, S., Alswailem, O., Alsubaie, N., Ferwana, I., & Alnajem, A. (2019). Prediction of hospital no-show appointments through artificial intelligence algorithms. *Annals of Saudi Medicine, 39*(6), 373-381.

Alshammari, A., Almalki, R., & Alshammari, R. (2021). Developing a predictive model of predicting appointment no-show by using machine learning algorithms. *Journal of Advances in Information Technology, 12*(3).

Asrah, N. M., Nor, M. E., Rahim, S. N. A., & Leng, W. K. (2018, April). Time series forecasting of the number of Malaysia Airlines and AirAsia passengers. *Journal of Physics: Conference Series, 995*(1), 012006.

Berg, B. P., Murr, M., Chermak, D., Woodall, J., Pignone, M., Sandler, R. S., & Denton, B. T. (2013). Estimating the cost of no-shows and evaluating the effects of mitigation strategies. *Medical Decision Making, 33*(8), 976-985.

Chang, P. C., Hsieh, J. C., Yeh, C. H., & Liu, C. H. (2006). A case-based seat allocation system for airline revenue management. In J. C. Wang, G. Wittenberg, & P. W. Cheng (Eds.), *Intelligent Computing: International Conference on Intelligent Computing, ICIC 2006, Kunming, China, August 16-19, 2006. Proceedings, Part I* (pp. 1003-1011). Springer Berlin Heidelberg.

Chin, W. S., Ting, C. Y., & Cham, C. L. (2023). No-show passenger prediction for flights. *JOIV: International Journal on Informatics Visualization, 7*(3-2).

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modeling. *Family Medicine and Community Health, 8*(1).

Fan, G., Deng, Z., Ye, Q., & Wang, B. (2021). Machine learning-based prediction models for patient's no-show in online outpatient appointments. *Data Science and Management, 2*, 45-52.

Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: A comparative study using the airline data. *Journal of the Royal Statistical Society Series C: Applied Statistics, 47*(2), 231-250.

Fildes, R., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and operational research: A review. *Journal of the Operational Research Society, 59*, 1150-1172.

Future of Aviation. (n.d.). Retrieved December 10, 2023, from https://www.icao.int/Meetings/FutureOfAviation/Pages/default.aspx

Garrow, L. A., & Koppelman, F. S. (2004a). Multinomial and nested logit models of airline passengers' no-show and standby behavior. *Journal of Revenue and Pricing Management, 3,* 237-253.

Garrow, L. A., & Koppelman, F. S. (2004b). Predicting air travelers' no-show and standby behavior using passenger and directional itinerary information. *Journal of Air Transport Management, 10*(6), 401-411.

Gorin, T., Brunger, W. G., & White, M. M. (2006). No-show forecasting: A blended cost-based, PNR-adjusted approach. *Journal of Revenue and Pricing Management, 5*(3), 188-206.

Idrus, N., & Mohamed, N. (2020). Forecasting the number of airplane passengers using Box-Jenkins and artificial neural network in Malaysia. *University Malaysia Terengganu Journal of Undergraduate Research, 2*(4), 89-100.

Kunnumkal, S., Talluri, K., & Topaloglu, H. (2012). A randomized linear programming method for network revenue management with product-specific no-shows. *Transportation Science, 46*(1), 90-108.

Lawrence, R. D., Hong, S. J., & Cherrier, J. (2003, August). Passenger-based predictive modeling of airline no-show rates. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 397-406).

Lim, S. T., Yuan, J. Y., Khaw, K. W., & Chew, X. (2023). Predicting travel insurance purchases in an insurance firm through machine learning methods after COVID-19. *Journal of Informatics and Web Engineering, 2*(2), 43-58.

Neuling, R., Riedel, S., & Kalka, K. U. (2004). 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.

Perez, A. (2021). Models for fitting correlated non-identical Bernoulli random variables with applications to an airline data problem. *Temple University.*

Petrović, B., & Milošević, P. (2021). Decision support system for predicting the number of no-show passengers in airline industry. *Tehnički vjesnik, 28*(1), 123-134.

Smith, B. C., Leimkuhler, J. F., & Darrow, R. M. (1992). Yield management at American Airlines. *Interfaces, 22*(1), 8-31.

Somboon, M., & Amaruchkul, K. (2017). Applied two-class overbooking model in Thailand's passenger airline data. *The Asian Journal of Shipping and Logistics, 33*(4), 189-198.

Suzuki, Y. (2002). An empirical analysis of the optimal overbooking policies for US major airlines. *Transportation Research Part E: Logistics and Transportation Review, 38*(2), 135-149.

Tissera, P. K., Waduge, K. T., Perera, M., Nawinna, D. P., & Kasthurirathna, D. (2020, December). Predictive analytics platform for airline industry. In *2020 2nd International Conference on Advancements in Computing (ICAC)* (Vol. 1, pp. 108-113). IEEE.

Topuz, K., Uner, H., Oztekin, A., & Yildirim, M. B. (2018). Predicting pediatric clinic no-shows: A decision analytic framework using elastic net and Bayesian belief network. *Annals of Operations Research, 263*, 479-499.

Ulaşan ve Erişen Türkiye 2002-2022. (n.d.). Retrieved December 13, 2023, from https://www.uab.gov.tr/uploads/pages/havacilik/03-sektor-raporlari-havayolu.pdf

Zenkert, D. (2017). No-show forecast using passenger booking data.