REPUBLIC OF TÜRKİYE

AKDENIZ UNIVERSITY

HYBRID PREDICTIVE MODELING APPROACH FOR PREDICTING FLIGHT DELAYS AND CANCELLATIONS

Gülmira AHMETOĞLU

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

OCTOBER 2024

ANTALYA

REPUBLIC OF TÜRKİYE

AKDENIZ UNIVERSITY

HYBRID PREDICTIVE MODELING APPROACH FOR PREDICTING FLIGHT DELAYS AND CANCELLATIONS

Gülmira AHMETOĞLU

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

OCTOBER 2024

ANTALYA

REPUBLIC OF TÜRKİYE

AKDENİZ ÜNİVERSİTESİ

HYBRID PREDICTIVE MODELING APPROACH FOR PREDICTING FLIGHT DELAYS AND CANCELLATIONS

Gülmira AHMETOĞLU

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERIG

MASTER THESIS

This thesis unanimously accepted by the jury on 14/10/2024.

Assoc. Prof. Dr. Alper ÖZCAN (Supervisor)
Assoc. Prof. Dr. Taner DANIŞMAN
Assoc. Prof. Dr. Emrah DÖNMEZ

REPUBLIC OF TÜRKİYE

AKDENİZ ÜNİVERSİTESİ

HYBRID PREDICTIVE MODELING APPROACH FOR PREDICTING FLIGHT DELAYS AND CANCELLATIONS

Gülmira AHMETOĞLU

INSTITUTE OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER THESIS

This thesis unanimously accepted by the jury on 14/10/2024

Assoc. Prof. Dr. Alper ÖZCAN (Supervisor)

Assoc. Prof. Dr. Taner DANIŞMAN

Assoc. Prof. Dr. Emrah DÖNMEZ

ÖZET

HYBRID PREDICTIVE MODELING APPROACH FOR PREDICTING FLIGHT DELAYS AND CANCELLATIONS

Gülmira AHMETOĞLU

Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Alper ÖZCAN

Ekim 2024; 52 sayfa

Bu tez, uçuş operasyonlarını optimize etmek için tahmine dayalı modeller geliştirmeyi hedeflemiştir. Özellikle, uçuş gecikmelerini ve iptallerini tahmin etmeye odaklanarak havacılık endüstrisindeki acil zorlukları ele almıştır. Bu araştırmada, uçuş gecikmeleri hafif gecikmeler (15 dakika ile bir saat arası) ve ağır gecikmeler (bir saatten fazla) olarak sınıflandırılmıştır ve uçuş gecikmelerinin ve iptallerinin olasılığı tahmin edilmiştir. Lojistik regresyonu karar ağaçlarıyla birleştiren hibrit bir modelleme yaklaşımı kullanan çalışma, tahmin doğruluğunu ve güvenilirliğini artırmayı amaçlamıştır.

Tahminler, planlanan uçuş saatlerinden 4, 12, 24, 48, 72 ve 168 saat önce olmak üzere üretildi ve en iyi performans 4 saat öncesinde gözlemlenmiştir. Bu yaklaşım, modelin uçuşlar kalkışa yaklaşırken optimum doğrulukla eyleme geçirilebilir içgörüler sağlamasına olanak tanımıştır.

Bu metodolojide, uçuş verileri Ulaştırma İstatistikleri Bürosu'ndan, havaalanlarının geçmiş hava durumu verileri ise Open-Meteo web sitesinin sunduğu API ile alınmıştır. Uçuş ve hava durumu verileri birleştirilerek analiz için kapsamlı bir veri seti oluşturulmuştur. Bu süreci özenli veri ön işleme ve özellik mühendisliği adımları takip etmiştir.

Hibrit modelleme yaklaşımında hem lojistik regresyonun hem de karar ağaçlarının güçlü yönlerinden yararlanılmıştır. Lojistik regresyon, geçmiş verilere dayanarak gecikmeler ve iptaller gibi ikili sonuçları tahmin etmek için olasılıksal bir çerçeve sağlamaktadır. Karar ağaçları ise özellikler arasındaki doğrusal olmayan ilişkileri ve etkileşimleri yakalayarak bu ilk tahminleri ayrıntılı karar kurallarıyla geliştirmektedir. Hibrit model, bu teknikleri birleştirerek verilerdeki hem doğrusal hem de karmaşık ilişkileri işleyebilmekte ve bu sayede daha yüksek bir tahmin performansı sağlamıştır.

Bu araştırmanın önemli bir katkısı, tüm uçuş verilerini evrensel olarak analiz etmek yerine, belirli bir havayolundan ve kalkış havaalanından gelen verilere odaklanan yenilikçi yaklaşımı olmuştur. Bu yaklaşım, gecikme ve iptallerin ağırlıklı olarak havayolu operatörüyle ilgili faktörlerden kaynaklandığı öngörüsüne dayanmaktadır. Çalışma, her bir operatörün kendi verilerini kullanarak özel olarak uyarladığı eğitim modelleri aracılığıyla, bireysel havayollarının benzersiz operasyonel özelliklerini ve zorluklarını yakalamayı amaçlamıştır. Algoritmayı tek bir taşıyıcıya ve birden fazla kökene odaklanarak test eden araştırma, özel modellerin genelleştirilmiş modellere göre faydalarını göstermiştir. Daha doğru ve eyleme dönüştürülebilir öngörüler sunan bu özel

yaklaşım, tahminleri her havayoluna göre optimize ederek daha kesin tahminler yapılmasına olanak tanımıştır.

ANAHTAR KELİMELER: Hibrit modelleme, Karar ağacı, Lojistik regresyon, Uçuş Gecikmeleri Tahmini, Uçuş İptalleri Tahmini

JÜRİ: Doç. Dr. Alper ÖZCAN

Doç. Dr. Taner DANIŞMAN

Doç. Dr. Emrah DÖNMEZ

ABSTRACT

HYBRID PREDICTIVE MODELING APPROACH FOR PREDICTING FLIGHT DELAYS AND CANCELLATIONS

Gülmira AHMETOĞLU

MSc Thesis in Computer Engineering

Supervisor: Assoc. Prof. Dr. Alper ÖZCAN

October 2024; 52 pages

This thesis addresses the pressing challenges in the aviation industry by developing predictive models to optimize flight operations, specifically focusing on predicting flight delays and cancellations. The research categorizes flight delays into light delays (between 15 minutes and an hour) and heavy delays (more than an hour), as well as predicting the probability of flight cancellations. By employing a hybrid modeling approach that combines logistic regression with decision trees, the study seeks to enhance prediction accuracy and reliability.

Predictions were generated at several time intervals, specifically 4, 12, 24, 48, 72, and 168 hours before scheduled flight times, with the best performance observed at the 4-hour mark. This approach allows the model to provide actionable insights with optimal accuracy as flights approach departure.

The methodology encompasses comprehensive data collection from the Bureau of Transportation Statistics and historical weather data for airports in the flight dataset collected from Open-Meteo, an open-source weather API website. This is followed by careful data preprocessing and feature engineering.

The hybrid modeling approach incorporates strengths from both logistic regression and decision trees. Logistic regression brings a probability formulation in predicting binary outcomes, such as delays and cancellations, from the historical data. Then, decision trees capture non-linear relationships and interactions between features and further refine initial predictions with detailed decision rules. The hybrid model can be executed with both linear and complex patterns in the data, hence improving the predictive performance by incorporating these techniques.

One of the major contributions of this research is its novelty in anchoring research on data from one airline and origin airport, instead of trying to analyze all flight data universally. This is influenced by the intuitive understanding that delays or flight cancellations take place primarily because of factors related to the airline operator. By training individual operator-specific models on their own data, this study hopes to capture what is unique in the operational characteristics and challenges of individual airlines. The benefits of the tailored models over the generalized quantified by testing the algorithm by focusing on a single carrier and its origins. Only this tailored approach allows for more

precise predictions, hence optimization of predictions for each airline, which has the potential for more accurate and actionable insights.

KEYWORDS: Decision trees, Flight Cancellations Prediction, Flight Delays Prediction Hybrid modeling, Logistic regression

COMMITTEE: Assoc. Prof. Dr. Alper ÖZCAN

Assoc. Prof. Dr. Taner DANIŞMAN

Assoc. Prof. Dr. Emrah DÖNMEZ

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Assoc. Prof. Dr. Alper Özcan, for their unwavering support, guidance, and encouragement throughout the course of this research. Their invaluable insights and expertise have been instrumental in shaping this thesis.

I would also like to extend my heartfelt thanks to the other professors in the Computer Science Department of Akdeniz University. Their constructive feedback and generous sharing of knowledge have significantly contributed to my academic and personal growth.

Furthermore, I am profoundly grateful to my husband, Hayal Ahmetoglu, for his endless patience, understanding, and encouragement. His unwavering support has been a constant source of strength and motivation throughout this journey.

To my family and friends, thank you for your continuous love and support. Your belief in me has been a driving force behind my perseverance.

Thank you all for being an essential part of this journey.

LIST OF CONTENTS

TEXT OF OATH

I hereby declare that this thesis, titled "HYBRID PREDICTIVE MODELING APPROACH FOR PREDICTING FLIGHT DELAYS AND CANCELLATIONS " is the result of my own original work and has not been submitted previously or concurrently in whole or in part for any degree or diploma at any other institution. I have fully acknowledged and referenced all sources of information and data used in the thesis, adhering to the ethical standards and guidelines of academic integrity.

14 / 10 / 2024

Gülmira Ahmetoğlu

SYMBOL AND ABBREVIATIONS

Abbreviations

AdaBoost	: Adaptive Boosting Algorithm
ANOVA	: Analysis of Variance
BiLSTM	: Bidirectional Long Short-Term Memory
BTS	: Bureau of Transportation Statistics (BTS)
CatBoost	: machine learning gradient-boosting algorithm
EDA	: Exploratory Data Analysis
GCN	: Graph Convolutional Networks
GOGCN	: Geographical and Operational Graph Convolutional Networks
IATA	: International Air Transport Association
KNN	: The k-nearest neighbors (KNN) algorithm
LightGBM	: Light Gradient-Boosting Machine
LSTM	: Long Short-Term Memory
MCC	: Mathews Correlation Coefficient
PCA	: Principal Component Analysis
RMSE	: The root mean square error (RMSE)
RNNs	: Recurrent neural networks
XGBoost	: eXtreme Gradient Boosting Algorithm
XGBR	: Regression-specific implementation of XGBoost.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The aviation industry plays a pivotal role in global transportation, facilitating the movement of people and goods across vast distances. Aviation benefits report has shown that as the popularity of flights grows, efficiency and reliability in flight operations become binding (2019). Unfortunately, aviation is a complex dynamic system plagued with lots of challenges, which include delays and cancellations of flights. These, however, demand an innovative approach that leverages a lot of data available in the modern era.

Big data analytics and machine learning techniques have actually caused huge disruption in many industries, especially over the last couple of years, by offering unprecedented opportunities in optimization and predictive analytics (Zhang et al. 2021). Aviation is one such segment that will really benefit from these modern technologies, considering the enormous amount of historical flight data, weather data, and passenger information. Airlines can tap into the power of data-driven predictive modeling to achieve better operational efficiency, reduction in costs, and improvement in passenger satisfaction.

This research will seek to find the application of data-driven predictive modeling in optimizing flight operations by prediction of flight delays and cancellations. By utilizing advanced analytical techniques such as machine learning algorithms and time series analysis, we will create robust predictive models that will help airlines with better resource management, planning, and informed decision making.

According to estimates from the FAA's Office of Aviation Policies and Plans, the total cost of delayed flights increased by 9.3%, from $30.2 billion to $33.0 billion (2022). The $2.8 billion increase is attributed to passenger delay costs. During the same comparative period, the total cost of delays increased from $19.2 billion to $33.0 billion. $8.4 billion of this change was reflected in the cost of delays to passengers. These results provide a sense of the incredible economic toll that flight delays take, not only directly affecting airlines and their passengers, but also affecting overall economic activity and consumer behavior.

Flight delays have deep implications for passengers in many directions, influencing their experience of traveling and general feeling of satisfaction with the carrier (Wang et. Al 2022; Kešel'ová and Hanák 2019). Delays, defined as the time difference by which the flight is late or postponed, are among the most remembered performance indicators in any system of transportation. For commercial aviation, delay is usually represented by the difference between scheduled and actual times of departure or arrival of an aircraft (Carvalho et al. 2020). This has been emphasized through the numerous flight delay tolerance threshold levels set by the national regulatory authorities.

The frustration caused to passengers by delays in flight is due to the inconveniences and changes of plans resulting from the uncertainty of the situation. This often causes missing connecting flights, delays to reach destinations, and subsequently interfering with other matters that require one's attention. Passengers might have to wait at the airports for long hours, thus causing them discomfort, dissatisfaction, and fatigue, especially when they are accompanied by children or older members of the family (Fleurquin et al. 2014). Besides, delays may involve financial costs, as passengers have to pay for extra meals, accommodation, transport, and other needs attached to waiting for flights. Sometimes, they even forfeit vital events they had hoped to attend, such as business meetings, family occasions, or vacations, which means further losses in addition to disappointment (Britto, Dresner and Voltes 2012).

Besides these immediate impacts, flight delays have the potential to create safety hazards as well as have environmental repercussions. Long waits can yield restless and tired crew and passengers, which may lead to less safe precautions and higher levels of accidents or incidents during flight. Besides, delays often lead to increased fuel consumption and greater emission of greenhouse gases due to the protracted idling of aircraft, exacerbating environmental degradation and climate change concerns. Therefore, mitigating flight delays benefits not only passengers and airlines in enhancing safety and satisfying customers but also plays a key role in reducing the aviation industry's ecological footprint. Airlines must minimize delays and disruption to passengers for an enjoyable, safe journey with minimal hassle.

The main motivation of this thesis is developing predictive models that respond to the immediate challenges of flight delays and cancellations in the aviation sector through efficient and optimized flight operations. These are very expensive, as reported by the FAA, due to huge cost increases because of delays, thus requiring effective predictive analytics. It enables airlines to further improve their resource utilization, develop more realistic schedules, and enhance general passenger satisfaction.

This thesis proposes a hybrid modeling approach that combines logistic regression with decision trees to predict flight delays and cancellations. The methodology involves extensive data collection from the Bureau of Transportation Statistics and historical weather data gleaned from Open-Meteo. Advanced machine learning methods are then applied, focusing on particular airlines and origin airports, in order to come up with strong models representing the specific operational characteristics of each airline. With this tailored approach, it is expected that the predictive performance and reliability of the models will increase.

The novelties and contributions of this thesis are manifold. First, this is a new approach in focusing on data from specific airlines and origin airports, rather than analyzing all flight data universally. Besides, it respects the fact that delays and cancellations in many cases are related to specific airline factors. Then again, this study is based on a hybrid modeling approach that leverages strengths from both logistic regression and decision trees. Logistic regression provides a probabilistic approach for binary outcomes, and decision trees capture nonlinearities in relationships and interactions among features. Combined, the hybrid will be able to learn both linear and complex patterns of data and hence provide improved performance. Finally, the high accuracy achieved by the Airline-Origin airports model, trained in this paper, further proves that domain-tailored machine learning models are truly effective for aviation industry applications. This discussion framework gives an airline the strength to provide for more precise and actionable predictive analytics. This is foreseen to increase efficiency at the level of operation and passenger satisfaction.

2. LITERATURE REVIEW

Flight delays and cancellations are one of the most studied topics in the aviation industry in terms of predictive modeling. Over the years, various works have been performed to improve the accuracy and reliability of predictions by adopting different methodologies, data sources, and feature engineering. Thus, this literature review will provide a broad overview of the key developments in the topic by categorizing the significant contributions into machine learning, deep learning, and ensemble/hybrid approaches. A greater emphasis on hybrid approaches is used here since they are most relevant to the hybrid modeling approach adopted in this work. What is emphasized in these areas of focus are seminal works, major contributions, and ongoing debates.

2.1. Machine Learning Approaches

General machine learning techniques applied to the prediction of flight delays include the following: Rebollo and Balakrishnan (2014) compared decision trees and random forests, where weather data were combined with flight schedules and historical delay records, underlining the importance of weather variables in enhancing predictive accuracy. Giu et al. (2015) showed that machine learning models are effective against flight delay prediction problems, where feature selection and engineering significantly improve the performance of a model.

Recently, a study published by Vidyapeetham (2023) employed several machine learning models in a flight delay prediction approach, mainly focusing on weather variables in conjunction with other relevant factors. They found that the best model was the XGBoost Regression with a root mean square error (RMSE) of 0.81, followed by the Random Forest model with an RMSE of 1.34. Applying Principal Component Analysis (PCA) resulted in poorer model performance due to PCA's failure to maintain high data variance. Normalizing the data did not affect non-tree models but reduced accuracy in tree models. Another study addresses flight delay prediction by framing it as a machine learning problem, specifically using binary classification. Seven algorithms were tested, and four evaluation metrics (accuracy, precision, recall, and f1-score) were used to assess performance. Due to the imbalanced dataset, measures were weighted to mitigate the dominance of non-delayed flights. The Decision Tree model outperformed others, achieving the highest scores across all metrics (accuracy: 0.9778), while KNN had the lowest f1-score of 0.8039 (Tang, 2021).

Choi et al. (2016) developed a supervised machine learning model that employs data mining and various supervised machine learning algorithms, including decision trees, random forest, AdaBoost, and k-Nearest Neighbors, to predict airline delays caused by inclement weather conditions.

2.2. Deep Learning Approaches

Deep learning approaches have shown promise in handling the complex and nonlinear nature of flight delay prediction.

Chai et al. (2021) examined flight delays issue from a network perspective, considering multiple airports. They usesd graph convolutional neural networks (GCN) to model dynamic interactions within airport networks. Since GCNs can't directly process both delay time-series and evolving graph structures, a temporal convolutional block is used. An adaptive graph convolutional block addresses incomplete graph inputs from unexpected routes. These experiments showed that this approach outperformed the benchmarks, and it was hereby confirmed that deep learning with graph-structured inputs serves efficiently in the flights delay prediction task. In a similar context, Kim et al. (2016) proposed a deep learning approach using RNNs which modeled the sequential information of each flight and achieved an improved performance in predictions. Giusti et al. (2019) also worked on flight delay prediction concerning concept drift. They showed how models can adapt to evolving patterns over time to sustain their accuracy.

Bisandu and Moulitsas (2023), proposed an advanced deep learning for flight delay prediction, assessing both techniques of BiLSTM and LSTM. Their results showed that the BiLSTM model was better than the LSTM, with an MCC value that reached a remarkable 0.99, which suggests that actual classes and that of predicted classes were well aligned. Using both forward and backward sequences with a deep neural network churns out high accuracy, recall, and F1-Score in the flight delay prediction using the model, BiLSTM.

2.3. Ensemble and Hybrid Approaches in Flight Delay Prediction

Ensemble and hybrid approaches have considerably proved to be effective ways in flight delay prediction, attempting to bring together the strengths of various machine learning algorithms and, hence, enhancing prediction accuracy and robustness (Guo et al. 2021; Khan et al 2022). Various advantages and applications are discussed in this section with references to real-world studies that have used these approaches.

2.3.1. Ensemble approaches

Ensemble methods have been developed to combine multiple base models to arrive at a single predictive model that outperforms individuals. The major types of ensemble methods are bagging, boosting, and stacking (Mtimkulu and Maphosa 2023).

In the process of bagging, several instances of the same algorithm are trained for different portions of the training data; their predictions are then combined by averaging. Random Forest bagging algorithm, as this is commonly used in flight delay prediction: Wang et al. (2021) illustrated such a case with Random Forest among other algorithms that have resulted in great improvements of accuracy in flight delay prediction (Guo et al. 2021).

Boosting models like AdaBoost, Gradient Boosting, and XGBoost train models incrementally to correct errors in previously trained models. Rahemeen et al. (2022) applied gradient boosting ensemble models like CatBoost, LightGBM, and XGBoost to the task of flight delay prediction. It was shown from the results that CatBoost improves the accuracy in predictions without affecting its stability, hence assuring the efficiency of ensemble learning techniques on flight delay prediction tasks (Khan et al 2022).

The basic idea of stacking is to involve a meta-model using base models for making a prediction. This technique leverages strength from several algorithms in enhancing overall performance. Mtimkulu and Maphosa utilized stacking with random-forest methods to construct a very accurate flight delay prediction model. They have come up with an efficient methodology of combining multiple best-of-breed algorithms, taking complete advantage of each to make their approach comprehensive and robust (Mtimkulu and Maphosa 2023).

2.3.2. Hybrid approaches

Hybrid approaches combine various machine learning methods to benefit from strengths of individual algorithms and to reduce weaknesses of those. In practice, the most common use is made of traditional machine learning algorithms together with deep learning and ensemble methods (Zhou et al. 2023)

Mamdouh et al. (2023) in their study proposed the Flight Delay Path Previous-based Machine Learning (FDPP-ML) algorithm, where RNNs combined with LSTM in order to model sequential dependencies in flight delays. This hybrid model meaningfully improved prediction accuracy by merging the strengths of machine learning and deep learning techniques (Mamdouh et al. 2023).

Dou (2020) combined machine learning models such as XGBoost, LightGBM, and CatBoost while predicting flight delays. The combination outperformed individual models and demonstrated the power of hybrid approaches in prediction tasks.

Cai et al. (2023) developed a hybrid approach using Geographical and Operational Graph Convolutional Networks (GOGCN) to predict flight delays. This model combined graph convolutional networks with spatial-temporal modeling to capture complex relationships in flight data. The GOGCN model, integrated operational and geographic aggregators to analyze global and spatial airport patterns, achieved improvement the prediction accuracy.

A hybrid machine learning model, using a voting aggregation technique to combine different classifiers, has been proposed by Bisandu and Moulitsas (2023). Compared to the other approaches, their model yielded impressive results in predicting arrival delays with a 0.99 degree of accuracy, precision, recall, and F1-score.

Ensemble and hybrid methods improve flight delay prediction by making it more accurate, reliable, and stable. The cited studies demonstrate their effectiveness, supporting their continued use and development in aviation applications.

3. MATERIAL AND METHOD

3.1. Data Source

The flight data used in this project were downloaded from the website of the Bureau of Transportation Statistics, part of the U.S. Department of Transportation. The BTS offers several sets of data on a wide range of topics in the U.S. transportation system. Among others, detailed records on airline operations include multi-year data on flight dates, carriers, origins, destinations, scheduled and actual departure and arrival times, and length of delay. This broad dataset forms a very strong foundation for developing predictive models that help predict flight delays, cancellations, and on-time performance-all contributing to enhanced airline resource management and improved passenger satisfaction.

For our Flight Delay Prediction model, we need accurate historical weather data. We have procured such data from the Open Meteo website. Open-Meteo integrates with national weather services to provide open data for your location, ensuring timely and precise forecasts. Open-Meteo has a Python API, which helps us fetch historical weather data based on latitude and longitude coordinates, date ranges, and desired weather parameters. Open-Meteo provides free access to weather data with a robust API, allowing users to retrieve detailed weather information for any location. This makes Open-Meteo a valuable resource for integrating weather factors into models like our flight delay prediction model.

Information about all the airports used here includes its unique name or 3-letter IATA code and coordinates taken from the OpenFlights site.

3.2. Data Structure

Data used in this research was collected from the BTS website. The dataset consisted of 21 operating airlines, 375 origin airports, and 376 destination airports for the years from 2022 to 2023. This large dataset, with a shape of (13.766.900, 44), contains features like flight date, airline, origin, destination, scheduled departure, actual departure, scheduled arrival, actual arrival, and delay time. Columns present in our flight dataset along with their descriptions can be observed in Table 3.1.

Table 3.1. Flight Data Structure

COLUMN NAME	**DESCRIPTION**
YEAR	Year in which the flight took place
QUARTER	Quarter of the year in which the flight took place (1 to 4)
MONTH	Month in which the flight took place
DAY_OF_MONTH	Day of the month on which the flight took place
DAY_OF_WEEK	Day of the week on which the flight took place(1=Monday..7)
FL_DATE	Flight date
MKT_UNIQ_CARRIER	Marketing carrier code
OP_UNIQUE_CARRIER	Operating carrier code
TAIL_NUM	Aircraft tail number
OP_CARRIER_FL_NUM	Flight number assigned by the operating carrier
ORIGIN	Origin airport code, 3-letter IATA code
ORIGIN_CITY_NAME	Origin city name
DEST	Destination airport code
DEST_CITY_NAME	Destination city name
CRS_DEP_TIME	Scheduled departure time
DEP_TIME	Actual departure time
DEP_DELAY_NEW	Departure delay in minutes (0 if not delayed)
DEP_DEL15	Departure delay indicator (1 if 15 min or more, 0 otherwise)
DEP_TIME_BLK	Departure time block
CRS_ARR_TIME	Scheduled arrival time
ARR_TIME	Actual arrival time
ARR_DELAY_NEW	Arrival delay in minutes (0 if not delayed)
ARR_DEL15	Arrival delay indicator (1 if 15 minutes or more, 0 otherwise)
ARR_TIME_BLK	Arrival time block
CANCELLED	Flight cancellation indicator (1 if canceled, 0 otherwise)
CANCELLATION_CODE	Reason for canc(A=carrier, B=weather, C=NAS, D=security)
DIVERTED	Flight diversion indicator (1 if diverted, 0 otherwise)
CRS_ELAPSED_TIME	Scheduled elapsed time of the flight
ACTUAL_ELAPSED_TIME	Actual elapsed time of the flight
FLIGHTS	Number of flights (always 1 in this dataset)
DISTANCE	Distance of the flight route in miles
DISTANCE_GROUP	Distance group (1:<250 miles, 2:250-499 miles,.., 12>=2750)
CARRIER_DELAY	Delay caused by the airline in minutes
WEATHER_DELAY	Delay caused by weather in minutes
NAS_DELAY	Delay caused by National Airspace System (NAS) in minutes
SECURITY_DELAY	Delay caused by security issues in minutes
LATE_AIRCRAFT_DELAY	Delay caused by late arrival of aircraft in minutes

This project used airport data taken from the Open Flights website. It incorporates detailed information on various airports. Some of the columns included in this dataset are Airport Name, City, Country, and IATA, known as a standardized three-letter code used to identify an airport. The dataset also includes geographical attributes such as latitude, longitude, and altitude, as well as operational details like the timezone and daylight saving time (DST) observance.

Table 3.2. Airports Data Structure

COLUMN NAME	DESCRIPTION
AIRPORT	Unique OpenFlights identifier for this airport.
NAME	Name of airport.
CITY	Main city served by airport.
COUNTRY	Country or territory where airport is located.
IATA	3-letter IATA code. Null if not assigned/unknown.
LATITUDE	Decimal degrees, usually to six significant digits.
LONGITUDE	Decimal degrees, usually to six significant digits.
ALTITUDE	In feet.
TIMEZONE	Hours offset from UTC. Fractional hours are expressed as decimals.
TZ DATABASE	Database timezone in "tz" (Olson) format, eg. "America/Los_Angeles

To predict flight delays and cancellations 4 hours before the scheduled departure time, we utilized the Open-Meteo API, which takes latitude, longitude, and date parameters to retrieve weather data for specific dates and locations. Hourly weather data for the years 2022-2023 was fetched to align with the origin airport and departure time of each flight, focusing specifically on weather conditions known from 4 hours before each flight's departure.

Table 3.3. Weather Data Structure

COLUMN NAME	DESCRIPTION
DATETIME	Date and time YYYY-MM-DD HH24:MI:SS
ORIGIN	Airport 3-letter IATA code
TEMPERATURE	Temperature at 2 meters, expressed as decimals
PRECIPITATION	Precipitation, expressed as decimals
WIND_SPEED_10M	Wind speed at 10 meters, expressed as decimals

To achieve this, we created an `Adjusted_Datetime` column by subtracting 4 hours from the planned departure time, aligning with the time the prediction would be made rather than the actual departure. The flight dataset was then enhanced by merging it with weather data, incorporating variables such as temperature, precipitation, and wind speed. This was done by performing an inner join on the adjusted date and Origin (departure airport) columns, ensuring that each row in the flight dataset was linked with weather conditions known at the prediction time-specifically 4 hours before the scheduled departure-rather than the actual departure weather conditions

3.3. Data Preprocessing

The original dataset exhibited a relatively low percentage of missing values, highlighting its overall completeness. However, despite this, addressing these missing values through data cleaning is crucial for enhancing the dataset's integrity and bolstering the efficacy of subsequent analyses or modeling endeavors.

In order to make informed decisions and make clear interpretations, we chose the basic columns 'OP_UNIQUE_CARRIER', 'ORIGIN', 'DEST', 'CRS_DEP_TIME' and 'CRS_ARR_TIME' as the core of our analytical model. Although only about 1.95% of the original dataset had missing values, it was important to perform data cleaning as even this small amount could affect the accuracy of the analysis. In order to maintain the quality and reliability of our analysis, data with empty columns, especially the columns selected for the basis, were removed from the dataset.

3.4. Exploratory Data Analysis (EDA)

In the “Data Analysis” section of this thesis, we aim to address a number of questions regarding flight delays and thereby enrich our understanding of the complexity of aviation operations.

The airlines report the causes of delay in broad categories established by the Air Carrier On-Time Reporting Advisory Committee. These categories include Air Carrier, National Aviation System, Weather, Late-Arriving Aircraft, and Security. The causes of cancellation are similar, except there is no category for late-arriving aircraft. Given that only 1.99% of the two-year data consists of cancellations, it is difficult to derive valuable statistics from this subset. Therefore, as shown in Figure 3.1. by statistical office, we will use the same features for modeling both delays and cancellations.

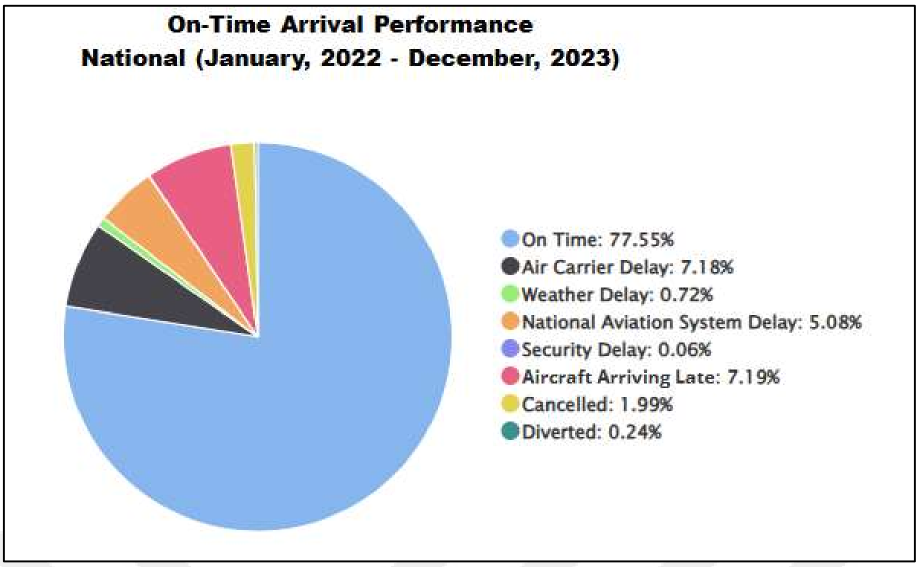

Figure 3.1. Airline On-Time Statistics and Delay Causes (Bureau of Transportation Statistics 2023)

Our research begins with a comprehensive review of key statistical descriptors for airlines and aims to uncover key performance metrics and variability in the industry.

According to the Bureau of Transportation Statistics (2023), air carrier delays and aircraft arriving late are the leading causes of flight delays, accounting for significant percentages of total delays. In 2022, air carrier delays constituted 39.8% of all delays, while aircraft arriving late accounted for 37.7%. In 2023, these percentages were 36.4% and 40.0%, respectively. This further underlines the importance of carrier-related features as delay predictors, given carriers are the largest contributors to overall delay statistics.

Adding carrier-specific features in our models offers the opportunity to better predict and manage dominant sources of delay for more accurate, actionable predictions. This highlights the urgent need for tailored models that can meet the unique challenges and operational realities of each airline.

We first conduct an in-depth analysis of flight delays by examining differences between departure and arrival delays to identify noticeable patterns or trends. Then, we investigate how delay events vary by departure and arrival airports, studying potential factors contributing to airport-specific delays. Our research also explores temporal changes in delay patterns, examining the impact of the day of the week and the day of the month on delay occurrences. Additionally, by analyzing delays according to seasonal variations, we attempt to explain the temporal dynamics that shape delay events throughout the year. Finally, we examine the distribution of delays between different departure time blocks to understand how changes in departure times affect delay frequencies and durations.

With this detailed examination, we will try to determine the answers to the following questions.

- Basic statistical description of airlines
- Analyzing Flight Delays: Departure versus Arrival
- Analyzing Flight Delays: Origin Airport versus Destination
- Do delays vary based on the day of the week and the day of the month?
- Analysis of Delays by Seasonal variations
- Does distance affect delay?
- How are delays distributed across different departure time blocks?
- Cancellations by reason

3.4.1. Basic statistical description of airlines

First, all flights from all airlines were examined. The goal here is to categorize the airlines based on their punctuality, and to do this, we calculate a few basic statistical parameters.

Table 3.4. Flight Stats per Airline

CARRIER	DESCRIPTION	TOTAL FLIGHTS	DELAYED FLIGHTS	% DELAYED	MAX DELAY	MEAN DELAY	MIN DELAY
WN	Southwest Airlines	2780551	708890	25.5	775	47.6	15
DL	Delta Air Lines Inc	1897275	303992	16.0	1311	70.3	15
AA	American Airlines	1820181	379386	20.8	4413	83.4	15
OO	SkyWest Airlines	1415203	205098	14.5	5995	86.7	15
UA	United Air Lines	1373669	258133	18.8	1622	71.8	15
YX	Republic Airline	612910	77158	12.6	7223	73.7	15
B6	JetBlue Airways	541959	166196	30.7	2508	82.4	15
NK	Spirit Air Lines	501824	132545	26.4	1452	68.8	15
MQ	Envoy Air	480413	71657	14.9	5327	61.2	15
AS	Alaska Airlines Inc	474104	80168	6.9	280	52.0	15
9E	Endeavor Air Inc	433072	56345	3.0	973	81.1	15
OH	PSA Airlines Inc	413079	65273	5.8	225	80.2	15
F9	Frontier Airlines Inc	339820	99547	9.3	393	76.2	15
G4	Allegiant Air	233405	59245	25.4	2065	74.7	15
YV	Mesa Airlines	200496	41876	00.9	2186	97.9	15
PT	Piedmont Airlines	190518	21493	11.3	1512	86.2	15

Table 3.4. Continued

QX	HORIZON AIR	158664	21063	13.3	873	54.6	15
HA	Hawaiian Airlines	155993	28458	18.2	1847	47.1	15
C5	CommuteAir	146449	21867	14.9	1464	89.2	15
ZW	Air Wisconsin	122558	22130	18.1	1498	92.8	15
G7	GoJet Airlines	100524	16925	16.8	1445	82.5	15

We analyzed the statistics provided in Table 3.4. and found insights below.

- Variation in Delay Rates: There is significant variation in the percentage of delayed flights among different airlines, ranging from as low as 11.3% for Piedmont Airlines to as high as 30.7% for JetBlue Airways. This suggests that some airlines may have more effective strategies for minimizing delays than others.
- Impact of Maximum Delay: The maximum delay experienced by each airline varies widely, with values ranging from 775 minutes for Southwest Airlines Co. to 7223 minutes for Republic Airline. This suggests that while most airlines experience delays within a certain range, outliers can significantly impact overall performance metrics.
- Average Delay Time: The average delay time provides insight into the typical length of delays experienced by passengers. Airlines like JetBlue Airways and American Airlines Inc. have relatively high average delay times (82.4 minutes and 83.4 minutes, respectively), indicating that delays, when they occur, tend to be lengthy.
- Delay statistics across different airlines strongly point to differences in the pattern of delays according to various airlines. While a few airlines have lesser flights delayed and longer waits whenever such delays do take place, other airlines can be noticed to consistently have a higher number of their total flights delayed, often related to having more total flights.

When developing a hybrid approach to delay prediction, the analysis should focus on individual airline data rather than generalized data. Fundamental to this strategy will be the recognition that delay patterns vary across airlines due to influential factors such as fleet size, routes flown, and location of operating centers. Focusing on data from each airline allows the model to address specific factors that affect delays. This tailored methodology will allow for more accurate and targeted modeling by taking into account the operating context of each airline rather than trying to generalize delay predictions for the entire airline industry.

3.4.2. Analyzing Flight Delays: Departure versus Arrival

When analyzing flight delays, it is important to determine whether the delay occurs at departure or arrival. So far, we have focused on delays at departure. In this section, we will examine whether these delays at departure lead to similar delays at landing.

Python code using Matplotlib and Seaborn libraries used to generate a bar plot comparing the mean delay at departure and arrival for each airline in the dataset:

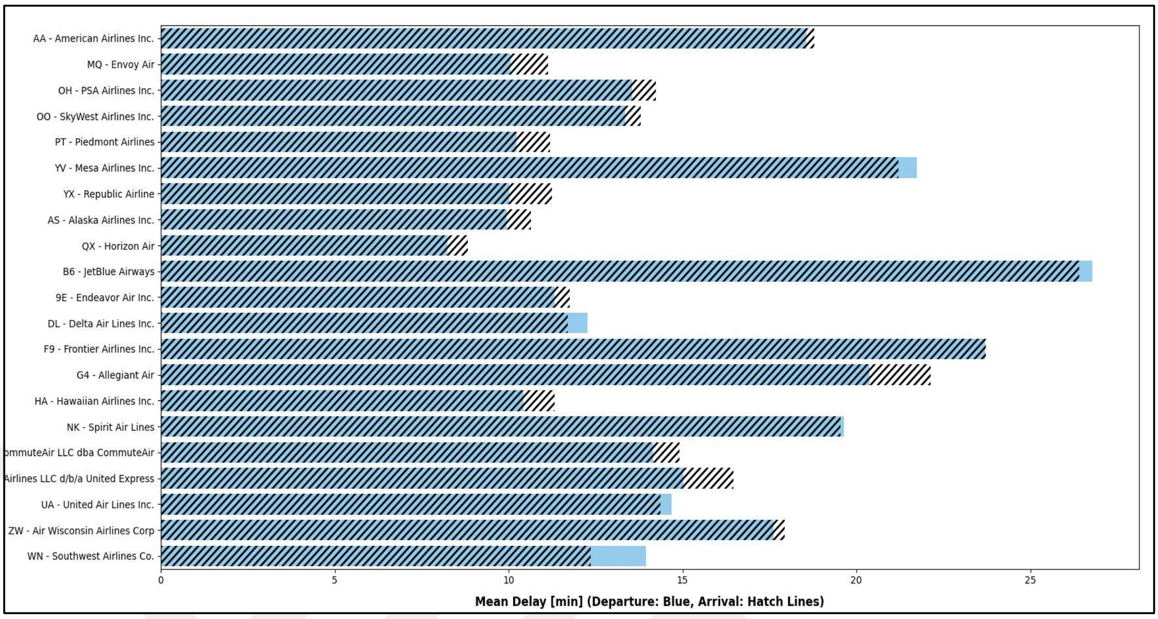

Figure 3.2. Departure Delays versus Arrival Delays

The data indicates that the mean delays at arrival are generally very similar to the mean delays at departure across different airlines. This observation suggests a few important points:

Propagation of Delays: The similarity in departure and arrival delays implies that delays incurred at the point of departure tend to propagate through to the point of arrival. This means that if a flight is delayed at takeoff, it is likely to arrive late as well, maintaining a consistent delay duration throughout its journey.

Focus on Departure Delays: Given that departure delays appear to be a primary driver of arrival delays, it makes sense to prioritize investigations and mitigation strategies at the departure stage. Addressing the root causes of delays at the departure point could have a significant impact on reducing overall flight delays.

Given the strong correlation between departure and arrival delays, our flight delay prediction model will focus on departure delays. Since these delays are likely to propagate, accurately predicting departure delays will inherently improve the prediction of arrival delays.

3.4.3. Analyzing Flight Delays: Origin Airport versus Destination

Which origin airport experiences the highest number of delays? Which destination airport experiences the highest number of delays?

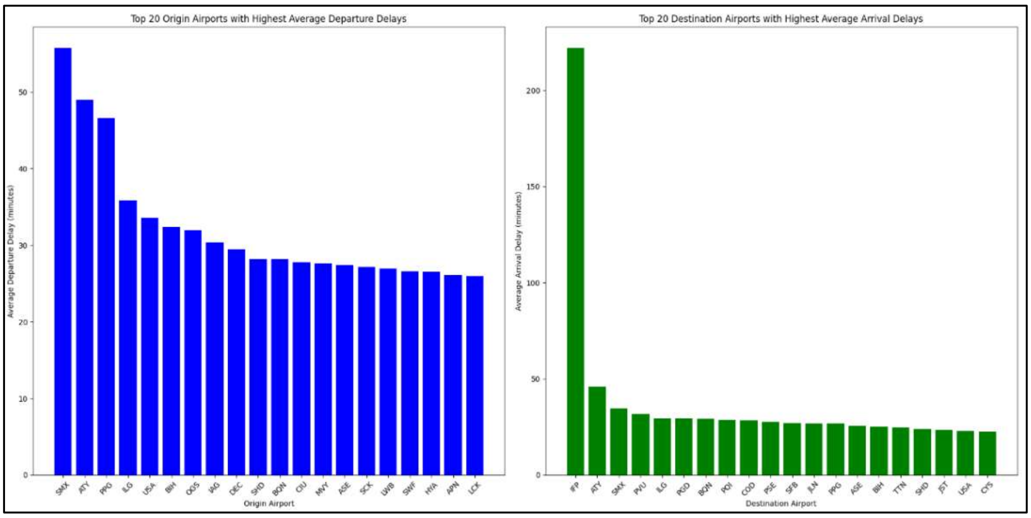

Figure 3.3. Top 20 Origin and Destination airports with the highest average delays

The analysis shows that delays are not evenly distributed across airports. The extremely high average arrival delay at Laughlin/Bullhead International Airport (IFP) stands out as an outlier. Given that its average delay (222.00 minutes) is significantly higher than other airports, we examined this data to see whether this data point represents an anomaly or a consistent problem. When we checked the number of flights to and from IFP, we saw that there were 2 extremely delayed flights. Because the sample size was so small, we excluded this data from the review because a few extreme lags could significantly skew the average.

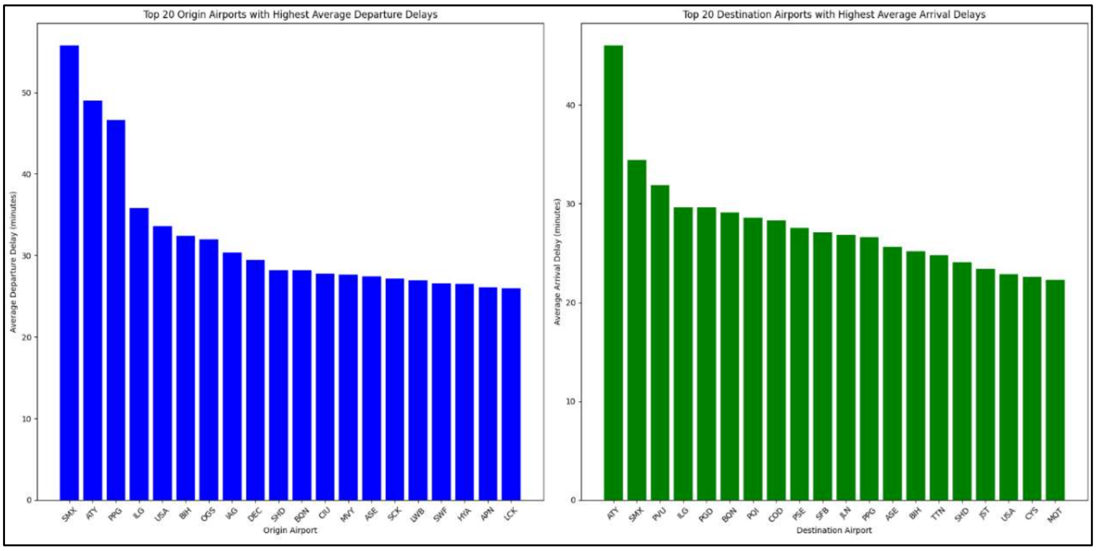

Figure 3.4. Top 20 Origin and Destination airports with the highest average delays (without outliers)

The graph in Figure 3.4. provides valuable information on the performance of various airports in terms of departure and arrival delays. It is noteworthy that the top 6 airports are included in the list of both highest average departure and arrival delays. It suggests that delays from these airports tend to persist throughout the entire flight, affecting both departure and arrival times. This consistency shows that the problems causing delays at these airports may be systemic and deeply rooted in factors such as Airport Infrastructure and Air Traffic Control. Limited facilities, overcrowded airspace, or inefficient air traffic management can create bottlenecks, making it difficult to handle high traffic volumes smoothly.

3.4.4. Temporal Analysis for Flight Delay Prediction

Investigate delays change depending on the day of the week, time of month, or season can really improve the accuracy of predicting flight delays. By detecting these time-related patterns, the model can better determine when delays are more likely to occur. For example, certain days may have more delays due to higher travel demand, while other seasons may have more issues due to bad weather or holiday traffic. The inclusion of such information helps the model be more accurate and reliable in terms of predicting delays and cancellations.

We calculated the daily average latency by organizing the data by day of the week. We then created bar charts as shown in Figure 3.5. to visually display these average latencies, making it easier to spot any patterns or trends.

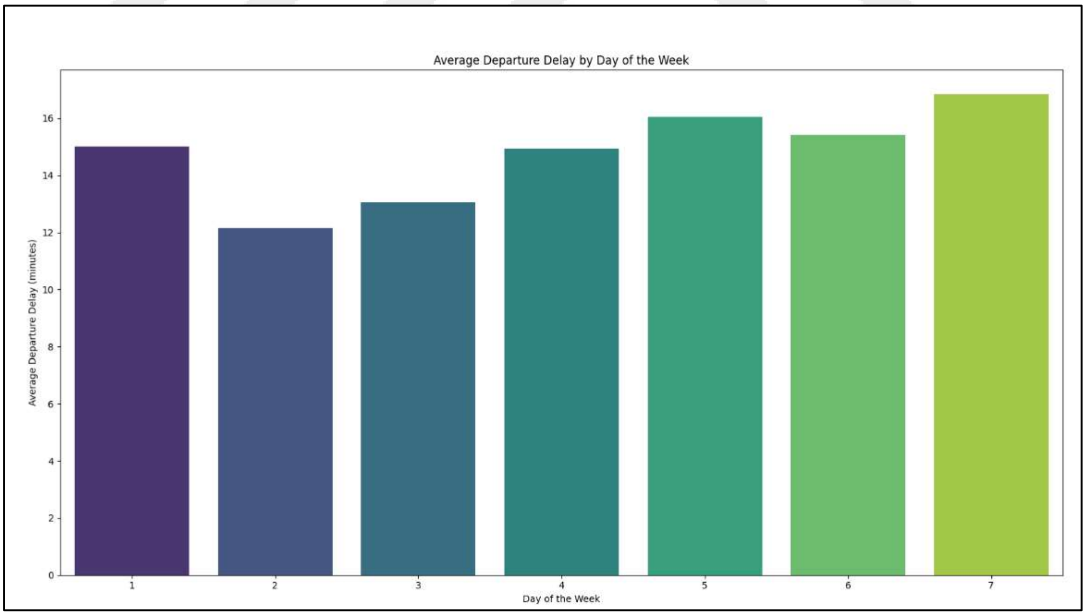

Figure 3.5. Average delays by day of the week

In Figure 3.6, we organized the data by day of the month and calculated the average delay for each day. We then drew bar charts to visually represent these averages so that we could easily identify any patterns or trends. This method helped us understand how delays fluctuated throughout the month and revealed specific days when more or less delay was likely.

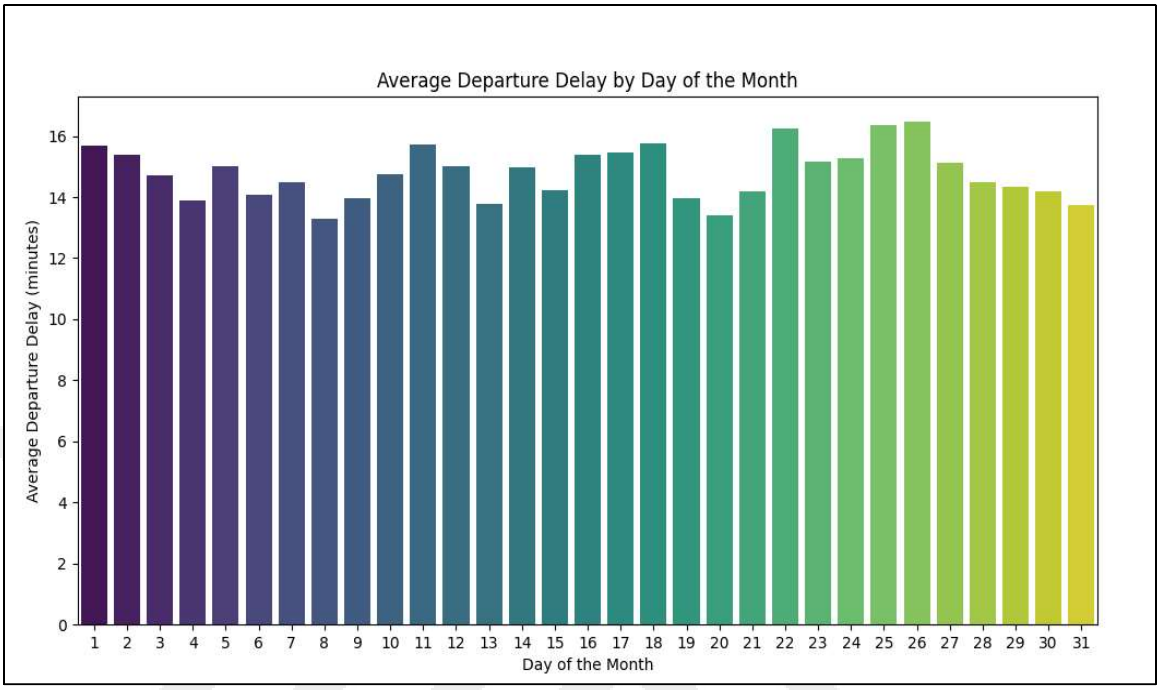

Figure 3.6. Average delays by day of the month

When analyzing departure delays, the average variation trends were relatively consistent when we considered single days of the month. However, knowing that more general temporal patterns play a role in aviation operations, our research continued to examine the distribution of delays by quarter. The distribution by quarter is consistent with standard practice in aviation and provides a more detailed look at seasonality. Fortunately, our data source already included quarterly data, and therefore we were able to directly input the data into our analyses. For a more detailed view of the temporal nature of flight delays, we aimed to gain more insight into the temporal patterns of flight delays by using quarterly flight delay data, as shown in Figure 3.7, which provided greater precision in our predictive model.

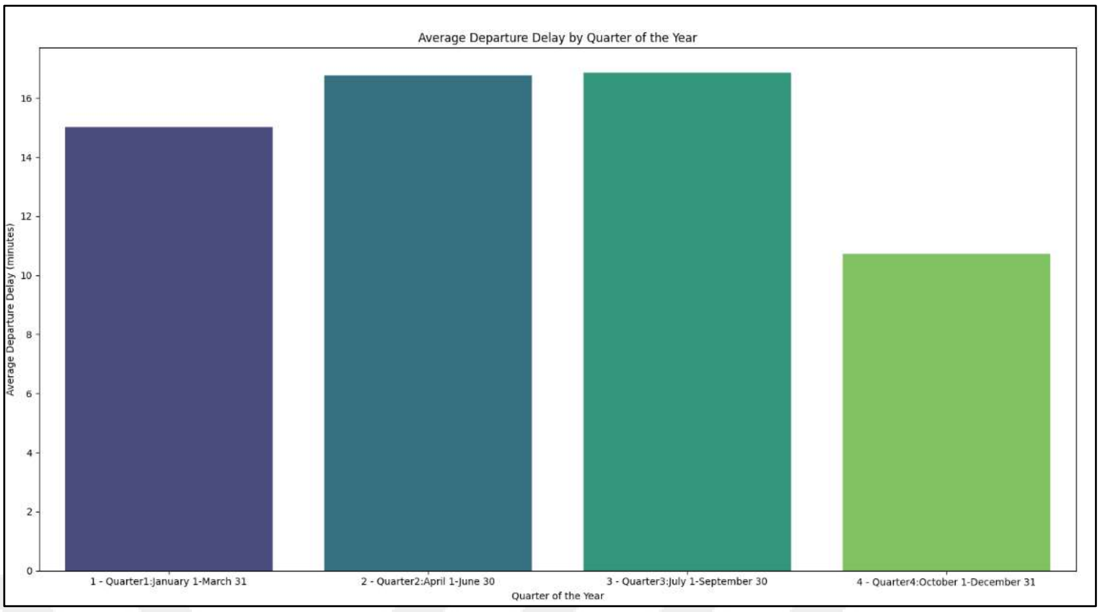

Figure 3.7. Average departure delay distribution by quarter of the year.

Analysis of Variance (ANOVA) is a statistical technique used to compare the means of three or more groups and determine if the differences are significant or just random. In flight delay estimation, using ANOVA on time-related features helps verify whether delays differ across periods. This increases accuracy by ensuring that the model is based on solid data. ANOVA is a meaningful tool for selecting the most important variables for the model.

Table 3.5. ANOVA Test Results

ANOVA RESULTS	F-STATISTICS	P-VALUE
DAY OF THE WEEK	1990.99	0.0
DAY OF THE MONTH	122.44	0.0
QUARTER OF THE YEAR	10958.84	0.0

The ANOVA test results in Table 3.5 show that flight delays vary significantly across different time periods, such as quarter of the year, day of the week, and day of the month. For a quarter of the year, the extremely high F-statistic of 10,958.84 and p-value of 0.0 indicate that the delays vary significantly across seasons, making it an important feature for the prediction model.

Day of the week also shows significant differences in the delay patterns; F-statistic of 1,990.99 and p-value of 0.0 indicate that it is useful to capture the weekly fluctuations in delays.

Although day of the month is also statistically significant, the F-statistic (122.44) is much lower, meaning it has less impact on the delays. It can be removed from the model to simplify things without losing too much accuracy.

The results shown in Table 3.5. confirm that the inclusion of quarter of the year and day of the week will increase the accuracy of the model by accounting for the underlying time-based changes in delays.

3.4.5. Analyzing the Influence of Distance on Delay

In this section we aimed to gain insights into how flight distance affects the probability and magnitude of delays and reveal related features for the model, relationship between distance and delay were investigated.

If a relationship between distance and delay is detected, flight characteristics associated with long or short distances can be used to predict delays. For example, longer flights may experience delays due to factors such as air traffic congestion at busy airports, longer taxi times, or adverse weather conditions along the route. Shorter flights may experience delays related primarily to airport turnaround times, gate availability, or scheduling issues

To capture the potential impact of flight length on delay occurrences, distance can be included as a feature in the delay estimation model, allowing model to account for changes in delay patterns at different flight distances and adapt predictions accordingly. Additionally, understanding the relationship between distance and delay can help prioritize the selection of other relevant features that interact with distance to influence delay outcomes, such as flight time of day, airline operations, or airport infrastructure.

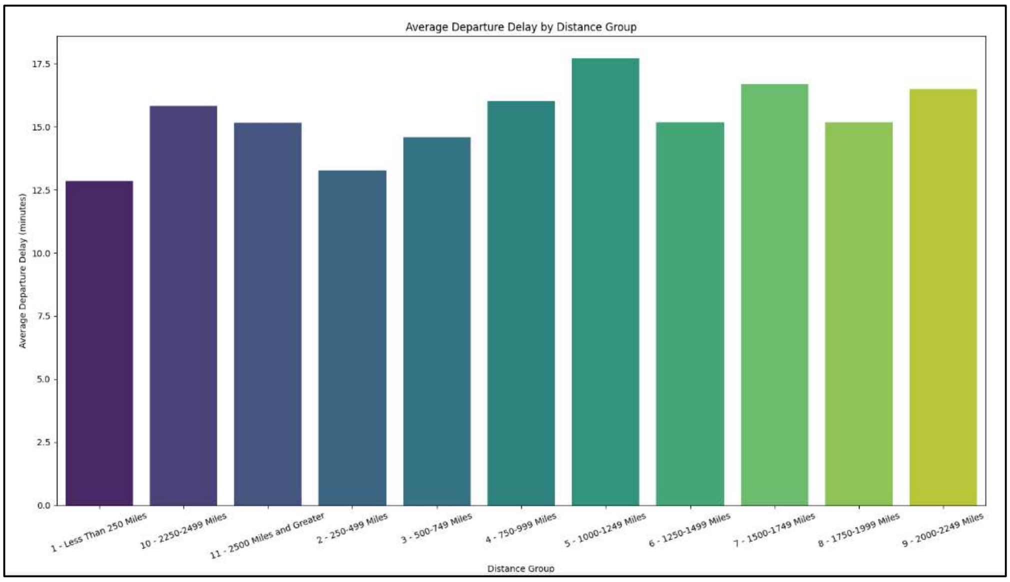

Figure 3.8. Average Departure Delay by Distance Group

The average departure delay by distance group was calculated and plotted as shown in Figure6. However, it did not yield any meaningful insights. Consequently, we opted to utilize the correlation coefficient between distance and departure delay to ascertain the strength and direction of their linear relationship.

3.4.6. Correlation Coefficient

The Pearson correlation coefficient, commonly represented as *r* evaluates the strength and direction of the linear relationship between two variables, X and Y, providing a measure of how one variable changes in relation to the other (Profillidis and Botzoris 2019). The formula for calculating the Pearson correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^{n}(Xi - X)(Yi - Y)}{\sqrt{\sum_{i=1}^{n}(Xi - X)^2 \sum_{i=1}^{n}(Yi - Y)^2}} \tag{3.1}$$

Where:

- *n* is the number of data points.
- Xi and Yi are the individual data points for variables *X* and *Y*, respectively.
- *X* is the mean (average) of the *X* values:

$$X = \sum_{i=1}^{n} Xi \tag{3.2}$$

- *Y* is the mean (average) of the *Y* values:

$$Y = \sum_{i=1}^{n} Yi \tag{3.3}$$

It ranges from -1 to 1, where:

- **1** signifies a perfect positive linear relationship,
- **-1** signifies a perfect negative linear relationship,
- **0** signifies no linear relationship.

First, we look at Pearson correlation coefficient between the continuous variable DISTANCE and the continuous variable DEP_DELAY_NEW. Calculated correlation of 0.03 (as found previously) indicates a very weak positive linear relationship between distance and departure delay.

After we calculated the Pearson correlation coefficient between the categorical variable DISTANCE_GROUP (which has been assigned numeric codes) and the continuous variable DEP_DELAY_NEW. By calculating the correlation between DISTANCE_GROUP and DEP_DELAY_NEW, we aimed to find how strongly and in what direction these two variables are linearly related. The calculated correlation value was 0.02, indicating that there is little to no linear relationship between the distance groups and the departure delay. This suggests that categorizing distances into groups does not reveal a significant trend in delays.

Thus, DISTANCE_GROUP might not be a particularly useful feature for our delay prediction model. However, it might still be worth considering in combination with other features or non-linear models.

3.4.7. Analysis of Delay Distribution by Departure Time Blocks

To determine how the departure time block (DEP_TIME_BLK) affects departure delays, we analyzed both; the number of delays and the average of delays within each time block.

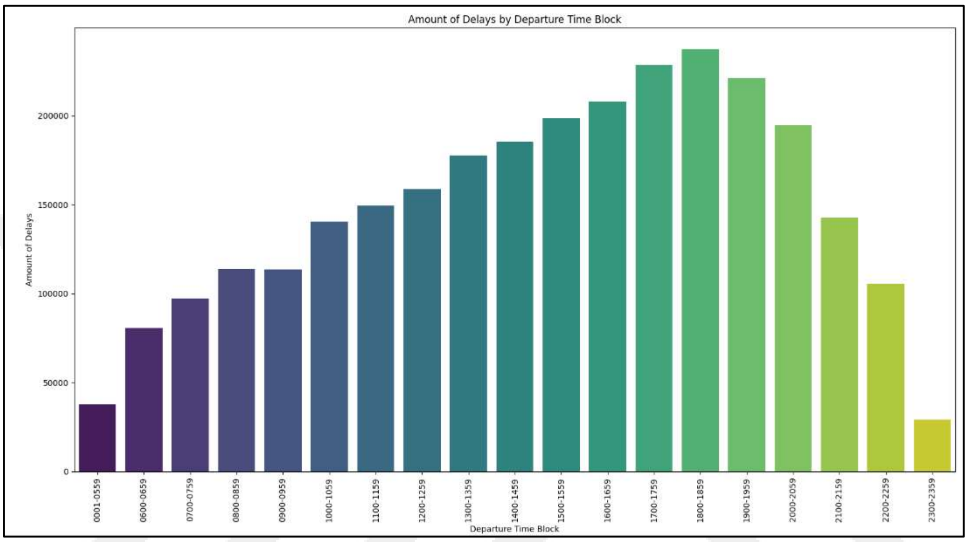

Figure 3.9. Number of Delays (DEP_DEL15) by Departure Time Block

Pros:

- Provides a count of how many flights were delayed within each time block.
- Useful for understanding the frequency of delays in different time periods.
- Helps in identifying peak periods with high numbers of delayed flights.

Cons:

- Does not provide information on the severity or duration of the delays.
- A large number of small delays might be less impactful than a few severe delays, which this metric would not capture.

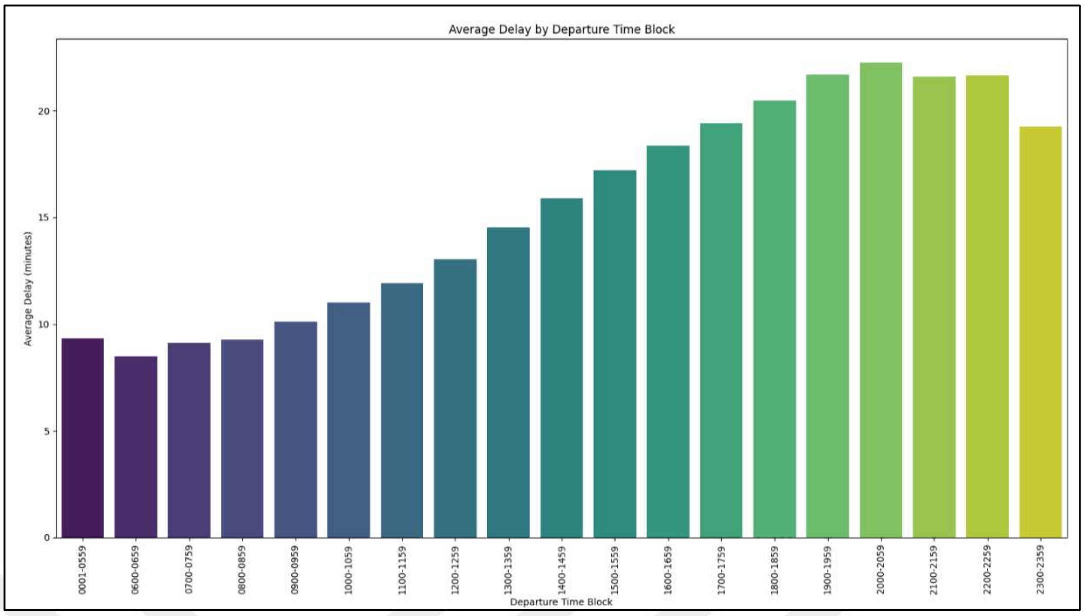

Figure 3. 10. Average of Delays (DEP_DELAY_NEW) by Departure Time Block

Pros:

- Provides insight into the average duration of delays within each time block.
- Useful for understanding the severity of delays during different times of the day.
- Helps in identifying periods with longer average delays, which could indicate more significant operational issues.

Cons:

- Averages can be skewed by extreme values (very long delays).
- Does not provide information on the frequency of delays, only the average duration.
- Both graphs provide valuable insights into the pattern of delays across different departure time blocks, but they highlight different aspects of the delays.

The "Number of Delays (DEP_DEL15)" in Figure 3.9 illustrates the frequency of delays in each time block, highlighting the periods with the highest number of delayed flights. The peak periods for delays are from 1700-1759 and 1800-1859, indicating these times have the most frequent flight delays.

The Average Delay (DEP_DELAY_NEW) graph in Figure 3.10 illustrates the average duration of delays within each time block, providing insight into the length of delays at different times of the day. There is a noticeable upward trend in average delays as the day progresses, with delays peaking in the evening. The longest average delays occur between 20:00 and 20:59.

The comparison of the two graphs in Figure 3.9 and Figure 3.10 shows that during the morning hours (0001-1159), both the number and severity of delays are relatively low, though the frequency of delays gradually increases. By midday to early afternoon (1200-1459), both the frequency and average duration of delays rise, indicating that more flights are delayed and that the delays are getting longer. The late afternoon to early evening period (1500-1959) represents the peak for both the number and severity of delays, with the highest frequency and longest average delay duration occurring during this time. In the evening hours (2000-2359), the number of delays decreases, but the average delay remains high, meaning that although fewer flights are delayed, those that are delayed experience significant delays. While both graphs highlight similar trends, the amount of delays graph focuses on the frequency of delays, identifying peak times, while the average delay graph emphasizes the severity of delays. Together, these graphs provide a comprehensive understanding of delay patterns, which is essential for feature selection to build an effective delay prediction model to help to improve flight scheduling and operational efficiency.

3.4.8. Cancellations by reason

As the Bureau of Transportation Statistics requires causes for delays and cancellations of airlines, the system designed by the Air Carrier On-Time Reporting Advisory Committee classifies the causes of flight cancellations into the following broad categories: Air Carrier, National Aviation System, Weather and Security (Bureau of Transportation Statistics 2023).

Air Carrier: Cancellations due to the internal problems of an air carrier, such as crew availability, maintenance problems, and other logistical challenges.

National Aviation System: Cancellations because of the problems in air traffic control in addition to airport security among other system-wide issues within the national aviation system.

Weather: Flight cancellations due to extremely bad weather conditions, such as storms and fog, or other natural unnamed phenomenon that could interfere with a flight.

Security: It is a cancellation for security reasons that may involve strict measures concerning the safety of the passengers.

In this section, a detailed overview will be shown of the distribution of these categories by specifying the main causes for flight operation consequences. By analyzing this, we'll look at ways to further enhance our flight cancellation prediction model by identifying the most important features.

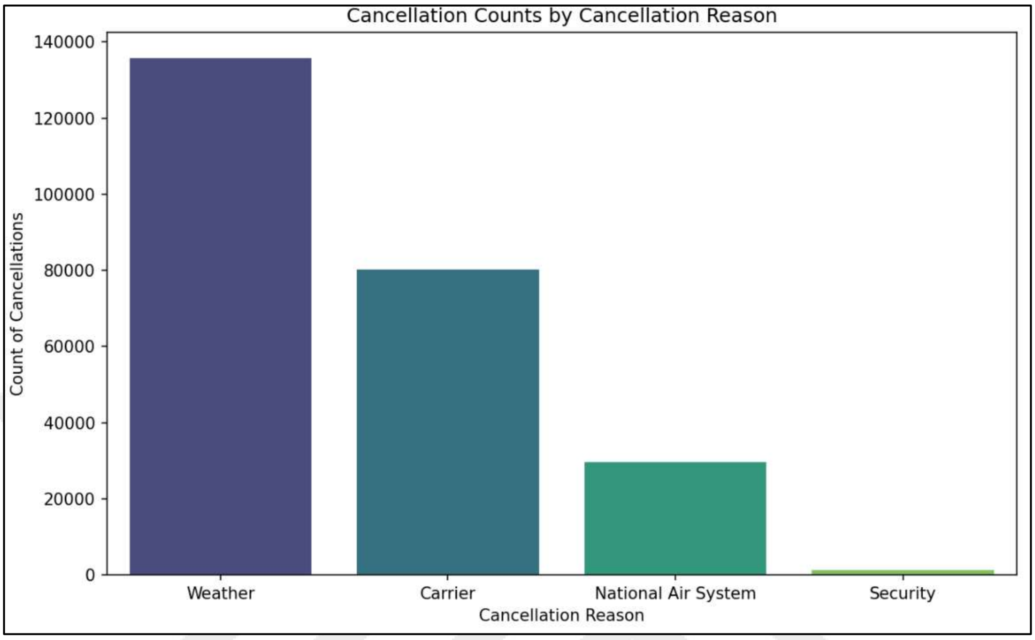

Figure 3. 11. Cancelation Counts by Reason

As shown in Figure 3.11, the top reason for flight cancellations was due to weather conditions, with a total of 135,751 flights canceled due to poor conditions. This is a high number, considering that weather conditions are a major factor affecting flight schedules, considering storms, fog, and other conditions. The second most common reason for cancellations in the same figure was the airline itself, with a total of 80,244 cancellations. This includes cancellations due to operational issues of the airline, crew availability, maintenance issues, and other logistical capacities, which further dictate that improvements in internal operations at the airline are needed to reduce such cancellations. Air traffic control issues, security at airports, and other system-wide issues accounted for 29,641 cancellations in the National Air System. The smallest of these reasons was security-related cancellations, with a total of 1,272 flights canceled due to security concerns.

The distribution of cancellation reasons provides valuable insights for the flight cancellation prediction model. Given that weather is the predominant cause, incorporating detailed weather data into the model will likely enhance its predictive accuracy. Similarly, understanding operational challenges within airlines and systemic issues within the National Air System would significantly enhance the model's ability to predict cancellations and delays. However, obtaining detailed and reliable data on these factors poses substantial challenges due to their complexity and variability.

The distribution of the cancellation reasons gives much valuable information to the flight cancellation prediction model. Since most of the data tends to point at the weather, incorporating comprehensive data about weather conditions is most likely to enhance the predictive capability of the model in this system. Likewise, any operational challenge understanding within the airlines and systemic issues within the National Air System will significantly enhance the predictive ability of cancellations and delays. However, obtaining detailed and reliable data on these factors presents considerable challenges due to their complexity and variability.

3.5. Feature Selection

Feature selection is one of the important steps in any model-building activity, mainly because it underlines the variables which provide the larger amount of power to the model and helps improve the performance of any predictive model (Wang et al. 2016). In the context of a hybrid flight delay/cancellation estimation model, feature selection shall ensure that the model shall capture the underlying pattern and its relationships.

Based on our analyses for these results, various important features can be identified which are likely to influence flight delays and cancellations. These features have been derived from various aspects of flight operations, scheduling, and other external factors such as weather. The selected features include:

- Temporal Features:
    - **QUARTER**: The quarter of the year when the flight is scheduled.
    - **DAY_OF_WEEK**: The day of the week when the flight is scheduled.
    - **DEP_TIME_BLK**: The time block during which the departure occurs.
- Operational Features:
    - **OP_UNIQUE_CARRIER**: The unique carrier code for the operating airline.
    - **ORIGIN**: The origin airport code.
- Delay Indicators:
    - **DEP_DEL15**: Departure delay indicator (1 if 15 minutes or more, 0 otherwise).
    - **DEP_DELAY_NEW**: The new departure delay time in minutes.
- Cancellation Indicators:
    - **CANCELLED**: Flight cancellation indicator (1 if canceled, 0 otherwise)
    - **CANCELLATION CODE:** Reason for flight cancellation (A = carrier, B = weather, C = NAS, D = security.
    -
- Weather and External Factors:
    - **CARRIER_DELAY**: Delay caused by the carrier.
    - **WEATHER_DELAY**: Delay caused by weather conditions.
    - **LATE_AIRCRAFT_DELAY**: Delay caused by late arrival of the aircraft.
    - **Temperature**: Temperature at Origin Airport
    - **Wind speed**: Temperature at Origin Airport
    - **Percipation**: Percipation at Origin Airport

3.6. Feature Engineering and Time Series

The time series analysis in this research helps improve the prediction of flight delays and cancellations. Time series data captures information by integrating time into series. Therefore, it remains an important tool in modeling patterns, trends, and seasonality in flight operations. Time series models make predictions informed by historical data points about the occurrence of delays and cancellations over time, driven by temporal factors such as weather, traffic, and operational bottlenecks. Since most real-world systems evolve moment by moment, enabling our model with time series techniques allows such systems to learn from their past and make more accurate and reliable predictions, not only by taking into account the current situation (Richardson et al. 2024). This aspect is even more serious in aviation, since the usual external factors of the previous days affect current operations. Therefore, time series analysis provides a powerful framework for understanding and mitigating flight disruptions.

This section of the research applies the time series in order to build meaningful features that enhance the predictive power for flight delays using the model developed here (Richardson et al. 2024). The model will use the entire dataset-from the past 2 years-along with the lagged feature from the column DEP_DELAY_NEW, representing the minutes of delay for its departure.

To better align this predictive model with real-world needs, approach refined to incorporate data from the hours leading up to each flight's departure. This adaptation ensures that the model bases its predictions on conditions known 4 hours before the scheduled departure time, providing a practical advantage by allowing actionable insights well before the flight takes off.

`Adjusted_Datetime` feature, which represents the prediction time, set to 4 hours before each flight's scheduled departure. This adjusted timeframe enables the model to integrate recent weather conditions and other variables as they stood 4 hours prior. This realistic shift allows predictions to reflect the latest trends that may affect departure, such as sudden weather changes or operational adjustments at the airport.

To capture delay patterns and trends relevant to the prediction window, several carefully designed time-series features were included into model. Recent Airport (Origin) Delay Trends feature ['AVG_DELAY_ORIGIN_24H_BEFORE'], ['AVG_CANCELLED_ORIGIN_24H_BEFORE'] calculates the average delay and cancellations at a specific airport within the 24 hours leading up to the prediction time, using the `Adjusted_Datetime` column. By filtering for flights from the same airport in this period, the model captures recent patterns at the airport itself, helping it make more informed predictions based on the airport's short-term history.

Carrier Delay and Cancellation Trends Across All Airports as ['AVG_DELAY_ORIGIN_24H_BEFORE'], ['AVG_CANCELLED_ORIGIN_24H_BEFORE'] desined to calculate an average delay and cancellations for each carrier across all airports, again using a 24-hour window leading up to the prediction time. Grouping by carrier and applying a rolling average excludes the current row's delay, allowing the model to recognize each carrier's recent performance without bias from the current instance.

Carrier Delay Trends at Origin Airport were added to model as ['AVG_DELAY_CARRIER_ORIGIN_24H_BEFORE'], ['AVG_CANCELLED_CARRIER_ORIGIN_24H_BEFORE'] for a more granular view and calculated a carrier-specific average delay at each origin airport, focusing on the 24 hours prior to prediction. This tailored feature considers both the specific carrier and airport dynamics, giving the model a precise understanding of how each airline operates at individual airports in the short term.

Moving beyond static weekly and monthly averages, dynamic 7-Day and 30-Day Average Delay feature ['WEEKDAY_AVG_DEP_DELAY'], ['MONTHLY_AVG_DEP_DELAY'] as well as in Cancellation Model ['WEEKDAY_AVG_CANCELLED'], ['MONTHLY_AVG_CANCELLED'] were introduced to capture average delays and cancellations over the past 7 and 30 days relative to the `Adjusted_Datetime`. Using a rolling 7-day, 30-day window, the model derives insights from the most recent weekly and month's trends, accounting for seasonal fluctuations without introducing bias from the current flight.

By focusing on conditions and patterns leading up to 4 hours before scheduled departure, this enhanced methodology creates a more responsive model that captures both recent operational realities and larger patterns affecting flight delays. This approach achieves a balance between precision and practicality, making it more relevant for real-world applications and robust in its predictive power.

Feature engineering is the process of generating new features or transforming existing features in a way that increases the performance of the machine learning model. A new feature called WEEKEND is created based on the DAY_OF_WEEK column. This feature indicates whether the flight is on a weekend or weekday, by representing state by 1 if the day is Saturday or Sunday, 0 otherwise.

3.7. Data Prepocessing

This section is a series of steps applied to the flight dataset to prepare it for the Data Preprocessing phase of model training. First comes optimization of memory usage, converting the data types of columns, which decreases the total memory footprint of this dataset considerably, hence making it far more efficient to process.

Categorical variables such as QUARTER, OP_UNIQUE_CARRIER, ORIGIN, DEP_TIME_BLK, and CANCELLATION_CODE are changed into one-hot encoded vectors using the pd.get_dummies function. This process converts each unique categorical value into a separate binary column (0 or 1), enabling machine learning algorithms, which require numerical inputs, to interpret these variables properly. One-hot encoding ensures that the categorical information is represented without introducing any unintended ordinal relationships.

Feature selection was done in detail, where only the most relevant features to predict flight delays and cancellations are considered. The final feature set developed shall address important information about flights: QUARTER, OP_UNIQUE_CARRIER, ORIGIN, DEP_TIME_BLK, and WEEKEND. In addition, new time series features were introduced to capture recent trends and patterns in delays and cancellations within the 24 hours prior to the prediction time. The features ['AVG_DELAY_ORIGIN_24H_BEFORE'] and ['AVG_CANCELLED_ORIGIN_24H_BEFORE'] calculate the average delay and cancellations at a specific airport within the 24 hours leading up to the prediction time. ['AVG_DELAY_CARRIER_24H_BEFORE'] and ['AVG_CANCELLED_CARRIER_24H_BEFORE'] compute the average delay and cancellations for each carrier across all airports, also using a 24-hour window prior to the prediction. ['AVG_DELAY_CARRIER_ORIGIN_24H_BEFORE'] and ['AVG_CANCELLED_CARRIER_ORIGIN_24H_BEFORE'] calculate a carrier-specific average delay and cancellation rate at each origin airport within the 24-hour period before prediction time. This approach combines recent performance trends with the dynamics of each specific airport, giving the model a nuanced understanding of how each airline operates at individual airports in the short term.

Features such as MONTHLY_AVG_DEP_DELAY and WEEKDAY_AVG_DEP_DELAY modeled seasonal patterns in delays, while the same seasonal patterns in the cancellations were modeled using features such as MONTHLY_AVG_CANCELLED and WEEKDAY_AVG_CANCELLED. In addition, WEATHER_DELAY, temperature_2m, precipitation, and wind_speed_10m were some of the added weather-related variables to capture further environmental factors that may influence flying. Lastly, LATE_AIRCRAFT_DELAY and CANCELLATION_CODE captured operational and cancellation-related factors, respectively.

Missing data, especially in the columns of delay or cancellation, are because of the flight not being delayed or not cancelled, so it represents null values in columns which represent delay in minutes or cancellation codes. In the case of these features representing delays and cancellations, the SimpleImputer strategy fills these null values with a constant of 0 to make the model understand these situations correctly. Where there is no delay or cancellation, it keeps the dataset intact and consistent hence allowing the model to interpret such instances correctly without many problems during the training phase because of missing data (Das et al. 2019)

At the end of these transformations and imputations, the data was converted to a sparse matrix format using sparse.csr_matrix. This format is really efficient in terms of memory usage and computational speed, especially when having large datasets that contain many zero values (El Abbadi 2014). Lastly, the dataset is divided into training blanks-the first 2.5 years-and testing sets-the last 6 months-to ensure the model is tested on unseen data for better performance evaluation.

3.8. Models

Predictive modeling leverages various statistical and machine learning methods to make predictions from historical data. Regression analyses, decision trees, ensemble techniques, and neural networks are some of the methods used. Each one of them has advantages and limitations, making them suitable for different types of data and different tasks of prediction.

3.8.1. Logistic regression

Logistic regression is a statistical technique of binary classification that is meant to estimate the probability of one of two possible outcomes (Peng et al. 2002). In contexts involving flight delay and cancellation prediction, logistic regression will be applied to estimate the likelihood of a flight being delayed or canceled.

It predicts the probability that a given input belongs to a specific class by applying the logistic function to a linear combination of the input features. The coefficients of the features are learned during the training process using optimization techniques that minimize the difference between the predicted probabilities and the actual class labels.

Probability scores indicating the probability of delay or cancellation for each flight are provided by the model. These probability scores are not only useful for making initial predictions but also provide valuable insights into the factors contributing to these outcomes.

The logistic regression model predicts for any given input, X, the probability p that belongs to the positive class, flight being delayed or canceled. It uses a probability model based on the logistic function:

$$p(X) = \frac{1}{1 + e^{-(\beta 0 + \beta 1X1 + \beta 3X2 + \cdots + \beta nXn)}} \tag{3.4}$$

Here:

- $X=(X1,X2,\ldots,Xn)$ represents the input features (e.g., weather conditions, time of day, flight route, etc.).
- $B0$ is the intercept term.
- $\beta 1,\beta 2,\ldots,\beta n$ are the coefficients for each feature, which are learned from the training data.

The logistic regression model is based on the log-odds (or logit) transformation of the probability (Peng et al. 2002):

$$logit\big(p(X)\big) = \log\left(\frac{p(X)}{1 - p(X)}\right) = \beta 0 + \beta 1X1 + \beta 3X2 + \cdots + \beta nXn \tag{3.5}$$

The logistic function $\sigma(z)$ is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.6}$$

Where

$$z = \beta 0 + \sum_{i=1}^{n} \beta i X i \tag{3.7}$$

The coefficients βi quantify the change in log-odds of the outcome for a one-unit change in the corresponding feature Xi, while keeping all other features constant.

The output from the logistic function is constrained to values between 0 and 1 and represents the predicted probability of the outcome-fly delay or fly cancellation-occurring.

Because it provides clear interpretation with robust performance for the classification task, logistic regression was chosen to perform the flight delay and cancellation predictions throughout this thesis. It interprets the logistic regression coefficients that quantify how much influence each predictor variable has on the probability of the outcome; hence, stakeholders would understand and trust these predictions. For aviation, this interpretability is going to be quite crucial since decision-makers will need to know not just what it is predicting, but also why that is the case-so as to take informed actions to avoid delays and disruptions. Further, logistic regression is computationally efficient and handles large data sets quite well, making it an ideal model to work with when it comes to big flight data. It gives probabilistic predictions, so useful in understanding the chance of delays and cancellations, thus helping aviation professionals manage disruptions.

The preprocessing steps performed before applying logistic regression include the transformation of categorical variables by one-hot encoding, filling missing values with SimpleImputer, and creation of time series features like lagged delays and rolling statistics to capture the temporal structure in the data. It then splits the data into training and testing sets, where the first 2.5 years of data are used to train the model and the last 6 months are kept for testing the model.

The target variable for the delay prediction was categorized into three classes: Class 0 delays less than or equal to 15 minutes, Class 1 from 15 to 60 minutes, and Class 2 above 60 minutes. This multiclass classification enables a logistic regression model to predict whether flights are going to face small, moderate, or significant delays, therefore making more effective and accurate predictions based on multiple levels of delay severity.

LogisticRegression is instantiated with max_iter = 5000 to ensure the model actually converges. Then, it is fitted to the preprocessed training data: X_train is the feature set used, and y_delay_class_train and y_cancel_train are the labels for delay classes and cancellations, respectively. It predicts both delay classes and cancellations for which the model has been trained on test data, X_test. The predicted values will be matched against the actual delay classes y_delay_class_test and cancellation outcomes y_cancel_test. Accuracy gives the general performance of the model, reflecting the rate at which the model correctly predicted the delay classes and their cancellations. On top of that, precision, recall, and F1-score give a detailed review of the model's performance in each of the delay and cancellation classes. This is important, since accuracy metrics will be utilized to evaluate the balance between false positives versus false negatives in predicting flight delays or cancellations by the model.

Logistic regression takes in input features such as flight details, weather data, and historical delays to compute the probability of a flight falling into one of the three delay classes. At training time, the model weights each feature with a coefficient-a measure of how the feature weighs in on the prediction for a particular class of delay. For example, a high coefficient on AVG_DELAY_ORIGIN_24H_BEFORE indicates that recency of delays is a strong predictor of future delays. Another important reason for logistic regression is that the algorithm provides interpretability through the clear weighting of each feature in a manner understandable even by decision-makers; hence, the impact of factors like past delays, weather conditions, and time of day will be easily comprehensible. Also, logistic regression is computationally efficient and suitable to process large datasets regarding flights. These also provide probabilistic predictions, which are helpful in decision-making by taking into account the probability of delay rather than absolute classifications.

3.8.2. Decision trees

Decision trees are nonlinear models that split data into branches depending on the values of certain features, thereby effectively capturing complex feature interactions (Peng et al. 2002).

Decision Trees are non-parametric models which divide the feature space into rectangular regions. They make predictions by means of recursive partitioning where the data is split depending on the feature that provides the greatest information gain or, equivalently that most decreases impurity. Each split leads to a branch in the tree and keeps going until an stopping criterion is reached, which could be a maximum depth or minimum number of samples per leaf. This makes decision trees interpretable and able to model nonlinear relationships between features.

As a supervised learning algorithm, decision trees are applicable to both classification and regression tasks (Niu 2018). They function by recursively partitioning the data into subsets based on input feature values, creating a tree-like structure. For predicting flight delays and cancellations, decision trees can model the intricate, non-linear relationships between different factors and the probability of delays or cancellations.

A decision tree is made up of nodes and branches, starting with the root node at the top, which represents the entire dataset and is the point where the initial split occurs (Peng et al. 2002). Internal nodes follow, representing decisions based on specific feature values. The process continues until leaf nodes, also known as terminal nodes, are reached; these represent the final decision or prediction, such as whether a flight is delayed or canceled.

The process of building a decision tree involves the several key steps. At each node, it chooses that feature which will best segregate the data into different classes. It is generally done using impurity criteria, including Gini impurity, entropy, or sometimes variance reduction.

Gini Impurity The probability that an element chosen at random from the set would be mislabeled if it were labeled according to the distribution of labels in the subset.

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{3.8}$$

Where pi is the proportion of samples belonging to class i in subset S.

Entropy and Information Gain refer to the degree of disorder or impurity in a subset. The information gain is simply the reduction of entropy from parent node to child nodes.

$$Entropy(S) = -\sum_{i=1}^{C} p_i \log(p_i) \tag{3.9}$$

$$Information(Gain) = Entropy(parent) - \sum_{k=1}^{m} \frac{|Sk|}{|S|} Entropy(Sk) \tag{3.10}$$

The first step in building a decision tree is to first split the dataset into subsets based on a selected feature and its values, grouping similar examples together and then creating purer nodes. This splitting process is then repeated iteratively on each subset until the stopping criterion is met, creating internal nodes. The stopping criterion can be anything from reaching the maximum tree depth, to having too few examples at a node, or further splits that provide no additional information.

The DecisionTreeClassifier in sklearn by default uses Gini impurity or entropy to find the best feature to make a split on data at each node. Gini impurity is quantified as the frequency of an element being labeled incorrectly if an element is randomly labeled based on the distribution class labels of the subset. It tries to minimize impurity. The entropy, in turn, means the extent of disorder in the data set. Gini impurity will be directly used in our code's decision tree algorithm for choosing features for splitting, to ensure that only the most optimal tree that leads to very accurate predictions is created. As it happens, the explicit calculation of these metrics is neatly done by the scikit-learn library.

Decision trees are chosen for this analysis because of their flexibility and their capability to model very complex nonlinear associations which may exist between input features and target variables: flight delay and cancellation. Unlike linear models, decision trees have the capability to handle variable interactions in a hierarchically dimensional space and, therefore, befit the algorithm for the variety of factors entailed within this dataset, ranging from weather conditions to the previous history of delays down to time-based patterns. This even makes decision trees interpretable intuitively, since every decision is based on a set of if-then conditions-easy to be visualized and captured. Such transparency enables decision-makers in the trenches of the aviation sector to understand the reasoning behind the predictions of models and thus gain real actionable insight into precisely why some flights would eventually be delayed or canceled. Other strengths of decision trees are efficiency in handling numerical and categorical data. The use of a decision tree can also be fruitful in cases with missing or imbalanced data-a very relevant aspect in flight operations data.

This decision tree model is similar to the logistic regression; therefore, similar preprocessing steps of one-hot encoding of the categorical variables will be performed, followed by SimpleImputer imputation of missing values, and created time series features such as lagged delay and rolling statistics. Then, similarly, the same dataset is divided into a training and testing set: the first 2.5 years are used for training, the remaining 6 months for testing, with a complete set of features, including flight-related, weather, and historical features. The delay classification is done in three classes Class 0: 15 minutes or less, Class 1: 15 to 60 minutes, and Class 2: more than 60 minutes.

The DecisionTreeClassifier is applied to model both flight delays and cancellations. First, let's train the decision tree models for each event on pre-processed training data: X_train will be used for features, while y_cancel_train will be used for cancellations or y_delay_class_train for delays. Decision trees use a recursive mechanism of data splitting, where on each node, they decide to split the data based on which feature gives them the most information or reduction of impurity. In this process, a tree is formed in which each path shows the number of decisions taken towards the final result. For predicting cancellation, it learns from features like the details about flights, times of weather changes, and previous delays to arrive at results if any particular flight is liable to be cancelled. It also uses the same feature set for delay prediction tasks, classifying flights into small, moderate, or significant delay.

The trained decision tree is used to make predictions of cancellations and delays on the test data, X_test. Predictions are made by this model and are matched against actual outcomes, y_cancel_test and y_delay_class_test, while performance is evaluated with a number of different metrics. These include accuracy, the proportion of correctly classified predictions, and precision, recall, and F1-score, allowing the model performance-evaluating balance between false positives and false negatives for each class of delay or cancellation predictions.

3.8.3. Hybrid model approach

Hybrid machine learning can be defined as finding an astute amalgamation of various machine learning algorithms to capitalize on their relative strengths and diminish their relative weaknesses (Zhou et al. 2023). Hybrid models mix different models to result in an improved predictive accuracy, robustness, and generalizability. By blending various techniques, hybrid models become more powerful in capturing those complex patterns and interactions in data that could not be discovered by separate models (Rudolph 2024).

Because hybrid approaches leverage the strengths of various models in making their predictions, they generally give higher predictive accuracy than any individual models do (Zhou et al. 2023; Rudloph et al. 2024). This becomes crucial because better forecasts ensure better and more efficient operational planning, thereby allowing for passenger satisfaction. With hybrid models, a greater range of scenarios and larger variation in the data can be dealt with. When different models make different assumptions and mechanisms, the whole system becomes robust against the change in data distribution or outliers. Hybrid models are better equipped to generalize on new, unseen data (Zhou et al. 2023). This aspect is of prime importance for flight delay prediction, as the models have to work on different kinds of airports, airlines, and weather conditions. Different models are able to capture different features of the data. For instance, logistic regression has linearity, while in decision trees, interactions can be nonlinear. Such a combination helps in handling linear and nonlinear relationships in the data efficiently.

In this study, we adopted the hybrid model strategy of integrating logistic regression and a decision tree into one to yield a robust model for flight delay prediction. Logistic regression is an integrated method normally used for modeling probability when the outcome is binary. Everyone happens to be particularly effective when relationships between features and outcomes are well-structured and linear. While decision trees are much better to model such complex nonlinear patterns by recursively partitioning the data depending on some feature values, by combining the two, we leverage the interpretability of logistic regression and the flexibility of decision trees, making a comprehensive model capable of performing well across various scenarios and capturing both the linear and nonlinear relationships in the data.

In our study combining logistic regression and decision trees, the evidence for its effectiveness is consistent with recent studies (Mtimkulu and Maphosa 2023; Bisandu and Moulitsas 2023; Zhou et al. 2023). Therefore, the hybrif method chosen for this domain applies soft voting for each class, where the average of the predicted probabilities for that class across all base models is determined and the class with the highest average probability across all base models is selected as the final prediction (Mahajan et al. 2023; Branco et al. 2023). Although various studies have employed ensemble techniques for disease prediction, voting has emerged as the second-best ensemble method (Mahajan et al. 2023). This method is beneficial because it combines the probabilistic strength of Logistic Regression with the decision-making capabilities of Decision Trees, resulting in a more balanced and accurate prediction.

Firstly, logistic regression and decision tree models were applied separately to predict flight delays and cancellations. These models were used separately to take full advantage of various strengths each model provides. Logistic regression was chosen because it is 'interpretable', providing clear coefficients in the explanation of how each predictor variable affects the likelihood of a delay or cancellation. This can help stakeholders in the aviation industry understand which elements are driving predictions. Decision trees were further applied to grasp more complex nonlinear relationships within these data that are not so well-suited by logistic regression. The decision tree model uncovers patterns involving the interactions of variables-say, weather and historical delays-by recursively partitioning the data into more homogeneous subsets based on the most informative features.

The next stage of work was to combine both models into hybrid, where their predictions would be combined by means of a Voting Classifier. Instead of manual setting, GridSearchCV automated the process of weighing to find the best combination for logistic regression and decision tree models. This is indeed a grid search, comprehensive over the pre-defined parameter grid, that enables the model to automatically experiment with multiple combinations of hyperparameters to identify the best performing configuration. In the case of Voting Classifier, the optimal weights identified were indeed weights = [4,1], in order to prioritize logistic regression's predictive stability and interpretability without losing the advantage of nonlinear relationships captured by the decision tree.

This would be based on an automated tuning process using the class GridSearchCV in the module sklearn.model_selection of the Python library scikit-learn. The method bettered the model by systematically searching for the best weighting with which two such models could be combined. In so doing, the hybrid model would be able to take advantage of the probabilistic interpretation coming from logistic regression and the flexibility of a decision tree when handling complex structures of data. This was an especially appropriate approach in the prediction of flight delays and cancellations in view of the fact that the underlying patterns of interest are linear and nonlinear in nature. Logistic regression is quite efficient in capturing the simpler factors, like weather, whereas decision trees capture such interactions among the variables quite well, such as past delays interacting with other conditions. We used Voting Classifier, with optimized weights, giving us increased resilience and accuracy within a wide range of patterns.

The final prediction y^, based on the soft voting mechanism, can be represented by the formula below, which reflects the combination of predicted probabilities from both models, logistic regression and decision tree.

The final prediction, y^, could be presented as the formula below, using the mechanism of soft voting. It reflects the combination of predicted probabilities from logistic regression and decision tree models.

$$\hat{y} = argmax(w_{LR}.p_{LR}(y = k|X) + wp_{DT}.p_{DT}(y = k|X)) \qquad (3.11)$$

Where y^ is the final predicted class, PLR(y = k | X) is the predicted probability from the logistic regression model for class k and PDT(y = k | X) is the predicted probability from the decision tree model for class k, wLR is the weight assigned to logistic regression and wDT is the weight assigned to the decision tree. k represents the possible classes-for example, delay or not a delay; argmax selects the class with the maximum combined weighted probability.

This formula, therefore, shows that the Voting Classifier combines the outputs of both models by taking the weighted sum of their respective predicted probabilities and selects the class that had the maximum weighted combination of the probability.

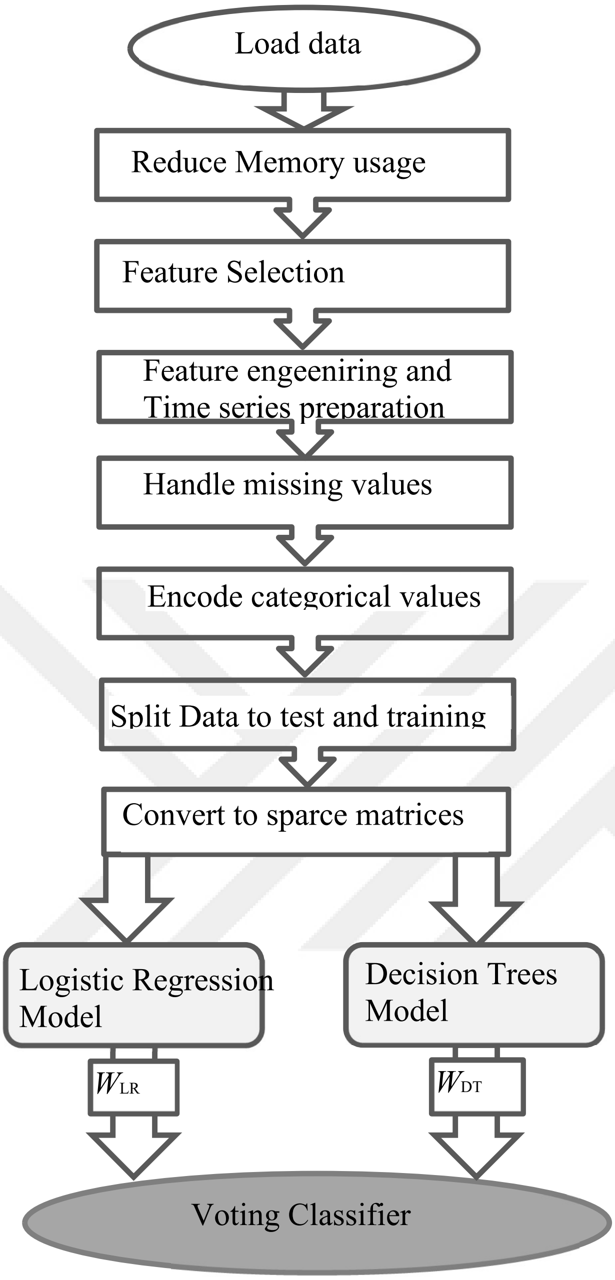

Figure 3.12. Hybrid Prediction Flowchart

The hybrid model is tested with major key performance indicators for accuracy, precision, and recall over all classes. This ultimately provides the best of both models and helps in canceling biases and variances of these individual models for reliable and more accurate prediction (Mahajan et al. 2023).

3.9. Performance Metrics

The key metrics to measure performance include accuracy, precision, recall, and ROC AUC for cancellation. In this case, all of these provide an overall view of the model's performance in predicting flight delays and cancellations: accuracy, precision, recall, and ROC AUC for cancellation. Among them, ROC AUC, which stands for Receiver Operating Characteristic-Area Under Curve, is a very important metric about evaluating the model's cancellation prediction ability when there is an imbalanced dataset.

The cancellation and delay prediction performances were evaluated based on important metrics such as accuracy, precision, recall, and ROC AUC. These will give a full view as to how the models will perform on flight delays and cancellations. Among them, the ROC AUC is really important when there is an imbalance in the dataset; thus, this is crucial in evaluating model performance for cancellation prediction.

ROC AUC is particularly useful for binary classification problems, such as cancellation prediction, where the goal is to distinguish between two classes: canceled and non-canceled flights. It plots the balance between the true positive rate (recall) and the false positive rate at various decision thresholds. The AUC represents the area under the ROC curve and gives a single number summarizing the model's performance. A perfect model would have an AUC of 1.0, indicating it can perfectly distinguish between canceled and non-canceled flights, while an AUC of 0.5 suggests no discriminative power (random guessing).

Given that canceled flights represent only a small fraction of the total (around 2%), ROC AUC becomes a crucial metric for cancellation prediction. Traditional metrics like accuracy might give misleading results, as a model can achieve high accuracy simply by predicting that most flights will not be canceled. However, ROC AUC captures the model's ability to rank predictions, ensuring that it performs well in identifying the rare cancellations (Richardson et al. 2024). This makes it ideal for assessing performance in highly imbalanced datasets like flight cancellations, where it is essential to detect rare events while minimizing false positives.

Accuracy is the ratio of correctly predicted instances to the total instances. It measures the overall correctness of the model (Hossin and Sulaiman 2015)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.11}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Accuracy gives an overall idea of how many predictions are correct, but it can be misleading if the classes are imbalanced.

Precision is the proportion of true positive predictions to the total positive predictions made by the model. It assesses how accurate the positive predictions are.

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive} \tag{3.12}$$

Precision is important when the consequences of false positives are significant. A high precision value indicates that the model has a low rate of false positive predictions (Hossin and Sulaiman 2015; Sokolova and Lapalme 2009).

Recall (Sensitivity or True Positive Rate Detection) is the proportion of true positive instances correctly identified by the model compared to all actual positive instances. It measures how well the model captures positive cases (Hossin and Sulaiman 2015; Sokolova and Lapalme 2009).

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \tag{3.13}$$

Recall is important when the cost of false negatives is high. High recall indicates a low false negative rate.

F1-Score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall.

$$F1\ score = \frac{Precision \text{ X } Recall}{Precision+ Recall} \tag{3.14}$$

The F1-Score is useful when the class distribution is imbalanced, as it considers both false positives and false negatives (Hossin and Sulaiman 2015; Sokolova and Lapalme 2009).

4. RESULTS AND DISCUSSION

In this study, a hybrid model combining logistic regression and decision tree models that can predict flight delays and cancellations is implemented. Each model will be trained separately on data from each airline to learn the operating model and then combined using a voting classifier. The hybrid method combines the strengths of both models; logistic regression handles imbalanced data well, while decision trees capture nonlinear relationships, thus making more robust and balanced predictions. The approach outperformed individual models in better recall of cancelled flights without sacrificing much precision.

The original dataset had two years of flight data for 21 airlines. Given the limited hardware and computational resources, we couldn't process all data in one go. Instead, we did the calculation for each airline separately and restricted our consideration to the 50 airports with the highest number of departures to keep the computation manageable.

For example, the dataset of one of the largest airlines, WL, included 2,357,714 rows and 20 features. When time series features were included in this set after application of `create_dummies` function for conversion of categorical variables into one-hot encoded variables, the dimensions of the dataset increased considerably. The Python function create_dummies(), using the method pd.get_dummies() from the pandas library, turned categorical columns such as origin airport, time block, carrier into separate binary columns representing each category. This step is necessary because most machine learning algorithms require numerical input. Thus, the WL dataset increased in size from 2,357,714 rows and 20 columns to 2,357,714 rows and 87 columns upon the creation of this feature engineering step for. This increase in dimensionality added to the scale of the dataset significantly increased both its memory usage and computational time. To find an optimal balance between preserving accuracy and reducing computation complexity, we selected the top 50 airports based on the frequency of departures for each airline. This sampling was done based upon the rationale that these busy airports account for a large percentage of operational activity occurring and would be more apt to show meaningful patterns regarding delay and cancellation predictions. By narrowing the analysis to the most relevant airports, we reduced the dataset size without compromising the representativeness of the data, allowing for more efficient processing while preserving the integrity of the model's predictions.

Logistic Regression model was regularized using L2 regularization to prevent overfitting, with the C=1.0 parameter controlling the strength of the regularization. The lbfgs solver was used to handle larger datasets efficiently. To manage convergence during training, we set the maximum iteration count to 5000. The class weights were used to compensate for the imbalance in the dataset so that Class 1, representing the cancelled flights, was not left out.

Table 4.1. Logistic Regression Model Parameters

Parameter	Value/Description
penalty	L2 regularization
C	1.0 (Inverse of regularization strength)
solver	'lbfgs'
max_iter	5000 (Increased iterations)
class_weight	'balanced' (Handles class imbalance)
random_state	42 (Ensures reproducibility)

In order to prevent overfitting, the depth of the Decision Tree was limited to 10; min_samples_split=20 and min_samples_leaf=10 helped to further control overfitting by limiting the number of samples required for the split and at the leaf nodes. This uses max_features='sqrt' and ensures that only a fraction of features is considered in each split, which encourages too much diversity across the splits of this model. As in Logistic Regression, the class weights were balanced by setting the reproducibility to a random case of 42 to balance an unbalanced dataset.

Table 4.2. Decision Trees Model Parameters

Parameter	Value/Description
max_depth	10 (Limits the depth of the tree)
min_samples_split	20 (Minimum samples required to split)
min_samples_leaf	10 (Minimum samples required at leaf nodes)
max_features	'sqrt' (Number of features considered for split)
class_weight	'balanced' (Handles class imbalance)
random_state	42 (Ensures reproducibility)

4.1. Cancellation Model Results

In the Cancellation Model, we consider two different classes of flights: Class 0 (Non-cancelled flights) and Class 1 (Cancelled flights). Class 0 represents all flights that were not cancelled and completed their journeys, while in this context Class 1 flights are cases which were cancelled before taking off. Precision, recall, and F1-score for each class give a view regarding the performance of the model in distinguishing whether flights are cancelled or not. Precision for Class 0 measures the precision in identifying flights that are not cancelled, while recall makes sure all flights not cancelled are captured. For Class 1, precision indicates how well the model identifies canceled flights, and recall focuses on capturing as many canceled flights as possible.

Initially, the model is trained using Logistic Regression with L2 regularization to handle class imbalance. Regularization helped prevent overfitting, which can occur when a model captures noise in the data rather than general patterns. Logistic regression, combined with regularization and balanced class weights, allowed for an optimized trade-off between precision and recall, ensuring that canceled flights (a minority class) received the necessary attention. To improve the model's generalization, I trained a Decision Tree with parameters such as limited depth and balanced class weights, which ensured the model could generalize well while maintaining a focus on the minority class.

To further enhance the model's performance, both the logistic regression and decision tree models combined using a Voting Classifier with soft voting, where the predicted probabilities were averaged. We used Grid Search to fine-tune the model's weightings, optimizing the combination of the two models. This approach ensured a well-regularized and balanced model that performed well across both classes, avoiding overfitting while providing robust predictions for flight cancellations.

Table 4.3. Cancellation Model Results

Models	**Airline**	**Accuracy**	**Precision (Class 0)**	**Recall (Class 0)**	**F1-Score (Class 0)**	**Precision (Class 1)**	**Recall (Class 1)**	**F1-Score (Class 1)**	**ROC AUC**
Logistic Regression		0.995	1.0	1.0	1.0	1.0	0.75	0.86	0.874
Decision Tree		0.996	1.0	1.0	1.0	0.95	0.79	0.85	0.884
Hybrid Model	WN	0.997	1.0	1.0	1.0	0.97	0.81	0.87	0.895
Logistic Regression		0.991	0.98	1.0	1.0	0.83	0.40	0.58	0.704
Decision Tree	DL	0.990	00.97	1.0	1.0	0.82	0.41	0.57	0.708
Hybrid Model		0.993	0.99	1.0	1.0	0.90	0.42	0.58	0.712
Logistic Regression		0.995	0.99	1.0	1.0	0.94	0.71	0.81	0.859
Decision Tree	AA	0.993	0.99	1.0	1.0	0.93	0.72	0.80	0.860
Hybrid Model		0.995	0.99	1.0	1.0	0.95	0.73	0.82	0.868
Logistic Regression	OO	0.997	1.0	1.0	1.0	0.97	0.87	0.92	0.941
Decision Tree		0.996	1.0	1.0	1.0	0.96	0.88	0.92	0.942

Table 4.3. Continued

Hybrid Model		0.998	1.0	1.0	1.0	0.98	0.89	0.94	0.944
Logistic Regression	UA	0.993	0.99	1.0	1.0	0.88	0.53	0.64	0.763
Decision Tree		0.992	0.99	1.0	1.0	0.87	0.57	0.66	0.787
Hybrid Model		0.994	0.99	1.0	1.0	0.9	0.64	0.68	0.792
Logistic Regression	YX	0.993	1.0	1.0	1.0	1.0	0.85	0.91	0.937
Decision Tree		0.993	1.0	1.0	1.0	0.94	0.87	0.89	0.940
Hybrid Model		0.995	1.0	1.0	1.0	0.97	0.89	0.93	0.950
Logistic Regression	B6	0.985	0.97	1.0	0.96	1.0	0.55	0.67	0.776
Decision Tree		0.984	0.96	0.99	0.97	0.76	0.6	0.66	0.792
Hybrid Model		0.988	0.99	1.0	0.99	0.79	0.65	0.67	0.796
Logistic Regression	NK	0.995	1.0	1.0	1.0	1.0	0.76	0.80	0.878
Decision Tree		0.994	1.0	1.0	1.0	0.86	0.79	0.82	0.890
Hybrid Model		0.996	1.0	1.0	1.0	0.89	0.78	0.86	0.892
Logistic Regression	MQ	0.998	1.0	1.0	1.0	0.97	0.90	0.94	0.953
Decision Tree		0.997	1.0	1.0	1.0	0.96	0.89	0.93	0.953
Hybrid Model		0.999	1.0	1.0	1.0	0.98	0.91	0.95	0.953
Logistic Regression	AS	0.991	0.96	1.0	1.0	0.99	0.2	0.33	0.599
Decision Tree		0.987	0.97	1.0	0.99	0.43	0.25	0.32	0.623
Hybrid Model		0.988	0.99	1.0	0.99	0.45	0.32	0.34	0.637

Based on the results in Table 4.1. Logistic Regression consistently delivered high accuracy across all airlines, ranging from 0.985 for B6 to 0.998 for MQ. Precision for Class 0 (non-cancelled flights) was perfect in most cases, indicating that nearly all predicted non-canceled flights were correctly identified. However, the model's performance for Class 1 (canceled flights) varied significantly, with recall values ranging from 0.20 for AS to 1.0 for airlines like NK and YX. This indicates that while Logistic Regression effectively avoided false positives, it struggled to capture all cancellations in certain cases, such as for AS and UA, where the recall was particularly low.

The ROC AUC scores for Logistic Regression reflected the model's ability to distinguish between canceled and non-canceled flights, with values ranging from 0.599 (AS) to 0.953 (MQ). The lower AUC for some airlines indicates that the model was less effective in capturing the nuances of cancellation prediction for those carriers, especially when the class imbalance was more pronounced.

The Decision Tree model also performed well across most airlines, with accuracy generally similar to Logistic Regression, though slightly lower in some cases (e.g., 0.984 for B6). While the precision for both classes was generally high, the recall for canceled flights was more variable, ranging from 0.25 (AS) to 0.89 (MQ, OO). The Decision Tree tended to perform better than Logistic Regression in capturing more canceled flights, as seen in the increased recall values for Class 1 in several airlines.

The ROC AUC scores for Decision Tree were generally close to those of Logistic Regression but slightly improved in most cases, reflecting its ability to better capture the complexity of flight cancellations, especially in airlines with a more significant cancellation rate, such as MQ (0.953) and OO (0.942). However, the Decision Tree struggled more with airlines like AS, where the AUC was lower (0.623), indicating that further tuning would be required to handle imbalanced datasets for certain carriers.

The Hybrid Model, which combines the strengths of Logistic Regression and Decision Trees, consistently outperformed the individual models. It gave the highest accuracy, hence large values such as 0.999 for MQ and 0.988 for B6. For Class 0, which referred to non-canceled flights, all airlines had perfect precision and recall using the Hybrid Model, while in Class 1, corresponding to canceled flights, the recall using the Hybrid Model was higher than when either Logistic Regression or Decision Trees was used alone.

Also, the ROC AUC scores of the model were significantly superior, with OO standing out at 0.944, MQ at 0.953, and YX at 0.950. The fact that this is so proves that the Hybrid Model had better performance in recognizing canceled flights from non-canceled ones; hence, it is the best choice when dealing with such an imbalanced dataset. Its performance was much better at recalling airlines like AS and UA, where the recall score increased a lot while the AUC reached more acceptable levels compared to the single models.

On the whole, the Hybrid Model balanced everything for all airlines with very high accuracy, precision, and recall, and very strong ROC AUC scores. Whereas Logistic Regression had strength in its precision with low false positive rates, it had poor recall in showing canceled flights of some airlines, a very important factor in predicting cancellations. Although the Decision Tree improved recall compared to Logistic Regression, the Hybrid Model was still better in terms of overall balance and predictive power.

The Hybrid Model was, in turn, the most reliable method of solving the problem of prediction cancellations on imbalanced datasets and managed to combine precision and recall into a wholesome insight.

4.2. Delay Model Results

In the developed delay prediction model, flights have been categorized into three classes: Class 0 (no or minimal delay), Class 1 (moderate delay), and Class 2 (severe delay). The performance of the model in terms of correct classification for no/minimal delay, moderate delay, and severe delay has been identified based on key metrics such as accuracy, precision, recall, and F1-score.

Table 4.4. Delay Model

Models	Airline	Accuracy	Precision (Class0)	Recall (Class0)	FScore (Class0)	Precision (Class1)	Recall (Class1)	FScore (Class1)
Logistc Regression	WN	0.835	0.85	0.95	0.90	0.75	0.51	0.60
Decision Tree		0.736	0.87	0.8	0.83	0.49	0.55	0.52
Hybrid Model		0.837	0.86	0.97	0.91	0.76	0.57	0.62
Logistic Regression	DL	0.830	0.83	0.97	0.90	0.70	0.24	0.36
Decision Tree		0.743	0.87	0.84	0.85	0.33	0.36	0.34
Hybrid Model		0.837	0.85	0.99	0.91	0.72	0.45	0.37
Logistic Regression	AA	0.847	0.85	0.96	0.90	0.74	0.35	0.5
Decision Tree		0.749	0.87	0.83	0.85	0.4	0.44	0.42
Hybrid Model		0.848	0.86	0.98	0.92	0.76	0.38	0.51
Logistic Regression	OO	0.892	0.9	0.99	0.94	0.84	0.21	0.33
Decision Tree		0.822	0.91	0.89	0.9	0.31	0.32	0.33
Hybrid Model		0.894	0.9	0.99	0.95	0.82	0.35	0.37
Logistic Regression	A	0.846	0.85	0.99	0.92	0.75	0.26	0.38
Decision Tree		0.757	0.87	0.85	0.86	0.35	0.36	0.37
Hybrid Model		0.847	0.86	0.99	0.92	0.74	0.37	0.4
Logistic Regression	YX	0.913	0.92	0.99	0.96	0.76	0.29	0.42
Decision Tree		0.855	0.93	0.91	0.92	0.33	0.39	0.36
Hybrid Model		0.914	0.92	0.99	0.96	0.77	0.41	0.44
Logistic Regression	B6	0.770	0.76	0.97	0.87	0.69	0.22	0.33
Decision Tree		0.673	0.81	0.77	0.79	0.35	0.37	0.36
Hybrid Model		0.773	0.79	0.96	0.87	0.62	0.25	0.35
Logistic Regression	NK	0.752	0.76	0.96	0.84	0.75	0.2	0.32
Decision Tree		0.658	0.79	0.78	0.78	0.36	0.35	0.34
Hybrid Model		0.759	0.77	0.98	0.86	0.71	0.22	0.36
Logistic Regression	MQ	0.910	0.93	0.99	0.95	0.80	0.46	0.59
Decision Tree		0.848	0.93	0.9	0.92	0.44	0.51	0.47

Table 4.4. Continued

Hybrid Model		0.917	0.93	0.99	0.96	0.83	0.47	0.62
Logistic Regression	AS	0.882	0.89	0.96	0.91	0.81	0.4	0.53
Decision Tree		0.807	0.91	0.87	0.89	0.42	0.49	0.46
Hybrid Model		0.885	0.8	0.99	0.94	0.85	0.5	0.64

Overall, the performances of the Logistic Regression model were high, ranging from 0.752 for NK to 0.913 for YX. It performed very well in Class 0, where the precision was very high and correctly flagged flights with no or very little delay. However, the model performed poorly in recalling flights in Classes 1 and 2, which had moderate and severe delays, respectively. For example, NK had a recall of 0.20 for Moderate Delays and OO had a recall of 0.21 for severe delays, indicating that Logistic Regression was failing to capture a significant portion of delayed flights, especially the more severe ones. The low F1 scores for delayed flight classes illustrated the riskiness of these harder-to-capture categories.

The Decision Tree model, posting accuracy scores like 0.673 for B6 and 0.658 for NK, while generally less accurate than Logistic Regression, performed better at identifying delayed flights, particularly moderate class 1 and severe class 2 delays. Using MQ as an example, recall for moderate delays reached 0.51, while Logistic Regression only managed 0.46. Similar to Logistic Regression, Decision Tree struggled with both precision and recall for severe delays, with low F1 scores. The model was not effective at dealing with imbalances across delay categories, making it difficult to predict these rare but significant delays.

The Hybrid Model, which combines predictions from both logistic regressions and decision trees, always gave the highest results for each airline. Accuracy increased for most carriers, with values such as 0.917 for MQ and 0.914 for YX. While precision for no delays Class 0 remained nearly perfect for all airlines, there were significant improvements in recall for flights experiencing delays, Class 1 and Class 2. For example, in MQ, recall for moderate delays increased to 0.47, and recall for severe delays even increased to 0.62, significantly better than anything achieved with Logistic Regression or Decision Trees separately.

The Hybrid Model provided a better balance between precision and recall, which actually made it more reliable in capturing delays across all classes. This actually saves the weaknesses of Logistic Regression, which tends to miss these delays, and Decision Tree, as it is skewed across classes. The Hybrid Model has become the most robust solution for the flight delay estimation problem.

In summary, the performance of Hybrid Model was better compared to Logistic Regression and Decision Trees, especially in terms of recall improvement for delayed flights. Although Logistic Regression had much better performance in terms of accuracy with good precision, it has very poor recall, especially in detecting severely delayed flights. The Decision Tree promised better recall but overall had much lower accuracy.

Therefore, by combining both, the Hybrid Model thus provided a balanced and effective solution to flight delay prediction, hence the scores are better in handling such complexities and imbalances amongst various airlines while predicting delays.

4.3. Impact of Forecast Horizon on Prediction Performance

In this section, we analyze the effect of the forecast horizon on the accuracy and F1 score of our predictive models. Predictions were made over five more different intervals prior to the scheduled departure, namely, 12, 24, 48, 72, and 168 hours. These represent distinct points in time and let us observe how the predictive accuracy and quality of classification develop as the model forecasts further from the time of departure.

Graphical visualizations are provided for comparing the results obtained across these different forecast horizons with average accuracies and F1 scores of each interval.

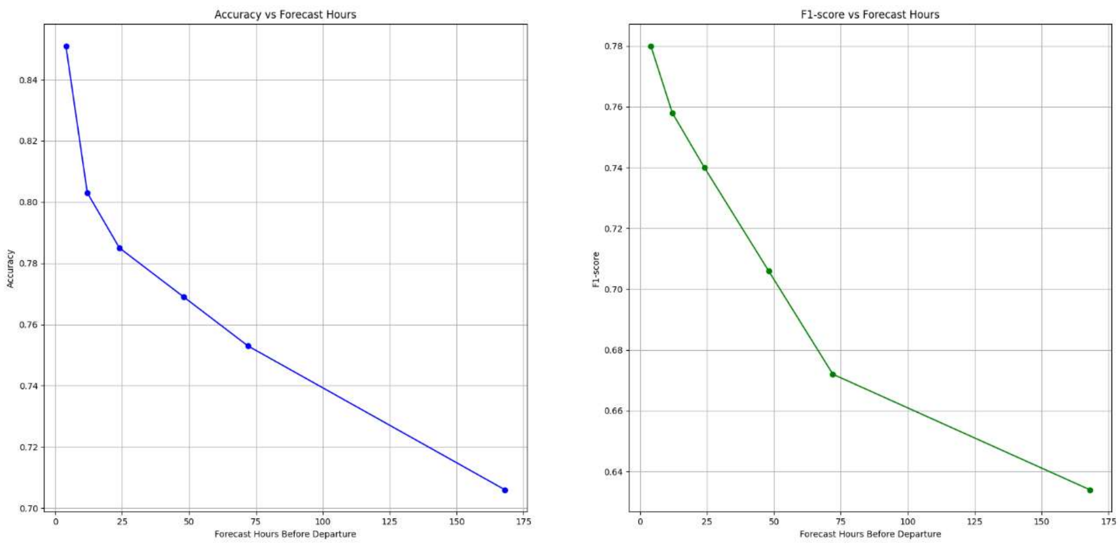

Figure 4.1. Effect of Forecast Time on Delay Model Accuracy and F1-Score

It is expected that the nearest intervals to departure time, especially at 4 hours, receive the highest performance in capturing the conditions nearer to real time. For longer horizons, small degradations in both accuracy and F1 scores appear with rising uncertainty over time due to changes in operation or shifting weather patterns.

These findings give evidence of the model's ability to adjust its predictive power according to proximity to the time of departure, where the best performance occurs at a forecast horizon of 4 hours.

The following comparative graphs outline how Cancel model performances change when forecast horizons increase and point toward both the robustness of the 4-hour predictions and the model's adaptability over time.

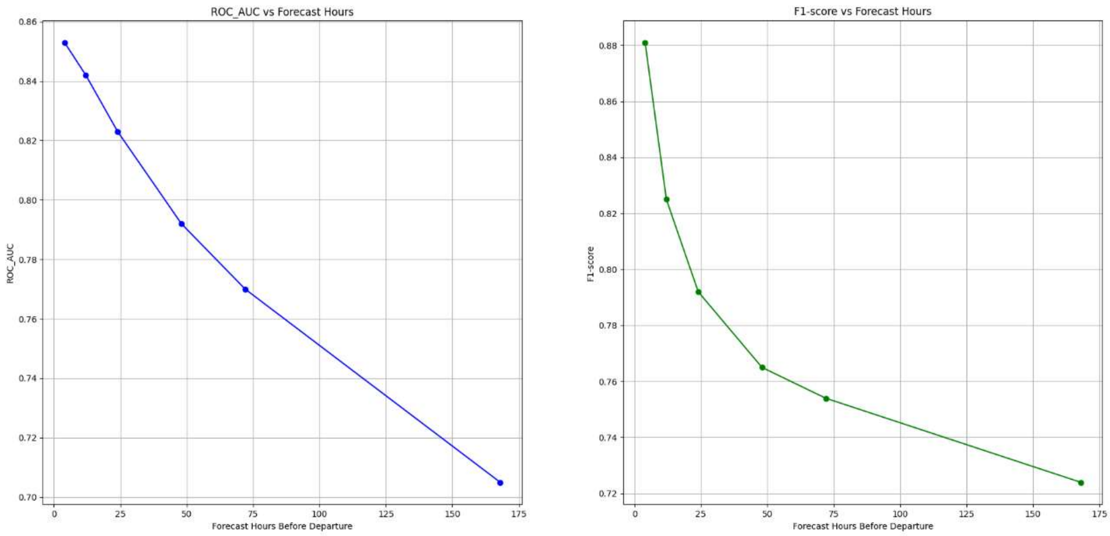

Figure 4.2. Effect of Forecast Time on Cancel Model Roc AUC and F1-Score

When evaluating the performance of a cancellation prediction model, analyzing the changes in F1 score and ROC AUC over different prediction horizons is crucial for understanding the model's effectiveness at various lead times. The F1 score provides a balance between precision and recall, highlighting the model's ability to accurately identify both canceled and non-canceled flights, which is particularly important in imbalanced datasets where cancellations are rare. On the other hand, ROC AUC offers a comprehensive view of the model's discriminatory power across all classification thresholds, illustrating how well it distinguishes between the two classes regardless of the specific threshold chosen.

In this analysis, as the prediction horizon increases from 4 to 168 hours, both the F1 score and ROC AUC show a decreasing trend, with F1 scores dropping from 0.881 to 0.724 and ROC AUC values declining from 0.853 to 0.705. This indicates that the model's predictive performance diminishes over time, suggesting that earlier predictions yield more reliable outcomes compared to those made further in advance.

5. CONCLUSION

5.1. Summary of Findings

This thesis focused on predicting flight delays and cancellations using a hybrid modeling approach combining logistic regression and decision trees. The study analyzed 10 airlines and their 50 most frequent destinations, allowing the models to adapt to the specific operational environments of each carrier. We observed that the hybrid method, which combines the strengths of both logistic regression and decision trees, outperformed the individual models.

The results showed that this combined approach provided more balanced and accurate predictions, especially for flights that were less frequently canceled. The hybrid model improved recall for cancellations without sacrificing precision and provided better insights tailored to each airline's unique operations and challenges. As a result, this approach significantly increased the ability to predict both flight delays and cancellations, making it a valuable tool for optimizing airline operations and resource allocation.

5.2. Contributions to the Field

This study introduces a new hybrid prediction model algorithm which connects the strengths of logistic regression with decision trees to improve the accuracy in its predictions of flight delays and cancellations. Unlike previous models, most of which use a single algorithm, this hybrid outperforms each technique with greater emphasis on strengths. In logistic regression, it offers strength in handling imbalanced data, especially for the forecast of no usual events such as cancellations, while decision trees contemporarily offer a non-linear, complex relationship. It then follows that the proposed hybrid methodology using Voting Classifier outperforms each model's accuracy and balanced predictions.

This hybrid approach was tested against tailored models developed for each individual airline, showing superior predictive performance, both for precision and recall. The combined method developed here resulted in an overall higher performance than logistic regression and decision trees taken individually by balancing the capability of detecting cancellations with a minimum of false positives. This is the hybrid approach that manages to handle unique operational characteristics both from airlines and airports, while improving prediction reliability.

This research conclusively establishes value in air-carrier-specific models but goes on to prove that the hybrid is able to generalize its predictions across different carriers more effectively. Combining both in a voting ensemble allows for a more comprehensive and actionable insight; hence, this will be a great addition in the field of predictive modeling in aviation.

5.3. Limitations of the Study

One limitation of this study is that data have been taken from certain airlines and airports, which might make the generality of models in other contexts difficult in the absence of further modifications. Indeed, all the models obtained within this study are highly fitted to specific operational environments, and their applications would have to be checked for accuracy by additional calibration and validation in settings different from those surveyed.

Additionally, the study identified not only the most relevant factors leading to flight delays and cancellations but also did not provide data on such relevant factors like real-time road traffic and more detailed weather. The absence of such variables in the feature space may reduce the ability of the models to capture all relevant factors and hence impact their performance.

Handling various scenarios and adapting the models to specific contexts added considerable complexity. Hence, the computational workload and expertise might become an obstacle for real-life applications. Moreover, the authors did not test the models with all available data due to limitations in hardware: 21 airlines and 375 origins. These models are not considering evolving operational practices and changes over time, which can likely affect their long-term reliability and will be in need of periodic updating if they are to remain effective.

5.4. Recommendations for Future Research

Future research should thus be directed at integrating the models presented here with real-time data feeding, such as but not limited to real-time weather conditions and air traffic situations, to arrive at more realistic predictions of delay occurrences. These dynamic points will go a long way in offering a comprehensive breakdown of the variables responsible for flight delays and cancellations with more accurate and timely predictions.

Also needed for model performance improvement is the enlargement of the feature set. Additional features can be included at a more detailed and diverse level, such as aircraft maintenance schedules and crew rotations, which in turn would provide more specific insight into operational challenges. In addition, mechanisms need to be developed for the updating of models on a regular basis to keep them relevant and their accuracy ongoing over time, reflecting changes in both operational environments and practices.

Further validation of models can be done by testing them across different geographical and operational contexts. This kind of cross-context validation will help assess the generalizability and spell out what changes are necessary for generalization. The investigation of more advanced hybrid and ensemble methods will promote further predictive accuracy and robustness, especially in handling diverse and complex datasets. However, this does presuppose close cooperation with airline and airport operators who co-develop the models with a view to overcoming particular operational challenges, while at the same time remaining as near as possible to everyday practice and needs.

6. REFERENCES

Cheevachaipimol, W., Teinwan, B., & Chutima, P. 2021. Flight Delay Prediction Using a Hybrid Deep Learning Method. *Engineering Journal*, 25(8):99-112.

Yu, Y., Mo, H., & Haifeng, L. 2019. A Classification Prediction Analysis of Flight Cancellation Based on Spark. *Procedia Computer Science*, 162: 480-486.

International Civil Aviation Organization. 2019. Aviaotion Benefits Report. https://www.icao.int/sustainability/Documents/AVIATION-BENEFITS-2019-web.pdf [Accesed: 01.06.2024].

Zhang, J. Z., Srivastava, P. R., Sharma, D., and Eachempati, P. 2021. Big Data Analytics and Machine Learning: A Retrospective Overview and Bibliometric Analysis. *Expert Systems with Applications*, 184, 115561.

Federal Aviation Administration. 2022. Air Traffic by the Numbers. https://www.faa.gov/sites/faa.gov/files/air_traffic/by_the_numbers/Air_Traffic_by_the_Numbers_2022.pdf.

Carvalho, L., Sternberg, A., Gonçalves, L. M., Cruz, A. B., Soares, J. A., Brandão, D., and Ogasawara, E. 2020. On the Relevance of Data Science for Flight Delay Research: A Systematic Review. *Transport Reviews*.

Wang, T., Zheng, Y., & Xu, H. 2022. A Review of Flight Delay Prediction Methods. 2nd International Conference on Big Data Engineering and Education (BDEE), pp. 135-141, August 5-7, Chengdu, China.

Fleurquin, P., Campanelli, B., Eguiluz, V., & Ramasco, J. 2014. Trees of Reactionary Delay: Addressing the Dynamical Robustness of the US Air Transportation Network. SESAR Innovation Days, November 25-27, Madrid.

Britto, R., Dresner, M., & Voltes, A. 2012. The Impact of Flight Delays on Passenger Demand and Societal Welfare. *Transportation Research Part E: Logistics and Transportation Review*, 48(2): 460-469.

Kešel'ová, M., & Hanák, P. 2019. Risk and Opportunities in the Process of Flight Delay. New Trends in Aviation Development (NTAD) Conferrence, pp. 87-91, September 26-27, Chlumec nad Cidlinou, Czech Republic.

Rebollo, J. J., & Balakrishnan, H. 2014. Characterization and Prediction of Air Traffic Delays. *Transportation Research Part C: Emerging Technologies*, 44: 231-241.

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. 2020. Flight Delay Prediction Based on Aviation Big Data and Machine Learning. *IEEE Transactions on Vehicular Technology*, 69(1): 140-150.

Vidyapeetham A., V. 2023. Flight Delay Prediction Using Machine Learning Models. IEEE 8th International Conference for Convergence in Technology, April 07-09, Lonavla, India.

Tang, Y. 2021. Airline Flight Delay Prediction Using Machine Learning Models. Proceedings of the 5th International Conference on E-Business and Internet, October 15-17, Singapore, Singapore.

Kim, Y., Choi, S., Briceno, S., and Mavris, D. 2016. A Deep Learning Approach to Flight Delay Prediction. IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pp. 1-6, September 25-30, Sacramento, CA, USA.

Giusti, L., et al. 2019. Analyzing Flight Delay Prediction Under Concept Drift. *Evolving Systems* 13: 723–736.

Cai, K., Li, Y., Fang, Y. and Zhu, Y. 2021. A Deep Learning Approach for Flight Delay Prediction Through Time-Evolving Graphs. *IEEE Transactions on Intelligent Transportation Systems*, 23(8): 11397-11407.

Choi, Y., et al. 2016. Prediction of Weather-Induced Airline Delays Based on Machine Learning Algorithms. *Journal of Air Transport Management*, 55: 15-24.

Bisandu, D. B., & Moulitsas, I. 2023. A Deep BiLSTM Machine Learning Method for Flight Delay Prediction Classification. *Journal of Aviation/Aerospace Education & Research*, 32(2): 1-32.

Mtimkulu, Z., and Maphosa, M. 2023. Flight Delay Prediction Using Machine Learning: A Comparative Study of Ensemble Techniques. International Conference on Artificial Intelligence and its Applications, November 9-10, Mauritius.

Guo, Z., Yu, B., Hao, M., Wang, W., Jiang, Y., & Zong, F. 2021. A Novel Hybrid Method for Flight Departure Delay Prediction Using Random Forest Regression and Maximal Information Coefficient. *Journal of Aerospace Information Systems*, 116, 106822.

Khan, R., Akbar, S., & Zahid, T. 2022. Flight Delay Prediction Based on Gradient Boosting Ensemble Models. *EasyChair Preprint* № 8511.

Mamdouh, M., Ezzat, M., & Hefny, H. A. 2023. A Novel Intelligent Approach for Flight Delay Prediction. *Journal of Big Data*, 10: 179.

Dou, X. 2020. Flight Arrival Delay Prediction and Analysis Using Ensemble Learning. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 836-840, June 12-14, Chongquing, China.

Cai, K., Li, Y., Fang, Y., Zhu, Y., Yang, Q., Yang, Y., & Du, W. 2023. A Geographical and Operational Deep Graph Convolutional Approach for Flight Delay Prediction. *Chinese Journal of Aeronautics*, 36(3): 357-367.

Bisandu, D. B., & Moulitsas, I. 2023. A Hybrid Ensemble Machine Learning Approach for Arrival Flight Delay Classification Prediction Using Voting Aggregation Technique. AIAA Aviation and Aeronautics Forum and Exposition, 12-16 June, San Diego, USA.

About the Bureau of Transportation Statistics. 2019. https://www.bts.gov/about-BTS [Accessed 01.05.2024].

Weather API Documentation. https://open-meteo.com/en/docs [Accessed 01.05.2024].

Airport and Airline Data. https://openflights.org/data.html [Accessed 01.05.2024].

Profillidis, V.A. and Botzoris, G.N. 2019. Chapter 5 - Statistical Methods for Transport Demand Modeling. In: Profillidis, V.A. and Botzoris, G.N. (Eds.), Modeling of Transport Demand, Elsevier, pp. 163-224.

Zhou, F., Fan, H., Liu, Y., Zhang, H., & Ji, R. 2023. Hybrid Model of Machine Learning Method and Empirical Method for Rate of Penetration Prediction Based on Data Similarity. *Applied Sciences*, 13(10): 5870.

Rudolph, M., Kurz, S., & Rakitsch, B. 2024. Hybrid Modeling Design Patterns. *Journal of Mathematics in Industry*, 14(3).

Das, D., Nayak, M., & Pani, S. K. 2019. Missing Value Imputation-A Review.*International Journal of Computer Sciences and Engineering*, 7(4):548-558.

El Abbadi, N. K. 2014. An Efficient Storage Format for Large Sparse Matrices Based on Quadtree. *International Journal of Computer Applications*, 105(13): 25-30.

Wang, S., Tang, J., and Huan, L. 2016. Feature Selection. In: Sammut C. and Webb G.I. (Eds.), Encyclopedia of Machine Learning and Data Mining, Springer, pp.1-9.

Peng, C. Y. J., Lee, K. L., and Ingersoll, G. M. 2002. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1): 3-14.

Niu, L. 2018. A Review of the Application of Logistic Regression in Educational Research: Common Issues, Implications, and Suggestions. *Educational Review*, 72(1): 41-67.

Song, Y. Y., and Lu, Y. 2015. Decision Tree Methods: Applications for Classification and Prediction. Shanghai Archives of Psychiatry, 27(2):130-135.

Yang, F.J. 2019. An Extended Idea about Decision Trees. International Conference on Computational Science and Computational Intelligence (CSCI), pp. 349-354, May 29-31, Honolulu.

Mahajan, P., Uddin, S., Hajati, F., and Moni, M. A. 2023. Ensemble Learning for Disease Prediction: A Review. *Healthcare*, 11(12):1808.

Branco, P., Torgo, L., and Ribeiro, R. P. 2016. A Survey of Predictive Modeling on Imbalanced Domains. *ACM Computing Surveys*, 49(2).

Hossin, M., and Sulaiman, M. N. 2015. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 5(2):1-11.

Sokolova, M., and Lapalme, G. 2009. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, 45(4): 427-437.

Bureau of Transportation Statistics. 2023. Understanding the Reporting of Causes of Flight Delays and Cancellations. https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations [Accessed 01.05.2024].

Flight Data from Bureau of Transportation Statistics. https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr [Accessed 01.05.2024].

Kolambe, M., & Arora, S. 2024. Forecasting the Future: A Comprehensive Review of Time Series Prediction Techniques. *Journal of Electrical Systems*, 20(2):575-586.

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., and Peters, B. 2024. The Receiver Operating Characteristic Curve Accurately Assesses Imbalanced Datasets. *Patterns*, 5(6): 100994.

CURRICULUM VITAE

Gülmira AHMETOĞLU

EDUCATION

Master of Science 2021-2024	Akdeniz University Institute of Natural and Applied Sciences Computer Engineering, Antalya
Bachelor of Science 2004-2009	Istanbul Technical University Faculty of Engineering, Computer Engineering, Istanbul

WORK EXPERIENCE

Expert Database Engineer 2023-Present	Turkish Airlines, Istanbul
Senior Database Engineer 2021-2023	Sunexpress Airlines, Antalya
Database Engineer 2013-2019	Vakıflar Bankası T.A.O., Istanbul
Technology Consultant 2009-2013	ORACLE Türkiye, Istanbul