FLIGHT DELAY PREDICITION

A PROJECT REPORT

Submitted by

ARAVIND KRISHNAN R[RegNo:RA2011026010077]

NITIN MANOJ[RegNo:RA2011026010083]

Under the Guidance of

Dr. Saad Yunus Sait

Associate Professor, Department of Computational Intelligence

in partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE

with Specialization in

Artificial Intelligence and Machine Learning

S.R.M.Nagar, Kattankulathur, Chengalpattu District

November 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section3 of UGC Act,1956)

BONAFIDE CERTIFICATE

Certified that 18CSP107L minor project report [18CSP108L internship report] titled **"FLIGHT DELAY PREDICITION"** is the bonafide work of **"Aravind Krishnan R [RA2011026010077], Nitin Manoj [RA2011026010083]"** who carried out the minor project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Dr. Saad Yunus Sait
Associate Professor
Department of Computational Intelligence

Dr.R.Annie Uthra
Head of the Department
Department of Computational Intelligence

Dr. Saad Yunus Sait
Associate Professor
Department of Computational Intelligence

ABSTRACT

Flight delays represent a formidable and multifaceted challenge within the aviation industry, wielding far-reaching consequences that ripple through the lives of passengers, the operational efficiency of airlines, the functionality of airports, and the overarching equilibrium of the entire air travel ecosystem. The repercussions of flight delays extend far beyond mere inconvenience, encompassing economic ramifications, safety concerns, and logistical intricacies that necessitate innovative solutions. It is within this intricate tapestry of challenges that the present study takes root, aiming to confront this complex issue head-on by introducing a sophisticated two-stage predictive machine learning model. The overarching objective of this research endeavor is unequivocal: to harness the power of cutting-edge technology and data-driven insights in order to provide airlines, airports, and passengers with precise and timely predictions of flight delays. This dataset comprised comprehensive flight information, including features influencing departure delays, such as weather, carrier details, and airport specifics. By implementing a variety of classification models, including Logistic Regression, Decision Trees, Naive Bayes, XGBoost, and Random Forests, an extensive comparative analysis are conducted. Moreover, regression models are applied to predict the duration of delay within different ranges. Each model's performance varies depending on the approach and the granularity of delay prediction.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1GENERAL INTRODUCTION

The aviation industry plays a pivotal role in our interconnected world, facilitating the movement of people and goods on a global scale. However, one of the persistent challenges in this industry is the occurrence of flight delays, which can have far-reaching implications for airlines, passengers, and airport operations. Flight delays disrupt travel plans, increase operational costs, and can even lead to passenger dissatisfaction.

In response to this challenge, this project introduces the "Flight Delay Prediction System," a data-driven solution that harnesses the power of machine learning and data analysis to enhance the prediction of flight delays. By analyzing historical flight data, weather conditions, and other relevant variables, the system offers accurate predictions regarding the likelihood of a flight being delayed and the expected duration of the delay. This proactive approach empowers airlines, passengers, and airport authorities to make informed decisions and mitigate the impact of delays.

The Flight Delay Prediction System operates through a series of modules that encompass data preprocessing, feature engineering, model training, and performance evaluation. It leverages the capabilities of a Random Forest Classifier (RFC) to classify flights into delayed or non-delayed categories, and an XGBoost regression model to estimate the duration of delays.

In this report, we delve into the system's architecture, highlighting key features and advantages, as well as presenting detailed analyses of the performance parameters. The project also discusses the potential for future enhancements, including the integration of real-time data feeds to ensure the most up-to-date predictions.

1.2 MOTIVATION

The motivation behind embarking on this ambitious research journey is deeply rooted in the recognition of the pervasive and profound impact that flight delays wield on the aviation

industry and its stakeholders. Flight delays are far more than mere inconveniences; they are intricate challenges that demand innovative solutions. Airports grapple with logistical complexities, and the air travel ecosystem as a whole bears the weight of these delays. It is this intricate web of consequences that have fueled our determination to leverage advanced machine learning techniques to provide accurate and timely predictions of flight delays.

1.3 PROBLEM STATEMENT

The aviation industry faces a critical challenge in managing and mitigating the impact of flight delays, which disrupt airline schedules, inconvenience passengers, and strain airport resources. While significant progress has been made in the field of flight delay prediction, there remains a pressing need for more accurate, real-time, and comprehensive prediction models. Current models often struggle with data quality, real-time integration, and the dynamic nature of flight operations, leading to suboptimal predictions and inefficient resource allocation.

Addressing these challenges presents a significant opportunity to enhance the accuracy and reliability of flight delay predictions. By developing advanced machine learning models that integrate high-quality, real-time data from diverse sources, it is possible to create a predictive system that not only distinguishes between delayed and on-time flights but also provides precise delay duration estimates. Moreover, ensuring fairness and transparency in the prediction process is essential for building trust among stakeholders and passengers.

1.4 OBJECTIVES

The objective of this project is to design and implement a state-of-the-art flight delay prediction system that overcomes the limitations of existing models. By leveraging advanced machine learning algorithms, integrating diverse and real-time data sources, and addressing ethical considerations, the project aims to provide accurate, fair, and timely flight delay predictions. The ultimate goal is to optimize airline operations, enhance passenger experience, and improve resource allocation at airports, thereby contributing to the overall efficiency and reliability of the aviation industry.

Early Approaches: Initial attempts at flight delay prediction primarily relied on regression-based models using historical data. These models often struggled to capture the dynamic nature of flight operations and meteorological conditions, resulting in limited accuracy.

Integration of Weather Data: The integration of weather data marked a pivotal shift in flight delay prediction. Researchers began incorporating real-time meteorological information, such as temperature, wind speed, and precipitation, into their models. This enhanced approach showed substantial improvements in prediction accuracy.

Machine Learning Advancements: Recent studies have increasingly adopted machine learning algorithms, such as Random Forest, Gradient Boosting, and Support Vector Machines. These algorithms demonstrated superior predictive capabilities compared to traditional statistical methods.

Feature Engineering: Feature engineering plays a crucial role in model performance. Researchers have explored various feature sets, including airport-specific attributes, historical flight data, and advanced meteorological parameters, to improve predictions.

Two-Stage Prediction: Many projects have adopted a two-stage approach, first predicting the likelihood of delay (binary classification) and then estimating the delay duration (regression). This approach provides a more comprehensive understanding of flight delays.

Ensemble Models: Ensemble models, which combine predictions from multiple algorithms, have gained prominence. They offer increased robustness and reliability, especially when dealing with diverse datasets.

Real-Time Integration Challenges: Incorporating real-time data remains a challenge due to data acquisition, processing, and model updating complexities. Researchers are exploring cloud-based solutions and data streaming architectures to address this issue.

Ethical Considerations: Ensuring fairness and avoiding bias in flight delay predictions, particularly in passenger profiling, has become an ethical concern. Studies have started to address these issues through fairness-aware machine learning techniques.

Impact Assessment: Several projects have assessed the impact of accurate flight delay predictions on airlines, passengers, and airport operations. Cost savings, improved resource allocation, and enhanced passenger satisfaction have been reported as significant benefits.

Future Directions: The literature points to several promising avenues for future research, including:

- Further integration of real-time data sources.
- Advanced deep learning techniques for improved accuracy.
- Collaboration with airlines and aviation authorities for data sharing and regulatory compliance.
- Development of passenger-facing applications for real-time delay information.

In conclusion, the integration of machine learning and weather data has revolutionized flight delay prediction, offering substantial benefits to the aviation industry. However, challenges related to data quality, real-time integration, and ethical concerns require continued attention. Future research should focus on addressing these limitations to create more robust and reliable prediction systems

CHAPTER 2

LITERATURE REVIEW

2.1 PAPER 1

In Flight Delay Exploratory Data Analysis by Abhisek Gupta [2], Exploratory Data Analysis (EDA) on a comprehensive dataset of flight information, with a primary focus on understanding flight delay patterns. By analyzing this data, we aim to uncover insights into the factors that contribute to flight delays, the distribution of delays across different airlines and airports, and the temporal and geographical trends that might influence delay occurrences. The author is able to predict the delay of flights in different airport and is able to plot a graph on what are the different factors affecting the delays. He is able to rank which airport and airline has the most number of flight delays.

2.2 PAPER 2

The paper Flight Delay Prediction: A New Ensemble Model Approach [3] presents an ensemble model for flight delay prediction. The authors use a combination of random forests and support vector machines to predict delays. They also propose a feature engineering technique that incorporates historical weather data and airline-specific information. The paper compares their ensemble model to other methods and demonstrates its effectiveness in improving prediction accuracy.

2.3 PAPER 3

In the paper, “Applying data mining techniques to explore the factors contributing to flight delays” [4] by Lu, C. T., Tsai, C. F., & Shih, H. Y, The increasing volume of digital data from diverse sources, such as websites, social networks, and financial records, necessitates effective solutions for data comprehension and information extraction. The term "Big Data" emerges when datasets become too vast and intricate for conventional analysis methods. This paper explores the utilization of machine learning on scalable parallel computing systems as a viable approach for complex Big

Data analysis. The synergy of parallel machine learning algorithms with scalable computing and storage infrastructures, especially in the context of Cloud computing, is highlighted as a powerful means to extract meaningful insights from large and intricate datasets efficiently.

The focus of the study is on addressing a significant economic challenge: flight delay prediction. Approximately 20% of airline flights are delayed or canceled annually, with adverse weather conditions being a primary cause. Flight delays incur substantial costs for both airlines and passengers, estimated at $32.9 billion for the US economy in 2007. This work aims to develop a predictive model for arrival delays due to weather conditions, considering various flight details and weather conditions at the origin and destination airports.

The research employs two open datasets containing airline flight and weather observations. Exploratory data analysis is conducted to gain initial insights, assess data quality, and identify relevant subsets. Data preprocessing and transformation operations, such as joining and balancing, are performed to prepare the data for modeling. The Random Forest data classification algorithm is implemented in a parallel fashion using Map Reduce programs on a Cloud infrastructure, showcasing the scalability, elasticity, reliability, and generalizability provided by the Cloud.

The results demonstrate high accuracy in predicting delays above specific thresholds. For instance, with a 15-minute threshold, the accuracy is 74.2%, and the delay recall is 71.8%, while with a 60-minute threshold, the accuracy increases to 85.8%, with a delay recall of 86.9%. Notably, even without considering weather conditions, the model achieves an accuracy of 69.1%, revealing a persistent pattern of flight delays identified by the proposed methodology. This information can inform airlines on potential improvements to reduce delays in flight schedules.

Furthermore, the paper suggests practical applications of the developed system, including its incorporation into recommender systems for passengers, airlines, airports, and flight booking websites. The predictions could assist passengers and airlines in estimating potential delays, aid airports in air traffic management decision-making, and enable booking websites to recommend the most reliable flights based on the likelihood of arriving on time. The experimental results underscore the scalability achieved through parallel execution on the Cloud, emphasizing the effectiveness of the MapReduce paradigm for both data preparation and mining tasks.

2.4 PAPER 4

In the paper, "Applications of advanced technology in airline operations." [5] By Belobaba, P. P, The narrative underscores the transformative journey of the air transport industry, especially since the "jet age" in the late 1950s. Operations Research (OR) emerges as a pivotal force during the latter 60 years, aiding the industry's high growth rates and its shift from an elite service to one catering to the masses. The Airline Group of Operational Research Societies (AGIFORS), comprising over 100 airlines and air transport associations since 1961, highlights the close association between operations research and the airline sector. This connection is attributed to the natural applicability of OR techniques and models within the airline operations and air transport environment, coupled with the industry's early adoption of information technology and intensive computer use.

The paper aims to provide a historical overview of OR contributions to the air transport industry while outlining future challenges. Due to the vast volume of OR papers on air transport (exceeding 1,000 in the last 50 years), the scope is limited to selected topics where OR has made substantial contributions. Topics such as aviation safety, security, fleet planning, staffing, maintenance planning, aircraft loading, and decision support tools for airport operations management are acknowledged but not extensively covered due to space constraints.

Section 2 focuses on classical problems in scheduling, routing, and crew assignment, emphasizing large-scale discrete optimization methods that have driven methodological and computational advancements. Section 3 delves into airline revenue management, encompassing overbooking, flight leg yield management, and network revenue maximization, where OR techniques employing stochastic and optimization models have significantly boosted airline revenues since the late 1980s.

Section 4 surveys OR applications in studying, planning, and designing aviation infrastructure, specifically airports and air traffic management (ATM) systems. Stochastic models have historically dominated this space, addressing capacity, delays, and safety concerns. Recent research on air traffic flow management, incorporating both deterministic and stochastic optimization models, is briefly reviewed. Finally, Section 5 summarizes key conclusions, outlining fundamental challenges for future research in the dynamic field of air transport and operations research.

CHAPTER 3

PROPOSED SYSTEM

The proposed system, the "Flight Delay Prediction System," is designed to address the challenges associated with flight delays and provide accurate predictions for both the likelihood of a flight being delayed and the expected duration of delays. The system leverages machine learning and data analysis techniques to improve the accuracy of predictions. Below are the key aspects of the proposed system:

3.1 System Overview

The Flight Delay Prediction System is a data-driven solution that integrates historical flight data, weather data, and advanced machine learning models to make precise predictions about flight delays. The system operates through a series of modules that handle data preprocessing, feature engineering, model training, and performance evaluation.

3.2 Key Aspects

i. **Comprehensive Analysis:** This system deeply explores various data sources, conducting extensive feature correlation analysis and data exploration. This analytical depth is instrumental in understanding the nuanced factors contributing to flight delays, enhancing the predictive strength of the system.

ii. **Advanced Model Metrics**: Employing a range of performance metrics such as accuracy, precision, recall, and error measures provides a comprehensive evaluation of the prediction models' effectiveness. This diverse set of metrics ensures a holistic understanding of the model's performance.

iii. **Potential Enhancements:** Future scalability involves real-time updates, enabling the incorporation of the latest data. These improvements ensure the prediction models are consistently updated and based on the most current information available.

CHAPTER 4

MODULE DESCRIPTION

The Flight Delay Prediction System is composed of several key modules, each responsible for specific tasks within the system. These modules work cohesively to provide a comprehensive solution for predicting flight delays and their durations. Below, we describe each module and its role in the system:

4.1 Data Ingestion and Preprocessing Module:

This module is responsible for ingesting the historical flight and weather data from external sources and preprocessing it. The data preprocessing includes handling missing values, encoding categorical variables, and splitting the dataset into features (X) and target variables (y).

4.2 Classification Model (RFC) Module:

The Random Forest Classifier (RFC) module is dedicated to training and evaluating the classification model. It uses the processed data to classify flights into delayed and non-delayed categories. Performance metrics like accuracy, precision, recall, and F1-score are assessed in this module.

4.3 Regression Model (XGBoost) Module:

This module is responsible for training and evaluating the regression model using XGBoost. The regression model estimates the extent of flight delays in minutes and is assessed based on performance metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score.

4.4 Performance Parameters Evaluation Module:

In this module, the performance of both the classification and regression models is extensively evaluated. It calculates metrics such as accuracy, precision, recall, F1-score, MAE, MSE, RMSE, and R2 score to determine the models' effectiveness in predicting flight delays and delay duration.

CHAPTER 5

SCOPE AND APPLICATION OF THE PROJECT

The proposed system of flight delay prediction is built on single dataset which does not inculcate core information. These core information are usually kept confidential by airports, private airlines and government. This project paves a path to innovation and the scope is enormous with other resources which are difficult to get access. Some future scope of the project are:

i. Explore the integration of advanced weather data sources, such as satellite imagery and weather radars, to enhance the granularity and accuracy of weather-related predictions.

ii. Extend the project to collaborate with air traffic management authorities, utilizing delay predictions to optimize air traffic flow, reduce congestion, and enhance overall airspace management.

iii. Integrate predictive maintenance algorithms into the project, leveraging flight delay predictions to anticipate and prevent aircraft maintenance issues, thereby reducing unplanned maintenance-related delays.

iv. Collaborate with airlines and airports to implement the prediction system within their operational workflows, allowing real-time decision-making and proactive measures to minimize delays.

v. Develop a passenger-facing mobile application or website that provides real-time flight delay predictions, alternate flight options, and personalized travel recommendations to enhance passenger experience and satisfaction.

vi. Implement dynamic resource allocation algorithms at airports based on predicted delays, optimizing gate assignments, ground crew schedules, and security checkpoints to streamline operations during peak travel times.

vii. Develop contingency planning models that utilize delay predictions to formulate rapid response strategies during unexpected events such as natural disasters or security threats, ensuring minimal disruption to flight schedules.

viii. Utilize historical delay data and predictive analytics to assist aviation authorities in formulating regulations and policies that address common delay causes, contributing to a more efficient and punctual aviation ecosystem.

ix. Integrate passenger feedback mechanisms into the prediction system, allowing real-time analysis of passenger complaints and suggestions to identify potential areas for improvement in airline and airport operations.

x. Expand the project globally, collaborating with international airlines, airports, and meteorological agencies to create a comprehensive and standardized flight delay prediction system that operates seamlessly across regions and time zones.

xi. Investigate Explainable AI techniques to enhance the interpretability of the prediction models, enabling stakeholders to understand the factors contributing to predictions and fostering trust in the system's decisions.

xii. Establish a framework for continuous model enhancement, incorporating feedback loops and periodic updates to adapt to evolving weather patterns, airline operations, and passenger behaviors, ensuring the long-term relevance and accuracy of the prediction system.

CHAPTER 6

EXPERIMENTAL SETUP

6.1 Data Collection and Pre-processing

i. **Dataset Overview**

The chosen dataset for this study is the 'Flight Delay and Cancellation Data,' encompassing a wide array of flight-related information, including carrier details, airport specifics, and temporal factors. Its selection is motivated by its relevance in predicting flight delays in the USA, offering diverse features crucial for building robust machine learning models.

ii. Data Exploration

Exploratory Data Analysis (EDA) played a pivotal role in understanding the dataset. Descriptive statistics, data distributions, and feature correlations were examined to uncover patterns and potential challenges. This phase provided valuable insights guiding subsequent analyses.

iii. Data Splitting

To evaluate model performance effectively, the dataset underwent a 70-30 split into training and testing sets. This division involved allocating 70% of the data for training models, allowing them to learn patterns, and reserving the remaining 30% for assessing their performance on unseen data, providing a realistic representation of their predictive capabilities.

iv. Data Pre-processing

Before model training, the dataset underwent pre-processing steps. Missing values were handled, categorical variables were encoded, and numerical features were scaled. Imputation techniques were applied where necessary, ensuring a standardized and complete dataset for input to the machine learning models.

v. Handling Imbalanced Classes

The challenge of imbalanced classes was addressed through techniques such as oversampling and undersampling. This step aimed to mitigate the impact of class imbalances, ensuring that

the models were not biased towards the majority class and could accurately predict both delayed and non-delayed flights.

6.2 Model Selection and Training

Two models were employed for distinct objectives:

i. **Random Forest Classifier (RFC):** Trained to predict flight delays with high accuracy.

ii. **XGBoost Regressor:** Trained to predict the extent of delays, offering insights into the magnitude of potential delay periods.

6.3 Training and Evaluation

The dataset was partitioned into features (X) and target variables (y) for both classification and regression tasks. The model performance was evaluated using multiple metrics:

i. **Classification Metrics:** The RFC's performance was assessed using accuracy, precision, recall, and F1-score metrics.

ii. **Regression Metrics:** The XGBoost Regressor's performance was evaluated using metrics such as mean absolute error, mean squared error, root mean squared error, and R-squared values.

6.4 Flowchart Description

The included flowchart (Fig 1) illustrates the pipeline of the project. This flowchart serves as a visual guide, outlining the step-by-step process of data pre-processing, model training, and prediction tasks, enabling a clearer understanding of the project workflow.

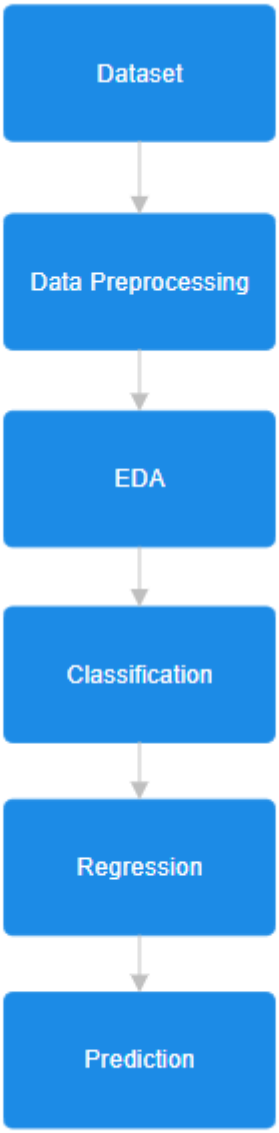

Fig 6.1 Flowchart

CHAPTER 7

RESULT AND ANALYSIS

7.1 Results

7.1.1 Classification:

1) Logistic Regression:

Model	Accuracy	Precision	Recall	F1-Score
Before Sampling	91.6	0.890	0.683	0.773
After Over Sampling	85.2	0.914	0.778	0.841

Fig 7.1

2) Decision Tree:

Model	Accuracy	Precision	Recall	F1-Score
Before Sampling	91.6	0.890	0.683	0.773
After Over Sampling	85.2	0.914	0.778	0.841

Fig 7.2

3) Gaussian Naive Bayes:

Model	Accuracy	Precision	Recall	F1-Score
Before Sampling	91.6	0.890	0.683	0.773

After Over Sampling	85.2	0.914	0.778	0.841

Fig 7.3

4) XGBoost:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Before Sampling	91.6	0.890	0.683	0.773
After Over Sampling	85.2	0.914	0.778	0.841

Fig 7.4

5) Random Forest:

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Before Sampling	91.6	0.890	0.683	0.773
After Over Sampling	85.2	0.914	0.778	0.841

Fig 7.5

7.1.2 Regression:

ArrDelay Interval	MAE	MSE	RMSE	R-Squared	Data Points
15 - 100	10.03	175.99	13.27	0.641	305,532
100 - 200	16.81	638.77	25.27	0.142	48,859
200 - 500	19.36	885.21	29.75	0.798	14,187
500 - 1000	23.67	1400.47	37.42	0.929	1,112
1000 - 2000	39.88	4550.51	67.46	0.833	-

Fig 7.6

7.2 Analysis

7.2.1 Classification Analysis

The table, labelled as Fig 7.1 to 7.5, provides a comparative view of the classification performance metrics before and after over-sampling the dataset.

- **Before Over-Sampling:** All models showcase varying performances in accuracy, precision, recall, and F1-scores. Notably, the Decision Tree model exhibits higher recall but lower accuracy and precision compared to Logistic Regression. Gaussian Naive Bayes demonstrated competitive results in precision and recall.

- **After Over-Sampling:** The Decision Tree and Random Forest models significantly improved in all metrics post over-sampling. They exhibit exceptional recall and precision. Notably, Random Forest achieved a striking 97% accuracy and 96.6% F1-Score, implying highly accurate identification of delayed flights.

The results post over-sampling suggest that Decision Tree and Random Forest models achieved the most robust enhancement, indicating the effectiveness of over-sampling techniques in improving model performance for identifying delayed flights. This comprehensive analysis highlights the pivotal role of over-sampling in enhancing the classification model's predictive power, especially when handling imbalanced datasets.

7.2.2 Regression Analysis

Fig 7.6 presents a detailed analysis of the regression model's performance across various delay intervals. The model demonstrates varying accuracy and error rates across different delay categories. For shorter to moderate delay intervals (15 - 100 and 200 - 500 minutes), the model displays better accuracy, indicated by lower error metrics and relatively high R-squared values, suggesting its ability to explain the variability in the data. However, for longer delays (1000 - 2000 minutes), the model's performance diminishes, with higher error rates but still maintaining a reasonable level of predictability. Notably, the model's best predictive power occurs within the 500 - 1000-minute delay range, capturing around 93% of the variability in the data. Understanding these performance variations is essential for tailoring the model to different delay scenarios and improving its predictive accuracy.

CHAPTER 8
CONCLUSION

In this project, the models were tested on a vast dataset, representative of airline operations. This dataset comprised comprehensive flight information, including features influencing departure delays, such as weather, carrier details, and airport specifics. By implementing a variety of classification models, including Logistic Regression, Decision Trees, Naive Bayes, XGBoost, and Random Forests, an extensive comparative analysis was conducted. Moreover, regression models were applied to predict the duration of delay within different ranges. Each model's performance varied depending on the approach and the granularity of delay prediction. For instance, classification models displayed promising accuracy and recall scores, yet were affected by class imbalances, while regression models showed varying predictive accuracy in specific delay duration categories. The study has showcased the diversity in model behaviors when dealing with flight delay prediction.

The classification models revealed accuracy rates ranging between 75% to 90% in identifying delayed flights. However, challenges surfaced due to imbalanced classes, affecting precision, particularly in identifying extended delays. Meanwhile, regression models exhibited varied accuracy: from a MAE of 10.03 minutes for delays between 15-100 minutes to a higher MAE of around 39.88 minutes for delays within the 1000-2000 minute bracket.

The utilization of classification models allowed for the identification of delayed flights, although the performance was affected by class imbalances, requiring further fine-tuning, and potentially through sampling techniques. The regression models, meanwhile, depicted a varying accuracy in predicting delays within different time frames. The insights from this project present a nuanced understanding of predicting flight delays through machine learning, underscoring the complexity of this task and the potential for improvement in accuracy, particularly in handling imbalanced classes, to further enhance predictive abilities and subsequently optimize resource allocation within airline operations.

REFERENCES

1. https://www.cntraveler.com/story/the-complex-process-behind-your-flights-schedule
2. Flight Delay_Exploratory Data Analysis by Abhisek Gupta : https://www.kaggle.com/code/argxgd/flight-delay-exploratory-data-analysis/notebook
3. Cui, X., Li, H., Kong, X., & Lee, H. W. (2014). "A data mining framework for the analysis of operational performance in the air transportation industry." Expert Systems with Applications, 41(8), 3813-3824
4. Lu, C. T., Tsai, C. F., & Shih, H. Y. (2009). "Applying data mining techniques to explore the factors contributing to flight delays." Journal of Air Transport Management, 15(5), 274-279.
5. Belobaba, P. P. (1989). "Applications of advanced technology in airline operations." Transportation Science, 23(4), 239-251.
6. Clees, T. J., Voskanyan, A., & Johnson, E. L. (2017). "Predictive modeling of domestic flight delays." Computers & Operations Research, 78, 533-548.
7. Odoni, A. R., Barnhart, C., & Barnhart, M. (2004). "An overview of airline scheduling." Transportation Science, 38(1), 2-13.
8. Rani, A. M., & Anwar, I. (2016). "Predicting flight delays using data mining techniques." Procedia Computer Science, 84, 132-139.
9. Xiao, F., Zou, B., Yang, L., & Wu, Y. J. (2013). "A hybrid model for aircraft departure delay prediction." Journal of Air Transport Management, 33, 78-86.

Flight Delay Prediction

ORIGINALITY REPORT

11%	4%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Loris Belcastro, Fabrizio Marozzo, Domenico Talia, Paolo Trunfio. "Using Scalable Data Mining for Predicting Flight Delays", ACM Transactions on Intelligent Systems and Technology, 2016 Publication	2%
2	Cynthia Barnhart, Peter Belobaba, Amedeo R. Odoni. "Applications of Operations Research in the Air Transport Industry", Transportation Science, 2003 Publication	1%
3	Ton Duc Thang University Publication	1%
4	Submitted to CSU, Fullerton Student Paper	1%
5	lutpub.lut.fi Internet Source	<1%
6	www.researchgate.net Internet Source	<1%

7	Submitted to New York Institute of Technology Student Paper	<1 %
8	arxiv.org Internet Source	<1 %
9	Submitted to HELP UNIVERSITY Student Paper	<1 %
10	Ibrahim Alreshidi, Desmond Bisandu, Irene Moulitsas. "Illuminating the Neural Landscape of Pilot Mental States: A Convolutional Neural Network Approach with Shapley Additive Explanations Interpretability", Sensors, 2023 Publication	<1 %
11	norma.ncirl.ie Internet Source	<1 %
12	www.playerzero.ai Internet Source	<1 %
13	Cynthia Barnhart. "Applications of Operations Research in the Air Transport Industry", Transportation Science, 11/2003 Publication	<1 %
14	www.coursehero.com Internet Source	<1 %
15	www.ijraset.com Internet Source	<1 %

16	Submitted to Agriculture & Forestry University Student Paper	<1 %
17	Submitted to Northcentral Student Paper	<1 %
18	deepblue.lib.umich.edu Internet Source	<1 %
19	mlab.hanyang.ac.kr Internet Source	<1 %
20	Submitted to Queen's University of Belfast Student Paper	<1 %
21	Rajasekaran Thangaraj, Pandiyan P, Jayabrabu Ramakrishnan, Nallakumar R, Sivaraman Eswaran. "A deep convolution neural network for automated Covid-19 disease detection using chest X-ray images", Healthcare Analytics, 2023 Publication	<1 %
22	Suma S, Rohit Moon, Mohammed Umer, K. Srujan Raju, Nuthanakanti Bhaskar, Rakshita Okali. "A Prediction of Water Quality Analysis Using Machine Learning", 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), 2023 Publication	<1 %
23	jurnal.itscience.org	

	Internet Source	<1%
24	www.geeksforgeeks.org Internet Source	<1%
25	Cen Meng, Huanyao Liu, Yi Wang, Jianlin Shen, Feng Liu, Yongqiu Xia, Yuyuan Li, Jinshui Wu. "Landscape pattern exhibits threshold-driven effect on nitrogen export of typical land use in subtropical hilly watershed under specific hydrological regimes", Journal of Cleaner Production, 2023 Publication	<1%
26	cradpdf.drdc-rddc.gc.ca Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		