MACHINE LEARNING-BASED FLIGHT DELAY PREDICTOR

MENAKA M, YUVAGANESH. G. K, RAKESH. S, SANJAI KUMAR. S

DEPARTMENT OF ELECTRONICS AND COMMUNICATION

VELTECH HIGHTECH DR. RANGARAJAN AND DR. SAKUNTHALA ENGINEERING COLLEGE
m.menaka@velhightech.com , Yuvaganesh090@gmail.com , raki8295@gmail.com , sanjaisago03@gmail.com

ABSTRACT:

Flight delays pose a significant challenge for the aviation industry, resulting in substantial financial implications each year. Prior studies have explored machine learning approaches to predict flight delays. However, relying solely on a single airport route may not suffice due to the dynamic nature of the aviation industry. Accurately estimating flight delays is crucial for airlines to enhance customer satisfaction and maximize revenue. This study proposes a novel Deep Learning (DL) model utilizing Support Vector Machine (SVM) for flight delay prediction. DL is a state-of-the-art technique capable of handling complex problems with large datasets and automatically extracting essential features. To address the presence of noisy flight delay data, a stack denoising autoencoder technique is incorporated into the proposed model. The results demonstrate that the proposed model achieves higher accuracy compared to the previous RNN model when forecasting flight delays for imbalanced and balanced datasets. The primary objective is to assess delays and analyse the underlying factors influencing them. The developed system leverages Support Vector Machine, Random Forest, and K-Nearest Neighbours (KNN) algorithms.

Flight delays- Deep Learning -blockchain based framework- Support Vector Machine (SVM)) - K-Nearest Neighbours (KNN) - departure delay

I INTRODUCTION

A flight delay occurs when an airline's departure or arrival time exceeds its scheduled time by more than 15 minutes. Causes of delays include adverse weather conditions, air traffic congestion, late arriving aircraft from previous flights, maintenance issues, and security concerns. These delays impose significant costs on airlines, disrupt scheduling and operations, and result in customer dissatisfaction. Predicting flight delays is crucial for mitigating losses and improving customer experience. However, accurate predictions are challenging, as neither customers nor airline ground staff receive advance delay forecasts based on various conditions. To address this issue, we can leverage statistical techniques like supervised machine learning and data mining, specifically regression analysis, to analyze current and historical data and make predictions about future delays. This analysis enables a detailed assessment of individual airlines, airports, and facilitates informed decision-making. Flight delays have wide-ranging negative impacts, including economic losses for commuters, airlines, and airport authorities. Additionally, delays contribute to environmental harm through increased fuel consumption and gas emissions. Therefore, accurate delay prediction across the diverse network of airline operations has become increasingly important. By combining live weather data with various metrics, we can calculate delays before departure, enhancing safety measures and improving decision-making processes for all stakeholders in the air transportation system. This comprehensive approach not only addresses customer concerns but also helps airlines avoid operational challenges and maintain a positive reputation.

II Related works

Rahul Nigam and Prof. Govinda K(2017)., In this project the focus is on logistic regression, a commonly used approach when the dependent variable is binary or dichotomous. This means that the dependent variable can only have two possible values, such as "Yes or No," "Default or No Default," "Living or Dead," "Responder or Non-Responder," or "Yes or No." The independent factors or variables can be either categorical or numerical. While acknowledging that the model may not be perfect, it still offers valuable insights and predictions regarding the likelihood of flight delays.

Yong Tian(2020), the author discusses the significant impact of flight delays on various aspects of air travel, including passenger travel plans, airline fuel plans, timetables, and air traffic control. The paper introduces the concept of a delay time window, which represents a specific time slot in a day determined by the actual off-block times of flights, along with minimum and maximum delay values. The proposed method utilizes historical flight operation data to establish delay time windows at departure airports and applies a flight block time reliability model to assess the reliability of each window. The case study focuses on Shanghai Hongqiao International Airport (ZSSS) and analyzes flight block time reliability under different delay time windows. The paper also presents a methodology that employs the DBSCAN algorithm to set up the delay time window based on departure and arrival delay times. Furthermore, the author defines a simple model for flight block time reliability and compares two analytical equations for estimating this reliability.

Priyanka Meel and Mukul Singhal(2020)., the authors discuss the importance of data preprocessing before applying algorithms to their dataset. Data preprocessing is necessary to convert the data into a suitable format for analysis and to improve data quality, considering that real-world data often contains missing values, noise, and inconsistencies. The authors obtained a dataset from the Bureau of Transportation for the year 2015, which consisted of 25 columns and 59,986 rows. The dataset had numerous rows with missing and null values. To address this, the authors used the dropna() function in the pandas library to remove rows and columns containing null values. After preprocessing, the dataset was reduced to 54,486 rows. The authors then applied various algorithms, including Logistic Regression, Decision Tree Regressor, Bayesian Ridge, Random Forest Regressor, and Gradient Boosting Regressor, to their preprocessed dataset. The performance of each model was evaluated using different metrics, and the results were compared and presented using bar graphs.

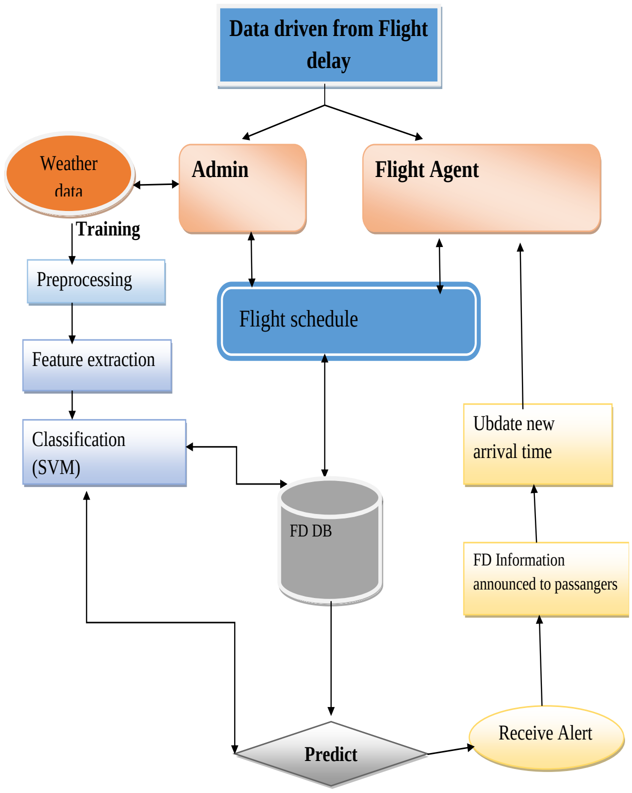
Qinggang Wu, Minghua Hu, and Xiaozhen Ma (2018)., Is based on Markov theory. This model demonstrates promising results when the flight normal rate is high, with an error controlled within 5%. In cases where the normal rate is moderate or extremely low, it is recommended to use this model for short-term forecasts, with an error tolerance of 10%. However, it should be noted that relying solely on the initial two hours' data to forecast the entire day's delay situation may not yield satisfactory results due to increased data fluctuations. This approach is more suitable for situations where data exhibits relatively large fluctuations. Nonetheless, this model holds practical value as it can serve as a forecasting tool for airport managers and passengers.

Engin Demir and Vahap Burhan Demir (2017)., They discussed various perspectives on identifying the factors contributing to flight delays in their 2017 article. In the past year, the Civil Aviation Administration of China (CAAC) has identified ten primary factors that can lead to flight delays, which include weather conditions, airline-related issues, air traffic control (ATC) factors, airport-related causes, joint inspection unit causes, fuel-related concerns, departure control system inefficiencies, passenger-related issues, air force-related factors, and public safety considerations. To address flight delays, the Air Traffic Management Bureau (ATMB) of the CAAC has implemented several measures in recent years. These measures can be broadly classified into three main categories: strengthening organization and leadership, upgrading operational and management practices, and enhancing supervision and inspection processes.

Jingyi Qu and Jinjie Zhang (2021)., In this project, a flight delay spread assessment method is proposed based on a flight chain model. Additionally, a visual analysis system for flight delay is designed. The system incorporates a large-screen interface that enables users to perform comprehensive analysis of flight delays. Through interactive features such as the time axis and map clicking, users can conduct comparative analysis of flight delays across multiple dimensions and regions. The visual analysis of flight delay facilitates a clearer understanding of the patterns present in the vast amounts of civil aviation big data for relevant departments. This, in turn, can contribute to the improvement of low-quality flights and the reduction of the propagation of flight delays. In the future, if data availability allows, the researchers plan to conduct data analysis specifically focused on domestic flight delay spread.

III Methodology

The project initiates with the collection of data from aviation companies, followed by data processing and feature extraction. Machine learning techniques are then applied to predict flight delays based on accuracy. The primary goal is to determine if a flight will experience delays, focusing on 13 key features including month, day, day of the week, flight number, origin airport, destination airport, scheduled departure, departure delay, taxi-out, distance, and scheduled arrival. To accommodate computational limitations, the models are trained and tested on laptops utilizing a smaller subset of 100,000 examples from a larger dataset of 5 million entries. The dataset is divided into training and testing sets, with an 80-20 ratio. The training set is used to train the models, while the testing set is employed to evaluate their accuracy and mitigate overfitting.



System Architecture of flight delay

In the experiment, a comparison is made between the differential evolution algorithm and the genetic algorithm. The results indicate that the differential evolution algorithm has a higher probability of achieving the optimal prediction model. Furthermore, the proposed model's performance is compared to that of the single factor prediction model and the relevance vector machine prediction model, demonstrating its superior effectiveness and accuracy in predicting flight delays.

Among the classifiers, Random Forest attains the highest accuracy of 77%. Additionally, an enhanced Support Vector Machine (SVM) model is implemented to predict flight departure delays, surpassing other algorithms in terms of effectiveness and accuracy.The learning process of neural networks revolves around optimizing network parameters, such as weights and biases, through back-propagation along the loss function. The classifiers yield a commendable accuracy of approximately 91%, with the decision tree classifier specifically possessing a tree depth of 7 and 127 leaf nodes, which represents a minimal fraction of the training examples.

IV Proposed System

The proposed system aims to provide airline passengers with real-time information about potential flight delays. To ensure scalability, it is important to select an algorithm that considers all relevant parameters as independent variables. Supervised learning involves training the machine using labeled data, where each data point is already tagged with the correct answer. By analyzing the labeled training data, the supervised learning algorithm generates accurate predictions.

In implementing the system, the completeness of the dataset was verified. While most of the dataset was complete, there were some missing values. For features such as arrival delay and departure delay, missing data could be calculated based on the scheduled and actual departure and arrival times. However, for features like tail number and flight number, the missing values were not calculable, resulting in the removal of examples with missing values from the dataset. Additionally, labels indicating arrival and departure delays were added to facilitate classification.

The dataset used for this project was obtained from The U.S. Department of Transportation (DOT), which provides a monthly report summarizing arrival delays, departure delays, on-time arrivals, and other relevant information. By utilizing this dataset, the system can predict potential departure delays for flight carriers, allowing for proactive measures to avoid delays and mitigate losses.

The project utilized a flights dataset with information from multiple carriers in the year 2015. A subset of the dataset was used to build the prediction model, while an airport dataset provided origin and destination information. The "DEPARTURE_DELAY" column played a crucial role in predicting departure delays, with negative values indicating early departures and positive values indicating delays. The "CANCELLED" column helped filter out rows associated with canceled flights. Unnecessary columns were removed, and the dataset was processed by handling numerical and categorical data separately, checking for null values, and applying one-hot encoding to categorical columns.

V Results and future scope

The training and test accuracies for the three classifiers were observed to be around 91%. The decision tree classifier had a tree depth of 7 and 127 leaf nodes, which accounted for less than 1% of the training examples.

By analyzing a 3D scatter plot, it was evident that the top-3 features identified by the decision tree classifier showed clear separation between on-time and delayed flights. This linear separability explained the high accuracy achieved by logistic regression and a simple single-layer neural network.

Currently, the dataset only includes flight and weather data from the United States. However, future plans involve expanding the project to include datasets from other international countries and domestic flights. Additionally, incorporating real-time weather data will enhance the accuracy and relevance of the results.

VI Conclusion

In this project, we successfully implemented machine learning algorithms to predict flight arrival delays. We found that simple classifiers like decision trees and logistic regression were effective in accurately determining if a flight would be delayed or not. To further enhance our models, we aim to increase the amount of training data and explore the use of deeper neural networks. Another aspect we plan to address is the prediction of taxi delays, considering factors such as airport runway and taxiway configurations, which have received limited research attention. Flight delays not only result in financial losses for the industry but also negatively impact airline reputation and reliability. Additionally, they contribute to sustainability issues by increasing fuel consumption and emissions. Our analysis not only predicts delays based on historical data but also provides statistical insights into airline performance, rankings, and time-related delay patterns, highlighting peak delay hours. This project can serve as a prototype for aviation authorities and be used for delay analysis using real datasets, including in the Indian context. We employ machine learning and deep learning algorithms, specifically Support Vector Machines, Random Forest, and K-Nearest Neighbors, and evaluate our system's precision, recall, and accuracy.

REFERENCES

[1] Guan Gui and Fan Liu, “Flight Delay Prediction Based on Aviation Big Data and Machine Learning “, IEEE 2020.

[2] N Lakshmi Kalyani and Bindu Sri Sai U , “Machine Learning Model – based Prediction of Flight Delay” , IEEE 2020.

[3] Jiage Huo and K.L. Keung , “The Prediction of Flight Delay: Big Data-driven Machine Learning Approach” , IEEE 2020.

[4] Vijayarangan Natarajan and Shubham Sinha, “A Novel Approach: Airline Delay Prediction Using Machine Learning”,IEEE 2018.

[5] Suvojit Manna and Sanket Biswas,” A statistical approach to predict flight delay using gradient boosted decision tree”, IEEE 2017.

[6] Tianyi Wang, Samira Pouyanfar, Haiman Tian, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen “A Framework for Airfare Price Prediction: A Machine Learning Approach”, IEEE 2020.

[7] Viet Hoang Vu, Quang Tran Minh, Phu H. Phung , “An Airfare Prediction Model to Developing Market”,IEEE 2020.

[8] K. Tziridis, Th. Kalampokas, G.A. Papakostas, K.I. Diamantaras “Airfare Prices Prediction Using Machine Learning Techniques”, IEEE 2020.

[9] Tao Liu, Jian Cao, Yudong Tan, Quanwu Xiao, “ACER: An Adaptive Context- Aware Ensemble Regression Model for Airfare Price Prediction”, IEEE 2017.

[10] Hao li, Yu Xiong, “Dynamic Pricing of Airline Tickets in Competitive Markets”, IEEE 2008.