Flight Delay Predictor

Machine Learning Project

Mariam ElGhobary
CSCE Department
The American University In Cairo
m.elghobary@aucegypt.edu
SID: 900211608

Youssef Elhagg
CSCE Department
The American University In Cairo
yousseframi@aucegypt.edu
SID: 900211812

INTRODUCTION AND MOTIVATION

In a time where global connectivity is continuously on the rise, it's only natural for the aviation field to follow. This, of course, opens the door for airlines worldwide to seize such an opportunity and increase their flights and revenues accordingly. However, it's not all roses as air-traffic congestion often surges as a direct consequence leading to an increase in duration and occasionally frequency of flight delays. The spectrum of issue intensity sparked by flight delays is anything but a narrow one. From an airline customer's perspective, consequences range from inconvenience due to late arrival to more serious ramifications such as missing connector flights. But the scale of repercussions exceeds those falling upon the individual, as significant economical and environmental drawbacks also arise.

PROBLEM SPECIFICATION

Based on what was discussed above, we can deduce that being one of the economic pillars of many countries, and with its growing scale of operation, flight scheduling may spark severe consequences as it becomes more error prone and embedded with greater uncertainty. This is primarily a result of the variety of flight delay causes with some of which being quite difficult to foresee. For example, delays caused by sudden weather fluctuations or emergency flight divergence etc are quite difficult to account for or quantify. Thus, since prevention is better than cure, one of the best and most cost efficient ways by which an airline and its customers can avoid flight delay ramifications is through delay-adapted flight scheduling. In other words, this is whereby sufficient leeway is created before, after, and/or between delay-prone flights accordingly as a precautionary measure to enable the resumption of normal operation despite unexpected crises. And in an age where technological advancements lead our world, what better tool can we utilize to predict possible flight delay periods than machine learning.

LITERATURE REVIEW

1. Supervised Learning Models:

Supervised learning models have been widely used in flight delay prediction. These models are trained using labeled data, i.e., flight records with known delay statuses. Once trained, they can predict the delay status of new, unseen flights.

For instance, a study by Yuemin Tang [1] at the University of Southern California implemented seven different supervised learning algorithms: Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Tree. The Decision Tree algorithm had the best performance with an accuracy of 0.9777, while the KNN algorithm had the worst performance with an f1-score of 0.8039.

2. Github Repositories:

Several GitHub repositories have been developed to predict flight delays using machine learning. These projects employ various machine learning techniques and feature engineering to enhance accuracy. For instance, the project by Tomer Yosef [2] focuses on predicting flight delays using machine learning techniques. They employ feature engineering and advanced regression algorithms to enhance accuracy. The dataset includes flight info, weather conditions, and other relevant factors. Their model achieves 94% accuracy. These repositories provide a wealth of resources for researchers and practitioners interested in flight delay prediction, offering code, datasets, and detailed explanations of the methods used.

3. **Ensemble Classifiers:**

Ensemble methods, which combine multiple machine learning models to obtain better predictive performance, have shown promise in flight delay prediction. The study by Yuemin Tang [1] found that tree-based ensemble classifiers generally have better performance over other base classifiers. This suggests that combining the predictions of multiple models could lead to more accurate flight delay predictions. Ensemble methods are particularly useful when the individual models have complementary strengths and weaknesses, as they can leverage the strengths of each model to improve overall performance.

4. **Regression Models:**

Regression models have also been used to predict flight delays. These models are used to understand the relationship between the dependent variable (flight delay) and the independent variables (factors influencing the delay). Regression models can provide insights into the factors influencing air travel disruptions and enhance predictions. For instance, the project by Khushi Bhadange [3] explores predictive modeling with regression-based models applied to a comprehensive dataset on flight delays and cancellations. Regression models are particularly useful for understanding the relationships between variables, as they can provide insights into the factors that influence flight delays.

5. **Big Data-Driven Approach:**

With the increasing amount of data in the aviation industry, big data-driven approaches have been proposed for flight delay prediction. For instance, a study by Jiage Huo [4] and K.L. Keung presented a big data-driven machine learning approach for flight delay prediction. This approach could be beneficial in handling the large amount of data involved in flight delay prediction. Big data-driven approaches can leverage the power of modern computing infrastructure to process large amounts of data quickly and efficiently, making them particularly suitable for applications like flight delay prediction that involve large datasets.

Each of these approaches has its strengths and weaknesses, and the choice of model depends on the specific requirements of the problem, the available data, and the desired trade-off between model complexity and predictive accuracy.

APPROACH TO SOLVING THE PROBLEM

The first step in solving the problem of flight delay prediction is for us to understand the factors that contribute to flight delays. These can include weather conditions, air traffic, mechanical issues, and many others. Once we have identified these factors, we can collect data on them for use in our machine learning model.

Next, we would need to preprocess this data. This could involve cleaning the data, dealing with missing values, and transforming the data into a format that can be used by a machine learning algorithm. We might also need to perform feature engineering, which involves creating new features from the existing data that might be more informative for the task of predicting flight delays.

Once our data is prepared, we can train a machine learning model on it. There are many different types of models we could use, including supervised learning models, ensemble classifiers, regression models, and big data-driven approaches. The choice of model would depend on the specific requirements of the problem, the available data, and the desired trade-off between model complexity and predictive accuracy.

After training our model, we would need to evaluate its performance. This could involve splitting our data into a training set and a test set, training our model on the training set, and then evaluating its performance on the test set. We might also want to use cross-validation, which involves splitting our data into several subsets and training and testing our model on different combinations of these subsets.

Finally, once we are satisfied with our model's performance, we can deploy it. This could involve integrating it into a flight scheduling system, where it could provide real-time predictions of flight delays.

POSSIBLE APPLICATIONS

The primary application of flight delay prediction is in the aviation industry, where it can be used to improve flight scheduling and reduce the impact of delays. By accurately predicting flight delays, airlines can better plan their schedules, resulting in fewer delays and cancellations. This can lead to improved customer satisfaction and potentially significant cost savings for airlines.

In addition, flight delay prediction can also be useful for passengers. For example, a mobile app could use a flight delay prediction model to provide passengers with real-time updates on their flights. This could help passengers plan their travel more effectively and reduce the stress associated with flight delays.

Furthermore, flight delay prediction could also be used by logistics companies that rely on air transport. By predicting flight delays, these companies could better plan their logistics operations, resulting in improved efficiency and cost savings.

DATASET ANALYSIS

In order to develop an efficient and accurate machine learning model, the choice of an appropriate dataset is nothing short of crucial. And after thorough research for the ideal dataset, the breakdown of the most fitting candidates is as follows:

1. **US Domestic Flights Delay Prediction (2013 - 2018) [5]**

This dataset was put together as part of the AWS Academy's Machine Learning Foundations course and consists of 1.64 million instances and 20 features/columns. The main features of interest include the CRSDepTime which is the expected departure time, components of the flight date such as the day of the week, distance covered in miles as well as the reporting airline. Post-flight-arrival, the ArrDelay column calculates the signed difference between expected and actual arrival times. Based on that the boolean classifying label is_delay indicates whether or not the flight is delayed with 1 indicating delay and 0 indicating otherwise.

Strengths

- Data representative of the real world as it is extracted from the Bureau of Transportation Statistics
- Having a large number of samples correlates with an improved model as training error decreases as the dataset points increase.
- Some preprocessing has already been done such as breaking down the date column into its subcomponents and binning the distance covered into 250 mile intervals

Limitations

- The large number of sample points may make the running time of the model excessively large
- There is some apparent redundancy/repetition in the features. For example ArrDelay and ArrDelayMinutes both show the difference between expected and actual arrival times but one shows the signed difference accounting for early arrivals and the second disregards negative differences. Moreover the original date column in the date format remains alongside the preprocessed date components.
- Despite expected departure time being present, there is no column indicating actual departure time hence only arrival delay can be computed
- The Cancelled feature has 100% false values and so can also be deemed as redundant
- The margin for the is_delay column is not specified i.e. the minimum minutes required for the entry to be set to true remains unknown. We also are not sure whether or not departure or arrival delay is what is being determined

2. **Flight Delay Prediction [6]**

This is a subset of 2015 flight delay and cancellation data put together by DOT's Bureau of Transportation Statistics. It consists of 837K samples and 32 features/columns.
When it comes to features, similar features to the previous dataset are present such as scheduled and actual times of arrival, distance covered and airline. However, this dataset also has an actual departure time feature as well as a consequent departure delay outcome. There are also additional features breaking down causes of delays such as late aircraft delay, weather delay and so on.

Strengths

- The large but reasonable size of this dataset balances between the minimized training error and acceptable model running time
- The capability of predicting both the departure and arrival delays is more fitting for the purpose of our project

Limitations

- No data dictionary available and so the meaning and units of some features becomes ambiguous for example the unit for distance covered
- Taking a subset of a dataset increases potential skewing/biasing of data especially since we don't know what the available rows were selected based on
- This dataset also has a high null percentage for several delay features such as airline delay and security delay. However, the cancellation reason feature reached a staggering 98% null rows. Such high volumes of missing/nullified data can result in preprocessing difficulty.

3. **Flight Delay and Causes [7]**

This is the dataset we have chosen to build our model with. It is a combination of multiple datasets making up a net total of 485K sample points and 29 columns. Once again the standout and impactful features are mainly actual and expected arrival time, actual departure time, arrival and departure airports as well as distance between them in

miles. Moreover, this is the only dataset with both actual and expected elapsed time as well as multiple features with a specific breakdown of delay causes and so we are offered great flexibility in our categorization alongside start and endpoints over which we compute our sought delay. Hence this dataset is easier to customize to fit our project. The two main labels here would be arrival and departure delay.

Strengths

- Once again dataset size despite being the smallest out of the three sets but it remains large enough to build a reliable model with acceptable running time
- Most features directly impacting the fields we seek to predict
- No null value presence enabling easier and smoother preprocessing

Limitations

- May not be as reflective of the real world as it only covers a six month period
- Cancellation and Diverted columns have a 100% false rate and so are quite redundant.

REFERENCES

[1] Tang, Y. 2018. Flight Delay Prediction with Supervised Learning.University of Southern California. DOI: https://doi.org/10.1145/3497701.3497725

[2] Yosef, T. 2020. Flight Delay Prediction. GitHub repository. Available at: https://github.com/topics/flight-delay-prediction

[3] Bhadange, K. 2021. Flight Delay Prediction. GitHub repository. Available at: https://github.com/topics/flight-delay-prediction

[4] Huo, J., & Keung, K. L. 2019. A Big Data-Driven Machine Learning Approach for Flight Delay Prediction. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 3453-3462). IEEE. DOI:https://doi.org/10.1109/BigData47090.2019.9006358

[5] Luiz, G. 2018. US Domestic Flights Delay Prediction 2013-2018. Kaggle dataset. Available at: https://www.kaggle.com/datasets/gabrielluizone/us-domestic-flights-delay-prediction-2013-2018

[6] Sankar, S. 2021. Flight Delay Prediction. Kaggle dataset. Available at: https://www.kaggle.com/datasets/nueve1122/flight-delay-prediction

[7] Trivedi, P. 2022. Flight Delay and Causes. Kaggle dataset. Available at: https://www.kaggle.com/datasets/undersc0re/flight-delay-and-causes