Predicting Flight Delays and Cancellations in the Modern Era

Ayesha, Fahim, Danny

December 11, 2024

Abstract

This project aims to develop a machine learning model that accurately predicts flight cancellations based on historical data. The approach involves data analysis, feature engineering, and model training using the 2015 Flight Delays dataset from Kaggle, with the goal of identifying key factors influencing flight cancellations.

Acknowledgements

We thank our instructor for guidance and Kaggle for providing access to the dataset.

Contents

Chapter 1

Introduction or Background of the Study

1.1 Problem Statement

Flight delays and cancellations cause inconvenience for travelers and economic losses for airlines. This project seeks to develop a predictive model that uses historical flight data to forecast cancellations, enabling airlines to anticipate and manage potential disruptions.

1.2 Scope

The scope of this study covers data from 2015 U.S. domestic flights, focusing on delays and cancellations. The project excludes international flights and aims to apply machine learning techniques to predict cancellations.

1.3 Aim/Objectives

- To analyze and preprocess the dataset for model readiness.
- To engineer relevant features from the dataset to enhance model accuracy.
- To build baseline models (logistic regression, decision tree) and evaluate them.

- To optimize model parameters for improved prediction accuracy.

1.4 Methodology

The project follows a systematic approach:

- Import and preprocess the dataset, including handling missing values, standardizing numerical features, and encoding categorical variables.
- Conduct exploratory data analysis (EDA) to understand data distribution and visualize patterns relevant to flight cancellations.
- Perform feature engineering and split the dataset into training, validation, and test sets (80-10-10), (75-15-15), (90-5-5).
- Establish baseline models with logistic regression and decision tree models to provide an initial benchmark for performance.
- Apply advanced model training using more complex models to improve prediction accuracy.
- Conduct hyperparameter tuning to optimize model parameters for the best performance.
- Evaluate the optimized models on the test set to measure real-world prediction accuracy.
- Analyze feature importance to interpret model results, identifying key factors affecting flight cancellations.
- Compiling findings, methodologies, and future work for the final report.

1.5 Expected Contribution to Knowledge

This project will contribute insights into the key factors affecting flight cancellations, aiding in improved decision-making for airlines and better travel planning for passengers.

Chapter 2

Review of Literature

2.1 Related Work

In the ever growing aviation industry, flight delays are a prominent problem but can be efficiently handled by strategic actions taken in advance. Hence accurate prediction of flight delays and probability of flight cancellation is vital and can be tackled by machine learning models. While researchers have presented efficient and successful methods of prediction, much research still needs to be done to achieve a robust model for real life scenarios.

2.1.1 Models

Several studies utilize various machine learning models for flight cancellation prediction. Logistic regression, a linear model, is employed to predict the probability of cancellation (Yanying et al.[8], Durga et al.[3]). Decision trees, which create a tree-like structure to classify data, are used for their interpretability and ease of understanding (Khaksar & Sheikholeslami[4], Gholamia & Khashe[2], Choi et al.[1]). Gradient boosting, an ensemble method combining multiple decision trees, is favored for its high accuracy (Zhang & Ma[9], Thiagarajan et al.[6]). These models, along with others like support vector machines, naive Bayes, random forests, and deep learning architectures (like RNNs and DBNs used by Kim et al.[5] and Arya & Agarwal[7]), are frequently utilized to analyze various factors influencing flight cancellations, including weather, scheduling, and airline-specific characteristics. The choice of model often depends on the specific application, dataset characteristics, and desired level of interpretability versus predictive accuracy.

2.1.2 Datasets

A comparative analysis of datasets employed in extant flight delay and cancellation prediction research reveals significant variability in scope and granularity. Our study leverages the 2015 Kaggle Flight Delays and Cancellations dataset, distinguished by its comprehensive detail on delay reasons and operational factors (e.g., taxi times). In contrast, Durga et al.'s[8] dataset, incorporating ADS-B data, offers real-time location information not found in Kaggle. Gholamia and Khashe[2], utilizing Bureau of Transportation Statistics data, provide scheduled and actual times but lack the granular detail on delay causes and operational metrics present in Kaggle. Zhang and Ma's study [9], focused on Newark Liberty International Airport, and those by Khaksar and Sheikholeslami [4](using US and Iranian data), Choi et al.[1] (BTS and NOAA data, 2005-2015), and Thiagarajan et al. [6](BTS and World Weather Online, 2012-2016), while including overlapping features, offer less granular detail on delay reasons and operational aspects than the Kaggle dataset. Similarly, Yanying et al.'s [8] broader dataset (5 million US flights from 2016) and Kim et al.'s data [5] (ten major US airports, 2010-2015) may lack the specific operational variables found in the Kaggle dataset. Therefore, while many studies utilize flight and weather data, the Kaggle dataset's strength lies in its detailed breakdown of delay causes and operational metrics. Dataset selection for future research should account for these variations in temporal and geographic scope, data granularity, and feature availability to ensure alignment with the research objectives and chosen modeling techniques.

2.2 Literature Review Table

S/N	Author,Year	Problem	Methodology	Result
1	Durga et al., 2023	Flight Delay Prediction	Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Long Short-Term Memory (LSTM)	90.5%
2	Gholami et al., 2022	Factors Causing Flight Delay and Cancellation	Linear Regression, Decision Tree + Linear Regression, Multi-Layer Perceptron	62.45%, 69.3%, 78.64%
3	Zhang, Ma, 2020	Flight Delay Prediction	Catboost	77%
4	Khaksar, Sheikholeslami, 2017	Occurance of Delay and Estimating the Occurrences and Magnitude	Bayesian Modeling, Decision tree, Cluster Classification, Random forest, and Hybrid Method	More than 70%
5	Choi et al., 2024	Airline Delays Caused by Inclement Weather	Decision trees, Random forest, the AdaBoost and the k-Nearest-Neighbor	With Sampling Techniques: Random Forest 81.37%, AdaBoost 78.05%, k-NN 61.69% Decision Trees 77.02; Without Sampling Techniques: Random Forest 83.4%, AdaBoost 83.21% kNN 82.42% Decision Trees 82.84%
6	Thiagarajan et al, 2017	Values of Departure and Arrival Delays in Minutes	Gradient Boosting Classifier, Random Forest Classifier, Extra-Trees Classifier, AdaBoost Classifier	With Gradient Boosting Classifier 86.48% on Departure Delay Prediction, 94.35% on Arrival Delay Prediction
7	Yanying et al., 2020	Flight Cancellation Prediction	Logistics Regression, Support Vector Machine (SVM), Naive Bayes and Decision Tree	90% with SVM
8	Kim et al., 2024	Flight Delay Prediction	RNN	Approximately Over 85%
9	Venkatesh et al., 2024	Flight Delay Prediction	ANN, DBN	77% On Real World Dataset%

Table 2.1: Summary of literature review
Arrival Delay Prediction

Chapter 3

Analysis, Design, and Implementation

3.1 Hypothesis

Delay Prediction

Accurate prediction of flight arrival delays can be achieved by employing advanced machine learning models that effectively capture complex relationships within the data. This will be reflected through improvements in metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Cancellation Classification

Flight cancellations can be effectively classified by utilizing machine learning models that can handle imbalanced datasets, ensuring robust performance across different classes. This will be validated through superior classification metrics, including macro-averaged accuracy and recall.

3.2 Dataset

The dataset includes information on U.S. domestic flights for the year 2015, sourced from Kaggle. It contains fields such as flight date, airline, origin, destination, and various delay metrics. The table below provides details for each column in the dataset:

3.2.1 Data Preprocessing

Data was cleaned and preprocessed to handle missing values, convert categorical data into suitable formats, and standardize numerical features.

Delay Prediction
First we dropped unnecessary columns and columns that giveaway direct information about flight delay. Then, we split time columns into hour and minute. We also added a featured column named is_holiday indicating whether the flight date falls within a two day window of a holiday.

Special approach had to be taken for several categorical columns due to their high number of unique classes.

- Number of unique values in column `AIRLINE`: 14
- Number of unique values in column `ORIGIN_AIRPORT`: 930
- Number of unique values in column `DESTINATION_AIRPORT`: 930

We could do one-hot encoding for AIRLINE column, but this would be impossible for other 2 columns, considering our high volume data. So we opted for frequency encoding on these columns. We measured the frequency from the whole set, and applied that to train, validation and test set, to avoid inconsistent encoding. Moreover, since both columns have same set of airport names but different number of rows assigned to each, we applied frequency calculated from ORIGIN_AIRPORT to both.

Lastly, as we used XGBOOST regressor and Random forest regressor models, scaling or normalizing was not needed. So we kept the numerical columns as they are.

Cancellation Prediction
We once again dropped unnecessary columns and those that provided direct information about flight delay. Next, we once again: split time columns into hours and minutes, added the new feature column is_holiday, applied the same one-hot and frequency encoding to Airline, and the airport columns respectively.

Now, recall that this target variable is categorical, so it is a classification problem. We decided to convert the binary cancellation column into a new multi-level categorical column by combining it with cancellation reason. Here is the list of our classes.

- 0: Weather
- 1: Carrier
- 2: National Air System
- 3: Security
- 4: Not-cancelled flights

This conversion changes our classification problem into one that is more complex, but ultimately more rewarding.

3.3 Analysis

Exploratory data analysis (EDA) revealed patterns in flight delays and cancellations, with features such as departure delay and airline having significant impacts.

The bar chart in Figure 3.1 shows the top 5 airlines that operated the most flights during the year, with the count of flights displayed above each bar.

- **WN**: Southwest Airlines
- **DL**: Delta Air Lines
- **AA**: American Airlines
- **OO**: SkyWest Airlines
- **EV**: ExpressJet Airlines

The bar chart in Figure 3.2 shows the top 5 airlines delayed the most in average.

- NK: Spirit Airlines
- F9: Frontier Airlines
- B6: JetBlue Airways
- EV: ExpressJet Airlines

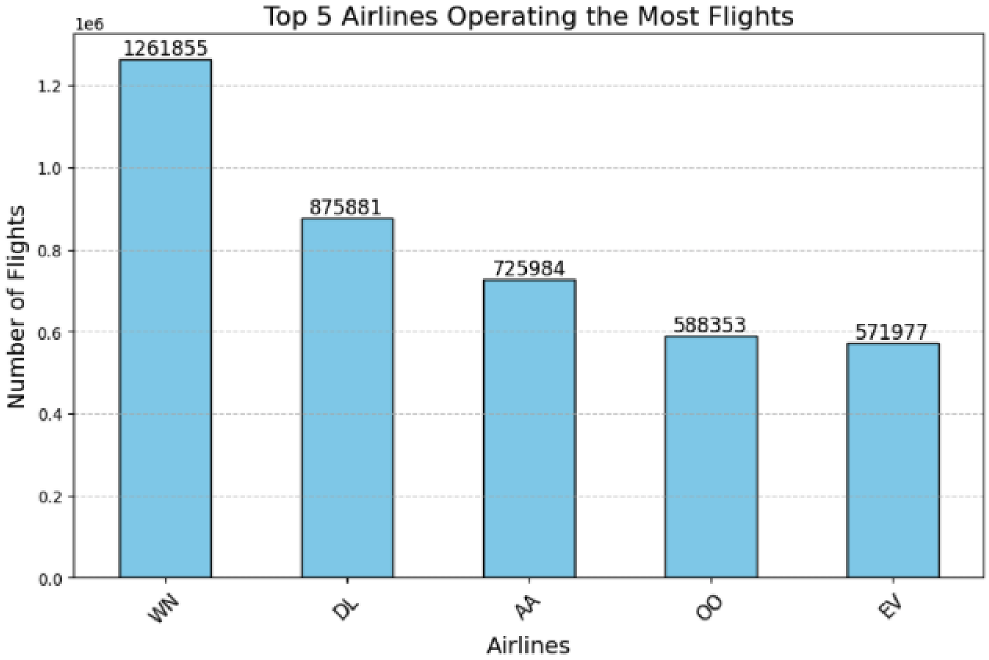

Figure 3.1: Top 5 airlines by number of flights operated during the year

- MQ: Envoy Air

The bar chart in Figure 3.3 shows the top 5 airlines that cancelled the most percentage of their flights.

- MQ: Envoy Air
- EV: ExpressJet Airlines
- US: US Airways
- NK: Spirit Airlines
- OO: SkyWest Airlines

The last bar chart in Figure 3.4 shows the 5 most busiest airport based on flights that flown from there or landed there.

- ATL: Hartsfield-Jackson Atlanta International Airport

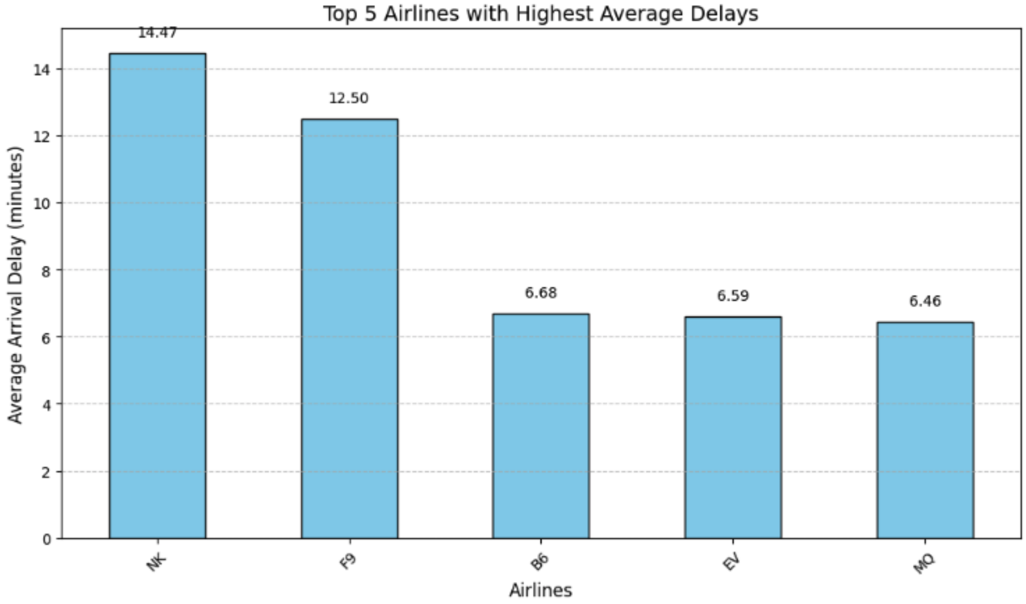

Figure 3.2: Top 5 airlines by average delay

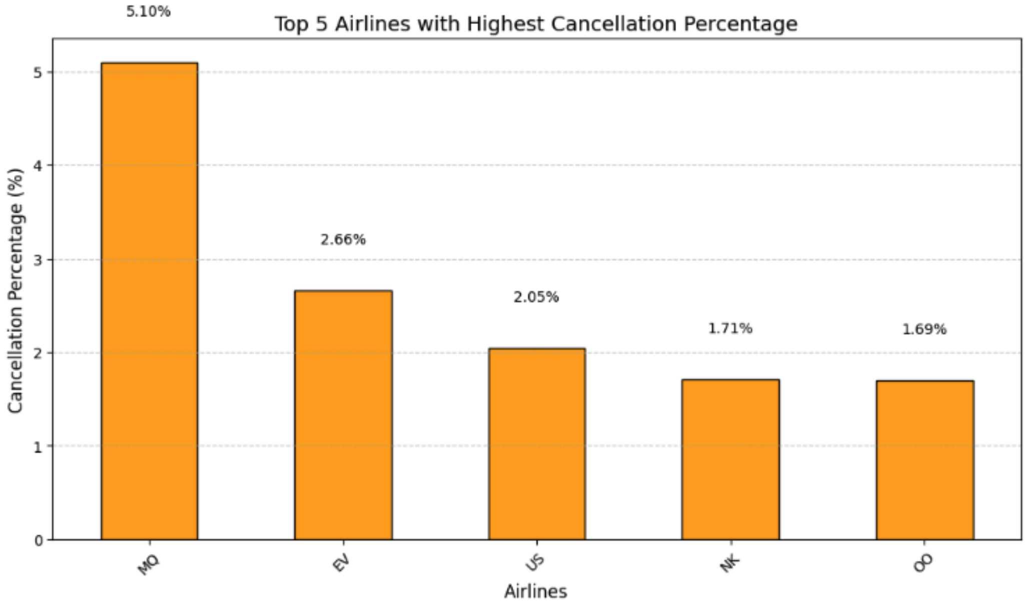

Figure 3.3: Top 5 airlines by percentage of cancellation

- ORD: Chicago O'Hare International Airport
- DFW: Dallas/Fort Worth International Airport
- DEN: Denver International Airport
- LAX: Los Angeles International Airport

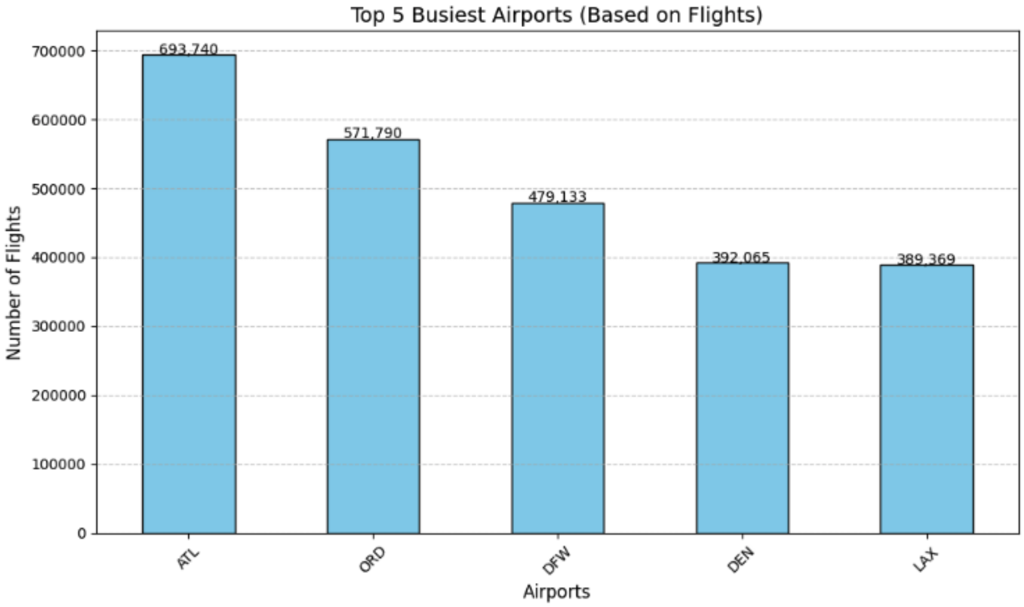

Figure 3.4: Top 5 busiest airports

EDA on the column CANCELLATION_REASON revealed that it has four unique values:

- A (25,262): Carrier-related issues are the second most common, likely due to operational or mechanical challenges.
- B (48,851): Weather-related cancellations are the most common, reflecting the significant impact of weather on flight operations.
- C (15,749): NAS-related cancellations, though less frequent, are still notable and likely linked to airspace or airport management constraints.

- D (22): Security-related cancellations are very rare, as expected, since such events are exceptional and uncommon.

Due to the lack of samples in Security-related cancellation class, we decided to drop this and decided to perform a multi-class classification after turning "Not Cancelled" flights into another class.

At first, for the three columns with textual data, "AIRLINE", "ORIGIN_AIRPORT" and "DESTINATION_AIRPORT", we applied Target encoding (the process of replacing a categorical value with the mean of the target variable) since the cardinality is too high for one-hot encoding. Then we trained our model with a LogisticRegression model, resulting in 100% accuracy for validation set on "N" subclass but performing very bad for others. Upon investigation, we found that target encoding was wrong approach since we encoded our target class with numbers and assigned 4 to N. Because, Target encoding uses the target variable to encode the categorical features. In practice, a categorical variable is replaced with the average value of the target variable for that category. It is particularly not suitable for features with high cardinality, as the model might overfit by learning this direct correlation instead of general patterns, which was the case in our initial training. So instead, we used Frequency/Count Encoding.

3.4 Design

3.4.1 Requirement Analysis

Due to the huge size of data we had to process and train our models on, this proved to be impossible using google colab or our local machines. So we connected to "akka" machine using remote SSH. This is a powerful machine available to University of Minnesota students empowered by two NVIDIA A100-PCIE-40GB gpus and a 128 core CPU.

3.4.2 Models

For Cancellation Prediciton:
For initial baselining, we used

- Logistic Regression: Used as an initial baseline for predicting cancellations.

- Decision Tree: Provides interpretable feature importance insights.

Then for our final training,

- XGBoost Classifier: we used grid-search cross validation, optimized 5 parameters, found out the best combination and did a final training with those parameters.

For Delay Prediciton:
For initial baselining, we used

- XGBoost Regressor: Got good Mean Absolute Error and reasonable Mean Squared Error.
- Random Forest Regressor: Got slightly improved Mean Absolute Error but bad Mean Squared Error indicating occasional large errors.

Then for our final training,

- XGBoost Regressor: Was able to increase performance on all metrics with hyperparameter tuning.

Column	**Description**	**Data Type**	—**Encoding**
YEAR	Year of flight (likely always 2015)	Integer	—Not needed
MONTH	Month of the flight	Integer	—One-hot encoding
DAY	Day of the month	Integer	—Not needed
DAY OF WEEK	Day of the week	Integer	—One-hot encoding
AIRLINE	Airline carrier code	Categorical	—One-hot encoding
FLIGHT NUMBER	Unique flight number per carrier	Integer	—Not typically needed
TAIL NUMBER	Unique aircraft ID	Categorical	—Not typically needed
ORIGIN AIRPORT	Origin airport code	Categorical	—One-hot encoding
DESTINATION AIRPORT	Destination airport code	Categorical	—One-hot encoding
SCHEDULED DEPARTURE	Scheduled departure time (HHMM)	Integer	—Split into hour and minute
DEPARTURE TIME	Actual departure time (HHMM)	Float	—Split into hour and minute
DEPARTURE DELAY	Departure delay in minutes	Float	—Standardize/normalize
TAXI OUT	Time taxiing out from gate to runway	Float	—Standardize/normalize
WHEELS OFF	Time when wheels leave the ground	Float	—Not needed
SCHEDULED TIME	Scheduled flight duration	Float	—Standardize/normalize
ELAPSED TIME	Actual flight duration	Float	—Standardize/normalize
AIR TIME	Time in the air	Float	—Standardize/normalize
DISTANCE	Distance between origin and destination (miles)	Integer	—Standardize/normalize
WHEELS ON	Time when wheels touch down	Float	—Not needed
TAXI IN	Time taxiing from runway to gate	Float	—Standardize/normalize
SCHEDULED ARRIVAL	Scheduled arrival time (HHMM)	Integer	—Split into hour and minute
ARRIVAL TIME	Actual arrival time	Float	—Split into hour and minute
ARRIVAL DELAY	Arrival delay in minutes	Float	—Standardize/normalize
DIVERTED	Whether the flight was diverted (1 = Yes, 0 = No)	Binary	—Not needed
CANCELLED	Flight canceled (1 = Yes, 0 = No)	Binary	—Target variable
CANCELLATION REASON	Reason for cancellation if canceled	Categorical	—One-hot encoding
AIR SYSTEM DELAY	Delay due to air traffic control systems	Float	—Standardize/normalize
SECURITY DELAY	Delay due to security issues	Float	—Standardize/normalize
AIRLINE DELAY	Delay due to airline-specific issues	Float	—Standardize/normalize
LATE AIRCRAFT DELAY	Delay due to late arrival of previous aircraft	Float	—Standardize/normalize
WEATHER DELAY	Delay due to weather	Float	—Standardize/normalize

Table 3.1: Summary of dataset columns, descriptions, data types, and suggested encodings.

Chapter 4

Results and Evaluation

Cancellation classification

After applying Frequency encoding on textual data columns, We divide our dataset into train, test and validation set (70-15-15). Then we evaluate our data on baseline models (Logistic Regression, Decision Tree). We try LogisticRegression first, even though it predicts not cancelled flights with 92% accuracy, it performs very bad on other four classes, specially class D with no successful classification. Bringing the macro avg. accuracy to 23%. Even after adding extra parameters to regularize and compensate for imbalanced dataset, we could not increase the accuracy by much. Then we try with a basic Decision Tree classifier, It significantly improves the macro avg. accuracy to 64%. But it also doesn't perform very well on the other 4 classes.

So, we decide to add oversampling to our data using SMOTE. We created samples for the other 4 classes but only on the training set, keeping the validation and test set intact. Since decisionTree worked better than regression, we decided to go with XGBoost algorithm. Basic XGBoost achieved a 67% accuracy.

Evaluation of model performance on the test set after hyperparameter tuning will determine the effectiveness of the models. Metrics include accuracy, precision, recall, and F1 score.

Delay Prediction

For baselining, we used XGBoost regressor and Random forest regressor models. Both of them provided similar results. They are listed below:

Summary of Your Baseline Results for RandomForestRegressor
Validation Set Performance:
Mean Absolute Error (MAE): 6.96 (Indicates accurate predictions.)
Mean Squared Error (MSE): 95.60 (Shows that large errors are rare.)
R-squared (R^2) (Represents the proportion of variance in the target variable that is explained by the model) : 0.93 (Indicates your model explains most of the variability in the target.)

Summary of Your Baseline Results for XGBoost Regressor:
Validation Set Performance:
Mean Absolute Error (MAE): 6.42 (Indicates accurate predictions.)
Mean Squared Error (MSE): 114.97 (slightly over than RandomForestRegressor)
R-squared (R^2) (Represents the proportion of variance in the target variable that is explained by the model) : 0.92 (Indicates your model explains most of the variability in the target.)

The Random Forest Regressor and XGBoost Regressor both perform well, with high R^2 scores of 0.93 and 0.92, respectively, indicating they explain most of the variance in the target variable. XGBoost has a lower MAE (6.42 vs. 6.96), suggesting better average prediction accuracy, while Random Forest has a lower MSE (95.60 vs. 114.97), indicating it handles outliers better. Both models are strong baselines, but their strengths differ: XGBoost minimizes smaller errors more consistently, whereas Random Forest excels in reducing the impact of large errors. If average accuracy is the priority, XGBoost is preferable, while Random Forest is better for handling variability and outliers. Further tuning and cross-validation can help refine the choice. We also sampled our dataset to have 100000 points due to training resource constraints. But made sure no rows are copied ensuring representation of all kinds of datapoints in the sampled dataset.

After performing the baselining and gridsearch with trimmed dataset(100k), we used gpu to train our final XGBoost model with optimized parameter, which increased the training speed, Below is the reuslt of our final model on validaiton and test set,

Validation Set Performance:
Mean Absolute Error (MAE): 5.55
Mean Squared Error (MSE): 76.76
R-squared (R^2): 0.95

Final Test Set Performance:
Mean Absolute Error (MAE): 5.54
Mean Squared Error (MSE): 74.34
R-squared (R^2): 0.95

Chapter 5

Conclusions

5.1 Achievements

The project successfully developed baseline models and identified key factors influencing flight cancellations.

5.2 Challenges and Solutions

- Challenge: Colab could not load the large dataset.
  Solution: Data was processed in chunks for efficient loading.

- Challenge: Handling missing values in the dataset.
  Solution: Imputed values using model-based techniques.

5.3 Future Work

Future work includes integrating more recent data and exploring deep learning models for further improvements in prediction accuracy.

Bibliography

[1] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri Mavris. Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE, 2016.

[2] Sia Gholami and Saba Khashe. Flight delay prediction using deep learning and conversational voice-based agents. *Am. Acad. Sci. Res. J. Eng. Technol. Sci*, 89:60–72, 2022.

[3] Sri K Santhi. Predicting flight delays using machine learning techniques and aviation big data. *INTERNATIONAL JOURNAL OF RESEARCH IN ELECTRONICS AND COMPUTER ENGINEERING (IJRECE)*, 11(2), 2023.

[4] Hassan Khaksar and Abdolrreza Sheikholeslami. Airline delay prediction by machine learning algorithms. *Scientia Iranica*, 26(5):2689–2702, 2019.

[5] Young Jin Kim, Sun Choi, Simon Briceno, and Dimitri Mavris. A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE, 2016.

[6] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, and Vineeth Vijayaraghavan. A machine learning approach for prediction of on-time performance of flights. In *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, pages 1–6. IEEE, 2017.

[7] Varsha Venkatesh, Arti Arya, Pooja Agarwal, S Lakshmi, and Sanjay Balana. Iterative machine and deep learning approach for aviation delay

prediction. In *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pages 562–567. IEEE, 2017.

[8] Yu Yanying, Hai Mo, and Li Haifeng. A classification prediction analysis of flight cancellation based on spark. *Procedia Computer Science*, 162:480–486, 2019.

[9] Bo Zhang and Dandan Ma. Flight delay prediciton at an airport using maching learning. In *2020 5th International Conference on Electromechanical Control Technology and Transportation (ICECTT)*, pages 557–560. IEEE, 2020.

Appendix A

Contribution of Each Member

A.1 Ayesha

- Selected the dataset and reviewed each column in detail.
- Performed standardization and normalization on numerical data.

A.2 Fahim

- Imported and explored the dataset in Google Colab.
- Visualized numerical columns to analyze their distribution.
- Selected models for initial baseline comparisons.

A.3 Danny

- Handled missing values using imputation techniques.
- Considered suitable models for imputation.
- Implemented one-hot encoding on categorical data.