Deciphering Airline Performance Data

Shraddha Ramprakash Gupta Radhika Ganesh Subhadra Mishra

project-gupta37-rganesh-sumishra

Abstract

Reliability and timeliness are crucial to the aviation sector. In addition to causing annoyance to travelers, flight delays and cancellations cause airlines to suffer significant financial losses. In order to determine causes, develop predictive models, optimize scheduling, evaluate airline performance, and enhance the overall customer experience, this research analyzes a dataset of airline delays and cancellations. Airlines may improve their operating efficiency, guarantee safety and compliance, and help travelers make better decisions by utilizing this dataset. The report also discusses the potential future uses of such data for emerging technologies and industry advancements.

Keywords

Machine learning, data science, prediction, classification, exploratory data analysis, data visualization, airline delays, flight cancellations, compliance, ANN.

1 Introduction

The aviation industry's relentless pursuit of efficiency and reliability is continuously challenged by flight delays and cancellations. These disruptions, caused by an array of factors such as weather, air traffic, and technical issues, not only inconvenience passengers but also impact airlines' operational and financial performance. Analyzing a comprehensive dataset of airline delays and cancellations is pivotal to understanding the root causes and consequences of these disruptions. This report focuses on the methods and approaches for analyzing such a dataset and explores the potential insights and applications that can arise from this analysis. By identifying the primary reasons behind delays and cancellations, airlines can refine their operations and optimize passenger satisfaction. Predictive modeling can enable proactive resource management and improved passenger notifications. Evaluating airline performance facilitates informed passenger choices, and the dataset can play a critical role in enhancing safety and regulatory compliance within the industry.

Previous work

The research papers that are cited below, collectively provide a comprehensive overview of flight delay prediction and management within the aviation industry. They explore various methodologies, including data mining, machine learning, and predictive modeling, to shed light on the factors influencing aviation disruptions. The papers emphasize the significance of using advanced technologies to enhance flight delay prediction and offer insights into the practical implementation of these methodologies for improving aviation data analysis. Together, they contribute valuable perspectives and approaches to address the challenges of flight delays in the aviation sector.

2 Methods

Dataset

We prepared the dataset by extracting information from the data 'Reporting Carrier On-Time Performance (1987 – present)' on the Bureau of Transportation Statistics (BTS) website, a division of the US Department of Transportation (DOT). After extracting the data for individual months from January 2022 to August 2023 from the BTS website, we merged the csv files to prepare the final dataset with 70,69,408 rows and 34 columns. There were several features available on the website, but we chose the ones that were most relevant for our project.

EDA

We pre-processed the dataset to ensure data quality and consistency. We performed some preliminary checks on the dataset such as checking the shape of the data, datatypes of columns, missing values, we performed Exploratory Data Analysis (EDA) on the data.The absence of outlier indices in the empty array (int64) is coupled with a t-test outcome (t-stat=0.9907, p-value=0.3331) indicating a lack of significant difference in delayed flight percentages between 2022 and 2023. This suggests that we lack adequate evidence to reject the null hypothesis, as the p-value (0.3331) surpasses the conventional significance threshold of 0.05. We studied the correlation between the different variables (Figure 1), and we found that the departure delay is highly correlated with arrival time, carrier delay, and late aircraft delay. The same goes for arrival delay.

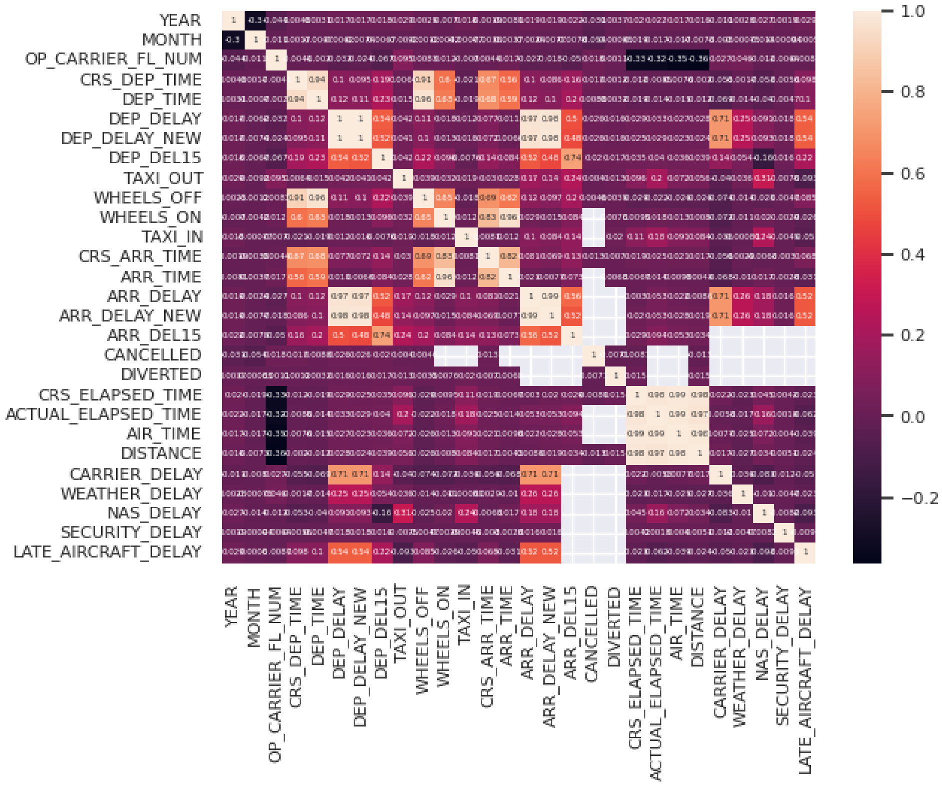

Figure 1: Correlation between the attributes

We chose ARRDEL15 column as our delay flag. Its value is 1 when arrival delay is more than 15 minutes and 0 otherwise. The line graph of month-wise percentage of delayed flights shows a spike in July 2023 which is in accordance with news articles stating that there were several flight delays due to the inclement weather conditions in several parts of the US. Similar line graph for month-wise percentage of cancelled flights shows highest percentage in January 2022, which can be attributed to the fact that there were many cancellations due to the spike in COVID-19 cases, linked to the Omicron variant (Figure 2).

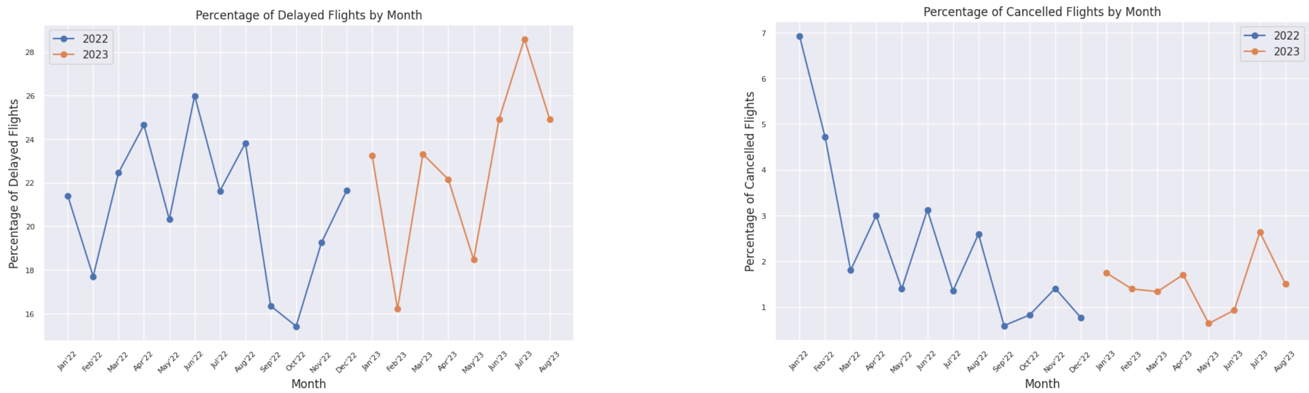

Figure 2: Delayed and Cancelled flights by month

We observed that there is not much variation in average flight delay percentages between January 2022 to August 2022 (22.24 percent) and January 2023 to August 2023 (22.72 percent). However, the average flight cancellation percentages between January 2022 to August 2022 and January 2023 to August 2023 show a drop from 3.11 percent to 1.5 percent. The arrival delay distribution of the overall dataset shows that in 21.6 percent cases, there was a delay of more than 15 minutes. The cancellation distribution of the overall dataset shows that only 2 percent of the flights were cancelled, which was expected (Figure 3).

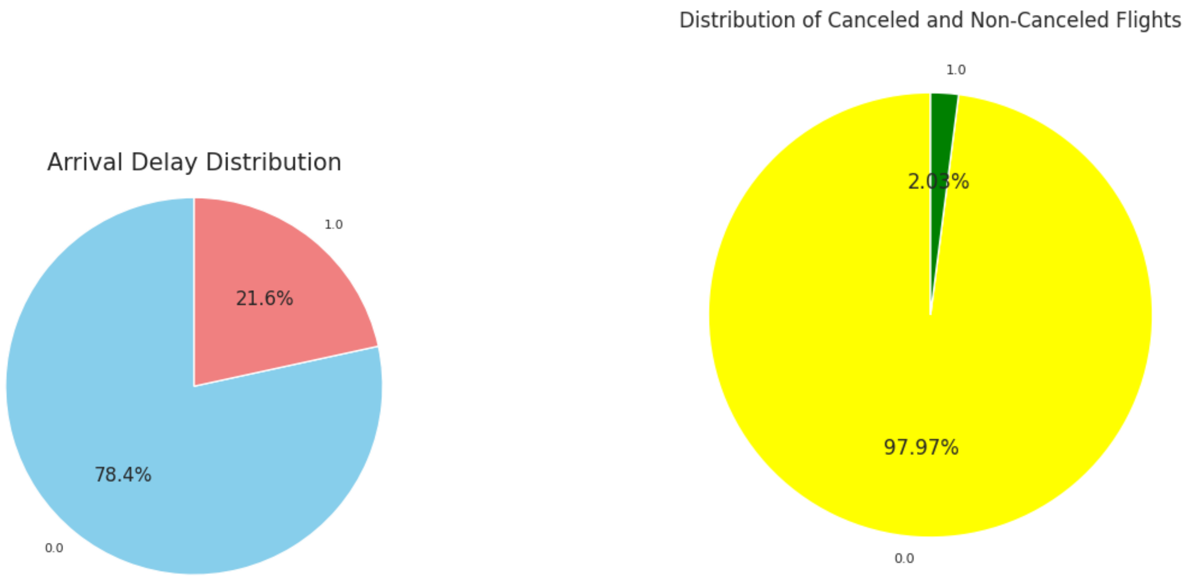

Figure 3: Delayed and Cancelled flights distribution

We replaced the carrier codes with their actual names prior to checking the carrier-wise distribution of delayed flights. We observed that approximately 34 percent of the flights operated by Frontier Airlines were delayed. Through similar analysis on the cancellation data, we observe that close to 3.75 percent of the flights operated by Republic Airways were cancelled(Figure 4).

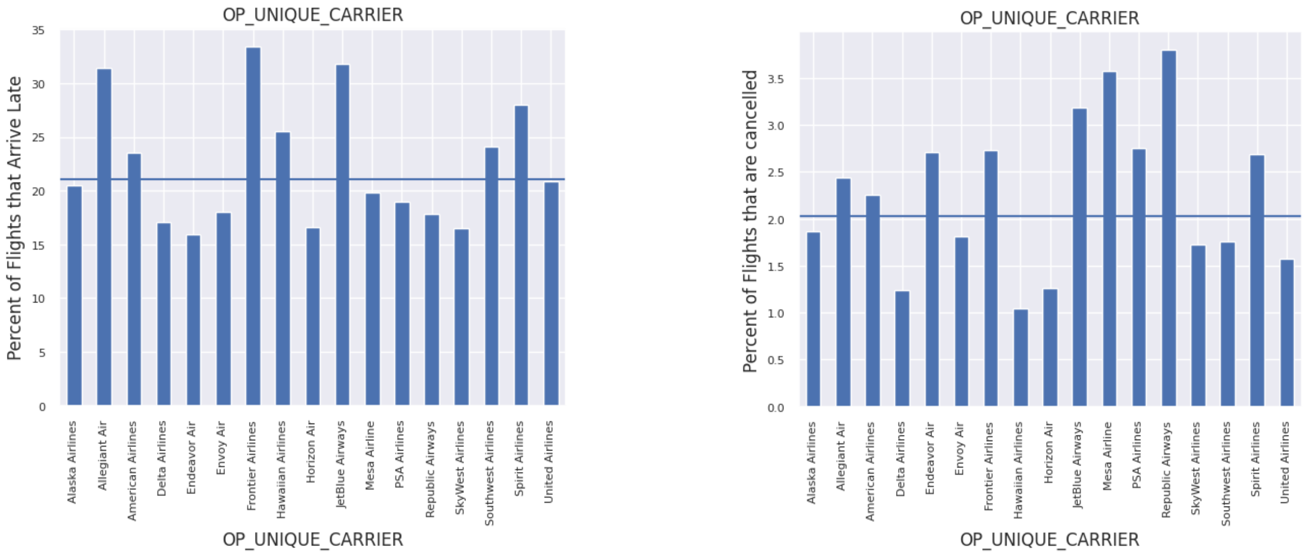

Figure 4: Carrier-wise distribution of Delayed and Cancelled flights

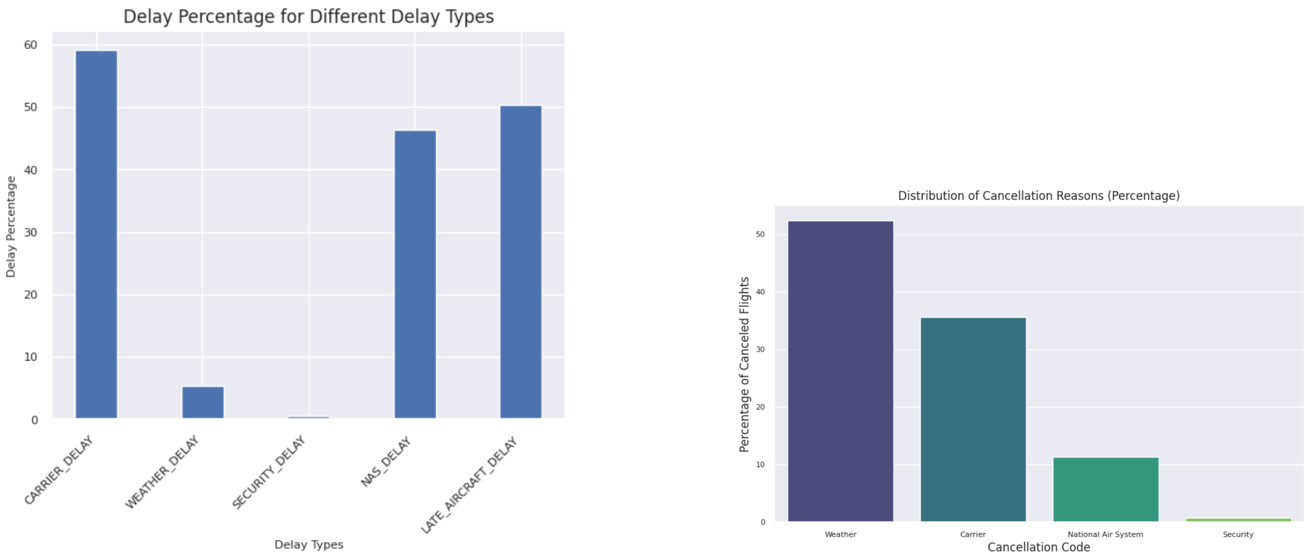

Figure 5: Delayed and Cancelled flights based on reasons

Further, we analyzed the data based on delay reasons and found that carrier delay is the top contributor followed by NAS delay (Figure 5). There are 5 different delay reasons in the dataset,
Carrier Delay: Delays caused by the specific airline's operational issues.
Weather Delay: Delays attributed to adverse weather conditions affecting flight operations.
Security Delay: Delays resulting from security-related issues or concerns.
NAS Delay: Delays caused by issues within the National Airspace System, such as congestion or air traffic control.
Late Aircraft Delay: Delays caused by the late arrival of the aircraft from a previous flight.
Similar analysis on the cancellation data showed that as expected, weather is the primary cause

for cancellation of flights. Also, for the delay reason columns we used fillna(0) as these columns represent delay durations, and a value is only present when there is a delay of that specific reason. Therefore, filling NaN values with 0 makes sense in this context. It implies that if there is no delay for a specific reason, the delay duration is considered zero minutes.

Classification Models

As the cancelled flights percentage is very low, we have focused more on building models for delayed flights.

Classification Models for prediction of delayed flights

We converted the categorical variables to numeric columns using the labelEncoder function, before building the classification models. ARRDEL15 was our target variable. We considered the test dataset size as 33 percent.
We implemented the following models for delayed flights:
Decision Tree
Logistic Regression
Naïve Bayes Classifier
Random Forest
Decision Tree Bagging
Logistic Regression Bagging
Adaboost
Histogram-based gradient boosting classifier
Artificial Neural Networks

Classification Models for prediction of cancelled flights

We implemented the following models for cancelled flights:
Histogram-based gradient boosting classifier
Artificial Neural Networks

3 Results

Classification Models for prediction of delayed flights

Model	Accuracy	Precision	Recall	F1 Score	Time period
Decision Tree	0.8843	0.7336	0.7283	0.731	Jan'22 to Aug'23
Logistic Regression	0.9244	0.8983	0.7328	0.8071	Jan'22 to Aug'23
Naïve Bayes	0.921	0.8529	0.7661	0.8072	Jan'22 to Aug'23
Random Forest	0.9215	0.8542	0.7207	0.7818	Jan'22 to Feb'22
Decision Tree Bagging	0.913	0.8553	0.7182	0.7808	Jan'22 to Aug'23
Logistic Regression Bagging	0.9244	0.8984	0.7327	0.8071	Jan'22 to Aug'23
Adaboost	0.8824	0.7349	0.712	0.7233	Jan'22 to Aug'23
Histogram based gradient boosting	0.9315	0.9309	0.9315	0.9289	Aug'23
ANN	0.931				Aug'23

Figure 6: Classification Models for prediction of delayed flights

The Decision Tree model exhibited high accuracy, yet there is room for improvement in precision, recall, and F1 score. Logistic regression demonstrated commendable performance, achieving a fine balance between accuracy, precision, and recall, resulting in a high F1 score, indicative of its effectiveness in handling both positive and negative instances. Naïve Bayes, with its elevated accuracy, showcased robust precision and recall values, leading to a high F1 score, signifying its adeptness in handling diverse instances. Due to RAM constraints, Random Forest was limited to a two-month dataset, yielding high accuracy but less balanced precision, recall, and F1 score compared to logistic regression and Naive Bayes. Decision Tree Bagging exhibited satisfactory accuracy but failed to surpass standalone logistic regression. Logistic Regression Bagging maintained the performance of the individual logistic regression model. Adaboost, with 50 estimators, achieved good accuracy but exhibited lower precision, recall, and F1 score compared to logistic regression and Naive Bayes models.

To summarize, considering the overall performance, the Logistic Regression model appears to be the best choice in this scenario. It has the highest accuracy and a good balance between precision and recall, as reflected in the F1 score. It seems to be effectively handling both positive and negative instances.

In addition to the above models, we also implemented a histogram based gradient boosting classifier and artificial neural network (ANN) on one month – August 2023. These are elaborated more in the discussion section.

Classification Models for prediction of cancelled flights

Model	Accuracy	Precision	Recall	F1 Score	Time period
Histogram based gradient boosting	0.999	0.999	0.999	0.999	Aug'23
ANN	0.9848				Aug'23

Figure 7: Classification Models for prediction of cancelled flights

For cancelled flights, we implemented a histogram based gradient boosting classifier and artificial neural network (ANN) on one month – August 2023. For ANN, the accuracy on the test set appeared high while in the case of histogram-based gradient boosting classifier (HistGradientBoostingClassifier), the best model obtained had high accuracy, precision, recall, and F1-score on both the training and test datasets.

4 Discussion

We implemented the histogram based gradient boosting classifier and artificial neural network (ANN) on August 2023 data for both delayed and cancelled flights. These are discussed in more detail below:

Delayed flights

The histogram-based gradient boosting classifier (HistGradientBoostingClassifier) was trained using grid search with cross-validation to find the best hyperparameters. The selected hyperparameters included a learning rate of 0.2, maximum depth of 7, minimum samples per leaf of 1, and a maximum iteration of 20. The provided evaluation metrics, including accuracy, precision, recall, and F1-score, indicated the performance of a HistGradientBoostingClassifier on both the training and test datasets. The classifier achieved high accuracy scores of approximately 93.1 percent for both training and test data. Precision, recall, and F1-score metrics provide a more nuanced evaluation, considering aspects of true positives, false positives, and false negatives. The similarity between the training and test metrics suggest generalization, and the high values across metrics indicate a well-performing model. However, the potential class imbalance, where there are more non-delayed flights compared to delayed flights, may have contributed to the high accuracy.

In the artificial neural network (ANN) implementation, a sequential model was created with three layers: an input layer with 64 neurons and ReLU activation, a hidden layer with 32 neurons and ReLU activation, and an output layer with one neuron using a sigmoid activation function. The model was compiled using the Adam optimizer and binary cross-entropy loss, with accuracy as the evaluation metric. During training over 10 epochs with a batch size of 32 and a validation split of 20 percent, the model exhibited consistent improvement on both the training and validation datasets, achieving an accuracy of approximately 93.0 percent on the test set. The decreasing trend in loss and increasing trend in accuracy throughout the epochs indicate effective learning and generalization. The evaluation on the test set yields an accuracy of 93.1 percent, indicating the model's ability to make accurate predictions on unseen data. The chosen architecture and training configuration seem to be suitable for the given binary classification task.

Cancelled flights

For the ANN, we oversampled the minority class using the RandomOverSampler from imbalanced-learn. The accuracy on the test set appeared high, likely influenced by the imbalanced nature of the dataset, where the majority class dominates, leading to a skewed accuracy metric.

In the case of histogram-based gradient boosting classifier (HistGradientBoostingClassifier), the best model obtained had high accuracy, precision, recall, and F1-score on both the training and test datasets. The imbalanced nature of the original dataset, where the majority class dominates, likely influenced these metrics. The oversampling technique was applied to balance the class distribution, resulting in improved model performance on the minority class. The tuned model demonstrated robustness in handling imbalanced datasets, achieving near-perfect performance metrics.

5 Author Contribution

At each stage of the project, we divided work equally amongst ourselves and worked together during team meetings. While preparing the data, we had to download the datasets month-wise which was a tedious task, so we each downloaded 6-7 months of data each and later appended them using Pandas. Similarly, during the EDA and analysis phase, we each developed the code for different graphs and trends like delay percentages, carrier-wise distribution, delay reason wise distribution and so on and later integrated everything into the final code. For the model building, we each ran the codes at our end for the different models like Decision tree, Naïve Bayes, ANN, etc. For the report part too, we collaborated and contributed to different parts of the report.

References

1. Sternberg, Alice Soares, Jorge Carvalho, Diego Ogasawara, Eduardo. (2017). *A Review on Flight Delay Prediction.*

2. Rebollo JJ, Balakrishnan H. *Characterization and prediction of air traffic delays.* Transportation Res Part C Emerg Technol. 2014;44:231–41.

3. C. Ariyawansa and A. Aponso. *Review on state of art data mining and machine learning techniques for intelligent Airport systems.* In Proceedings of 2016 International Conference on Information Management, ICIM2016, pages 134–138, 2016.

4. Kalliguddi, Anish Leboulluec, Aera. (2017). *Predictive Modeling of Aircraft Flight Delay.* Universal Journal of Management. 5. 485-491. 10.13189/ujm.2017.051003.

5. Somani, Saksham. (2021). *An Approach of Applying Machine Learning Model in Flight Delay Prediction- A Comparative Analysis*. Revista Gestão Inovação e Tecnologias. 11. 1233-1244. 10.47059/revistageintec.v11i3.2006.

6. Sahu, Prof Desale, Kunal Patil, Ashish Laishetty, Prithvi Patil, Bhuvaneshwar. (2023). *Flight Delay Prediction System in Machine Learning using Support Vector Machine Algorithm*. IJARCCE. 12. 10.17148/IJARCCE.2023.125188.

7. Yiu, Cho Yin Ng, Kam K.H. Kwok, Kin Lee, Wing Mo, Ho. (2021). *Flight delay predictions and the study of its causal factors using machine learning algorithms*. 10.1109/ICCASIT53235.2021.9633571.

8. https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data

9. https://www.transtats.bts.gov/Homepage.asp