FLIGHT DELAY/CANCELLATION PREDICTION

Gurleen Kaur[1] Kashmira Golatkar[2] Aditya Prajapati[3]

Abstract
A lot of companies rely on different airlines for connecting them with other regions of the world nowadays because the aviation industry is so important to the global transportation sector. However, problems like carrier delay, air traffic control issue, connecting flight delay, security concerns and overbooking can have a direct impact on airline services by causing delays in flights. In order to address this problem, precise forecasting of these aircraft delays helps airlines to respond to likely causes of the delays in advance to lessen the negative effects and lets passengers to be well prepared for the interruption caused to their journey. This project's goal is to examine the methods used to create models for forecasting aircraft delays.

Keywords
Time Series Data - Flights Delay/Cancellation - EDA - Machine Learning Algorithms - Delay Prediction - Deployment

[1] *Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1 **Problem and Data Description** 1
2 **Data Preprocessing & Exploratory Data Analysis** 2
2.1 Handling Missing Values 2
2.2 Exploratory Data Analysis 2
3 **Algorithm and Methodology** 3
4 **Experiments and Results** 4
5 **Deployment and Maintenance** 4
6 **Summary and Conclusions** 5
Acknowledgments 5
References 5

1. Problem and Data Description

The goal of this project is to predict flight delays and cancellations using historical data on flight schedules, cancellations, weather conditions, and other relevant factors. In this problem, we aim to identify patterns and relationships between the data to accurately predict if the flight will be delayed or not. The project will involve tasks like collecting, restructuring, and preprocessing data from various sources, performing feature selection and evaluating different machine learning models to predict delays. Flight Delay Prediction is an important problem for both travelers and airlines. Flight delays can cause increased operating costs, decreased revenue and potential loss of customers. The cost of flight delays to US airlines was estimated to be 2.6 billion in 2019, according to a study by masFlight. The total cost of flight delays to the US economy, including the impact on travelers and businesses, was estimated to be 32.9 billion in 2019, according to the US Travel Association. An accurate flight delay system will help airlines prepare for the forthcoming schedule change in advance and will be able to minimize the impact on their customers. For travelers, flight delays can cause a number of issues like missed connections, missed meetings/events and other inconveniences. A system that accurately informs about future delays will help travelers to modify their plans beforehand and avoid unnecessary hassle. Overall, Airline delay prediction is beneficial for both airlines and passengers in a number of ways. Flight delays cause a lot of inconveniences for passengers. Having faced the effects of flight delay when we missed our connecting flight as the previous flight was delayed, we understand the problems that come along with it. Missing connecting flights, not reaching their destination on time, and staying at an airport or any unknown place are a few of the many problems that travelers face along with the increased financial strain. A better flight delay prediction system will help passengers in multiple ways along with improving the sustainability of the aviation industry and improving the traveling experience of millions of customers worldwide.

Data: We have extracted the data from Federal Aviation Administration (FAA) Federal Aviation Administration (FAA), consisting of flight details with its arrival and departure locations, duration of the flight, date-time along with the weather forecast, as it might play an important role in this analysis. Similarly for the weather data we have extracted from Bureau of Transportation statistics. Our final data-set contains the following attributes namely - Flight number, departure and arrival airport details, scheduled departure time, actual departure time, schedule arrival time, actual arrival time, weather delay in total containing 5689512 rows × 61 columns.

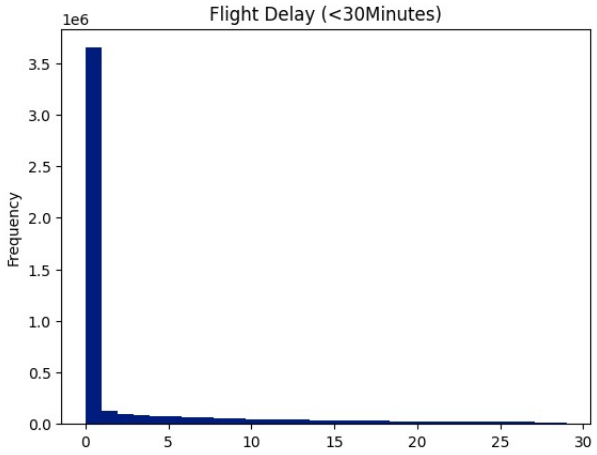
2. Data Preprocessing & Exploratory Data Analysis

2.1 Handling Missing Values

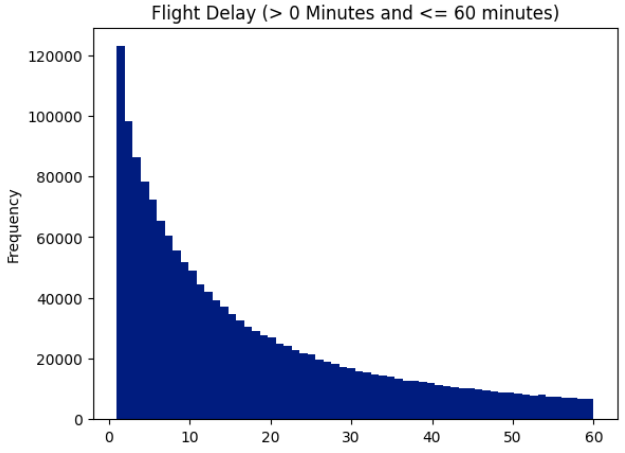
For handling the null values we need to first find the total number of null values in each attributes. The following are the attributes containing null values - DepTime, DepDelayMinutes, TaxiOut, TaxiIn, ArrTime and ArrDelayMinutes. Since there less number of null values in these columns, we can assume that the delay is 0. Hence substituted null values in DepTime, DepDelayMinutes, TaxiOut, TaxiIn, ArrTime and ArrDelayMinutes with 0.

2.2 Exploratory Data Analysis

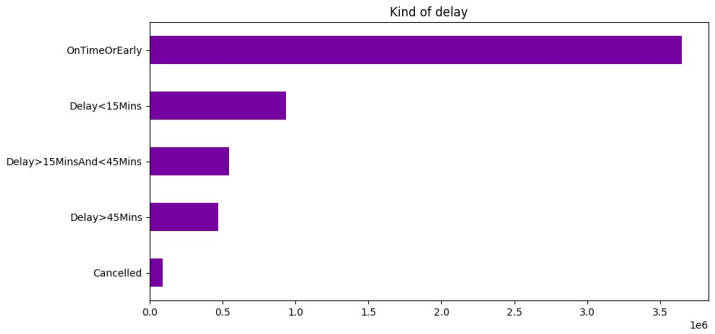
Exploratory data analysis (EDA) is a crucial initial step in studying data on aircraft delays ArrTime,and cancellations. EDA aids in comprehending the data's features, spotting any patterns or trends, and providing information for additional study.



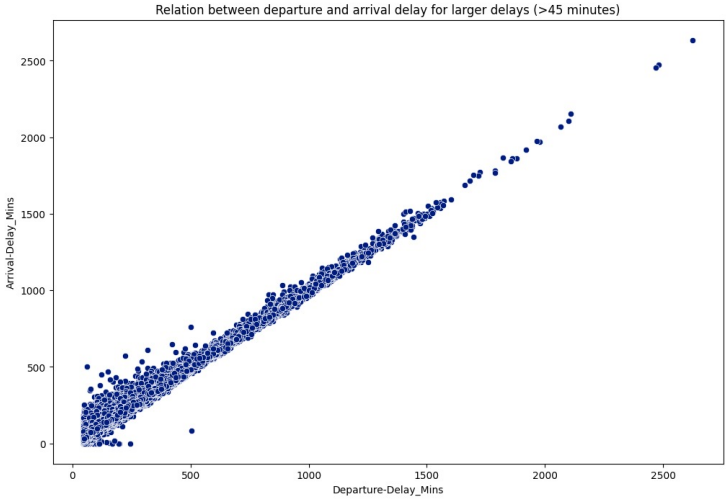
The above histogram shows the distribution of number of flights delayed (where delay was less than 30 minutes). We can see that the distribution is skewed towards left, meaning majority of the flights were not delayed (since delay time is 0).



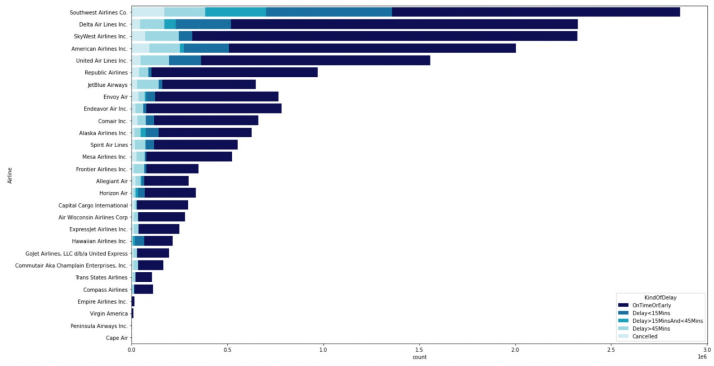
This histogram shows the distribution of number of flights delayed (where delay was greater than 0 minutes and less than or equal to 60 minutes.). We have kept the upper bound to 60 minutes to ignore the outliers with unusually large delay times. From the graph we can see that as the delay time increases the frequency of flights falling in that category also decreases implying that there are less number of flights that are delayed for larger amount of time.



This bar graph shows the comparison between different kinds of delay (based on the amount of time delayed). We can see that majority of flights are on time, and as the delay time increases the number of flights in that category decreases with the least number of flights in falling in the cancelled category.



We can see from this graph that arrival-delay and departure-delay are following a positive linear correlation, where if there is a delay in departure, it's highly likely that there's a delay in arrival.



This plot shows the distribution of different kinds of delays among various airlines. From the graph we can see that Southwestern airlines has the most delayed flights (in all

kinds of delays) and Cape Air has the least delayed flights.

KindOfDelay	OnTimeOrEarly	Delay<15Mins	Delay>15MinsAnd<45Mins	Delay>45Mins	Cancelled
Month					
1	68.169907	13.961832	7.878093	7.008976	2.981192
2	64.960611	15.300969	8.828322	7.786858	3.123240
3	66.363322	14.172305	7.716007	6.173113	5.575252
4	64.436275	14.051844	7.824258	6.853866	6.833756
5	65.023841	16.224252	8.965554	7.875154	1.911198
6	59.488265	17.349791	10.754222	10.442093	1.965629
7	61.996702	16.361900	10.052315	9.906083	1.683000
8	64.622096	15.009025	9.096512	9.137345	2.135022
9	72.924881	13.140776	6.725733	5.813787	1.394822
10	69.570264	14.859330	7.970955	6.487423	1.112028
11	70.182986	15.208750	7.826532	5.941371	0.840361
12	64.138827	16.782343	9.731652	7.916722	1.430456

From the given correlation matrix between months and delays (grouped by kinds of delays) we can observe that the month of June has the strongest correlation with the delayed flights followed by July and August. We can also observe that for the month of September has the strongest correlation with the on time flights followed by October and November. This implies that flights are more likely to get delayed in the months of June, July and August whereas flights are more likely to be on time in the months of September, October and November. Also we can observe that flights are most likely to get cancelled in April and March. Add subsections if needed.

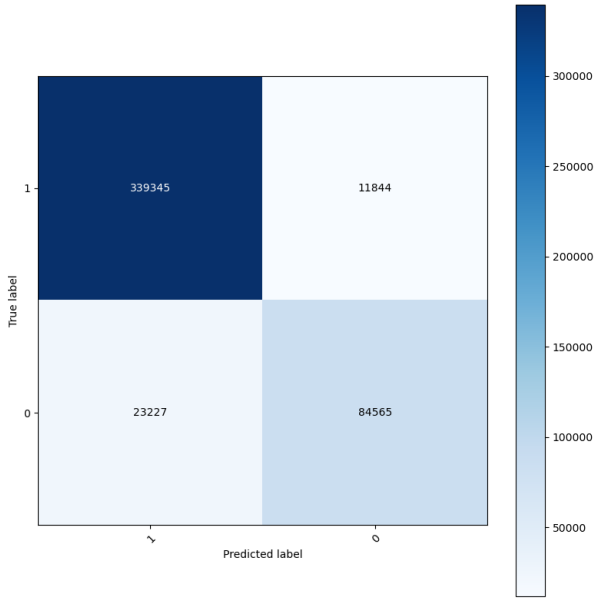
3. Algorithm and Methodology

Classification

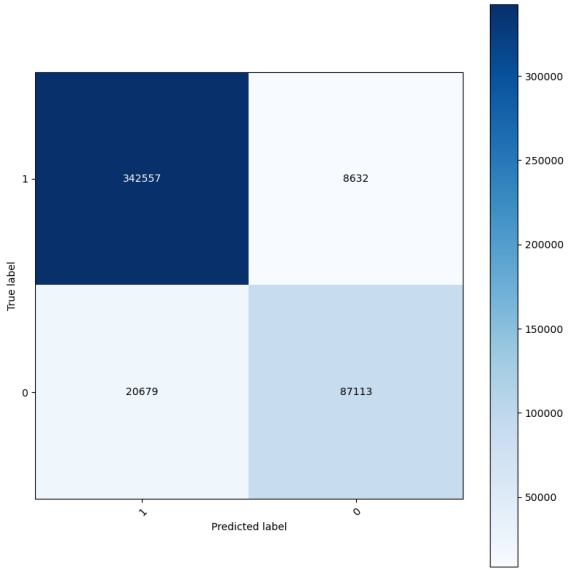
Logistic Regression : In Logistic Regression, the sigmoid function is employed to convert the output into a probability score, enabling the classification of the output. The hypothesis function for logistic regression is given in

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\beta^\top \mathbf{x}}} \tag{1}$$

The accuracy score for Logistic Regression is 0.71

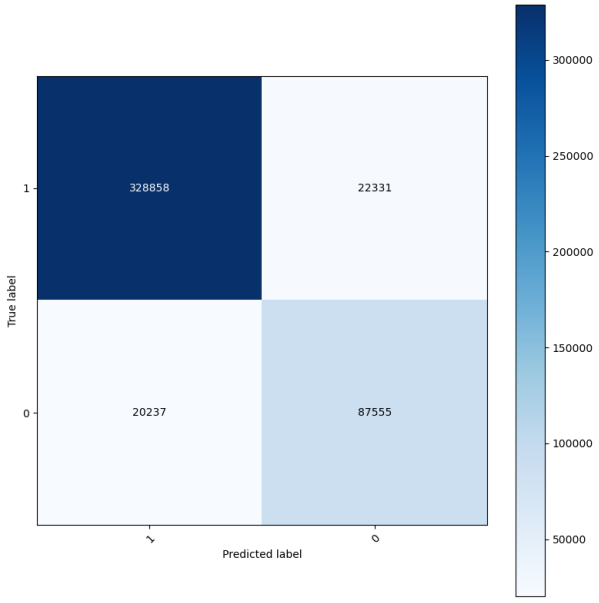


Linear SVC : Linear SVC is a supervised learning algorithm primarily used for binary classification problems. It is a variant of the Support Vector Machine (SVM) algorithm that employs a linear kernel function to separate the classes. The fundamental principle underlying Linear SVC is to identify a hyperplane that optimally separates the two classes within the feature space.

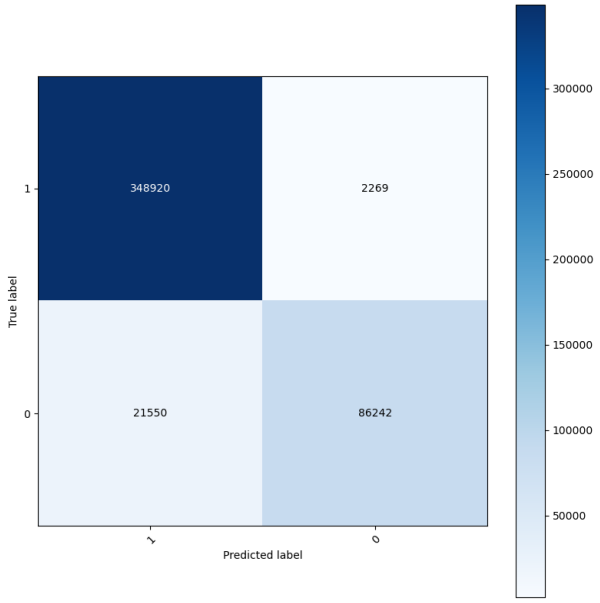


The hyperplane is represented as a linear equation in the form of: y = w0 + w1x1 + w2x2 + ... + wn*xn In this equation, y represents the prediction or output, while x1 through xn are the input features. The coefficients or weights assigned to each feature, w0 to wn, help determine the optimal placement of the hyperplane, which partitions the feature space into two distinct regions, one for each class. The accuracy score for logistics regression is 0.81

Decision Tree : We applied the default parameters to the Decision Tree model using the provided data, resulting in an accuracy of 0.806.



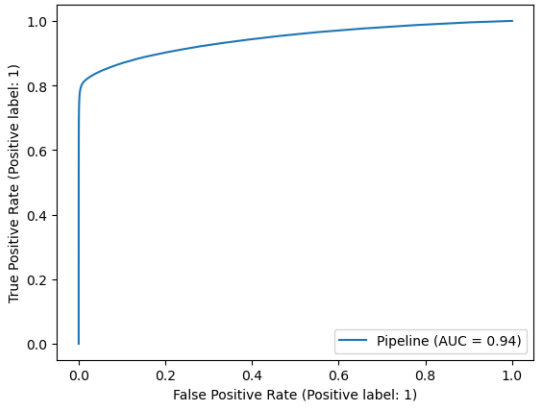
Random Forest Classifier: After initial testing, the Random Forest Classifier demonstrated a baseline accuracy of 0.873. To optimize its performance, we utilized Grid-Search CV to experiment with various values for Hyper-parameters such as class weight, criterion, and n estimators.



Methodology

In order to assess the effectiveness of our models, we divided our dataset into three subsets: training, validation, and testing, with a ratio of 70:10:20 respectively. We utilized several classifiers from the Sklearn library, including Random Forest, Linear SVC, Decision Tree, and Logistic Regression, to train our models on the training set. Additionally, we employed Grid Search Cross Validation to determine the optimal hyper-parameters for each model. To analyze and compare the performance of the models, we generated appropriate graphs and metrics.

4. Experiments and Results



Based on the Receiver Operating Curves shown in the figure above, we can observe that the Random Forest Classifier had the highest Area Under Curve (AUC) scores, indicating better overall performance than the other classifiers such as Linear SVC, Logistic Regression, and Decision Tree. To further evaluate the models, we can refer to the classification report table which presents a combined Confusion Matrix for the different classifiers along with Precision, Recall, and F1-score. This enables us to analyze the models based on various parameters and identify the optimal model for our problem statement.

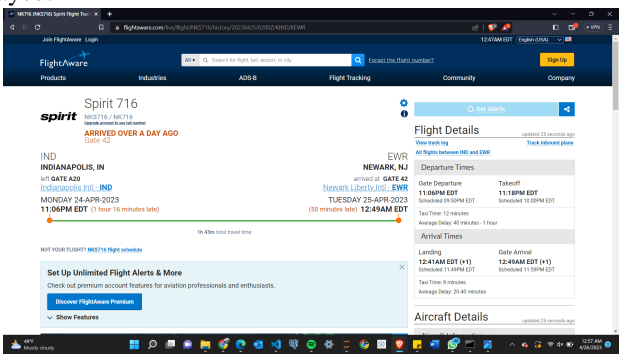
In our case, we need a model with high Recall value to accurately inform about flight delays so that concerned professionals and people can make adjustments accordingly. Therefore, we can conclude that Random Forest Classfifer with a Recall value of 0.75 is the best choice among the four classifiers.

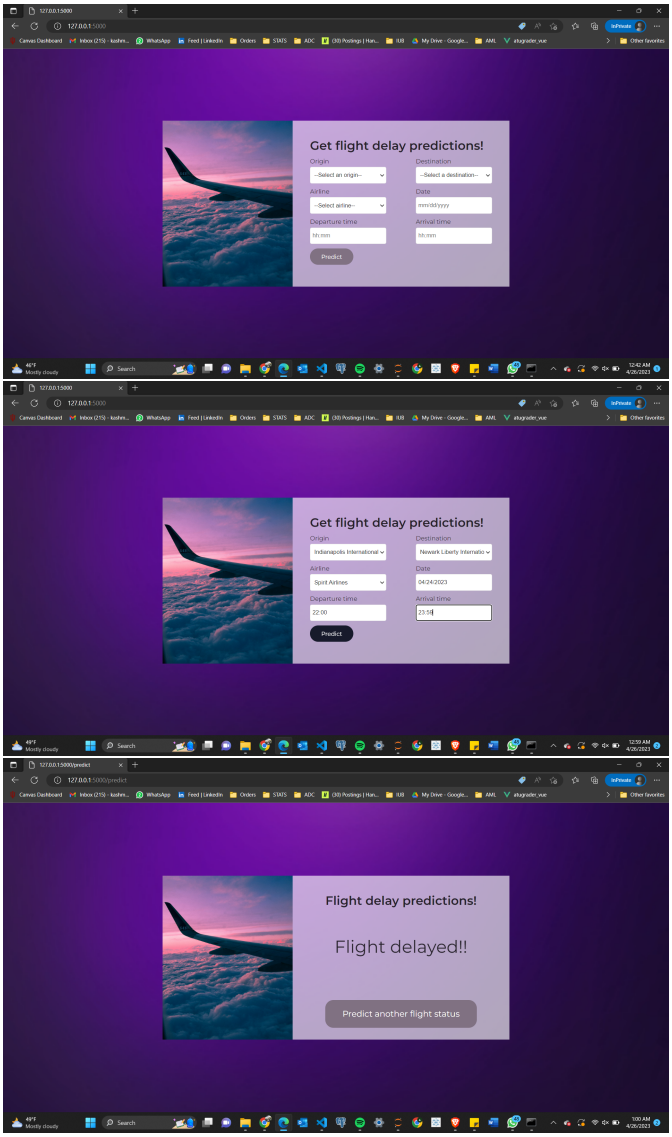
We can also obtain insights into feature importance from the Random Forest Classification model. By doing so, we observed that the most important features for predicting flight delays were Departure Delay and Taxi-Out. Interestingly, the distance from origin to destination did not contribute much to the delay, highlighting the importance of other factors in predicting flight delays.

5. Deployment and Maintenance

After achieving satisfactory performance of our machine learning models in predicting flight delays, we have decided to deploy the application as a web service.

We have utilized HTML and CSS for our front-end, and Flask, an open-source Python framework, for our back-end. This choice was based on the fact that the Python libraries used in our model designs are easily integrated with Flask, and Flask provides a convenient way to handle and manipulate multidimensional data. The combination of HTML, CSS, and Flask allows us to create a responsive and user-friendly interface for our application, while leveraging the powerful machine learning capabilities of Python in the back-end. With Flask, we can easily manage routes, request and response handling making it an ideal choice for building web applications that require data processing and analysis. Overall, the use of Flask as the back-end framework and HTML/CSS for front-end development provides a robust and efficient architecture for our application. To check if the model is trained accurately, lets check for a latest flight status, where the flight is actually delayed:





6. Summary and Conclusions

In summary, our analysis demonstrates that the Random Forest Classifier was highly effective in accurately predicting flight delays. The Random Forest Classifier performed well, achieving an accuracy score of 0.88 and an AUC score of 0.93, making it the preferred model for this problem statement.

We believe that the results of our analysis could have significant implications for society, and we hope that interested parties will find our work useful. In future work, we plan to leverage additional data sources to further improve our predictions and expand the scope of our analysis.

In addition to predicting whether a flight will be delayed or not, we also plan to explore the possibility of predicting the length of the delay. This would provide even greater value to stakeholders by allowing them to better prepare for potential disruptions and manage their time more effectively.

Our web app utilizes state-of-the-art machine learning algorithms that are continually refined and optimized through ongoing model updates. We employ a range of open-source technologies, including Python, Flask to ensure the highest levels of performance, scalability, and flexibility.

Moving forward, we plan to integrate our platform with other carrier companies to acquire robust APIs and microservices, allowing us to provide a more comprehensive range of services to our users.

Acknowledgments

We would like to express our sincere appreciation to our faculty advisor, Dr. Hasan Kurban, for his invaluable guidance and support throughout the development of this project. His expertise and mentorship have been instrumental in helping us achieve our goals with accuracy and efficiency.

Additionally, we would like to express our gratitude to the Luddy School of Informatics, Computing, and Engineering for giving us access to state-of-the-art computing tools like the Big Red 500, JetStream and Carbonate. These tools gave us liberty and assurance to thoroughly investigate the scope of our project by enabling us to fully utilize the capabilities of our models and test them with real-time data.

References

1. Roberto Henriques and Ines Feiteira. Predictive modelling: ˆ flight delays and associated factors, hartsfield–jackson atlanta international airport. Procedia computer science, 138:638–645, 2018

2. Joint Economic Commitee Majority Staff. Your flight has been delayed again. Technical report, Tech. rep, 2008