Frank Seely, Emily Zhao, Nick David, Ryan Cheng
Professor Kevin Gold
December 8 2023
Data Science 110

Exploring Factors Influencing Flight Delays in the United States: A Comprehensive Analysis Using Visualizations and Machine Learning

Introduction

As the use of air travel becomes more prevalent in America, the number of delays also rises. Flight delays inconvenience passengers and create significant financial challenges for airlines. The cause of these delays typically depends on various factors, including the airline, airport, location, and flight time. For our project, we aimed to determine the primary factors that affect flight delays by using a variety of visualizations and machine-learning techniques. We used a dataset from Kaggle that tracked airline, flight, time, airport from and to, and day of the week of flights in the U.S. to help us make these predictions. By accurately predicting flight delays, we hope to contribute to developing more effective strategies for mitigating air travel disruptions.

Previous work

Understanding the causes of flight delays is pivotal for optimizing air travel efficiency. Many studies have drawn insights from various perspectives and used different factors to analyze the elements.

In one study, Yuemin Tang used supervised machine learning models to predict flight delays. She used a dataset that recorded information on flights departing from JFK airport for one year for her predictions and used Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Trees. Her results showed that the Decision Tree algorithm had the best performance and the KNN performance was the worst.

Martina Zámková et al. analyzed the causes of flight delays of an airline operating in Europe. Mainly, she used independence and correspondence analysis and created correspondence maps to represent the results. She concluded that higher-capacity aircraft were somewhat prone to longer delays.

In the study by Miguel Lambelho and Mihaela Mitici, they conducted a generic assessment of strategic flight schedules implemented in European airports by using predictions about

arrival/departure flight delays and cancellations. They used a machine learning-based approach to assess the impact of these strategic flight schedules and developed a generic ranking of the strategic programs.

A study by Meysam Kazemi Asfea and Majid Jangi Zehib conducted an empirical investigation to determine important influences on flight delays. They used an analytical hierarchy process to compare the different factors. They decided that technical defects and delayed entry were among the most critical factors for flight delays.

Another study done by K. Sreenivasulu et al. focuses on predicting flight delays using intelligent algorithms and big data technology. They used a system that predicted delays using aviation data and employed Random Forest (R.F.), K-Nearest Neighbor (KNN) Linear Regression (L.R.), logistic regressions, and Support Vector Machine (SVM). The data used considered climate, airline information, airport terminal information, etc.

Most prior studies analyzed flight data from Europe or a specific city like New York. In our research, we wanted to look at flights across the U.S. to determine the most prominent factors. Moreover, the previous studies analyzed different factors than those in our dataset. Including these additional variables can help provide a better understanding of flight delay predictions overall.

Methodology

Data Preparation

As our dataset maintains a high usability rating by Kaggle, the raw dataset was exceptionally clean. The set contained 539,383 entries and initially had zero null values. We believe the ‘I.D.’ and the ‘Flight’ columns (containing flight numbers for each entry) were unnecessary for our analysis, so they were dropped. For readability, we added a column containing String values representing each day of the week, corresponding to the 'DayOfWeek' column containing integer values from 1 to 7 (1 being Monday, 2 being Tuesday, etc.) Similarly, we added a separate column, 'DayType,' containing binary values – 1 for weekdays and 0 for the weekends. Additionally, we added another column representing the hour of the day for each flight (0 being midnight, 1 being 1 AM, etc.) This would make it easier to visualize a bar graph of the relationship between the hour of day and the count of delays in the dataset.

We also created a 'FlightType' column, which further classified flights into three different types based on the 'Length' column (containing the length of flights in minutes). Flights were determined to be 'Short-haul' if the flight was less than 3 hours ('Length' < 180), 'Medium-haul' if flights were between 3 and 6 hours (180 ≤ 'Length < 360), and 'Long-haul' if flights were more

significant than 6 hours ('Length' ≥ 360). This column will come in handy when determining if the length of a flight may affect the likelihood of delays when generating visualizations.

Another factor we were interested in testing was the distinction between Western U.S. flights versus Eastern U.S. flights on flight delays. Our dataset provided three-digit letter codes designated to each airport by the International Air Transport Association (IATA), including departure and landing locations. Using this information, we found a separate dataset (which we named 'iata-icao.csv') containing the airport codes as long as the corresponding longitude. All other columns were dropped for convenience before merging this DataFrame with our original dataset. A 'left' merge was performed from the second dataset onto the first. It was essential to use this specific type of merge as we knew the 'iata-icao' DataFrame contained more IATA code entries than our original DataFrame, preventing unnecessary IATA codes and NaN values. Finally, we utilized simple lambda functions to create two more rows, which converted longitudes into 'WEST' or 'EAST' string values for readability. West Coast airports were classified as having a longitude to the left of the 100th (less than -100) meridian, and East Coast airports were to the right of the 100th meridian (greater than -100).

To better understand what statistical and machine learning methods to use, we performed some exploratory data analysis by creating various visualizations:

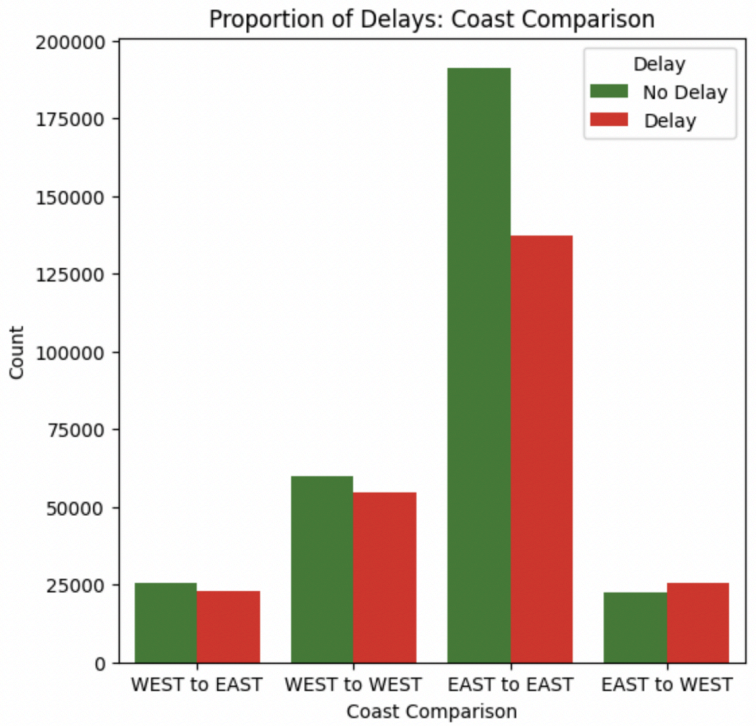

Fig. 1. Proportion of Delays: Coast Comparison. For this visualization, we created a bar chart to compare the combinations of possible flights flying from coast to coast. The flights from east to west have a more considerable proportion of delayed flights than nondelayed flights.

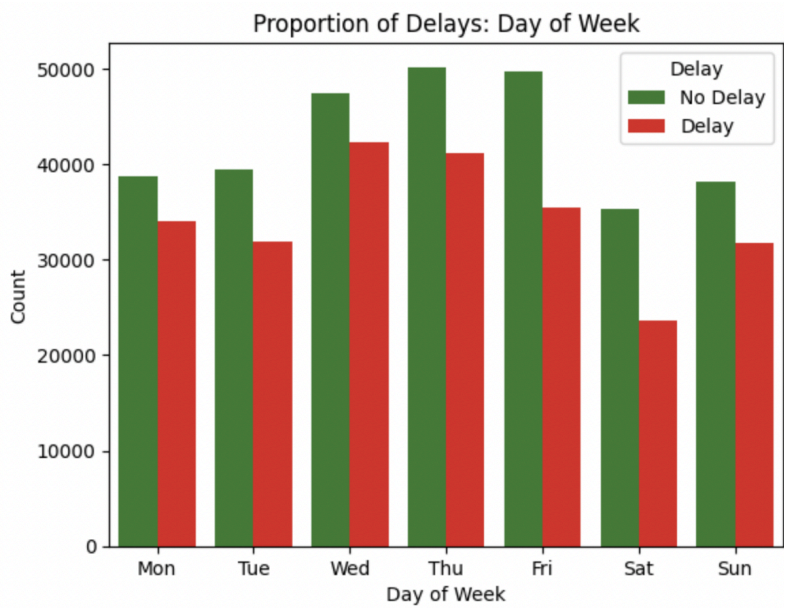

Fig. 2. Proportion of Delays: Day of Week Comparison. This visualization uses a bar chart to compare the proportion of delayed flights to nondelayed flights for each day of the week. We wanted to see if a particular day had a more significant proportion of delayed flights compared to the rest of the days.

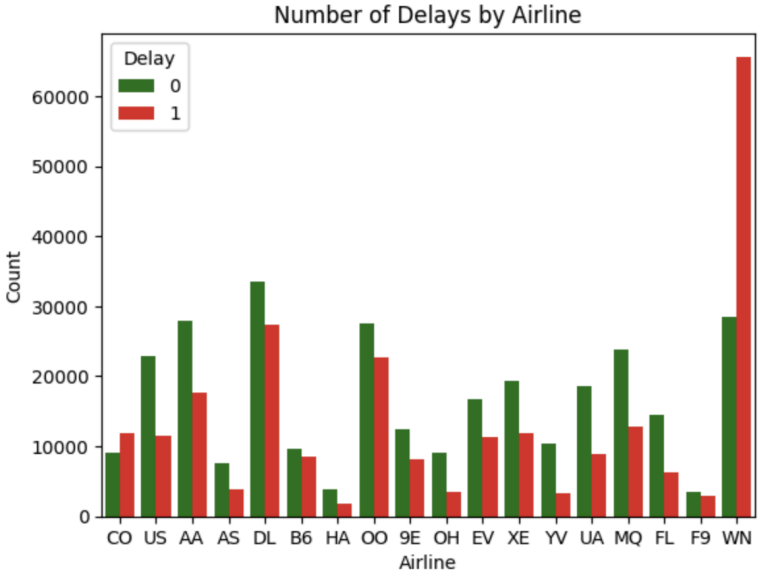

Fig. 3. Proportion of Delays: Airline Comparison. In this visualization, we compared the number of delays between each airline in the dataset. Although most airlines did not show a significant delay trend, Southwest (W.N.) stands out with a considerable disparity in delay occurrences.

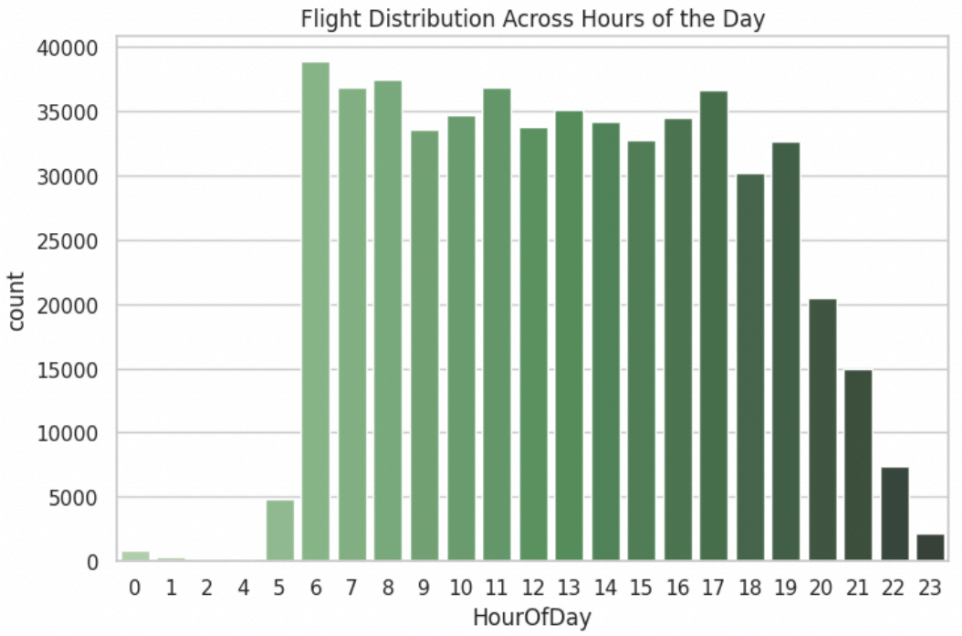

Fig. 4. Flight Distribution Across Hours of the Day In this count plot, we were curious whether a particular hour had more flights than another. As we can see, it is pretty even from 6 AM to 6 PM. The highest number of flights happens at 6 AM.

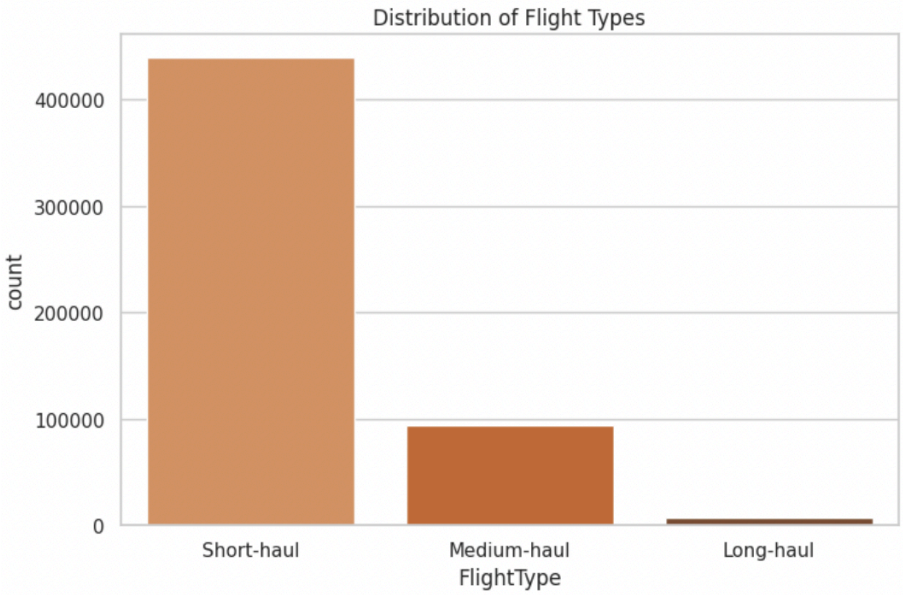

Fig. 5. Distribution of Flight Types. In this visualization, we compare the number of flight types. The bar chart shows that most flights are short-haul, less than 3 hours long. Less than a quarter of the flights are medium-haul or between 3 to 6 hours. Then, only a tiny fraction of the flights are long-haul flights or longer than 6 hours. The visualization shows that most of our data comes from short-haul flights, which will influence our analysis.

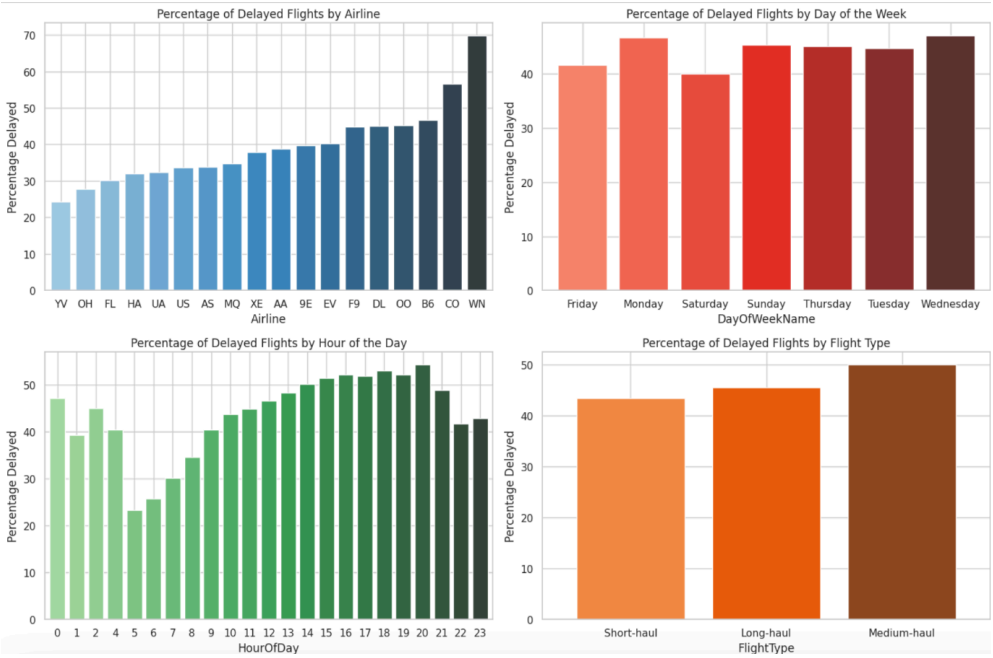

Fig. 6: Distribution of Flights by Percentage: In the figure above, we created subplots further to visualize our data as a percentage for each factor. The distribution helps us better understand the distribution and contribution of each element to the overall dataset. It also allows us to compare each factor's relative importance or prevalence easily.

Statistics

For our statistical methods, we used a correlation analysis to determine what factors had the most substantial relationship. We compared 'Airline,' 'DayOfWeekName,' and 'FlightType' against 'Delay' to determine which factor most influenced delays. We also used chi-squared tests for airline, day of the week, and flight type. By running a chi-square test, we wanted to see if there was a significant difference between the variables that cause delays. Finally, we ran a T-test between West and East Coast flights in the United States against 'Delay' to determine if there was a significant statistical difference in flight delays depending on the location of the plane's departure.

Machine learning

We used a decision tree for our first machine learning method because it would be optimal to predict a binary classification since it can take numerical and categorical data. Using the decision tree, we are trying to predict if the flight will be delayed based on the features in our data frame, including airline, day of the week, hour of the day, flight length, and flight type. To utilize categorial features ('Airline,' 'DayOfWeek,' 'FlightType'), we needed to apply one-hot encoding, which turns categorical features into numerical values that can be processed and interpreted by our machine learning models. Next, we split the data into training and testing sets (30% of the data in the test set). We utilized a classifier pipeline rather than creating the decision tree classifie

The pipeline's job is to encapsulate a series of data transformations into a single entity. It consists of two main components: the preprocessor and the classifier. The preprocessor handles the categorial features through the previously mentioned one-hot encoding techniques. The remaining parts ('HourOfDay' and 'Length') remained untouched as they were already fit to be interpreted by the model. The decision tree classifier is trained on the preprocessed data and the trained model, and then the remaining data is used to predict the target variable, 'Delay.' We varied the max depth by choosing different values to optimize the model's performance. The original decision tree was generated with a depth of 20. The tree was then tuned with depths of 3, 5, 10, and 15.

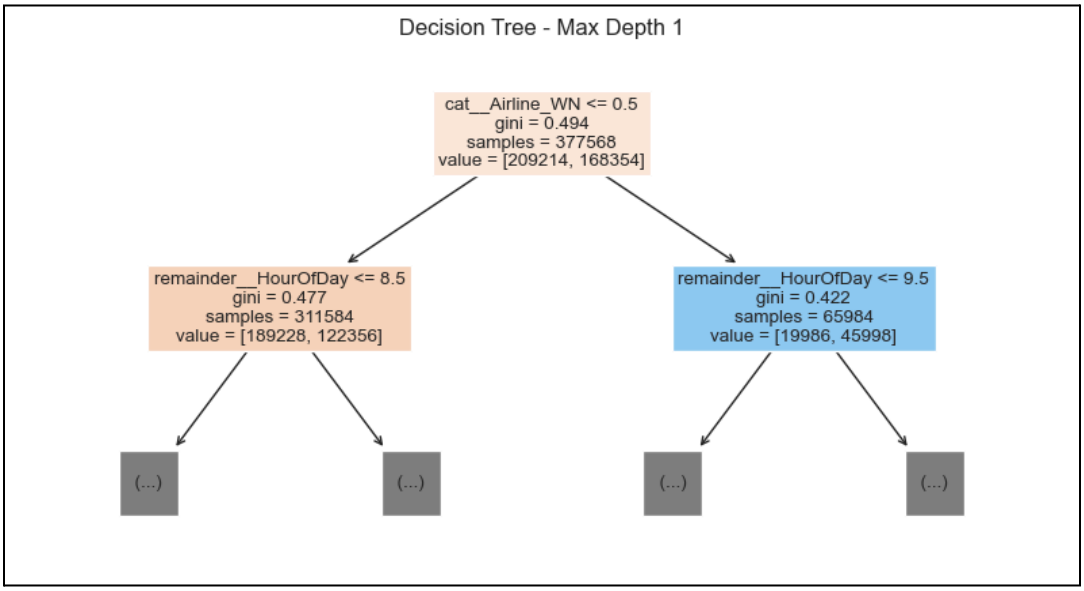

Fig. 7. The basic structure of the Decision Tree. It only depicts a depth of one, while the actual model's depth was 20.

We performed a logistic regression in our secondary machine learning model to analyze the relationship between our binary classification labels (delay versus no delay). Logistic regression is also able to take numerical and categorical values as input. Hence, we used the same information as we did for the decision tree. We varied the *C* parameter with 0.01, 0.1, 1, 10, and 100 values to further optimize the model's performance. Testing with various parameters may be a helpful way to control overfitting/underfitting; larger values of *C* reduce the strength of regularization (the model fits data more closely), while smaller values of *C* tend to increase the strength (leads to simpler models).

Results

Statistics

To look for correlations between the different numerical features, we created a heat map to determine the most correlated features. However, the map showed that most variables had weak correlations since most of the values were less than 0.0. Time (flight departure time) and delay were the only variables with a positive correlation with a value of 0.15. Furthermore, our correlation analysis ultimately came to a similar conclusion. We compared the relationship of length of flight and hour of day to delays. The R-value for distance and delay was 0.040, which suggests a weak correlation. The correlation between delay and hour of the day had a slightly stronger correlation with an R-value of 0.15, which means it has a somewhat more significant impact on delays.

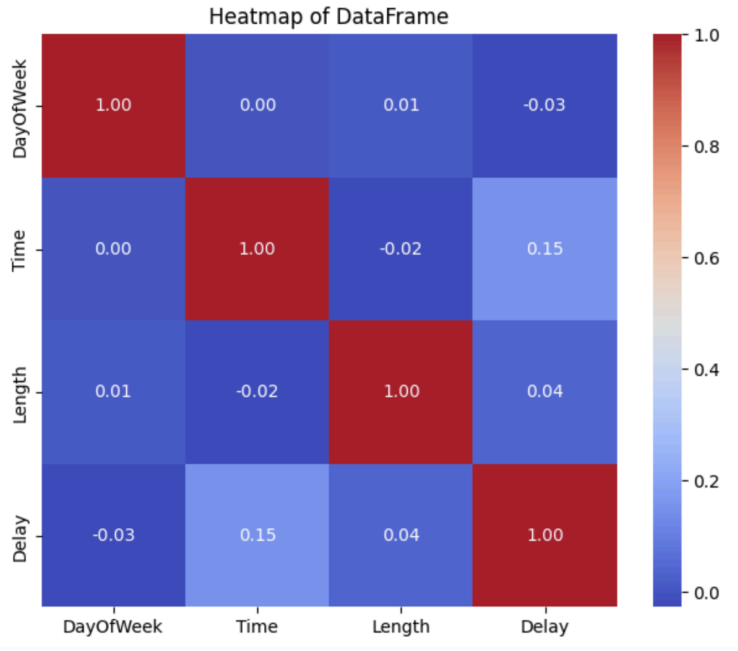

Figure 8. Heatmap Comparing Various Features. The heatmap above compares the variables 'Delay,' 'Length,' 'Time,' and 'DayOfWeek' and their corresponding correlation.

We also did chi-squared tests for our categorical variables to test if there was a considerable difference between them. For the chi-squared test on the airlines, we got a p-value of 0.00, which is less than the p-value threshold of 0.05 and shows a significant difference between the airlines. This would make sense since we saw that Southwest had a considerable difference in delays compared to the other airlines. The p-value we got from our chi-square test reaching days of the week was 2.60e-251 – approximately 0 – which is less than 0.05 and shows a significant statistical difference compared to delay. Similarly, the chi-squared test for flight type had a

p-value of 9.34e-294 – approximately 0. This is less than the p-value threshold of 0.05, suggesting a significant difference between flight types and delay.

Additionally, a T-test was performed between delays on flights departing from the East Coast and those from the West Coast. After setting up this test, we found a T-statistic of approximately 29.55 and a P-value of 1.17e-191. Because the P-value is nearly zero, we reject the null hypothesis and conclude that there was some statistical significance between delays in the east versus west coast of the United States.

Machine learning

The decision tree model's performance was based on an accuracy score representing the percentage of times it could correctly classify whether a flight was delayed. Initially, we achieved an accuracy score of 63%. When we varied the maximum depth of the tree, we found that we could improve the accuracy to about 64.4%. This improvement showed that fine-tuning the tree's complexity captured more nuanced patterns in the data, resulting in slightly higher precision in classifying flight delays. Still, this difference in the model's accuracy is insignificant, leading us to believe the dataset needed to fit the model better.

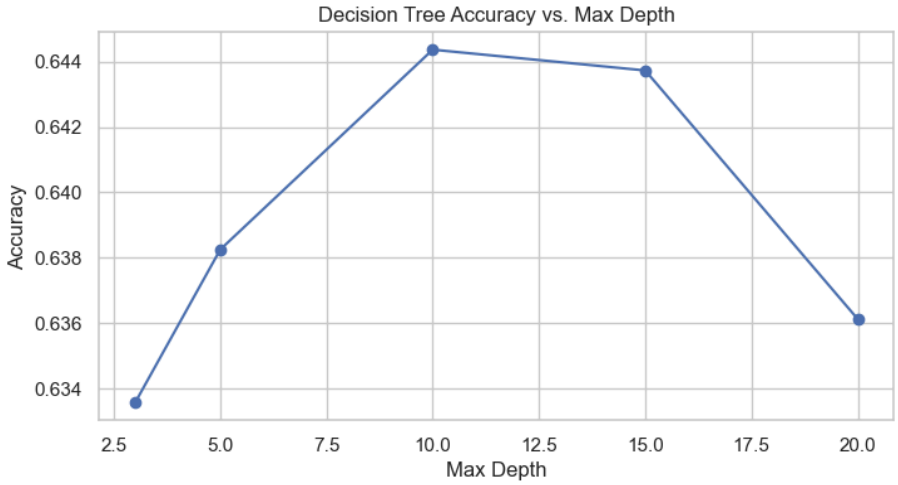

Figure 9. Decision Tree Accuracy vs. Max Depth. Shows the various Max Depths used when tuning the decision tree classifier and their corresponding scores.

The logistic regression model's performance was also based on accuracy. Our original accuracy score was 63.4%. Although we tried varying the C parameter to increase performance, it stayed around the same accuracy of approx. 63.4 percent, meaning the model correctly labels only about three of every five instances. Unfortunately, this score is not very impressive in correctly classifying delayed flights from nondelayed flights in our dataset.

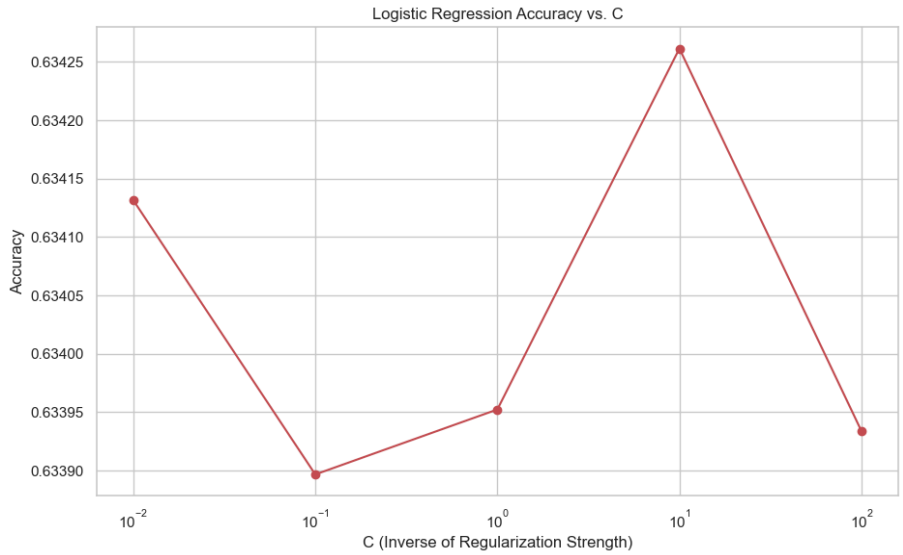

Fig. 10. Logistic Regression Accuracy vs. C. Depicts the relationship between various *C* values chosen for the logistic regression and their corresponding accuracies when tuning the model.

The recall (sensitivity score) was 0.78 for the classification 'No Delay' (value = 0) and 0.45 for the classification 'Delay' (value = 1). The recall represents the ratio of accurate optimistic predictions to actual positive instances; therefore, results closer to 1 represent more accurate predictions. Similarly, the f1-score was 0.70 for 'No Delay' and 0.52 for 'Delay.' A high f1-score indicates a good balance between precision and recall. This suggests the model can sufficiently identify positive instances while minimizing false positives/negatives. While a score of 0.70 is moderately acceptable, a score of 0.52 could be more impressive, signaling unreliability in the model.

To better understand our results with the logistic regression model, we plotted a Receiver Operating Characteristic (ROC) curve, which is a representation that illustrates the performance of our binary classification model. It plots the true positive rate (sensitivity) against the false positive rate at various threshold settings. A true positive rate may be found with the following equation: *# Delayed predicted as Delayed / (# Delayed predicted as Delayed + # Delayed predicted as not Delayed)*. On the other hand, the false positive rate is plotted with the following: *# Not delayed predicted as Delayed / (# Not delayed predicted as Delayed + # Not delayed predicted as delayed).*

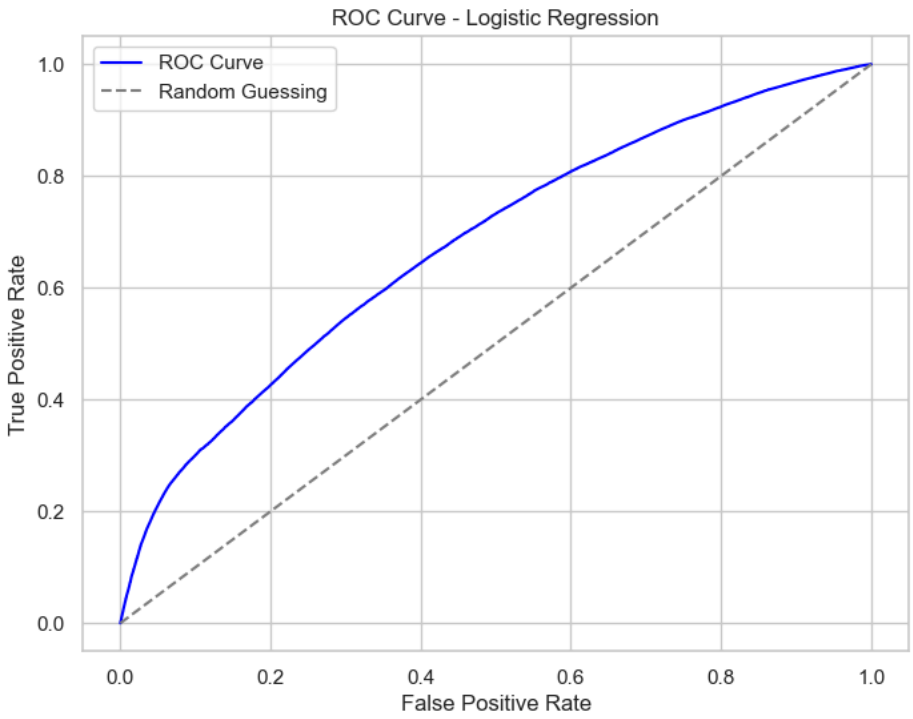

Fig. 11. ROC Curve - Logistic Regression.

This curve tells us that our model is better than random guessing (depicted by the diagonal line), though it is not performing exceptionally well either. A model that performed exceptionally well will be able to reach the top left corner of the graph. Meanwhile, a model that performed poorly will be towards the bottom right corner, meaning it performed worse than random guessing.

Conclusion

Our analysis of flight delays in the United States utilized machine-learning techniques and visualizations to determine the primary factors influencing disruptions in air travel. By examining a dataset encompassing airline, flight, time, airport details, and day of the week, we sought to contribute valuable insights to flight delay predictions. While other studies often focused on specific regions or factors, our research took a holistic approach, considering a nationwide perspective and additional variables. Using airport codes and longitudes, our exploration into the distinction between coast-to-coast flights provided a better understanding of regional influences on delays. Moreover, the application of decision trees and logistic regression models revealed patterns in flight delays based on factors such as airline, day of the week, hour of the day, and flight length. While the decision tree demonstrated some degree of improved accuracy through fine-tuning, the logistic regression model maintained a consistent accuracy level. Despite our findings, the complexity of factors contributing to flight delays emphasizes the need for additional research and the development of optimized strategies to enhance air travel efficiency. Our work aims to reduce disruptions in the aviation industry and develop accurate predictive models for flight delays.

Bibliography

[1] Asfe, Meysam & Jangizehi, Majid & Tash, Mohammad & Yaghoubi, Nour Mohammad. (2014). Ranking different factors influencing flight delay. Management Science Letters. 4. 1397-1400. 10.5267/j.msl.2014.6.030.

[2] Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020). Machine Learning Approach for Flight Departure Delay Prediction and Analysis. Transportation Research Record, 2674(8), 145-159. https://doi.org/10.1177/0361198120930014

[3] K. Sreenivasulu, B. Sowjanya, V. R. Motupalli, S. H. Yadav, K. K. Baseer and M. J. Pasha, 'Prediction of Flight Delay through Intelligent Algorithms and Big Data Technology,' 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 1074-1080, doi: 10.1109/ICAAIC56838.2023.10141246. https://ieeexplore.ieee.org/document/10141246/figures#figures

[4] Lambelho, Miguel, et al. 'Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions.' Journal of Air Transport Management, vol. 82, 2020, p. 101737, https://doi.org/10.1016/j.jairtraman.2019.101737.

[5] Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In 2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15-17, 2021, Singapore, Singapore. ACM, New York, NY, USA, 7 Pages. https://doi.org/10.1145/3497701.3497725 https://dl.acm.org/doi/fullHtml/10.1145/3497701.3497725

[6] Zámková, M.; Rojík, S.; Prokop, M.; Stolín, R. Factors Affecting the International Flight Delays and Their Impact on Airline Operation and Management and Passenger Compensations Fees in Air Transport Industry: Case Study of a Selected Airlines in Europe. Sustainability 2022, 14, 14763. https://doi.org/10.3390/su142214763