Predicting Flight Delays

How Weather & the Taylor Swift Era's Tour Impacts Flight Delays

Hope Husemann

Data Science Capstone Project, Nov 3, 2024

Springboard

The problem

Goal

Predict the likelihood or extent of USA flight delays based on historical flight data, weather conditions, airline/airport characteristics, and other external factors including Taylor Swift's Era's Tour dates and NFL game dates.

Context

Flight timelines are often unpredictable to the average person, because delays can be caused by so many different factors. Our goal is to provide meaningful data in regards to their upcoming flights to predict likelihood or extent of delays.

Problem statement

Develop a machine learning model to predict flight delays to provide valuable insights and decision-making support for airlines, airports, and passengers to better manage delays.

Challenges deep-dive

Challenge 1

Limited to free, accessible data available online

Challenge 2

Scope of Datasets due to Limited Memory: While I'd like to be able to hone in on and train this dataset further with a larger scope (more years), but due to size of datasets I am only submitting what time has allowed.

Challenge 3

Once the dataset has been better trained with more years, make it available to larger audience.

Scope of Solution space

https://www.bts.gov/

Bureau of Transportation Statistics (BTS): We will utilize on-time performance data from the BTS https://www.bts.gov/ to acquire historical flight information like departure time, arrival time, origin/destination airports, airlines, and delay information (canceled, diverted, etc.). The Bureau of Transportation Statistics (BTS) typically provides flight data in a standardized format, and the times are generally recorded in Coordinated Universal Time (UTC). This means that all departure and arrival times are expressed in a single time zone, allowing for consistent comparison across different flights and time zones.

Scope of Solution space

https://ourairports.com/.

OurAirports.com: This website provides airport data like location, runway length, and passenger traffic, offering insights into potential bottlenecks
https://ourairports.com/.

Scope of Solution space

https://www.ncei.noaa.gov/access/search/.

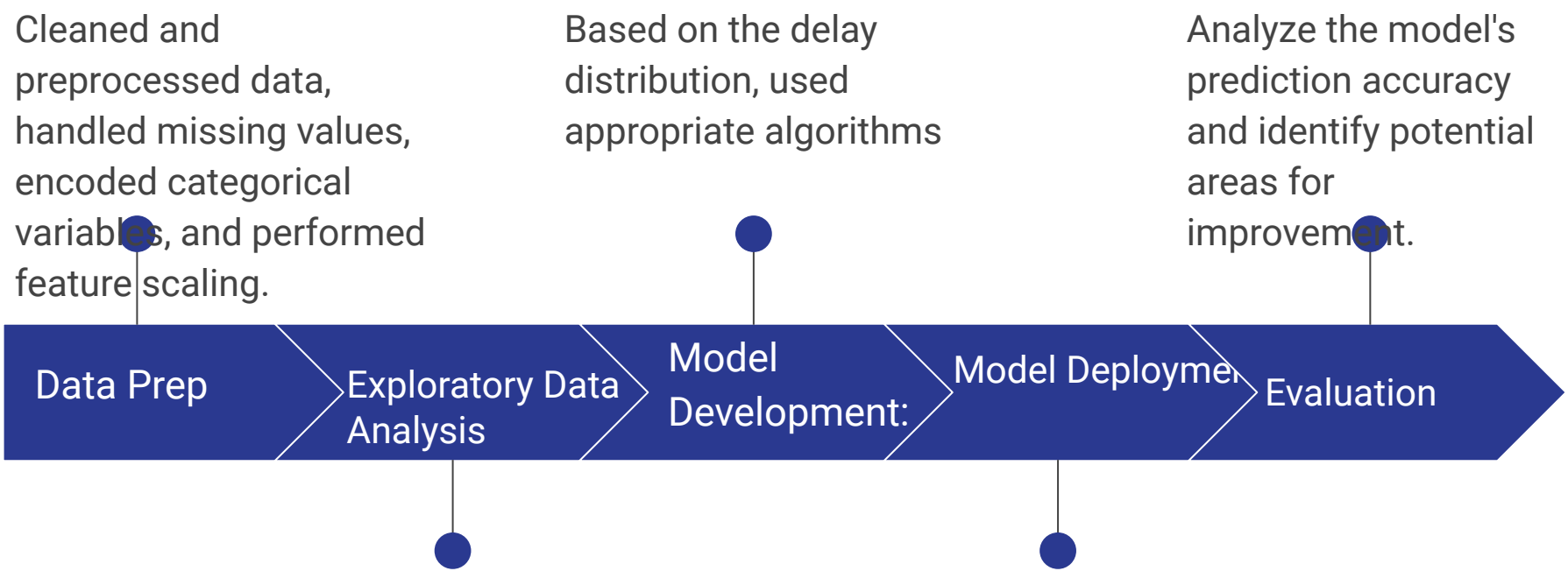
NOAA's Global Historical Climatology Network (GHCN) Daily Data: We will extract weather data like temperature, precipitation, wind speed, and visibility from NOAA's GHCN database for the corresponding flight times https://www.ncei.noaa.gov/access/search/.

Scope of Solution space

www.nfl.com
https://www.taylorswift.com/tour/

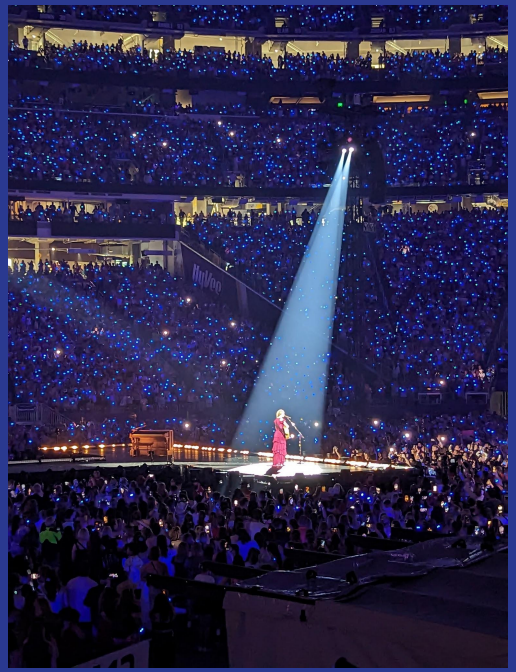
Explore NFL schedule data and Taylor Swift Eras Tour dates to analyze potential impact of games and tour dates on nearby airports (e.g., increased congestion).

Implementation


Cleaned and preprocessed data, handled missing values, encoded categorical variables, and performed feature scaling.
Based on the delay distribution, used appropriate algorithms
Analyze the model's prediction accuracy and identify potential areas for improvement.
Data Prep
Exploratory Data Analysis
Model Development:
Model Deployme
Evaluation
Create visualizations to explore how features correlate with flight delays.
Deploy the best performing model

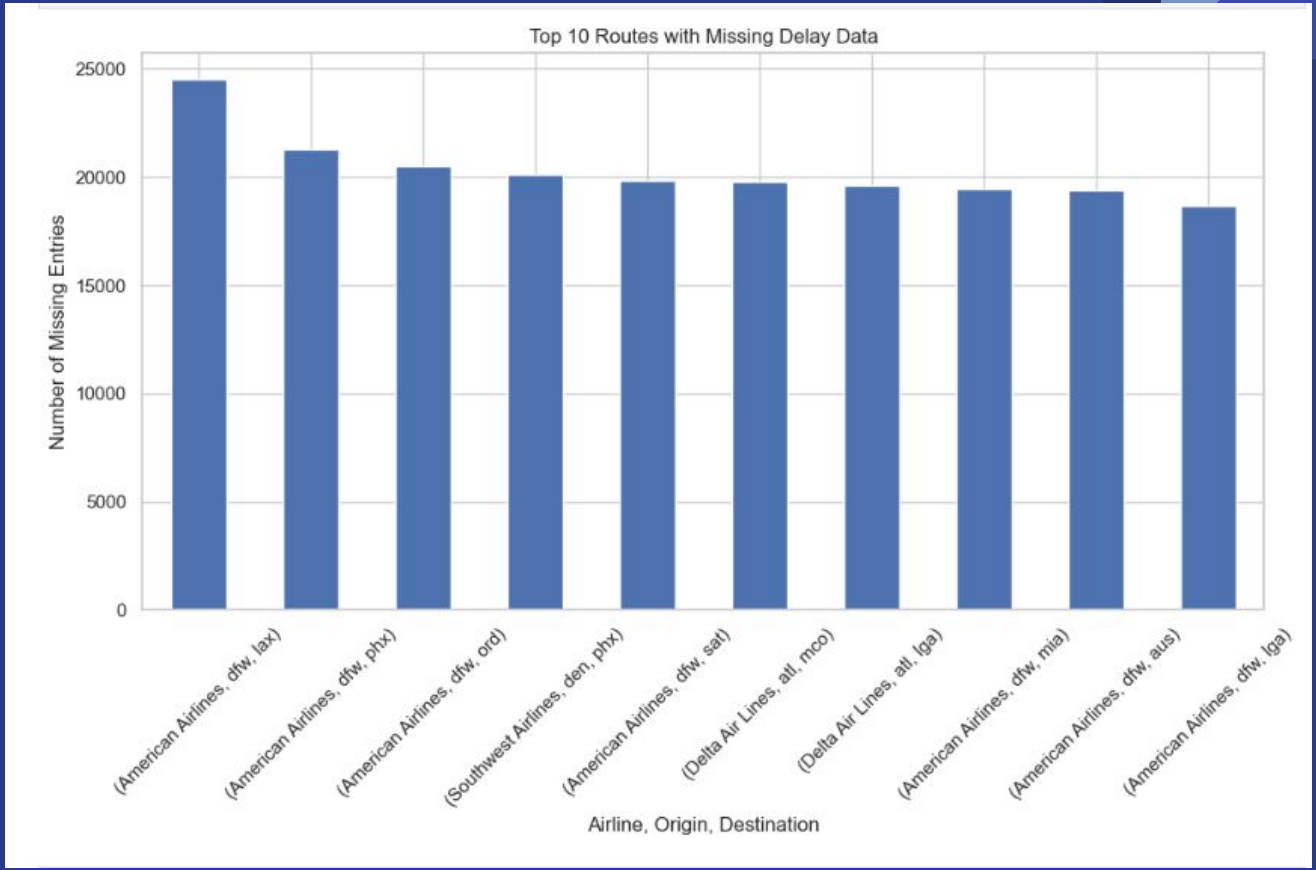
Questions and Exploration of Data

What factors might affect flight delays?

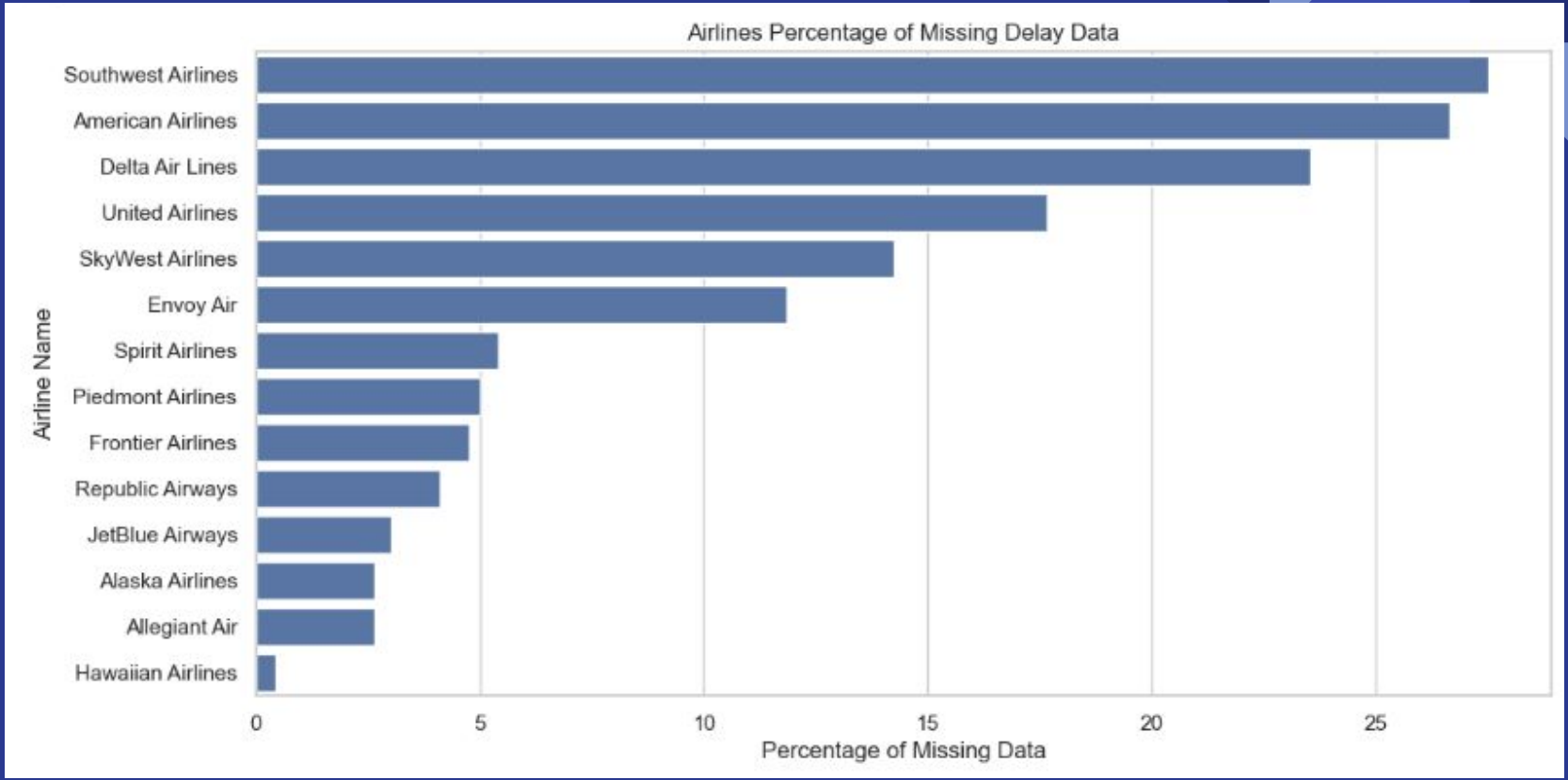


- Flight schedules
- Flight length
- Airport locations (both origin and destination)
- Weather conditions (both origin and destination)
  - Includes max temp/min temps, precipitation
- Do large events impact air traffic?
  - Taylor’s Swift Era’s Tour Dates
  - ~~NFL Dates~~

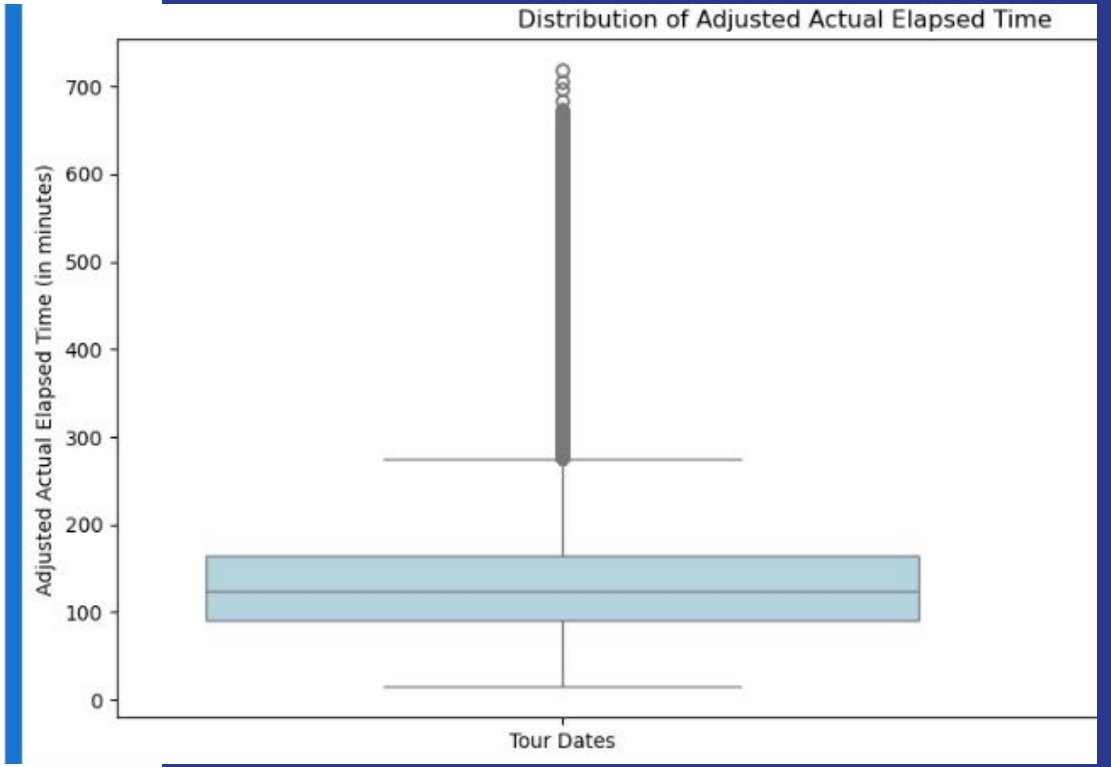
Missing Data by Airline Route



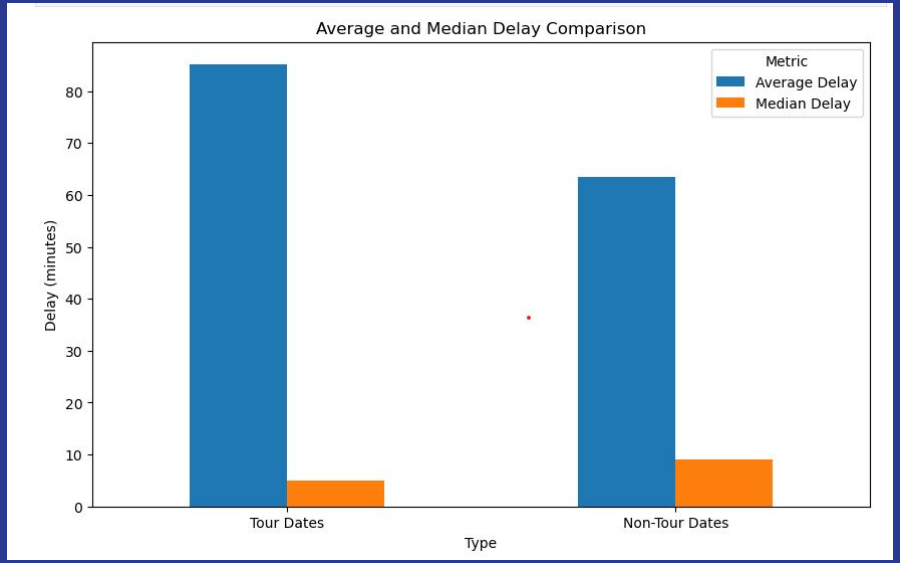
Missing Data by Airline



Do the Eras Tour Dates affect flight delay times?

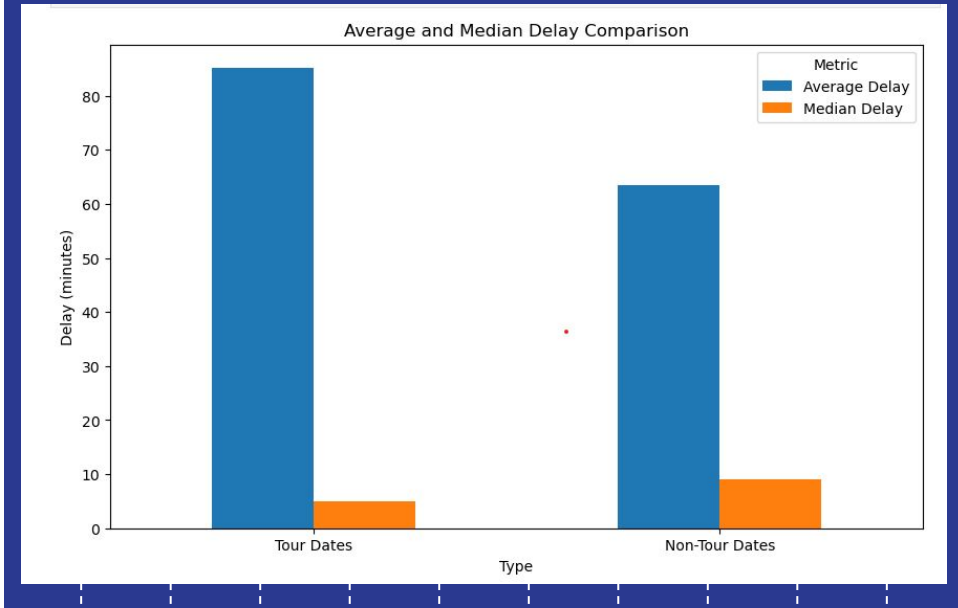


Do the Eras Tour Dates affect flight delay times?



Impact: Swifties, Plan ahead!

Would increased traffic at airports result in longer delays for tours where the Eras Tour perform?
According to results, yes!



Do NFL Games affect flight delay times?

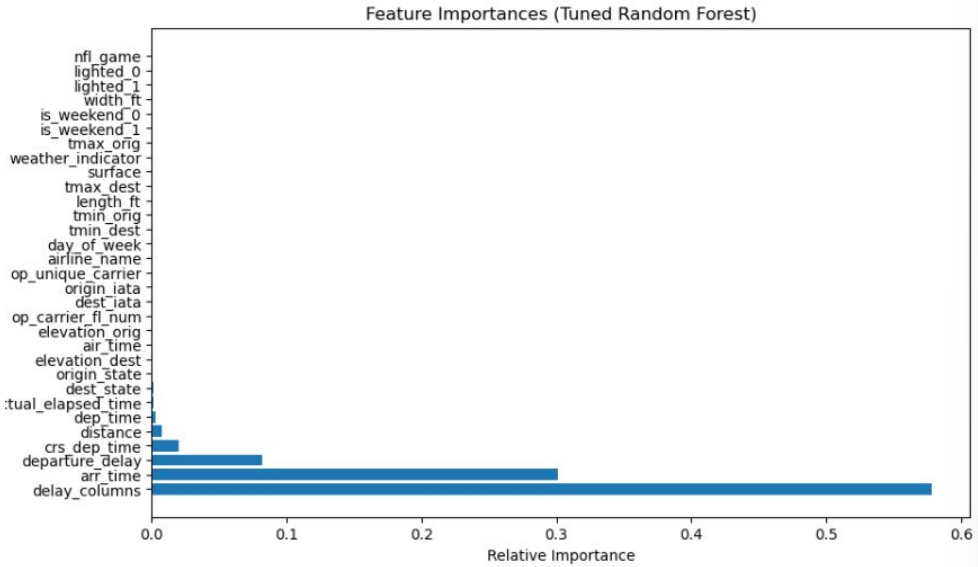
While we guessed NFL games may bring enough extra airline traffic to a city to see noticeable delays, the results were inconclusive. Correlation was not strong enough to suggest one way or another.

Fun Findings:

JFK, LAX, ORD are most likely to be late on arrival if you're departing from here.

DFW, SFO, and MIA are most likely to be late on arrival if they are your destination airport.

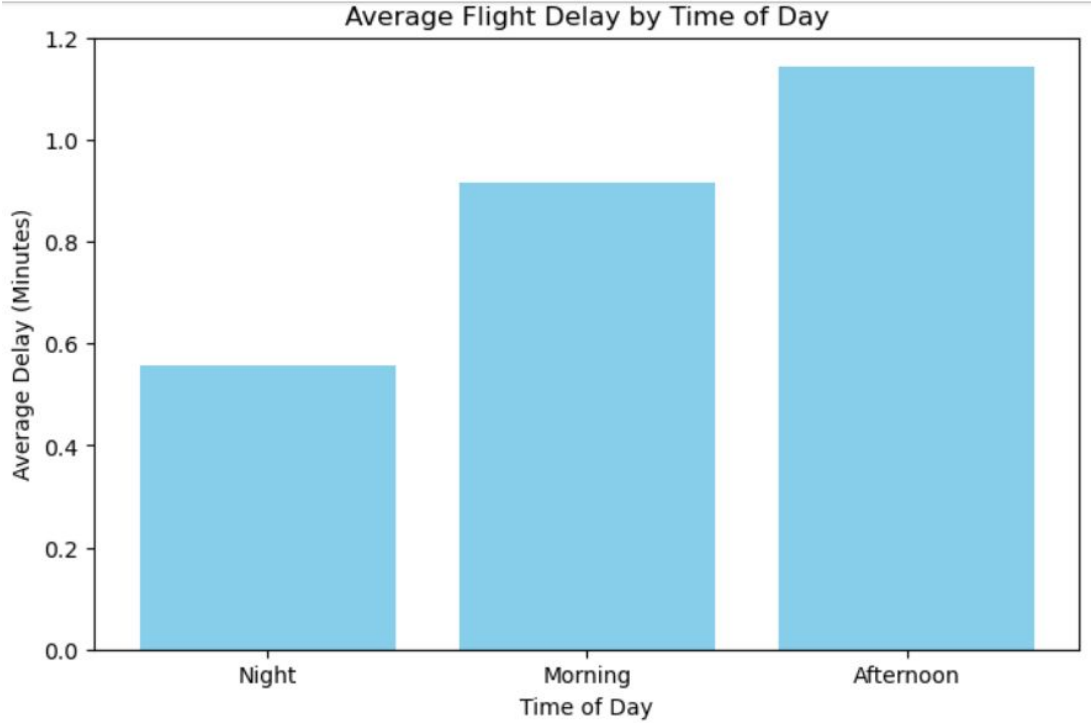
Weather Impact:



Due to what appears to be non-reported weather data, the data alone was too sparse to have any relevant value.

However, I combined any weather data that was reported for a flight and combined to be counted into a single column, ‘delay_columns’, which can be seen here as the feature with most relative importance!

Time of Day:



	time_of_day	delay_minutes
0	Night	0.558512
1	Morning	0.916216
2	Afternoon	1.142897

Model Development:

- Based on the delay distribution, choose appropriate algorithms
- Train multiple machine learning models, such as:
    i. Regression Models:
        1. Linear Regression,
        2. Random Forest Regression
        3. Gradient Boosting
- Use metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), to evaluate model performance.

Final Model:

Random Tree Regressor

Model Performance Comparison:

Model	RMSE	MAE	Accuracy	R^2
Linear Regression	3.1926	0.70541	N/A	N/A
Random Forest Regression	0.00582	0.00291	N/A	.9997
Gradient Boosting	0.08038	0.01078	N/A%	0.9983

Conclusion

This project demonstrates the power of machine learning in predicting flight delays based on a variety of factors, including weather, airline characteristics, and external events. Future work will involve additional testing and model tuning, as well as the inclusion of more external variables to increase the model's accuracy and robustness.

Challenges and Future Work

Model Interpretability: While Random Forest performed well, it is important to improve the interpretability of the model. Techniques like SHAP (SHapley Additive exPlanations) values could be used to understand which features most influence flight delays.

External Factors: More granular data on external events (e.g., large conventions, holidays) could enhance the model's predictive power. This can be investigated further by merging other event-related data.

~~Build~~ing out for specific use cases.

Appendices

1. **Code for Data Preprocessing and Feature Engineering**
   a. **https://github.com/hopehusemann/Capstone**
2. **Please reach out for access to some of the .csv files if needed. The compressed files are still too large to load to GitHub.**

Thank you!

For any questions, please contact me.