Predicting Flight Delays

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

Hope Husemann
Data Science Capstone Project, Nov 3, 2024
Springboard

Introduction

Predicting flight delays is a complex but valuable task that can offer insights for a variety of stakeholders, including airlines, airports, passengers, logistics companies, and businesses relying on corporate travel. This project focuses on predicting flight delays using historical flight data, weather conditions, and external factors such as NFL games and the Taylor Swift Era's Tour.

The goal of this analysis is to develop a machine learning model that can predict the likelihood and extent of flight delays, providing valuable decision-making support for airlines, passengers, and other stakeholders in managing delays and mitigating their impact.

Problem Statement

Flight delays are influenced by multiple factors such as:

- **Weather conditions** (e.g., temperature, precipitation, snow),
- **Airline and airport characteristics** (e.g., size, location),
- **External events** (e.g., major events like NFL games or large concerts, such as the Taylor Swift Era's Tour),
- **Operational delays** (e.g., air traffic, security, mechanical issues).

The challenge is to integrate all these data sources and develop a robust predictive model that can forecast delays, enabling better planning and response.

Predicting flight delays has broad applications beyond the immediate scope of the airline industry including, but not limited to the following:

- Logistics and Supply Chain Management: Freight companies can use delay predictions to manage logistics and adjust delivery schedules based on anticipated disruptions in air cargo transport. By anticipating delays, logistics companies can optimize routing and scheduling to minimize the impact on their supply chains.
- Large Businesses and Corporate Travel: Businesses may need to partner with airlines for corporate travel needs. Predicting flight delays helps businesses choose airlines that offer more reliable service, improving the efficiency of corporate travel and reducing risk of disruptions..
- Travel Insurance: Insurance companies can use delay predictions to better manage claims related to flight delays, helping to streamline the claims process and improve customer service. They can also use predictive insights to price travel insurance policies more accurately, taking into account the likelihood of flight delays.
- Travel and Hospitality Industry: Travel agencies and booking platforms can use delay predictions to offer more accurate travel itineraries and recommendations. By understanding patterns in flight delays, these platforms could consider implementation of dynamic pricing and availability in real time to mitigate the impact of delays.

Data Sources

The following datasets were used:

1. **Flight Data**: Includes information about each flight's schedule, delays, and characteristics (e.g., airline, route, origin and destination airports, and actual vs. scheduled times).
   a. **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.

2. **OurAirports.com:** This website provides airport data like location, runway length, and passenger traffic, offering insights into potential bottlenecks https://ourairports.com/.
3. **Airport Information**: General data on airport locations, size, and services.
4. **Weather Data**: Includes weather information for the origin and destination airports, such as temperature, precipitation, and snowfall.
    a. **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/.
5. **External Event Data**: Includes NFL game schedules and Taylor Swift Era's Tour dates, which could potentially affect flight schedules due to increased air traffic and congestion around those events. Data from nfl.com and taylorswift.com

Context:
Flight timelines are often unpredictable to the average person, because delays can be caused by so many different factors. Our goal is to review historical flight data, weather conditions, airline/airport characteristics, and potentially external factors such as NFL games to provide meaningful data in regards to their upcoming flights to predict likelihood or extent of delays.

Goal:
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**.**

This proposal submission shows:

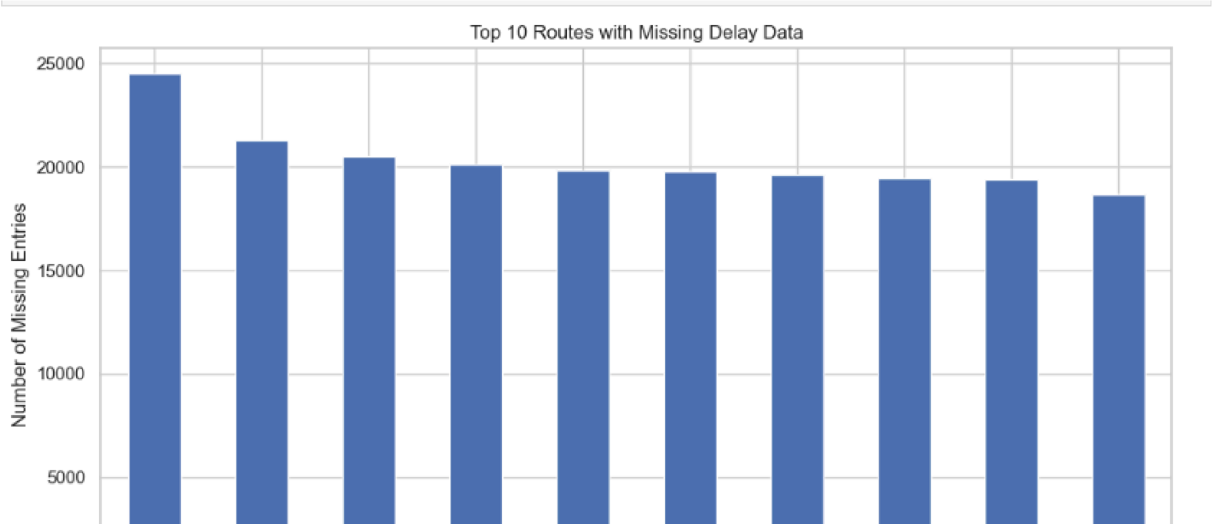
- All data wrangling steps were applied where necessary.
- EDA figures and visualizations were created and reviewed for feature relationship understanding.
- the create dummy features steps were completed.
- the features were magnitude standardized.
- the split into train and test data subsets was completed. ○
- three different models were built.
    - Linear
    - Regression
    - SVC
    - Other
- the model performance comparison table is filled in.

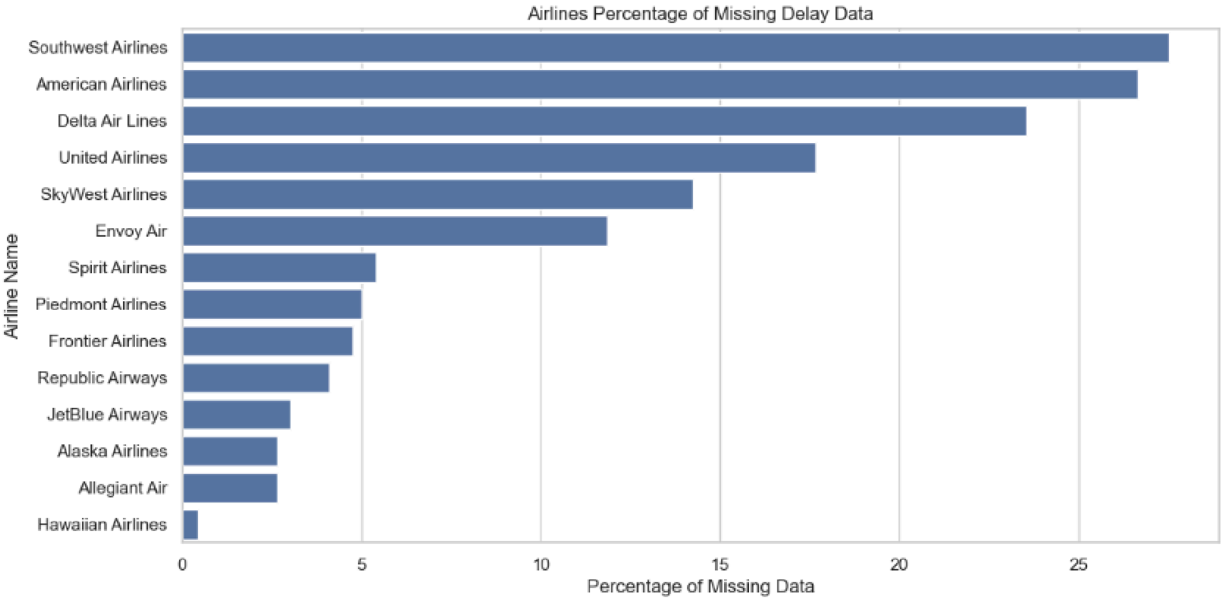
- Upload datasets:
    - flights_data = pd.read_csv(r"C:\Users\hopeh\Desktop\data_science_bootcamp\flight_times_capstone\flights_airport_iata.csv", low_memory=False)
    - weather_data = pd.read_csv(r"C:\Users\hopeh\Desktop\data_science_bootcamp\flight_times_capstone\weather_iata.csv", low_memory=False)
    - IATA Codes - Universal Code for each Airport (used to merge all data)
    - NFL Game Schedule
    - Taylor Swift Era's Tour Schedule

- Merged airport data, flight data, airport data based on the date, location, and nearest medium/large airport nearby (via IATA Codes.)
- Merged flights data and weather data.
    - At origin airport and destination airport
- Merged Era's Tour Dates by nearest IATA Code
- Merged NFL Game Day by nearest IATA Codes to corresponding closest NFL Stadiums

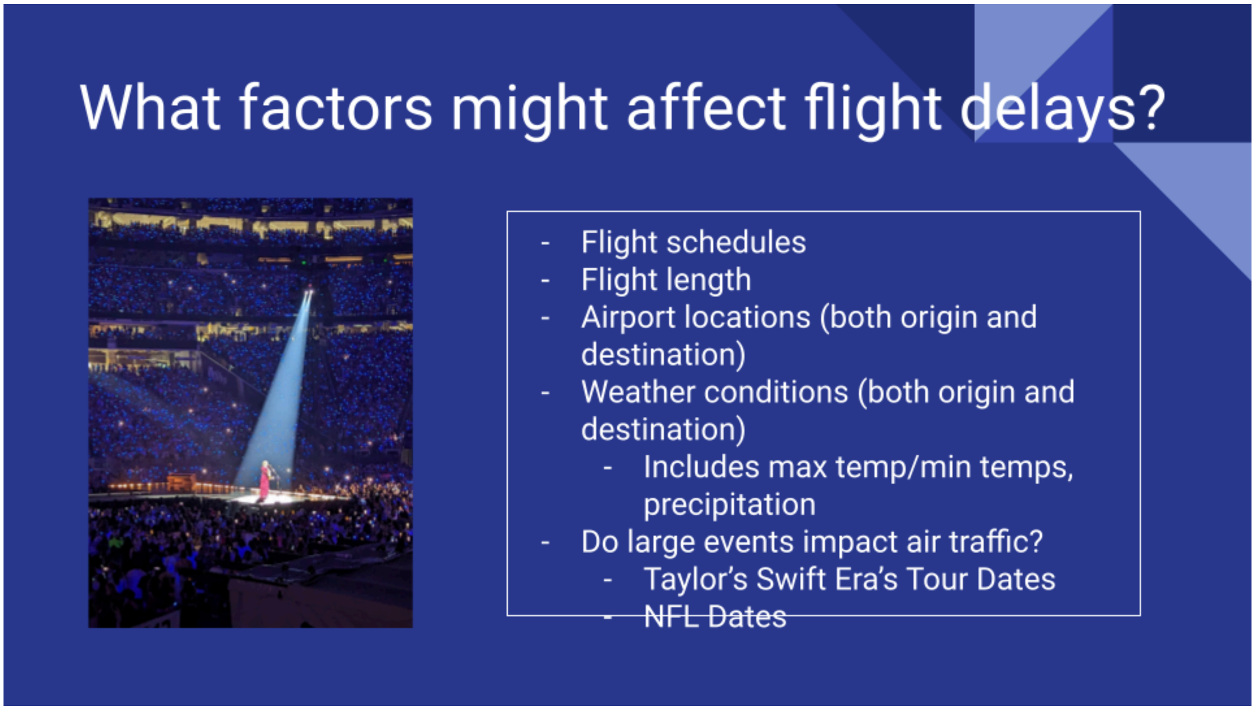
Data Cleaning and Preprocessing

1. **Formatting and Standardization**:
    - Standardized all columns to lowercase and removed extra whitespace in string columns.
    - Ensured consistent datetime formatting for all date and time-related columns.
2. **Handling Missing Data**:
    - **Weather Data**: For weather-related columns with high missing values (e.g., snow, snow depth), we treated null values as zeros unless missing data could not be explained by a lack of significant weather events.
    - **Flight Data**: Removed flights with missing critical delay data, and addressed missing values for airlines with large amounts of missing data (e.g., United Airlines, Southwest Airlines).
    - **Airport Information**: Replaced missing airport characteristics with median values based on state or region.





3. **Outlier Detection and Removal**:
   - Identified outliers in departure and arrival times, particularly for very early or late departures, and removed these to avoid skewing model performance.
4. **Feature Engineering**:
   - Created new features such as `eras_tour` (flights on or near Taylor Swift Era's Tour dates) and `nfl_game`flights on NFL game days), to explore how these events correlate with delays.
   - Created lagged features and aggregate statistics (e.g., average weather conditions at origin/destination, average delays by airline, etc.) to enrich the dataset.



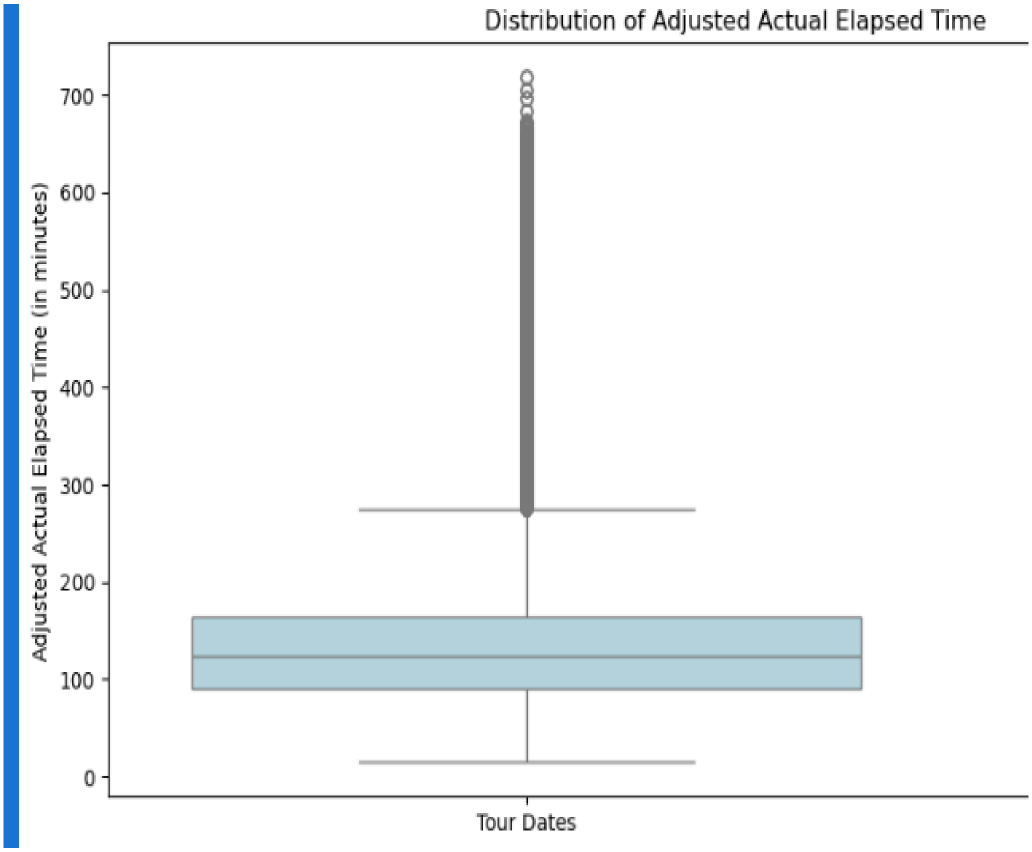
What factors might affect flight delays?

- 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

Exploratory Data Analysis (EDA)

During the EDA phase, several key relationships were explored:

- **Delay Distribution**: We observed a large imbalance in delay data, with many flights experiencing no delay (0 minutes). This created a class imbalance that may affect model performance.
- **Weather Conditions and Delays**: Weather factors like snow, temperature extremes, and precipitation were analyzed to see how they influenced delays.
- **Event Impact**: Flights occurring on NFL game days or during Taylor Swift’s Era’s Tour were shown to have higher delay frequencies.



Visualizations like histograms, boxplots, and heatmaps were used to better understand feature relationships and identify potential issues like multicollinearity.

Key Findings from EDA:

- Not a strong correlation between adverse weather conditions and flight delays, but the state/airport at both the origin and destinations impact delay times and frequencies.
- Event days (e.g., NFL games) can correlate to higher delay frequencies, especially in airports near major stadiums.

Model Development

Given the distribution of delays and the nature of the data, we built and compared several machine learning models:

Models Trained, Delay is Continuous (Regression Problem):

1. **Linear Regression: A simple starting point.**
2. **Random Forest Regressor: Non-linear model that performs well in many cases.**
3. **Gradient Boosting: Powerful algorithms like XGBoost, LightGBM, or CatBoost can often outperform random forests.**

Evaluation Metrics:

- **For regression**: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)
- K-fold cross validation

Model Performance Comparison:

Model	**RMSE**	**MAE**	**Accuracy**	R^2	**Recall**	**F1-Score**
Linear Regression	3.1926	0.70541	N/A	N/A	N/A	N/A
Random Forest Regression	0.00582	0.00291	N/A	.9997	N/A	N/A
Gradient Boosting	0.08038	0.01078	N/A%	0.9983	N/A	N/A

Model Evaluation and Insights

- Linear Regression is underperforming here, likely because the data has complex, non-linear patterns that are not well captured by a simple linear model. Random Forest and XGB perform better because they can capture non-linear relationships. XGB has built in regularization helping it avoid overfitting and giving it more flexibility to model complex patterns. Random Forest Regression performed best. Both MSE and MAE are quite low, indicating that the model is making accurate predictions.

Challenges and Future Work

1. **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.
2. **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.
3. Building out for specific use cases.

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.

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.**