Project Proposal: Predicting Flight Delays

Hope Husemann

Problem Statement: Predict the likelihood or extent of flight delays based on historical flight data, weather conditions, airline/airport characteristics, and potentially external factors such as NFL games. 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.

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.

1. **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.

2. **Criteria for Success:**
    - Develop a machine learning model capable of predicting flight delays with reasonable accuracy.
    - Identify key factors influencing flight delays including weather, airlines, airports, and potentially NFL games (if data allows).

3. **Scope of the solution space:**
    - **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.).

- **OurAirports.com:** This website provides airport data like location, runway length, and passenger traffic, offering insights into potential bottlenecks https://ourairports.com/.
- **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/.
- **Potential Additional Data Source:** Explore NFL schedule data to analyze potential impact of games on nearby airports (e.g., increased congestion).

4. **Constraints within solution space:**
    - Limited to free, accessible data available online

5. **Stakeholders to provide key insight**

6. **Key data sources**
    - **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.).
    - **OurAirports.com:** This website provides airport data like location, runway length, and passenger traffic, offering insights into potential bottlenecks https://ourairports.com/.
    - **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/.
    - **Potential Additional Data Source:** Explore NFL schedule data to analyze potential impact of games on nearby airports (e.g., increased congestion).

Steps to achieve goals using the Data Science Method:

Data Preparation: Clean and preprocess data, handle missing values, encode categorical variables, and perform feature scaling.

- Align datasets based on common attributes, such as Date, Airport
- For weather data, match using date and airport location
- Create new features that combine information from multiple datasets, such as weather conditions at the departure and arrival airports or the operational performance of specific airlines.
- Normalize weather data and encode categorical variables from the airline operations and airport information datasets.
- Address any missing values by imputation or exclusion based on the dataset's context.

Exploratory Data Analysis:

- Create visualizations to explore how weather conditions, airline operations, and airport features correlate with flight delays.
- Analyze correlations between combined features and flight delays to identify influential factors.

Model Development:

- Based on the delay distribution, choose appropriate algorithms
- Train multiple machine learning models, such as:
    1. Regression Models: Linear Regression, Random Forest Regression.
    2. Classification Models: Logistic Regression, Random Forest Classifier, Gradient Boosting.
- Use metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Accuracy, Precision, Recall, F1 Score to evaluate model performance.

Model Deployment and Evaluation:

- Deploy the best performing model on a separate hold-out test data set for final real-world performance evaluation.
- Analyze the model's prediction accuracy and identify potential areas for improvement.
- Explore incorporating additional data sources to further enhance prediction performance.