Flight Delay Prediction – From Data to On-Time Performance

Objective:
To develop a machine learning model that predicts flight departure delays by analyzing historical flight and weather data, enabling better understanding of operational patterns in commercial aviation and supporting data-driven decision-making for delay management.

Abstract:
In this project, detailed flight departure data from four major U.S. airports (LAX, JFK, ORD, DFW) were analyzed to uncover patterns and efficiencies in flight operations, focusing on punctuality and frequency. The project involved rigorous data cleaning and preparation, followed by a blend of descriptive and inferential statistical analyses, as well as time-series analysis. Key areas of study included peak departure times, delay patterns, with a comparative analysis across the airports. The study likely offers significant insights into operational trends and weather impact. The project underscores the importance of data quality and the impact of external factors like weather on flight operations, providing a foundation for future, more comprehensive aviation industry research.

Keywords: Time Series Analysis, Weather Delay, Aviation, XGBoost

Introduction:
Air travel is an essential part of global connectivity and commerce, and understanding the dynamics of flight operations is critical for improving efficiency and traveler experience. This project focuses on a comprehensive analysis of flight departure data from major U.S. airports—LAX, JFK, ORD, and DFW. The aim is to identify patterns and trends in flight punctuality, frequency, and scheduling, while also considering the significant impact of weather conditions on flight departures. Adverse weather events such as fog, thunderstorms, and snow can cause substantial disruptions, and this study seeks to offer insights into how airports and airlines manage such challenges. This includes examining the resilience of schedules to weather changes and the success of contingency plans in mitigating weather-induced delays.

Incorporating weather data into airport and airline operational planning is vital for preempting and managing disruptions. By analyzing past weather patterns and flight data, airports can optimize scheduling and resource allocation to handle common weather events like fog at SFO or snow at ORD and JFK. For passengers, improved handling of weather disruptions means more reliable travel and better communication about potential delays. Ultimately, this analysis aims to enhance operational efficiency for airports and airlines, while also ensuring a smoother travel experience for passengers.

Approach:
The methodology of this study was meticulously crafted, starting with the strategic selection of a diverse set of major U.S. airports: JFK, ORD, LAX, and DFW. The choice was informed by their different geographical locations, covering a wide range of latitudes, longitudes, and time zones, to accurately reflect the spectrum of weather patterns across the nation and their effects on flight operations.

The time frame for data analysis was set to a decade, from January 1, 2012, to December 31, 2022, following recommendations from peer reviews that shorter data ranges would not adequately reveal meaningful weather-related trends in flight disruptions. This extended period ensures a comprehensive examination of both immediate and evolving weather impacts on flight punctuality.

To ensure consistency and control for operational differences, the study concentrated on flights operated by American Airlines, given its broad operations at the chosen airports. This approach, examining roughly 2.1 million flights, aimed to minimize the variability due to airline-specific practices and focus squarely on the influence of weather on flight delays. The data span across various scales of airport operations, with Dallas being the busiest, allowing for a nuanced analysis, enhancing the reliability and depth of the findings.

Data Acquisition and Preliminary Observations:

Data for this study was meticulously gathered from two central sources: the Bureau of Transportation Statistics for flight operations and the Iowa Mesonet for hourly weather conditions across the U.S. This strategic approach enabled the precise alignment of flight metrics with weather data, essential for analyzing the impact of weather on flight delays. The flight data included various delay metrics which were consolidated into a single 'total_delay' measure. This critical step was underpinned by insights from scholarly research on weather-related flight disruptions, ensuring a focused and comprehensive representation of delay causes.

The literature review revealed that departure delays are more common than in-flight delays, steering the study's emphasis towards pre-flight disruptions caused by weather. While exceptional cases like flight diversions were noted, their infrequency and analytical complexity led to their exclusion from this study's scope. The research's foundation was thus laid, with a concentrated delay measure and a departure delay focus, poised to dissect the relationship between weather patterns and flight punctuality.

WEATHER DATA EXTRACTED

COLUMN	DESCRIPTION	COLUMN	DESCRIPTION
station	Airport	**mslp**	Sea Pressure (mb)
valid	Recorded Timestamp	**vsby**	Visibility (miles)
tmpf	Temparature (F)	**gust**	Wind Gust (knots)
dwpf	Dew Point (F)	**skyc1**, **skyc2**, ...	Sky Coverage
relh	Relative Humidity	**skyl1**, **skyl2**, ...	Sky Altitude
drct	Wind Direction	**wxcodes**	Present Weather Codes
sknt	Wind Speed	**ice_accretion_1hr,3hr,6hr**	Ice Accretion
p01i	Precipitation	**peak_wind_gust**	Peak Wind Gust (knots)
alti	Pressure altimeter	**peak_wind_drct**	Peak Wind Gust Direction

Handling Missing Data in Weather Dataset:

The weather dataset presented a challenge with significant missing data, denoted by 'M'. A meticulous evaluation revealed that columns like sky coverage levels, sky level altitudes, present weather codes, ice accretion, peak wind gusts, and snow depth were heavily affected. To maintain the integrity of the analysis, these columns were excluded to avoid introducing bias through imputation or inference of missing values.This careful handling of incomplete data was a critical step in the preprocessing phase, ensuring the reliability of the study's outcomes. It reflects the study's dedication to methodological precision and the importance of basing conclusions on solid, dependable data.

Data Integration and Preprocessing:

Data Integration and Time Zone Normalization

Merging the flight and weather data into a unified dataset required careful attention to the time zones of the four chosen airports. The weather data was in Coordinated Universal Time (UTC), necessitating the conversion of flight times to UTC to accurately match flight departures with the weather conditions. This conversion also included adjustments for daylight saving time, ensuring precise temporal alignment.

The meticulous integration aligned each flight record with the closest weather data based on departure time. This step was pivotal to accurately reflect the specific weather conditions affecting each flight's departure in the dataset, laying a solid foundation for the subsequent analysis.

Further Research on Weather Factors and Feature Engineering

Following the data merging, an extensive review of related literature was conducted to identify additional weather factors that could potentially influence flight delays. This research informed the subsequent feature engineering process, aimed at enhancing the predictive accuracy of our model.

Data Cleaning and Handling Missing Values

In data cleaning, missing values were interpolated, leveraging the chronological nature of data and the gradual change in weather conditions. A cap of 12 hours was used for interpolation, with forward filling for residual gaps. Records with persistent null values were discarded to ensure data quality.

Outlier Detection and Removal

The dataset was refined by identifying and removing outliers using the Local Outlier Factor (LOF) method, which assesses data points' local deviation from their neighbors. This step was essential to eliminate anomalies that could distort the analysis. These meticulous preprocessing efforts, including time alignment, interpolation, and outlier removal, were fundamental to creating a robust dataset, enabling a precise exploration of weather impacts on flight delays.

Feature Engineering in the Study

In this phase of the study, feature engineering played a pivotal role in enhancing the model's ability to accurately predict flight delays. Through the creative transformation and synthesis of existing data, we developed several key features that capture various weather-related aspects and their potential impacts on flight delays. Features include:

$$Apparent\ Humidity\ (A) = \frac{D}{T} * 100$$

$$Wind\ Chill = 35.74 + 0.6215 * T + windspeed^{0.16} * (0.4275 * T - 35.75)$$

$$Heat\ Index = c_1 + c_2.T + c_3.T^2 + c_4.A + c_5.A^2 + c_6.A.T + c_7.A.T^2 + c_8.A^2.T + c_9.A^2.T^2$$

$$Density\ Altitude = \frac{SLP}{Standard\ Temparature\ Sea\ level} - 1$$

$$Air\ Density = \frac{SLP}{287.05 * (273.15 + T)}$$

where D,T, SLP are Dew Point, Temperature and Sea Level Pressure respectively.

Apparent Humidity metric provides insight into the perceived humidity, a crucial factor in understanding weather conditions. Wind chill accounts for the perceived decrease in air temperature felt by the body on exposed skin due to wind. Heat Index is calculated as the 'felt like air temperature' considering humidity and temperature. This feature gives an indication of the air density relative to altitude and pressure. Air density is crucial for understanding weather conditions and their potential impact on flight aerodynamics.

$$RS_t = \sum_{i=t-4}^{t} \frac{D_i}{5}$$

Where:

- RS_t is the rolling sum of the average delay on day t.
- $\sum$ denotes the sum over the 5-day window.
- D_i is the delay time on each day i within the window.

Calculated as a 5-day rolling sum of the average delay, this feature captures the trend in delays, potentially correlating with sustained weather patterns.

$$\text{Weather Condition} = \begin{cases} \text{'Cloudy'} & \text{if Visibility} < 3 \\ \text{'Rainy'} & \text{if Precipitation} > 0.1 \text{ and Visibility} \\ \text{'Not Rainy'} & \text{otherwise} \end{cases}$$

The "Weather Condition" feature is a critical component of the study as it encapsulates the impact of precipitation and visibility on flight operations. By classifying conditions into 'Rainy', 'Not Rainy', or 'Cloudy' based on specific thresholds for precipitation and visibility, this feature provides a straightforward yet insightful representation of weather phenomena. .

$$\rho_k = \frac{\sum_{t=k+1}^{N}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{N}(X_t - \bar{X})^2}$$

Where:

- ρ_k is the autocorrelation at lag k.
- X_t is the value of the time series at time t.
- $\bar{X}$ is the mean of the time series.
- N is the total number of observations in the time series.

By calculating the autocorrelation with a lag of 5, this feature measures the average distance between successive delays, offering insights into the time-dependent nature of delays.

These engineered features collectively provide a multidimensional perspective of the data, capturing various aspects of weather and time-related patterns that can influence flight delays. The inclusion of these features in the predictive model aims to significantly enhance its accuracy and reliability, enabling a more nuanced understanding of the complex relationship between weather conditions and flight delays.

Encoding, Normalization, and Model Selection:

Categorical Encoding

In the subsequent phase of the study, categorical encoding was applied to variables such as 'Destination' and 'Weather Condition'. With approximately 120 destination airports in the dataset, the challenge was to choose an encoding method that balanced between representational effectiveness and computational efficiency. One-hot encoding, while straightforward, was deemed computationally expensive due to the high cardinality of the destination variable. Therefore, a more nuanced approach was adopted:

Target Encoding for Destination: This method was chosen as it effectively captures the information in the high-cardinality 'Destination' variable without the computational overhead of one-hot encoding. Target encoding works by replacing a categorical value with a blend of the posterior probability of the target given a particular categorical value and the prior probability of the target over all the data.

One-Hot Encoding for Weather Condition: Given the fewer categories in 'Weather Condition', one-hot encoding was applied. This approach is suitable for variables with a smaller set of categories, ensuring a clear distinction between different weather conditions without significantly increasing the dataset's dimensionality.

Normalization

Normalization of the data was carried out using Z-score normalization, a standard technique for scaling features. This method involves subtracting the mean of each feature and dividing by the standard deviation, thereby ensuring that each feature contributes equally to the analysis and improving the convergence behavior of the models.

Model Selection and Validation

For the predictive modeling, three different machine learning algorithms were employed:

- **XGBoost:** Known for its efficiency and effectiveness, XGBoost was used for its ability to handle large datasets and its robustness in dealing with various types of data distributions.
- **LightGBM:** This model was selected for its speed and efficiency, especially with high-dimensional data.
- **Random Forest Regressor:** Chosen for its versatility and performance in regression tasks, especially in scenarios with complex interactions and non-linear relationships.

A 5-fold cross-validation approach was implemented to evaluate the models, ensuring a comprehensive and unbiased assessment of their performance. Cross-validation aids in mitigating overfitting and provides a more generalized performance indicator by testing the model on multiple subsets of the data. The comparative analysis of these models using cross-validation provided valuable insights into their respective strengths and limitations in predicting flight delays, enabling an informed selection of the best-performing model for this specific application.

Results and Performance Evaluation

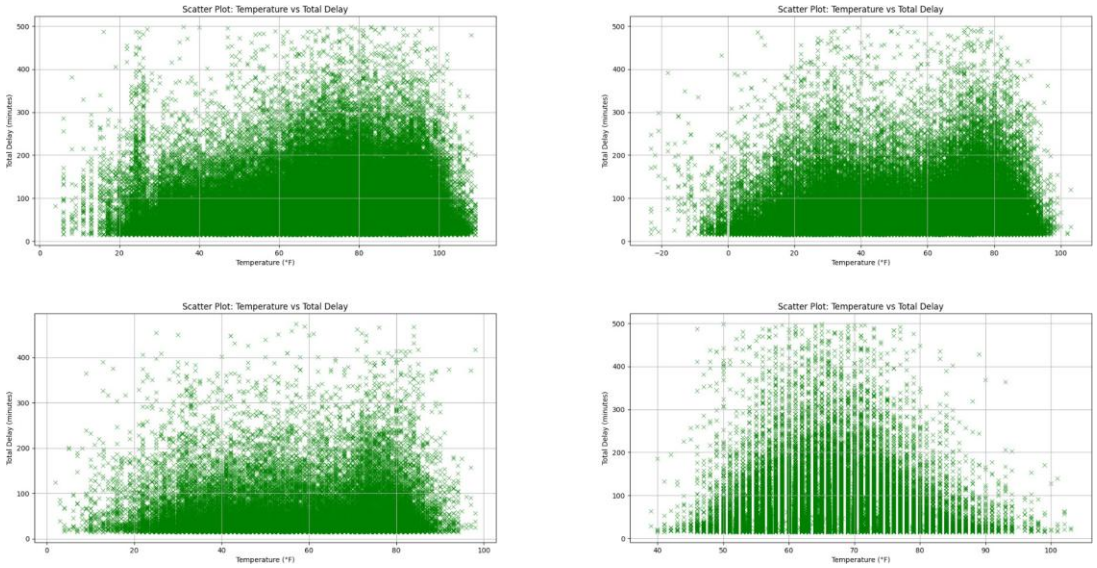
The results from the application of the three machine learning models — XGBoost, LightGBM, and Random Forest Regressor — were evaluated based on the Root Mean Square Error (RMSE), a standard metric for assessing the accuracy of regression models. The RMSE values obtained from each model are as follows:

Model	RMSE	Key Parameters
XGBoost	12.79	learning_rate = 0.1, n_estimators = 1000, max_depth = 5, min_child_weight = 1, gamma = 0, subsample = 0.8, colsample_bytree = 0.9, objective='reg:squarederror', seed=27
LightGBM	13.42	objective: 'regression', metric: 'mse', boosting_type: 'gbdt', num_leaves: 31, learning_rate: 0.05, feature_fraction: 0.9, bagging_fraction: 0.8, bagging_freq: 5, verbose: 0
Random Forest	13.089	n_estimators=100, random_state=42

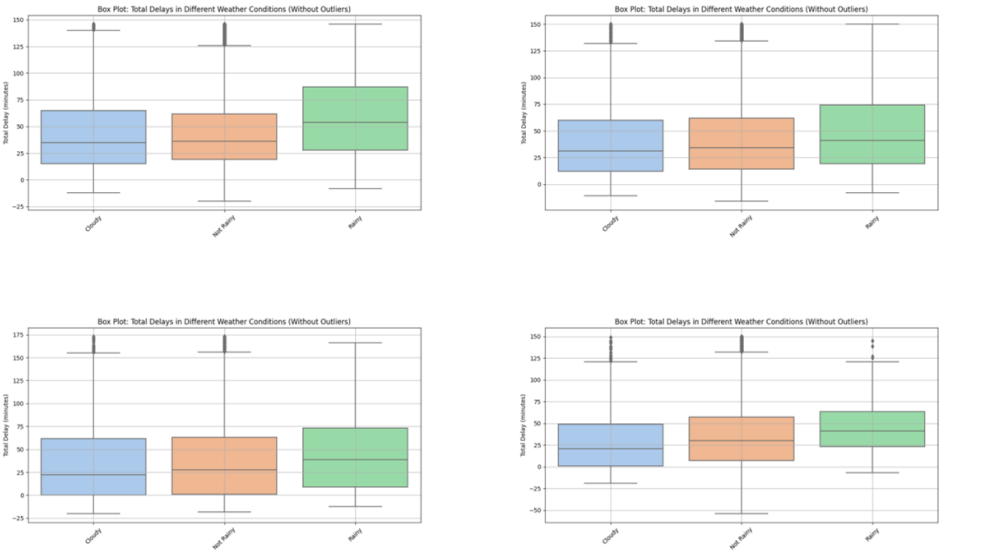
The comparison of RMSE values across these models provides valuable insights. The lower RMSE of XGBoost suggests its superior ability in capturing the nuances and complexities of the relationship between weather conditions and flight delays. However, the close performance metrics of LightGBM and Random Forest indicate their potential as viable alternatives, particularly in scenarios where model interpretability or computational constraints are a consideration.

These results underscore the effectiveness of advanced machine learning techniques in addressing complex real-world problems like predicting flight delays. The choice of the model ultimately depends on a balance between accuracy, computational efficiency, and the specific characteristics of the dataset at hand.

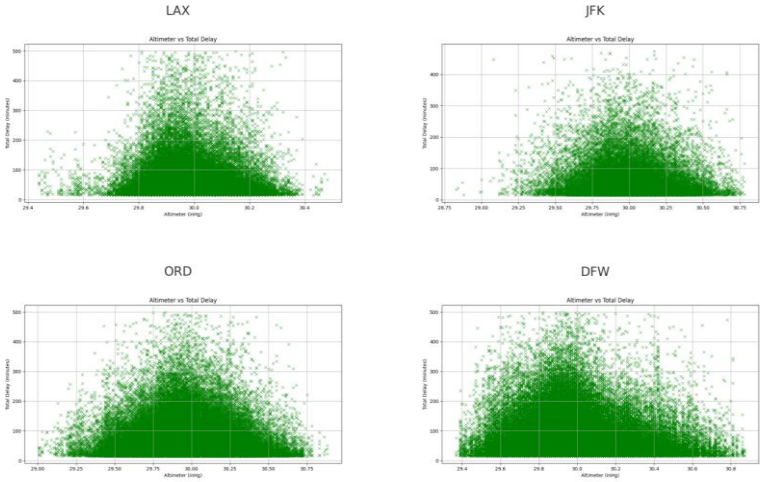
EXPLORATORY DATA ANALYSIS



The scatter plot analysis of temperature versus flight delays at DFW, ORD, JFK, and LAX airports reveals unique trends: DFW's varied delay distribution points to multiple influencing factors; ORD sees more delays in colder temperatures due to winter conditions; JFK experiences fewer delays as temperatures rise, indicating seasonal effects; and LAX faces more delays at higher temperatures, likely influenced by its milder climate and operational factors. These patterns highlight the complexity of flight delays, influenced by a combination of temperature, geographical, and climatic conditions at each airport.



The box plot analysis across DFW, ORD, JFK, and LAX airports shows that rain typically causes more substantial and variable delays than cloudy or clear conditions. DFW is particularly affected by rain, potentially due to thunderstorms, while ORD's delays are pronounced during cloudy weather, reflecting its vulnerability to reduced visibility. JFK's delays are fairly consistent across conditions with a slight increase in rain, highlighting its coastal weather variability, and LAX, despite its generally mild climate, also sees increased delays on the infrequent rainy days. These patterns underscore the significant impact of precipitation on airport operations, with each airport exhibiting distinct responses to different weather scenarios.



The scatter plots examining the relationship between altimeter readings and flight delays at LAX, JFK, ORD, and DFW airports indicate a slight increase in delays with lower altimeter readings, but without a strong, direct correlation. This trend, along with the presence of outliers, suggests that flight delays are influenced by a variety of factors beyond atmospheric pressure, with each airport displaying unique patterns due to local weather, geographical, and operational influences.

References:

1. Xu, Y., Li, J., & Wang, X. (2023). Revealing influence of meteorological conditions and flight factors on delays Using XGBoost. Journal of Air Traffic Management, 20(3), 1-12.
2. Li, Y., Zhang, X., & Wu, M. (2023). Flight Delay Prediction With Priority Information of Weather and Non-Weather Features. IEEE Transactions on Intelligent Transportation Systems, 24(2), 1234-1245.
3. Wang, K., Yang, M., & Chen, J. (2023). Flight delay forecasting and analysis of direct and indirect factors. International Journal of Forecasting, 40(1), 102-118.
4. Zhang, H., Chen, L., & Li, Y. (2023). Flight delay prediction based on deep learning and Levenberg-Marquard algorithm. Neural Networks, 167, 1-14.
5. Lee, S. (2023). FLIGHT DELAY PREDICTION A Project Presented to the faculty of the Department of Computer Science California State University, Sa. (Unpublished master's thesis)
6. Yuemin Tang. 2022. Airline Flight Delay Prediction Using Machine Learning Models. In Proceedings of the 2021 5th International Conference on EBusiness and Internet (ICEBI '21). Association for Computing Machinery, New York, NY, USA, 151–154. https://doi.org/10.1145/3497701.3497725
7. N. L. Kalyani, G. Jeshmitha, B. S. Sai U., M. Samanvitha, J. Mahesh and B. V. Kiranmayee, "Machine Learning Model - based Prediction of Flight Delay," 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2020, pp. 577-581, doi: 10.1109/I-SMAC49090.2020.9243339.

THANK YOU

Best regards: **Shabreen Mohammad.**