Arrival Delay Prediction

1. Project Statement

When planning flights, travellers often rely on complex itineraries involving multiple connections across various airports and carriers. Flight delays and cancellations – affected by a broad range of factors such as time of day, season, weather, and overall traffic – can have significant repercussions, including missed connections and additional costs. Travellers may have strict deadlines or may desire a certain level of confidence in arriving on time.

Key Question:
How can travellers predict the likelihood and duration of flight delays with a specified level of accuracy to meet particular arrival-time requirements?

Data Source:
The U.S. Domestic Flights Delay (2013-2018) dataset from the U.S. Bureau of Transportation Statistics provides extensive flight-level data, including scheduled and actual departure/arrival times, origin and destination airports, airline, flight distance, and delay status. Focusing on flights from 2014 to 2018, this dataset serves as the foundation for a predictive model to estimate both the probability and extent of flight delays.
(Source: Kaggle)

Project Goal:
This project aims to leverage historical data and machine learning techniques to predict the likelihood and magnitude of flight arrival delays. Such predictions can enable travelers to make more informed decisions, potentially adjusting their itineraries and layovers to minimize the risk of disruptions.

2. Data Collection and Cleaning

Data Scope and Features:
The dataset originally spanned 60 files (12 per year) from 2014 to 2018, each with around 500,000 records and 110 columns, totaling approximately 81 GB of raw data. After initial exploration, 20 original features were selected, and 12 additional features were engineered to improve predictive power. Memory optimization steps included assigning appropriate data types and converting object/string columns to categorical variables where possible.

Target Variable:
The chosen dependent variable for the regression task was the Actual Arrival Delay (ActArrDelay). For a separate classification task, the continuous target was discretized into multiple classes (0 to 13), each representing a range of arrival delays

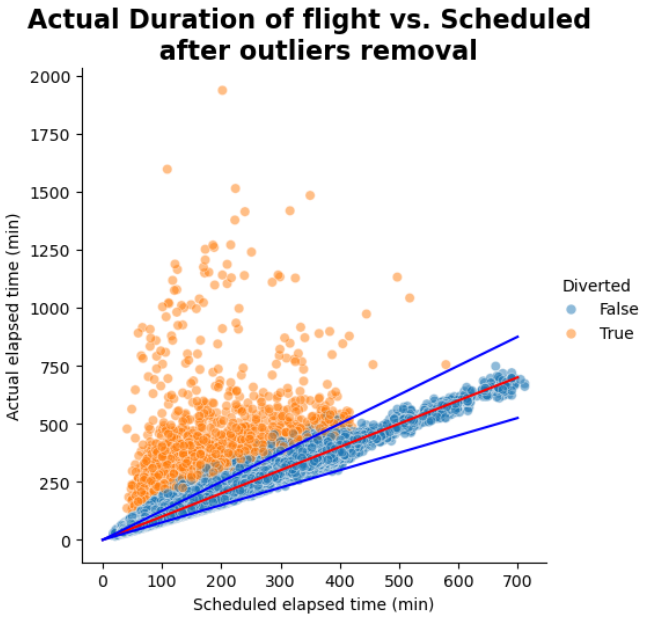
Time and Consistency Adjustments:
To ensure temporal consistency, times – originally recorded in 'hhmm' format – were converted to minutes from the start of the day. Scheduled and actual date/times were calculated or verified based on flight durations and delays. All times were then standardized to UTC, using airport IATA codes, to ensure uniformity across different locations.

Key Lessons in Data Handling:

- **Timezone Management:** Pandas datetime columns can only store data with a single time zone. Mixing zones can force conversions to the inefficient 'object' type, increasing memory usage.

- **Data Types:** To improve memory and processing efficiency, all 'object'-type columns were eliminated, and categorical data types were used where possible.

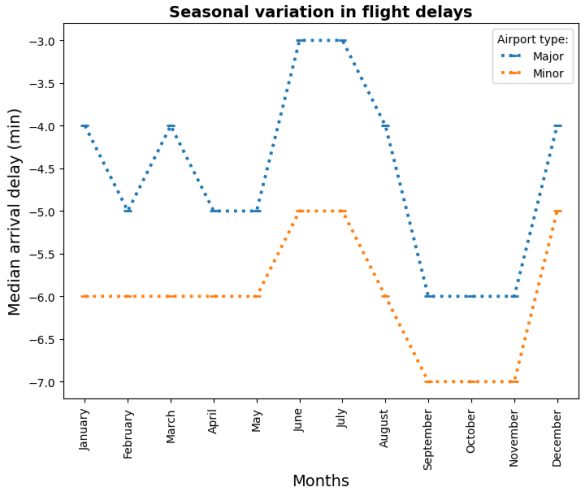
- **Consistency Checks:** Flight times and delays were cross-verified. Records with severe inconsistencies, such as large mismatches between expected and actual flight durations, were removed – fewer than 2% of the total dataset.

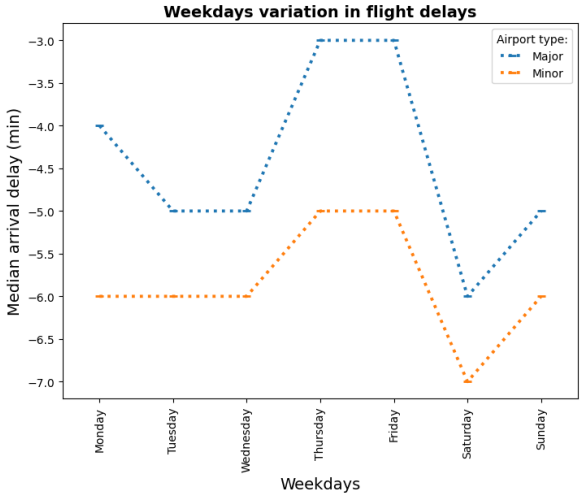


Outcome of Data Cleaning:
After thorough preprocessing, the dataset was reduced from ~81 GB to under 3 GB and stored in a .pickle format. Approximately 30 million records remained, ready for modeling.

3. Exploratory Data Analysis (EDA)

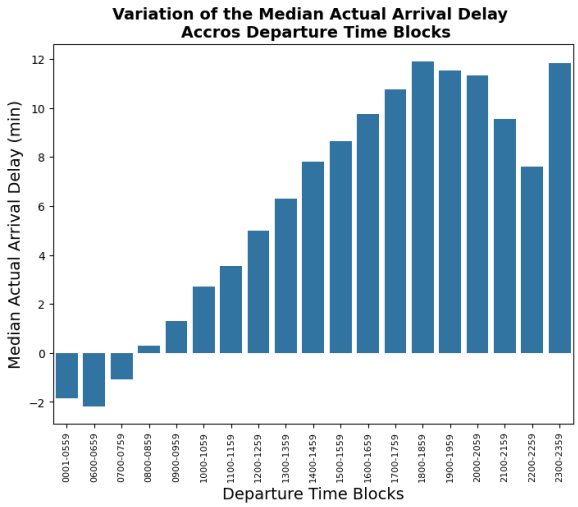
Temporal Patterns:
While the mean Actual Arrival Delays showed visible variations across months and weekdays, statistical hypothesis testing using the Chi-square test indicated no significant differences attributable solely to these time periods. The p-value was 10.8% for major airports and 48.2% for minor airports.

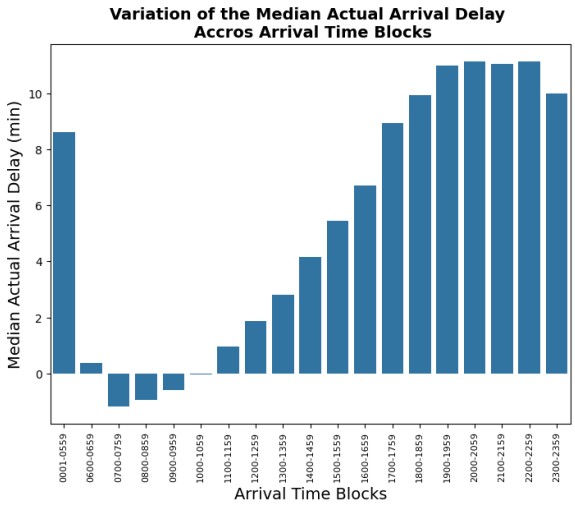




Airport and Airline Effects:
Chi-square tests indicated no statistically significant differences in delay frequencies across different airports and airlines. This suggests that, based on the current analysis, neither the carrier nor the origin/destination airports have a strong or meaningful relationship with the likelihood of delays – all P-values exceeded the 0.05 threshold.

Departure/Arrival Time Blocks:
Both visualizations and Chi-square tests highlighted a strong relationship between Actual Arrival Delay and departure/arrival time blocks. Certain times of day appear more prone to delays, justifying the inclusion of time-related categorical features.





Cancellations and Diversions:
No significant correlation was found between cancellations and specific temporal or location factors. Similarly, diverted flights' elapsed times and delays provided no predictive value, as they are less predictable and often influenced by external, unmodeled factors.

Distribution of the Target Variable:
The Kolmogorov-Smirnov test confirmed that Actual Arrival Delay is right-skewed and non-normally distributed. Attempts at normalizing this distribution (log, Box-Cox transformations) were unsuccessful. Given that all predictive features with a relationship to the target are categorical, it was decided to model the problem using categorical approaches and non-linear methods where appropriate.

4. Data Preprocessing and Feature Engineering

Categorical Feature Handling:
A large number of categorical variables – airports, airlines, month, weekday, time blocks – were encoded into dummy variables. To control the dimensionality explosion, minor airports were grouped into an 'OTHER' category using a cumulative share threshold. This step reduced the number of dummy variables from 728 to 258, significantly improving model efficiency.

Scaling and Standardization:
Since the only numeric feature in the final dataset was the target itself (for regression), and all predictors were categorical, standardization was deemed unnecessary.

5. Training and Testing Split

A 70/30 train-test split was used to ensure a robust evaluation of model performance on unseen data. Additionally, cross-validation was applied within the training set to provide more reliable estimates of model generalization and to guide hyperparameter tuning.

6. Modeling Approach and Potential

Modeling Framework:
The curated dataset and feature sets enabled the implementation of multiple machine learning models to predict arrival delays or classify cancellation likelihood. Both regression and classification techniques were explored:

- **Regression:** Predicting the magnitude of arrival delays.
- **Classification:** Categorizing delays into discrete classes.

Future steps included model selection, hyperparameter tuning, and performance evaluation using metrics suited to each task (e.g., R-squared for regression, F1-macro for classification).

7. Regression Modeling

7.1. Models Used:

- **Linear Regression**: A baseline model to establish a performance benchmark.
- **Regularized Regression (Lasso)**: To handle the large number of features and potentially remove less informative predictors.
- **Lasso with PCA**: Dimensionality reduction using Principal Component Analysis to address feature sparsity and collinearity.
- **Random Forest Regression**: A non-linear model capable of capturing more complex relationships.

7.2. Hyperparameter Tuning and Validation:

For models requiring parameter tuning (Lasso, Random Forest), a two-stage approach was used:

1. **Randomized Search CV** to identify a promising parameter range.
2. **Grid Search CV** to fine-tune the selected hyperparameters.

7.3. Results and Observations:

- **Linear Regression:**

With default parameters, this model served as a baseline. The resulting R-squared was around 0.0173 (training) and 0.0197 (test), indicating that the chosen features explained less than 2% of variance in arrival delays.

- **Lasso Regression:**

Regularization provided only minor improvements over the baseline. Even with tuned alpha values, the R-squared remained low, not substantially improving predictive power.

- **Lasso with PCA:**

PCA reduced the feature set from 330 to 15 components. However, this approach did not improve performance. In fact, results worsened, suggesting that key delay-related signals were not preserved in the principal components or that the underlying relationships are not linear.

- **Random Forest Regression:**

Despite being non-linear and robust to various data complexities, the Random Forest model showed the lowest R-squared scores. Evidence of overfitting (high training R-squared, very low test R-squared) was noted, and tuning did not mitigate this problem.

Conclusion for Regression:
All regression models struggled to explain arrival delay variance, achieving R-squared values below ~2.2%. This result suggests that the selected features – primarily categorical proxies for time, location, and airline – contain minimal predictive information for forecasting actual arrival delays.

8. Classification Modeling

When direct regression proved unsuccessful, an alternative was to treat arrival delay as a classification problem by binning the target into discrete categories. This approach intended to simplify the prediction task into identifying delay "buckets" rather than predicting exact minute values.

8.1. Target Transformation:

The continuous target was discretized into multiple classes (0 to 13), each representing a range of arrival delays. The class distribution was imbalanced, necessitating careful choice of metrics such as F1-macro to ensure fair evaluation across all classes.

8.2. Models Used:

- **K-Nearest Neighbors (KNN)**: A simple, instance-based classifier.
- **Random Forest Classifier**: A powerful ensemble method often effective in handling categorical and imbalanced datasets.

8.3. Hyperparameter Tuning:

For both models, Randomized Search CV was employed to find optimal hyperparameters (e.g., number of neighbors in KNN, number of estimators and maximum depth in Random Forest).

8.4. Results and Observations:

- **KNN Classifier:**

The F1-macro scores were extremely low (~0.0740 on training and 0.0564 on test), indicating almost no predictive power.

- **Random Forest Classifier:**

Despite expectations, the Random Forest Classifier also failed to produce meaningful discrimination among classes. F1-macro scores remained below 5%, showing the model's inability to leverage the provided features for accurate classification.

Conclusion for Classification:

The classification approach did not yield significant improvements. Both KNN and Random Forest classifiers performed poorly, suggesting that discretizing arrival delays did not help uncover predictive patterns in the chosen features.

9. Summary and Reflections

9.1. Objectives and Methods:

This project set out to predict flight arrival delays using historical U.S. domestic flight data. Data cleaning, extensive verification of time zones, and feature engineering were performed to ensure data quality. Both regression and classification frameworks were explored using a variety of machine learning models.

9.2. Model Evaluations:

- **Regression Models (Linear, Lasso, Lasso with PCA, Random Forest):**

Achieved R-squared scores no higher than ~2.1%, indicating negligible predictive capability.

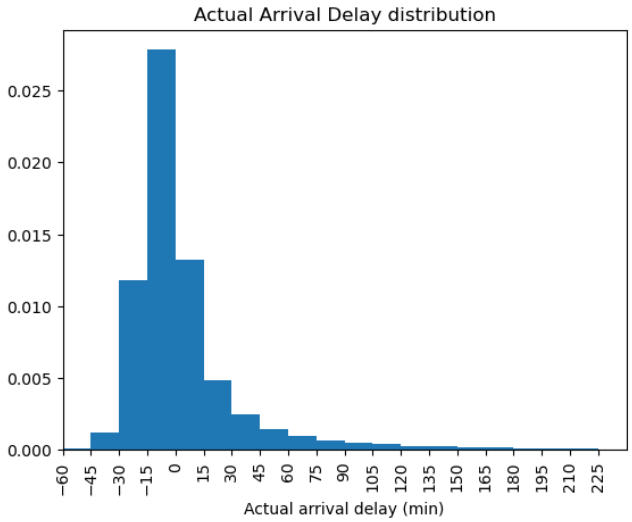
- **Classification Models (KNN, Random Forest):**

Produced extremely low F1-macro scores, showing minimal ability to categorize delays effectively.

Model Type	Parameters	Train Score	Test Score
Regression modelling			
Linear Regression		R^2: 0.0251	R^2: 0.0197
Regularized Regression (Lasso)	Alpha: 0.17	R^2: 0.0220	R^2: 0.0207
Lasso with PCA Reduction	15 PCA factors, Alpha: 0.000001	R^2: 0.0121	R^2: 0.0122
Random Forest Regression	150 estimators, Max Depth: 3	R^2: 0.0111	R^2: 0.0051
Classification modelling			
K-Neighbors Classification	19 neighbors	F1-macro: 0.073963	F1-macro: 0.056446
Random Forest Classifier	50 estimators, Max Depth: 20, Criterion: 'gini'	F1-macro: 0.049368	F1-macro: 0.037726

9.3. Insights and Limitations:

The consistently poor results indicate that the selected features—limited to temporal, airline, and airport categorical variables—lack the necessary predictive signals to accurately forecast delays. Flight delays are often driven by factors absent from the dataset, such as real-time weather conditions, mechanical issues, air traffic control restrictions, and crew scheduling constraints. Under these circumstances, the best possible predictions may rely on the distribution of actual arrival delays:



9.4. Future Directions:

To improve model performance, consider incorporating additional explanatory variables:

- **Weather Data:** Historical weather conditions at origin and destination airports.
- **Aircraft and Maintenance Information:** Aircraft type, maintenance schedules, or fleet utilization rates.
- **Real-Time Air Traffic Data:** Information about current air traffic control hold times or runway availability.

While these enhancements could potentially improve the model's predictive power, they also introduce significant challenges for practical usage. Incorporating such data would require access to specialized, often real-time information, which may not be available to an ordinary user planning a flight. This added complexity could limit the model's applicability to general audiences.

Additionally, experimenting with more sophisticated modeling techniques, such as gradient boosting methods, neural networks, or specialized time-series analyses, could help. However, even advanced models may struggle without richer and more relevant features, and the increased computational and data requirements may further complicate their use.

10. Conclusion

This project demonstrated a comprehensive end-to-end process – from data collection and cleaning, through exploratory data analysis, to model training and evaluation. Despite rigorous efforts, the predictive models failed to explain or accurately categorize flight arrival delays using the given features. The lack of meaningful predictive power underscores the complexity of airline operations and the multifaceted nature of delays. Future research should focus on acquiring more relevant and detailed data sources to provide the predictive insights travelers and stakeholders need.