Predicting Flight Arrival Delay

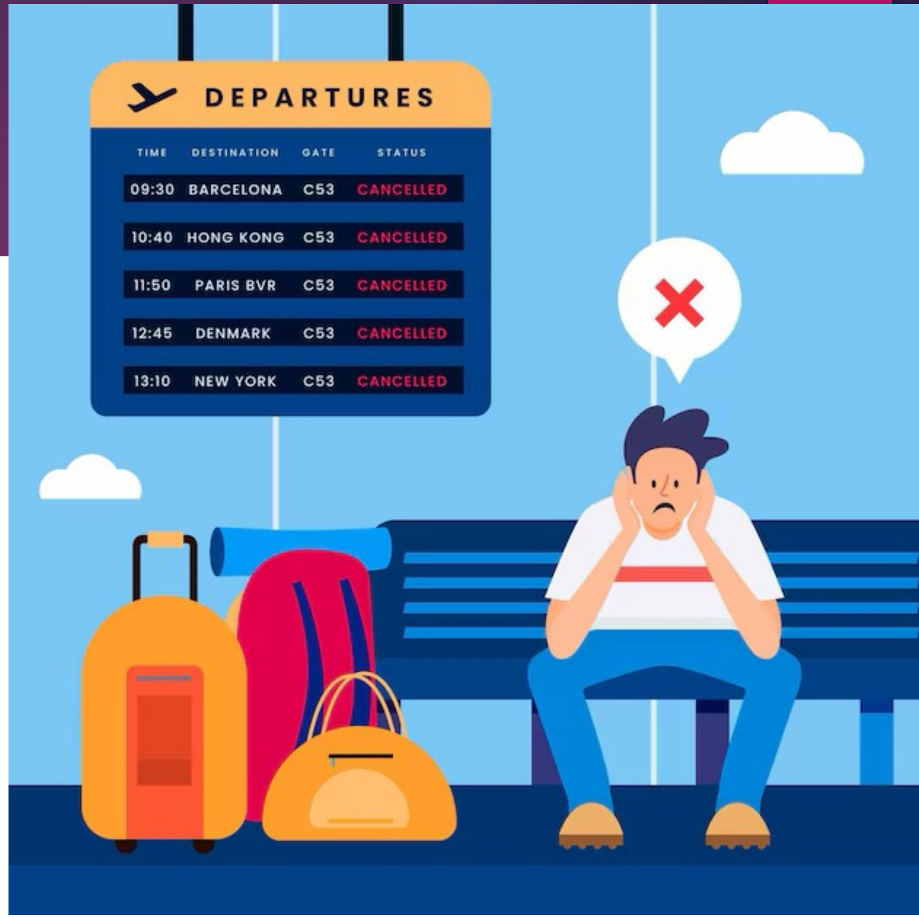
A DATA-DRIVEN APPOACH TO ENCHANCE TRAVEL PLANNING

Alexey Kholodov

13 December, 2024

Project Overview

- **Problem Statement:** Flight delays impact travel plans, causing missed connections and financial costs.
- **Key Question:** How can travellers predict the likelihood and duration of flight delays with sufficient accuracy?
- **Data Source:** U.S. Domestic Flights Delay (2014-2018) dataset from the U.S. Bureau of Transportation Statistics.
- **Goal:** Build predictive models to estimate the likelihood and extent of delays.



Data Overview

Key Features:

- Flight month, Weekday
- Departure and Arrival Time Block
- Airports of origin and destination
- Airlines

US Domestic Flight from 2014-2018

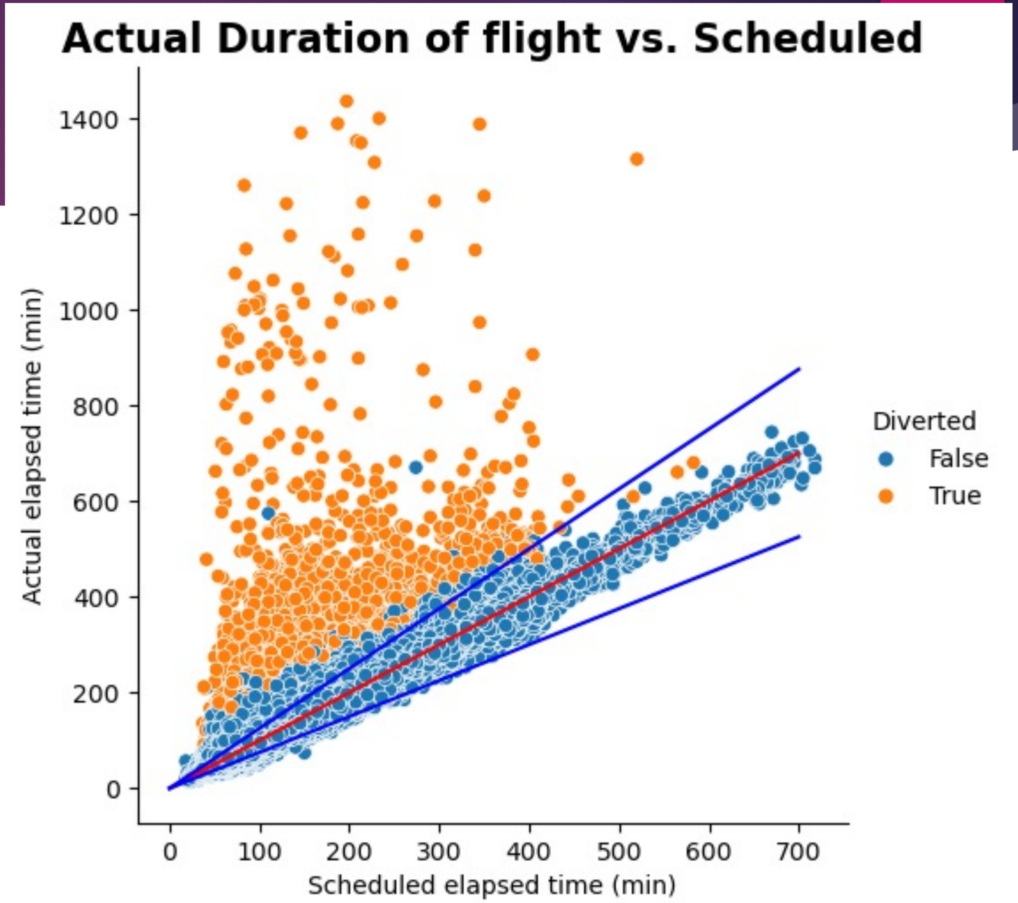
30 million records and 110 features

20 features selected
12 features engineered

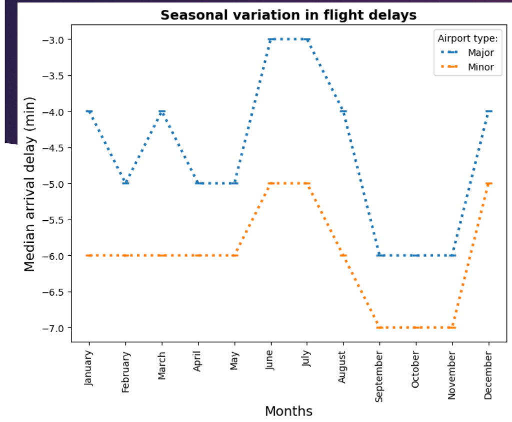
Dateset reduced from 81 GB to under 3 GB

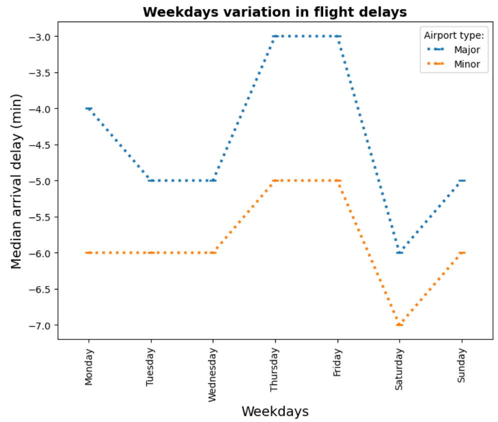
Data Cleaning

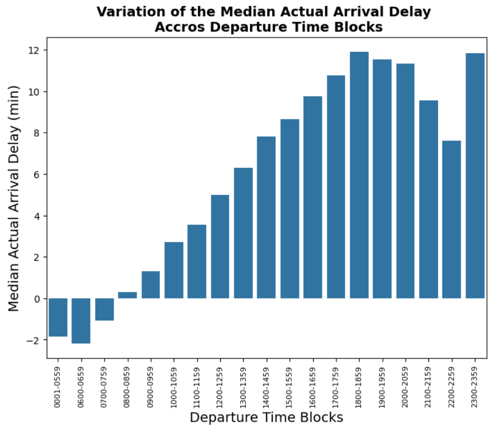
- Addressed data inconsistencies:
  - Time zone adjustments and standardization to UTC.
  - Eliminated severe mismatches and outliers (<2% of data).
- Optimized memory usage
  - Converted object data types to datetime or categorical

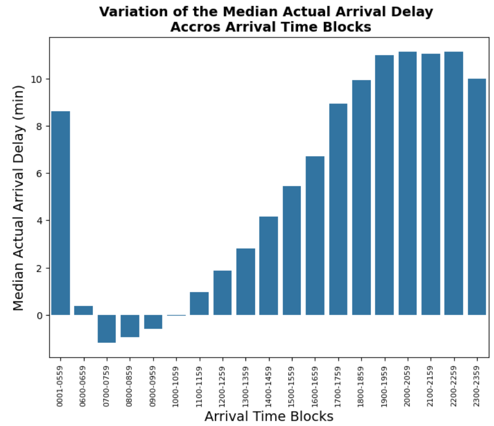


Exploratoty Data Analysis









Insights:

- Graphs indicated a relationship between delays and months, weekdays, and departure/arrival time blocks.
- Statistical significance was confirmed only for the relationship between delays and departure/arrival time blocks.

Non-significant findings:

- No significant delay patterns by airline or airport.
- Diversions and cancellations were unpredictable.

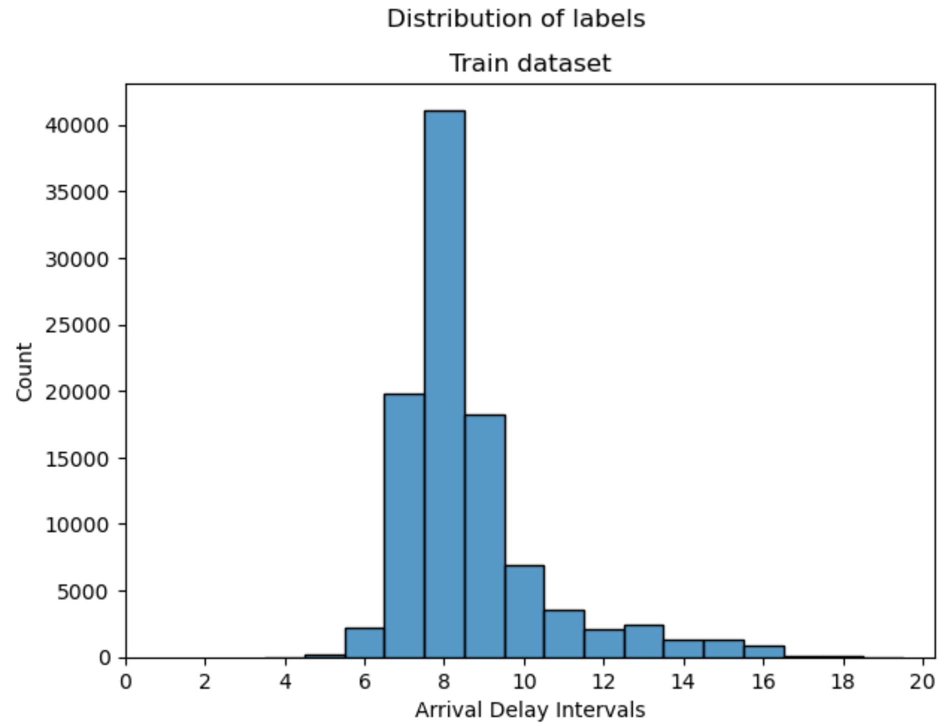
Target Variable

Regression Task: Predict actual arrival delay (minutes).

Classification Task: Group delays into 14 classes (0 to 13).

Challenges:

- Target variable was right-skewed and non-normal.
- Imbalanced class distribution for classification.



Modeling Approach

Regression Models:

- Linear Regression
- Lasso
- PCA with Lasso
- Random Forest.

Evaluation: R-squared

Classification Models:

- K-Nearest Neighbors (KNN)
- Random Forest Classifier.

Evaluation: F1-macro

Data spliting:

- Testing data – 30%
- 5 folds

Hyperparameters tunning:

- Random Search Cross-Validation
- Greed Search Cross-Validation

Model Performance (Regression)

Linear Regression:

- Training R-squared: 0.0251.
- Test R-squared: 0.0197.

Lasso Regression:

- Alpha: 0.17.
- Training R-squared: 0.0220.
- Test R-squared: 0.0207.

PCA with Lasso:

- 15 PCA factors, Alpha: 0.000001.
- Training R-squared: 0.0121.
- Test R-squared: 0.0122.

Random Forest Regression:

- Number of Estimators: 150, Max Depth: 3.
- Training R-squared: 0.0111.
- Test R-squared: 0.0051.

Conclusion: Minimal predictive signal in selected features

Model Performance (Classification)

KNN:

- Number of Neighbors: 19.
- Training F1-macro: 0.073963.
- Test F1-macro: 0.056446.

Random Forest:

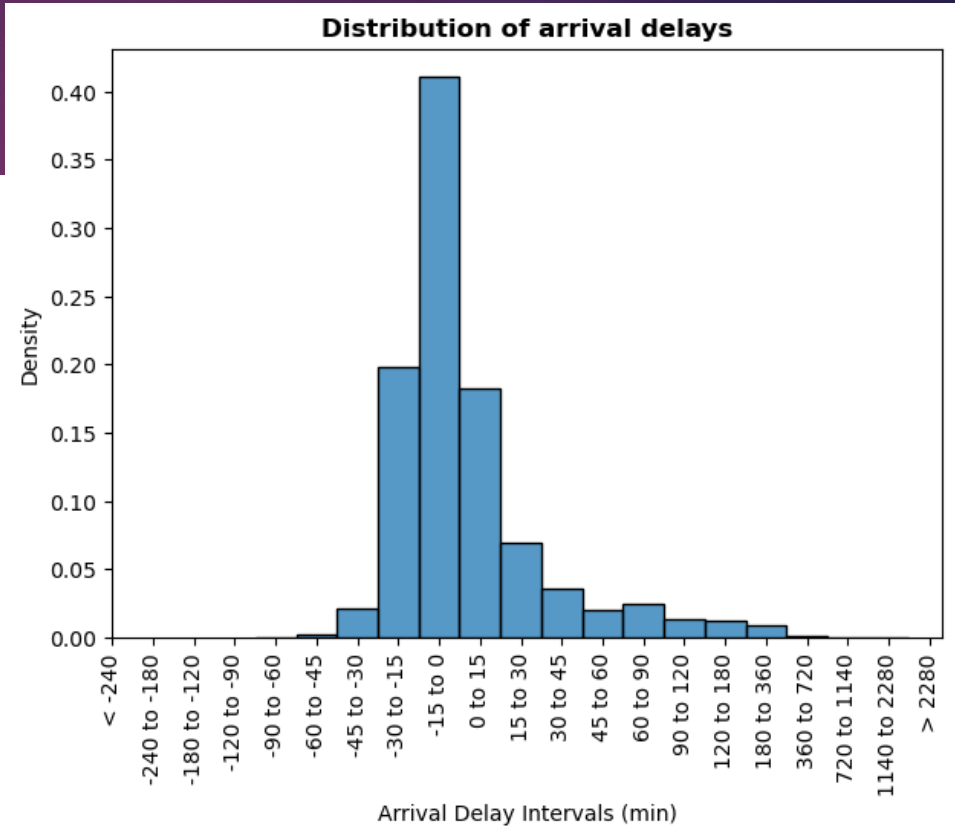
- Number of Estimators: 50, Max Depth: 20, Criterion: 'gini'.
- Training F1-macro: 0.049368.
- Test F1-macro: 0.037726.

Conclusion: Minimal predictive signal in selected features

Key Challenges

- Predictive features lacked strong relationships with target.
- Significant external factors (e.g., weather, maintenance) were missing.
- Imbalanced class distribution limited classification effectiveness.

Under these circumstances, arrival delay predictions can be estimated using the actual distribution of arrival delays.



Lessons Learned

- Data quality and feature selection are critical for predictive power.
- Understanding domain-specific complexities (e.g., airline operations) is essential.
- Modeling frameworks must align with data characteristics.

Future Directions

Enhance Data:

- Incorporate weather, maintenance, and air traffic data to improve predictions, though this may limit the model's applicability to business or professional users.
- Add real-time variables for dynamic predictions, further limiting applicability and potentially shortening the forecast period.

Advanced Techniques:

- Explore Gradient Boosting, Neural Networks, or Time-Series Models to improve performance.
- Mitigate class imbalance using oversampling or cost-sensitive learning techniques.

Conclusions

Summary:

- Predicting flight delays is complex due to multifaceted influences.
- Current models showed limited success due to data constraints.

Impact:

- Insights inform data acquisition and methodology for future projects.

Thank you!

Contact information:

Alexey Kholodov

Email: a.kholodov@me.com

LinkedIn: linkedin.com/in/a-kholodov