Project Report: Data Analytics with Tools

Mika Pärssinen

February 10, 2025

Abstract

This study investigates flight delay prediction using machine learning techniques applied to the 2008 American Airlines dataset. By using Apache Spark for big data processing and linear regression modeling, I developed a predictive model achieving an R^2 value of 0.9739 and RMSE of 17.45 minutes. The analysis revealed that departure delays, taxi times, and ground operations are the most significant predictors of total flight delays, while weather and carrier-specific factors showed lesser influence. The model demonstrated strong performance for shorter delays, with 50% of predictions having errors under 8.39 minutes, though accuracy decreased for longer delays. Temporal analysis identified peak delay periods, with highest averages occurring on Day 5 and during December. These findings provide valuable insights for airlines and passengers, contributing to better delay prediction and management. The implementation methodology and results suggest potential for practical applications in flight scheduling and operations management, while also highlighting areas for future improvement in handling extreme delay cases and incorporating additional predictive factors.

1 Introduction and Problem Formulation

Air travel is one of the most popular and efficient form of transportation for long distances. It not only enables people to explore new destinations and experiences but also keep the connection between families, friends, and communities. Millions of passengers rely on airplanes every day to transport them safely and efficiently from one location to another.

However, flight delays are a continuos issue that disrupt this otherwise reliable form of travel. From the traveler's perspective, delays can be inconvenient and frustrating, leading to missed connections, appointments or events. For airlines, delays result in significant economic consequences, including increased operational costs, dissatisfied customers, and damage on their reputation. Airplane companies estimates that these losses counts to billions of dollars annually. Fixing flight delays would not only fix passengers frustration but also save airlines substantial financial resources, contributing to a more efficient and customer friendly air travel experience.

The aim of this project is to analyze historical flight data from American Airlines 2008 dataset, and then develop predictive models to predict flight delays using machine learning techniques. By identifying key factors that contribute to delays, such as departure times, distances, and weather conditions, this research seeks to provide insights for mitigating delays. Ultimately, the goal is to enhance both efficiency for airlines and convenience for passengers.

2 Implementation

This project was implemented by using Apache Spark, a distributed computing framework which is great for processing large scale data. The implementation followed a systematic approach comprising several key stages of data preprocessing, model development, and analysis.

2.1 Data Preprocessing

The initial preprocessing stage involved critical steps to ensure data quality and consistency. I began by removing columns that i thought was unnecessary, additionally i removed cancelled flights to focus on actual delay patterns, as these represent anomalies rather than predictable delays. Missing values in key fields including arrival time, elapsed time, and delay related columns were filled by a 0. I combined different types of delays into one "TotalDelay" feature to better predict and analyze flight delays in my model.

2.2 Feature Selection and Engineering

Based on correlation analysis and domain knowledge, i selected six key features for my model:

- Departure Delay
- Distance
- Taxi-In Time
- Taxi-Out Time
- Carrier Delay
- Weather Delay

These features were chosen after a analysis of their correlations with arrival delays, where we found particularly strong relationships with taxi-out time (correlation: 0.322) and moderate relationships with taxi-in time (correlation: 0.123). Interestingly, distance showed only a weak positive correlation (0.007), challenging some common assumptions about flight delays.

2.3 Model Development

The dataset was split using an 80-20 ratio, with 80% allocated for training and 20% for testing and evaluation. I implemented a Linear Regression model using PySpark's ML library, with features assembled into vectors using VectorAssembler. The model's hyperparameters was tuned using the CrossValidator with a 3-fold cross-validation approach. The hyperparameter optimization explored different combinations of:

- RegParam: [0.01, 0.1, 1.0]
- ElasticNetParam: [0.0, 0.5, 1.0]

The final model utilized the optimal parameters of RegParam = 0.01 and ElasticNetParam = 0.0, as determined by the grid search optimization process.

2.4 Initial Data Analysis

My preliminary analysis revealed several important patterns in flight delays:

2.4.1 Temporal Patterns

Day-of-week analysis showed highest average delays on Day 5 (10.95 minutes), while monthly analysis revealed December as having the highest delays (16.68 minutes). These patterns suggest significant temporal variations in delay probabilities that could be valuable for both airlines and passengers in planning.

2.4.2 Carrier Performance

Significant variations were observed among carriers, with average delays ranging from -2.89 to 12.61 minutes. This variance indicates that carrier-specific factors play a substantial role in delay patterns, though these differences might also reflect variations in route networks and operational environments. The implementation used Spark's distributed computing capabilities to efficiently process and analyze the large dataset, while maintaining the flexibility to handle various data quality issues and model optimization requirements. This approach provided a robust foundation for my subsequent evaluation and analysis phases.

3 Evaluation

The evaluation of my flight delay prediction model revealed several important insights about its performance and reliability. The model achieved strong overall performance metrics, with an R^2 value of 0.9739 indicating that it explains approximately 97.39% of the variance in flight delays. The Root Mean Square Error (RMSE) of 17.45 minutes and Mean Absolute Error (MAE) of 11.88 minutes suggest reasonable prediction accuracy given the natural difference in flight delays.

Analysis of feature importance revealed that departure delay was the most significant predictor of total delay, with a coefficient of 2.88. This was followed by taxi-out time (1.41) and taxi-in time (1.27), indicating that ground operations have a substantial impact on overall delays. Weather delays (0.15) and carrier delays (0.10) showed relatively smaller influences, while distance surprisingly had a minimal negative impact (-0.004). The residual analysis showed great results with a mean residual close to zero (0.03), indicating unbiased predictions. The standard deviation of residuals was 17.45 minutes, with 50% of predictions having errors less than 8.39 minutes. Notably, 90% of predictions had errors less than 25.74 minutes, and only 1% of predictions exceeded errors of 61.03 minutes. The model's performance varied across different delay ranges, showing better accuracy for shorter delays:

- Very Low Delays: Mean absolute error of 8.95 minutes (SD: 8.20)
- Low Delays: Mean absolute error of 8.01 minutes (SD: 6.85)
- Medium Delays: Mean absolute error of 8.93 minutes (SD: 7.61)
- High Delays: Mean absolute error of 14.45 minutes (SD: 12.20)
- Very High Delays: Mean absolute error of 19.53 minutes (SD: 20.16)

Visual assessment through actual vs. predicted delay plots showed a strong linear relationship, with points closely following the prediction line. The residual distribution appeared approximately normal, centered near zero, supporting our linear regression assumptions. However, the presence of some outliers and increasing spread of residuals for larger delays suggests potential areas for model refinement, particularly in handling extreme cases.

4 Future Work

This study has revealed several promising areas for future research and improvement.

- **Enhanced Feature Engineering:** Incorporating additional features such as detailed weather data, airport congestion metrics, and seasonal patterns could improve prediction accuracy, particularly for extreme delays.
- **Advanced Modeling Techniques:** Exploring non-linear models or ensemble methods might better capture the complex relationships in flight delays, especially for longer delay predictions where the current models accuracy is worse.
- **Real-time Prediction:** Developing a system for real-time delay prediction by incorporating live data feeds would increase the practical utility of the model for both airlines and passengers.

These improvements could enhance the model's practical applicability and provide more comprehensive insights into flight delay patterns and prevention strategies.

5 Discussion

The results of the flight delay prediction model reveal several interesting patterns and insights about the nature of flight delays. The high R^2 value of 0.9739 suggests that the selected features effectively capture the majority of factors influencing flight delays. However, this strong performance must be considered alongside some notable limitations and observations. The model's varying performance across different delay ranges is particularly noteworthy. Its higher accuracy for shorter delays (mean absolute error of 8-9 minutes) compared to longer delays (mean absolute error of 19.53 minutes for very high delays) suggests that extreme delays may involve complex, interconnected factors not fully captured by the current feature set. This pattern aligns with industry observations that major delays often result from compound effects of multiple factors. The feature importance analysis revealed some surprising findings. While the strong influence of departure delays and taxi times was expected, the minimal impact of distance on delays challenges common assumptions. This suggests that operational factors at airports, rather than flight duration, play a more crucial role in determining delays. Additionally, the relatively low importance of weather delays in our model might indicate that weather impacts are often captured indirectly through other features like departure delays and taxi times.

6 Conclusion

This study has demonstrated the effectiveness of machine learning techniques in predicting flight delays using big data analytics. The linear regression model, implemented using Apache Spark, achieved high accuracy with an R^2 value of 0.9739 and RMSE of 17.45 minutes, providing reliable delay predictions for most scenarios. The analysis highlighted the critical role of departure delays and ground operations in determining total flight delays, while challenging some common assumptions about factors like flight distance. The model's great performance in predicting shorter delays, combined with its decreased accuracy for extreme delays, suggests both the potential and limitations of current predictive approaches. These findings provide valuable insights for airlines and airport operators in identifying and addressing key delay factors, particularly in ground operations and departure scheduling. While the model shows strong practical potential, opportunities remain for enhancement through additional features and advanced modeling techniques. The methodology and findings presented here establish a foundation for future work in flight delay prediction and management, contributing to the ongoing improvement of air travel efficiency and reliability.

7 References