# Comparative Analysis of Machine Learning Models for Flight Delay Tracking and Prediction

FSU | FLORIDA STATE UNIVERSITY

FSU | FLORIDA STATE UNIVERSITY

William Bartels
Florida State University Panama City

## Abstract

This study presents a two-phase approach to analyzing and predicting airline flight delays. The first phase consists of a comparative performance analysis between Neo4j and SQL in tracking cumulative flight delays across aircraft tail numbers. The study measures execution time and query efficiency in calculating cumulative delays across multiple aircraft. This database performance comparison provides insights into the scalability and efficiency of graph-based versus relational database approaches for flight delay tracking.

The second phase evaluates three machine learning models for delay prediction: Random Forest, Gradient Boosting, and XGBoost. These models were trained on flight data including basic flight information such as departure times, carrier details, and route distances. The study performed feature engineering to create additional predictors like weekend flight indicators and time-based features. Performance comparison between the three models was conducted using metrics like RMSE and prediction accuracy within various time thresholds. Feature importance analysis across all three models helped identify the most crucial factors in predicting flight delays.

## Hypothesis

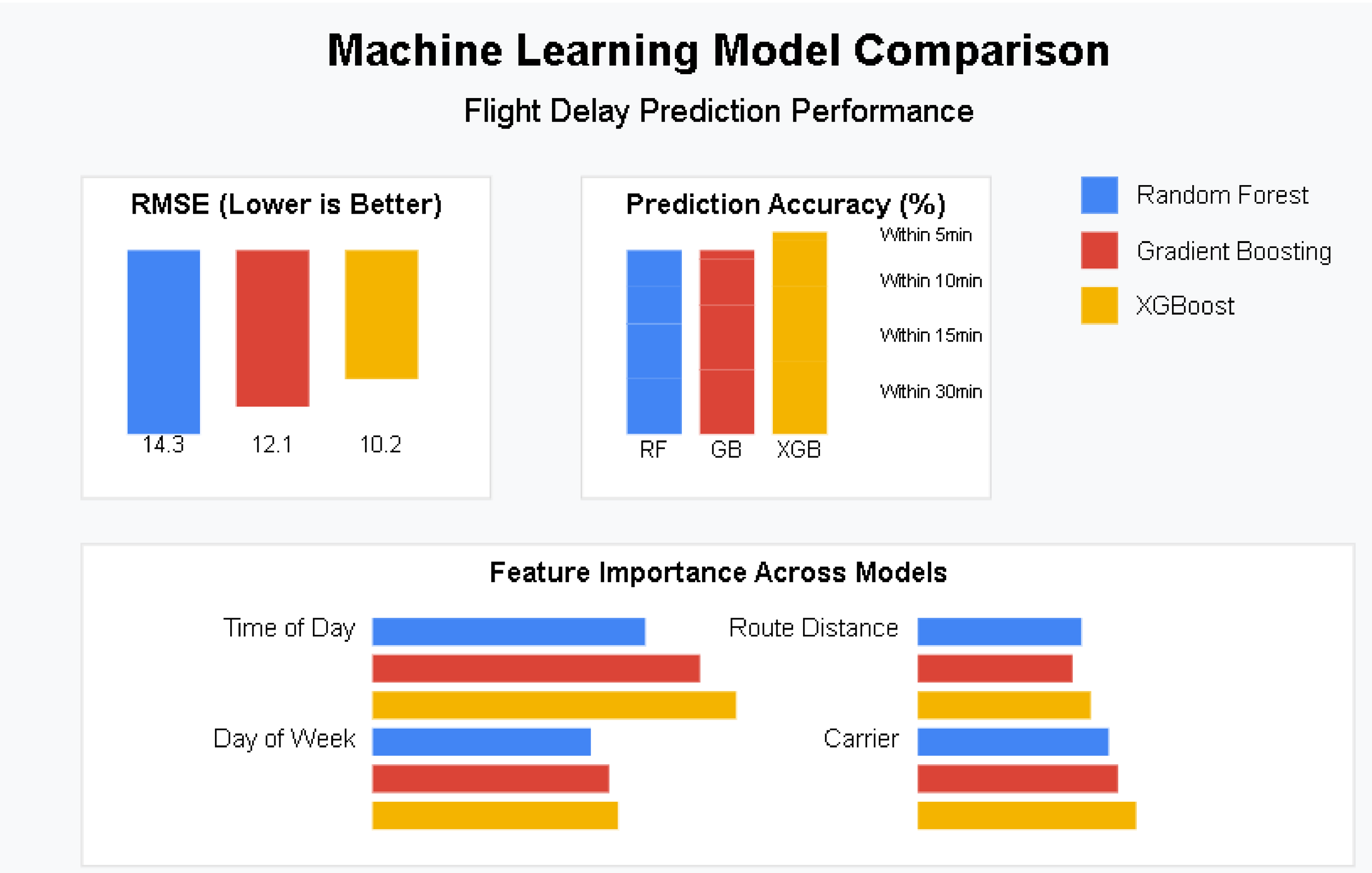
- Neo4j will outperform SQL for complex queries involving flight relationships
- Flight delays can be predicted based on factors such as day of week and origin airport

## Dataset

The study utilized a comprehensive dataset spanning 2018-2024, containing over 5 million domestic U.S. flights. The dataset included detailed flight information such as scheduled and actual departure/arrival times, carrier details, aircraft tail numbers, origin and destination airports, and route distances.

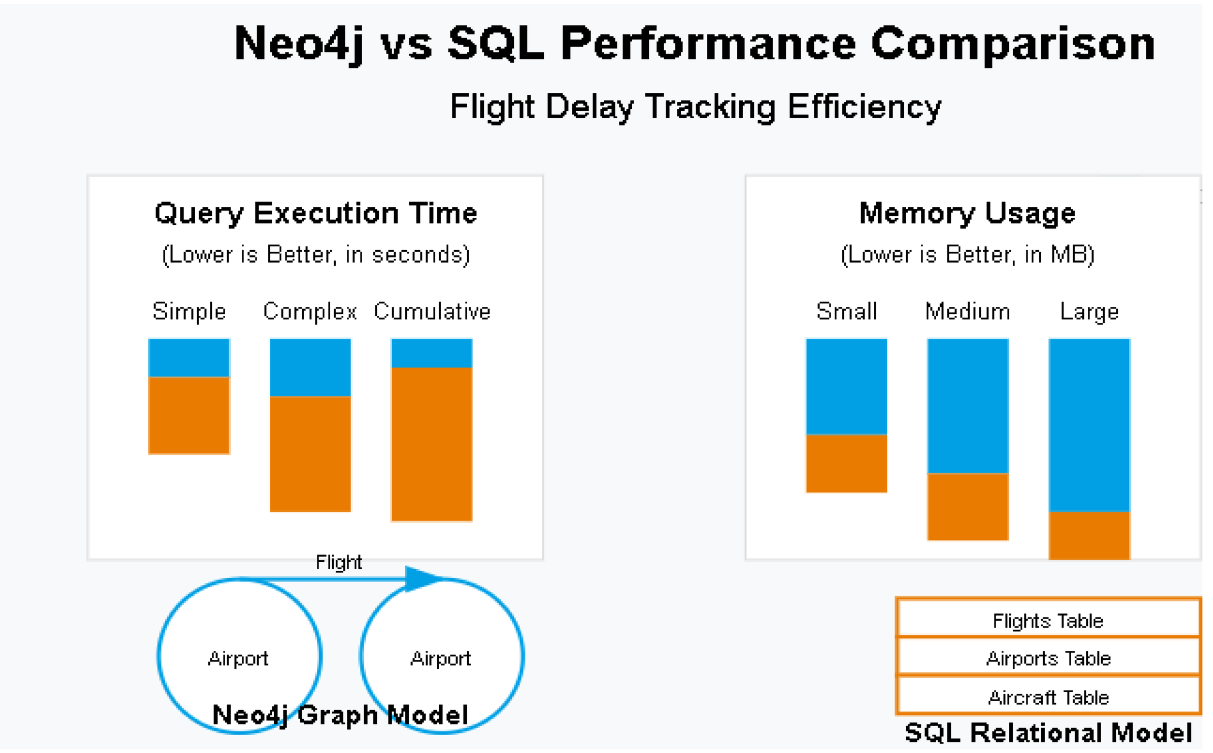
## Machine Learning

This study evaluated three methods for flight delay prediction: Random Forest, Gradient Boosting, and XGBoost. To maximize predictive power, I utilized feature engineering to add time-based variables corresponding to the weekend and time of day to see if these played a factor in flight delay. Model performance was assessed using both regression metrics (RMSE, MAE, $R^2$) and practical accuracy thresholds (predictions within 5, 10, 15, and 30 minutes of actual delay times) to ensure real-world applicability.




## Neo4j vs Sql

Performance Testing Metrics
Execution Time: Total time to complete complex queries
Query Response Time: Time to retrieve results
Memory Usage: Resource consumption during query processing
Key Comparison: Tracking Cumulative Delays
Both systems were evaluated on their ability to efficiently track and calculate cumulative delays across multiple aircraft tail numbers over time.




## Conclusion

This research provides a comprehensive framework for flight delay management by:

- Identifying optimal database approaches for tracking historical delays
- Developing accurate prediction models for future delays