Predicting Flight Delays: A Comprehensive Analysis of US Flight Performance Data

Xiaodan Lu, Hongxin Wu, Yunhong Yang, Xinran Zhang
Georgetown University

Abstract

The airline industry faces the challenge of providing timely services, especially during the holiday season when passenger traffic peaks. In this study, we aim to predict flight delays in December 2022, using data from the US Transportation Department's performance dataset for on-time, delayed, or canceled flights. We employ machine learning techniques to identify variables contributing to delays and explore patterns in flight performance across airlines, days of the week, and city pairs. Complementary to our predictive modeling, we use Tableau to create dashboards and other visualization tools to facilitate a comprehensive understanding of the data for both technical and non-technical audiences. Our project offers insights that can help airlines optimize their operations, enhance the travel experience for passengers, and promote efficient resource allocation during the busy holiday season.

Introduction

Flight delays and cancellations cause significant inconvenience for passengers and disrupt the smooth functioning of the airline industry, particularly during the busy holiday season.

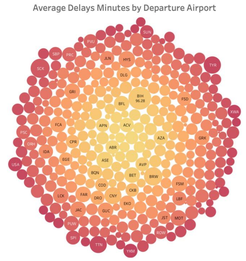
Objective: Our project aims to predict flight delays in December 2022 using machine learning models and real data from the US Transportation Department's performance dataset. We will identify key variables that influence flight performance and create informative data visualizations to provide valuable insights into the airline industry.

Dataset & Exploratory Analysis

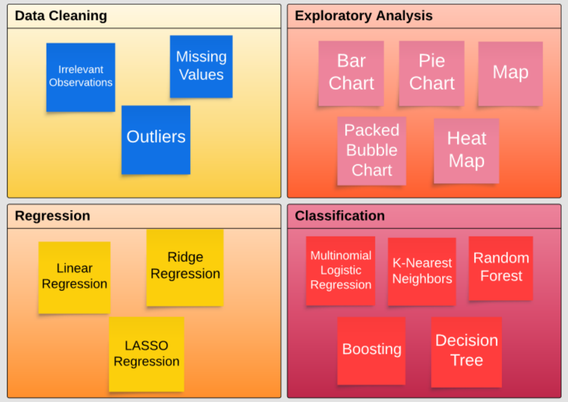
Source: On-Time Reporting Carrier On-Time Performance (1987–present) from the US Department of Transportation (Dimensions: 576,827 x 48)

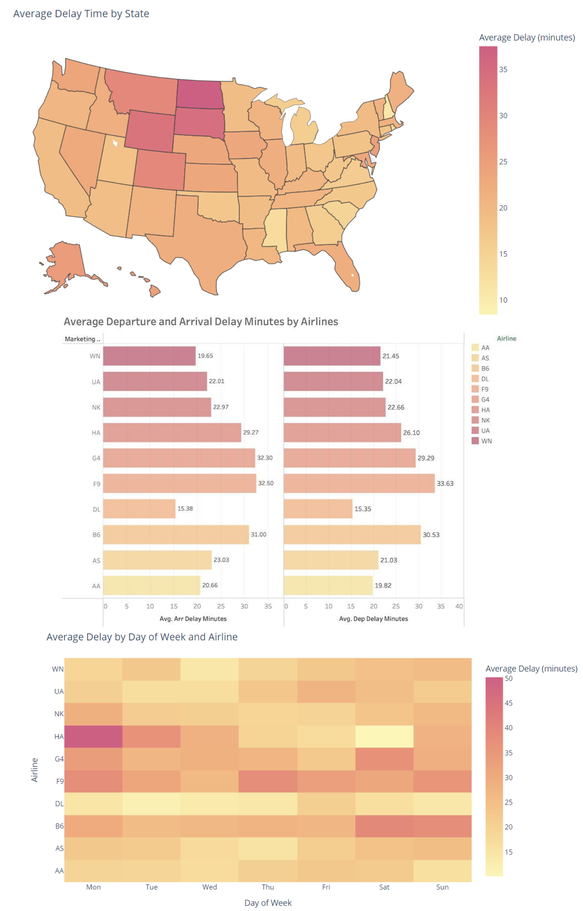
Dataset Overview:

576,827	355	20.81 min	20.89 min	5.385%
Flights	Destinations	Avg. Departure Delay	Avg. Arrival Delay	Cancellations



Methodology





Modeling & Results

The dataset is divided based on the day of the month for training and testing purposes. Data collected between December 1, 2022, and December 25, 2022, constitutes the training set, while data from December 26, 2022, to December 31, 2022, serves as the testing set. Key variables utilized for model training include "DayOfWeek", "Marketing_Airline_Network", "TaxiOut", "WheelsOff", "WheelsOn", "TaxiIn", "Distance", and others.

Variable Name	Description
DepDelayMinutes	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DayOfWeek	Day of Week (i.e.: 01, 02, ...)
Marketing_Airline_Network	Unique Marketing Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2).
TaxiIn	Taxi In Time, in Minutes
TaxiOut	Taxi Out Time, in Minutes
WheelsOn	Wheels On Time (local time: hhmm)
WheelsOff	Wheels Off Time (local time: hhmm)
CRSDepTime	CRS Departure Time (local time: hhmm)
Distance	Distance between airports (miles)

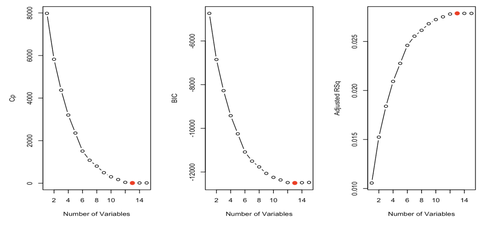


Fig. Output of Best Selection Subset

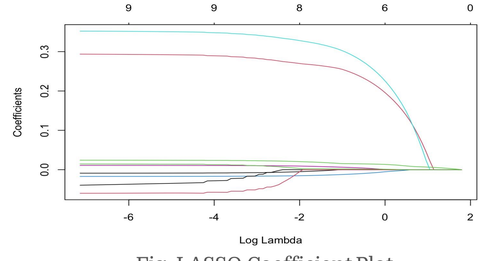
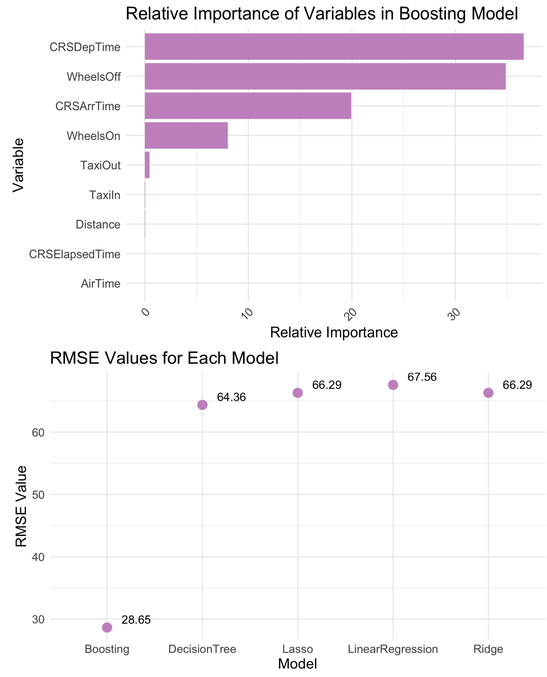


Fig. LASSO Coefficient Plot



Conclusion

- Boosting provides a significantly better method for the prediction.
- Estimated departure time and arrival time contribute to a major factor of flights delay.
- The map plot reveals more flight delays in the Midwest than the East Coast, suggesting a possible link between regional economic development and flight delays.

Future work

- Dataset is too large for local computer to run. Use cloud or computer with more GPU to train the model
- Include more factor variables such as departure airport and arrival airport.
- Many factor variables are highly unbalance, Resampling and Class weight adjustment may apply to improve the models.

References

- "Bureau of Transportation Statistics." *United States Department of Transportation,* https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvzr.