Data-Driven Prediction of Flight Delay Duration

Carnegie Mellon University

By: Ghada Abdelhady, Vaibhav Khandelwal, Isabella Salas-Allende, Kamron Soldozy, Ezra Sutter, Jack Waters

Background & Introduction

- Our globalized world is highly dependent on air travel for a variety of purposes
- Consequently, flight delays have significant socio-economic costs, especially unpredictable delays
- Song et al. (2024) report that flight delays severely impacted passenger satisfaction nationwide
- Accurate predictions of delays could help airlines optimize operations to improve passenger satisfaction and minimize economic costs

We aim to predict the length of flight delays using data collected by the United States Bureau of Transportation Statistics on commercial airline flights.

Data Distribution and Analysis

- The original dataset contained 34,314 flights originating in either Dallas (DFW) or Chicago (ORD), of which we retain 15,054 flights which were delayed (58% from ORD, 42% DFW)
- We are interested in predicting the logarithm of a flight's arrival delay at its destination **(Fig 1)**, using 22 predictor variables (5 categorical, 17 quantitative)

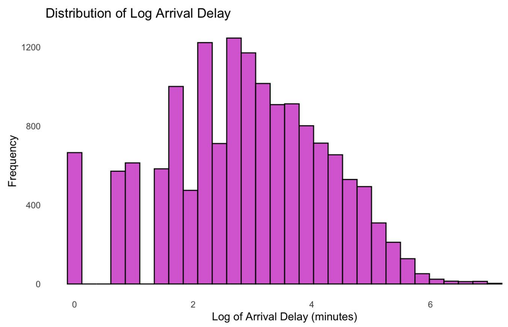


Figure 1: Distribution of the response variable

- Initial qualitative analyses revealed that the delay is greater when ORD is the origin airport **(Fig 2a)**
- Mean arrival delays vary between airline carriers, with Frontier airlines (F9, **Fig2b**) most delayed and Alaskan Airlines (AS, **Fig 2b**) least delayed
- Arrival delays were similar by destination, but notably higher for Rhode Island and Wyoming, possibly due to fewer flights to these destinations skewing the numbers **(Fig 2c)**

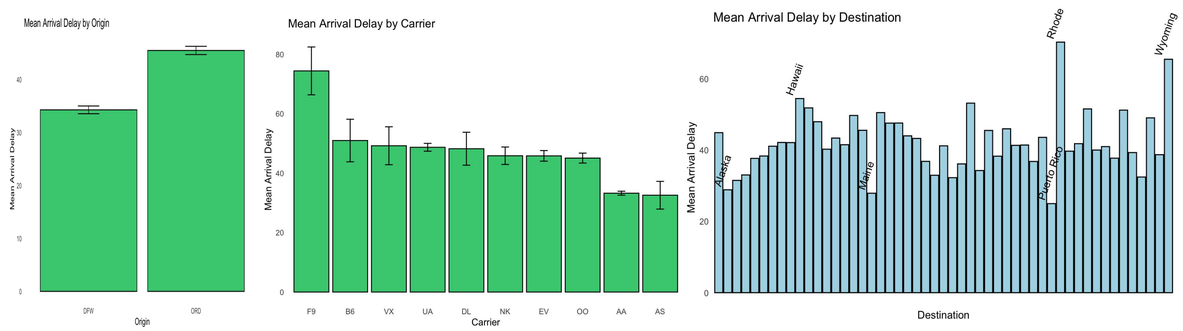
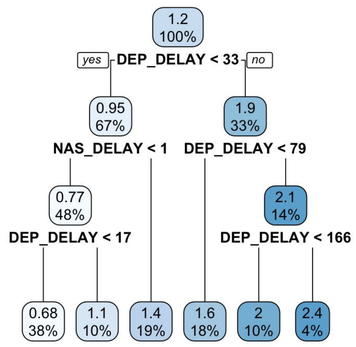
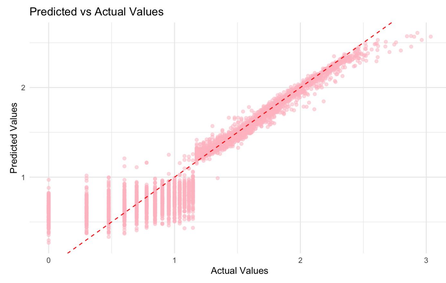


Figure 2: Arrival delay by origin airport (left), by airline carrier (center), and by destination (right)

Predictive Modeling

We used a random subset of 70% of our data to train four models **(Fig 3-5)** to predict the duration of flight delays and tested each of them on the same 30% of withheld data: regression tree, multiple linear regression (with and without variable selection), and random forest





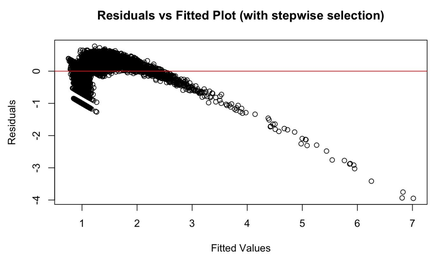


Figure 3: Residuals vs fitted values for multiple linear regression with best variable selection. R^2 = 0.5582, indicating poor model fit.

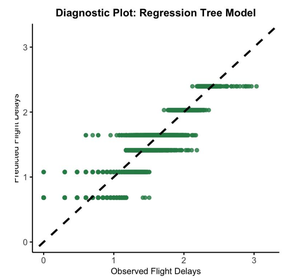


Figure 4. Structure of learned regression tree model (top) and diagnostic plot of model predictions against observations (bottom)

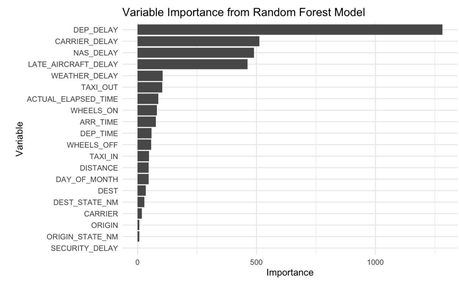


Figure 5: Model predictions against observations for the random forest (top) and the relative importance of each predictor to model accuracy (bottom)

Conclusion

Model	Test MSE
Regression tree	0.070
Linear Regression	0.149
Best Subset Linear Regression	0.147
Random Forest	0.042

Table 1: Test set MSE for our 4 predictive models

- We were able to reliably predict the duration of delay of flights with low test set MSE. Random forest RMSE was 0.206, equating to *1.2 minutes.*
- Our diagnostic plots suggest that the relationship between our response variable and predictors may be nonlinear
- Intuitively, and based on Figure 5, using the departure delay to predict arrival delay is likely the reason behind our model's great predictive ability
- Future work should try to predict delay using variables other than delay

References

Song, C., Ma, X., Ardizzone, C. & Zhuang, J. The adverse impact of flight delays on passenger satisfaction: An innovative prediction model utilizing wide & deep learning. *J. Air Transp. Manag.* (2024)

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L. & Bärnighausen, T. Regression Discontinuity Designs in Epidemiology. *Epidemiology* (2014)