Big Data Energy 2020 TAMIDS Competition

Johnathan Lo & Isaac Ke

3/28/20

Contents

Chapter 1

Introduction

Reliable transportation supports a strong economy by facilitating the rapid and timely exchange of goods and services and bolstering tourism revenue. In the US, the transportation industry accounts for XXX billion dollars per year, which is XXX% of GDP [cite]. Of that economic product, XXX% is accounted for by the airline industry [cite]. A key metric for evaluating the efficiency of airline industry production is flight delay time. In 2018, flight delays led to an economic loss of XXX billion dollars[cite]. For individual companies, delays can influence consumer choice, and for the industry itself, unmitigated delays can impel consumers to switch to substitute goods, such as automotive or rail-based transport.

Therefore, a major goal of this project is to analyze flight delays and diagnose areas for improvement. We intend to create models using publicly available data that can accurately predict future delays. In doing so, we can hopefully uncover significant and controllable covariates that can help guide airline companies to reduce flight delays.

Chapter 2

Executive Summary

2.1 Problem and approach

2.2 Data preprocessing

2.3 Exploratory analysis

2.4 Model formulation

2.5 Model selection

2.6 Applications and conclusions

Chapter 3

Motivation, data description, and software

3.1 Motivation

As stated in the introduction, flight delays can have a wide-ranging effect on the economy. Most airline companies have already done everything in their power to mitigate and reduce delays. We are interested in finding whether delays can be further mitigated, and whether those variables can be controlled by airline companies. To the extent that some delays are unavoidable or difficult to predict, we are also interested in devising methods to minimize the impact of those delays, whether by reducing the number of passengers affected, offering alternate routes to affected passengers, or discounting tickets. Overall, for the benefit of airline companies, consumers, and society-at-large, we should minimize flight delays, or the impact thereof.

3.2 Data collection

Our data was provided to us as csv files by the competition organizers. The primary dataset was composed of over 10,000,000 observations of 50 variables. Each observation was a distinct flight that occurred between 1/1/2018 and XX/XX/2019, and the 70 covariates included origin, destination, quarter, arrival delay, departure delay, distance, and many more variables pertaining to each flight. An auxiliary dataset included pricing data given for each route, by quarter.

In addition to the data provided to us by the competition organizers, we also sought out additional data to enhance our dataset. We obtained geographic coordinates for each

airport from XXXXX [cite], and weather data from NOAA databases through the NCDC API [cite]. The geographic coordinates are given in decimal format, and our weather data describes meteorological events near the origin and destination of each flight. Importantly, data *along* the flight path was not obtained, due to time constraints and complexity. A full list of covariates along with brief descriptions can be found in Supplementary Table 1.

3.3 Software

All analyses were performed in R v3.6.3 [cite]. Packages used include, but are not limited to, *ggplot2*, *dplyr*, *caret*, *rnoaa*, and *Isaac put stuff here*. Individual datasets were loaded as *data.frame* objects and combined using *merge*. The final dataset can be found as a csv file in Supplementary Data 1.

Chapter 4

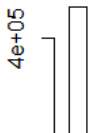
Exploratory data analysis

4.1 Data wrangling

Our dataset was drawn from 4 different sources - flight delays and fare data, provided to us by the competition organizers, geographic coordinates from XXXXX, and weather data from NOAA. Flight delays and geographic coordinates were combined by merging on common origin and destination names. The resulting dataframe was then combined with fare data by common routes, year, and quarter. Adding weather data was more challenging in that the observations relate information collected by weather stations, and not the airports themselves. Thus, weather station coordinates were cross referenced with airport coordinates to find the closest active weather station to each airport. Due to this constraint, 10 airports, corresponding to XXXXX observations were dropped, due to the lack of NOAA weather stations nearby. Weather data was then merged with the rest of the data on common dates and airports, with separate variables for weather at origin and weather at destination.

4.2 Distribution of flight delays

A histogram of all arrival delays is shown in Fig 1. Clearly, the data is strongly right-skewed. To correct the skewness, a cube-root transformation was performed, but subsequent Shapiro-Wilk test provided strong evidence against normality for this transformation, so it was abandoned. To evaluate the



Chapter 5

Model formulation

Chapter 6

Model selection

Chapter 7

Forecasting Flight Delays for 2019 Q3

Chapter 8

Business recommendations

Chapter 9

Closing thoughts

Chapter 10

Appendix

10.1 References

10.2 Additional figures, tables, and data

Bibliography

[Mangel u. Clark 1988] Mangel, Marc ; Clark, Colin W.: *Dynamic Modeling in Behavioral Ecology*. Princeton, New Jersey : Princeton University Press, 1988

[Sandholm 2010] Sandholm, William H.: *Population Games and Evolutionary Dynamics*. Cambridge, Massachussetts : The MIT Press, 2010

[Sarah P. Otto 2007] Sarah P. Otto, Troy D.: *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution*. Princeton University Press, 2007. – ISBN 0691123446,9780691123448

[Sigmund 1993] Sigmund, Karl: *Games of Life: Explorations in Ecology, Evolution and Behavior*. Dover Edition. Mineola, New York : Dover Publications, 1993, 2017

[Smith 1982] Smith, John M.: *Evolution and the Theory of Games*. Oxford, Great Britain : Cambridge University Press, 1982