Manipulate Flight Delay in an Efficient way by Predictions with Machine Learning

A Preprint

Mengchen (Veronique) Wang
mw5ew
mw5ew@virginia.edu

Sara Feng
sjf8yqz
sjf8yqz@virginia.edu

Minjun (Elena) Long
ml6vq
ml6vq@virginia.edu

February 16, 2020

Abstract

As flight delay is a concerning problem nowadays, it would be beneficial for both travelers and airports if patterns and predictions of these delays could be revealed. Effective predictions would not only help travelers to better prepare for the delays, but also help airports to control and reduce chained delays which is caused by the inefficient arrangement of airport facilities. Thus, we decide to work on a dataset from U.S. Department of Transportation's Bureau of Transportation Statistics, using Regression in Machine Learning to model, compare models, and ultimately find a good way to predict future flight delays. This can benefit travelers and airports, including Virginia airports and its residents.

***K*eywords** Regression · Transportation · Virginia · Machine Learning · Cross validation

1 Motivation

Nowadays, while air travel becomes a common transportation, constant flight delays seem to be a problem. Delays are especially annoying since people do not expect or well-prepare for it, which usually ends up with wasting hours in the airports. Delays usually are caused by several factors, including weather, previous flights, crew issues, and air traffic control, etc. Many of these factors could be utilized to find out a pattern of delayed fights and let airports predict the departure delay for future flights. For example, delay time could be reduced if we plan ahead and control air traffic in an efficient way. These predictions also benefit travelers, as they could prepare for it accordingly; for example, bring additional entertainment devices. We believe accurate predictions on future flights would benefit the community, including Virginia and its residents. With these information ahead, travelers could be better prepared before they come to the airport and adjust their schedule accordingly. Besides, airports in Virginia could also use these predictions to better prepare for chained delays. Virginia has several airports, including Dulles International Airport (IAD), Norfolk International Airport (ORF), Richmond International Airport (RIC), and Charlottesville Albemarle Airport (CHO), etc. With predictions, they could prepare snacks and drinks for customers and also arrange boarding gates, runways, shuttle bus, and jet bridges in a more efficient way. Lastly, with these information, airports could also reduce delay time since they could assign jet bridge, control air traffic, and even adjust flights efficiently. The result of this application would benefits not only residents but also airport staffs.

2 Dataset

URL: https://www.kaggle.com/usdot/flight-delays

This dataset is from Kaggle that is original collected from the U.S. Department of Transportation's Bureau of Transportation Statistics. It contains a summary about year of the flight trip, month of the flight trip, day of the flight trip, day of week of the flight trip, airline identifier, flight identifier, aircraft identifier, starting airport, destination airport, planned departure time and other related variables.[1][1]

3 Related Work

A conference paper called "Chained Predictions of Flight Delay Using Machine Learning" written by Jun Chen and Meng Li was presented in January 2019. The main purpose of their model is to predict flight delay along the same aircraft's itinerary given the initial departure delay. In their research, the methodology used is a combination of classification and propagation model. Instead of using all features available in the dataset, they implemented an optimal feature selection process to improve the performance of the model. From all the features, they found departure delay and late arriving aircraft delay are the most influential factors. Thus, a delay propagation model was proposed to serve as a connection between these two main features to build a chained delay prediction model. Apart from the propagation model, they also used the Random Forest classifier, which is an ensemble method based on multiple decision trees. In conclusion, they adopted a mixed approach with three modules: the arrival delay prediction module, the departure delay prediction module, and the delay propagation module, with a Random Forest Classifier with multi-labels. [2][2]

4 Intended experiments

We plan to implement several regression models, including Linear regression, polynomial regression, Decision Tree regression, and Random Forest regression, etc., to figure out the pattern of departure delay. We plan to start with looking at the big picture and using dataset to frame our question. Next, we plan to explore and visualize the data and check correlations to help us understand our data better. Then, we will perform data cleaning step, so that we can get a neat data set with less bias. We plan to build models and choose to use RMSE to measure the performance. After that, we will perform model selection and fine-tuning this model.

One of the techniques we decide to use is cross-validation. This technique is used for testing if our final model could be applied into the practical world. This technique requires us to split the data into two sets: a training set to estimate the parameters and build the model and a testing set to check the model productivity. We will use two specific methods in cross validation: 5-fold and Leave-one-out cross validation. 5-fold cross validation will be similar to what we learned in class. We split the data into 5 equal parts. One of the five parts will be used as a testing set, and the other four parts will be considered as a training set. We repeat this procedure until each data will be used in the testing set exactly once, and used in the training set four times. Similarly, leave-one-out cross validation will take one observation in data as testing set and leave all others as a training set. We will perform both cross validation methods to evaluate our model to get a better RMSE.

References

[1] U.S. Department of Transportation. 2015 flight delays and cancellations. 2017.

[2] Jun Chen and Meng Li. Chained predictions of flight delay using machine learning. In *AIAA Scitech 2019 Forum*, page 1661. AIAA, 2019.

[1] U.S. Department of Transportation. 2015 flight delays and cancellations. 2017.

[2] Jun Chen and Meng Li. Chained predictions of flight delay using machine learning. InAIAA Scitech 2019 Forum,page 1661. AIAA, 2019.