ML4VA FINAL PROJECT PROPOSAL (RICH'S ENTHUSIASTS)

Corneel Hollebrandse
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 229033
jh7jss@virginia.edu

Trent Bilyeu
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
trb7ap@virginia.edu

Spencer Hernandez
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
slh3mm@virginia.edu

June 15, 2023

1 Abstract

The goal of our project is to find what features are most important in determining flight delays in order to make a model that can predict if a given flight will be cancelled or delayed by 15 minutes or more. Our data set consists of 66,661 domestic US flights to and from Virginia from the last 35 years recorded by the United States Bureau of Transportation Statistics. After creating a baseline method of linear regression, we will build an ensemble model using multiple other machine-learning models.

2 Introduction

Our project is focused on airline data delays. According to the Federal Aviation Administration in 2018, the annual cost of flight delays was 28 billion dollars. This is calculated due to lost time, missed meetings, and general opportunity costs. We would like to use machine learning to find which features correlate to Virginia airports having more delays/cancellations, such as what times of day and days of the week. The results of this data would be useful for Virginian airline passengers who seek to avoid delays as well as airlines for planning routes to avoid delays. This impacts Virginia directly because all of the flights used in the data set are either from or to a Virginian airport. Our hypothesis is that there is not a uniform distribution of flight delays since there are higher densities of delays due to external factors such holidays. We propose that the magnitude of these delays will vary depending on our data's features.

3 Method

First, we cleaned the original data set of 2 million domestic flights from 1987 to only include flights with an origin or destination within Virginia. This was done to focus the project more specifically on Virginia to better fit the project requirements. This reduced the size of the data set to 66,661 flights. We also chose to drop canceled flights because they necessarily do not have an arrival time and therefore delay.

Furthermore, we continued feature engineering by removing features from the data that would make our results less meaningful. For example, we removed arrival delay, taxi on, and taxi off times for two reasons. One, a passenger using our model to predict if a flight was delayed would not have access to that data because their flight has not yet occurred. Second, these features are extremely correlated with departure delay and could potentially dwarf meaningful

information we would want our model to learn about the importance of other features, such as the day of the week and holiday features.

Then, we split the data into training, validation, and test sets for use in all of our models. We ran the training set through a pipeline using a simple imputer using the median strategy for numerical data and ordinal encoding for categorical data.

Next, we visualized the data using heat maps and scatter matrices to get a sense of how all of the features are correlated with the label of arrival delay. The result of this visualization is shown below in Figure 1:

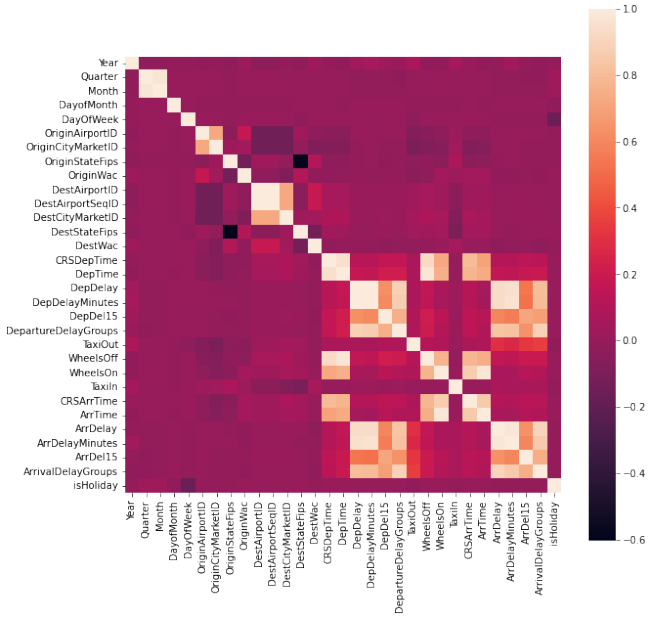

Figure 1: Heatmap of Flight Dataset Features

Then, we created baseline regression models to predict how long a flight's departure delay would be. We chose to run linear regression, gradient-boosted regression, and XGB Regression. Using Root Mean Squared Error as our error function, our most accurate baseline model was able to predict values on the test set.

4 Experiments

Our linear regression model is able to predict values on the test set with an RMSE of 30.1608. It predicted that the mean delay time for all flights in the test set was 27.5408 minutes, with a standard deviation of 9.4949.

Our gradient-boosted regression model was able to predict values on the test set with an RMSE of 29.0735. It predicted that the mean delay time for all flights in the test set was 27.5278 minutes, with a standard deviation of 9.5248.

The XGBoost regression model was able to predict values on the test set with an RMSE of 28.5570. This is by far our best model so far, however, concern has to be taken as to whether this model is overfitting the data.

After completing all of the baseline regression models, we moved forward with implementing Ensemble Machine Learning. While our regression models focused on predicting the departure delay minutes for a given flight, our final machine-learning model focuses on the classification of flights. Instead of departure delay minutes, our classifier was whether or not a flight was delayed by over 15 minutes.

We implemented a voting classifier that took the votes from a Random Forrest Classifier, a Support Vector Classifier, and a Logistic Regression model to implement ensemble learning. The performance of each of these classifier models is discussed in the following section.

5 Results

Our results provided a fascinating insight as to what features were most important in regard to delayed flights.

With our cancellation prediction, we achieved an accuracy of 99.9691% using ensemble learning. In addition, Random Forrest Classifier, Support Vector Classifier, and Logistic Regression were able to achieve an accuracy of 99.98%, 99.98%, and 99.98% respectively. The voting classifier also achieved an accuracy of 99.98%.

The feature that was most correlated as to whether a flight is canceled was the reporting airline with an R-value of 25.36%. This was followed by the departing time block, day of the month, and day of the week, with R-values of 10.34%, 8.53%, and 8.38% respectively. Surprisingly the origin state for inbound flights, the destination state for outbound flights, and whether the flight was on a holiday had the lowest correlations. After further review, we can conclude that because there are fewer travelers on holidays, the chances of a flight being canceled are lower. As for why the destination state and origin state have a low correlation, we infer that because there are dozens of airports in a given state, cancellations average out over time making it a less important feature. It is likely that this feature was overfitting the data, but we were unable to pinpoint the specific cause as to why.

We performed the same ensemble learning method on classifying whether a flight was delayed by 15 or more minutes. For this classification, we were able to achieve an accuracy of 84.26% using ensemble learning. In addition, Random Forrest Classifier, Support Vector Classifier, and Logistic Regression were able to achieve an accuracy of 84.11%, 84.26%, and 84.26% respectively. The voting classifier also achieved an accuracy of 83.86%.

The feature that was most correlated as to whether a flight is delayed by 15 more minutes was the reporting airline with 16.33%. This was followed by the day of the month (meaning later days in the month were more correlated with delayed flights), department time block, and the Year (more recent years have had more delays), with R-values of 12.87%, 10.07%, and 8.11% respectively. The cancellations, origin, and whether the flight was on a holiday had the lowest correlations.

6 Conclusion

We discovered very quickly that our hypothesis was correct because flight delays are not uniformly distributed. More interesting was seeing the correlations of the various features in classifying whether a flight was delayed or canceled. In both cases, the day of the month was a highly correlating feature. This means that flights taking place later in the month were more likely to be delayed and canceled. We were a little shocked to discover that holidays had such a low correlation with delays and cancellations and can only speculate that most holiday-related travel delays happen before and after the holidays such that the holidays themselves are less busy. It was also interesting to see how there was a strong correlation between the reporting of airline and flight delays. This means that if you want to minimize your chances of having a delay, try to avoid Reagan (DCA) Airport, as this was the most frequent airport in the dataset, implying this is the most likely cause.

If we were to move forward with this project, it would make sense to work with airline companies so that when customers are purchasing tickets we could automatically run a prediction on their flight details to determine the likelihood of whether their flight will be delayed by 15 or more minutes and if it will be canceled. This would help Virginians save time and money and be a practical application of our machine learning project to the real world.

7 Member Contribution

The team worked well together with extensive collaboration via in-person and Zoom meetings to work on the project. To produce the video, we wrote a script, and then we booked a room in Clemons Library to use professional-grade audio equipment to ensure our voices were clean and crisp. Individually contributed the following to the project:

Spencer mutilated the data in Excel by eliminating flights that weren't to or from Virginia, cleaned the data in Collab, added extra features (ie: whether or not a flight fell on a holiday), found correlations between features, cleaned up code in Collab, assisted with Ensemble learning implementation, worked on the Preliminary Experiment and Next Steps sections of this document.

Corneel wrote the code for the pipeline, linear regression, and gradient boosted regression, and collaboratively worked with Spencer on implementing the Ensemble learning code. On this document, he worked on every section, but especially on Methods, Results, and Conclusion.

Trent assisted in the acquisition of the data set and the decision as to what features to keep. He also helped draft the next steps section and coded the XGBoost regression model. Trent also worked on the production of the final

video presentation. This presented many unique challenges. Then Trent used his editing magic to clip together relevant stock videos. He then scoured the interwebs to find the perfect score to set the video to. In the end, a magnificent video emerged from the ashes of the team's hard work.

8 Related Work

Due to the fact that airlines have always had delays, people have already spent tremendous resources researching the trends in airline delays. Some examples in our References section include a hobbyist who wrote a medium article analyzing airline delays, to a graduate student who devoted their Master's Thesis to this topic. There are several Machine Learning methods that our team plans on using, but just in researching this problem, we found a prior method that utilized linear regression. One of the biggest strengths of this project is the sheer amount of data sets available that contain relevant information. We were able to find several that had hundreds of thousands of rows (our IBM data set in particular had over 200 million), and some with around 30 features.

References

[1] Wang, Yuyang. "Modeling Flight Delays through U.S. Flight Data." Medium, Analytics Vidhya, 15 July 2020, https://medium.com/analytics-vidhya/modeling-flight-delays-through-u-s-flight-data-2f0b3d7e2c89.

[2] M. Abdel-Aty, C. Lee, Y. Bai, X. Li and M. Michalak, "Detecting periodic patterns of arrival delay", Journal of Air Transport Management,, Volume 13(6), pp. 355– 361, November, 2007.

[3] S. AhmadBeygi, A. Cohn and M. Lapp, "Decreasing Airline Delay Propagation By Re-Allocating Scheduled Slack", Annual Conference, Boston, 2008.

[4] A. A. Simmons, "Flight Delay Forecast due to Weather Using Data Mining", M.S. Disseration, University of the Basque Country, Department of Computer Science, 2015.

[5] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms", Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th, Sacramento, CA, USA, 2016.

[6] L. Schaefer and D. Millner, "Flight Delay Propagation Analysis With The Detailed Policy Assessment Tool", Man and Cybernetics Conference, Tucson, AZ, 2001.

[7] B. Liu "Sentiment Analysis and Opinion Mining Synthesis", Morgan & Claypool Publishers, p. 167, 2012.