SmartDelay: Making Your Next Flight Easier

William D. Lombardi, Dahai Liu, Adam Miller, Poorvi Varma
IST 718: Big Data Analytics

Problem and Objective

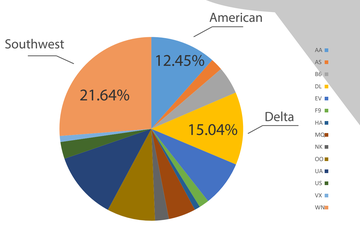
Everyone hates flight delays. This can cause a big inconvenience to the customer, which can cause a big loss to the airlines in the form of customer recompense. There is a current system in place to notify travellers of flight delays, but it isn't robust enough to allow people to plan accordingly. As a result, flyers who have their flights delayed end up waiting in the terminal for possibly hours on end, getting increasingly aggravated. This also results in decreased airport traffic efficiency, as a single flight delay can affect all the other scheduled flights at that airport.Our project will use a year's worth of flight delay data in order to build a predictive model that will be able to predict whether any given flight will be delayed, given that specific flight's history.

Our Data

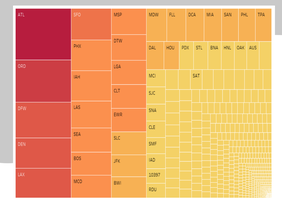
The dataset was obtained from Kaggle, but originated from the U.S. Department of Transportation. The data is composed of more than 6 million flights records, in the year of 2015. Due to the large number of records, only 100,000 randomly sampled ones were used in the final build.

Descriptive Statistics

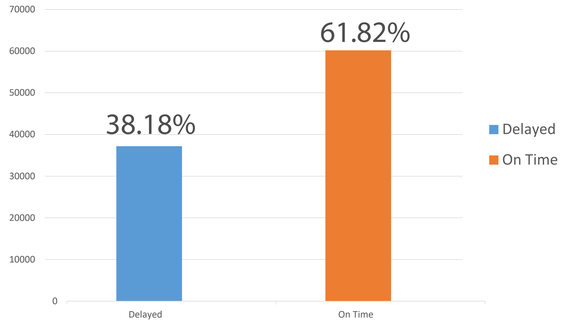
Distribution of Flight Delays per Airline



Tree Map of Origin Airports



Count of "Delayed" and "On Time" Flights



Models and Features

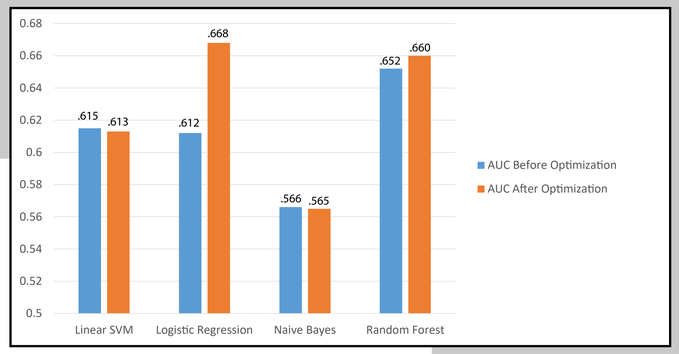
Models	Features Used	Techniques Applied	Evaluation Metric
Random Forest	Tail Number, Origin Airport, Flight Number, Scheduled Departure Time, Destination Airport, Airline, Scheduled Arrival Time, Diverted Classifier	Data Discretization, String Indexing, Vector Assembling, Grid Optimization	AUC Accuracy Recall
Logistic Regression, Naive Bayes, Linear SVM		Data Discretization, String Indexing, OneHotEncoding, Vector Assembling, TuneGrid	

Table 1:
Displays the models we tested, the features we utilized, the techniques we applied, and the evaluation metrics we chose to use

Model Comparison and Validation

We ended up choosing AUC as our primary evaluation metric due to its ability to show the robustness of each of our models at different probability thresholds.

AUC Validation Comparison



The model with the highest AUC after optimizing using a grid-based method was the Logistic Regression model, with a value of **.668**. It was also the model that saw the most appreciable difference in performance once we had optimized its parameters.

Default Parameters:

- Regularization Parameter: **0**
- Elastic Net Parameter: **0**
- Maximum Iterations: **100**

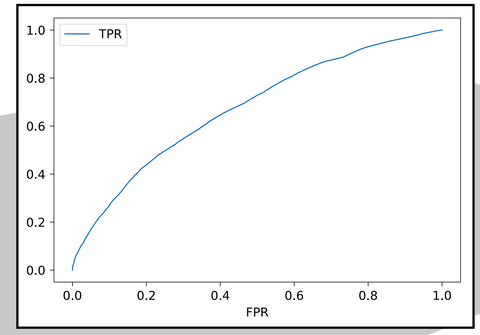
Optimized Parameters:

- Regularization Parameter: **0.01**
- Elastic Net Parameter: **0.5**
- Maximum Iterations: **10**

Test Performance

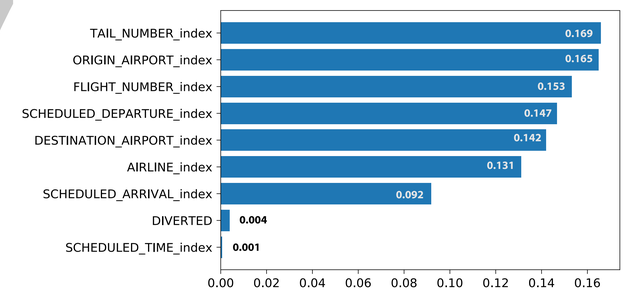
After applying our optimized parameters to the model and transforming it against the test data, the final value for our model's AUC was **.675**, which is relatively close to that of our validation dataset. Below is the ROC curve for this AUC, which we generated in Matplotlib.

ROC Curve for Optimized LR Model (Test Data)



Feature Importances

Random Forest Feature Importances



Conclusion

The economic impact of flight delays for domestic flights in the US is estimated to be more than $19 Billion per year to the airlines and over $41 Billion per year to the national economy. (H. Balakrishnan, Annual Reviews in Control). By using machine learning and modeling technique, we were able to improve the ability to foresee potential flight delays and take actions hour before the delay occurs to a minor degree. This overall lack of accuracy can be attributed to our preprocessing methodology. Given how we had to take out many features in order to accurately replicate a practical situation where this model might be applied, the lack of data could have curtailed the strength of our models.

However, that isn't to say that we don't think this application is useful. The information provided by these models can help airline companies save on resources, and improve customer satisfaction. Also, a positive impact can be made to society and environment by reducing customer wait time, plane fuel consumption, and gas emissions.

Finally, we generated inferential conclusions that the factors which have the most influence on a plane's delayed status is its airline and its origin airport.

Source: U.S. Department of Transportation (https://www.kaggle.com/usdot/flight-delays#flights.csv)