Predicting Airline Delays: A Comparison of Models and Features

Dorian Goldman[1]

[1] *Conde Nast, Data Scientist,*
Columbia University, Adjunct Professor of Data Science

In this note, we compare different methodologies for predicting airline delays. We focus on all airline delays which are made publicly available on the website of The Bureau of Transportation Statistics [1]. We compare different modeling methodologies including, classification and regression, before and after an enhanced set of features is introduced. The features which we add to provide supporting predictive accuracy are: weather data from the NOAA [3], plane data from the Federal Aviation Administration [2], publicly available holiday data and data about the passenger traffic at USA airports [4]. We find that three of the above improve predictive performance for regression and classification, with features ranking in the top 10-15 most predictive for the nonlinear models. Motivated by an analysis of the time series, we also implement Poisson regression time series model to predict the mean delay on days and hours prior to the day and hour respectively being evaluated. The performance is compared to the standard auto regressive time series model and shown to have significant improvements in the RMSE. The author conjectures that the ideal model would be a contextual Poisson regressive time series, but due to time constraints, this was not implemented. This note contains only the results of the classification analysis, as the results of the regression were unremarkable.

I. INTRODUCTION

In this note we briefly summarize the main results obtained in attempting to predict airline delays. We model this problem as

$$Y^t = f(X_{t-1}) \text{ for } t > 0, \tag{1.1}$$

where t is measured in hours, X_{t-1} represents any collection of features we know at time $t-1$ (ie. one hour before the flight) and f is the model we wish to find. We believe this is a reasonable assumption in making in app that would be used to predict delays - expecting to predict delays a day in advance would rely only on general trends of the airline and unreliable weather forecasts, and any probabilistic forecast wouldn't change whether or not an individual needs to be at the airport at the designated time. We focus only on the classification problem and we define a delay to be if $Y^t > 20$ minutes. The results of the regression analysis can be found in [7]. To be more precise, X_{t-1} represents:

- Weather data [3]
- Plane data [2]
- Holiday data
- Airport capacity data [4]
- Time series data on the delays up to time t-1. [6]

We introduce two time relevant variables based on our exploratory analysis of the distributions of the data each hour and each day. In particular, we consider

$$Y_t^d \sim \mathcal{P}(\mu_t^m) \tag{1.2}$$
$$Y_t^h \sim \mathcal{P}(\mu_t^h) \tag{1.3}$$

where $\mu_t^d = \beta^d \cdot X_{t-1}^d$, $\mu_t^h = \beta^h \cdot X_{t-1}^h$ and $\mathcal{P}$ is a Poisson distribution. We justify this assumption in [5] and develop the coefficients in [6].

Model	AUC
Random Forest (Enhanced)	0.76
Gradient Boosted (Enhanced)	0.75
Logistic Regr. (Enhanced)	0.75
Random Forest (Original)	0.70
Gradient Boosted(Original)	0.70
Logistic Regr. (Original)	0.65

TABLE I. Model performance comparison

Our paper is outlined as follows. In Section II we discuss the main results of the models with enhanced (all features mentioned above) and the original variables. We then discuss the most important features in Subsection II B. We then discuss our methodology in II C.

II. DISCUSSION OF MAIN RESULTS

A. Performance

In this section we display the performance of the models compared. We find that including the features above, our performance improves the ROC by 7-8% overall. In addition, at least one variable from each dataset above appears in the top 15 features which are most predictive, indicating that each choice of data set did indeed have a meaningful contribution. The comparison of results is summarized in Table I and the ROC curves are plotted and compared in Figure 1. With more time, the Gradient Boosted Classifier would most likely out perform the Random Forest model, but time wasn't spent on this due to time constraints.

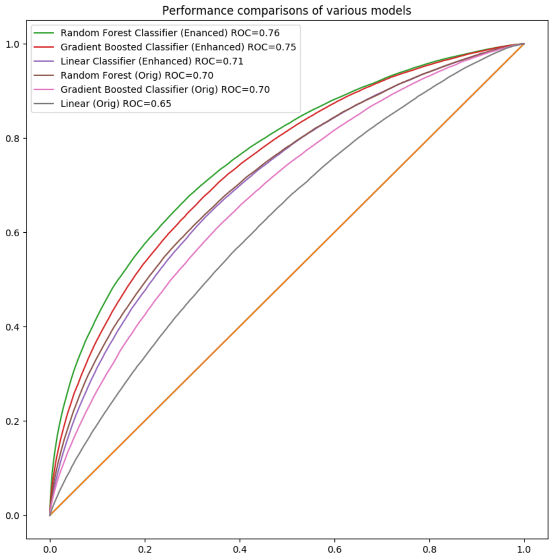

FIG. 1. ROC Comparison Between Models..

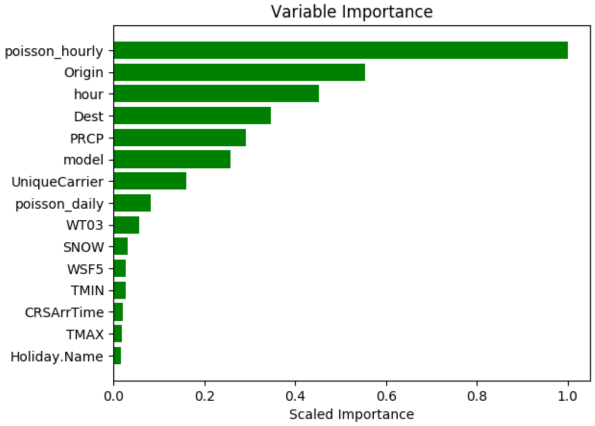

FIG. 2. GBT Variable Importances.

B. Most Predictive Features

For the reader who wishes to see the feature importance plots for the models with original variables only, we refer to [8]. For the enhanced models, we see that features from 4 out of 5 of our original data sets contribute to the top 15 features. Figure 2 refers to the Gradient Boosted model and 3 refers to the Random Forest model. The variables PRCP and SNOW refer respectively to precipitation and snow in mm per hour [9], model and issue_date refer to details of the plane [2] and poisson_hour and poisson_day are as defined above, constructed in [6]. Finally Holiday.Name is the name of the holiday on any given day.

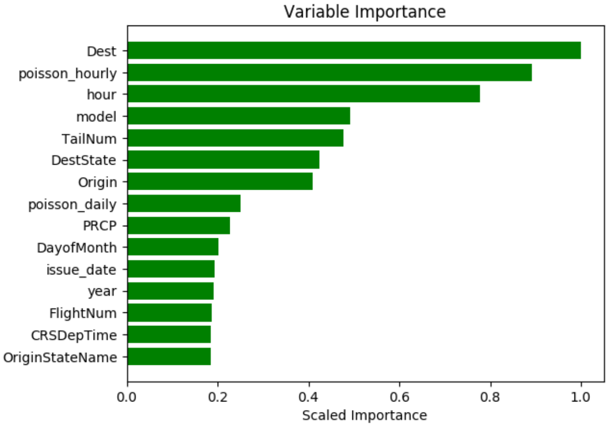

FIG. 3. RF Variable Importances.

C. Methodology

We use Python for all of the data processing, and to deal with the very large amount of data (2450509 rows and over 1200 columns when categorical features are expanded), we use the h2o library which performs the model building and analysis in a paralelized way. The work is summarized in the following notebooks:

- Airline Delays Part 0 - Processing of Weather Data [5]
- Airline Delays Part I - Feature Analysis, Preparation and Processing [6]
- Airline Delays Part II - Time Series Analysis [7]
- Airline Delays Part III - Model performance before and after feature enhancement [8]

For the reader wishing to see the comparison of models, they can safely go to notebook [8]. The variables poisson_hour and poisson_day are generated in [7]. Feature analysis and EDA which motivate the feature creation and filter out nonsensical values is done in [6].

III. CONCLUSION

We see by using a wide variety of data sets , utilizing the seasonality of time series and structure of the underlying distribution, we can improve the performance of the model by a significant amount.

[1] RITA/BTS. Bureau of Transportation Statistics.. *Airline On-Time Performance Data.* `https://www.transtats.bts.gov`. 2016

[2] Federal Aviation Administration. *Flight Standards Service - Civil Aviation Registry* http://stat-computing.org/dataexpo/2009/plane-data.csv. 2009

[3] National Centers for Environmental Information. *Local Climatological Data.* https://www.ncdc.noaa.gov/cdo-web/datatools/lcd. 2016

[4] Federal Aviation Administration. *Passenger Boardings at Commercial Service Airports.* https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger/media/cy14-commercial-service-enplanements.pdf. 2014

[5] Dorian Goldman. *Airline Delays Part I - Feature Analysis, Preparation and Processing* https://github.com/doriang102/Airline_Delays/blob/master/analysis/. 2017

[6] Dorian Goldman. *Airline Delays Part II - Time Series Analysis.ipynb* https://github.com/doriang102/Airline_Delays/blob/master/analysis/. 2017

[7] Dorian Goldman. *Airline Delays Part III - Model performance before and after feature enrichment* https://github.com/doriang102/Airline_Delays/blob/master/analysis/. 2017

[8] Dorian Goldman. *Processing of Weather Data* https://github.com/doriang102/Airline_Delays/blob/master/analysis/. 2017

[9] National Centers for Environmental Information. *Readme File For Daily Global Historical Climatology Network* ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt2016