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 all 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.

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. To be more precise, X_{t-1} represents:

- Weather data
- Plane data
- Holiday data
- Airport capacity data.
- Time series data on the delays up to time t-1.

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 $\mu_t^d = \beta^d \cdot X_{t-1}^d$ and $\mu_t^h = \beta^h \cdot X_{t-1}^h$.

Our paper is outlined as follows. In Section **??** 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 **??**.

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

II. DISCUSSION OF MAIN RESULTS

While the actual delays closely resemble a Tweedie distribution [] (a Poisson distribution with a negative concentration), the means seem to closely resemble a Poisson distribution [].

We also find that the poisson variable mentioned above occurs as the first or second variable in both nonlinear models compared, justifying the initial hypothesis that timely information is essential for an accurate prediction.

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.

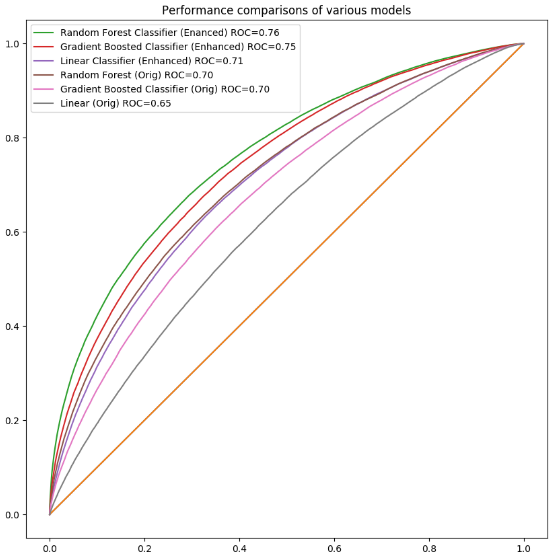

FIG. 1. ROC Comparison Between Models..

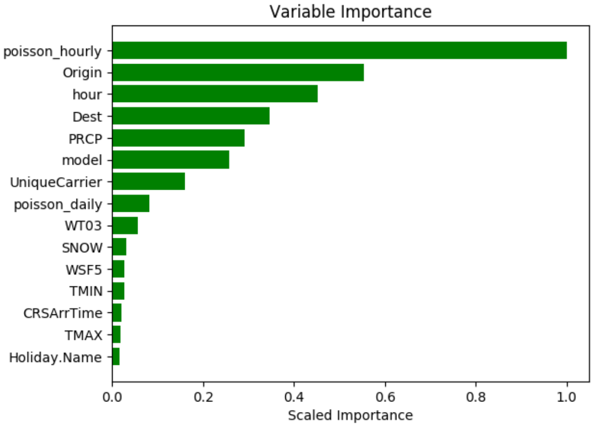

FIG. 2. GBT Variable Importances.

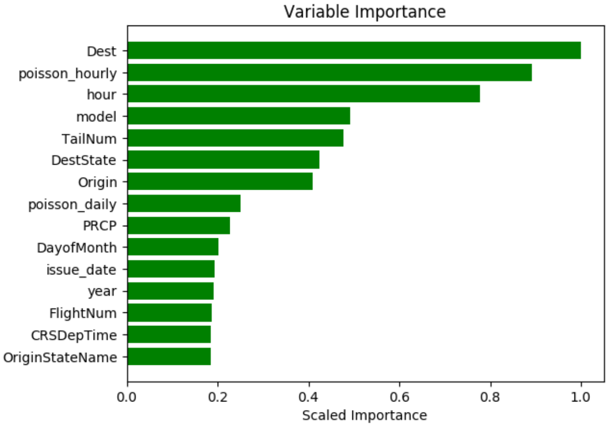

FIG. 3. RF Variable Importances.

B. Most Predictive Features

C. Methodology

- blah

III. CONCLUSION

[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