Predicting Airline Delays

Introduction

For this analysis, I will consider the 2016 airline delay data only, and my goal will be to predict delays using this one year of data.

I have several hypotheses which I will evaluate throughout this presentation.

Goal: *Given all of the airline data from 2016, can we predict delays?* I tried this using two methods:

1. Regression on the departure delay.
2. Classification above a 20 minute threshold (defined as a delay by most major airlines).

File Structure

`analysis/` : The notebooks used for all of the data processing and analysis.

`doc/` : The research paper and TeX file for the research paper.

`fig/` : Figures generated used in the research paper.

`json/` : JSON files used for processing some of the weather data.

`data/` : All data sets used. Note this was not included on Github due to size constraints, but is available upon request.

`src/` : Source files used for scraping weather data.

Research Results

A research paper written on the results is contained here:

Predicting Airline Delays: A Comparison of Models and Features

Analysis

The analysis is done in four iPython notebooks:

- 1) Airline Delays Part 0 - Processing of Weather Data

This notebook contains the processing of all of the weather data from 2016.

- 2) Airline Delays Part I - Feature Analysis, Preparation and Processing

This notebook includes analysis of the features involved including exploratory analysis. It also includes merging of the data sets.

- 2) Airline Delays Part II - Time Series Analysis

This notebook includes analysis of the time series, and creation of the Poisson variables used.

- 3) Airline Delays Part III - Model performance before and after feature enhancement

The final notebook includes the comparison of the models with and without the new features added below.

I used the following data in my analysis.

Data sets:

- Airline On-Time Performance Data. RITA/BTS. Bureau of Transportation Statistics.. 2016.
- Local Climatological Data. National Centers for Environmental Information.. 2016.
- Flight Standards Service — Civil Aviation Registry. Federal Aviation Administration.. 2009.
- Passenger Boardings at Commercial Service Airports. Federal Aviation Administration.. 2014.

Caveats of above data: Plane data is from 2009, but will have missing planes from 2010-2016.

Main Results:

Performance:

We compare performance of multiple models with and without the additional data used above. More precisely, we define:

Original: Only the Airline On-Time Performance Data.

Enhanced: All of the data included in the above links, in addition to a customized Poisson variable. More precisely, I define two time dependent variables with Poisson priors given by:

$$Y_t^d \sim \mathcal{P}(\mu_t^m) \tag{1.1}$$

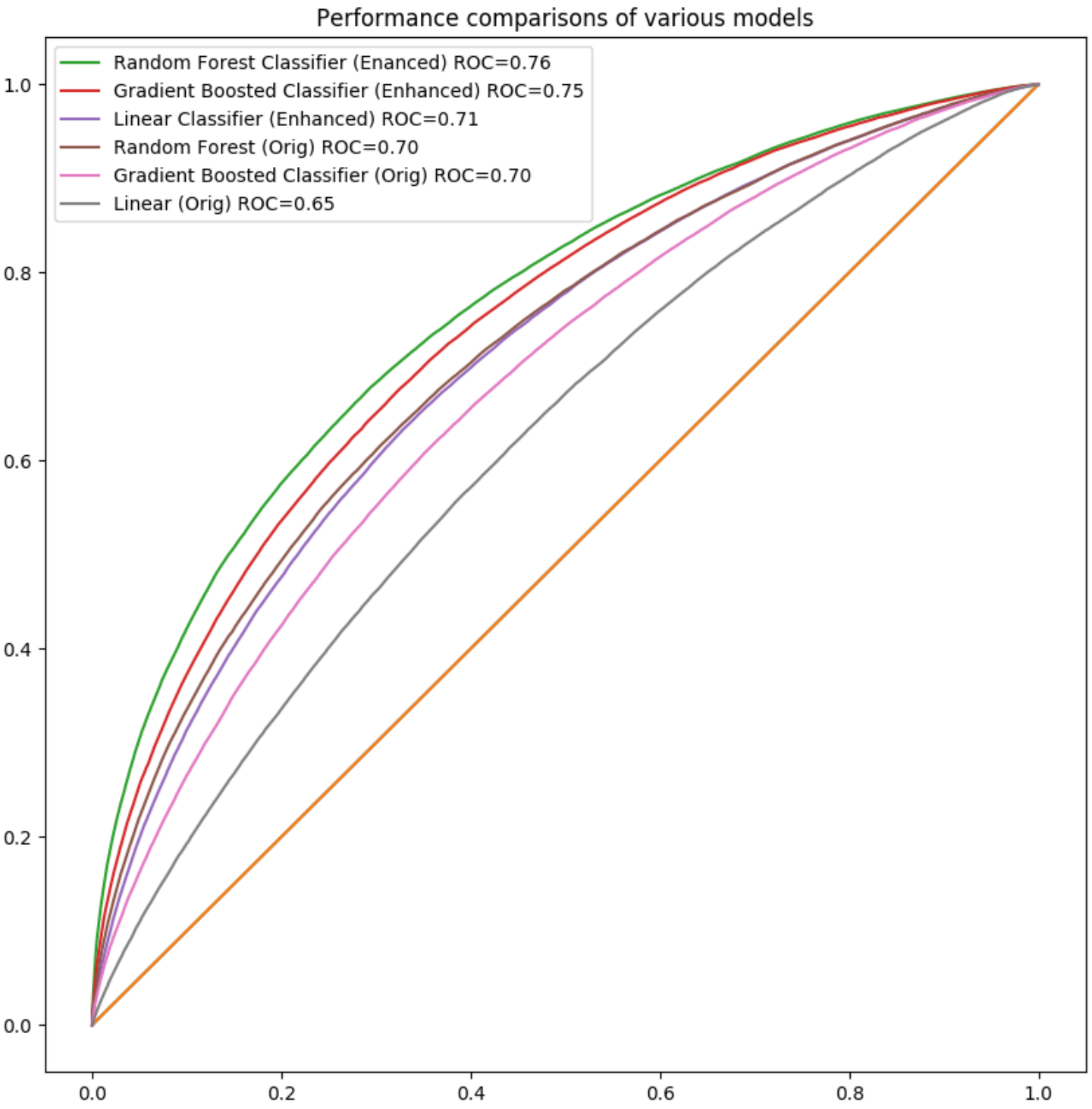
$$Y_t^h \sim \mathcal{P}(\mu_t^h) \tag{1.2}$$

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

where the P represents a Poisson distribution with time dependent mean depending on the delay of the previous hour or day respectively.

The hour variable (called **hourly_poisson** below) is seen to be the top variable for the Gradient Boosted Trees and second top variable for the Random Forest classifier.

Below we see a comparison of three different classification models for the original and enriched variable set. The best performance obtained was by the Random Forest Classifier with the enriched data set, which achieved an ROC of 0.76.



Variable Importances:

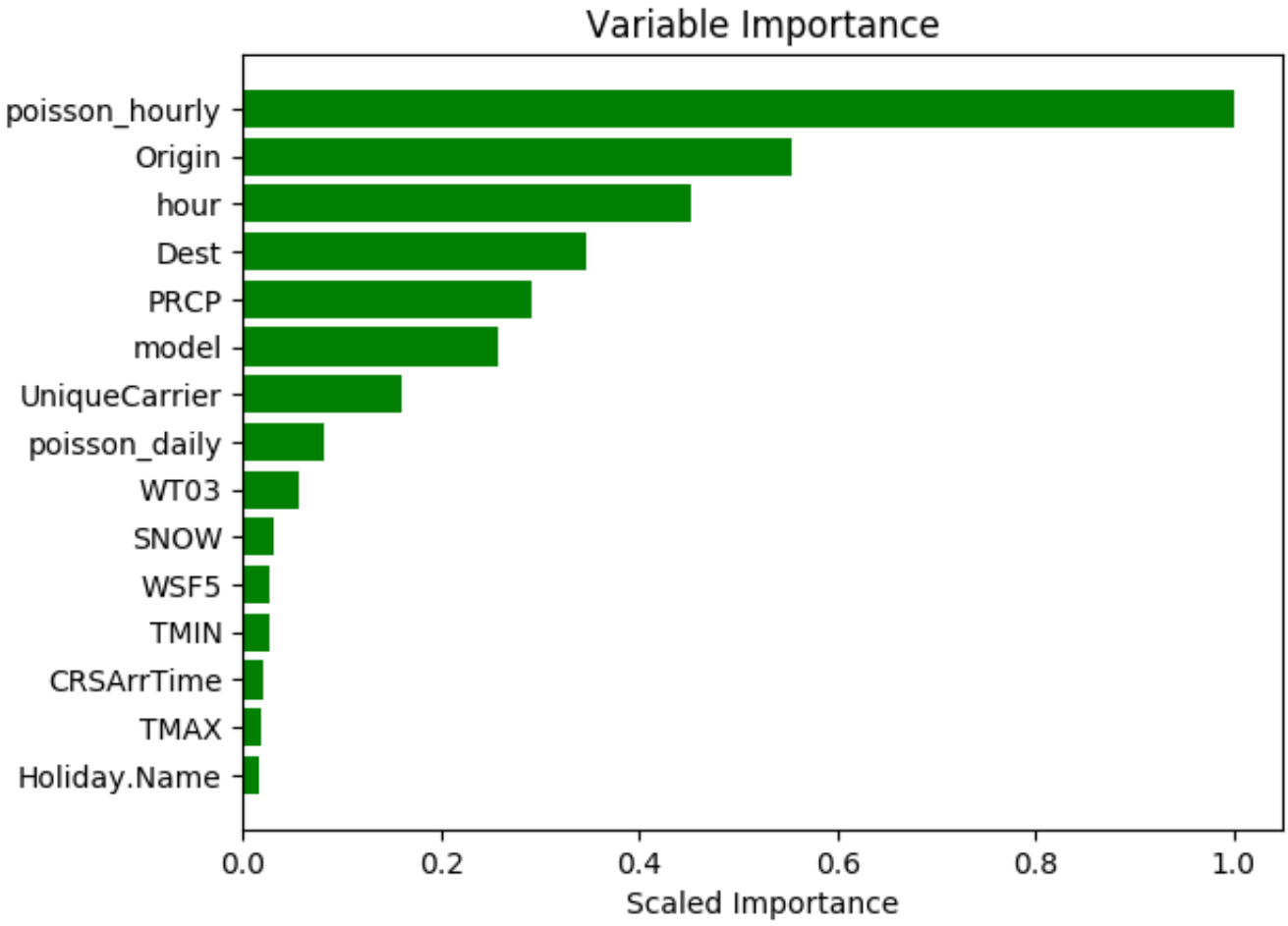
The variables are explained further in the notebook (Processing Weather Data), but for the reader's convenience here are some descriptions.

Weather: `PRCP` , `SNOW` and `WT03` are weather variables representing precipitation, snow and wind.

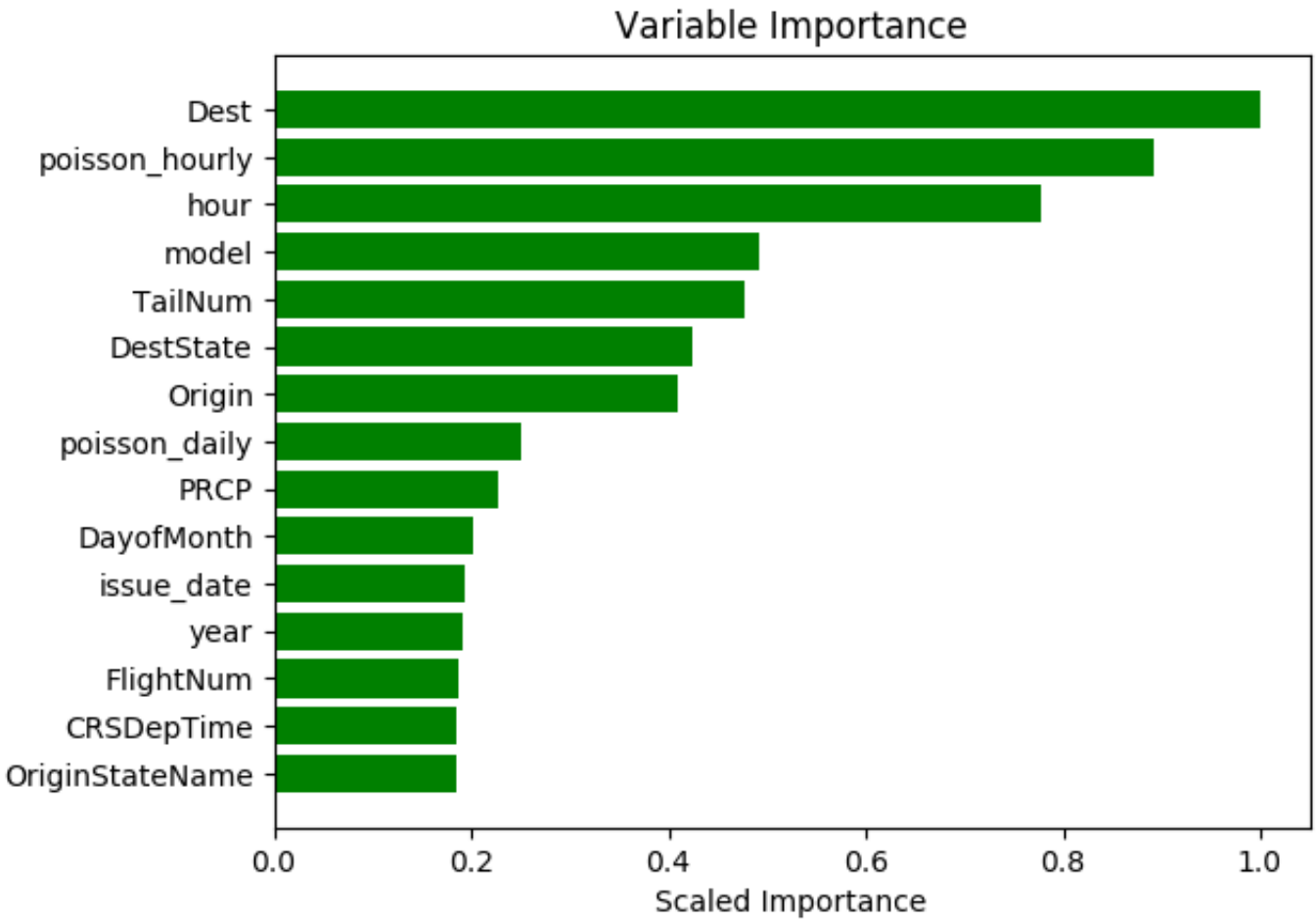
Plane Model: `model` and `issue_date` correspond to the make of the plane.

Poisson Regression Time Series: `poisson_hourly` and `poisson_daily` represent the variables I discussed briefly above.

Gradient Boosted Trees:



Random Forest:


Variable Importance
Dest
poisson_hourly
hour
model
TailNum
DestState
Origin
poisson_daily
PRCP
DayofMonth
issue_date
year
FlightNum
CRSDepTime
OriginStateName
0.0
0.2
0.4
0.6
0.8
1.0
Scaled Importance