Flight Delay Prediction

Praveen Kumar R

September 7, 2019

Abstract

The Project aims to predict the whether a scheduled flight will be delayed on departure or Arrival based on various parameters that define the travel and also the environmental conditions. Dataset of flights departing and arriving at 15 different stations were used for this task. The weather details were collected at an interval of one hour for the years 2016 and 2017. Based on these data I tried to predict weather the flight will be delayed or not (classification) and if it is delayed what will the amount of time (in minutes) it will be delayed(regression). Out of various algorithms Random Forest Classifier(0.8514, 0.9386, 0.9392) and Extra Tree Classifier(0.8407,0.9428,0.9418) proved to be better for the classification task and Gradient Boosting Regressor(0.9422) proved to be working for the regression task.

1 Introduction

Flight operators incur a considerable amount of loss due to flight delay. Delay in flight's departure or arrival can be related with various factors that define the scheduling of the flight and the environmental conditions. On a abstract level the flight delay prediction can be viewed a pipelined operation of two sequential tasks, first to predict whether a flight will be delayed or not(classification) and secondly, if the flight is delayed then to predict by how much time(in minutes) the fight will be delayed. For the classification task logistic regression, random forest classifier, support vector machine, Extra Tree Classifier and Extra Gradient Boost Classifier can be applied. For Linear Regression, Extra Tree Regressor, Support Vector Regressor, and Extra Gradient Boosting Regressor

2 Data Set

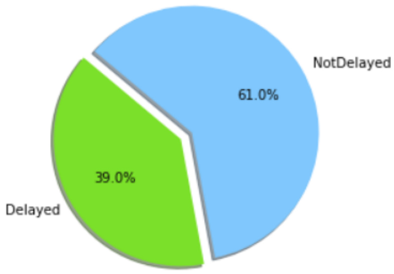
The flight data set, a comma separated value, contains the details the flight schedules in 15 airports of US. Each entry in the dataset is uniquely identified by the composite key (Flight ID, day of month, month, year). In total there are 18,77,667 datapoints the flight data set. The weather data set is a JSON file that contain weather data that is periodically recorded for every one hour over two years(2016-2017).

3 Pre-Processing

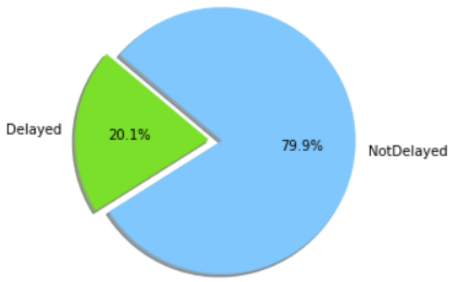
The following features were extracted from flight dataset.

Feature	Value	Feature	Value
Quarter	[1-4]	Month	[1-12]
Day of month	[1-31]	Day of week	[1-7]
Flight Number	Integer	Origin	15 air stations in US
Scheduled Departure Time	Integer	Actual Departure Time	Integer
Departure Delay	Integer	Departure Delay > 15 minutes	Integers
Destination	15 air stations in US	Year	$2016, 2017$
Scheduled Arrival Time	Integer	Actual Arrival Time	Integer
Arrival Delay	Integer	Arrival Delay > 15 minutes	Integers

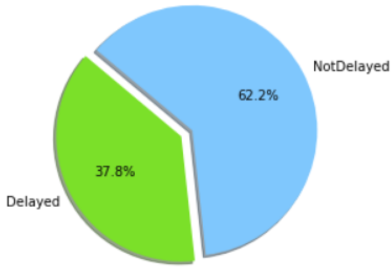
Out of these features 'Origin' and 'Dest' were label encoded to [1-15], each value representing each station.



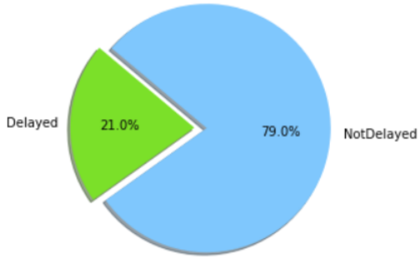
(a) Departure Delay



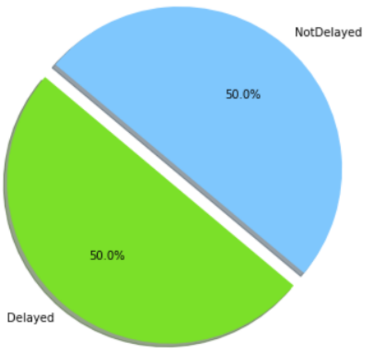
(b) Departure Delay > 15



(c) Arrival Delay



(d) Arrival Delay > 15



(e) Dataset after Sampling

Figure 1: Delay Analysis

The weather data set is a JSON file consisting of weather data of 2 year recorded every one hour for each of the 15 stations. Thus there are 731x24x15 = 263160 records in the weather data set. The following features were extracted from the weather dataset.

Feature	Value	Feature	Value
Quarter	[1-4]	Month	[1-12]
Day of month	[1-31]	Day of week	[1-7]
Time	Integer	Airport	15 air stations in US
Dew Point in Celsius	Float	Heat Index in Celsius	Float
Wind temperature (Celcius)	Float	Wind Speed (kmph)	Float
Percentage cloud cover	Float	Humidity	float
Precipitation (mm)	Float	Pressure	Float
Average Temperature	Float	Visibility	Float
Wind Direction (Degree)	Float	Weather Code	Integer

In ordered to merge these to datasets the fields time, day, month, year and airport code were used. Since the time in fight data set is not rounded off to the nearest hour. Based on what field is chosen for the 'time' from the flight data set we can generated two different datasets, snapshot before departure and snapshot after the departure.

We observe that the dataset is skewed towards the class 'Not-Delayed'. This skew can cause the model to not learn properly because the information is biased to one class. Thus we perform sampling to reduce this skew. The are two method to make the dataset even, Over-sampling or Under-sampling. In order to preserver the existing data make full use of it, Over-sampling was employed to balance the dataset.

4 Classification

Classification task is done to predict whether a scheduled flight will be delayed or not. Three kinds of classification were performed on the dataset obtained.

- Prediction of departure delay based on the weather data obtained during scheduled departure time.
- Prediction of arrival delay based on the weather data obtained during scheduled departure time
- Prediction of arrival delay based on the weather data obtained during actual departure time.

Random Forest Classifier, Gradient Boosting Classifier, Extra Tree Classifier, Support Vector Machine and Extra Gradient Boosting Classifier were used to perform the classification. The results obtained are as as follows,

Table 1: Departure Delay

Algorithm	Precision	Recall	f1 score	Accuracy
Random Forest Classifier	0.89	0.80	0.84	0.8514
Extra Tree Classifier	0.85	0.82	0.84	0.8407
Gradient Boosting Classifier	0.93	0.72	0.81	0.8296
XGB Classifier	0.93	0.72	0.81	0.8317

Table 2: Arrival Delay based on the weather data before departure

Algorithm	Precision	Recall	f1 score	Accuracy
Random Forest Classifier	0.96	0.91	0.94	0.9386
Extra Tree Classifier	0.96	0.92	0.94	0.9428
Gradient Boosting Classifier	0.96	0.91	0.94	0.8955
XGB Classifier	0.96	0.92	0.94	0.9413

Table 3: Arrival Delay based on the weather data after departure

Algorithm	Precision	Recall	f1 score	Accuracy
Random Forest Classifier	0.96	0.91	0.94	0.9392
Extra Tree Classifier	0.96	0.92	0.94	0.9418
Gradient Boosting Classifier	0.96	0.91	0.94	0.8944
XGB Classifier	0.90	0.69	0.78	0.9197

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$f1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

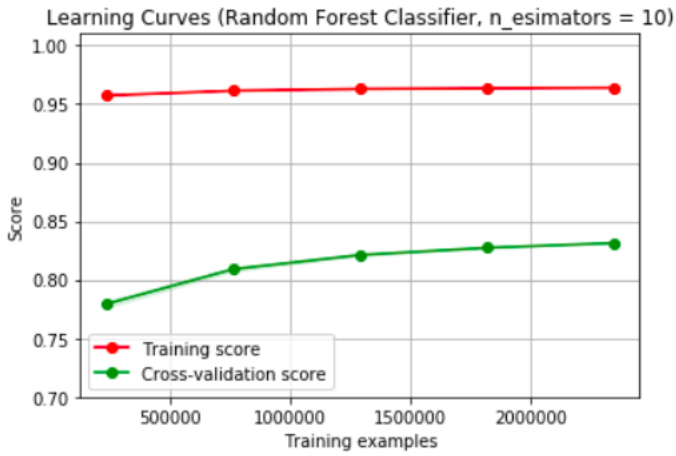


Figure 2: Learning history of the Random Forest Classifier

5 Regression

Regression algorithm have a continuous output unlike the classification task which produce discrete output. Arrival delay of the flight at destination was predicted based on the weather data of the origin airport during the actual departure time. Linear regression, Support Vector Machine, Extra Tree Regression and Gradient Boosting regression were used for the task. The performance for each algorithm is shown as below.

Table 4: Arrival delay prediction based on weather data after departure at the origin airport

Algorithm	MSE	RMSE	MAE	R2-score
Linear Regressor	415.9970	20.3960	15.1254	0.9178
Support Vector Machine	2485.3669	49.8535	16.9795	0.5057
Extra Tree Regressor	326.9334	18.0813	12.7911	0.9349
Gradient Boosting Regressor	292.9988	17.1172	11.8265	0.9422

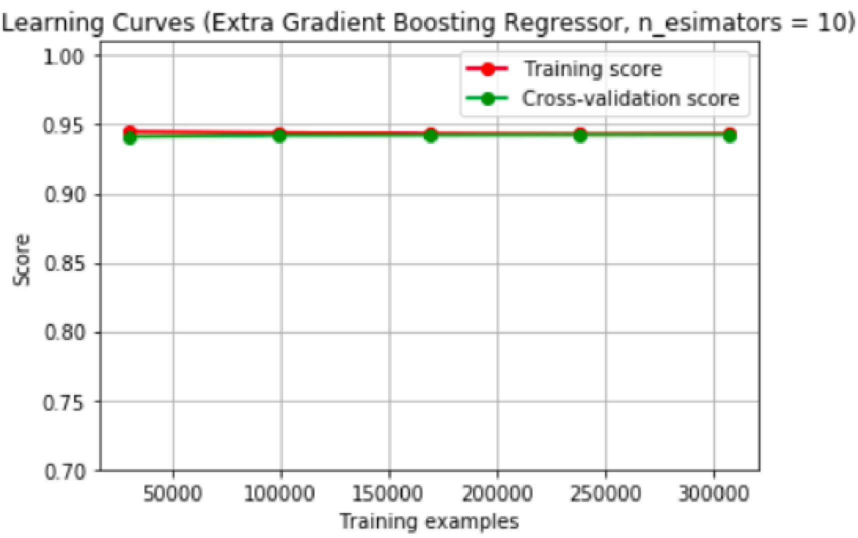


Figure 3: Learning History of XGB Regressor

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n}}$$

$$MAE = \sum_{i=1}^{n} \frac{|y_i - \bar{y}|}{n}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

6 Regression Testing

As we analyse the arrival delay value in the dataset we observe that, though the delay value range from 0 to 2142, only values in the range 0-200 occur with maximum frequency. The frequency distribution graph of the arrival delay is shown as below. The Linear Regressor model was tested on various ranges

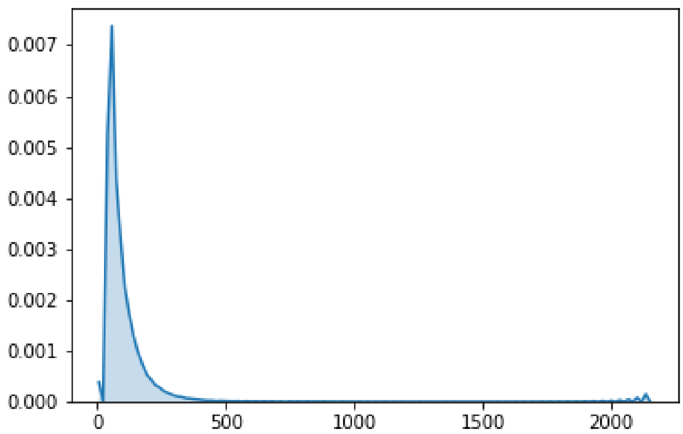


Figure 4: Frequency distribution of arrival delay

of input. It was obsewrved that as the range of input approached the maxima of the distribution plot the errors reduced. The testing information is as follows

Table 5: Testing of Linear Regression Model

Range	MSE	RMSE	MAE
> 2000 minutes	26194.1169	161.8460	161.7119
> 1000 minutes	9286.1945	96.3649	92.4357
< 200 minutes	379.2796	19.4751	14.7297
< 100 minutes	329.5595	18.1538	14.3314

7 Pipelining

In order to test the whole model a pipelined architecture was followed. The data was preprocessed to perform classification using Random Forest Classifier. The data points that were predicted to be delayed were selected to perform the regression. Linear regressor was used to perform the prediction. The overall flow of the program can be depicted as below.

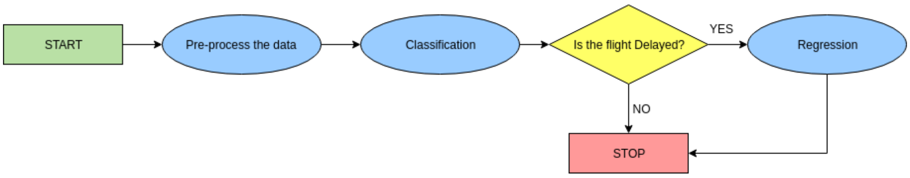


Figure 5: Flow the pipelined program

The model performed with a r^2 score of 0.9272. The Mean Squared Error was 408.9991, The Root Mean Squared Error was 20.2237 and the Mean Absolute Error was 14.9275.

8 Conclusions

The flight and weather datasets were preprocessed and merged in to one comma separated value file. We see that learning process of the model can be significantly affected by the skewness of the dataset. The algorithm tend to give less importance to the classes that are less dominant in the dataset. Thus to improve the performance of the algorithm the dataset was over-sampled to reduce the skew. Out of all classification algorithm employed Random Forest Classifier and Extra Tree Calssifier proved to be better performing on compared based on the validation accuracy for the test dataset for these algorithms. Gradient Boosting Regressor proved was better performing than on algorithm when compared based on the RMSE, MAE and R^2 score. The pipelined model constructed using the Random Forest Classifier and linear regressor was useful in evaluating the overall performance of the system.