Flight Delay Prediction For Aviation Industry Using Machine Learning

Modeling :

The first stage of the model generation is collection of data ,in this stage we have explored and obtained data

that will be used to feed the machine. Web scraping is one of the common technique to collect information

automatically from various sources such as APIs. And next the data is visualized and checked for the

correlations between different characteristics. The data is separated in the ratio of 80/20 for training and

testing of the data. After that we will choose the algorithms for the further steps. There are many algorithms

available some of them are random forest, decision tree, KNN algorithm, naïve bayes algorithm. according to

the project or processes such as image, text, sound and numerical values, we choose the model to process the

data.

The steps involved in the Model generation are:

1. Data Collection

2. Prepare the data

3. Choose the model

4. Train the model

5. Evaluation

6. Parameter tuning

7. Prediction

Now that the data has been cleaned and gone through a thorough EDA process done in two stages, its time to start with the modeling which will be a binary classification, where a "0" will correspond to a flight being on time, and a "1" to a flight being delayed.

This dataset consists of 28 features, out of which there are a series of them (listed above) that can affect the predictive model in a positive way in terms of predictions and therefore accuracy. However, when you use them, you are making the assumption that you are most probably already sitting in the plane, or in the best case scenario, your flight status on the departure boards has been changed to: "delayed". This is what the majority of the published models do, so I decided to do something slightly different by limiting the model to only features that won't directly indicate a delay.

Because I am not sure which Machine Learning algorithm will be the best for this type of binary classification I will be testing the following six:

1. Bagged Trees
2. Random Forest
3. AdaBoost
4. Gradient Boosted Trees
5. XGBoost
6. Deep Neural Network (MLP)

Through this notebook you will read that I am referring to ML and Neural Networks, now I know that strictly speaking a Neural Network is a type of ML model that is usually a supervised learning, so I am doing this only for practicality reasons due to the way how I will be using and evaluating both.

We already know that this dataset if severely imbalanced which will force me to do weighting with most of the ML algorithms. Putting that aside, the next step was to check for "categoricals", which I knew I had plenty such as the air carriers, days, months, weekdays, departure and arrival cities, and every column related with time and/or dates. So quite a few to deal with. At this point the decision to drop several columns was made to simplify the dataset, and the ones kept were dealt with using hot_encoding, you can follow all the details on the Cleaning and Preprocessing notebooks. So now let's separate the models into ML and Deep Neural Network models and go through what was done with each:

›

Various methodology can be applied to implement the system that predicts the delay in flight. Few of those methodology are discussed below.
Decision Tree: As the name suggest the main idea behind decision tree algorithm is to make a tree like structure and get the answers in form of true of false. The model begins from a root node and ends on the decision. Each node receives a Yes No question and answer is passed on to the next node. Root node gets all the input of the training dataset.

The challenge to assembling such a tree is that question to ask at a node and when. To do this, decision tree algorithmic program uses accepted indices like entropy or Gini-impurity to quantify an uncertainty or impurity related to an explicit node. Equations (1) and (2) show however entropy and Gini impurity are calculated, severally, for a setoff information. Within the equations, C is that the variety of classes[1].

$$H(s) = -\sum_{c \in C} p(c) \log p(c)$$

$$H(s) = 1 - \sum_{c \in C} p(c)2$$

Logistic Regression: Logistic regression an algorithm that performs classification using,

$$h\theta(x) = g(\theta Tx) = 1 + 1 / e^{-\theta Tx}$$

Which in turn represents the maximum likelihood of estimation and gradient ascent

Logistic regression is the applicable multivariate analysis to conduct once the variable is divided (binary). Like all regression analyses, the logistical regression is a predictive analysis. logistic regression is employed to explain data and to explain the relationship between one dependent binary variable and one or additional nominal, ordinal, interval or ratio-level independent variables. Sometimes logistic regressions are tough to interpret; the Intellects Statistics tool simply permits you to conduct the analysis, then in plain English interprets the output

Neural Network: Neural Network is made by stacking along multiple neurons in layers to provide a final output. Initial layer is that the input layer and therefore the last is that the output layer. All the layers in between is named hidden layers. every nerve cell has an activation function. a number of the popular activation functions are Sigmoid, ReLU, tanh etc. The parameters of the network are the weights and biases of each layer. The goal of the neural network is to search out the network parameters specified the expected outcome is that an equivalent as the ground truth. Back-propagation on loss-function is employed to search out the network parameters [1]

Algorithm
Decision Tree
Logistic Regression
Neural Network

Precision
.93

.92
.91
Classification Report of Decision tree, logistic regression and Neural Network Classifiers.
Number of test samples used to generate the reports is 15001.
Data patameters used for the algorithms are Month, Day, Day of the week, Flight Number, Origin airport, Destination Airport, Scheduled departure, departure delay, taxi-out, distance, Scheduled Arrival.
These data features are the ones which are usually known beforehand.
Fig 4[1]: Receiver Operating Curves for Decision Tree, Logistic Regression and Neural Network models
Figure 4 shows the receiver operating curves (ROC) for all three classifiers with an area under the Curve. The observation here is that decision tree classifier turns out to be better at predicting on time flights whereas performance of neural network has be better at delayed flight's prediction.The difference is, however, very small[1].
For a balanced dataset. whereas the accuracy of the most effective algorithmic rule for a two hour prediction with a sixty minute threshold is 93.7%, even a naïve classifier that continually predicts a delay below the edge can provide an accuracy of 93.5%.