Application of Machine Learning Algorithms to Predict Flight Arrival Delays

Sofya Torosyan

Email: sofyatorosian@gmail.com

Abstract – Flight delays in aviation have high impact on environment, passengers, airports and airlines. It is very essential to have some confidence/knowledge whether flight will be delayed or not. In this paper I did research from Data Science perspective, applied machine learning algorithms like logistic regression, decision tree, random forest, support vector machine and neural networks classifiers to predict arrival delays.

1 INTRODUCTION

Through years the role of air transportation system is increasing as more and more people prefer it. And this demand has led to serious capacity problems, so efficient management of existing airside resources seems to be the most effective approach to solve these problems [1]. Flight delays are responsible for large economic and environmental losses.

Flight delays are highest indicators for air transportation system. Delay may be represented as difference between scheduled and real times (e.g. flight can be considered to be delayed if that difference > 15 . The US FAA considers this calculation.). The economic impact of flight delays for domestic flights in US is estimated to be more than $19 Billion per year to the airlines and over $41 Billion per year to the national economy. [2]

My data contains information about airlines, airports, detailed description of flights including home and destination airports, scheduled/delayed times different delay reasons, etc. Before applying machine learning models I did data exploration using various visualization tools/libraries to get thoroughly insight of our dataset. After it machine learning algorithms like logistic regression, decision tree, random forest, support vector machine and neural networks classifiers to predict arrival delays and did comparative analysis based on different evaluation metrics.

2 DATASET AND FEATURES

I used publicly available Kaggle dataset for United States domestic air traffic for 2015 year. The flight delay and cancelation data was collected and published by the DOT's Bureau of Transportation Statistics [3]. Data consists from 3 datasets: "airlines", "airports" and "flights" in .csv file formats. Data contains 5 million examples with 30 features:

- Airlines: Airline identifier for airports, airport's names.

- Airports: Location identifier, airport's name, Latitude, Longitude, etc.
- Flights: Year, month, day of the trip, airline, flight number, origin airport, scheduled/delayed times etc.

First of all investigation was done for data completeness. As demonstrated in figure 1 flight data contained Nan values (~4 million) for the following features: Cancelation_reason, air_system_delay, security_delay, airline_delay, late_aircraft_delay and weather delay. As almost all values for this features were Nan instead of filling with appropriate values I removed these columns as they couldn't give any help to predict our arrival delays. For features like tail number missing values were impossible to retrieve, but for features like arrival/departure delay it was possible using scheduled and real times. But instead of filling these missing values I simple removed them as one of main stores was that my data was very large (~5 million), so for computational reasons I used representative sample from data, and missing values were not so much that it could impact on results.

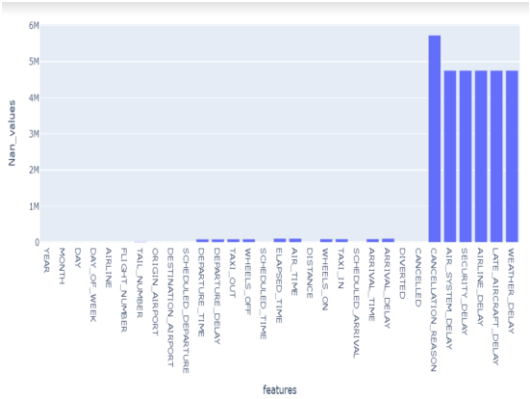


Fig. 1. Figure shows the fraction of Nan values for each feature in flights data

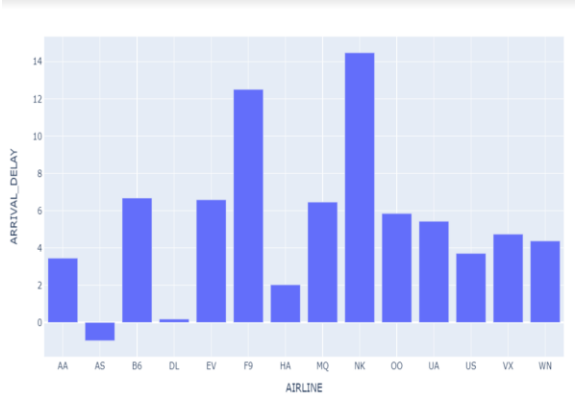


Fig. 2. Figure shows the fraction of delays (arrival) grouped by airlines.

Figure 2 shows fraction of flights delayed in the year 2015, grouped by airlines. For example DL is Delta Air Lines Inc., for which delayed flights were not so much because flights were very little. Figure shows that most of flight delays were from Spirit Air Lines, Frontier Airlines Inc.

2.1 TRAINING DATA AND FEATURE SELECTION

The main aim of my project is to classify whether flight was delayed or not. Dataset contained 30 features, but after investigation many features were dropped, I used 3 methods: ML algorithms were applied to data with some dropped features, to data created from PCA with 7 features and to data with 4 features which had highest importance score(given by Decision tree classifier). As shown I Figure 3 90% of variance is explained or retained just by using 7 features/dimensions. So instead of using 19 dimensions for modeling, I experimented ML models also on 7 features received by PCA analysis.

Also as mentioned above I took sample (100000 examples) from data so that for 50000 examples flights were delayed, and for 50000 examples not delayed. As I already had Arrival delay in flights, I considered flight delayed if it was negative, otherwise not delayed.

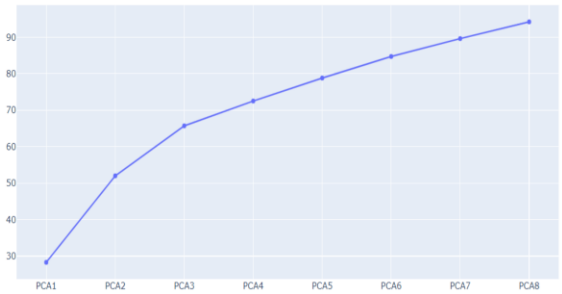


Fig.3. Figure shows marginal variances of PCA components

3 METHODS

At first dataset was split into 70:30 training and test datasets and 18 features were used.

In this section following algorithms will be introduced: logistic regression, decision tree, random forest, neural network, SVM.

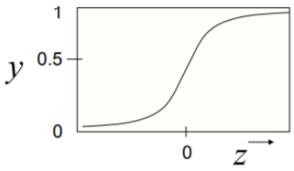
3.1 LOGISTIC REGRESSION

Logistic regression is a classification algorithm that where sigmoid is applied to a linear function of the data

$$Y(x) = \sigma(w^T x + w_0)$$

Where sigmoid is defined as

$$\sigma(z) = 1/(1 + e^{-z})$$



The output is a smooth function of the inputs and the weight. [4]

3.2 DECISION TREE

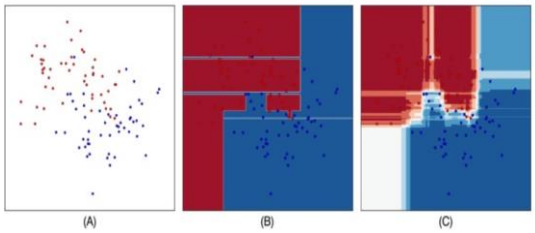
Decision tree learning is a supervised machine learning technique for inducing a decision tree from training data. A *decision tree* (also referred to as a classification tree or a reduction tree) is a predictive model which is a mapping from observations about an item to conclusions about its target value. In the tree structures, leaves represent classifications (also referred to as labels), nonleaf nodes are features, and branches represent conjunctions of features that lead to the classifications. Decision trees are grown using training data. Starting at the root node, the data is recursively split into subsets. In each step the best split is determined based on a criterion. Commonly used criteria are *Gini index* and *entropy*: [5]

$$Gini\ index: G(E) = 1 - \sum_{j=1}^{C} p_j^2,$$

$$entropy: H(E) = -\sum_{j=1}^{C} p_j \log p_j.$$

3.3 RANDOM FOREST

The prediction with decision trees is very fast and operates on high-dimensional data. On the other hand, a single decision tree has overfitting problems as a tree grows deeper and deeper until the data is separated. This will reduce the training error but potentially results in a larger test error. Random forests address this issue by constructing multiple decision trees. Each decision tree uses a randomly selected subset of training data and features. The output is calculated by averaging the individual decision tree predictions. As a result, random forests are still fast and additionally very robust to overfitting. An example of a decision tree and a random forest is shown in figure: [5]



3.4 NEURAL NETWORKS

Neural Network is built by stacking together multiple neurons in layers to produce a final output. First layer is the input layer and the last is the output layer. All the layers in between is called hidden layers. Each neuron has an activation function. Some of the popular activation functions are Sigmoid, ReLU, etc. The parameters of the network are the weights and biases of each layer. The goal of the neural network is to learn the network parameters such that the predicted outcome is the same as the ground truth. Back-propagation along loss-function is used to learn the network parameters. In my project I used ReLU for input and hidden layers and sigmoid for output layer and categorical_crossentropy and MSE loss functions. [5]

$$\sum_{i=1}^{N} \frac{(w^T x(i) - y(i))^2}{n}$$

Mean squared error loss function.

$$L(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}\left(y_{ij} * \log\left(\hat{y}_{ij}\right)\right)$$

Categorical cross entropy function.

3.5 SUPPORT VECTOR MACHINE

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N – the number of features) that distinctly classifies the data points. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. The loss function that helps maximize the margin is hinge loss.

$$c(x, y, f(x)) = \begin{cases} 0, & if\ y * f(x) \geq 1 \\ 1 - y * f(x), & else \end{cases}$$

$$c(x, y, f(x)) = (1 - y * f(x))$$

Hinge loss function (function on left can be represented as a function on the right)

$$\min_{w} \lambda\ \|w\|^2 + \sum_{i=1}^{n}(1 - y_i\langle x_i, w\rangle)$$

Loss function for SVM

4 RESULTS AND DISCUSSION

As mentioned above I used 100000 sample from initial dataset and split it into train and test sets (70:30). Used Keras API for neural networks and scikit learn for other models. In the end using decision tree classifier I took 4 features with highest importance scores as shown in below table and used only those for all models. It was interesting (and maybe evident) that departure delay has the highest impact on arrival delay, which can be seen both from Figure 4 and Table 1.

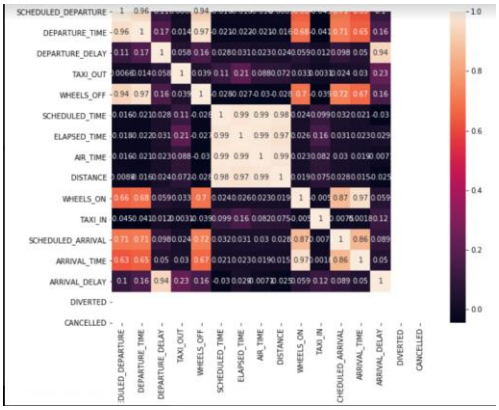


Fig. 4. Figure shows high correlation between departure and arrival delays

Table 1

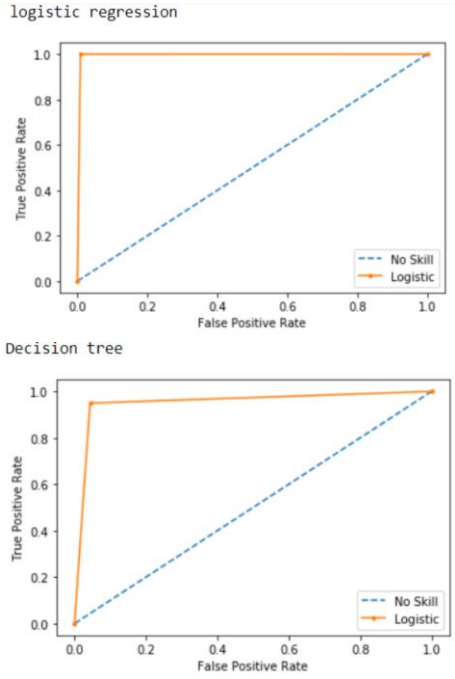
Top 4 features and their importance scores

Departure delay	0.377
Scheduled time	0.19
Elapsed time	0.16
Taxi out	0.09

For logistic regression accuracy was 0.99%. I used Grid Search with 4 fold cross validations to get best criterion and maximum depth for decision tree. Using ‘entropy’ criterion and 36 maximum depth got 0.95% accuracy for decision tree.

Table 2
Accuracies for all applied ML models

Logistic regression	0.9956
Decision tree classifier	0.9531
Random forest classifier	0.948
Support vector machine	0.98
Neural networks	0.5



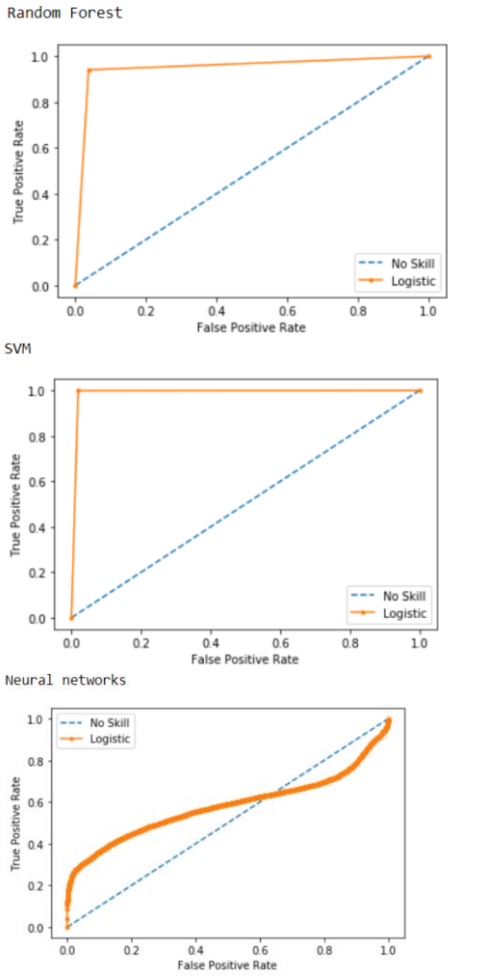


Fig. 5. Figure shows ROC curves for applied models

5 CONCLUSION AND FUTURE WORK

In this project I successfully applied 5 machine learning models to predict whether flight will arrive delayed or not using simple classifiers like logistic regression, decision tree, random forest, which gave very accurate results. As I used sample from initial data in future I will do the same on the whole data using Apache Spark unified analytics engine which is very convenient for big data preprocessing and applying machine learning models.

6 REFERENCES

[1] Cetek, C., Cinar, E., Aybek, F., & Cavcar, A. "Analysis of aircraft ground traffic flow and gate utilisation using a hybrid dynamic gate and taxiway assignment algorithm".

[2] K. B. Nogueira, P. H. Aguiar, and L. Weigang, "Using ant algorithm to arrange taxiway sequencing in airport," International Journal of Computer Theory and Engineering, vol. 6, no. 4, p. 357, 2014.

[3] Kaggle dataset https://www.kaggle.com/usdot/flight-delays

[4] Richard Zemel, Raquel Urtasun and Sanja Fidler "Logistic Regression"

[5] Vikrant A. Dev, Mario R. Eden, in Computer Aided Chemical Engineering, 2019 "Decision Trees"

[6] Nathalie Kuhn, Navaneeth Jamadagni "Application of Machine Learning Algorithms to Predict Flight Arrival Delays"