Application of Machine Learning Algorithms to Predict Flight Arrival Delays

Sofya Torosyan

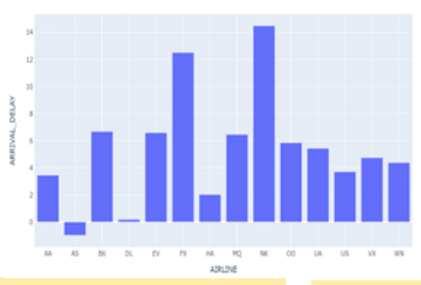
ABSTRACT

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. I applied machine learning algorithms like logistic regression, decision tree, random forest, support vector machine and neural networks classifiers to predict arrival delays.

DATA AND FEATURES

The flight delay data was collected and published by the DOT's Bureau of Transportation Statistics. Data contains 5 million examples with 30 features. In the end the following features with highest importance score were used for applying ML algorithms:

Departure delay, Scheduled time, Elapsed time, Taxi out

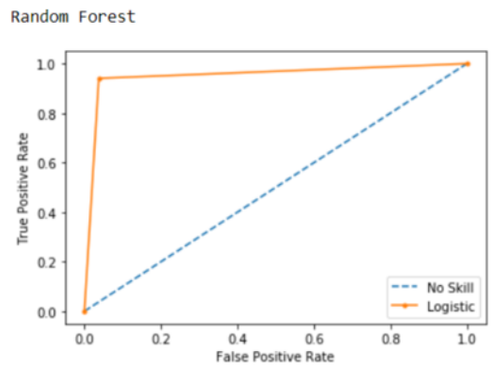


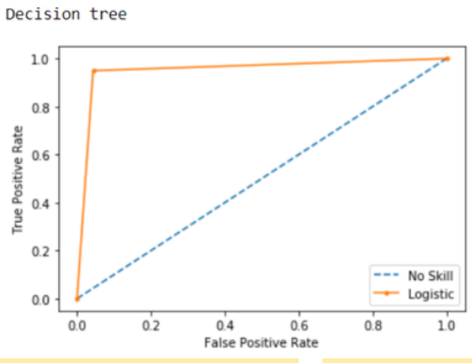
MODELS

Applied machine learning algorithms using scikit learn/keras API s.

Model	sklearn/keras method
Logistic Regres.	Log. Regression for binomial prediction
Decision Tree	DecisionTreeClassifier
Random Forest	RandomForestClassifier
SVM	SVC with gamma = 'auto' option
Neural Network	Neural network with SGD and ADAM optimizers

RESULTS and DISCUSSION





My final decision tree on 4 features with highest importance scores has 36 max depth and used 'entropy' criterion.

For all models except NN got more than 90% accuracy.

INTRODUCTION

Flight delays are responsible for large economic and environmental losses. They 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.

RESULTS

Input data: 100000 samples, 70% for train, 30% for split.

Features: Below are mentioned 4 features having highest score(received from DecTree)

DEPARTURE_DELAY	0.377
SCHEDULED_TIME	0.19
ELAPSED_TIME	0.16
TAXI_OUT	0.09

ACCURACIES

Logistic Regression	0.9956
Decision Tree	0.9531
Random Forest	0.948
SVM	0.98
Neural Network	0.5

FUTURE WORK

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.

REFERENCES

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".

Vikrant A. Dev, Mario R. Eden, inComputer Aided Chemical Engineering, 2019"Decision Trees"