Explanation of Results from Technical Perspective

Based on EDA machine learning and modeling supposed to be focused on determining time of flight delay versus 6 parameters: carrier, distance of flight, time of departure, airport of origin, tail number and day of week. All our variables are categorical (even distance and hour of departure were set to some abstract numbers based on the bin we applied for the sake of clarity). Thus way among 4 models of regression analysis two of them are suitable candidate for the regression: Decision Tree and Random Forest.

To apply corresponding Sklearn models we had to parse categorical values to dummy numerical ones using methods of pandas library. While this step variable responsible for tail number (that counts 5366 unique values) could not be processed with an “Memory error”. Thus way we had to omit this variable for further analysis. For the other 5 variables we applied DecisionTree Regressor and RandomForestRegressor with two different sets of attributes (see RegModels_Korchagina.ipynbv). To ensure that our choice of test and training parts of datasets are not luckily chosen but are random and independent we used both conventional splitting and training with cross validation. In the former choice first 80% of dataset were used for training and last 20% for testing; in the latter case we choose number of splits and then consequently use each part for testing and the rest for training.

For all our models (independently on chosen attributes) we obtained extremely similar results. All absolute errors and standard deviations are about 34 min and 53 min^2, correspondingly. Moreover, evaluation of applied models also revealed identical results between all methods applied. Whiskers plots indicates the same distribution of quartiles, mean and outliers and demonstrate positive skewness which shows that more than half of flights under study has smaller delay time that the mean value (there are more shorter delayed flights than longer delayed ones).

On the top of that order and distribution of feature importance obtained from each mode are the same and shows following: Carrier (0.72), Origin (0.17), Distance (0.06), DayOfWeek (0.05) and SchedDepTime (0.001) (Fig,1). In order to have better idea about feature importance we attempted to create bar plots for all parameter. However, variable “Origin” has 303 unique values and its plotting was restricted by Python settings. Thus we ended up with plotting remaining four graphs showing how each of them affects mean delay time (Fig.2). One can see that among four plots two demonstrate the most variance in data. One of the is “Departure Delay vs Carrier” which is the most influential attribute and another one is “Departure Delay vs Scheduled Departure Time”. The latter unusual behavior can be attributed to outliers that are clearly presented on Whiskers plot.

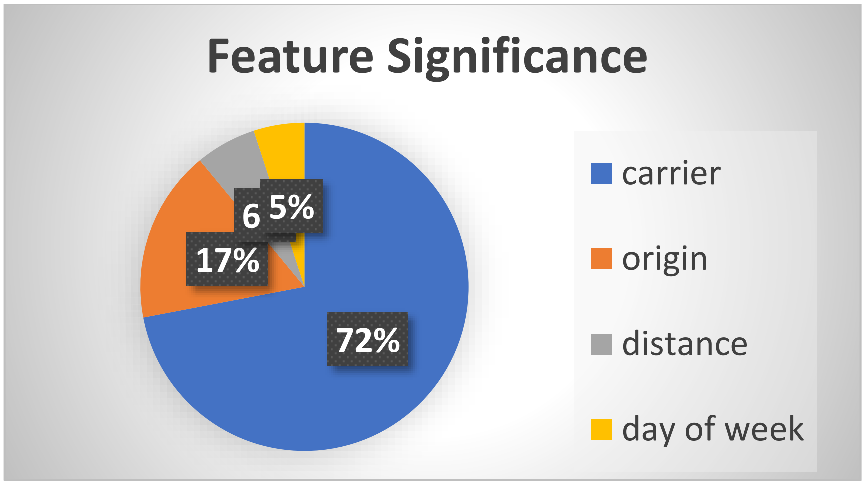

Fig1. *Pie chart representing feature significance distribution based on its contribution to influence on flight delay time. The least feature SchedDepTime (0.001) is not presented on the chart.*

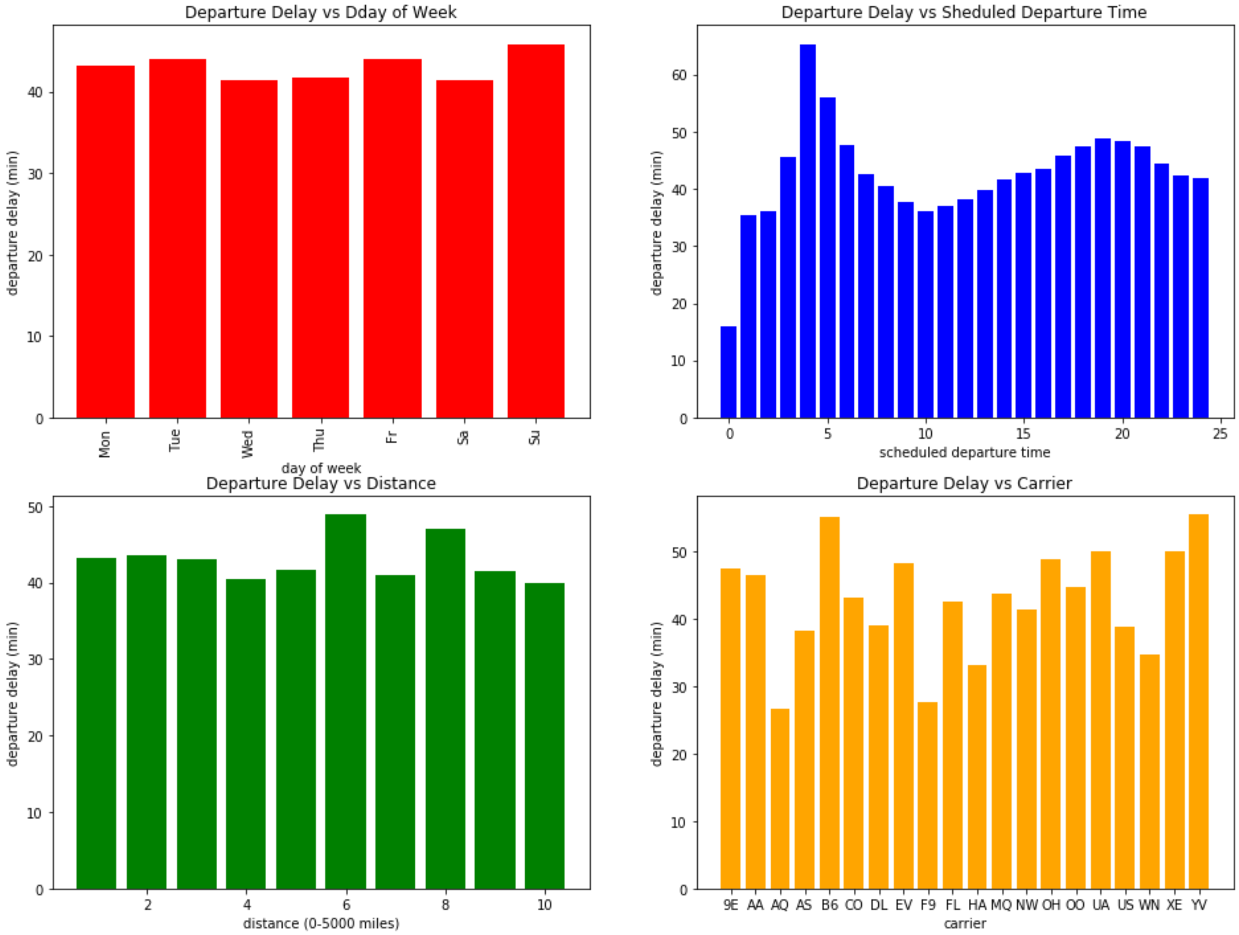

Fig2. *Bar plots showing dependence of flight delay time on flight' attributes ("Day of Week" – top left, "Distance" – bottom left, "Scheduled Departure Time" – top right, "Carrier" – bottom right.*

All in all, inspite of relatively high absolute error and standard deviation, we obtained good consistency of data observed by applying various regression models and different ways of dataset splitting that ensures us about stability and reproducibility of results obtained. Further step, creating an application, can involve any of the regression models to predict flight delay time. One can choose, for example, RandomForest10 (see RegModels_Korchagina.ipynbv).