Report for First Assignment of ML

Motivation:

Flight delay is inevitable and it plays an important role in both profits and loss of the airlines. An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and incomes of airline agencies. There have been many researches on modeling and predicting flight delays, where most of them have been trying to predict the delay through extracting important characteristics and most related features. However, most of the proposed methods are not accurate enough because of massive volume data, dependencies and extreme number of parameters. This report contains three models for predicting flight delay based on some important features. But the most of flight delay data are noisy. So these models are not accurate. we need to modify these models or use other regression models or using deep learning to solve this problem with high accuracy.

Task Definition:

First, we need to preprocess the training data before applying any machine learning model:

1. Reading the data set.
2. Removing the outliers using the plot box.
3. Sorting the data set.
4. Calculate Flight Duration.
5. Scaling using Min Max Scaler and Standard Scaler
6. Create new data frame by selecting important features (Flight Duration, Departure Airport, Destination Airport, Delay)
7. Splitting the data to train and test such that the data is split based on Scheduled departure time. The train data is all the data from year 2015 till 2017.All the data samples collected in year 2018 are to be used as testing set.
8. Reducing dataset to 2D and 3D using PCA and visualize it after that.

Second, I estimate the flight delay time using (Multiple Linear Regression, Polynomial Regression, Lasso, SVR) based on independent predictors (Flight Duration, Departure Airport, Destination Airport) and dependent target(Delay).

Third, I measure the performance of these models using(R2 Score, MSE, MAE) and visualized the training data, Predicted training data, testing data and Predicted testing data to see the training error and testing error.

Data Description:

The Dataset comes from Innopolis University partner company analyzing flights delays. Each entry in the dataset file corresponds to a flight and the data was recorded over a period of 4 years. These flights are described according to 5 variables. A sneck peek of the dataset can be seen in the table below:

Departure Airport	Scheduled departure time	Destination Airport	Scheduled arrival time	Delay (in minutes)
SVO	2015-10-27 09:50:00	JFK	2015-10-27 20:35:00	2.0
OTP	2015-10-27 14:15:00	SVO	2015-10-27 16:40:00	9.0
SVO	2015-10-27 17:10:00	MRV	2015-10-27 19:25:00	14.0
MXP	2015-10-27 16:55:00	SVO	2015-10-27 20:25:00	0.0
...	...	...	...	...

The description of the 5 variables describing each flight are:

Variable name	Description
Departure Airport	Name of the airport where the flight departed. The name is given as airport international code
Scheduled departure time	Time scheduled for the flight take-off from origin airport
Destination Airport	Flight destination airport. The name is given as airport international code
Scheduled arrival time	Time scheduled for the flight touch-down at the destination airport
Delay (in minutes)	Flight delay in minutes

Outlier Detection & Removal:

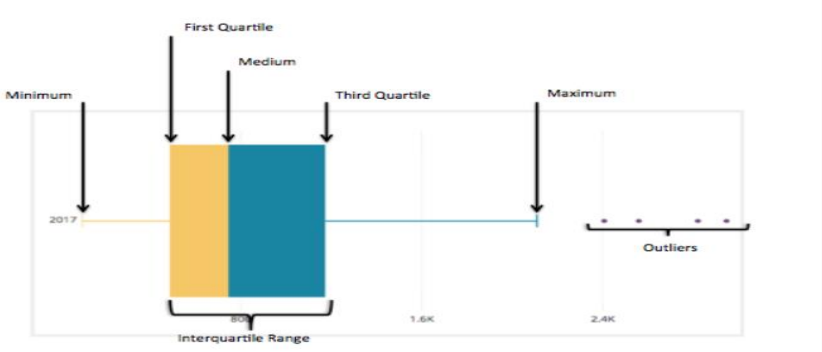
1. Outlier Detection

A Box Plot is the visual representation of the statistical five number summary of a given data set.

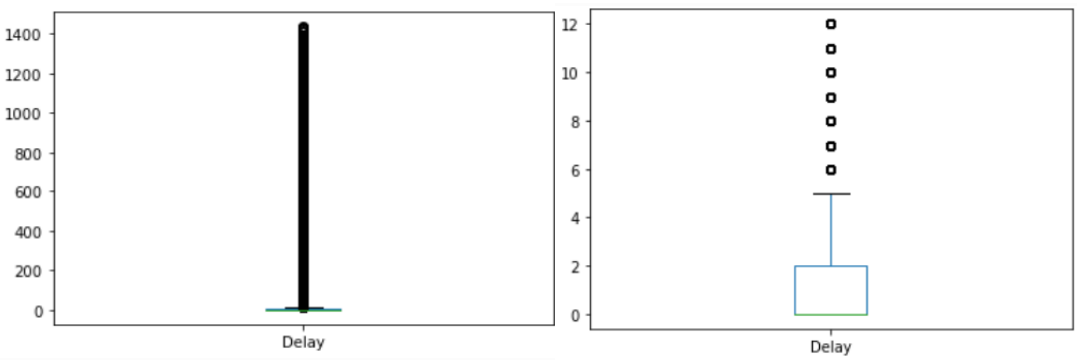
A Five Number Summary includes:

a. Minimum
b. First Quartile

c. Median (Second Quartile)
d. Third Quartile
e. Maximum



I use the box plot to show the outliers which are exist the delay feature.

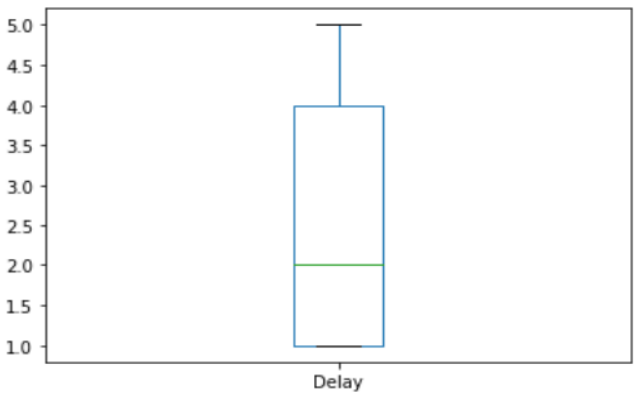


there are many outliers in the box plot. So I use this equation to remove the outliers from the data set.

The maximum is less than the third Quartile (Q3) + 1.5*IQR *and the minimum is greater than the first Quartile (Q1)-1.5*IQR*. *(IQR=Q3-Q1)*

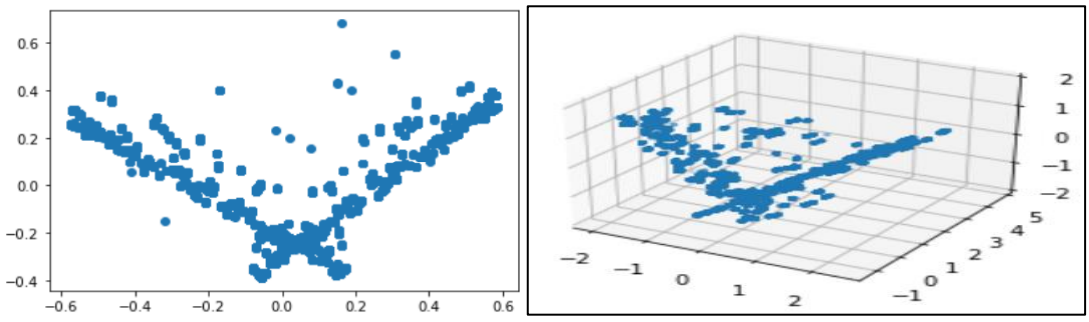
2. Outlier Removal

The box plot of the delay feature after removing the outliers.



Reducing dimensions and Visualizing the Data set:

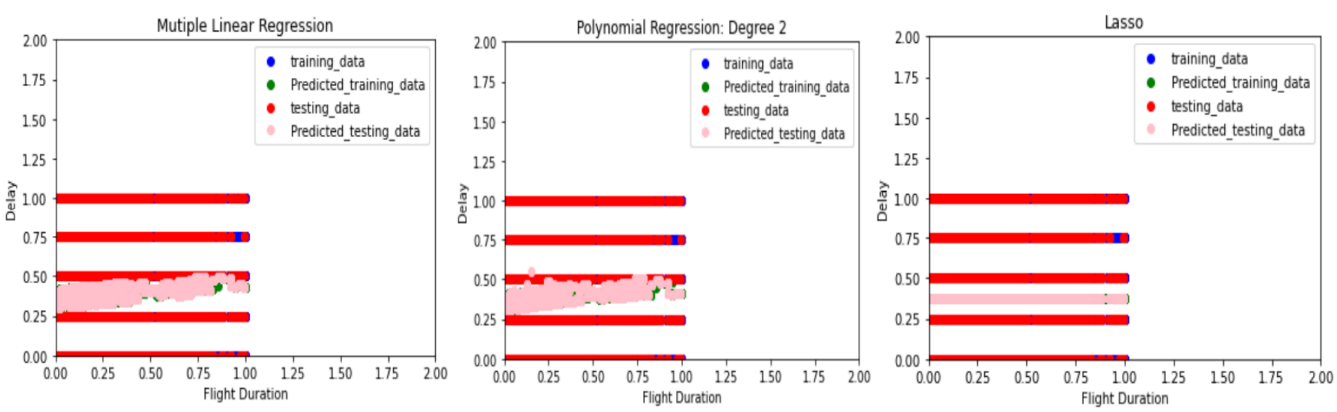
Using PCA to reduce the dimensions of the dataset to 2D or 3D.



2D 3D

Comparison between Machine Learning Models:

Model	Multiple Linear Regression	Polynomial Regression	Lasso
Results	Model intercept : 0.3614446130273279 Model coefficients : [-0.07807069 0.07231831 0.11238985] Accuracy using R2 Score: -0.004059079212486161 Mean Squared Error:0.1225169637073714 Mean Absolute Error:0.30227835804957337	Model intercept : 0.6160187046326249 Model coefficients : [0. -0.34679555 -0.32984052 0.22901239 -0.04324996 0.40079253 -0.06688352 0.08115379 -0.03810591 -0.06372956] Accuracy using R2 Score: -0.003509193970666402 Mean Squared Error:0.12244986579290595 Mean Absolute Error:0.3022287586930065	Lasso model Coefficients: [-0. 0. 0.] Accuracy using R2 Score: -0.015995762621510368 Mean Squared Error:0.1239734977284141 Mean Absolute Error:0.30522272126254846



The lasso is better than the multiple linear regression and polynomial regression such that the lasso have the highest accuracy.

But in general the accuracy is very bad in the three models because the training error and testing error is very high (shown in the graphs) Because the model can't learn from the data (Underfitting) in three models.