Flight Delay Forecasting

Siba Issa

September 24, 2021

Abstract

This report is part of Machine Learning [ML] course for first year master students at Innopolis University. This report is stating the methods that I used in order to solve the task of flight delay estimation using machine learning while the code is available on GitHub.

1 Introduction

In this assignment, I am going to solve the task of flight delay estimation using machine learning. The process of solving this task consists of three main stages:
- First of all, we need to understand our problem and to be able to read our data deeply and to know the importance of each feature and how it will contribute in the solution. In order to do that we need to derive our dataset and get all the information from it, then we will prepossess our data and visualize it.
- Second stage, is to apply our selected machine learning models on our data.
- And in the end, we have to analyze our results and evaluate the used models.

2 Dataset

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 as it shown in (Figure **??**) (you can find the complete dataset here.).

	Depature Airport	Scheduled depature time	Destination Airport	Scheduled arrival time	Delay
0	SVO	2015-10-27 07:40:00	HAV	2015-10-27 20:45:00	0.0
1	SVO	2015-10-27 09:50:00	JFK	2015-10-27 20:35:00	2.0
2	SVO	2015-10-27 10:45:00	MIA	2015-10-27 23:35:00	0.0
3	SVO	2015-10-27 12:30:00	LAX	2015-10-28 01:20:00	0.0
4	OTP	2015-10-27 14:15:00	SVO	2015-10-27 16:40:00	9.0

Figure 1: first five rows from flightdelay Dataset

3 Data Preprocessing and Visualization

This task is to to estimate the flight delay in which makes it a regression task. And according to our dataset We have 4 predictors (Depature Airport, ScheduScheduled depature time, Destination Airport and Scheduled arrival time) and one target (Delay). As we see in (Figure 3) all the predictors are objects and to be able to reprocess these data we need to represent it from string to the machine-readable format by using Label encoder.

```
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Depature Airport        675513 non-null  object
 1   Scheduled depature time 675513 non-null  object
 2   Destination Airport     675513 non-null  object
 3   Scheduled arrival time  675513 non-null  object
 4   Delay                   675513 non-null  float64
dtypes: float64(1), object(4)
```

Figure 2: Data types

We used LabelEncoder to transform Departure and Destination airports from strings to unique integers. Meanwhile we reprocessed the time columns, because it contains more than one feature. Furthermore, we might notice that there is specific months where the flights delay more than others (for instance, the weather in winter may affects on flight delay) and we can apply the same aspect on days and years. So we divided the 'scheduled Departure time' into (Departure year, Departure month and Departure day of the week) and we computed the duration of the flight (because delay depends on flight duration). After prepossessing our data, we visualize it and for simplicity we plotted the delay (target) against flight duration which is the time difference between departure and arrival. (Figure 3).

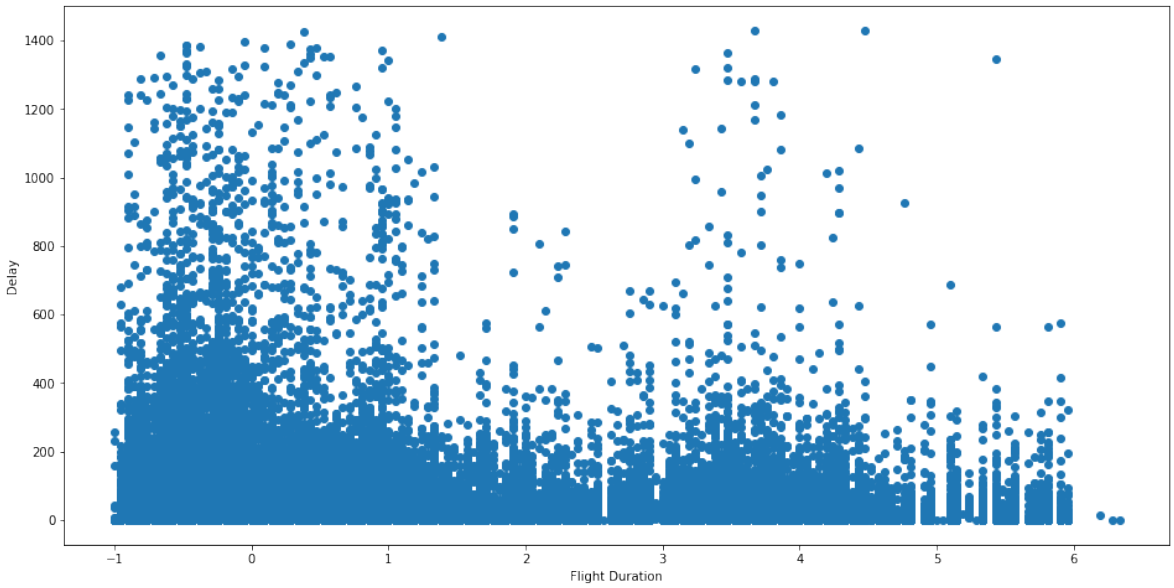

Figure 3: Delay against Flight Duration

As it is clear from above that there is a few number of flights with big delay and the more the flight lasts the less it delays. But it is clear also that the relation between Flight Duration and Delay is not linear and there is so many outliers.

4 Outlier Detection Removal

An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data. Wikipedia defines it as 'an observation point that is distant from other observations'. I used the mathematical function Z-score to discover outliers. The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. First we tested the existence of the outliers on the first month from the data then we generalized on the whole dataset. We detected 95 outliers in the first 2441 samples (around 3.89%-symbol) and

19970 outliers in the whole dataset (around 2.96 %-symbol). After detecting the outliers we removed it and plot it (Figure 4).

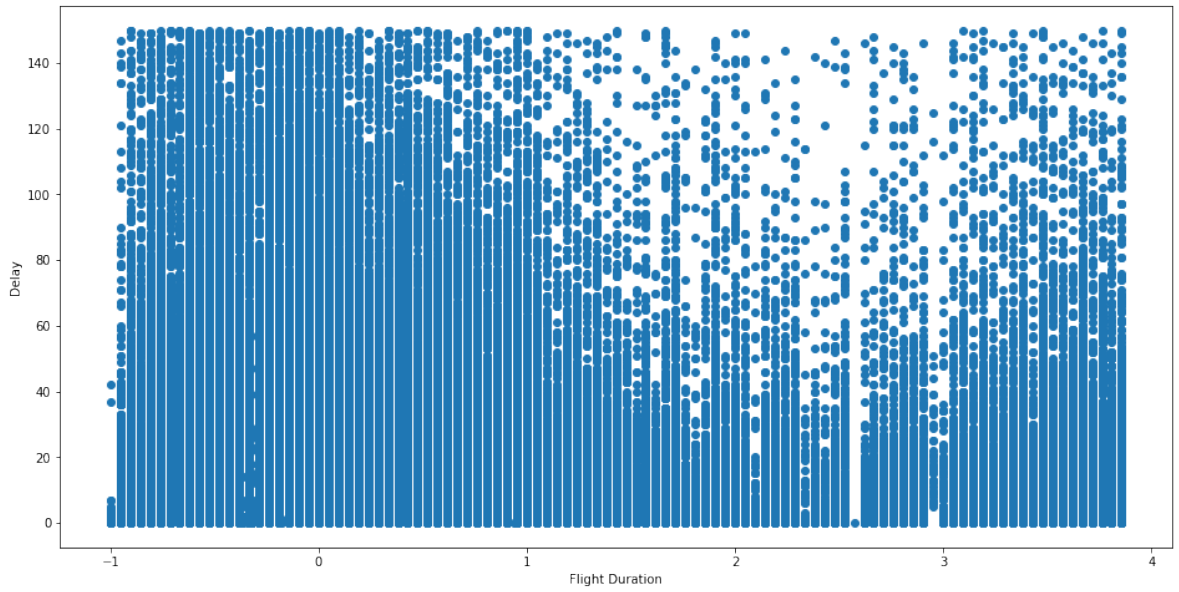

Figure 4: Delay against Flight Duration after Outliers removal

5 Machine learning models

Before using our regression models to train the data, the data should be split to train and test. The data is split based on 'Departure year'. The train data is all the data from year 2015 till 2017. All the data samples collected in year 2018 are used as testing set. Regarding that our task is a regression task we used the following three models to train our data:

5.1 Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

5.2 Polynomial Regression

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_2 x_1^3 + ... + b_n x_1^n$$

5.3 Lasso regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (ex: models with fewer parameters). Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from

the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. Lasso solutions are quadratic programming problems and the goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_j(x_{ij}\beta_j))^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

6 Performance Measurement

In order to measure the performance of our selected models we used these metrics:
- Mean Absolute Error
- Root Mean Squared Error
- $R_s^2 core$

Model	Linear Regression d	Polynomial Regression [degree 1]	Lasso Regression
Mean Absolute Error	11.400016	11.399956	11.397266
Root Mean Squared Error	39.98458	39.98458	39.98578
$R_s^2 core$	0.0093	0.002498	0.002439

Table 1: DH- Parameters

6.1 Conclusion

As we saw in the previous paragraph, all the three models are underfitting. And that is due to using Linear models on a non-linear relationship between the Delay and the Fly Duration. So, to enhance the performance we need to do better prepossessing and extract more useful features and off course we need to use more complex models.