INNOPOLIS UNIVERSITY

"Assignment 1"

FLIGHT DELAY FORECASTING

By:

Ahmed Mohsen Mohamed Abdelkhalek Elsayed Ali

Date

22-sep-2021

Contents

1) Motivation

This report aims to show the results of different Machine Learning algorithms applied to the flight delay forecasting problem. A brief description is given about the data and then more concise explanation about the ML algorithms is provided. First, Preprocessing data is explained which includes splitting, outlier detection, and extracting some additional feature such as flight duration. Second, different approaches were implemented such Linear Regression, Multi Linear Regression, and Polynomial Regression. Also, Lasso was used as Regularization technique to prevent overfitting. At the end, results will be shown and discussed along with some comments about the data itself.

2) Task Description

The task of this report is to predict flight delay given dataset from one of Innopolis University Partners. The task is implemented as follows:

- **Preprocess the data:**
  - Converting some string features into some useful features such as "Arrival Airport" and 'Departure Airport' are converted to "Flight Duration".
  - **Visualizing** the data by plotting single critical feature such as 'Flight Duration' against "Delay".
  - **Encoding** the categorical features using Label Encoder
  - **Splitting** the data into train and test set in which the train data is all the data from year **2015** till **2017** and test data are all data collected in year 2018.
  - **Features Scaling** using MinMaxScaler
  - No **imputation** is needed since the data has no missing values
  - Applying **Outlier detection** using LocalOultierFactor

- **Machine Learning models:**

  More details about the models are given later in the report but in general the ML models that were used are as follow:

  - **Linear Regression** (with 3 predictors)
  - **Polynomial Regression** (with 3 predictors)
  - **Lasso Regularization** (with 3 predictors)
  - **Ridge Regularization** (with 3 predictors)

- **Results and Performance Comparison:**

  Performance Measurement methods are applied to all ML models used in the report. The main metric used are:

- **MSE**
- **Precision**
- **Recall**
- **R score**

3) Data set Description and Exploration:

The dataset comes from one of IU partners. The data contains 4 features: Departure Airport, Scheduled Departure Time, Destination Airport, and Scheduled Arrival Time. It also contains one target which is Delay (in minutes). All predictors are strings and the target is float. The dataset comes in the following shape:

Departure Airport	Scheduled departure time	Destination Airport	Scheduled Arrival Time	Delay
SVO	2015-10-27 09:50:00	JFK	2015-10-27 20:35:00	2
OTP	2015-10-27 14:15:00	SVO	2015-10-27 16:40:00	9
MXP	2015-10-27 17:10:00	MRV	2015-10-27 19:25:00	14

4) Data Preprocessing and Visualization:

The dataset needs some preprocessing before building the model to avoid any outlier affect the model and ensure better results.

4.1) Feature Engineering:

First, the “Scheduled departure time” and “Scheduled Arrival time” are converted from string to pandas datetime in order to calculate the flight duration from these two features using pandas.series.dt. By doing so, two features are combined into one critical feature without losing much data and the original two features can be ignored. For this step, only “Scheduled Arrival time” is dropped since “Scheduled departure time” is needed in a later process (Data Splitting). The resulting dataset is as follows:

	Depature_Airport	Scheduled_depature_time	Destination_Airport	Delay	flight_duration
0	SVO	2015-10-27 07:40:00	HAV	0.0	785.0
1	SVO	2015-10-27 09:50:00	JFK	2.0	645.0
2	SVO	2015-10-27 10:45:00	MIA	0.0	770.0
3	SVO	2015-10-27 12:30:00	LAX	0.0	770.0
4	OTP	2015-10-27 14:15:00	SVO	9.0	145.0

Figure 1; Dataset after features engineering

4.2) Label Encoding:

Label encoding is used to encode categorical features "Departure Airport" ," Destination Airport". Label encoding is better in this case since it does not need too much computational time such as OneHotEncoding and does not give the features certain order as OrdinalEncoder does. Both features are encoded as shown below.

	Depature_Airport	Destination_Airport
0	144	56
1	144	68
2	144	94
3	144	82
4	113	144

Figure 2; Categorical Features after Label Encoder

4.3) Data Splitting:

The dataset is splitted into trainset and testset as indicated in the task:

- Train data is all the data from year **2015** till **2017**
- Test data are all data collected in year 2018

The train data has a shape of (499062, 4) while test set has a shape of (176451, 4). Afterwards, predictors are chosen to be in x_trian and x_test and target variable "Delay" is chosen to be in y_train and y_test.

4.4) Features Scaling:

In order to decrease the range of predictors values, Feature scaler is used. In this report, MinMaxScaler is used which has the following formula:

$$Value = \frac{Value - min}{\max - min}$$

4.5) Visualizing the data:

In order to better know the data, especially after all preprocessing, plotting was made between each predictor and the target. These plots are shown below.

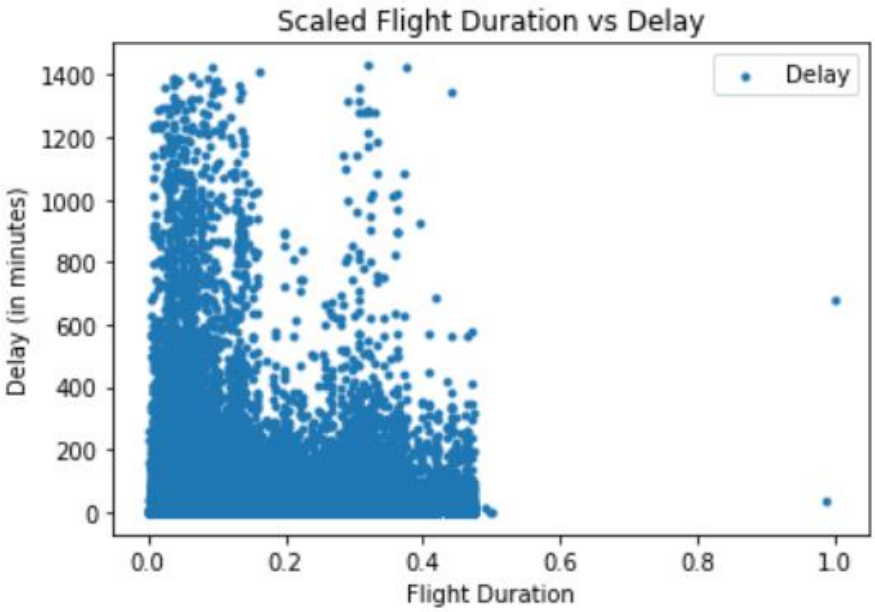

Figure 3; Flight Duration Vs Delay

This plot shows as the duration of flight increase, the probability of high delay decreases. Also, some outliers are clearly visible in the graph as the data in the far right are away from the mainstream data. This is why outlier detection is needed.

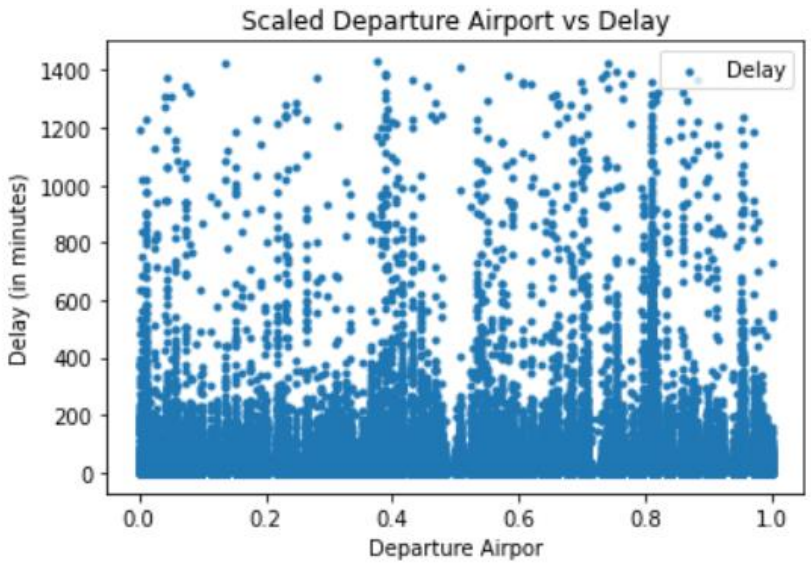

Figure 4;Departure Airport Vs Delay

This plot shows that departure airport has almost no effect on flight delay as the distribution is the same all over the plot. Outliers are visible also

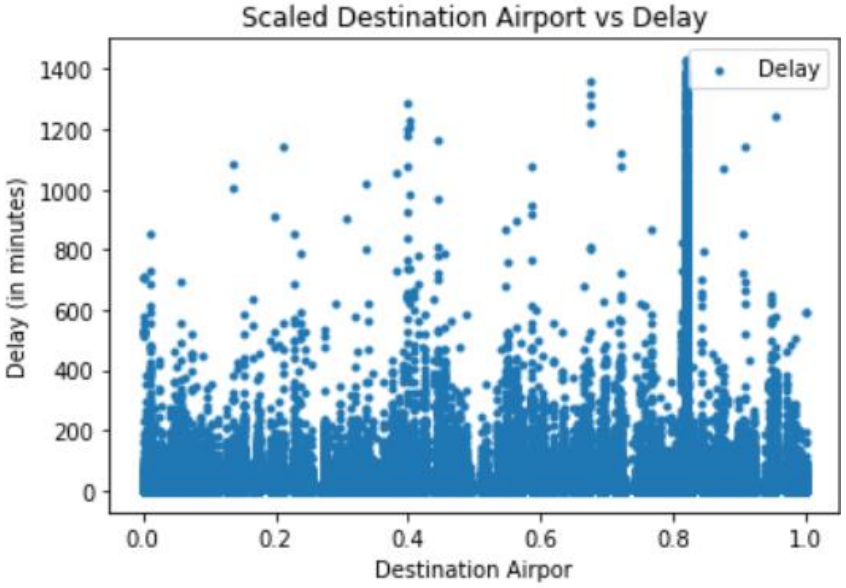

Figure 5;Destination Airport Vs Delay

This plot shows that Destination airport has almost no effect on flight delay as the distribution is the same all over the plot. There is high density about 0.81 which corresponds to SVO airport in Moscow. This is due to that half of the flights in x_train have a destination airport in 'SVO'.

5) Outlier Detection:

Before entering the dataset for ML models, outlier has to be detected and removed to get better results. LocalOutlierFactor is used to determine the outlier from x_train set and remove them both from x_train and corresponding values in y_train. The following table summarize the outlier detection on x_train.

Size Before Outliers Removal	499062
Size After Outliers Removal	498112
Number of Outliers	950

The following graph for flight duration (without Outliers) is clearly different form figure 3.

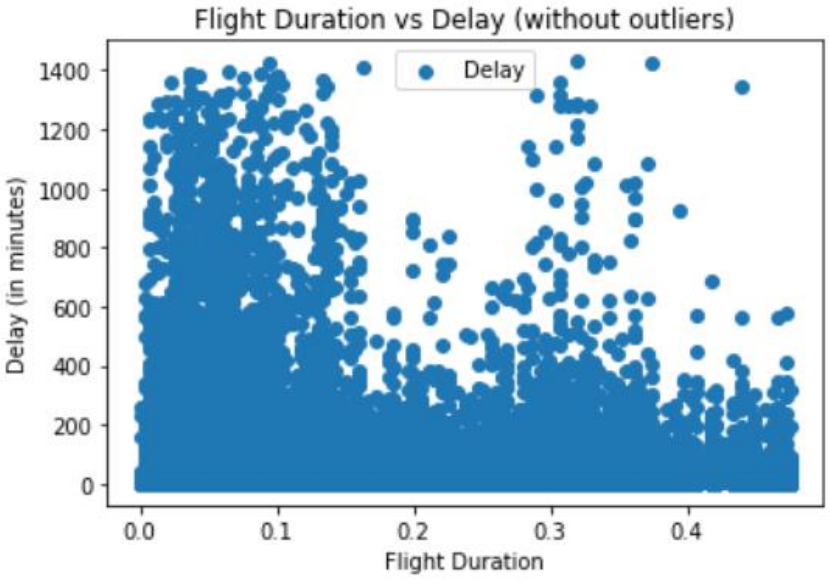

Figure 6; Flight Duration Vs Delay(Without outliers)

6) Machine Learning Models:

The ML models used in the report are: Linear Regression, Polynomial Regession, Lasso, and Ridge.

6.1) Linear Regression

The linear regression model is applied fitted the whole x_train. The model parameters are as follow:

Model Intercept	5.762
Model Coefficient	[-1.17 5.63 25.86]*

*These coefficients are for 'Departure Airport', 'Destination Airport', 'Flight Delay'

6.2) Polynomial Regression

This model is trained with different degrees. However, increasing the degrees result in overfitting (high test error). So the degrees that is shown in this report are 2nd, 3rd and 4th degrees.

6.3) Lasso Regression

The best alpha is chosen as 0.1

Best Alpha	0.1
Model Coefficient	[-.4058 4.21 10.31]

6.4) Ridge Regression

Best Alpha	0.1
Model Coefficient	[-1.1762 5.638 25.8524]

7) Results:

		Trian Error	Test Error
Linear Regression	Mean Absolute Error	15.3	14.31
	Mean Squared Error	2123	1616.86
	Root MSE	46	40.21
	R^2 Score	0.003	-0.008
Polynomial Regression (2^{nd} Degree)	Mean Absolute Error	15.3	14.31
	Mean Squared Error	2121	1615
	Root MSE	46	40
	R^2 Score	0.003	-0.008
Polynomial Regression (3^{rd} Degree)	Mean Absolute Error	15.28	14.41
	Mean Squared Error	2119	1631
	Root MSE	46	40.39
	R^2 Score	0.004	-0.0179
Polynomial Regression (4^{th} Degree)	Mean Absolute Error	15.27	80
	Mean Squared Error	2118	7368522
	Root MSE	46	2714
	R^2 Score	0.005	-4596
Lasso	Mean Absolute Error	15.38	14.4
	Mean Squared Error	2124	1619
	Root MSE	46	40
	R^2 Score	0.002	-0.01

Ridge	Mean Absolute Error	15.3	14.32
	Mean Squared Error	2123	1616
	Root MSE	46.07	40.21
	R^2 Score	0.003	-0.008

8) Comparison and Remarks:

- Using the performance results indicated above, all the models relatively yield the same results. Except for polynomial regression with degree >=4, it overfits the model, resulting in very high test error
- All the models (except polynomial with degree more than 4) exhibits underfit will high test and train errors
- R2 factor for all models are very low. This indicate the high variability of the data and that the data are scattered around the model function.
- Graphs 3,4,5 shows that each single predictor value can have multiple target values so Linear model should not be used for such data
- Graph 4,5 show that there is weak relation between some predictors and target
- Based on the coefficients of all models, Flight duration has the most effect on delay then the Destination Airport.
- Some other data are needed to establish strong model for flight delay prediction such as weather conditions.

9) Code Implementation:

The code implementation for this report can be found in the following github repository.
https://github.com/Ahmed-Mohsen-7/Machine_Learning_Assignment_1.git