

ML Assignment 1 Report

Alaa Aldin Hajjar

https://github.com/AlaaAldinHajjar/ML_A1

1 Introduction

Imagine that you booked a flight to an important event in specific time, unfortunately a delay for the departure time happened and you missed the event, how would you feel about that? That's why we tried to make a model to predict the flight delay using different ML algorithms, and the best possible results.

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. A sneak 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
MRV	2015-10-27 16:55:00	SVO	2015-10-27 20:25:00	0.0
...

Figure 1: Data Description.

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

Figure 2: Variables Description.

And we can add a new column which is the flight duration in hours and also converted the delay to hours because it is a very important feature. The data is splitted to train with years from 2015 to 2017, and test for year 2018.

3 Preprocessing Steps

3.1 Converting the dates:

These time features can be easily be extracted using `pandas.Series.dt`.

3.2 Add Flight Delay

We can calculate this from the date time.

3.3 Encode

We used Ordinal encoder.

3.4 Imputing

We used simple imputer most frequent.

3.5 Outlier Detection Removal

In statistics, If a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and about 99.7% lie within three standard deviations. Therefore, if you have any data point that is more than 3 times the standard deviation, then those points are very likely to be anomalous or outliers. We applied this for delay and flight duration. Using box plot from `sns` to visualize the outlier. from this figure we can see outlier in flight duration:

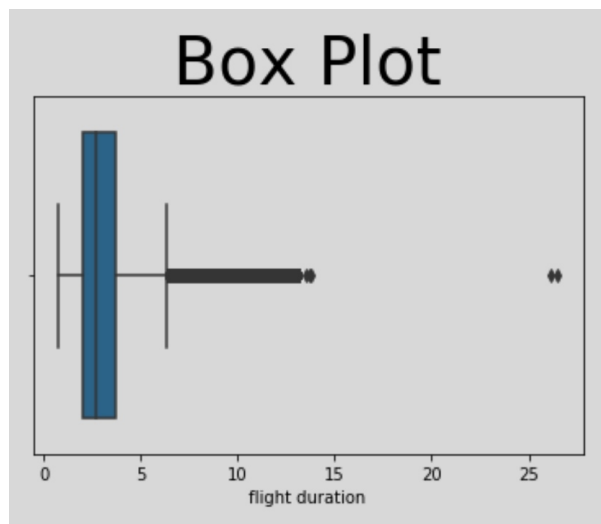


Figure 3: Outliers in Flight Duration.

And here outliers for delay:

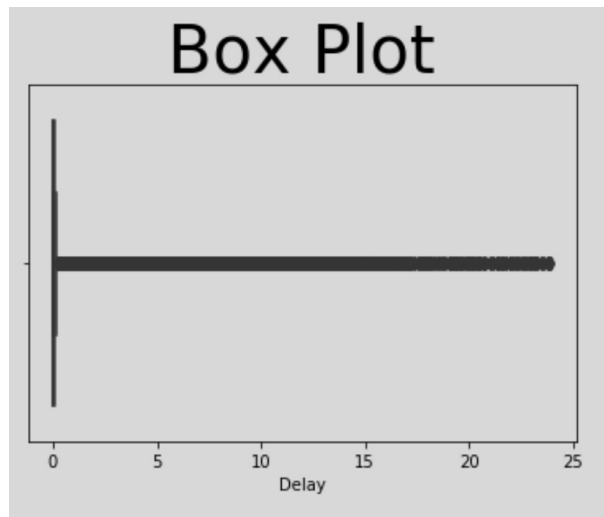


Figure 4: Outliers in Delay variable.

4 Visualization

Here, there is a visualization of the task in two dimensions taking flight duration vs delay, and it is not clear to set up in advance a type of relationship of the variables with just this chart.

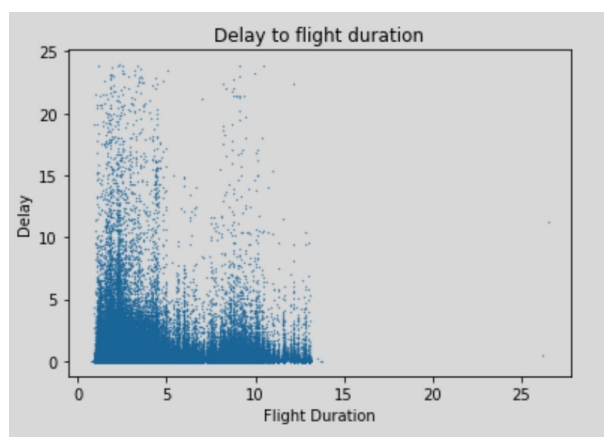


Figure 5: Delay vs FLight duration after outliers removal..

Using PCA to project the data to 2D

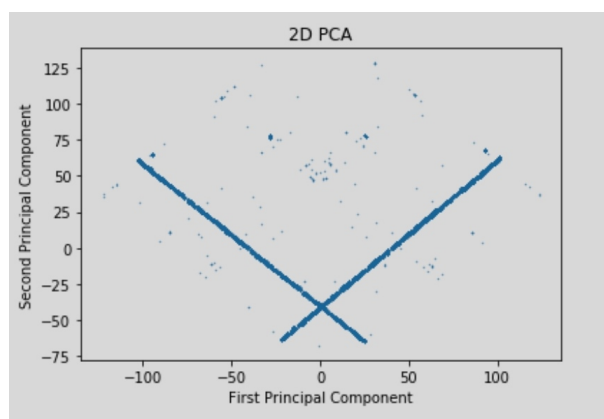


Figure 6: Delay vs Flight duration.

5 Models

Linear regression simple taking only flight duration as feature. multi linear regression with all features. lasso and ridge with regulations using different alphas. polynomial regression with one feature flight duration. polynomial regression with all features. both polynomials with degree 2. Neural network.

In this table we can see the applied algorithms and the metrics of results, the values obtained are found in the next table:

Table 1: Results obtained from the models.

	Simple Linear Regression	Multiple Linear Regression	Lasso with Regul.	Ridge with Regul.	Polynomial Regression with Flight duration as feature	Polynomial Regression all features	Using NNs
Train MSE	0.0766	0.0761	0.0772	0.0761	0.0766	0.0753	0.0909
Test MSE	0.0494	0.0516	0.0497	0.0516	0.0494	0.1258	0.0494
Train RMSE	0.2769	0.2759	0.2779	0.2759	0.2768	0.2744	0.3016
Test RMSE	0.2224	0.2272	0.2230	0.2272	0.2223	0.3548	0.2224
Train MAE	0.1535	0.1521	0.1552	0.1521	0.1535	0.1500	0.1171
Test MAE	0.1391	0.1506	0.1405	0.1506	0.1390	0.3283	0.0574
Train R2 Score	0.0074	0.0143	0.0	0.0142	0.0077	0.0253	-0.1776
Test R2 Score	-0.0711	-0.1177	-0.0771	-0.1177	-0.0699	-1.7253	-0.0715

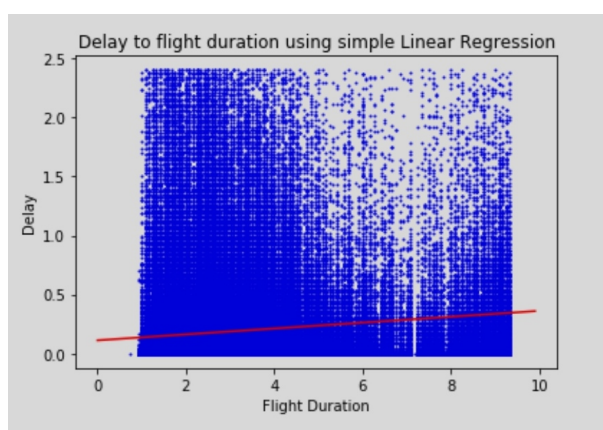


Figure 7: Linear regression.

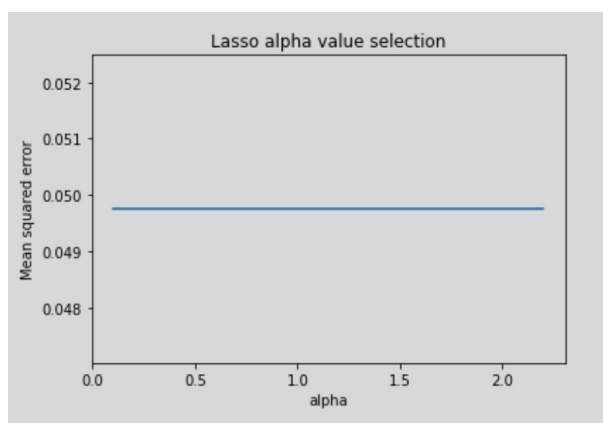


Figure 8: Lasso regression.

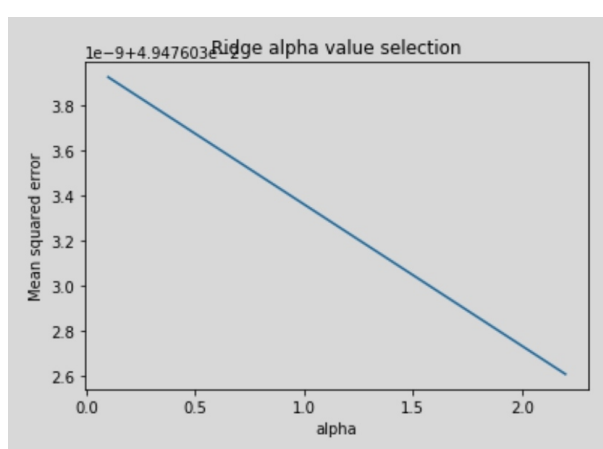


Figure 9: Ridge regression.

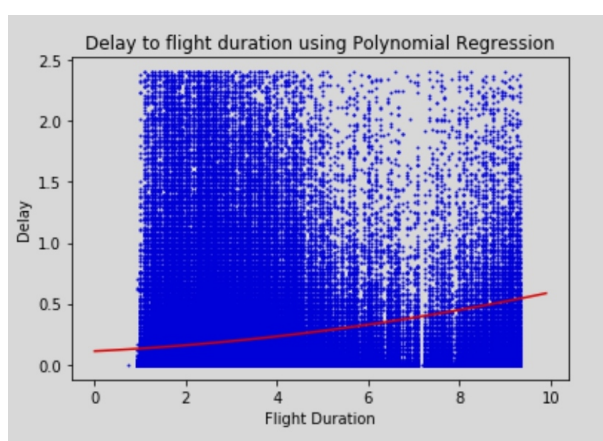


Figure 10: Polynomial regression.

We can see that most of the models does not overfit but underfit except Polynomial Regression with all feature somehow over fit, but at the end non of them have a positive R2 score, so I recommend to gather more data like the events and the plain type. For polynomial regression the degree was 2. I think the best model Is the NN because It has the best score overall it contain 3 layers first with 16 N second with 8 third with 1 and the activation is relu optimizer is adam.