Innopolis University

Machine Learning
Fall 2021
M21-DS-01
Ahror Jabborov

Assignment 1 - Report

24th September 2021

Github link: https://github.com/ahrorjabborov/ML_Assignment_1

MOTIVATION

In this report, first I read & analyze the given dataset of flight delays. Then preprocess it to make it suitable for computing. Preprocessing process includes adding new computed predictors from existing ones, deleting some of the columns, checking for nans, encoding into numerical values and splitting the dataset into train and test datasets. Then, I implement three machine learning models to predict the new delays. Afterall, I compare the efficiency of the models using some appropriate metrics. By reading this report, you can expect to gain knowledge on some practical data visualization, preprocessing and implementation of models.

TASK DEFINITION & DATA DESCRIPTION

The task is to implement three machine learning models to predict the flight delays using some predictors. The main idea behind the task is to prepare the given dataset, visualize, choose appropriate models, implement them and compare the results.

Dataset is given by the course instructor and it has the following structure:

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
...	...	...	...	...

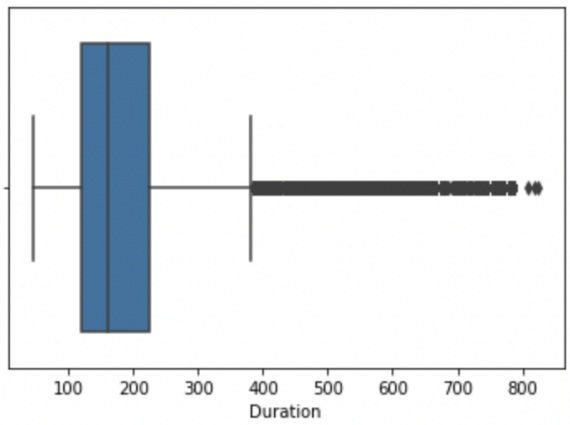
Delay is our dependent variable - the outcome.

The names of the airports (**Departure Airport** and **Destination Airport**) should be encoded into numerical values to make it possible to do computations on them.

Several new features can be obtained using **Scheduled departure time** and **Scheduled arrival time** such as departure/arrival time, day of the week, month, season and the duration of the flight.

Outlier detection and elimination

Before implementing models, I prepare the dataset by exploring the dataset, checking for nans, encoding labels, adding new computed features, getting rid of useless columns and getting rid of outliers in both outcome and predictor. Then I move to implementing models. In order for our model to be more precise, we need to eliminate outliers from the dataset.



Because I choose one feature - **flight_duration**, I drop all rows with a threshold less than 3. Using boxplot from seaborn library I first have a look at outliers. So here it seems like outliers are above 400.

After removing (using z-score and np.where) my dataset decreased by about 1.5 thousand records.

Then I split data to train and test comparing values of departure time. All the data samples collected in 2018 are to be used as a testing set.

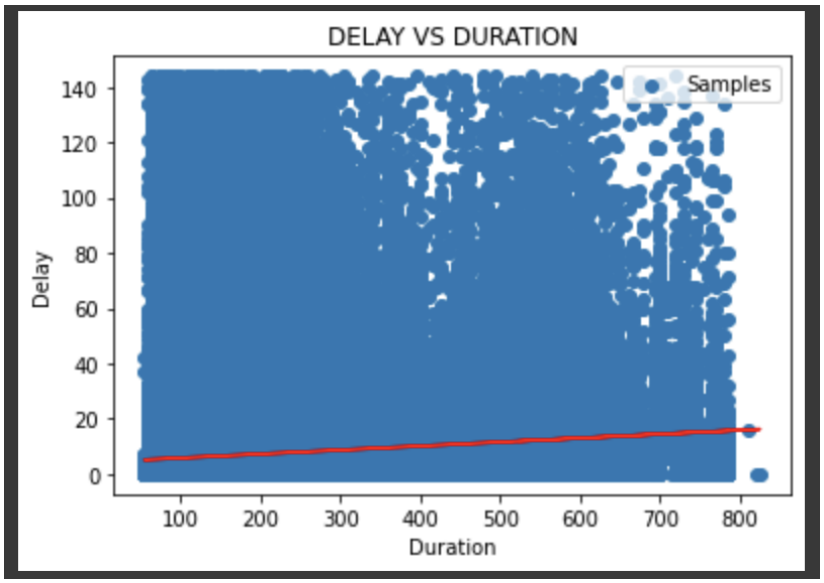
IMPLEMENTATION OF 3 MODELS

LINEAR REGRESSION

First I start with Linear Regression with one predictor:

Outcome: Delay

Predictor: Duration of the flight.



	Actual	Predicted
0	0.0	3.640779
1	2.0	3.640779
2	0.0	3.640779
3	0.0	3.640779
4	9.0	3.640779
...	...	...
7995	2.0	3.640779
7996	0.0	3.640779
7997	0.0	3.640779
7998	9.0	3.640779
7999	0.0	3.640779

R2-score: `-0.002788366634237427`

Model Coefficient = 0.016

Model intercept = 7.984

y = 7.984 + 0.016 * x

POLYNOMIAL REGRESSION

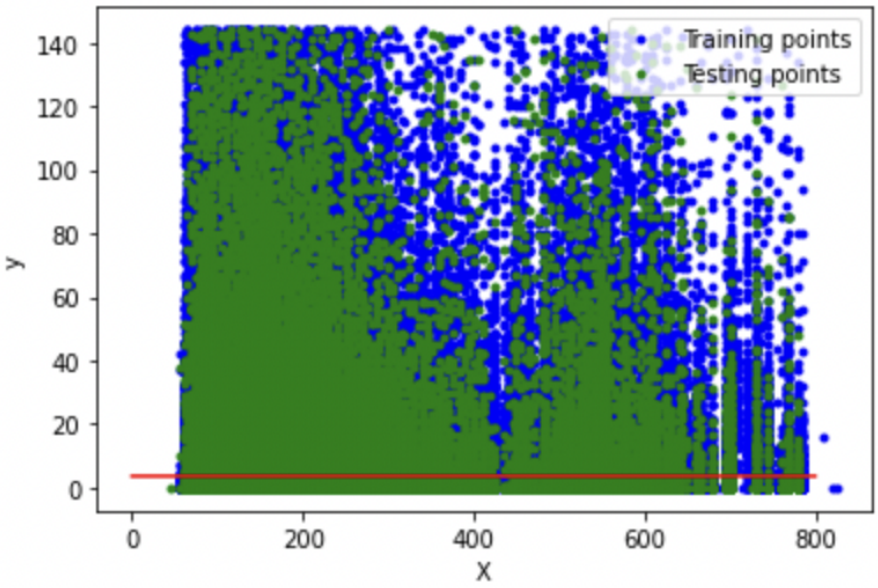
As for the result in the code, I left degree=4, because it gave the highest R2 score for .

```
The model performance for the training set
-------------------------------------------
RMSE of training set is 46.204105507567824
R2 score of training set is 0.001904771819981832


The model performance for the test set
-------------------------------------------
RMSE of test set is 39.6515358252226
R2 score of test set is -0.010448533929967807
```

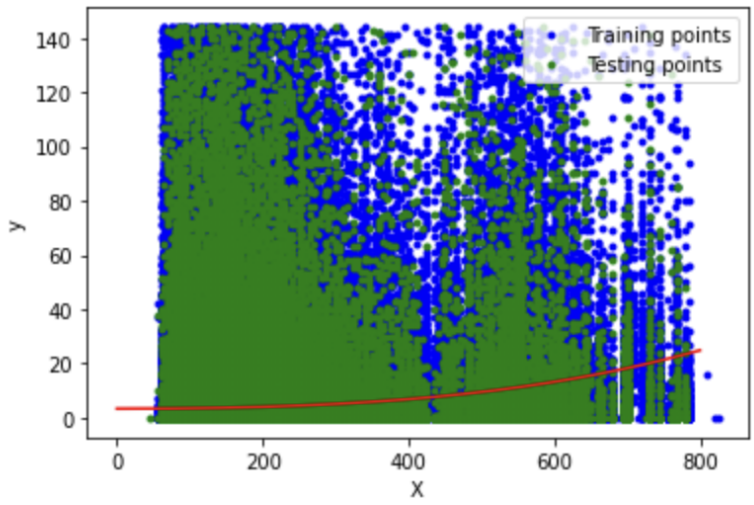
Here, blue points are test data and geen points are train data, whereas the red line is the polynomial model with above writer characteristics. The intercept is `3.64077902`

And the coefficients are: `6.22177947e-17, -3.66680580e-21, 3.84057137e-21, 8.43410594e-19, 1.45730861e-16, 1.54626631e-14, -8.78678124e-17, 1.93248469e-19, -1.92223038e-22, 7.21885752e-26`



As we can see R2-score is below 0, which means it is not recommended to use this model for this dataset.

LASSO with REGULARIZATION(L1)



R2: `-0.16224757138163315`

The intercept is `3.36060036`

And the coefficients are:

```
0.00000000e+00
-0.00000000e+00
1.28113350e-05
1.95631534e-08
8.10254589e-12
-3.83955864e-15
-1.50662944e-17
-2.44592979e-20
-3.16205395e-23 -3.66746457e-26 -4.00731054e-29
```

However, Lasso score is `0.014197654414266014.`

EVALUATION & RESULTS

R_2 score is below 0 in all models which means the models are not applicable using the above predictor. In my opinion, more appropriate predictors should be considered for predicting flight delays such as weather conditions or business of the airport.

REFERENCES

[1]: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[2]: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

[3]: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

[4]: https://scikit-learn.org/stable/modules/linear_model.html

[5]: https://www.kaggle.com/fabiendaniel/predicting-flight-delays-tutorial/notebook