**Aigerim Gilmanova**
**Machine Learning (F21)**
**Innopolis University, 2021**
**Assignment 1**
**Github: https://github.com/aigerimu/ML_A1**

## Introduction

The following report is directed to describe the Assignment 1 of flight delay estimation by using machine learning models. The flight delay estimation is based on the dataset which is obtained from Innopolis University partner company. The main parts of the report include Task and Data description, Data Preprocessing and Visualization, Outlier Detection and Removal, Machine Learning Models and Conclusion.

## Task and Data description

The assignment task is to measure the flight delay using machine learning methods. In particular, it is necessary to:

- Preprocess data, visualize it and split the dataset
- Apply 3 machine learning models to estimate the flight delay
- Compare these models

This task is a Regression task because flight delay is estimated in minutes.

| | 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 |
| **...** | ... | ... | ... | ... | ... |
| **675508** | SVO | 2018-08-31 23:50:00 | SVX | 2018-09-01 02:10:00 | 0.0 |
| **675509** | LED | 2018-08-31 23:50:00 | SVO | 2018-09-01 01:10:00 | 0.0 |
| **675510** | SVO | 2018-08-31 23:55:00 | EGO | 2018-09-01 01:20:00 | 0.0 |
| **675511** | SVO | 2018-08-31 23:55:00 | TSE | 2018-09-01 03:15:00 | 0.0 |
| **675512** | SVO | 2018-08-31 17:25:00 | IKT | 2018-08-31 23:05:00 | 379.0 |

675513 rows × 5 columns

Figure 1. Flight delay estimation dataset.

The figure above illustrates the given dataset for the assignment task. It can be seen that there are 4 predictors, which are Departure and Destination Airports, Scheduled departure time and Scheduled arrival time. Also, there is 1 target value, it is a Delay column. The whole dataset includes 675513 rows and 5 columns.

## Data Preprocessing and Visualization

The dataset contains categorical features (string representation), so it is necessary to transform them into integer format to get data from the ‘Departure Airport’ and ‘Destination Airport’ columns. It is done by using **LabelEncoder**. However, this encoder converts categorical features into numerical representation in a random way. Thus, the model might suggest them as an order (SVO > HAV), while these data is different and cannot be ordered. Also, there is another encoder that is **OneHotEncoder**. However, it was not used as this encoder increases the dimensionality. The figure below represents the dataset with encoded categorical features.

| | Depature Airport | Scheduled depature time | Destination Airport | Scheduled arrival time | Delay |
|---|---|---|---|---|---|
| **0** | 144 | 2015-10-27 07:40:00 | 56 | 2015-10-27 20:45:00 | 0.0 |
| **1** | 144 | 2015-10-27 09:50:00 | 68 | 2015-10-27 20:35:00 | 2.0 |
| **2** | 144 | 2015-10-27 10:45:00 | 94 | 2015-10-27 23:35:00 | 0.0 |
| **3** | 144 | 2015-10-27 12:30:00 | 82 | 2015-10-28 01:20:00 | 0.0 |
| **4** | 113 | 2015-10-27 14:15:00 | 144 | 2015-10-27 16:40:00 | 9.0 |

Figure 2. Dataset with encoded categorical features.

In order to estimate flight delay, it is necessary to understand its relation with other features. As there is no direct relation with ‘Departure Airport’, ‘Destination Airport’, ‘Scheduled departure time’, ‘Scheduled arrival time’, the new feature was added. This feature is ‘Flight duration’, that is the difference between ‘Scheduled arrival time’ and ‘Scheduled departure time’. However, according to the correlation matrix from the figure 3, that shows the correlations among values, there is no strong relations of the proposed features with flight delay. Thus, the following features are insignificant in flight delay estimation. The strongest correlation of flight delay is with flight duration.

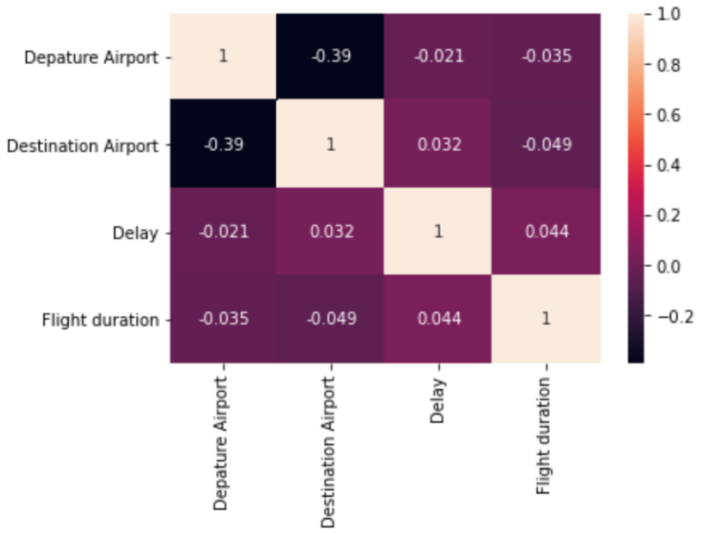


Figure 3. Correlation matrix

In order to visualize data, the graph ‘Flight duration vs Delay’ was plotted.

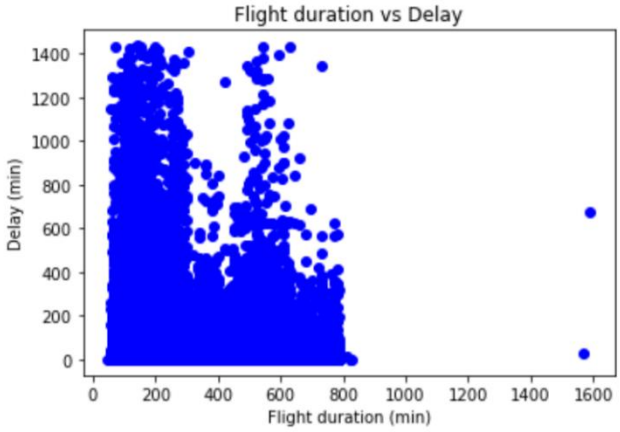


Figure 4. Flight duration vs Delay

After preprocessing, data was split into train and test set. The train set consists of data that was collected in 2015-2017 years, while test set includes data from 2018 year.

## Outlier Detection and Removal

Outliers are observation points which are significantly differ from the other observation points. They can be treated either as a mistake or as a new behavior. Thus, to detect outliers, different methods can be used. One of them is a Box plot that graphically illustrates sets of numerical data according to quartiles, and outliers are shown as individual points. According to figure 5, points that are larger than 900 min, are outliers.

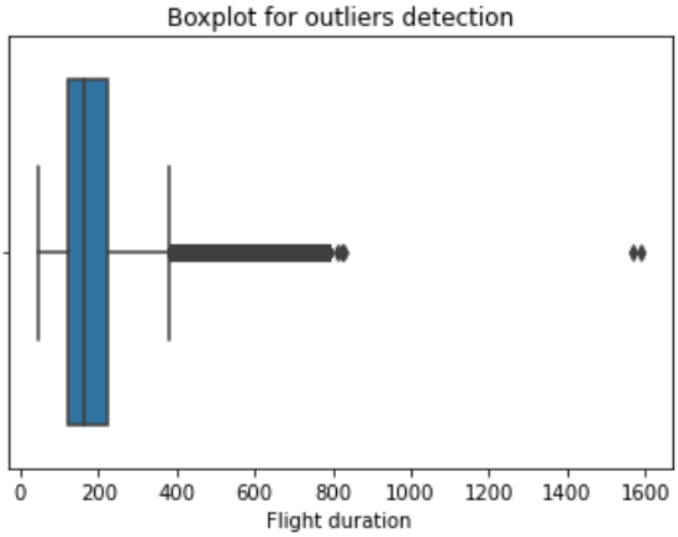


Figure 5. Outliers detection

After outliers observation, it is necessary to remove them from the train dataset. Then, the train dataset was split into predictors and target, *X_train and y_train*, respectively, without such features as ‘Scheduled departure time’ and ‘Scheduled arrival time’.

## Machine Learning Models

The regression models for this task were used. They are Linear Regression, Polynomial Regression and Linear Regression with Lasso regularization. The models were trained on the train and test datasets. Their performance can be seen from Tables 1 and 2. The performance of the following models can be evaluated through the metrics. The regression metrics were chosen, which are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE). These metrics represent errors in minutes. Also, R2 (coefficient of determination) score was used. This score shows the relation between predictors and target. R2 = 1 means the strongest relation (maximum), while R2 = 0 the weakest relation (minimum).

| **Model / Metrics** | Linear regression | Polynomial regression | Linear regression with Lasso regularization |
|---|---|---|---|
| MAE | 15.378918182116797 | 15.358906223983395 | 15.335765748398678 |
| MSE | 2148.3813857403484 | 2144.5207378790783 | 2130.7620652088576 |
| RMSE | 46.35063522477711 | 46.30897038241164 | 46.160178348971506 |
| R2 score | 0.003102281599158263 | 0.004893709820420145 | 0.003112191962348354 |

Table 1. Metrics for train data

| **Model / Metrics** | Linear regression | Polynomial regression | Linear regression with Lasso regularization |
|---|---|---|---|
| MAE | 14.355920147039628 | 14.405498917663062 | 14.339740132569611 |
| MSE | 1617.2888538653208 | 1620.4170470177985 | 1617.0759834246674 |
| RMSE | 40.21553000850941 | 40.2544040698381 | 40.21288330155732 |
| R2 score | -0.00905757032108534 | -0.011009310094960645 | -0.00892475636572665 |

Table 2. Metrics for test data

### Linear Regression

According to tables 1 and 2, linear regression predicts with large errors and its R2 score is 0. It is because the given data is not linear. Also, after applying this model to train and test sets, both gives huge errors, so it means that linear regression model **underfits**.

### Polynomial Regression

In polynomial regression, it is necessary to define the degree of the polynomial. In this case, it is 3 because it provides the best results and requires less computations comparing with higher orders. Polynomial regression provides better results comparing with the linear model, however, it is still low. Thus, the model **underfits.**

**Linear Regression with Lasso Regularization**

The Lasso regularization was chosen because the provided dataset has not many predictors. Lasso has a hyperparameter that was obtained from the graph below. It is 1.

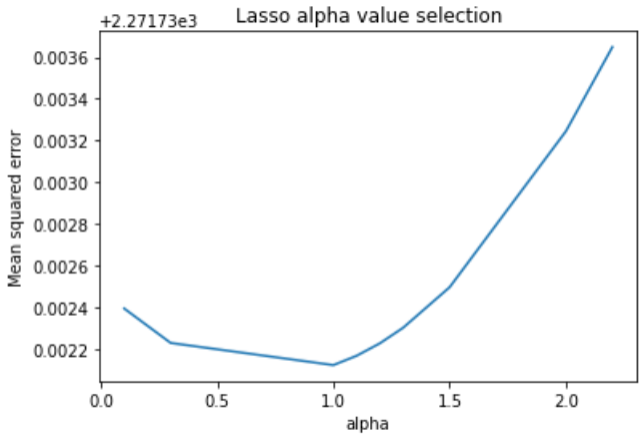


Figure 6. Detection of hyperparameter

Comparing with other models, linear regression with Lasso regularization has the best results. However, from the tables above it can be seen that results from train and test sets are almost the same, so the following model **underfits.**

## Conclusion

In order to make chosen models fitting, several steps were done. In particular, preprocessing, adding features and outliers removal. However, all 3 models are underfitting. One of the underfitting reason is choosing simple models for complex dataset. The second reason is selecting useful predictors. For instance, the weather more influences flight delay, comparing with the given features.