Machine Learning Assignment 1

Ali Shaheen MS21-RO

September 24 2021

1 Motivation

This report describes the solution of the first assignment of machine learning course at Innopolis University. It explains the approaches used in data preprocessing, representation, visualization. Moreover it discusses the details regarding the creation of various machine models and a comparative study of their results according to numerous metric scores. All of the source code and other details can be found here: `https://github.com/AliShaheen123/ML-Assignment-1`

2 Task Description

The task can be explained as estimating flight delay using machine learning.

2.1 Dataset

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 shown in the table below.

The data should were split to train and test. The data is split based on scheduled departure time. The train data is all the data from year 2015 till 2017. All the data samples collected in year 2018 were used as testing set. The delay is the target variable while all of the other variables are features.

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

3 Preproccessing and Visualization

3.1 Encoding Categorical Features

Since all the features columns represent non-numeral values, we need to find a way to convert them into numeral values.

- We decided to use one hot encoding with the **departure and destination airports**. Since there is no explicit relation or order between the different airports, i.e. the airports are non-ordinal, one hot encoding is the best strategy to use.

- For the **scheduled departure times**, we decided to do the following: the *year* part does not have that much of an effect, thus, it is disregard. Moreover, the month and day are represented in a single column called "day order" which ranges from 1 to 365. The time (hours and minutes) is represented again in a single column called (minute order) which ranges from 1 to 24 * 60 = 1440.

- Finally for handling the **scheduled arrival time**, it is enough to represent it as a single column (flight duration) which is simply the difference in minutes between the departure and arrival times.

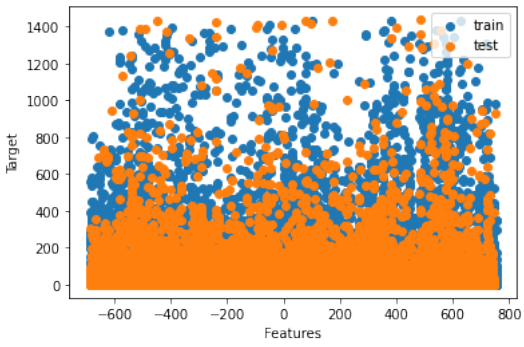
The resultant features' number are 338.

3.2 Visualization

Multiple ways were used to visualize the data.

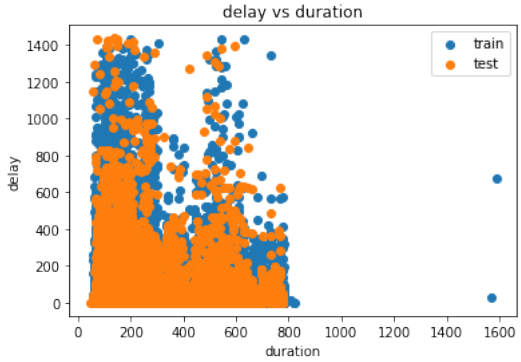
3.2.1 PCA

Here the features were centralized by subtracting the mean of all of the features. Then PCA with single component was used to reduce the dimensions from 338 to 1.



3.2.2 Delay vs Duration

This graph simply represents the delay with respect to a single feature–the duration of a flight which is the time difference between departure and arrival.

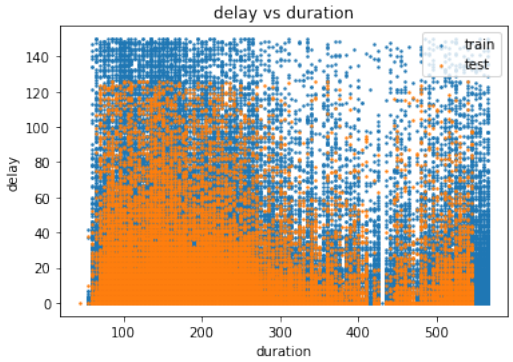


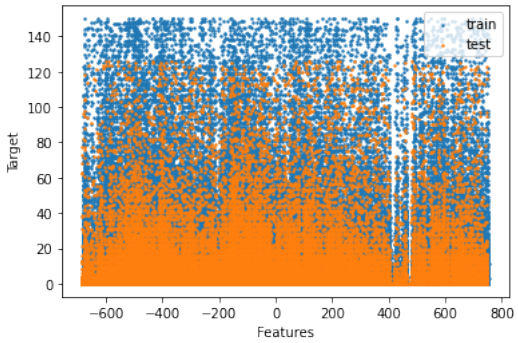
4 Outlier Detection and Removal

Considering the new features from the preprocessing, we can easily see that there cannot be any outliers in the features resultant from the one hot encoding since they are only zero or one. So we are left with three features. Both day of the year, and minute of the day cannot have outliers either because their ranges are known. Thus, the duration is the only feature that could have extreme values (outliers). Plotting the delay with respect to duration in the previous section has shown that there are obviously many extreme points in terms of the duration.

The Z-score was used to detect the outliers in duration. Then the rows which have duration with Z-score more than or equal to the threshold = 3 were eliminated from the training set. They represented only 2% of the original training set, thus they were removed entirely.

Below are the graph of Delay vs Duration and the graph resultant from PCA, respectively, after the removal of the outliers in training and testing sets. The outliers were less 3% of the over all dataset, thus it was reasonable to eliminate them entirely.





5 Machine Learning Models

Six models were introduced to predict the delay. Each one is described below with the corresponding metrics scores. The metric used is the mean square error (MSE).

5.1 Linear Regression Model

Here, we applied a regular Linear Regression model. The obtained results:

```
Testing set:
MSE: 155.10269114264923
Training set:
MSE: 285.06943410725785
```

5.2 Linear Regression Model on PCA

The features were reduced from 338 dimensions to only 3 dimensions. Then a linear model was used to predict the delay. The results were as the following:

```
Testing set:
MSE: 154.39314786223102
Training set:
MSE: 290.36325169731793
```

5.3 Linear Regression Model with Ridge Regularization

The results:

```
Testing set:
MSE: 154.2326504590595
Training set:
MSE: 289.5756141516386
```

5.4 Quadratic Regression Model

Here, we applied a regular Quadratic Regression model. The obtained results:

```
Testing set:
MSE: 150.7895840609102
Training set:
MSE: 284.31029725043
```

5.5 Quadratic Regression Model on PCA

The features were reduced from 338 dimensions to only 3 dimensions. Then a linear model was used to predict the delay. The results were as the following:

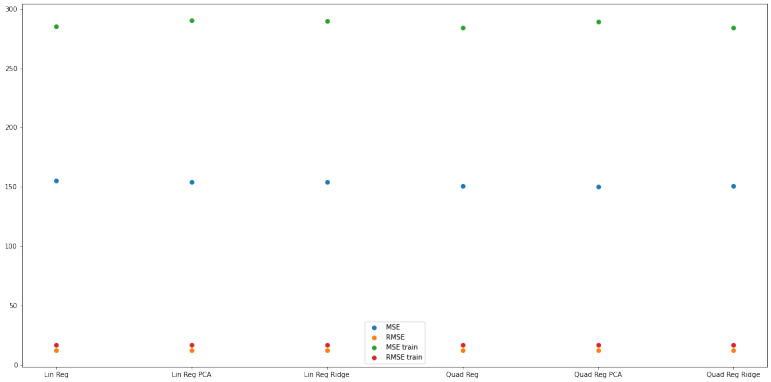
```
Testing set:
MSE: 149.97274342900184
Training set:
MSE: 289.3082915753967
```

5.6 Quadratic Regression Model with Ridge Regularization

The results:

```
Testing set:
MSE: 150.74196392469094
Training set:
MSE: 284.3317550840917
```

6 Comparison & Analysis



From the previous chart and the error values, we can conclude that quadratic models have achieved better performances than the linear models. Thus predicting flight delays problem is too complex for linear models to predict. Hence, for this problem it would be beneficial to utilize quadratic models.

Comparing the different types of quadratic models, quadratic regression with PCA is found to produce the best results according to the mean square error metric. However, the difference in the performances was insignificant between the models.

Generally, all models achieved satisfactory results having MSE of the range [149, 155].

7 Conclusion

This report demonstrated the solution of predicting flights delay problem. It explained the preprocessing and analyzed the datasets. It exhibited various manners to visualize the dataset and implemented outliers detection and removal. Moreover, this report has shown six different models to estimate flight delays relying on machine learning and provided a comparative study between them. Finally it presented the best model according to the metric considered.