Machine Learning Assignment 1

Guzel Safiullina

September 22, 2021

Contents

1 Introduction

In this assignment, we are going to be solving the task of flight delay estimation using machine learning. The project includes:

- Preprocess, visualize and split the dataset
- Outlier detection and removal
- Use 3 machine learning models to estimate the flight delays (Linear regression, polynomial regression, Linear Regression with Lasso regularisation)
- Compare the selected machine learning models performance using the appropriate evaluation metrics.

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 (fig. 1).

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

Figure 1: Dataset

The description of the 5 variables describing each flight are:

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 Time scheduled for the flight touch-down at the destination airport

Delay (in minutes) - flight delay in minutes

3 Data preprocessing

First of all, we have extracted new features from "Scheduled departure time" and "Scheduled arrival time". The new features are: month, day of the week, year,time, and the most important - Duration of the flight. After that, all features which are represented with strings have been encoded using LabelEncoder.

For data visualization on 2D plane dimension PCA reduction method was used. The one meaningful feature vs Delay was plotted (fig. 3). We have figured out that the most meaningful feature is flight Duration (fig 2).

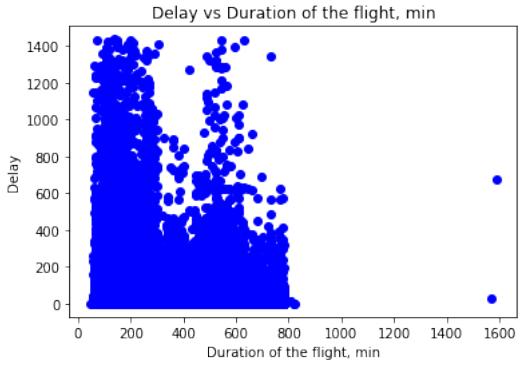

Figure 2: Flight duration vs Delay

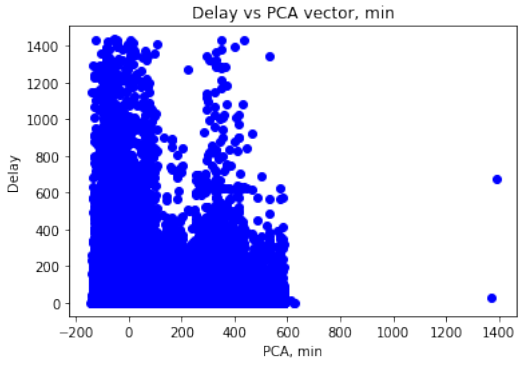

Figure 3: PCA vector vs Delay

The data was 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 are to be used as testing set.

4 Outlier Detection & Removal

During preparing datasets for machine learning models, it is really important to detect all the outliers and either get rid of them or analyze them to know why you had them there in the first place. In training machine learning models (especially supervised models), outliers can deceive the training process resulting in prolonged training times, or lead to the development of less precise models.

Outliers are not easily recognizable in the data collection stage however they can be detected in the analysis stage. Firstly, we have figured out that all targets are concentrated in a small interval from 0 minutes to approximately 100 minutes (fig 4). In this research Isolation Forest algorithm was used. After detecting and removal of outliers there are only targets with value less than 50 minutes (fig. 5). Only train data was proceeded.

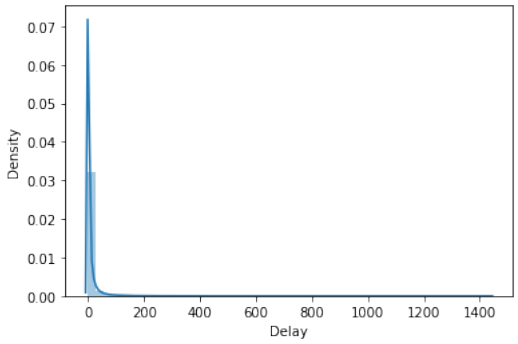

Figure 4: Delay histogram

5 Machine learning models

For estimating flight delay time, we have selected three appropriate machine learning algorithms:linear regression, polynomial regression, linear regression with Lasso regularization.

5.1 Results

Results of the models are in table 1.

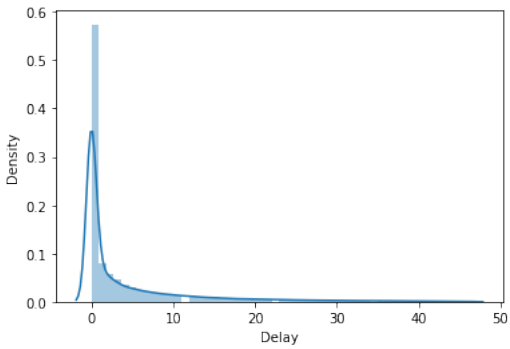

Figure 5: Delay histogram after removing outliers

	Linear regresion with Lasso	Linear regresion with Lasso	Polynomial regresion degree = 2	Polynomial regresion degree = 4	Polynomial regresion degree = 5
MAE	9.28	9.58	9.3	9.3	9.3
RMSE	40.02	40.02	40.04	40.03	40.03
R_2	0.0005	0.0006	-0.0001	2.7e-05	5.6e-05

Table 1: Errors in different models

The mean value of delay is ≈ 10 minutes, so the errors in each model are giant and $R^2 << 0.5$. There are two possible explanation: test data contains noisy points, which gives a big error and problems with data.

5.1.1 Novelty detection

The training data is not polluted by outliers and we are interested in detecting whether a new observation is an outlier. In this context an outlier is also called a novelty. In this case, We have used Isolation Forest algorithm to drop some noisy points from test set. Results are in table 2. R^2 is negative only when the chosen model does not follow the trend of the data, so fits worse than a horizontal line.

It is obvious that error in each model have decreased R^2 score has increased, but error is still too big in comparasion with mean delay and $R^2 < 0.5$. We can see, that noisy points have an impact on error, but there is problems with data in general. For example, this study did not take into account weather.

	Linear regresion with Lasso	Linear regresion with Lasso	Polynomial regresion degree = 2	Polynomial regresion degree = 4	Polynomial regresion degree = 5
MAE	5.3	5.02	5.05	5.05	5.05
RMSE	6.8	6.6	6.6	6.6	6.6
R_2	0.2	0.15	0.15	0.15	0.15

Table 2: Errors in different models after removing noisy points

6 Results

Results have shown that we have only one meaningful feature and this feature are not suitable to predict Delay of flight. And all models have small R^2 score and giant errors and they all are underfitted.