Machine Learning Assignemnt 1 - Report

Valeriy Novossyolov
Data Analysis and Artificial Intelligence
Innopolis University
Innopolis, Russian Federation
v.novossyolov@innopolis.university

***Abstract*—This paper tries to solve the problem of predicting the delay of airplanes using several machine learning approaches, which are related to Simple Linear Regression. The results show that the Third Degree Polynomial Regression model performs the best in comparison to others with MAE of 10.924 minutes and RMSE of 39.853 minutes. However, further investigation of the provided data and the models' performance show that this task requires a more complex model, with the number of predictor in the dataset the most obvious limiting factor.**

I. Task and Data Description

A. *Initial Setup*

The main task of this assignment is to train and test three different machine learning approaches to predict the possible delay of an airplane given its departure-arrival schedule. In this case, the main target was delay in minutes, which means that this is a regression problem. Along, side the delay the dataset of airlines schedules contained both information on scheduled departure and arrival date, time and airport.

Looking further into the data, the delay could range from 0 minutes to 1436 minutes (1 day). However, the majority of the data seems to show low delay, with 57.9% of the whole dataset having 0 minute delay and with more extreme values closer to 1436 minutes individually amounting to less than 0.1%. Looking at the predictors, there seems to be no class imbalance here, however, the departure and arrival airports seems to favor SVO, with 49.85% of all data having SVO as either departure or arrival airport.

B. *Predictor extraction and further data analysis*

After loading the dataset, I decided to extract more useful information from the given date-time pairs. From them, I extracted the following features: flight day, flight month, day of the week, time of departure and time of arrival (both converted into minutes), and finally flight duration in minutes. In addition, I extracted the year of the flight to later on split the dataset into train and test sets. In other words, the 4 original predictors were turned into 8, with year of flight being the data splitting parameter. It is worth to mention the distinction between time of departure and arrival and flight duration. The purpose of the departure and arrival times is to represent the specific part of the day because of which a given airport might be congested resulting higher delay. Additionally, the difference between arrival time and departure time will not always give the exact value of the flight duration, thus, the flight duration is not the linear combination of these two predictors.

After analysis of the final extracted data, I constructed the graphs of each predictor versus the delay.

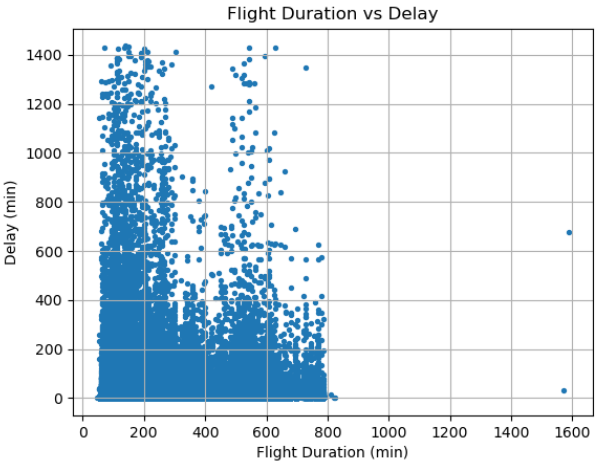

Fig. 1. Relation between Flight duration and Delay

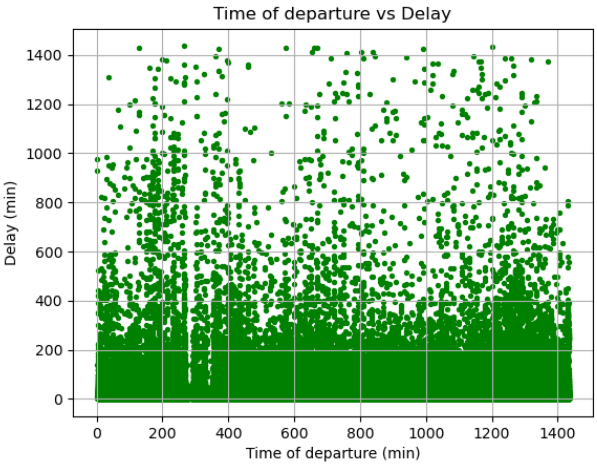

Fig. 2. Relation between Time of departure and Delay

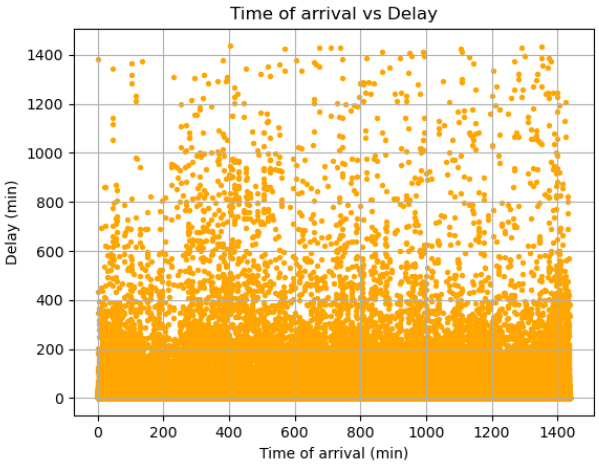

Fig. 3. Relation between Time of arrival and Delay

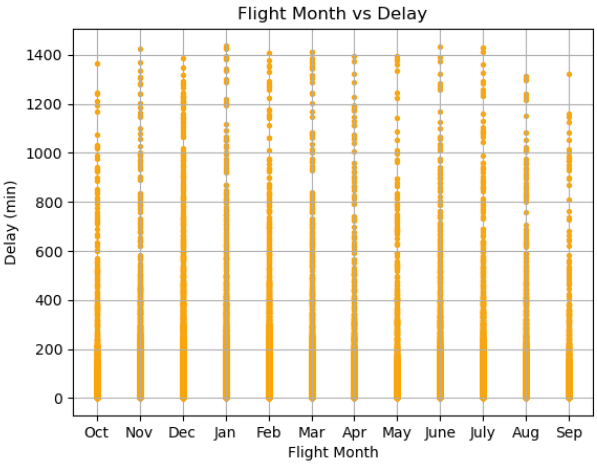

Fig. 6. Relation between Flight month and Delay

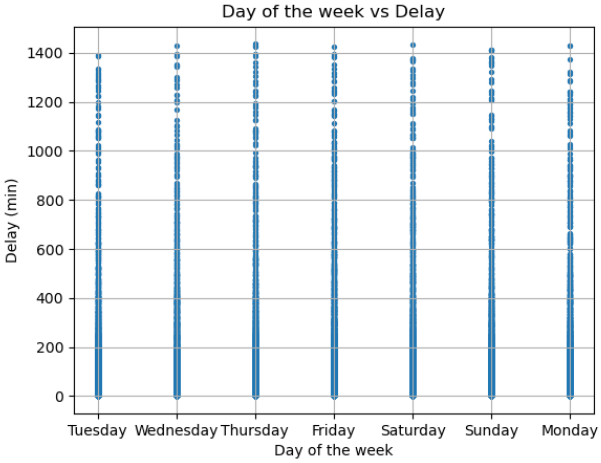

Fig. 4. Relation between Day of the week and Delay

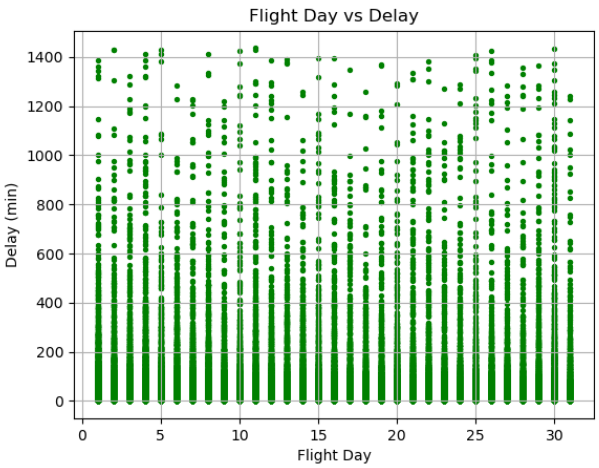

Fig. 5. Relation between Flight day and Delay

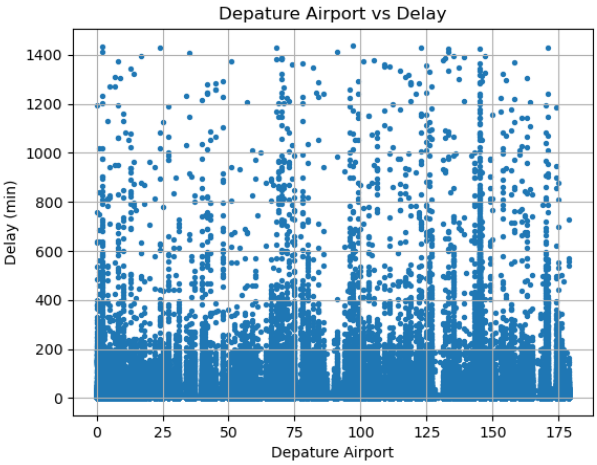

Fig. 7. Relation between Departure airport and Delay. The airport names were encoded into unique integers.

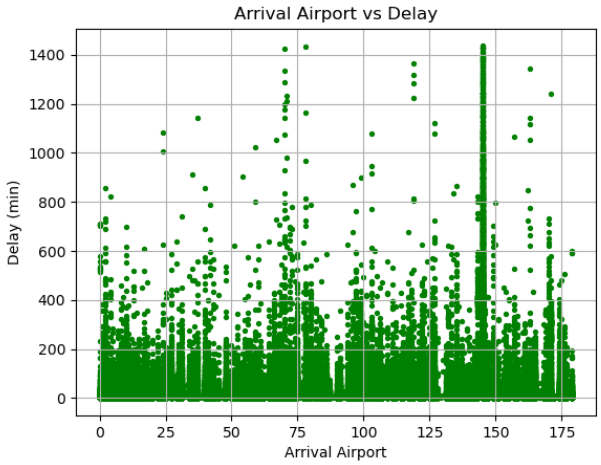

Fig. 8. Relation between Arrival airport and Delay. The airport names were encoded into unique integers.

As can be seen from the graphs, there is no clear linear dependence between each individual predictor and the target value. Furthermore, each predictor has wide range of delay for each of its values. Apart from the few points in the Fig. 1, it is hard to visually identify the possible outliers within the data. For example, it visually seems like that there are outliers present in the Fig. 8, but the value representing the most occurring airport SVO has all possible values of the delay. Thus, I cannot conclusively define any other outliers from visual inspection. The Fig. 9 show the only affected by the initial outlier reduction. After this outlier removal the dataset lost approximately 0.074% of its original data, thus, I can safely assume that these outliers will not significantly affect the performance of the machine learning models.

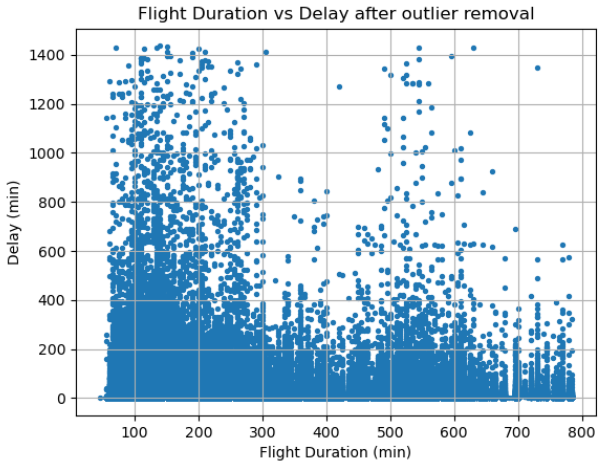

Fig. 9. Flight duration relationship graph after initial outlier removal

C. *Train and test datasets*

As was mentioned previously, division into train and test sets was done according to the year of flight. All flights that happened in between 2015 and 2017 went into train dataset, and the flights that happened in 2018 went into test dataset. After the datasets were separated, the parameter of year was completely removed from the datasets.

II. Machine Learning Models

A. *Models and their performance*

For this task I chose the 4 following machine learning approaches: Simple Linear Regression, Polynomial Linear Regression, Lasso Regression, and Ridge Regression. For each of the models, I trained them on train set, calculated their Mean Absolute (MAE) and Mean Square (MSE) errors on the test dataset, and calculated their cross-validated Mean Square error on the train set. Initially, all underfitted showing high error both on train and test datasets. For instance, the best performing model was Polynomial Linear Regression model with 2nd degree polynomial transformation. It showed MAE of 13.613 minutes and RMSE of 40.031 minutes on test dataset and RMSE of 46.368 minutes on train dataset.

In order to improve the performance of the models, I decided to explore the effect of removing outliers from the train dataset. I used z-score on Delay dataset with varying threshold levels and applied the reduced dataset to Second Degree Polynomial Regression model.

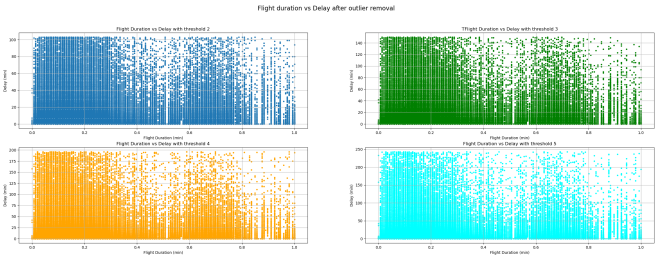

Fig. 10. Graphs showing the effect of outlier removal from Delay using z-score

Lowering the threshold of the z-score seemed to improve the error of the model, while also lowering the top most delay value. The best performance was shown by the threshold value of 2, with MAE of 10.256 minutes and RMSE of 39.891 minutes. However, looking at the previous threshold value of 3, the MAE stayed at 10.951 minutes and RMSE stayed at 39.878 minutes. In other words, going past threshold value of 3, the MAE continued to drop, but the value of RMSE started to grow. This is directly related to the amount of data left after the ”outlier” removal from Delay. Since the most occurring values of Delay were close to zero, lead to reducing the largest value of Delay in the train dataset and consequently the model started predicting lower Delay values more often. For that reason I decided to stop at the threshold value of 3, so that the models still were able to predict higher delay values.

B. *Final results*

After finishing working with train dataset, performance of all models significantly improved.

	MAE	MSE	RMSE	Train MSE
Simple Linear regression	11.268	1597.686	39.971	295.541
Polynomial degree: 2	10.951	1590.268	39.878	295.073
Polynomial degree: 3	10.924	1588.315	39.854	295.725
Polynomial degree: 4	11.221	1611.017	40.137	5.283939952345166e+17
Lasso Regression	11.27	1597.688	39.971	295.039
Ridge Regression	11.268	1597.686	39.971	295.04

Fig. 11. Final results of all models

As at could be clearly seen from the Fig. 11, the train error dropped to MSE of 295 minutes^2 and the new best performing model became Third Degree Polynomial Regression model. On the contrary, the Lasso and Ridge regressions didn't seem to improve in comparison to the Simple Linear Regression model. To explore that, I plotted the regularization parameter, alpha, against the model's error.

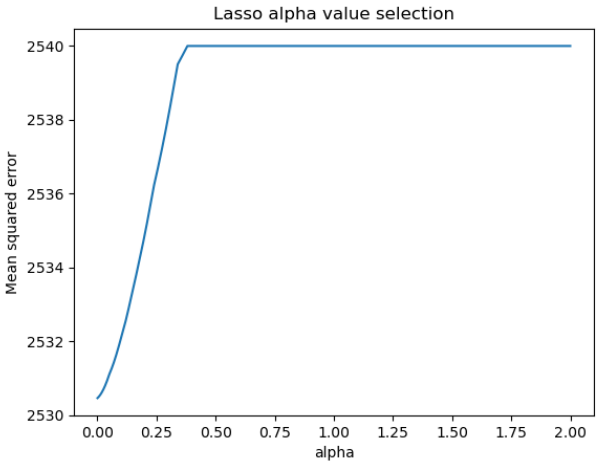

Fig. 12. Relation between alpha and Lasso's error

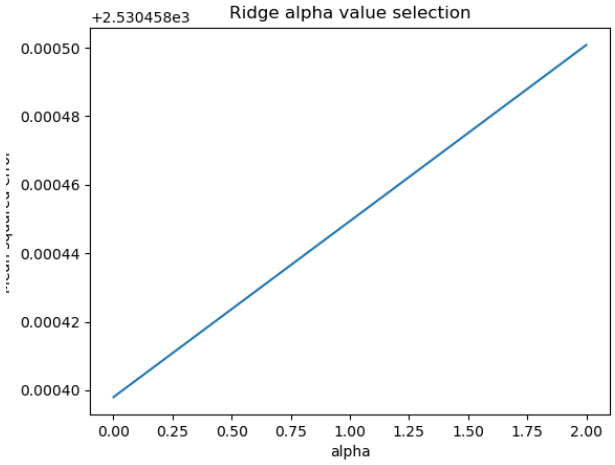

Fig. 13. Relation between alpha and Ridge's error

The graphs for both Lasso and Ridge showed the same trend, lower values of alpha lowered the final error value. However, that means that the best value of alpha is zero, which the same as the Simple Linear Regression without regularization. This the task still required more complex model to improve the result, but higher degrees of polynomial transformation seemed to decrease the performance of the model. These observations might be the result of absence of different predictors in the dataset.

III. Conclusion

By looking at the Fig. 11, the best performing model for this task is the Third Polynomial Linear Regression model with MAE of 10.924 minutes and RMSE of 39.853 minutes. In other words, the model seems to perform good on the majority of the data points, but it has a higher error for bigger delays. The potential source of this error is that the original dataset did not have enough data in order to accomplish this task. For instance, it is logical to assume that the delay would related to the weather conditions along flight path. Furthermore, the model is imbalanced for certain classes, for example, the majority of the Delay lies close to the zero. In order to improve the current results, the complexity of the model needs to be increased by increasing the number of predictors.