I. Abstract

In this assignment, we had to preprocess and apply recently studied machine learning models to the dataset. This is a task to test the student's knowledge of the material they have passed in the course [F21] Machine Learning at Innopolis University.

II. Introduction

This report presents a comparison of different models. All models were measured by several metrics. We have been given the dataset 'flight_delay.csv'. Our task is to preprocess, visualize and split the dataset. After that, we must apply the studied machine learning models and compare them. We need to compare using certain metrics and their results on models.

III. Metrics

1) Mean Squared Error (MSE) - Mean squared error regression loss.
2) Mean Absolute Error (MAE) - Mean absolute error regression loss.
3) Root Mean Squared Error (RMSE) - Under the root MSE.

Most common metrics for the regression-based model, which I will use further.

IV. Dataset description

The data set used is a sample of airports along with the departure, arrival and delay times. The data set consists of 675513 rows of data presented in 5 columns.
dataset feature vectors:

1. Departure Airport - the airport from which the plane took off (type: string)
2. Scheduled departure time - the exact time of departure of the aircraft (type: string)
3. Destination Airport - the airport where the plane should arrive (type: string)
4. Scheduled arrival time - exact arrival time (type: string)
5. Delay - the time of the plane's delay in minutes (type: integer)

V. Comparison of models

I used four models to train the model on training data. These are Linear Regression, Polynomial Regression, Lasso and Ridge.

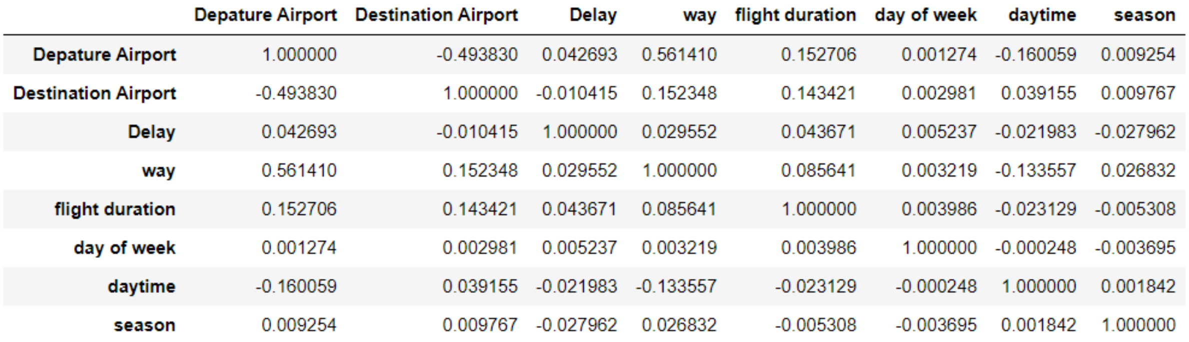
1. LinearRegression - Ordinary least squares Linear Regression.
2. Polynomial Regression (features) - Generate polynomial and interaction features.
3. Lasso - Linear Model trained with L1 prior as regularizer.
4. Ridge - Linear least squares with l2 regularization.

In general, all the models I selected showed a weak result (Fig. 5). A possible reason for this is the spread of data. And referring to the correlation(Fig. 1), we can confidently say that the data are absolutely independent of each other. The same can be seen in the visualization(Fig. 2) between the columns "Delay" and "Flight duration", which are the most correlated features. To train the models, I selected all the data from the dataset except the target column, which is "Delay". The reason for taking all the data was the overall low correlation with all the columns, despite the fact that the task description suggested choosing "Flight duration". During the training of the models, all the models behaved strangely. They showed relatively good results on the training data (Fig. 4) and very low results on the test data. Here we can assume

1. overfitting
2. non-uniform distribution of zeros in "Delay".

Referring to the picture, we can understand that more than 50% of the training data are just zeros, while in the test data the number of zeros is only about 28% (Fig. 3). Most likely, this could be the reason for the high indicators in the test data. If we consider each model, we can say that a simple Linear Regression, Lasso and Ridge are similar. The most interesting is the Polynomial Regression (linear regression after fitting polynomial features on the dataset). Because the Polynomial Regression with a second degree has the best indicator at the moment. This is because the (Pearson) correlation is very low, and the linear regression I used has the same coefficient (the standardized regression coefficient is the same as the Pearson correlation coefficient). According to my assumption, the polynomial regression shows good values precisely because it can correspond to such a distribution of training data. At the moment, the polynomial regression with a second degree is more suitable, because if you look at the data, then a parabolic curve is more suitable than a line. (Fig. 2)

VI. Tables



	Depature Airport	Destination Airport	Delay	way	flight duration	day of week	daytime	season
Depature Airport	1.000000	-0.493830	0.042693	0.561410	0.152706	0.001274	-0.160059	0.009254
Destination Airport	-0.493830	1.000000	-0.010415	0.152348	0.143421	0.002981	0.039155	0.009767
Delay	0.042693	-0.010415	1.000000	0.029552	0.043671	0.005237	-0.021983	-0.027962
way	0.561410	0.152348	0.029552	1.000000	0.085641	0.003219	-0.133557	0.026832
flight duration	0.152706	0.143421	0.043671	0.085641	1.000000	0.003986	-0.023129	-0.005308
day of week	0.001274	0.002981	0.005237	0.003219	0.003986	1.000000	-0.000248	-0.003695
daytime	-0.160059	0.039155	-0.021983	-0.133557	-0.023129	-0.000248	1.000000	0.001842
season	0.009254	0.009767	-0.027962	0.026832	-0.005308	-0.003695	0.001842	1.000000

Fig. 1 Correlation of dataset

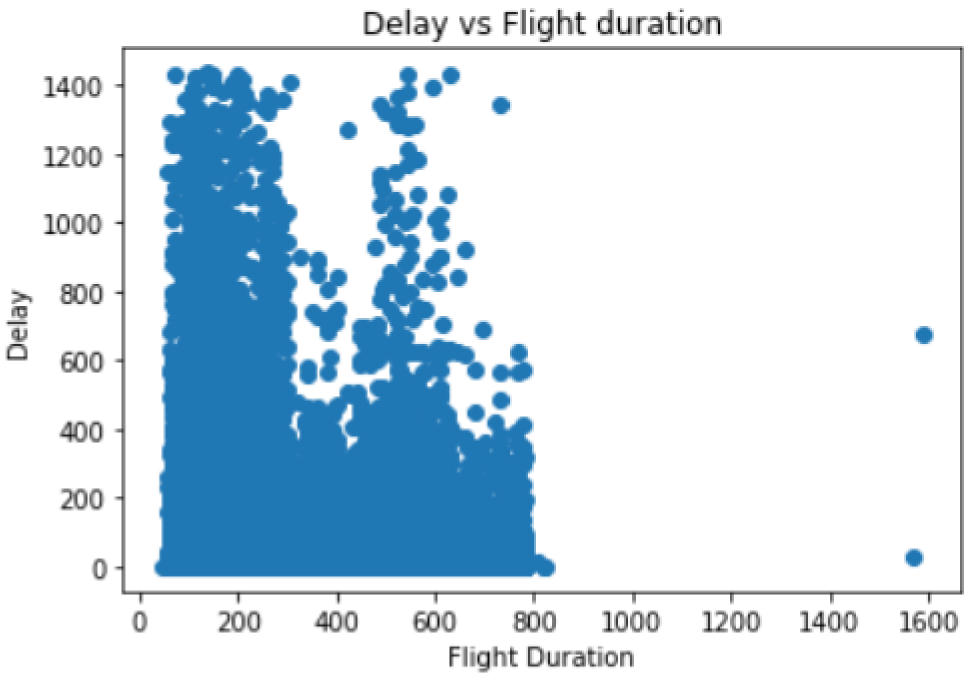
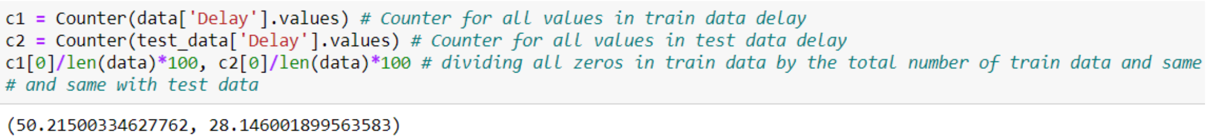
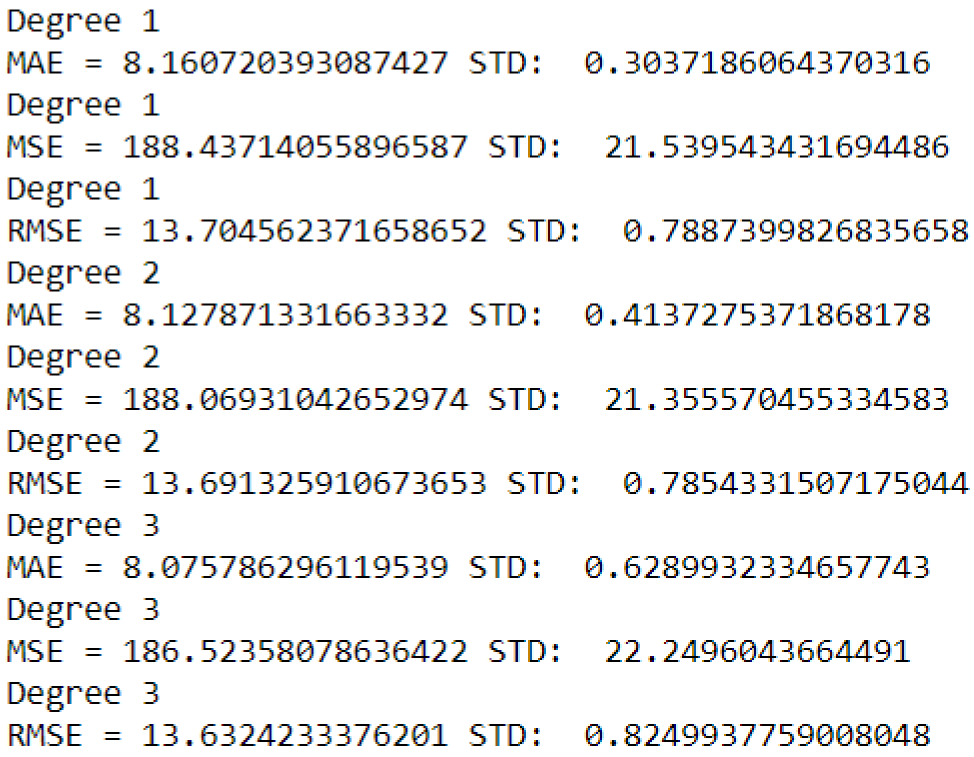


Fig. 2 Delay vs. Flight duration



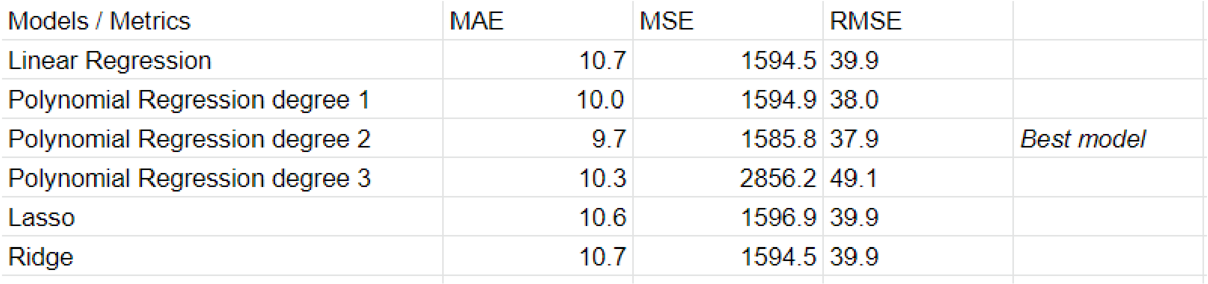
```
c1 = Counter(data['Delay'].values) # Counter for all values in train data delay
c2 = Counter(test_data['Delay'].values) # Counter for all values in test data delay
c1[0]/len(data)*100, c2[0]/len(data)*100 # dividing all zeros in train data by the total number of train data and same
# and same with test data

(50.21500334627762, 28.146001899563583)
```

Fig. 3 Number of zeros in train and test datasets (in percent)



```
Degree 1
MAE = 8.160720393087427 STD:  0.3037186064370316
Degree 1
MSE = 188.43714055896587 STD:  21.539543431694486
Degree 1
RMSE = 13.704562371658652 STD:  0.7887399826835658
Degree 2
MAE = 8.127871331663332 STD:  0.4137275371868178
Degree 2
MSE = 188.06931042652974 STD:  21.355570455334583
Degree 2
RMSE = 13.691325910673653 STD:  0.7854331507175044
Degree 3
MAE = 8.075786296119539 STD:  0.6289932334657743
Degree 3
MSE = 186.52358078636422 STD:  22.2496043664491
Degree 3
RMSE = 13.6324233376201 STD:  0.8249937759008048
```

Fig 4 Polynomial Regression results on train data



Models / Metrics	MAE	MSE	RMSE	
Linear Regression	10.7	1594.5	39.9	
Polynomial Regression degree 1	10.0	1594.9	38.0	
Polynomial Regression degree 2	9.7	1585.8	37.9	*Best model*
Polynomial Regression degree 3	10.3	2856.2	49.1	
Lasso	10.6	1596.9	39.9	
Ridge	10.7	1594.5	39.9	

Fig. 5 All models results

VII. References

1) https://stackoverflow.com/a/50392824 - Convert DateTime to minutes
2) https://stackoverflow.com/a/9847359 - How to get the day of the week by a given date
3) https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/ - Z score for outlier detection
4) https://www.kaggle.com/onotole/eda-flight-delays - idea to create new feature data['way']
5) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html- information about Linear Regression
6) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html#sklearn.linear_model.Ridge - information about Ridge
7) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html#sklearn.linear_model.Lasso- information about Lasso
8) https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html - information about Polynomial Features
9) https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/ - correlation vs. linear regression
10) https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics - Metrics

Note: All descriptions about preprocessing and code part you can see in .py file in the comments.