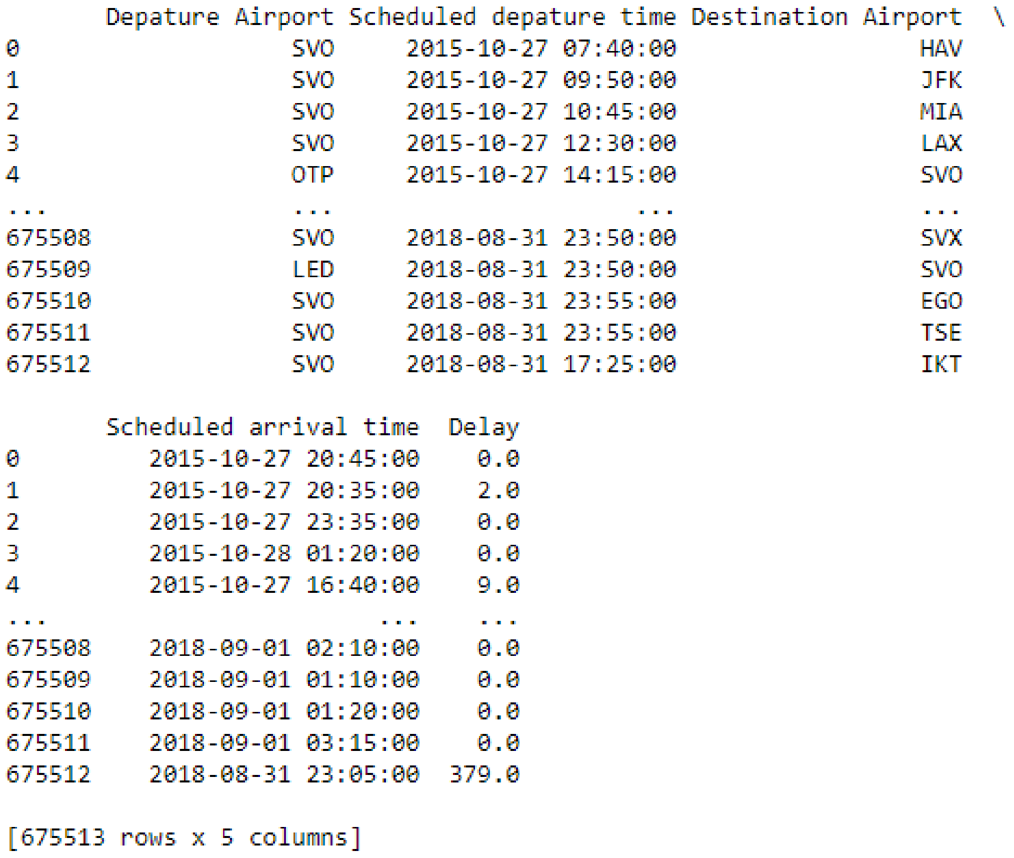
Report

The current report contains the machine learning task solution dedicated to prediction of flight delays by the data collected by an airline. The solution consists of the following steps: data preprocessing, outliers detection and removal, machine learning models creation and analysis of their representation. Also the report might contain explanation of reasons of bad output results.

Task definition and data description

The task is to create the machine learning model which is capable to predict the flight's amount of delay via the data provided by an airline.

The dataset consists of about 700 thousands rows of data and 5 named columns: Depature Airport, Scheduled depature time, Destination Airport, Scheduled arrival time and Delay. The last column contains labels, while the others contain features. A piece of data is represented on image 1 below.



```
       Depature Airport Scheduled depature time Destination Airport  \
0                   SVO     2015-10-27 07:40:00                 HAV
1                   SVO     2015-10-27 09:50:00                 JFK
2                   SVO     2015-10-27 10:45:00                 MIA
3                   SVO     2015-10-27 12:30:00                 LAX
4                   OTP     2015-10-27 14:15:00                 SVO
...                 ...                     ...                 ...
675508              SVO     2018-08-31 23:50:00                 SVX
675509              LED     2018-08-31 23:50:00                 SVO
675510              SVO     2018-08-31 23:55:00                 EGO
675511              SVO     2018-08-31 23:55:00                 TSE
675512              SVO     2018-08-31 17:25:00                 IKT

       Scheduled arrival time  Delay
0         2015-10-27 20:45:00    0.0
1         2015-10-27 20:35:00    2.0
2         2015-10-27 23:35:00    0.0
3         2015-10-28 01:20:00    0.0
4         2015-10-27 16:40:00    9.0
...                       ...    ...
675508    2018-09-01 02:10:00    0.0
675509    2018-09-01 01:10:00    0.0
675510    2018-09-01 01:20:00    0.0
675511    2018-09-01 03:15:00    0.0
675512    2018-08-31 23:05:00  379.0

[675513 rows x 5 columns]
```

Image 1 – Initial dataset

Data preprocessing

The initial dataset basic features were remade into another set of features. These new features contain ‘Month’ – month of departure (may reflect a bit weather conditions), ‘Path’ – combination of codes of departure and arrival airports, ‘Duration’ – difference between arrival and departure times in minutes.

Afterwards, the new dataset was splitted into train and test datasets by the ‘Year’ label which was initially embedded into the new dataset. The data about the 2018 year was defined as test.

Then the training features dataset was concatenated row-wise with the training label dataset in terms to identify the unique rows of data to prevent an issue of data imbalance. As a result, only 26% of training data were identified as unique. Then the features and the labels were divided again.

Outliers detection and removal

In terms to define the outliers it was decided to find outliers for every feature. Therefore, the correlations between each feature and delay were plotted. The graphs are shown on images 2 – 4.

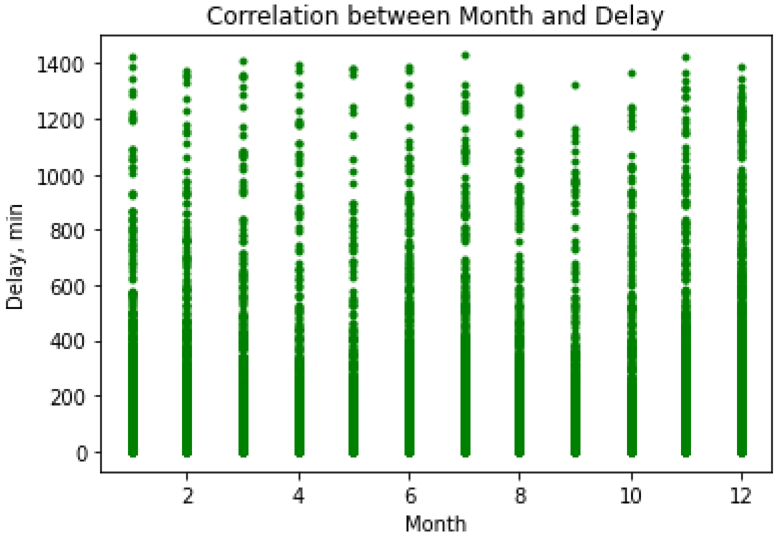


Image 2 – Correlation for ‘Month’ feature

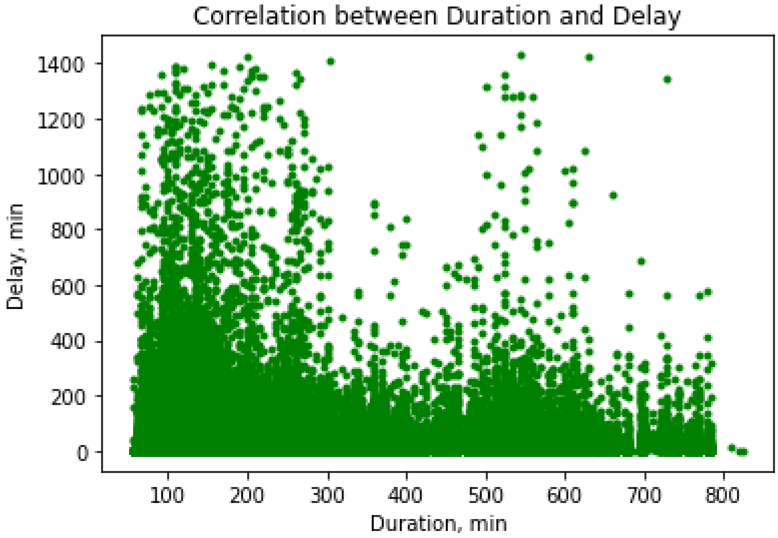


Image 3 – Correlation for ‘Duration’ feature

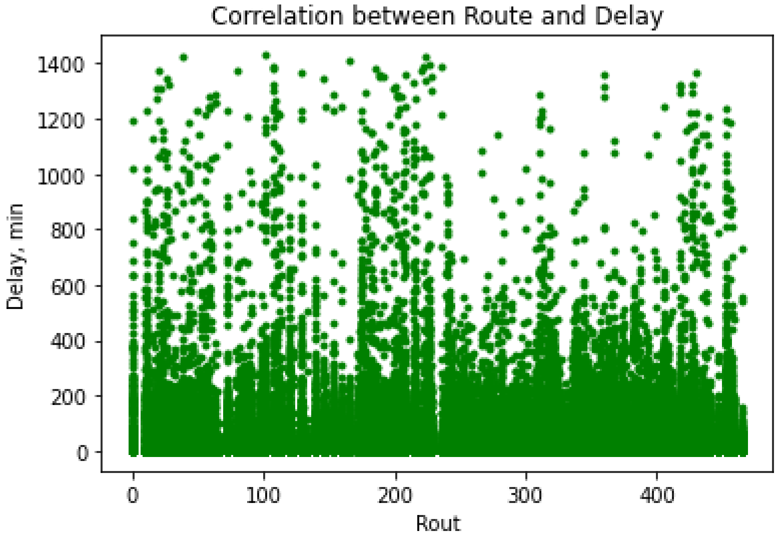


Image 4 – Correlation for ‘Path’ (‘Rout’) feature

Then each current value of feature was analyzed: every month, every duration, every rout according to possible values of delay. If the current delay value’s fraction in comparison to the overall number of delays for the current’s feature value (for example value of delay 2 mins for March has a fraction 16% of overall delays in March) is bigger, than the threshold value, then this delay value isn’t an outlier (16% > 5%-threshold → 2 mins value isn’t an outlier). Examples are shown on images 5 – 7.

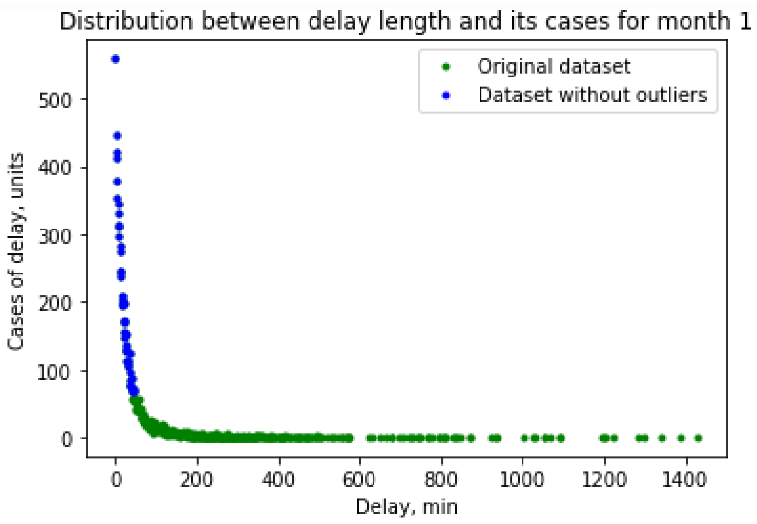


Image 5 – Outliers detection for January

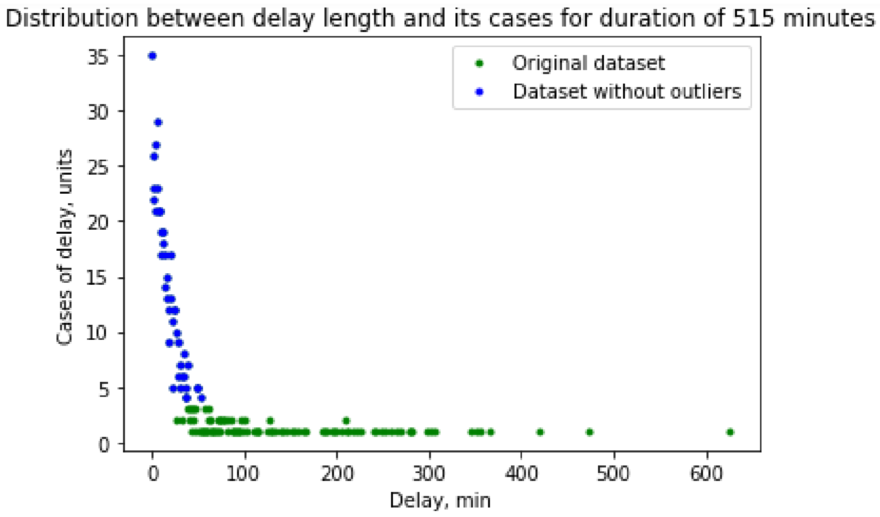


Image 6 – Outliers detection for duration of flight 515 minutes

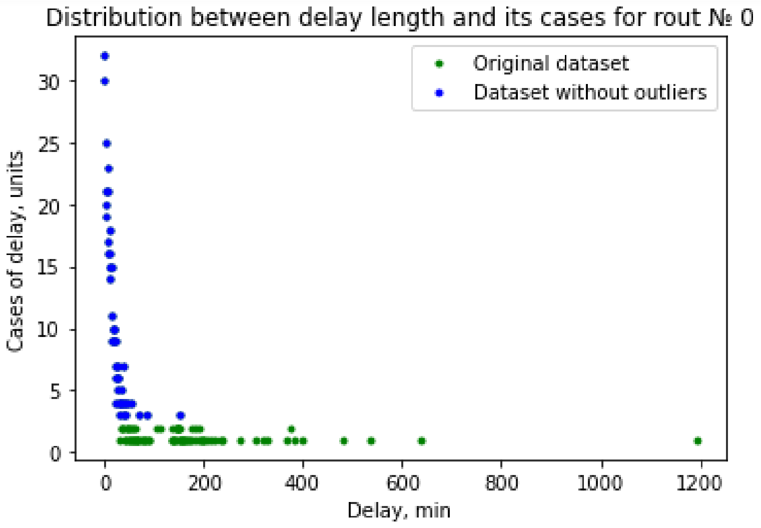


Image 7 – Outliers detection for rout № 0 (code of the path)

As a result, for every feature a list of non-outlier data was synthesized. The non-outlier data plotting with respect to initial data is shown on images 8 – 10.

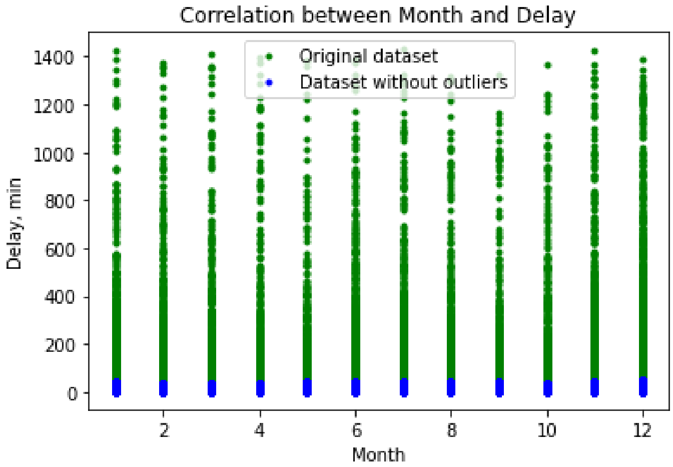


Image 8 – Non-outlier data of feature ‘Month’

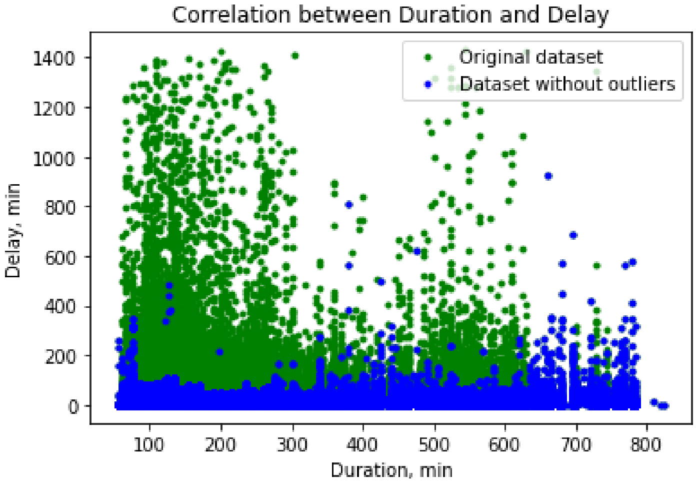


Image 9 – Non-outlier data of feature ‘Duration’

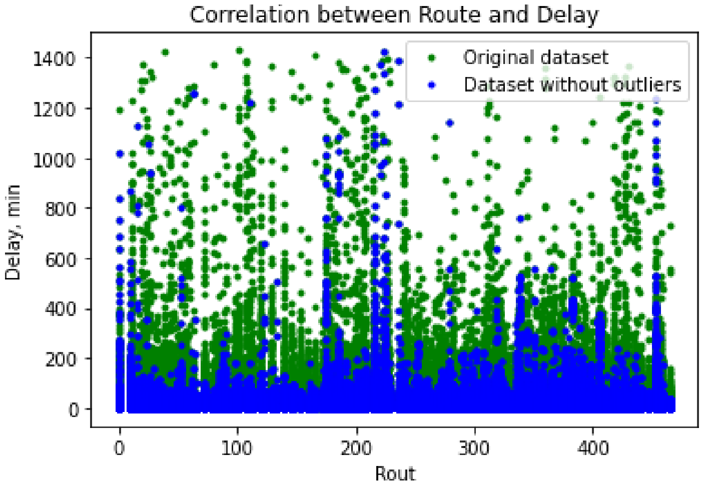


Image 10 – Non-outlier data of feature ‘Path’

To finish the process of outliers removal the intersection of non-outlier lists of every feature was found. The data defined as 'non-outlier data' took 77% of the 'unique data dataset'. It was named as generalized dataset without outliers. The results are depicted on images 11 – 13.

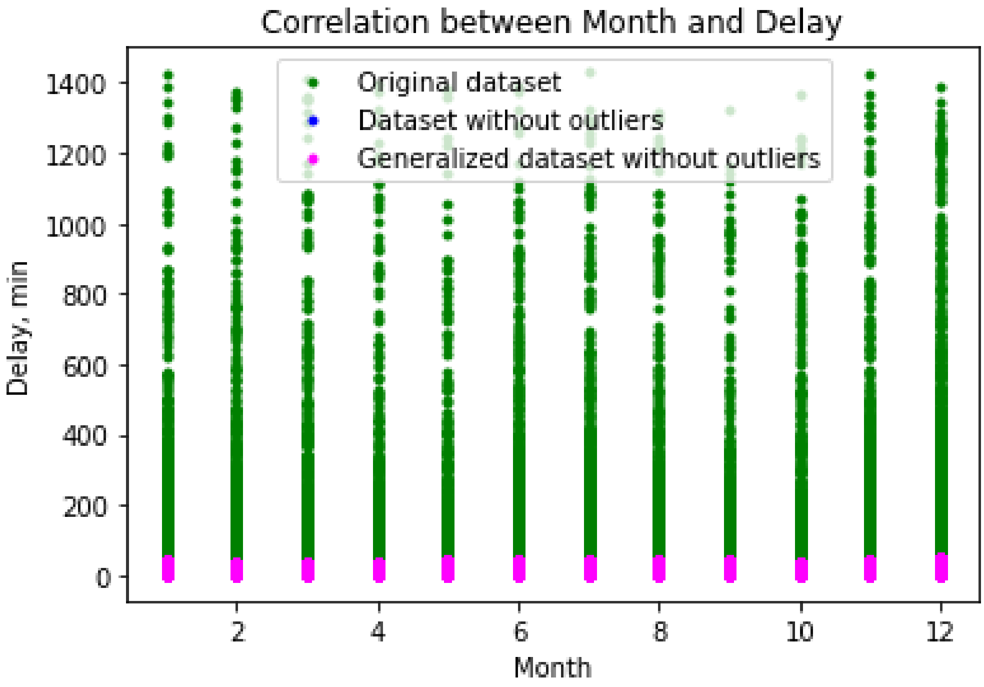


Image 11 – Generalized non-outlier data of feature 'Month'

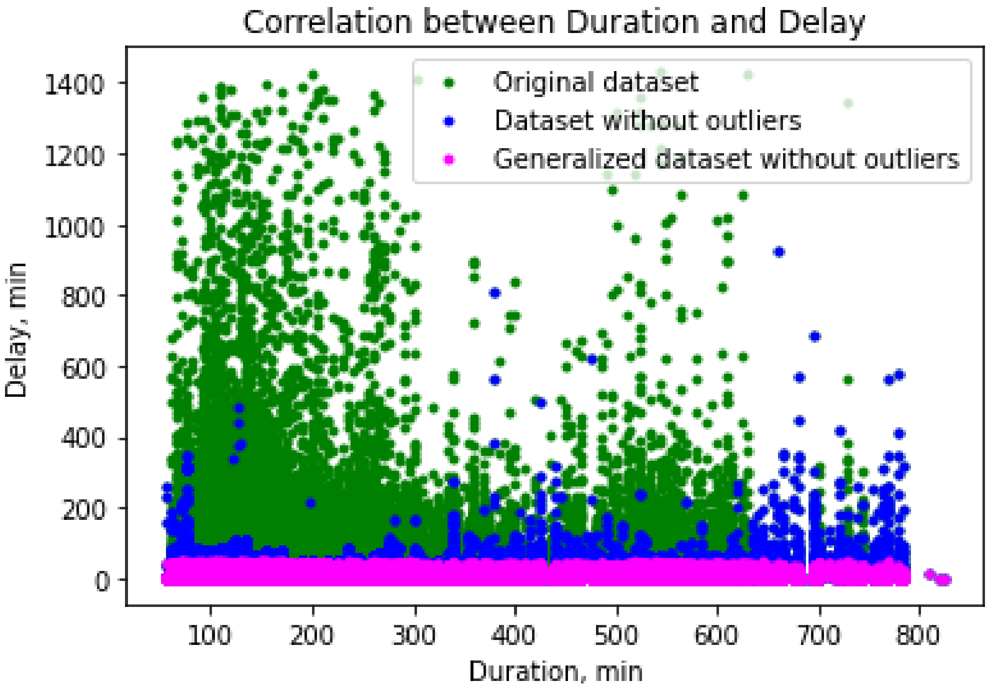


Image 12 – Generalized non-outlier data of feature 'Duration'

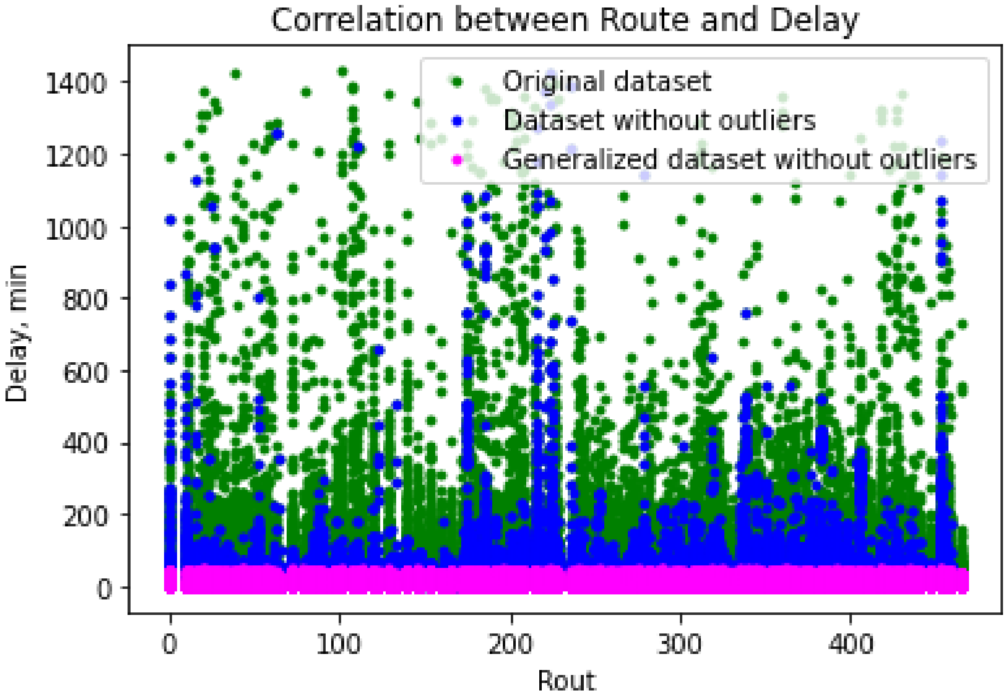


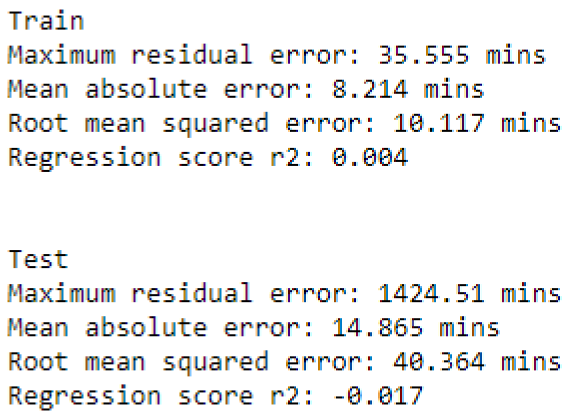
Image 13 – Generalized non-outlier data of feature 'Path'

It also worth to mention, that the PCA method was considered as not suitable as its nature is data projection, that leads to loss of some information. Therefore, the features were analyzed in terms of outlier issue separately.

Machine learning models creation and analysis

For the task solution research 7 models were created: Linear Regression model (L_R), Polynomial Regression model with degree of 3 (P_R), Linear Regression model with Lasso regularization (Lasso_R), Polynomial Regression model (3 degrees) with Lasso regularization (PL_R), Linear Regression model with Ridge regularization (Ridge_R), Polynomial Regression model (3 degrees) with Ridge regularization (PR_R) and Support Vectors Regression model with epsilon 0.5 and C 1 parameters (SV_R).

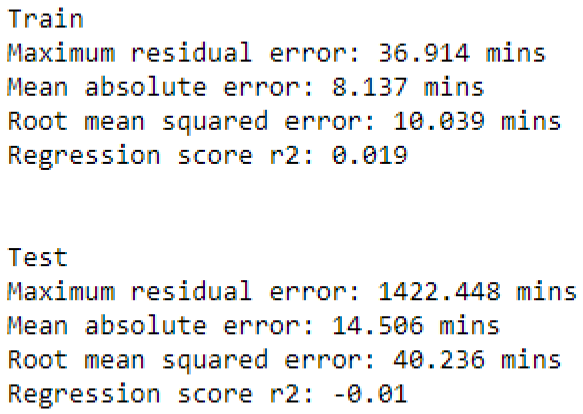
Each model has its representation results depicted on images 14 – 20.



```
Train
Maximum residual error: 35.555 mins
Mean absolute error: 8.214 mins
Root mean squared error: 10.117 mins
Regression score r2: 0.004


Test
Maximum residual error: 1424.51 mins
Mean absolute error: 14.865 mins
Root mean squared error: 40.364 mins
Regression score r2: -0.017
```

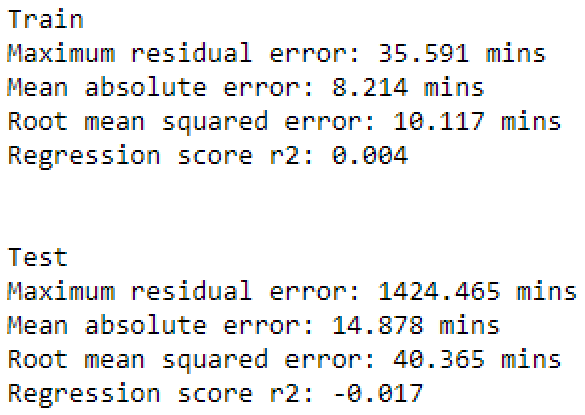
Image 14 – L_R model representation



```
Train
Maximum residual error: 36.914 mins
Mean absolute error: 8.137 mins
Root mean squared error: 10.039 mins
Regression score r2: 0.019


Test
Maximum residual error: 1422.448 mins
Mean absolute error: 14.506 mins
Root mean squared error: 40.236 mins
Regression score r2: -0.01
```

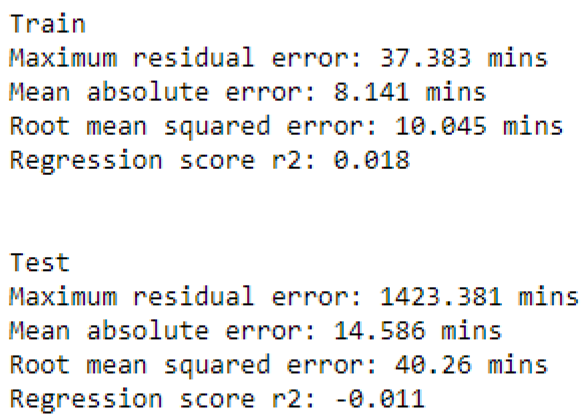
Image 15 – P_R model representation



```
Train
Maximum residual error: 35.591 mins
Mean absolute error: 8.214 mins
Root mean squared error: 10.117 mins
Regression score r2: 0.004


Test
Maximum residual error: 1424.465 mins
Mean absolute error: 14.878 mins
Root mean squared error: 40.365 mins
Regression score r2: -0.017
```

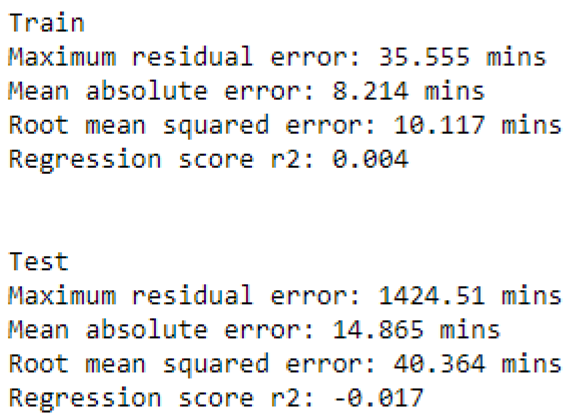
Image 16 – Lasso_R model representation



```
Train
Maximum residual error: 37.383 mins
Mean absolute error: 8.141 mins
Root mean squared error: 10.045 mins
Regression score r2: 0.018


Test
Maximum residual error: 1423.381 mins
Mean absolute error: 14.586 mins
Root mean squared error: 40.26 mins
Regression score r2: -0.011
```

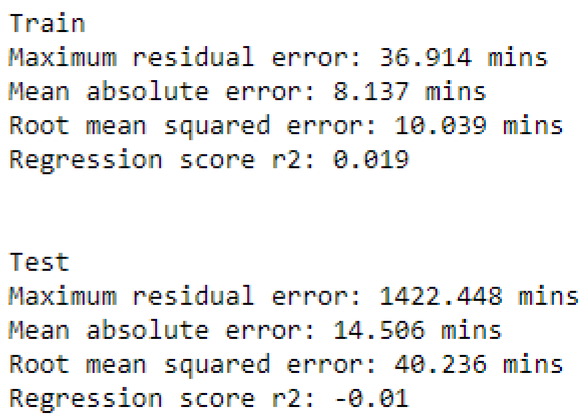
Image 17 – PL_R model representation



```
Train
Maximum residual error: 35.555 mins
Mean absolute error: 8.214 mins
Root mean squared error: 10.117 mins
Regression score r2: 0.004


Test
Maximum residual error: 1424.51 mins
Mean absolute error: 14.865 mins
Root mean squared error: 40.364 mins
Regression score r2: -0.017
```

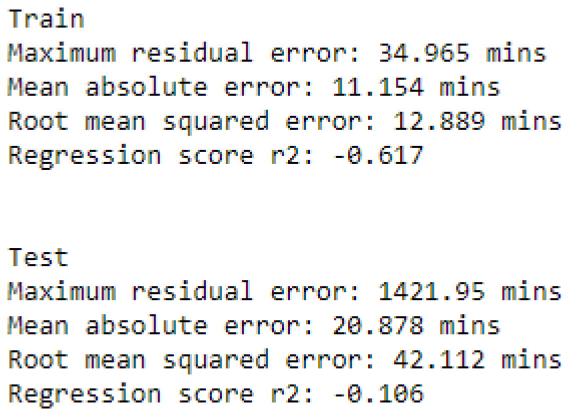
Image 18 – Ridge_R model representation



```
Train
Maximum residual error: 36.914 mins
Mean absolute error: 8.137 mins
Root mean squared error: 10.039 mins
Regression score r2: 0.019


Test
Maximum residual error: 1422.448 mins
Mean absolute error: 14.506 mins
Root mean squared error: 40.236 mins
Regression score r2: -0.01
```

Image 19 – PR_R model representation



```
Train
Maximum residual error: 34.965 mins
Mean absolute error: 11.154 mins
Root mean squared error: 12.889 mins
Regression score r2: -0.617


Test
Maximum residual error: 1421.95 mins
Mean absolute error: 20.878 mins
Root mean squared error: 42.112 mins
Regression score r2: -0.106
```

Image 20 – SV_R model representation

It can be easily seen that all the models underfit as both their train and test errors are big enough. The better performance on train dataset can be explained by outliers removal. Nevertheless, all the models weren't learned properly.

But there's still need to define which model performs better on test dataset. The comparison is provided in the table below.

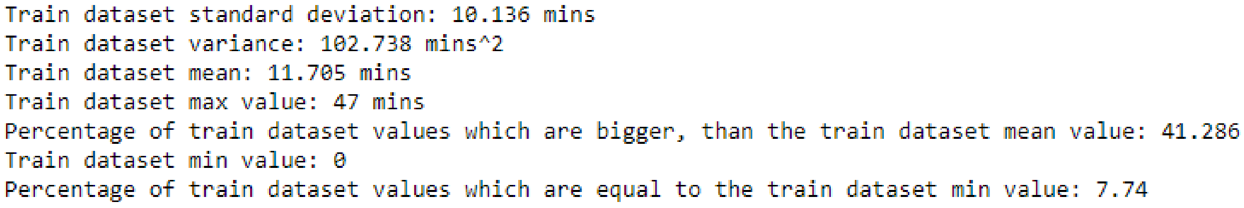
Table – Comparison of the models

Models / Errors	MRE, mins	Mean absolute error, mins	RMSE, mins	Regression score R^2
L_R	1424.51	14.865	40.364	-0.017
P_R	1422.448	14.506	40.236	-0.01
Lasso_R	1424.465	14.878	40.365	-0.017
PL_R	1423.381	14.586	40.26	-0.011
Ridge_R	1424.51	14.865	40.364	-0.017
PR_R	1422.448	14.506	40.236	-0.01
SV_R	1421.95	20.878	42.112	-0.106
* SV_R model results can significantly vary from one initialization to another				

In terms of maximum residual error metrics the SV_R model demonstrates the best result, while in terms of all the rest metrics P_R and PR_R models provide the best results.

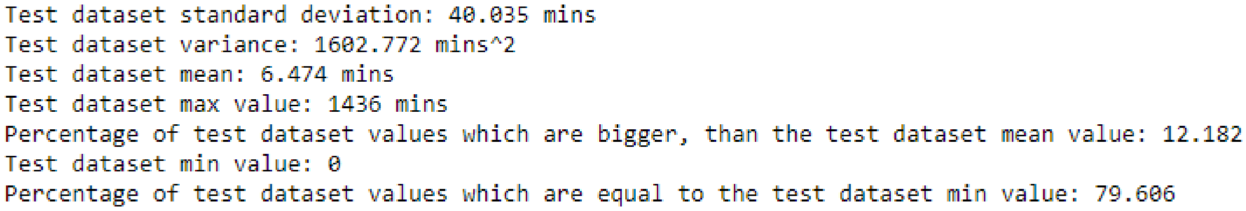
Data analysis

To explain bad results of an attempt to solute the problem the analysis of the train and test datasets was performed. Its results can be seen on images 21 – 22.



```
Train dataset standard deviation: 10.136 mins
Train dataset variance: 102.738 mins^2
Train dataset mean: 11.705 mins
Train dataset max value: 47 mins
Percentage of train dataset values which are bigger, than the train dataset mean value: 41.286
Train dataset min value: 0
Percentage of train dataset values which are equal to the train dataset min value: 7.74
```

Image 21 – Train dataset analysis



```
Test dataset standard deviation: 40.035 mins
Test dataset variance: 1602.772 mins^2
Test dataset mean: 6.474 mins
Test dataset max value: 1436 mins
Percentage of test dataset values which are bigger, than the test dataset mean value: 12.182
Test dataset min value: 0
Percentage of test dataset values which are equal to the test dataset min value: 79.606
```

Image 22 – Test dataset analysis

It can be easily noticed, that the datasets strongly differ in terms of their data distribution. Although the outliers removal also had an impact on train dataset properties, both datasets have completely different fractions of minimal values which commonly can’t be defined as outliers. Therefore, the train and test data doesn’t correspond each other.

Moreover, it’s strongly supposed, that the given dataset doesn’t content the crucial data to derive the delay function, such as weather forecast and summed delay for the departure gate (how much the airport doesn’t fit its own schedule) data.

Therefore, the task with such initial data can be considered as not able to be solved for reason of the needed data absence.

Conclusion

The task solution results can be called not satisfactory. The supposed reason is lack of efficient / suitable data (features).

Link to the solution on GitHub: github.com/Alcor2/ML-Home-Assignment-1