Machine Learning Assignment 1 Report

Klim Basargin
k.basargin@innopolis.university

1 Motivation

Detect unstable network which can be used to take steps towards adapting the data streaming and minimizing data packet loss using machine learning.

2 Data

The first dataset for regression research about bitrate prediction with features 'fps_mean', 'fps_std', 'rtt_mean', 'rtt_std', 'dropped_frames_mean', 'dropped_frames_std', 'dropped_ frames_max', 'bitrate_std', 'bitrate_mean' and target column 'target'. The second dataset for classification research about stream quality classification with features 'fps_mean', 'fps_std', 'rtt_mean', 'rtt_std', 'auto_bitrate_state', 'auto_fec_ state', 'auto_fec_mean', 'dropped_frames_mean', 'dropped_ frames_std', 'dropped_frames_max', 'fps_lags' and target column 'stream_quality'. Both datasets contain train and test parts.

3 Exploratory data analysis

I used data profiling to get more information about data. I found out that in bitrate dataset columns 'bitrate_std' and 'bitrate_mean' give a big hint to prediction, we can not know it, when we get new data. I deleted these columns. I also deleted columns 'dropped_frames_mean', 'dropped_frames_std', 'dropped_frames_max' in both datasets and 'fps_lags' from stream quality data, because of more than 93% zero values. Also I selected one meaningful feature and plot it against the target variable. Then I encoded several object columns from stream quality dataset woth OneHotEncoder to get better models performance in future analysis. Datasets do not have NaN values. Finally, I scaled datasets without target columns with MInMixScaler between 0 and 1.

4 Task

Using regression predict bitrate from the first dataset. Using classification divide streams into good(1) and bad(0) quality from the second dataset.

4.1 Regression

I used several regression models and regularizations to predict bitrate. I chose best degree for Polynomical Regression and alphas for regularizations. This is the comparison table:

Table 1. Regression results

Model	MAE	MSE	RMSE
Linear	4489.81	33699329.25	5805.11
Polynomical(4th)	4365.61	32911477.48	5736.85
Lasso(0.1)	4493.51	33722060.31	5807.06
Ridge(0.1)	4491.24	33707014.68	5805.77

4.2 Classification

I used Logistic Regression with L1 and L2 regularizations to classify streams. This is the comparison table:

Table 2. Classification results

Model	Acc.	Recall	Precision	F1-score
LogReg(L1)	0.94	0.10	0.64	0.17
LogReg(L2)	0.94	0.07	0.68	0.13

5 Results

We can see that the best performance has been achieved by Polynomical Regression 4th degree. The best performance in classification belongs to Logistic Regression with L1 regularization.

6 Data Imbalance

At this step I deleted outliers in both train datasets using Inter Quartile Range, after that I balanced classification data by making sizes of classes 1 and 0 the same.

7 Final Results

Table 3. Regression results

Model	MAE	MSE	RMSE
Linear	4643.64	59262194.01	7698.19
Polynomical(2th)	$5.5*10^5$	$2*10^9 * 10^5$	$1.4*10^7$
Lasso(1.1)	4439.84	33810735.34	5814.70
Ridge(2.2)	4450.76	33803938.92	5814.12

Table 4. Classification results

Model	Acc.	Recall	Precision	F1-score
LogReg(L1)	0.56	0.67	0.09	0.17
LogReg(L2)	0.59	0.63	0.09	0.17

8 Conclusion

After all data transformations we can see that deleting outliers did not help regression models. Ridge has the best result, but it is worse than Polynomical Regression before deleting outliers. If we talk about classification, results have changed, we have more recall and less precision and accuracy, f1-score stays the same. In my opinion it is bad, when we classify low quality stream as high (low precision), but not so bad when we classify high quality stream as low (low recall). In this case I think that results before deleting outliers and balancing were better.