Assignment 1

Ostapovich Oleg
o.ostapovich@innopolis.university

1 Motivation

With the appearance of cloud gaming, people have access to a high-quality gaming experience without the need to buy expensive PC accessories. The only problem that is difficult to solve is the quality of the Internet connection.By analyzing the data of the Internet connection, it is possible to reduce the number of problems and improve the gaming experience.

2 Data

Classification data have 12 columns: fps mean, fps std, fps lags, rtt mean, rtt std, dropped frames mean, dropped frames std, dropped frames max, auto bitrate state, auto fec state, auto fec mean, stream quality. First 3 represent Frames per second on peoples screen. Next 2 describe how fast signal goes from server to client. Dropped frames shows how much data was lost. Auto bitrate state and fec give information about how automatic modules try to maintain connection. The result of all these factors is expressed in the column "stream quality".

Regression data have 10 columns: fps mean, fps std, rtt mean, rtt std, dropped frames mean, dropped frames std, dropped frames max, bitrate mean, bitrate std, target. They have same meaning as in classification data. Target column shows value of resulting bitrate.

3 Exploratory data analysis

With the help of Pandas Profiling library gathered some insights of data. Columns describing dropped frames contain zero value in almost 97 percent of it's data and other 3 percent filled with outliers. RTT and bitrate columns also have outliers. This might cause some problems with machine learning. "Bitrate mean" column also have high correlation with target value. All of this data insights was used in data preprocessing and feature selection.

4 Task

From the point of view of machine learning(ML), the task of finding the target value is to find a function that can describe the distribution of data.

4.1 Regression

To predict bitrate of stream as value from 0 to all positive it is needed to define regression task. For this purpose polynomial, linear and ridge regression models was chosen.

4.2 Classification

Classification task was defined to find binar answer on question: "Was that stream bad or good". To answer this logistic regression with L2 regularization and ridge classifier models were used.

5 Results

The results are shown in the graph below. It looks like all models can deal with the data. In the classification task, all models show accuracy of about 90 percent. In the regression task, the MAE metric shows an error of about 1000. Taking into account that the average value of target is about 7000, we can assume that the average accuracy of the model is about 85 percent. Detailed results are presented below.

Table 1. Regression task test score

Model	MSE	MAE	r2 score
Lasso(L1)	3.8e+06	1078	0.89
Ridge(L2)	3.8e+06	1078	0.89
Linear	3.8e+06	1079	0.89
Polynomial	4.2e+06	1091	0.88

A comparison of the test and predicted target can be seen in Figure 1. Here you can see that the values closely repeat each other.

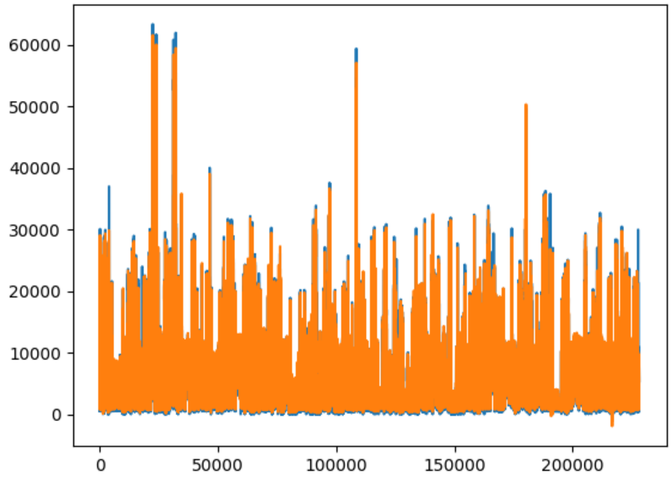

Figure 1. Predicted and test target value on one graph

Table 2 shows the results of using Logistic Regression and Ridge as ML models for classification task. Based on metrics and cross validation score of used models it is possible to say that models are underfitted because of difference between train and test predictions metrics is small. Fitting of models can be continued to avoid underfitting. If metrics will show overfit, we can try different learning rate or activation function to finish fitting.

Table 2. Classification task test score

Model	Acc	Prec	Recall
LogReg(L2)	0.93	0.52	0.20
Ridge	0.94	0.62	0.14

6 Outlier Detection

Outliers often indicate measurement error. This means that an error occurred during the measurement and this may confuse the model. To make predictions cleaner, outliers need to be removed. Isolation forest algorithm was used to detect and remove this anomalies. As a result, 39227 outliers were found in 406573 values. Comparison of the results of the models before and after outliers removal is presented in the Table 3. It seems that outliers do not affect the results of the models too much.

Table 3. Outlier detection score

Model	Acc	Prec	Recall
LogReg	0.93	0.52	0.20
Ridge	0.94	0.62	0.14
LogReg(Outliers removed)	0.93	0.52	0.20
Ridge(Outliers removed)	0.93	0.62	0.14

7 Data Imbalance

The problem of data imbalance was solved with the help of the "Imbalanced learn" library. The positive values in the "stream quality" column make up 6.8 percent of the entire training dataset. This may cause some problems, for example, the model may always assume that the stream is high-quality, and the accuracy of such a model will be 93.2 percent.To avoid this, it is needed to expand the dataset so that the number of positive quality streams is equal to negative. Comparison of the results of the models before and after data balancing is presented in the Table 4. Here it is possible to see that the Precision metric has dropped, while the Recall metric has increased

Table 4. Data imbalance score

Model	Acc	Prec	Recall
LogReg(imbalanced)	0.93	0.52	0.20
Ridge(imbalanced)	0.94	0.62	0.14
LogReg(balanced)	0.85	0.23	0.56
Ridge(balanced)	0.86	0.24	0.56

8 Conclusion

To conclude, all ML models used for classification and regression tasks cope with their tasks. We can say that the average error of classification models is 10 percent based on the accuracy metric. The difference between the real values and those predicted in the regression problem is approximately 15 percent based on the MAE metric. Based on cross validation score it is possible to say that Regression and Classification task solution models are underfitted.