ML. Assignment №1

Amir Zakirov
am.zakirov@innopolis.university

1 Motivation

Cloud gaming is rapidly gaining popularity nowadays, because it has a set of advantages over traditional ways of gaming. Firstly, the player will not have to buy an expensive console or a computer with powerful components. However, since the game runs on a remote server, real-time network analytics is required to ensure the best gaming experience. In addition to the speed of the internet-connection, the quality of the game is also affected by the set of parameters presented in the presented dataset.

2 Data

1)RTT (Round-trip time): The time taken to send the signal plus the time it takes to confirm that the signal has been received.The larger it is, the greater the delay between pressing the key and the server's response will be visible.
2)Dropped Frames : Lost video frames in the process of data transfer. If during the game there were a lot of dropped frames, then the game picture loses its smoothness.
3)FPS: Frames per second is the frequency of individual images that are displayed on a video device per second. It's also affects on the smoothness.
4)Adaptive bitrate: Its a mode on how a bitrate is calculated (categorical feature).
5)Auto FEC: Automatic Forward Error Correction mode (categorical feature).
6)The bitrate: The target variable in regression task.
7)Stream quality: The target variable in classification task.

3 Exploratory data analysis

Two datasets were presented to the student.
1)Regression Task: The dataset for the regression task did not contain categorical features, and therefore its encoding was not required. To show the relationship between target feature and the others the scatter plot was used (Figure 1).Judging by the correlation matrix, the FPS and RTT features are the most associated with the target feature. This relationship was confirmed using L1 regularization.
2)Classification Task: In the dataset for the task of classification to the main features, there were also two categorical features: Adaptive bitrate and Auto FE. Their data was encrypted using BaseNEncoder. However, based on the coefficients obtained using L1 regularization, these categorical features are slightly related to the target feature.
Also, to avoid overfitting of machine learning (ML) models, the values of various hyperparameters were considered, and its values for each dataset were taken into account.

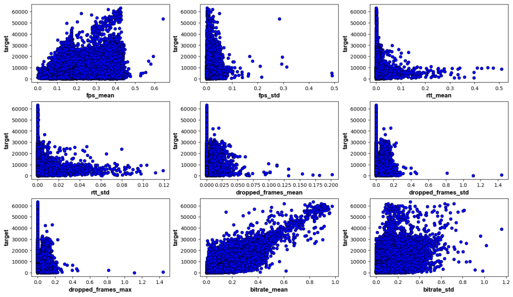

Figure 1. Relationship between target feature and the other features

4 Task

4.1 Regression

For a bitrate prediction, linear regression, polynomial regression and L1 regularization models were used. The Mean Squared Error, Mean absolute error and R-Squared metrics were calculated for each model.

4.2 Classification

For a stream quality prediction, Logistic regression with L1 regularization machine learning algorithm was used. Accuracy, Precision, Recall were calculated for each model. After that to visualize of the performance of an algorithm a confusion matrix was calculated.

5 Results

5.1 Regression

Designated metrics for each model are presented in Tables 1-4. Also, for polynomial regression, the optimal degree of the polynomial equal to 6 was calculated.

Table 1. Quantitative metrics for each ML-model in Train dataset. Independent variable : "FPS-mean"

Model	MSE, $[10^7]$	MAE, $[10^3]$	R-squared
Linear regression	3.57	4.67	0.04
Polynomial regression	3.47	4.60	0.06
Lasso	7.89	7.08	-1.14

5.2 Classification.Data Imbalance Problem.

Designated metrics for Logistic regression with L1-regularization are presented in Tables 5-6. The classification task dataset is imbalanced and contains outliers.The values in the table are

Table 2. Quantitative metrics for each ML-model in Train dataset. Independent variable : "RTT-mean"

Model	**MSE, [10^7]**	**MAE, [10^3]**	**R-squared**
Linear regression	3.64	4.70	0.01
Polynomial regression	3.58	4.63	0.03
Lasso	3.58	4.63	0.03

in the ratio 7:93 %. Capping the outlines was performed using Z-score method. And balancing data was performed according SMOTE-algorithm. A confusion matrices are presented in Tables 7-9.

Table 3. Quantitative metrics for each ML-model in Test dataset. Independent variable : "FPS-mean"

Model	**MSE, [10^7]**	**MAE, [10^3]**	**R-squared**
Linear regression	3.43	4.55	0.04
Polynomial regression	3.41	4.50	0.05
Lasso	7.66	6.72	-1.15

Table 4. Quantitative metrics for each ML-model in Test dataset. Independent variable : "RTT-mean"

Model	**MSE, [10^7]**	**MAE, [10^3]**	**R-squared**
Linear regression	3.54	4.52	0.008
Polynomial regression	3.45	4.44	0.03
Lasso	3.45	4.44	0.03

Table 5. Quantitative metrics for Logistic Regression in Train dataset

Model	**Accuracy**	**Precision**	**Recall score**
Logistic regression (LR)	0.94	0.84	0.24
LR with Z-score	0.96	0.56	0.02
LR with SMOTE	0.62	0.99	0.24

Table 6. Quantitative metrics for Logistic Regression in Test dataset

Model	**Accuracy**	**Precision**	**Recall score**
Logistic regression (LR)	0.94	0.71	0.13
LR with Z-score	0.91	0.34	0.46
LR with SMOTE	0.94	0.71	0.13

Table 7. Confusion matrix for a Logistic Regression. The actual values represented vertically, predicted values - horizontally

0	227054	848
1	13642	2052

Table 8. Confusion matrix for a Logistic Regression and Outlier capping. The actual values represented vertically, predicted values - horizontally

0	213918	13984
1	8446	7248

Table 9. Confusion matrix for a Logistic Regression and Balancing the data. The actual values represented vertically, predicted values - horizontally

0	227057	845
1	13645	2049

6 Conclusion

1)Regression task:

a.Three machine learning models were applied to predict the bitrate;

b.The corresponding metrics were calculated

1)Classification task:

a.Logistic Regression model was applied to predict the stream quality;

b.The corresponding metrics were calculated

c.Attempts to balance the dataset have been made