Machine Learning Assignment

Ahmad Aldaher
a.aldaher@innopolis.university

1 Motivation

Using a Cloud Gaming service, you no longer need to own the computational hardware since it will be owned and operated by the gaming service. All that is expected of you is to select the game you want to play and start pushing away on your buttons. Your input commands are sent to the server, calculated in the game, rendered to video audio streams, and then streamed back to you. And to do so you need a stable internet connection so you can play on real time and don't suffer from the difference between input-lag and response time. And in this homework we try to use machine learning algorithm to predict the bit-rate and signal quality for some user. and build a model which we could be used in adapting the packets sending from and to user and give more convenient experience.

2 Data

We have two data-sets, the first one contains a information about bit-rate and we will used to predict the bit-rate value this data-set contains a statistical info about some essential features of the signal like bit-rate ,frame per second , Round-trip time and Dropped or lost Frames and all these feature which will we use as predictors are numerical values. The second data-set will use to classify the quality of the signal (good or bad) this data-set contains statistical information about frame per second , Round-trip time and Dropped Frames. Also it has a categorical info about automatic forward error correction mode and numerical info form it's mean. And it has another categorical feature which is automatic bit-rate state. And the target we want to predict is the signal quality.

3 Exploratory data analysis

Investigations on the two data-sets leads to determine what kind of problem we deals with. for the first data-set it is clear that we should this data to predict the value of the signal bit-rate where as the second data should used to classify the quality of the signal. After studying the two data-sets and profiling them using pandas. we can infer form the correlation matrix which predictors are correlated with the target and have a more impact on its value. we will benefit from this information in selecting the most suitable features for each model.

4 Task

We have two task to do the first one is to predict the value of the bit-rate of the signal and for this task will use regression to find a function or estimator which will estimate the bite-rate value. we will use different model (simple linear regression, multiple linear regression, polynomial regression and also will test multiple regression with regularization (lasso)) Regrading the second data-set our task is to classify the signal quality so it clear that we need to use classification machine learning method and will use logistic regression for that.

4.1 Regression

we applied regressing on the first data-set to find a model to estimate the bit-rate. But before that we started by prepossessing the data first by remove duplicate rows and we did need to impute any data because there are no missing values as we can see from the data-set profile also there are no categorical data to encode. after that we removed all outliers that could corrupted our data. after the prepossessing the data we implemented three different regressing model, the first one is the simple linear regressing with one predictor and we chose the bit-rate feature from the data-set as it is the most correlated with target. after that we used multiple linear regressing with top correlated features from the data-set with target and final we used polynomial regressing on the multiple data and tried different degrees for the model. finally we implement the linear multiple model with with lasso regularization.

4.2 Classification

we applied logistic regressing on the second data-set to find a model to classify the signal quality. We started by prepossessing the data here we did not need to remove duplicate or impute any data because there are duplicated or missing values as we can see from the data-set profile. After we balanced the data so we could get a unbiased results, we used up-sampling for the small group of signal quality which is the good quality signals and up-sampled to much the size of bad quality signals group. Then we selected which features to use in our classification model and chose the correlated predictors with the target. After feature selection we used on-hot encoding to encode the categorical data and scale the data using Min-Max-Scaler. after that we removed all outliers from our data. finally after the prepossessing we logistic regressing on the data.

5 Results

Following we introduce the different results we got from our models. first the Regression part and then the classification.

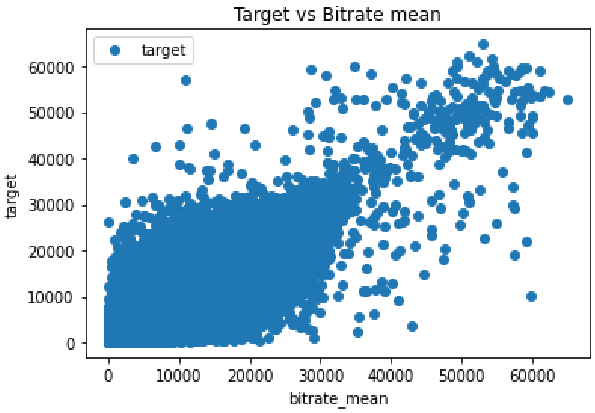

Figure 1. The data before removing outliers

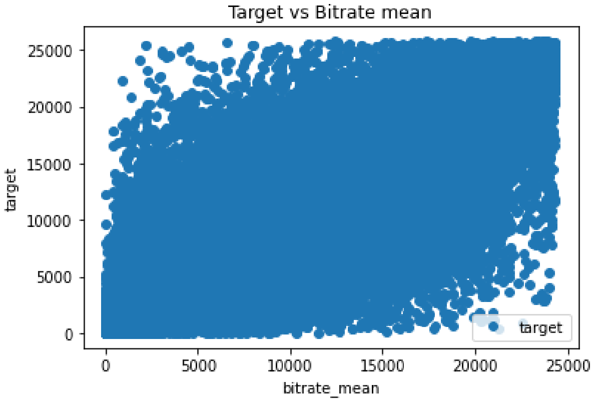

Figure 2. The data after removing outliers

5.1 Regression

here we discuss the results for the different regression models on the first data-set. the figures 1 and 2 show the target data plotted against the bitrate-mean before and after removing the outliers. and form these we graphs we can notice that it difficult to fit these data using linear regression and that we will have an under-fit as we can see from table 1 where we used different metric to evaluate the models we used. and we can see that non of these models was good for modeling such data, and we conclude form that we should for another method in machine learning arsenal to solve this problem.

Table 1. Regression Errors

Model	MAE	MSE	RMSE
Simple Linear	1079	3825402	1955
Multiple Linear	1072	3801263	1949
Polynomial degree 2	1055	3888569	1971
Polynomial degree 5	12105	2598439	518329
Multiple degree with lasso	1067	3801890	1949

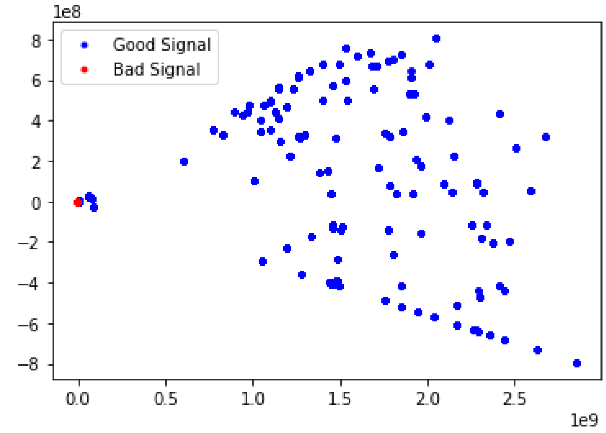

Figure 3. Classification data

5.2 Classification

we used logistic regression to classify the second data-set and we selected the correlated features with the signal quality as input to the classifier. Figure 3 shows the scatter for the labeled data after using PCA to reduce dimensionality.

we used different metric to evaluate the output of the classifier and we tested it with balanced and imbalanced data to investigate on the effect of data balancing. Also we tested with and without removing the outliers. Table 2 shows the metrics for different cases of the data. and we can notice that when we use imbalanced data we get high accuracy but that does not measures how good is our model because we have high recall, and know the imbalanced data could be based. So we should depend ion F1-score to evaluate our model, and we can see that when we used balanced data the accuracy of our model decreased but the F1-score increased. Also when can see from the table the impact of removing the outliers on the goodness of our model.

Table 2. Classification Metrics

Model	Acc.	Recall	Precision	F1-score
imbalanced with outliers	0.9334	0.4589	0.1842	0.2629
imbalanced without outliers	0.9399	0.6470	0.1505	0.2442
balanced with outliers	0.7868	0.1723	0.6072	0.2685
balanced without outliers	0.7911	0.1748	0.6024	0.2710

Figures 4 and 5 show the different metric with changing of the threshold for imbalanced data and balanced data. Where Figure 6 shows the confusion matrix for our classification model.

6 Outlier detection

Outliers in data are values which differ in nature form other points in data and might lead to errors in model estimation.

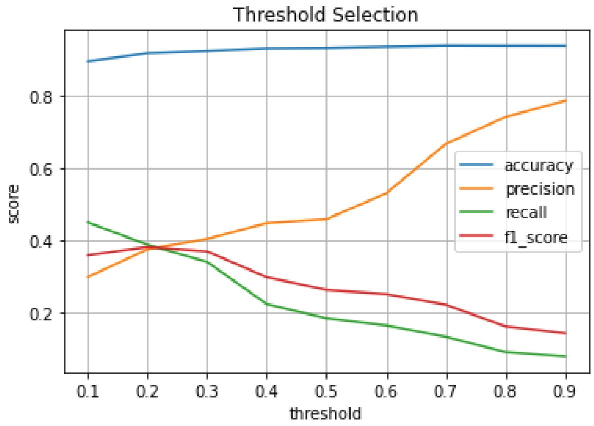

Figure 4. metric without balancing or remove outliers data

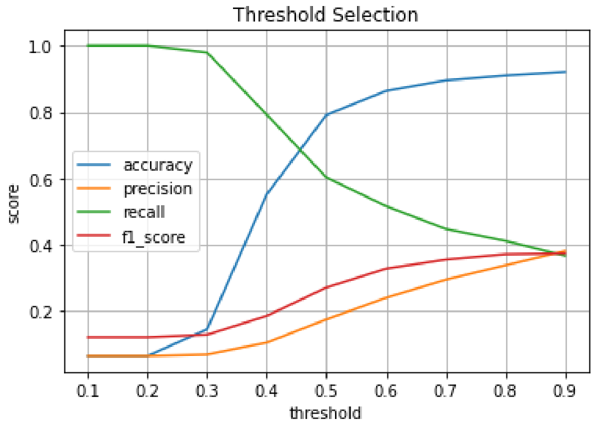

Figure 5. metric after balancing or remove outliers data

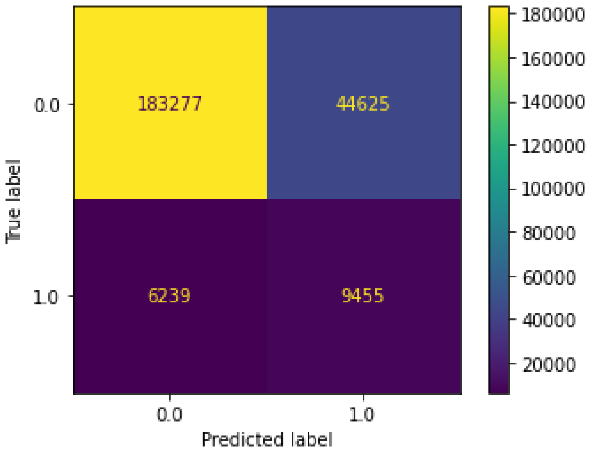

Figure 6. Confusion Matrix

as we have seen that removing outliers improved our results (in the classification part). There are different methods to remove outliers but the one we used is based on considering a normal distribution for the data and calculate the mean and std for the data, after that we excluded all points out of the range (mean-3*std, mean+3*std).

7 Data Imbalance

Imbalanced data as we noticed could lead to false results when used in classification problems. So to get unbiased results we should balance the data before classification. an examples of data balancing methods are up-sampling or under-sampling based on the size of the data.

8 Conclusion

we used regression for estimating a function to predict the value of bit-rate in the first data-set and we could not find any valid regression model due to the nature of the data. so we should try other method. Whereas for the second data-set we used logistic regression to classify the signal quality and tested the effects of balancing the data and removing outliers on improving our model.