Machine Learning Assignment 1

Ozioma Okonicha
o.okonicha@innopolis.university

1 Motivation

Cloud gaming is a relatively new form of online gaming. It operates by rendering the game data through a high-speed network on servers not the player's system; and these servers are cloud-based and high-speed. Since users do not always have stable internet for the required streaming quality, it leads to data packets loss. To enhance the quality of players' experience, adaptive algorithms need to be developed to improve the streaming and minimize data packets loss. With the help of Machine Learning algorithms, some of which have been taught to us in this course, given certain statistics about the user we can try to estimate two things:

- the bitrate (bits transferred per unit time): **Regression**
- the stream quality (good or bad): **Classification**

2 Data

For the regression task, in our training set we have 379021 rows and 10 columns: fps_mean, fps_std, rtt_mean, rtt_std, dropped_frames_mean, dropped_frames_std, dropped_frames _max, bitrate_mean, bitrate_std, target. All of them have type real number. The first 9 columns are all features and the 10th column [target] is the predictor.

For the classification task, in our training set we have 406572 rows and 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. The first 11 columns are all features and the 12th column [stream_quality] is the predictor. 9 of the features have type real number and 2 are categorical: auto_bitrate_state (off, full, partial) and auto_fec_state (partial, off).

3 Exploratory data analysis

The first insight noted is that there are no NaN values in both datasets, hence there was no need for imputation. Next in the bitrate dataset, pandas-profiling report showed that the 3 important columns in this dataset are: fps_mean, bitrate_mean and bitrate_fps. Furthermore, out of these three, when we perform feature selection using `SelectKBEst` and got that bitrate_mean is the only feature with any significant impact.

For the stream quality dataset, pandas-profiling report graphs showed the important features. After proceeding to perform feature selection using `SelectKBEst` again and got that fps_lag is the feature with the most significant impact for this dataset.

I also performed `PCA` with `n_components=1` after seeing the significant features in each of the datasets. However, it was not able to preserve a high variance for either of the datasets hence the reduced data were not used in training the models, they were used purely for visualization.

4 Task

The goal is to use machine learning to find the bitrate to send and then label is the stream quality is good or bad. To first visualize our dataset, I used the dominant feature after the feature selection against our predictors for each data set.

For both datasets, before training the data was scaled using a `StandardScaler` and in the case of the classification task, the categorical features were also encoded using `LabelEncoder`.

4.1 Regression

For the regression task, three estimators were used:

- Linear Regression
- Polynomial Regression (degree 2)
- SGD Regression

The linear regressor was fairly straightforward. The polynomial regressor was interesting because I checked various degrees and the 2nd degree was just slightly better than 1st and 3rd degrees. Degrees higher than 3 did not perform well.

As a requirement, one of the estimators had to make use of regularization and so the SGD Regressor made use of `l2` regularization as it is more stable and computationally efficient. It is important to note that at first I tried using SVR for the model but due to the nature of the dataset, it was not feasible.

4.2 Classification

For the classification task, three estimators were used:

- Naive Bayes Classifier
- Logistic Regression
- SGD Classifier

The `GaussianNB` classifier was fairly straightforward. The logistic regressor was interesting because I tried without penalty, with `l1` and with `l2` but different solvers. There was no significant difference in any of these cases. However, it was definitely the slowest estimator.

The SGD Classifier was used with default `l2` and an optimal learning rate. It is important to note that at first I tried using SVC for the model but due to the nature of the dataset, it was not feasible.

5 Results- Comparison of selected ML models

There was no underfitting in my models. There could have been some overfitting but I made use for regularization l1 and l2 in various models as well `crossvalidation` to avoid this

I used `cv=5` for all of the models. In terms of the cross validation, for the regression task, the polynomial model is better.

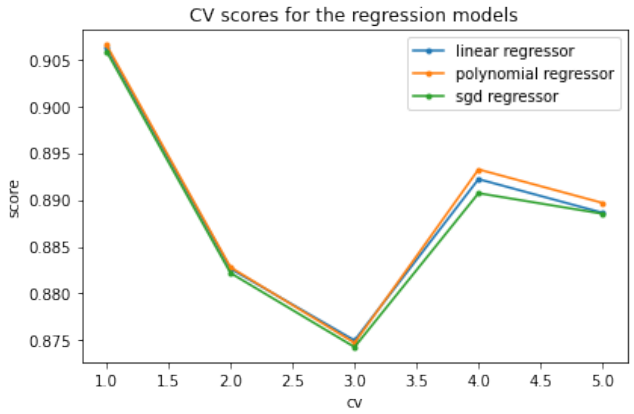

Figure 1. Cross validation for the regression models

Table 1. Regression model metrics

Model	MSE	MAE	R2 Score
`Linear Regression`	3798252.208	1076.826	0.893
`Polynomial Regression`	3775620.261	1052.145	0.894
`SGDRegressor`	3826342.088	1067.271	0.892

I also used `cv=5` for all of the models. In terms of the cross validation, for the classification task, the sgd regressor model is better.

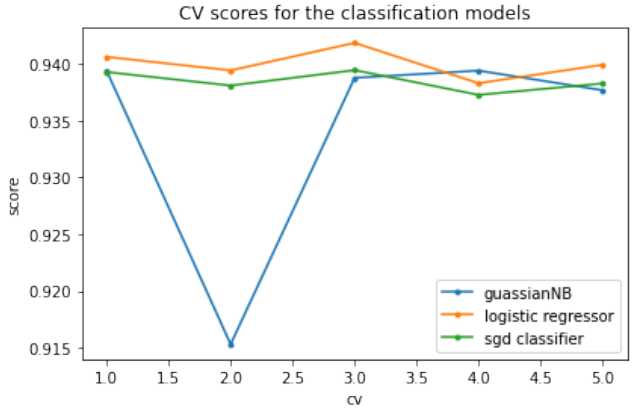

Figure 2. Cross validation for the classification models

Table 2. Classification model metrics

Model	Precision	Recall	Accuracy	F1-score
`GaussianNB`	0.632	0.113	0.938	0.192
`Losgistic Regression`	0.707	0.130	0.940	0.220
`SGDClassifier`	0.826	0.057	0.938	0.108

6 Outlier Detection & Data Imbalance

In order to detect outliers, `SGDOneClassSVM` was used. Its output is an array of −1 and 1s were −1 means it is an outlier. This way, it is easy to iterate and drop all outliers before training on the balanced dataset.

After dropping the outliers, I used the `LogisticRegression` on the newly balanced dataset. The main impact on removing the outlier is that precision dropped and f1 and recall rose; accuracy was more or less the same. By comparing them visually we get the figure below:

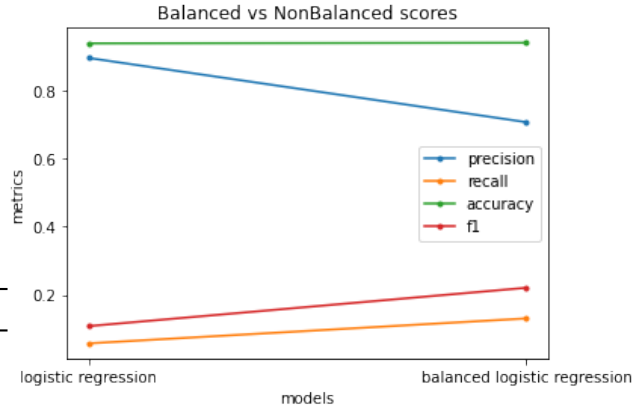

Figure 3. Logistic Regressor with the non balanced dataset vs balanced dataset

7 Conclusion

In conclusion, it was definitely an interesting dataset. For the regression task I would say my polynomial regressor worked best out of the three and for the classification task the logistic regressor worked best out of the three. In both cases, the performance differences were not vastly different, but they still did better in comparison.