ML-Assignment-1

Jameel Ahmed Syed
j.syed@innopolis.university

1 Motivation

Cloud Gaming was started during the start of the 20th Century where the user devices are just used to get the stream of the games which are rendered and run on an online server/Cloud. In the Cloud gaming world one of the challenges is to match the data rate or how fast the data is streamed to the end user because the user device will be not always be receiving the data at the same rate at which the server is streaming, if this problem is not controlled it will lead to loss of data. Therefore there should be a solution by which the Cloud gaming company will be able to predict the "Stream Quality" is good or bad and then predicts the "Bit Rate" by which the server shall send data to the end user's device by using Classification and Regression respectively.

2 Data

Regression and Classification data (features and predictor)

In Regression task we have the following features/ predictor and target: Regression Features and target: 1. Non categorical data in regression as Predictors 2. Frames per second mean, std as Predictors 3. Round trip time mean, std as Predictors 4. Dropped frames mean, std, max as Predictors 5. Bitrate mean, std as Predictors 6. Target as The main Target

---

In Classification task we have the following features/predictor and target Classification Features and Target: 1. Categorical data is Auto bit rate state, auto fec state as predictor 2. frames per second mean, std, lags as predictor 3. Round trip time mean, std as predictors 4. dropped frames mean, std, max as predictors 5. Stream quality as the main target

3 Exploratory data analysis

In Regression task: The insights from the data exploration that I got were, The dropped frames mean, std and max were having more than 95 percent of its rows as zeros, So I removed them and there was increase in the performance due to it. Current best regression score is 0.9098 score from polynomial regression of 3rd degree with MAE of 999.78 MAE.

In Classification task: The insights from the data exploration that I got were, again the dropped frames mean, std, max and fps lags were having more than 95 percent of its rows as zeros, So I removed them and there was increase in the performance due to it. Current best classification Precision is 0.9532 and Accuracy is 0.9675 from the Logistic regression with L2 Regularization

4 Task

4.1 Regression

---

1. Linear Regression: In Linear regression we have to find the estimates for the b0 and b1 as per the following equation of the straight line y = b0 + b1x . The graph of this estimated regression equation for simple linear regression is a straight line relationship between y and x. _____
2. Polynomial Regression: In Polynomial regression we have to find the estimates for the b0 and b1 so on as per the following equation of the quadratic line y = b0 + b1x^2... The graph of this estimated regression equation for 3rd degree polynomial regression is a curved relationship between y and x. _____
3. Lasso Regression: Lasso Regression is similar to Linear Regression except that the lasso regression contains the Regularization_____
4. Ridge Regression: Ridge regression is a method of estimating the coefficients of multipleregression models in scenarios where the independent variables are highly correlated._____

4.2 Classification

---

1. Logistic Regression: The Logistic regression is an example of supervised learning. It can be used to predict the probability of a binary (yes/no) event for example if a person has cancer or not, or if a person has covid19 or not._____
2. Decision Tree Classifier: A decision tree is a graph that uses a branching method to illustrate every possible output for a specific input._____
3. Random Forest Classifier: Random forests or random decision forests is an ensemble learning method for classification, regression. For classification tasks, the output of the random forest is the class selected by most trees.

5 Results

In Regression task from the results from Table: 1 we can say that the Polynomial regression of 3rd degree with score of 0.9084 which is 0.14 percent higher than the Linear, Lasso, Ridge scores for Validation set. Whereas for the Test set the score is 0.9077 for the polynomial regression which show that the model is optimally fitted, its not overfitted neither underfitter.

Finally, By this we can say that the polynomial regression with 3rd degree is the best fit for this regression task.]

Results of Classification task: In Classification task from the

Table 1. Results of Regression Models

Model	RMSE	MAE	Score
Linear Regression	1841.67	1028.40	0.9084
Polynomial Regression	1826.98	999.78	0.9098
Lasso Regression	1841.67	1028.65	0.9084
Ridge Regression	1841.67	1028.40	0.9084

results from Table: 2 we can say that the Logistic regression with L2 Regularization with Acc. of 0.9675 which is nearly 5 percent higher than the Logistic regression with Oversampling, Decision Tree Classifier, Random Forest Classifier for Test set. We can say by this that the model is not overfitted neither underfitter. Finally, By this we can say that the Logistic regression with L2 Regularization is the best fit for this Classification task.

Table 2. Results of Classification Models

Model	Acc.	Avg. Recall	Precision
Logistic Reg - L2 Reg	0.9675	0.5179	0.9532
Logistic Reg - Oversamp	0.9161	0.5590	0.5329
Decision Tree Classifier	0.9158	0.5480	0.5272
Random Forest Classifier	0.9666	0.5376	0.7480

6 Data Imbalance

For Data Balancling and Removal of Outlier: Removing the Outliers improved the model performance by nearly 2 percent. Whereas Balancing the data did not help improve the model in the case of the Classification task rather the model performed badly in my case. I followed an algorithm for the data preprocessing (As in the code) which already balanced and removed the outliers from the data set, That is the reason why balancing the data after the data preprocessing stage was not beneficial in my case.

7 Conclusion

In Conclusion, For regression task the Polynomial Regression is the best fit with MAE of 999.78 and R2 Score of 0.9098 and For the Classification task Logistic Regression with L2 Regularization is the best fit with an Accuracy score of 0.9675 and Precision score of 0.9532