Word Count: 903

CE802_Report

Name: Abdul
ID: 2202396

Machine Learning

—

De Feo, Vito

Introduction

In this report I have defined and recorded information on Classifications and Regressions methods on two different datasets that I applied on and evaluated the outcome of different models on the both datasets and also discussed come prone and cons of selecting few datasets.

This reports also contains tables of information on the performance of the datasets which also helps in evaluation of datasets. At the end of the report, it is justified that why a particular algorithm was selected in the proposal with the proof of accuracy.

THE PROCESS

Classification

Classification was performed on total of 6 different models in order to get the best possible model. In the preprocessing, the data was divided into training and testing on the ratio of 70:30 respectively. One of the empty row was filled with the mode after checking the skewness of the dataset. Before fitting the classification models, the parameters for the classifiers were hyper Para tuned using grid search in order to get the best efficiency and accuracy from the models. All the models were fitted on the same dataset and were evaluated using accuracy score. The models which were fitted on the dataset are as follows:

- Decision Tree

  Decision Tree had the best accuracy overall and was efficient in terms of time as well. It was able to outrun all the other models and quite efficiently. It had

overall accuracy of 85.6 % which was the best recorded one on this dataset as shown in Table 1.

- Random Forest

  Random Forest was right behind the Decision Tree in terms of accuracy but had a poor time complexity. That was the only and major drawback of the random forest as it is essentially combination of multiple decision trees. Decision Tree received accuracy score around 83% but costed around 50 times more in terms of efficiency as compared to the Decision Tree as shown in Table 1.

- KNN

  To be honest I was expecting to get more accuracy from the KNN model but it only received 64% accuracy as shown in the Table 1.

- SVM(Classifier)

  SVC has this major drawback where when it comes to large datasets it has never been a good choice. For this reason, it was only able to receive just 50% accuracy which is not good at all.

- Logistic Regression(Classifier)

  Logistic Regression is a classification model and it had great results. In terms of accuracy, it was came after Random Forest around 75% as shown in table 1. But it was much more time efficient as compared to random forest.

- Naïve Bayes Classifier

  In terms of accuracy, Naïve bayes Classifier performed exactly same as the KNN and both had the similar accuracy around 64% as shown in the table 1 which was second worst after SVM classifier.

Classifier	Accuracy
Decision Tree	85.6%
Random Forest	83%
KNN	64%
SVM	50%
Logistic Regression	75%
Naïve Bayes	64%

Table 1

Regression

In terms of performance, Regression algorithms were pretty bad on the dataset as expected because the type of the problem. Which proved the selection of classification the best choice. In order to use the dataset to perform the operation, some preprocessing was done which required using LabelEncoder from sklearn to convert the non-numeric data into numeric data. Two of the columns (F6, F9) were converted into

numeric relevant values using the built-in library. The models were evaluated on the basis of RMSE(Root mean square error) and MAPE(Mean Absolute Percentage Error).Total of six regression models were applied on data set.

- Linear Regressor

  Linear regressor had and accuracy of -0.01 and RMSE oh 884 and MAPE oh 2.26 as shown in Table 2. These figures were pretty bad as compared to classification model as shown in Table 1.

- Random Forest Regressor

  As compared to linear regressor, random forest did really good job. it had in RMSE oh 718.26 as shown in Table 2 and it had MA PE call 41.84. the accuracy was around 0.67, which is a bit better that's compared to linear regressor.

- Bayesian Linear Regressor

  Bayesian Linger regression was even worse then linear regression is the RMSC was above 1000 as shown in Table 2 and our Missy was around 2 .9. the accuracy was down to 0.1 as shown in Table 2 which is considerably better than linear regression but in terms of oral messy it was worst.

- Decision tree Regressor

  Decision Tree classifier was pretty much the same. The RMSE was around 1058 and MAPE around 2.95 as shown in the table. The MAPE value was quite great and the overall score was around 0.22

- KNN Regressor

  KNN had the worst results so far with RMSE above 1300 almost 1400 and MAPE around 3.4 around 3.5. It performed the worst if we compare it with rest of regression algorithms

- Support Vector Regressor

  Support Vector Regressor also had the RMSE value around 1300 but it had relatively good MAPE value as compared to KNN and worst score of all around -0.03 as shown in table.

Regressor	RMSE	MAPE	Score(Accuracy)
Linear	884	2.26	-0.01
Random Forest	718	1.84	0.67
Bayesian Linear	1195	2.9	0.1
Decision Tree	1058	2.95	0.29
KNN	1386	3.47	-0.2
SVR	1288	2.66	-0.03

Table 2

Conclusion

The classification Models worked best for this sort of problem especially Decision Tree and Random Forest. The selection of model still depends on the type and size of the dataset