Abstract

While implementing any machine learning algorithms it is good to have the descriptive knowledge of the dataset. In any dataset, in case having more than 90% of the data in target variable is from class 1 and the remaining data is from class 2. In such type of dataset, error evaluation metric accuracy is not going to help much. Having the unknown dataset with only class 1 itself gives more than 90% accuracy, which shows accuracy as evaluation metric should be ignored. Such a problem with highly skewed target outcome is known as an Imbalanced classification problem. There is a number of techniques to deal with imbalanced dataset. In this paper, we are interested to see how sampling techniques and XGBoost can be used while working with the Imbalanced dataset.

References

2. Richard G. Lyons, How Fast Must You Sample?, Test and Measurement World,


Index Terms

Computer Science          Data Mining

Keywords

Random Forest, XGBOOST, ROC curve, Anomaly detection, ROSE