Anomaly Detection in Highly Imbalanced Dataset

Sridhar Amirneni, Master of Science in Computer Science

University of Dublin, Trinity College, 2019

Supervisor: Dr. Bahman Honari

Credit Card Fraud is a major problem in todays financial world. There has been an exponential increase in the losses due to fraud in the recent years. These losses are expected to increase in the future, due to which need for research and investment is required in this field. The major problems in this are the dynamic nature of the fraud and the less financial data available. The data available has a high class imbalance with very less number of fraudulent cases present.

When this data is given as input to the current classification models, it is tending to be partial towards the majority class. Consequently, it is labeling a fraudulent transaction as a non fraudulent one. To overcome this, a Random undersampling and Synthetic Minority Over-sampling Technique (SMOTE) at the data level are implemented. The algorithms implemented were Logistic Regression, k Nearest Neighbour, Support Vector Machine, and Decision Tree Classifier. A neural network with one hidden layer was also constructed. All the models were then compared for performance between undersampling and SMOTE. The evaluation metrics used were precision, recall, f1-score, support, area under ROC curve, and confusion matrices. The results showed that Logistic Regression performed better than all the other classifiers but Neural Network with SMOTE implemented had the least number of misclassified transactions.