Credit Card Fraud Detection

Sachin Meena 200070069

Mentor
Ankit Yadav

Contents

1 Project Github link

- https://github.com/sachinmeena16/CREDIT_CARD_FRAUD_DETECTION

2 Motivation

- Cybersecurity is becoming increasingly important. When it comes to digital security, the most difficult task is detecting unusual activities.
- Credit limit in credit cards sometimes helps us make purchases even if we don't have the amount at that time.
- These features are misused by cyber attackers
- We need a system that can abort the transaction if it finds fishy

3 Data Processing Understanding

- The exact variables are not disclosed due to security concerns, however they have been modified versions of PCA. As a consequence, there are one time, 29 feature columns and one final class column to be found.
- The dataset is imbalanced towards a feature "legit transaction".
- Dataset has no null values.
- The mean amount of Fraudulent transactions is greater than the legit'.
- I removed duplicate transactions

4 Training data Test data - Splitting data

- Since dataset is significantly unbalanced, i first undersample the data from the majority class.
- We upsample the minority class using SMOTE and build a sample dataset containing similar distribution of normal transactions and Fraudulent Transactions
- We divide the data into two datasets - training data and testing data

5 Oversampling Technique .

- Data is unbalanced so i used Oversampling mathod for balancing the training data
- i used this technique for increasing the minor data points equal to majority data points.

6 Undersampling Technique .

- Data is unbalanced so i used Undersampling mathod for balancing the training data
- i used this technique for decreasing the majority data points equal to minority data points

7 SMOTE Technique .

- SMOTE starts by picking a minority class instance at random and then looking for its k closest minority class neighbours
- The synthetic instance is then constructed by randomly selecting one of the k nearest neighbours b and connecting a and b in the feature space to form a line segment.
- The synthetic instances are created by combining the two chosen examples a and b in a convex way.

8 Model Building - Random Forest

- Uses numerous decision trees to classify data
- It employs bagging and feature randomization in order to generate an uncorrelated forest of trees
- There needs to be some actual signal in our features
- The predictions made by the individual trees need to have low correlations with each other

9 Model Building - Decision Tree

- Most powerful and popular tool for classification and prediction
- Each internal node denotes a test on an attribute
- Each branch represents an outcome of the test
- Each leaf node (terminal node) holds a class label

10 Conclusion

- We find that the best model which gives highest accuracy in test data is Random Forest
- Found that the five variables most correlated with fraud are, in decreasing order, V14, V10, V12, V4, and V17
- Here we can see that RandomForest SMOTE give highest value of f1score,precision,recall compared two all models then we can say that RandomForest SMOTE is good model for this credit card fraud ditection.

THANK YOU