CSC 5825 Bella Su

Term Project

Section 1. Background/Introduction

Background

Credit Card Fraud is a commonly existing problem. The cost of fraud for financial institutions is significant. It is really important to ensure the transactions are protected. Using technologies like machine learning models to detect the fraud transactions is one of the ways. Therefore, in this project, we are going to discuss how machine learning model to detect the fraud transactions.

Introduction

The data set we are going to work with is Credit Card Fraud Detection. It contains transactions made by credit cards in September 2013 by European cardholders. It contains 30 numerical features in total, and they are V1-V28, Time, Amount. Feature 'Class' is the response labels, and it takes value 1 in case of fraud and 0 otherwise.

The dataset is highly unbalanced (Figure_1), the positive class (frauds) account for 0.172% of all transactions. Applying learning models on a highly unbalanced data will result a bias towards majority class. Therefore, we need to balance the data before we apply any learning models.

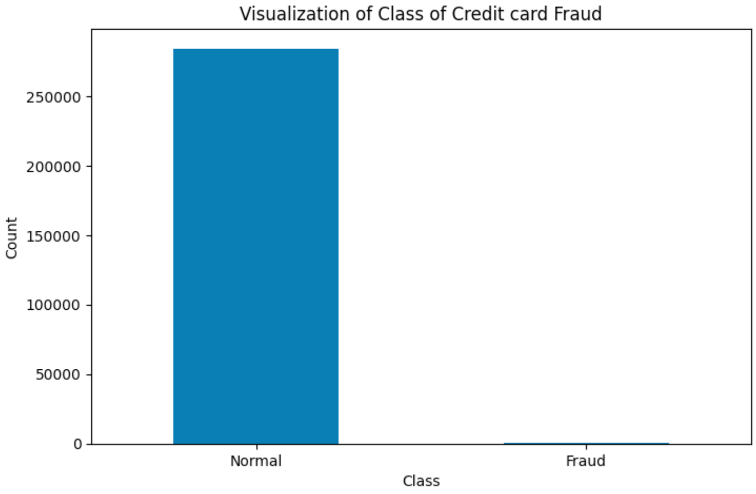

Figure_1

Section 2 Methods

Supervised Learning: Random Forest Classifier

Supervised learning is where we have input variables (features X) and an output variable (label Y). Through learning from the training data set, when we have new input data, then we can predict the output label.

The supervised learning model we are going to use in this project is Random Forest Classifier. Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction (Yiu, 2019).

Unsupervised Learning: Isolation Forest

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.

The unsupervised learning model we are going to use in this project is Isolation Forest. This algorithm builds a random forest in which each decision tree is grown randomly. At each node, it picks a feature randomly, then picks a random threshold value (between min and max value) to split the dataset in two (Sawant, 2021). The anomalies (frauds cases) are points with shortest average path length (Figure_2).

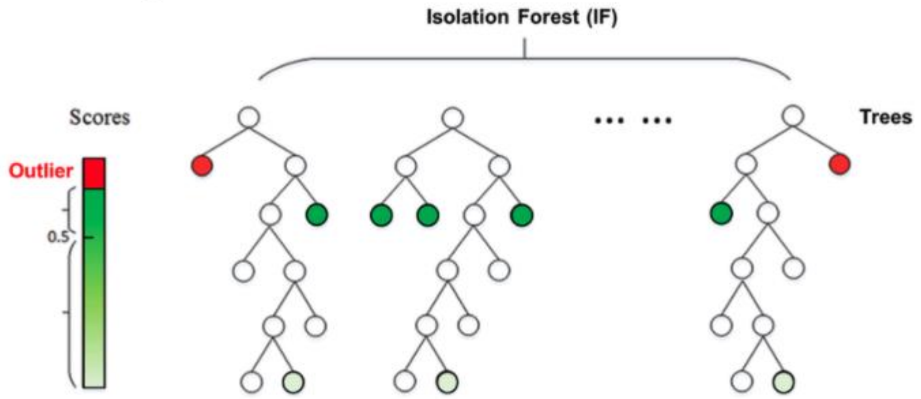

Figure_2 (Chen, Hansi, et al., 2020)

Section 3 Experiment, Result and Discussion

We have a highly unbalanced data set, and A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. One way to solve this problem is to oversample the examples in the minority class. Therefore, we apply the Synthetic Minority Oversampling Technique (SMOTE) on the data set. In this way, we can balance the class distribution but does not provide any additional information to the model (Brownlee, 2020).

The next step, we apply the Random Forest algorithm on the training data set for both before applying SMOTE and after SMOTE. We get the accuracy as 0.99964, and F1-score as 0.87843 before applying SMOTE. get the accuracy as 0.99957, and F1-score as 0.86545 after applying SMOTE. For the Isolation Forest, we get the accuracy as 0.99806, and F1-score as 0.3519.

Both of the algorithms give us high accuracy (Table_1) which is typically excellent for a binary classification prediction. However, the issue in this case is that we need to protect user’s finance by flagging as many as fraud transactions as possible, but at the same time we need to reduce the false flag to keep the normal transactions pass through smoothly. Therefore, F1-score is really important in this case.

Method	**Accuracy**	**F1-score**
Random Forest with unbalanced data	0.99964	0.87843
Random Forest	0.99957	0.86545
Isolation Forest	0.99806	0.38519

Table_1

Section 4 Conclusion

- It is possible to have a good accuracy with the unbalanced data set
- Random Forest Classifier offered us the better accuracy and F1-score compare to Isolation Forest.

Bibliography

Brownlee, Jason. "SMOTE for Imbalanced Classification with Python." *Machine Learning Mastery*, Machine Learning Mastery, 17 Jan. 2020, machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

Chen, Hansi, et al. "Anomaly detection and critical attributes identification for products with multiple operating conditions based on isolation forest." Advanced Engineering Informatics, vol. 46, no. 1011, Oct. 2020.

Yiu, Tony. "Understanding Random Forest." Towards Tata Science, Towards Tata Science, 12 June 2019, towardsdatascience.com/understanding-random-forest-58381e0602d2.

Sawant, Sumeet. "Credit Card Fraud Detection Using Unsupervised Learning." *The Startup*, The Startup, 5 Feb. 2021, medium.com/swlh/credit-card-fraud-detection-using-un-supervised-learning-8f3cfd6be765.