Statistical Methods

A Project Report on

[Credit Card Fraud Detection using Machine Learning]

Saumen Mondal (Roll No. CS2230) and,
Ritesh Kumar Tiwary (Roll No. CS2224)

On 21st April, 2023

Instructor

Prof. Shyamal Krishna De

Indian Statistical Institute, Kolkata

Abstract

The purpose of this project is to detect the fraudulent transactions made by credit cards by the use of machine learning techniques, to stop fraudsters from the unauthorized usage of customers' accounts. The increase of credit card fraud is growing rapidly worldwide, which is the reason actions should be taken to stop fraudsters. Putting a limit for those actions would have a positive impact on the customers as their money would be recovered and retrieved back into their accounts and they won't be charged for items or services that were not purchased by them which is the main goal of the project. Detection of the fraudulent transactions will be made by using three machine learning techniques Logistic Regression, Random Forest, XG-Boost, those models will be used on a credit card transaction dataset.

Keywords: Credit Card Fraud Detection, Fraud Detection, Fraudulent Transactions, Logistic Regression, Random Forest, XG-Boost

Contents

1. **Introduction**
2. **Aim of the Project**
3. **Project Methodology**
   - 3.1 Business Understanding
   - 3.2 Data Understanding
   - 3.3 Data Preparation
   - 3.4 Handling Imbalanced dataset
   - 3.5 Modeling
     - *3.5.1 Logistic Regression*
     - *3.5.2 Random Forest*
     - *3.5.3 XG-Boost*
   - 3.6. Evaluation and Deployment
4. **Conclusion**
5. **Recommendations**

1. Introduction

With the increase of people using credit cards in their daily lives, credit card companies should take special care in the security and safety of the customers. According to (Credit card statistics 2021) the number of people using credit cards around the world was 2.8 billion in 2019, in addition 70% of those users own a single card at least. Reports of Credit card fraud in the US rose by 44.7% from 271,927 in 2019 to 393,207 reports in 2020. There are two kinds of credit card fraud, the first one is by having a credit card account opened under your name by an identity thief, reports of this fraudulent behaviour increased 48% from 2019 to 2020. The second type is by an identity thief uses an existing account that you created, and it's usually done by stealing the information of the credit card, reports on this type of fraud increased 9% from 2019 to 2020 (Daly, 2021). Those statistics caught my attention as the numbers are increasing drastically and rapidly throughout the years, which gave me the motive to try to resolve the issue analytically by using different machine learning methods to detect the credit card fraudulent transactions within numerous transactions.

2. Aim of the Project

The main aim of this project is the detection of credit card fraudulent transactions, as it’s important to figure out the fraudulent transactions so that customers don’t get charged for the purchase of products that they didn’t buy. The detection of the credit card fraudulent transactions will be performed with multiple ML techniques then a comparison will be made between the outcomes and results of each technique to find the best and most suited model in the detection of the credit card transaction that are fraudulent, graphs and numbers will be provided as well. In addition, exploring previous literatures and different techniques used to distinguish the fraud within a dataset.

3. Methodology

Phase I: Business Understanding

As state before credit card fraud is increasing drastically every year, many people are facing the problem of having their credits breached by those fraudulent people, which is impacting their daily lives, as payments using a credit card is similar to taking a loan. If the problem is not solved many people will have large amounts of loans that they cannot pay back which will make them face a hard life, and they won't be able to afford necessary products, in the long run not being able to pay back the amount might lead to them going to jail. Basically, the problem proposed is the detection of the credit card fraudulent transactions made by fraudsters to stop those breaches and to ensure customers security.

Business Objective: Identification of fraudulent transaction to prohibit deduction from effected customers' accounts.

3. Methodology

Phase II: Data Understanding

In the Data understanding phase, it was critical to obtain a high-quality dataset as the model is based on it, the dataset was explored by taking a closer look into it which gave the knowledge needed to confirm the quality of the dataset, additionally to reading the description of the whole dataset and each attribute. It's also important to have a dataset that contains several mixed transaction types "Fraudulent and real" and a class to clarify the type of transaction, finally, identifiers to clarify the reason behind the classification of the transaction type. I made sure to follow all of those points during the search for the most suited dataset.

3. Methodology

Phase III: Data Preparation

After choosing the most suited dataset the preparation phase begins, the preparation of the dataset includes selecting the wanted attributes or variables, cleaning it by excluding Null rows, deleting duplicated variables, treating outlier if necessary, in addition to transforming data types to the wanted type, data merging can be performed as well where two or more attributes get merged. All those alterations lead to the wanted result which is to make the data ready to be modelled.

The dataset chosen for this project didn't need to go through all of the alterations mentioned earlier, as there were no missing variables but there were duplicate rows, so we must remove them. There is some column which need feature scaling, so we do that. Any unnecessary column we drop them.

3. Methodology

Phase IV: Handling imbalanced dataset

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e. one class label has a very high number of observations and the other has a very low number of observations.

In rare cases like fraud detection or disease prediction, it is vital to identify the minority classes correctly. So model should not be biased to detect only the majority class but should give equal weight or importance towards the minority class too. Here I use Under sampling and Oversampling technique.

3. Methodology

Phase V: Modeling

After making sure that the data is ready to get modelled the three models were created using Sci-Kit learn library in python. the model Logistic Regression, Random Forest, XG-Boost they were created using Sci-Kit learn library.

Logistic Regression

Logistic Regression model is statistical model where evaluations are formed of the connection among dependent qualitative variable (binary or binomial logistic regression) or variable with three values or higher (multinomial logistic regression) and one independent explanatory variable or higher whether qualitative or quantitative.

This model managed to :

(Before Handling Imbalanced dataset)

Score and Accuracy : (0.999093),
Precision : (0.578947),
Recall : (0.846154),
F1-Score : (0.687500)

(After Handling Imbalanced dataset using Under sampling)

Score and Accuracy : (0.952632),
Precision: (0.905263),
Recall : (1.0),
F1-Score : (0.950276)

(After Handling Imbalanced dataset using Oversampling)

Score and Accuracy: (0.945238),
Precision: (0.914604),
Recall: (0.974296),
F1-Score: (0.943507)

Random Forest

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees.

This model managed to :

(Before Handling Imbalanced dataset)

Score and Accuracy : (0.999347),
Precision : (0.673684),
Recall : (0.927536),
F1-Score : (0.780488)

(After Handling Imbalanced dataset using Under sampling)

Score and Accuracy : (0.952632),
Precision: (0.905263),
Recall : (1.0),
F1-Score : (0.950276)

(After Handling Imbalanced dataset using Oversampling)

Score and Accuracy: (0.999791),
Precision: (0.999964),
Recall: (0.999619),
F1-Score: (0.999791)

XG-Boost

The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage.

It’s no wonder then that CERN recognized it as the best approach to classify signals from the Large Hadron Collider. This particular challenge posed by CERN required a solution that would be scalable to process data being generated at the rate of 3 petabytes per year and effectively distinguish an extremely rare signal from background noises in a complex physical process. XG-Boost emerged as the most useful, straightforward and robust solution.

XG-Boost

This model managed to :

(Before Handling Imbalanced dataset)

Score and Accuracy : (0.999474),
Precision : (0.757895),
Recall : (0.923077),
F1-Score : (0.832370)

(After Handling Imbalanced dataset using Under sampling)

Score and Accuracy : (0.963158),
Precision: (0.926316),
Recall : (1.0),
F1-Score : (0.961749)

(After Handling Imbalanced dataset using Oversampling)

Score and Accuracy: (0.999700),
Precision: (0.1),
Recall: (0.999401),
F1-Score: (0.999700)

3. Methodology

Phase VI: Evaluation and Deployment

So after Under sampling Random Forest and XG-Boost classifier gives us almost similar kind of accuracy. But after Oversampling Random Forest gives us better accuracy then XG-Boost classifier for this dataset.

4. Conclusion

In conclusion, the main objective of this project was to find the most suited model in credit card fraud detection in terms of the machine learning techniques chosen for the project, and it was met by building the three models and finding the accuracies of them all, the best model in terms of accuracies is Random Forest which scored 99.97% . I believe that using the model will help in decreasing the amount of credit card fraud and increase the customers satisfaction as it will provide them with better experience in addition to feeling secure.

5. Recommendations

There are many ways to improve the model, such as using it on different datasets with various sizes, different data types or by changing the data splitting ratio, in addition to viewing it from different algorithm perspective. An example can be merging telecom data to calculate the location of people to have better knowledge of the location of the card owner while his/her credit card is being used, this will ease the detection because if the card owner is in Dubai and a transaction of his card was made in Abu Dhabi it will easily be detected as fraud.