International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Rajnish Kumar

B. Tech Scholar, Department of IT, MAIT (GGSIPU), Delhi

Abstract: Financial fraud is a developing threat with many consequences in the finance industries, corporate companies and government organizations. From many criminal activities occurring in the financial industry, credit card fraudulent activities are the most prevalent. It is important for the credit card companies to be able to detect the fraud transactions so that the customers won't get charged for the items they did not purchase. The credit card fraud detection becomes challenging for the following reasons, first, the profiles of the genuine users and fraudulent behaviours change constantly and second, the credit card fraud data sets are highly skewed. In this paper, the credit card fraud detection is done using several methods of anomaly detection or outlier detection with Probability densities. It is an interesting part of unsupervised machine learning.

Key Words: Credit card fraud, applications of machine learning, data science, logistic regression, SVM, random forest

1. INTRODUCTION

Fraud in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used.

Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated.

This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time.

These are not the only challenges in the implementation of a real-world fraud detection system. However, in real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent.

The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.

2. OBJECTIVE

Our objective is to create the best classifier for credit card fraud detection. To do it, I'll compare classification models from different methods:

* Logistic regression
* Support Vector Machine
* Bagging (Random Forest)

The datasets contain transactions made by credit cards in September 2013 by European cardholders.

3. LITERATURE REVIEW

Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule or policy with an aim to attain unauthorized financial benefit.

S P Maniraj [1] In this paper, they describe Random forest algorithm applicable on Find fraud detection. Random forest has two types. They describe in detail and their accuracy 91.96% and 96.77% respectively. This paper summaries second type is better than the first type.

Suman Arora [2] In this paper, many supervised machine learning algorithms apply on 70% training and 30% testing dataset. Random forest, stacking classifier, XGB classifier, SVM, Decision tree and KNN algorithms compare each other i.e. 94.59%, 95.27%, 94.59%, 93.24%, 90.87%, 90.54% and 94.25% respectively. Summaries of this paper, SVM has the

highest ranking with 0.5360 FPR, and stacking classifier has the lowest ranking with 0.0335.

It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

4. METHODOLOGY

The approach proposed for this project is:

- Data Collection
- Data Processing and under-sampling
- Modeling
- Classification

❖ Data Collection

The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. I decided to proceed to an under-sampling strategy to re-balance the class.

It contains only numerical input variables which are the result of a PCA transformation.

In our dataset we have 30 parameters, we have the time and the amount of the transaction as well as 28 other features that are result of a PCA (Principal component analysis) dimensionality reduction in order to protect the identity and the sensitive information involved in these credit card transactions.

- **Column Description**

Time	Number of seconds elapsed between new transaction and the first transaction
V1	Maybe a result of PCA Dimensionality Reduction
V2	
..	
..	
..	
V28	
Amount	Transaction Amount
Class	1 for fraudulent transactions and 0 for genuine

❖ Data Processing and under-sampling

Data processing is done to convert the raw data into a required format. In this project, the datasets are collected from different resources which have different formats and attributes. Hence, the data can be duplicate and they may contain some attributes which are not useful. So, we convert the data into our required format with required attributes which are used to train our model.

Time is not needed for classification so I simply remove the feature from the dataset.

We need to standardize the 'Amount' feature before modelling. For that, we use the StandardScaler function from sklearn. Then, we just have to drop the old feature

The dataset is highly imbalanced! It's a big problem because classifiers will always predict the most common class without performing any analysis of the features and it will have a high accuracy rate, obviously not the correct one. To change that, I will proceed to random undersampling.

The simplest undersampling technique involves randomly selecting examples from the majority class and deleting them from the training dataset. This is referred to as random under sampling.

Although simple and effective, a limitation of this technique is that examples are removed without any concern for how useful or important they might be in determining the decision boundary between the classes. This means it is possible, or even likely, that useful information will be deleted.

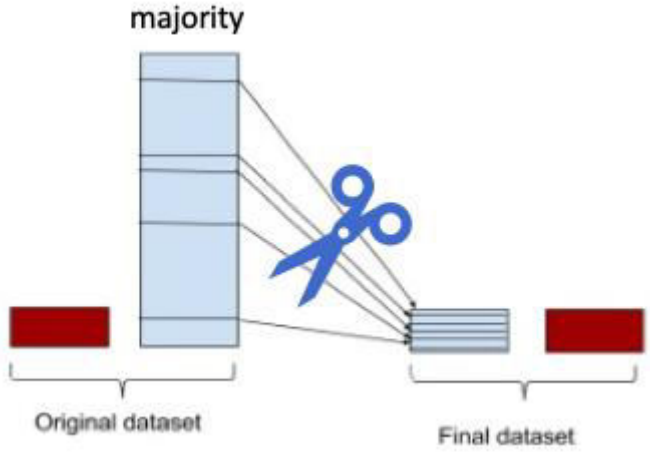

Fig1. *How undersampling works*

To undersample, we can use the package imblearn with RandomUnderSampler function.

❖ Modeling

I'll compare classification models from different methods:

- Logistic regression
- Support Vector Machine
- Bagging (Random Forest)

A. *LOGISTIC REGRESSION*

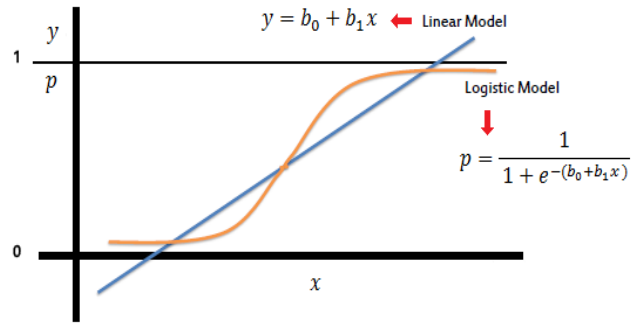

Fig1. How Logistic regression works

In Logistic Regression, input values (X) are combined linearly using weights or coefficient values to predict

an output value (y). A key difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric value. Logistic regression is a linear method, but the predictions are transformed using the logistic function.

B. *SUPPORT VECTOR MACHINE*

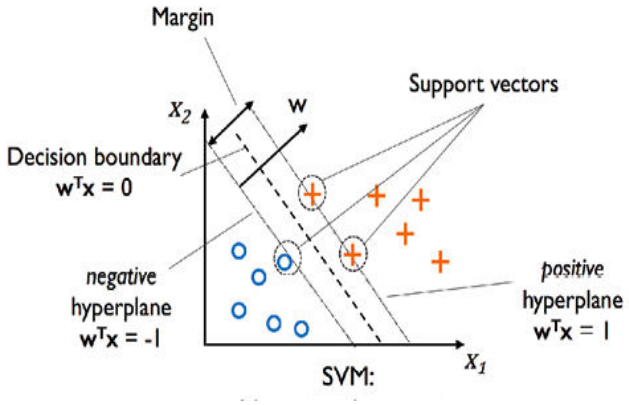

Fig 2. How SVM works

SVM Classifier uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary (hyper-plane) between the possible outputs. Support vector machines focus only on the points that are the most difficult to tell apart, whereas other classifiers pay attention to all of the points.

C. *RANDOM FOREST*

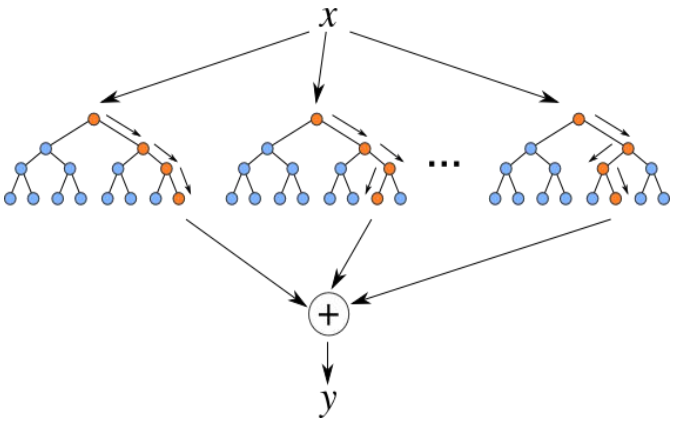

Fig3. How Random forest works

"A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models"

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 become our model's prediction.

❖ CLASSIFICATION

➢ *LOGISTIC REGRESSION*

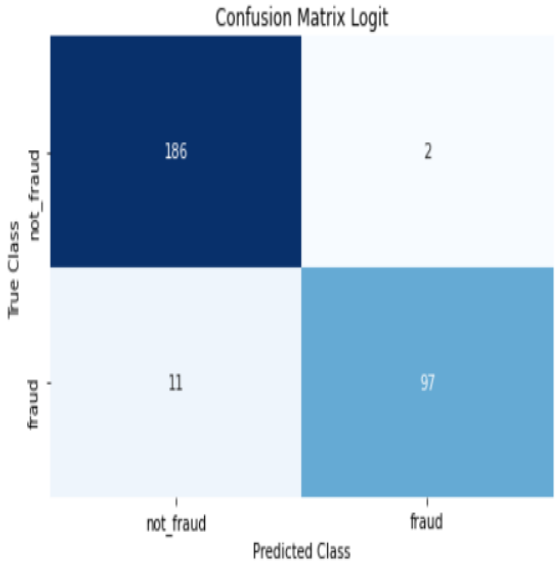

Fig 4. Confusion matrix for Logistic regression

➢ *SUPPORT VECTOR MACHINE*

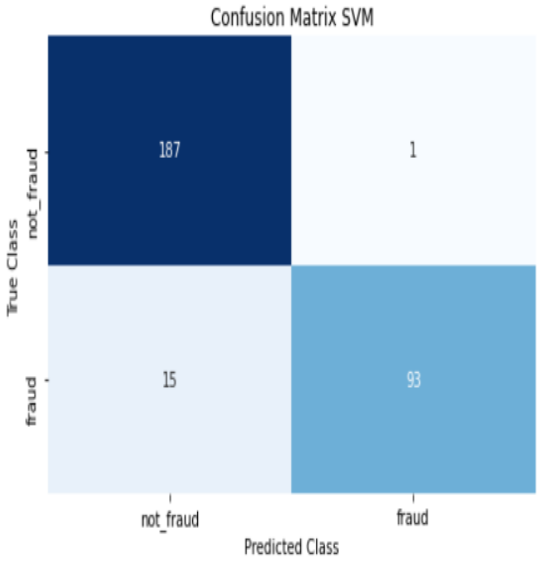

Fig 5. Confusion matrix for SVM

➢ *RANDOM FOREST*

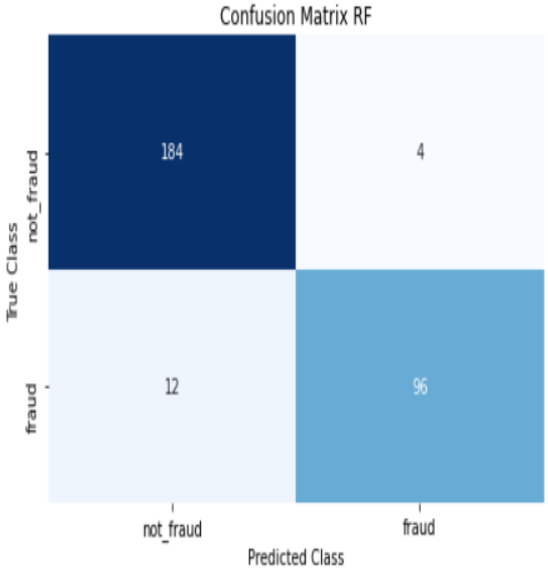

Fig 6. Confusion matrix for Random forest

5. CONCLUSION

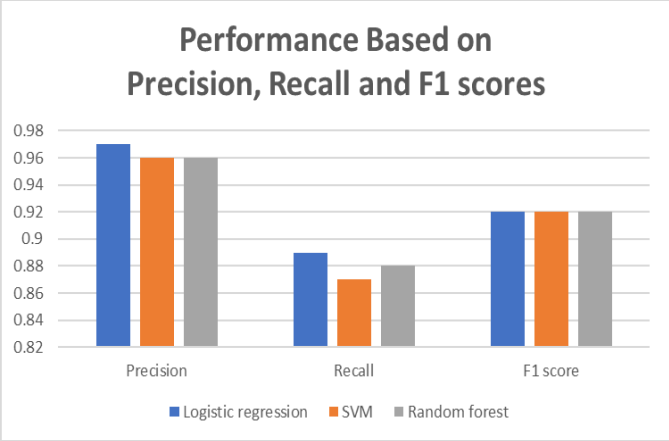

Fig 7. *Classification report*

Fraud detection is a complex issue that requires a substantial amount of planning before throwing machine learning algorithms at it. Nonetheless, it is also an application of data science and machine learning for the good, which makes sure that the customer's money is safe and not easily tampered with.

In this project three models are built using the relevant training dataset. To evaluate these models, we used the remaining transactions in the test sets. Accuracy rates were used to describe the usefulness of the models. Accuracy is probably the most commonly used metric to measure the performance of targeting models in the classification applications. However, the number of fraudulent transactions caught is also an important performance indicator.

From the chart in fig.7, it is clear that logistic regression outperform the other two models. Logistic regression from sklearn is the one that minimizes the most the false negatives so I decided to keep this model to predict credit card frauds. It's very important that a bank do not miss frauds so minimizing false negatives rate is essential.

REFERENCES

[1]. https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8

[2]. The Nilson Report. (2015). Global fraud loss reaches $16.31 Billion. Edition: July 2015, Issue 1068.

[3]. https://www.academia.edu/36810759/Machine_Learning_Approaches_for_Credit_Card_Fraud_Detection

[4]. https://www.kaggle.com/mlg-ulb/creditcardfraud

[5]. Credit Card fraud Detection Based on Transaction Behavior -by John Richard D. Kho, Larry A. Vea published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[6]. L.J.P. van der Maaten and G.E. Hinton, Visualizing High-Dimensional Data Using t-SNE (2014), Journal of Machine Learning Research