CREDIT CARD FRAUD DETECTION SYSTEM

DATA MINING FINAL PROJECT REPORT

(CSCI-B 565)

Submitted by:

BHUSHAN PATIL (bpatil)

SHUBHANGI MISHRA (shubmish)

Under the supervision of:

DR. YUZHEN YE

Luddy School of Informatics, Computing and Engineering

Indiana University Bloomington

DECEMBER 2021

Abstract

In today's time financial fraud is a growing issue with long term consequences in the financial industry. Several techniques have been discovered to get rid of this problem faced by various companies (or financial institutions). In order to analyse the huge volumes of data (or financial databases) to which these techniques are applied brings in the role of data mining. Data mining plays a crucial role in the detection of credit card fraud in online transactions. This problem is highly challenging at times primarily due to reasons like first the datasets for credit card fraud detection are highly skewed (i.e., highly imbalanced showing biasness towards one category) while the other being the behaviour of fraudulent and non-fraudulent data changes frequently. Machine learning and related methods are largely used for the given problem which comprises of decision trees, logistic regression and neural networks among others. Dataset of credit card transactions is sourced from Kaggle which consists of one csv file which is used both for training and testing purposes. The dataset contains 1296675 transactions. Different techniques are applied on the raw and pre-processed data. The performance of the techniques is evaluated based on accuracy and confusion matrix.

Table of contents

Introduction

In current time high dependency on internet technology can be observed which even contributes to increase in credit card transactions along with the escalation in credit card fraud for both offline and online transactions (predominantly for the online ones). Many computational methodologies (consisting of different software) are used for detecting fraud in various forms of businesses such as credit card, e-commerce, retail besides others. Data mining is one among the notable methodologies used for credit card fraud detection problem. These methodologies are chiefly used by banks to protect their customers from any kind of fraud as identifying any transaction as fraud is a tedious task and at times it leads to the delay in the purchase of required items by the card holder. Fraud detection in credit card is the process of identifying the authenticity of any given transaction as fraud or non-fraud. To make these predictions several machine learning techniques like Random Forest, XGBoost, Logistic Regression can be considered. Other than machine learning algorithms even rule induction techniques and deep learning methods can serve as a solution for the above-mentioned problem. While exploring different solutions for the fraud detection problem we come across a couple of challenges such as one being the unavailability of public data which basically suggests that there is a lack of proper datasets to evaluate the existing predictive models for the same besides the data present for credit card fraud detection is highly imbalanced. We can observe this for the dataset used by us in the project in which the actual number of fraudulent cases is way less than the non-fraudulent one. As a result, the models do not work as efficiently as they should be on these types of data. The most important aspect for getting efficient results is selection of optimal features for model selection along with the type of sampling approach (under sampling or oversampling) used for balancing the dataset. Under sampling has been used by us for balancing the dataset. Towards the end results have been evaluated with accuracy and confusion

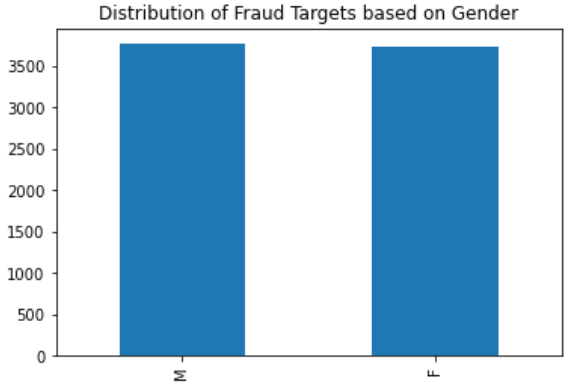
matrix considered as evaluation metric. In our project we have used Random Forest, XGBoost and KNN classifiers for predicting fraud in credit card transactions. The accuracy for both Random Forest and KNN is computed as 91.35% while for XGBoost the accuracy comes out to be 91.77%. Results suggest that XGBoost is the best model for our problem statement.

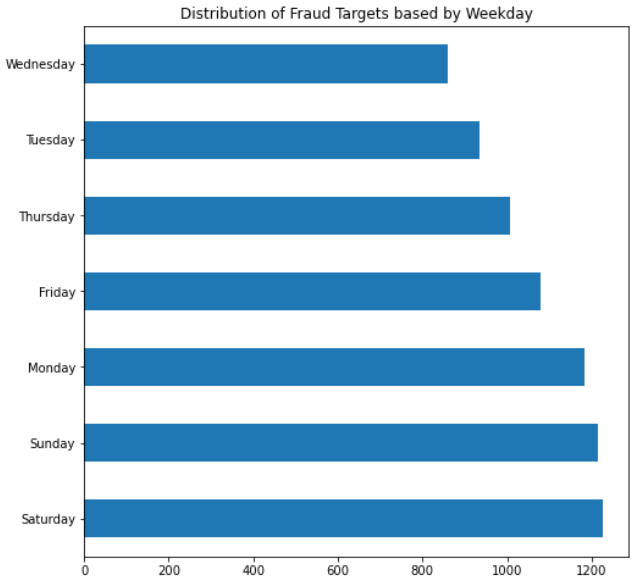
Methods

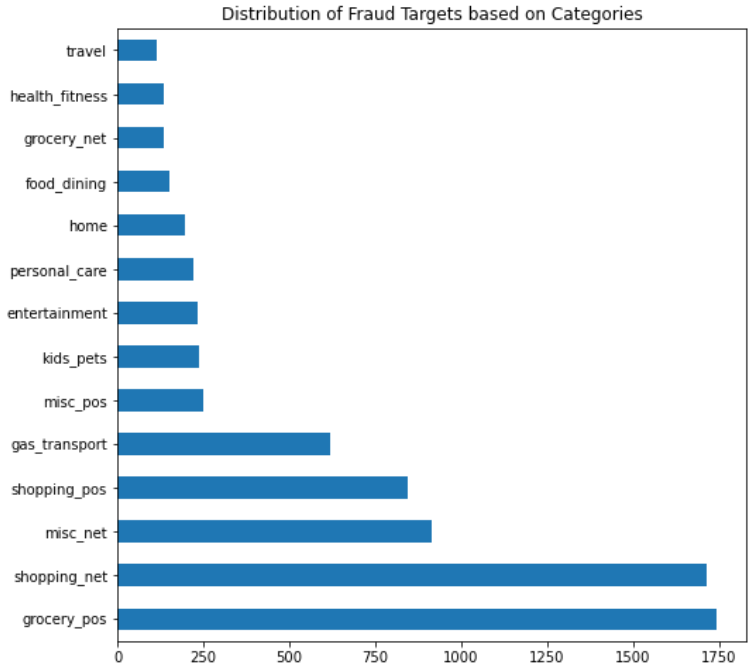
This describes the design and implementation methods for detecting fraud in credit card transactions. The steps involved are as follows:

1. **Data Pre-processing**
    - This method is used for converting the given dataset into useful format.
    - This method comprises of three steps namely data reduction, data cleaning and data transformation.
    - Data cleaning is mainly used for imputing missing values; for the given dataset we don't have any missing value.
    - The dataset used is highly imbalanced. Most of the transactions present are not fraud. If we use the imbalanced dataset as our base for the analysis of predictive models, then our algorithm overfits. To get rid of this problem we balance the data using sampling. Sampling is a part of data reduction which is performed on the subset of the entire dataset.

2. **Exploratory Data Analysis (EDA)**
    - This method is used for analysing and exploring the dataset.
    - EDA consists of several data visualization methods used for outlining the major characteristics present in the given data.
    - In our project we have used bar plots for predicting the target variable **is_fraud** based on different predictors (input variables) such as **gender**, **weekday** and **categories**.

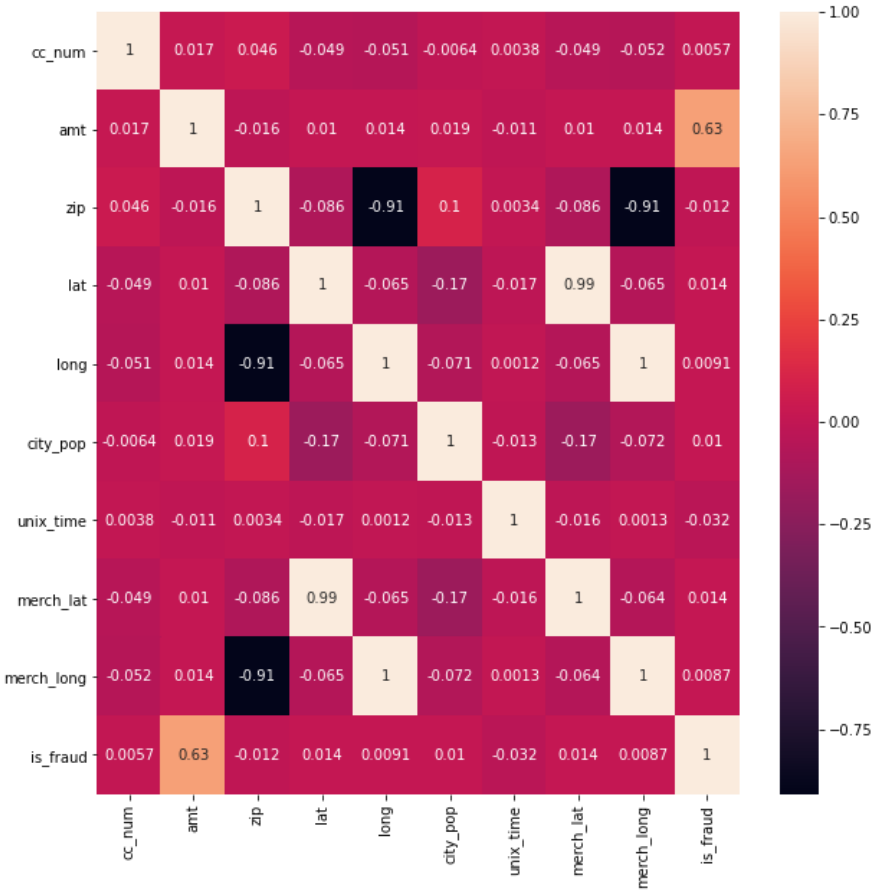

Distribution of Fraud Targets based on Gender
3500
3000
2500
2000
1500
1000
500
0
M
F


Distribution of Fraud Targets based by Weekday
Wednesday
Tuesday
Thursday
Friday
Monday
Sunday
Saturday
0
200
400
600
800
1000
1200


Distribution of Fraud Targets based on Categories
travel
health_fitness
grocery_net
food_dining
home
personal_care
entertainment
kids_pets
misc_pos
gas_transport
shopping_pos
misc_net
shopping_net
grocery_pos
0
250
500
750
1000
1250
1500
1750

3. Feature Selection

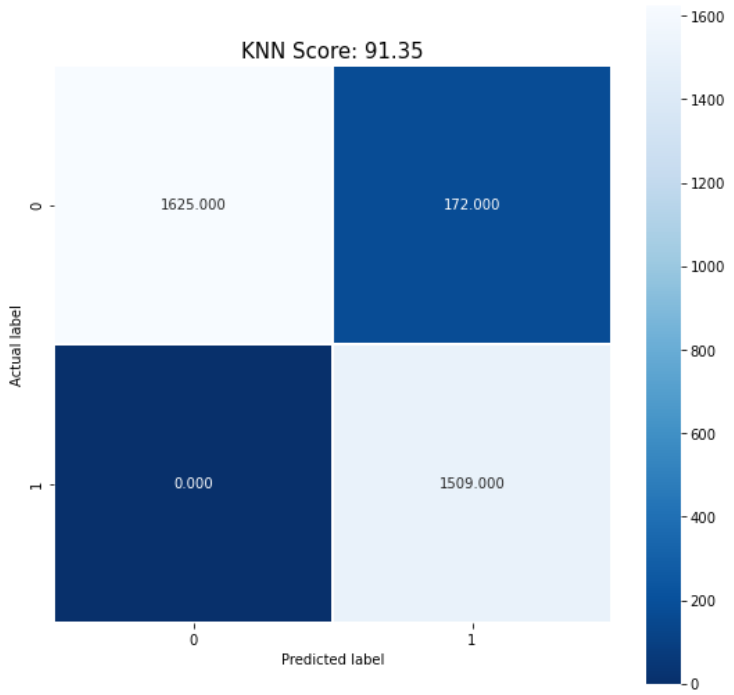
- This method reduces the number of predictors (or input variables) by selecting the most important features while discarding the least important features.
- Feature selection improves the overall performance of the model.
- In our project we have used correlation matrix as a metric for feature selection. Features having higher correlation between them are dependent on each other and are considered for model selection whereas lower correlation between two variables shows suggest that the variables are independent of each other.
- Negative correlation between two variables suggest that the behaviour of both variables are inversely related as in increase in one of the variables leads to the decrease in the other and vice-versa.



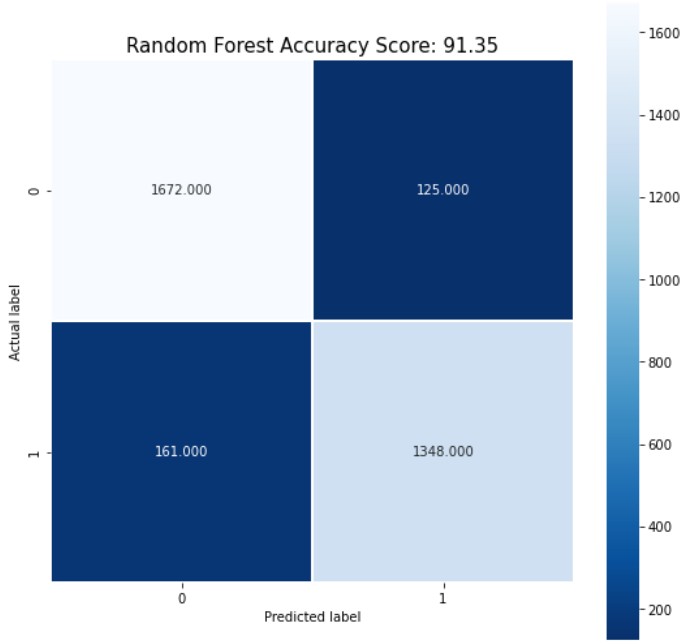
Correlation matrix

4. Modelling

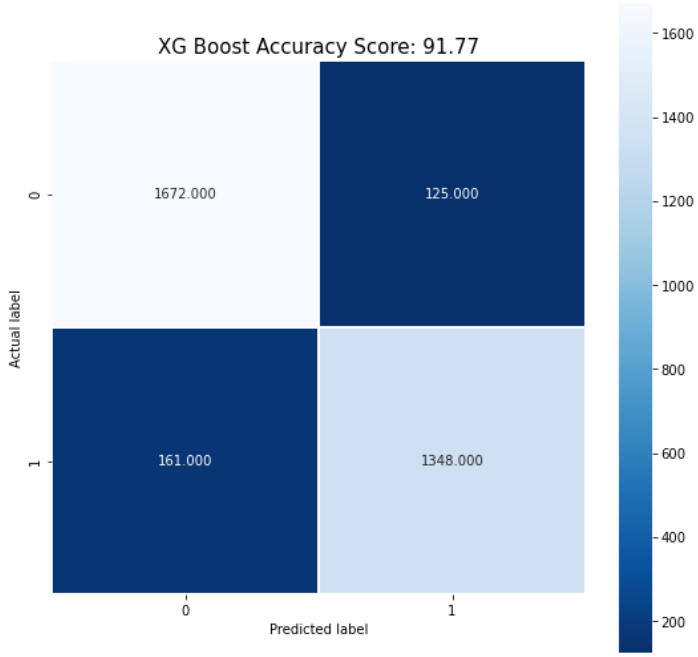
- This method involves training and testing on the given dataset.
- For training and testing the given dataset namely **fraudTrain** dataset into 70% for training and 30% for testing.
- Classifiers used for prediction are KNN, Random Forest and XGBoost. Performance is evaluated based on metrics like accuracy and confusion matrix.
- In the confusion matrix the loss function is used is mean squared error.



Confusion matrix for KNN



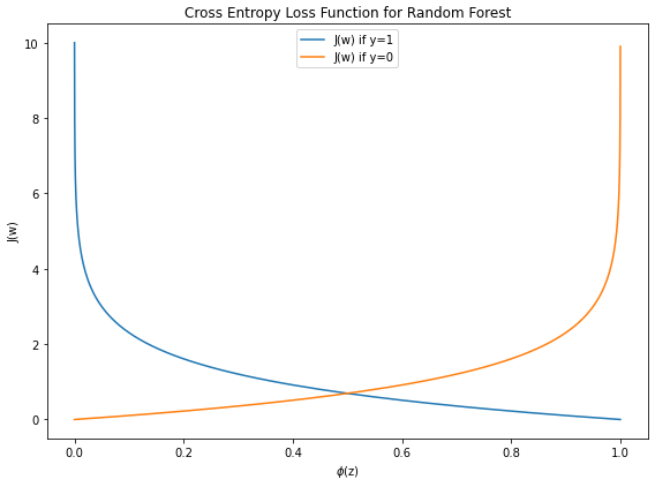
Confusion matrix for Random Forest



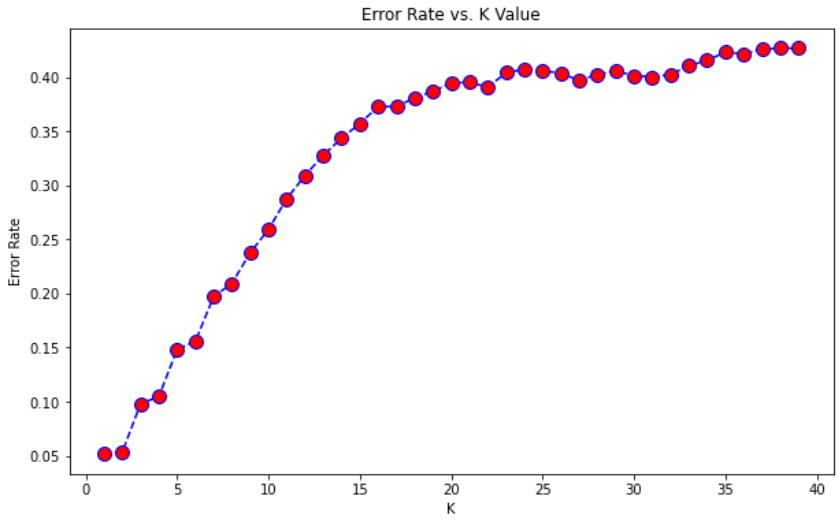
Confusion matrix for XGBoost

RESULTS

- We can observe that after sampling Random Forest has an accuracy of 91.35% while the entropy loss function value is quite small.
- Smaller loss function indicates that the errors made in the process of fitting the model is very less further stating that the predictive model used is highly efficient.



- The reason for the good performance in terms of accuracy is that Random Forest is very suitable for dealing with high dimensional noisy data (in other words highly imbalanced data).
- Other classifier used is KNN which too has an accuracy of 91.35%. Graph is plotted between k value and error rate which reflects with increasing value of k error rate also increases.



- The last classifier used is XGBoost with an overall accuracy of 91.77%.
- XGBoost performs the best for the reason being that it tunes the parameters by itself and works in a parallelized fashion distributed among clusters. We can say that balanced data in the clusters improves the overall performance.
- After comparing both the models we conclude that **XGBoost** is the best methodology for the proposed problem with an accuracy of **91.77%**.

DISCUSSIONS

Here we talk about the future scope of the project together with the challenges faced. The challenges faced are:

1. **Imbalanced Data:** The actual number of fraudulent cases is way less than the non-fraudulent ones. We can observe this for the dataset used in the given problem due to which the models don't perform efficiently.
2. **Unavailability of public data:** The data required for credit card fraud detection is highly sensitive; therefore, many times there is lack of proper (and accurate) datasets to evaluate the existing predictive models for the same.

The **future scope** of our project outlines:

- Using neural networks, deep learning methods and genetics algorithms along with other unsupervised algorithms for predicting the authenticity of fraud accurately without raising any false alarms.
- Deep learning techniques is the next big thing after machine learning which can be used as an optimization technique for understanding the pattern of credit card user in a much better way.
- Duplex verification system for both user's (debit) and seller's(credit) end should be developed. This can be used to assess even the past transactions present in the database and helps in determining whether certain transaction was fraudulent or not besides producing evidence.

REFERENCES

- https://datasciencecmu.wordpress.com/2014/04/18/the-future-of-fraud-detection-2/
- https://www.ripublication.com/ijaer18/ijaerv13n24_18.pdf
- https://www.ijcseonline.org/pub_paper/2-IJCSE-08130.pdf
- https://www.questjournals.org/jrhss/papers/vol8-issue2/B08020411.pdf
- https://www.ijsr.net/archive/v10i7/SR21712231418.pdf
- **http://www.iaeng.org/publication/IMECS2011/IMECS2011_pp442-447.pdf**
- **https://www3.nd.edu/~dial/publications/dalpozzolo2015calibrating.pdf**