International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 3101-3106 www.ijaem.net ISSN: 2395-5252

Comparative Analysis and Prediction of Credit Card Transactions

Project Guide: CH.NarayanaRao
Team Members:
S.Keerthana(173J1A05C8)
B.Khanitkar (173J1A05E9)
Y.B Varun (173J1A05F3)
M.SrikanthRaju(183JA0506)
Computer Science and Engineering

Submitted: 10-07-2021 Revised: 23-07-2021 Accepted: 26-07-2021

ABSTRACT:
The project titled “Comparative analysis and prediction of credit card transactions detects the fraudulent card during transactions and alerts the customer regarding the fraud. This project also aims in minimizing the number of false alerts. Here we implement different machine learning algorithm on an imbalanced dataset such as Light gradient classifier, XGB Classifiers. Financial fraud is an ever growing menace with far consequences in the financial industry. Data mining had played an imperative role in the detection of credit card fraud in online transactions. Credit card fraud detection, which is a data mining problem, becomes challenging due to two major reasons - first, the profiles of normal and fraudulent behaviour change constantly and secondly, credit card fraud data sets are highly skewed. The performance of fraud detection in credit card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection techniques used . This paper investigates the performance of Support vector classifier, Decision tree classifier, Random forest ,xgboost , LightGreadient, k-nearest neighbor and logistic regression on highly skewed credit card fraud data.

I. INTRODUCTION:

Credit card fraud detection is significantly difficult, but also a popular problem to solve. In our proposed system we built credit card fraud detection using Machinelearning. Withthe advancement of machine learning algorithms. Machine learning had been identified as successful measure for fraud detection. Large amount of data is transferred during online transaction process, resulting in a binary result: genuine or fraudulent. Within these sample fraudulent datasets, features are constructed. These are data points namely the age and value of the customer account, as well as the origin of the credit card.'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 itand 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.

II. LITERATURESURVEY:

[1]Fraud detection has been usually seen as a data mining problem where objective is correctly classify transactions legitimate . For classification problems many performance measures are defined in most of them which are related to correct number of cases classified

correctly.[2] A more appropriate measures are needed due to inherent structure of credit card transactions. When card is copied or stolen or lost and captured by fraudsters it is usually useduntil available limit is depleted. Rather than the number of correctly classified transactions, the solution which minimizes total available limit on cards subject to fraud is more prominent.

Since fraud detection problem has been mostly defined as classification problem, in addition to some statistical approaches many data mining algorithms has been proposed to solve that. Among these decision trees ,artificial neural networks are most popular ones.[3] The study of Bolton and Hand provides a good summary of literature on fraud detection problems.

[4] However, when the problem is approached as a classification problem with variablemisclassifications costs as discussed above, the classical data mining algorithms are not directly applicable; either some modifications should be made or new algorithms must be develop specifically for this purpose are needed.[5] An alternative approach would be trying to make use of the general purpose meta heuristic approaches like genetic algorithms.

[6]The datasets contains transactions made by credit cards in September 2013 by europeancardholders. 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.

III. EXISTINGSTATEMENT:

In most of the existing system they used machine learning approches like random fores, SVM , K-nearest algorithm ,Decision Tree algorithm which are less accurate.

IV. PROPOSED SYSTEM:

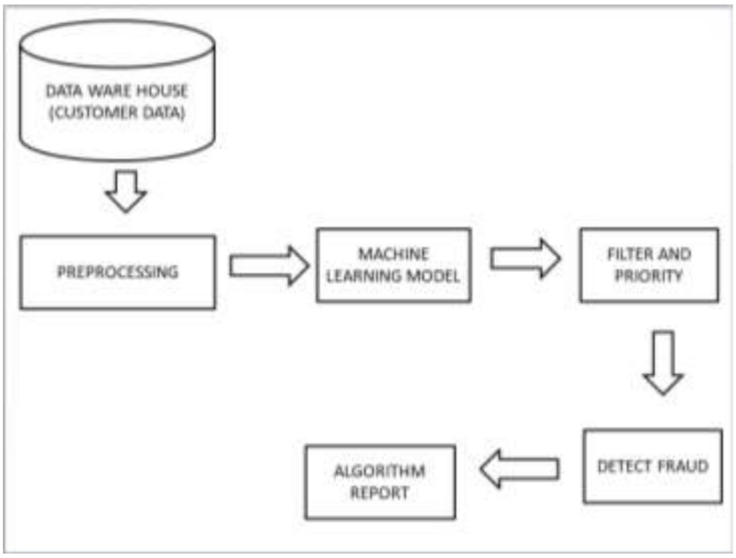
Here in this project we implement two machine learning models like Xgboost and LightGradient which can give better accuracy . principal component analysis for data prediction and also detect the fraud transactions.

V. OBJECTIVE:

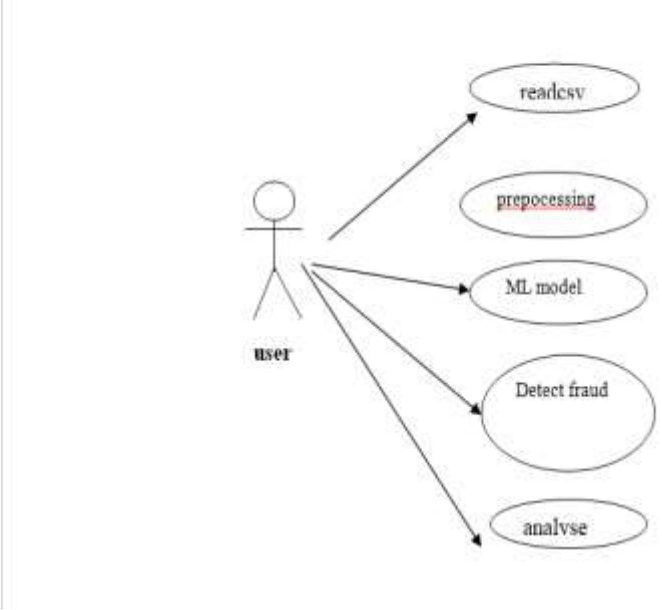
To develop a tools which can take the user inputs using principal component analysis data and identify the fraudulent transaction .

VI. METHODLOGY:

- The customer data in the data warehouse is subjected to the rules engine which consists of the fraud rule set.
- The filter and priority module sets the priority for the data and then sends it to the genetic algorithm which performs its functions and generates the output.



VII. USECASE DIAGRAM:



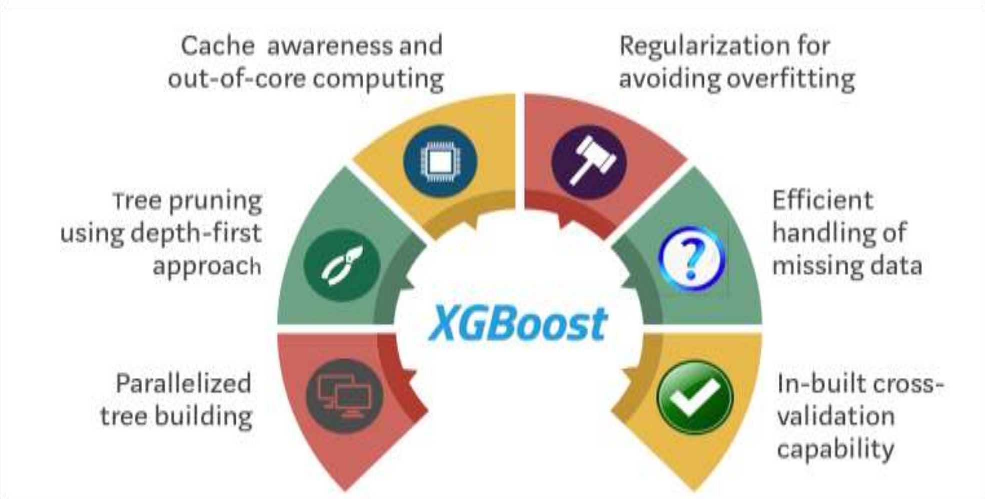
Principal Component Analysis:

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables. PCA is a most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover, PCA is an unsupervised statistical technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit

VIII. ALOGORITHMS:

1.XGBOOST:

When using gradient boosting for regression, the weak learners are regression trees, and each regression tree maps an input data point to one of its leafs that contains a continuous score. XGBoost minimizes a regularized (L1 and L2) objective function that combines a convex loss function (based on the difference between the predicted and target outputs) and a penalty term for model complexity (in other words, the regression tree functions). The training proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It's called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.


Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

2. LIGHT GRADIENT:

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSSand EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks

IX. SOFTWARE REQUIREMENTS:

Programming Language: PYTHON
Operating system:-windows 10

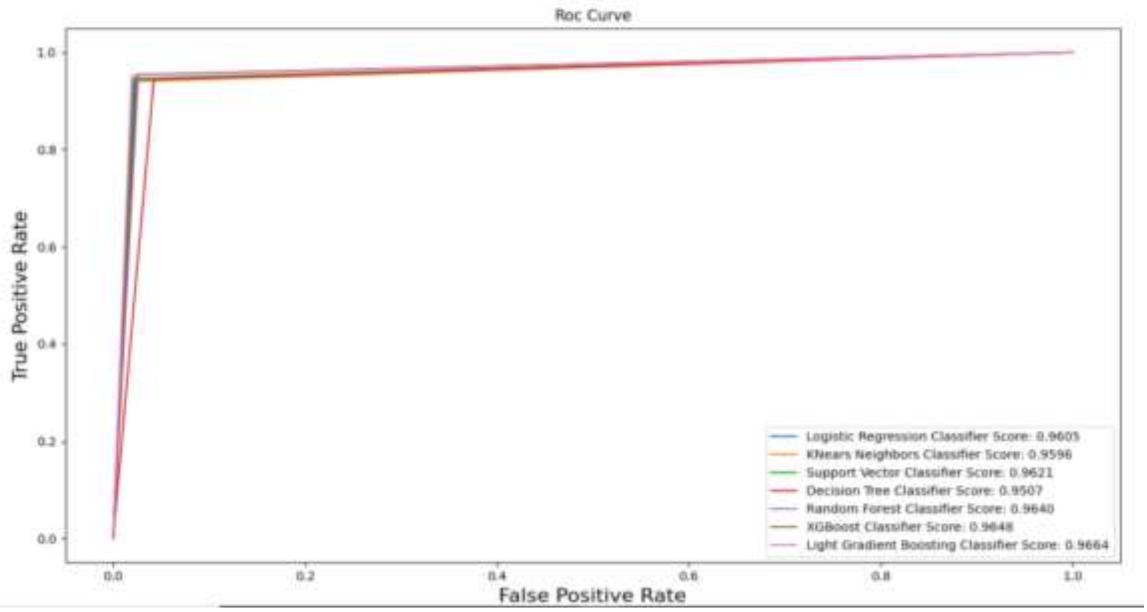
X. HARDWARE REQUIREMENTS:

RAM:4GB
System: Intel core
Hard disk:- 10gb(min)

XI. LIABRARIES:

NUMPY
Pandas
SKLearn
NLTK
Matplotlib
Seaborn

XII. RESULTS:



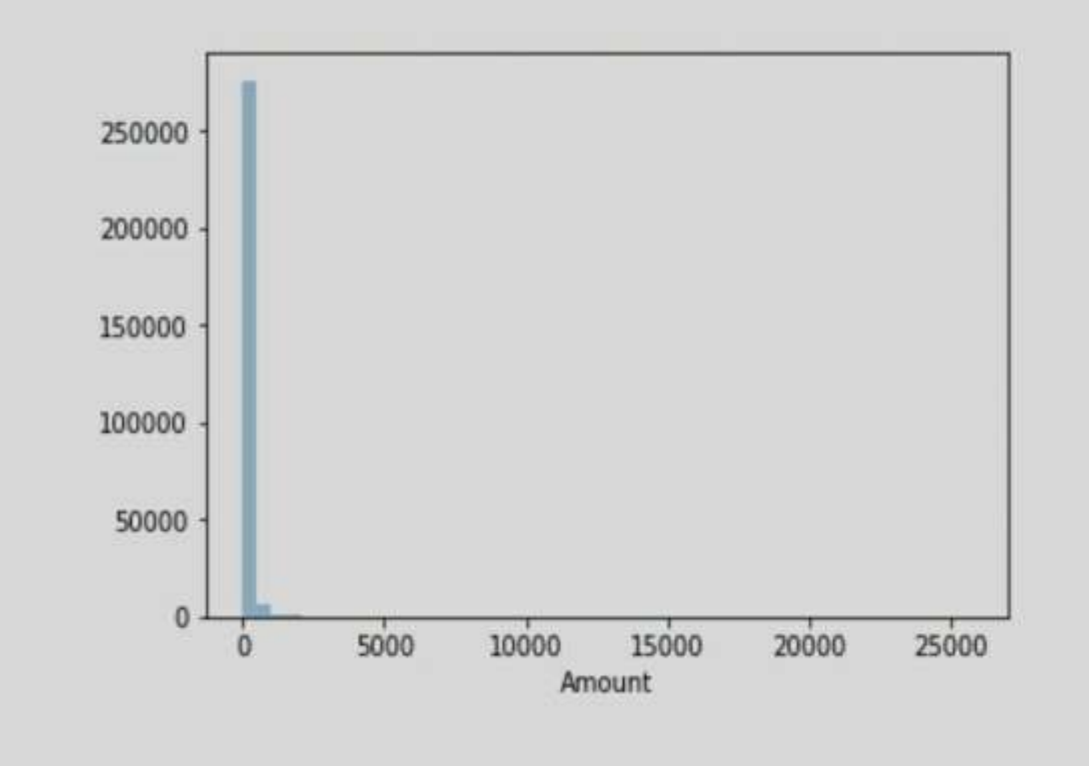
In this we can identify that XGBoost Classifier accuracy is 96.48 and Light Gradient Boosting Classifier accuracy is 96.64.

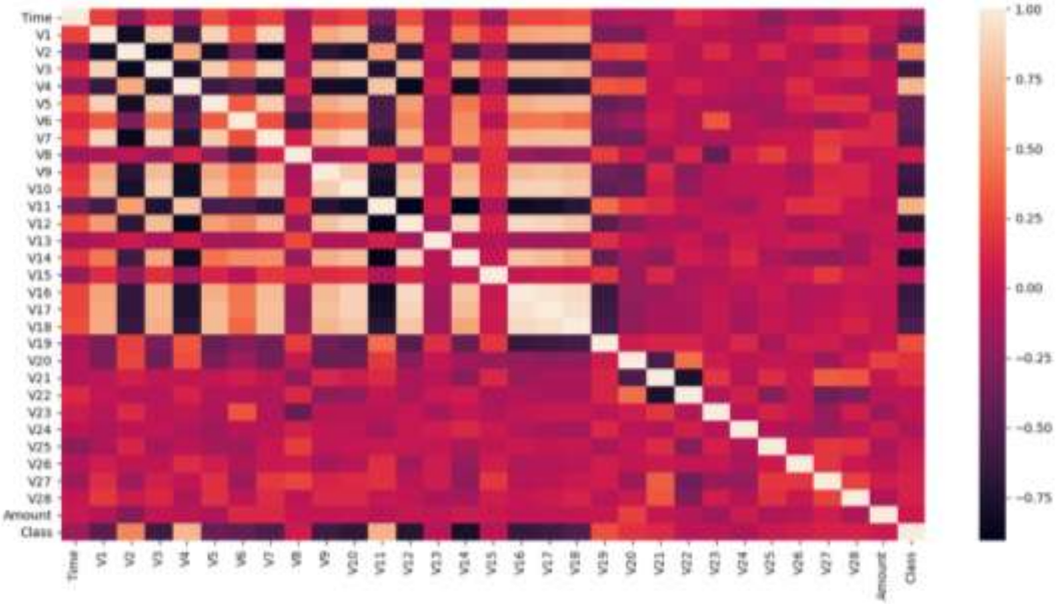
```
precision   recall  f1-score   support
0.0       0.98       0.98       0.98       632
1.0       0.98       0.98       0.98       568

accuracy
0.98       1200
macroavg       0.98       0.98       0.98       1200
weightedavg       0.98       0.98       0.98       1200
```

```
Confusion matrix:
[[618  14]
 [ 18 550]]

[[623   9]
 [ 19 549]]
```





XIII. CONCLUSION:

This method proves accurate in deducting fraudulent transaction and minimizing the number of false alert. If this algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions. And a series of anti-fraud strategies can be adopted to prevent banks from great losses and reduce risks.

The objective of the study was taken differently than the typical classification problems in that we had a variable misclassification cost. As the standard algorithms like random forest,Logistic regression, SVM,Decision tree are less acurate when compared to Xgboost and Light Gradient

REFERENCES:

[1]. M. HamdiOzcelik, EkremDuman, Mine Isik, TugbaCevik, Improving a credit card fraud detection system using genetic algorithm, International conference on Networking and information technology 2010.

[2]. Wen-Fang YU, Na Wang, Research on Credit Card Fraud Detection Model Based on Distance Sum, IEEE International Joint Conference on Artificial Intelligence 2009.

[3]. cliftonphua, vincent lee1, kate smith & ross gayler, A Comprehensive Survey of Data Mining-based Fraud Detection Research,2005.

[4]. Elio Lozano, Edgar Acu˜na, Parallel algorithms for distance-based and density-based outliers,2006.

[5]. Credit card fraud detection using hidden markov model – Abinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. majumdar

Websites:

[1]. http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/tcw2/report.html

[2]. http://www.kxcad.net/cae_MATLAB/toolbox/gads/f6691.html

[3]. http://java.sun.com/developer/onlineTraining/Programming/BasicJava1/front.html

[4]. http://www.easywayserver.com/blog/user-login-in-jsp/

[5]. http://www.faqs.org/patents/app/20100094765

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