IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

Credit Card Fraud Detection System

Amisha Jawanjal, Anushka Fulzele, Sejal Patil
Department of Information Technology
Priyadarshini College of Engineering, Nagpur, India

Abstract: *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 KNN, SVM and Logistic Regression, those models will be used on a credit card transaction dataset.*

Keywords: Credit Card Fraud Detection, Fraud Detection, Fraudulent Transactions, K Nearest Neighbors, Support Vector Machine, Logistic Regression, Naïve Bayes

I. INTRODUCTION

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 behavior 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. 1.2 Project goals. 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.

Data Preparation

The first figure bellow shows the structure of the dataset where all attributes are shown, with their type, in addition to glimpse of the variables within each attribute, as shown at the end of the figure the Class type is integer which I needed to change to factor and identify the 0 as Not Fraud and the 1 as Fraud to ease the process of creating the model and obtain visualizations

DOI: 10.48175/IJARSCT-26444

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

```
'data.frame':   284807 obs. of  31 variables:
 $ Time  : num  0 0 1 1 2 2 4 7 7 9 ...
 $ V1    : num  -1.36 1.192 -1.358 -0.966 -1.158 ...
 $ V2    : num  -0.0728 0.2662 -1.3402 -0.1852 0.8777 ...
 $ V3    : num  2.536 0.166 1.773 1.793 1.549 ...
 $ V4    : num  1.378 0.448 0.38 -0.863 0.403 ...
 $ V5    : num  -0.3383 0.06 -0.5032 -0.0103 -0.4072 ...
 $ V6    : num  0.4624 -0.0824 1.8005 1.2472 0.0959 ...
 $ V7    : num  0.2396 -0.0788 0.7915 0.2376 0.5929 ...
 $ V8    : num  0.0987 0.0851 0.2477 0.3774 -0.2705 ...
 $ V9    : num  0.364 -0.255 -1.515 -1.387 0.818 ...
 $ V10   : num  0.0908 -0.167 0.2076 -0.055 0.7531 ...
 $ V11   : num  -0.552 1.613 0.625 -0.226 -0.823 ...
 $ V12   : num  -0.6178 1.0652 0.0661 0.1782 0.5382 ...
 $ V13   : num  -0.991 0.489 0.717 0.508 1.346 ...
 $ V14   : num  -0.311 -0.144 -0.166 -0.288 -1.12 ...
 $ V15   : num  1.468 0.636 2.346 -0.631 0.175 ...
 $ V16   : num  -0.47 0.464 -2.89 -1.06 -0.451 ...
 $ V17   : num  0.208 -0.115 1.11 -0.684 -0.237 ...
 $ V18   : num  0.0258 -0.1834 -0.1214 1.9658 -0.0382 ...
 $ V19   : num  0.404 -0.146 -2.262 -1.233 0.803 ...
 $ V20   : num  0.2514 -0.0691 0.525 -0.208 0.4085 ...
 $ V21   : num  -0.01831 -0.22578 0.248 -0.1083 -0.00943 ...
 $ V22   : num  0.27784 -0.63867 0.77168 0.00527 0.79828 ...
 $ V23   : num  -0.11 0.101 0.909 -0.19 -0.137 ...
 $ V24   : num  0.0669 -0.3398 -0.6893 -1.1756 0.1413 ...
 $ V25   : num  0.129 0.167 -0.328 0.647 -0.206 ...
 $ V26   : num  -0.189 0.126 -0.139 -0.222 0.502 ...
 $ V27   : num  0.13356 -0.00898 -0.05535 0.06272 0.21942 ...
 $ V28   : num  -0.0211 0.0147 -0.0598 0.0615 0.2152 ...
 $ Amount: num  149.62 2.69 378.66 123.5 69.99 ...
 $ Class : int  0 0 0 0 0 0 0 0 0 0 ...
```

Figure 1 - Dataset Structure

Figure 1 - Dataset Structure

The second figure shows the distribution of the class, the red bar which contains 284,315 variables represents the non-fraudulent transactions, and the blue bar with 492 variables represents the fraudulent transactions.

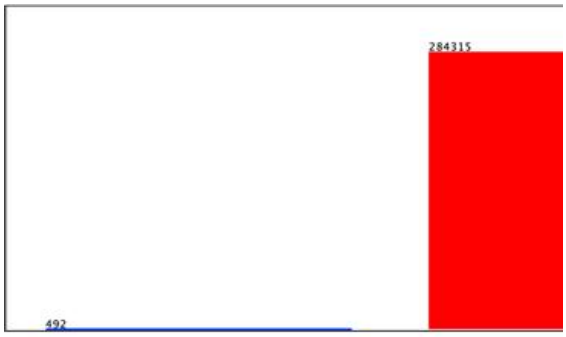

Figure 2 - Class Distribution

Figure 2 - Class Distribution

Correlation between attributes “Image from R” The correlations between all the of the attributes within the dataset are presented in the figure below.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

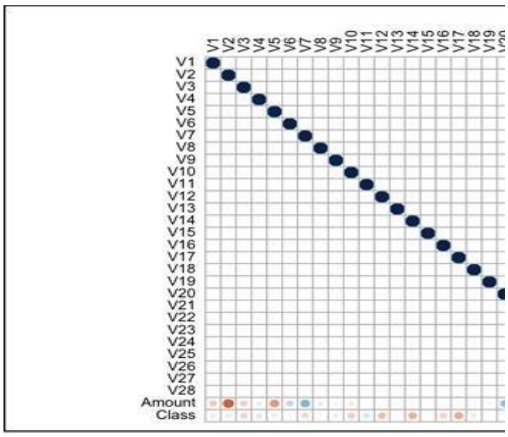

Figure 3 - Correla

Figure 3 - Correlations 12 4.1.2 Attribute with the most fraud Figure 4 below shows attribute 18 the attribute with the most credit card fraudulent transactions, the blue line represents the variable 1 which is the fraudulent transactions.

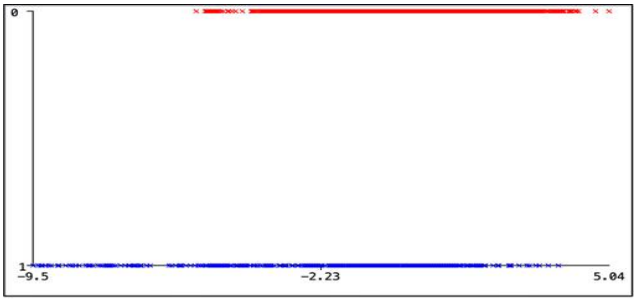

Figure 4 – Variable 18

Figure 4 – Variable 18 4.1.3 Attribute with the less fraud The figure below shows the variable that have the lowest number of fraudulent transactions, as mentioned earlier the blue line represents the fraudulent instances within the dataset.

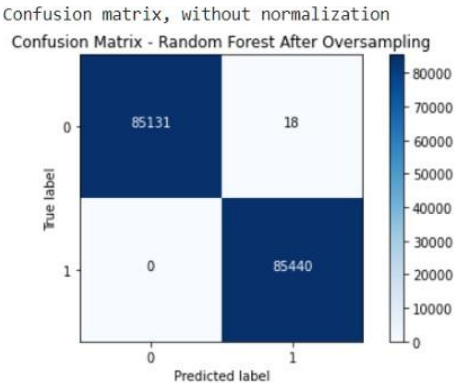

Figure 5 - Variable 28 13 4.2 Data Preprocessing As there are no NAs nor duplicated variables, the preparation of the dataset was simple the first alteration that was made to be able to open the dataset on Weka program is changing the type of the class attribute from Numeric to Class and identify the class as {1,0} using the program Sublime Text. Another alteration was made on the type as well on the R program to be able to create the model and the visualization.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

Data Modeling

After making sure that the data is ready to get modeled the four models were created using both Weka and R. the model SVM was created using Weka only, as for KNN, Logistic Regression and NaïveBayes they were created using R and Weka.

KNN The K-Nearest Neighbor algorithm (KNN) is a supervised ML technique that can be applied in both scenario instances, classification instances along with regression instances (Mahesh, 2020).To figure the best KNN model two Ks where used K=3 and K=7, both are presented with figures from both Weka and R.

II. RESULTS AND DISCUSSION

Easily, Random Forest model works better than Decision Trees. But if we observe our dataset suffers a serious problem of class imbalance. The genuine (not fraud) deals are further than 99 with the fraud deals constituting of 0.17. With similar kind of distribution, if we train our model without taking care of the imbalance issues, it predicts the label with more significance given to genuine deals (as there are more data about them) and hence obtains further fragility. The class imbalance problem can be resolved by reasonable number of ways. Over slice is one of them. Finally, after oversampling the confusion matrix and the accuracy scores are calculated.

Accuracy: 0.99989
Precision: 0.99979
Recall: 1.00000
F1-score: 0.99989

III. CONCLUSION

Credit card fraud is the biggest frauds that are being happened right now around the whole ground. This paper has explained how credit card frauds have been happening and we studied these frauds using a dataset that consists of transactions made in the real world. We saw how different machine learning algorithms are used to predict the fraud transactions on our dataset and we also addressed the class imbalance issue of our dataset and used oversampling to finally use Random Forest classifier that got a good accuracy score.

REFERENCES

[1], 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

[2] CLIFTON, PHUA1,, VINCENT, LEE1,, KATE, SMITH1, &, ROSS, GAYLER2, ", , A, Comprehensive, , Survey, of, , Data, Mining-based, , Fraud, Detection, Research", published, by, School, of, Business, Systems,, Faculty, of, Information, , Technology,, Monash, , University,, Wellington, , Road,, Clayton,, Victoria, 3800,, Australia,

[3] "Survey, Paper, on, Credit, Card, Fraud, Detection, by, Suman", Research, Scholar, ,GJUS&T, Hisar, , HCE,, Sonepat, , published, by, , International, Journal, of, Advanced, Research, in, Computer, Engineering, &, Technology, (IJARCET), Volume, 3, Issue, 3,, March, 2014

[4] "Research, on, Credit, Card, Fraud, Detection, Model, , Based, on, Distance, Sum, –, by, Wen-Fang, YU, and, Na, Wang", published, by, 2009, International, Joint, Conference, on, Artificial, Intelligence, \ The evaluation metrics for the Random Forest model (after oversampling) are as follows: As we can see the accuracy scores of the Random Forest model after the oversampling which is done to avoid the class imbalance issue, is quite good and better than the different algorithm approaches. So we can say that the [6]

[5] "Credit, Card, Fraud, Detection: ,A, , Realistic, , Modelling, , and, a, , Novel, Learning, Strategy", published, by, IEEE, TRANSACTIONS, ON, NEURAL, NETWORKS, AND, LEARNING, SYSTEMS,, VOL., 29,, NO., 8,, AUGUST, 2018