

Credit Card Fraud Detection in Transaction By Using Random Forest Algorithm With Data Sampling

Shraddha Jain
M.tech Scholar
Department of CSE
LNCT
Bhopal, India

Dr. Sunil Phulre
Associate. Professor
Department of CSE
LNCT
Bhopal, India

Dr. Bhupesh Gour
Professor
Department of CSE
LNCT
Bhopal, India

Abstract— There are various fraudulent activities detection techniques has implemented in credit card transactions have been kept in researcher minds to methods to develop models based on artificial intelligence , data mining, fuzzy logic and machine learning. Credit card fraud detection is significantly difficult, but also popular problem to solve. This paper we proposed an classification techniques which takes dataset of credit card transactions containing 284,807 transactions records which has a data imbalance problem which is the total number of a class of data (positive) is far less than the total number of another class of data (negative). In these we proposed various data sampling techniques applied on random forest algorithms and calculate the accuracy of the model based on various sampling techniques such as oversampling, undersampling and ROSE sampling.

Keywords: *data mining, random forest, decision tree, machine learning, classification, comparative analysis, credit card fraud.*

I.INTRODUCTION

Credit card fraud can be defined as “Unauthorized account activity by a person for which the account was not intended. Operationally, this is an event for which action can be taken to stop the abuse in progress and incorporate risk management practices to protect against similar actions in the future”.

Types of Frauds : Various types of frauds in this article include credit card frauds, telecommunication frauds, computer intrusions, Bankruptcy fraud, Theft fraud/counterfeit fraud, Application fraud, Behavioral fraud

Credit Card Fraud: Credit card fraud has been divided into two types:

- Offline fraud is committed by using a stolen physical card at call center or any other place .
- On-line fraud is committed via internet, phone, shopping, web, or in absence of card holder.

Brief of Dataset

Context It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Content The datasets contains 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.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original

features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are „Time“ and „Amount“. Feature „Time“ contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature „Amount“ is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature „Class“ is the response variable and it takes value 1 in case of fraud and 0 otherwise.

II. LITERATURE REVIEW

According to [1], during this paper they mainly specialise in mastercard fraud detection in world . Here the mastercard fraud detection is predicated on fraudulent transactions. Generally mastercard fraud activities can happen in both online and offline. But in today's world online fraud transaction activities are increasing day by day. So as to seek out the web fraud transactions various methods are utilized in existing system. In proposed system they use Random Forest Algorithm(RFA) for locating the fraudulent transactions and therefore the accuracy of these transactions. This algorithm is predicated on supervised learning algorithm where it uses decision trees for classification of the dataset. After classification of dataset a confusion matrix is obtained. The performance of Random Forest Algorithm is evaluated supported the confusion matrix. The results obtained from processing the dataset gives accuracy of about 90%.

According to [2], Credit card transactions became common place today then is that the frauds related to it. one among the foremost common routine to hold out fraud is to get the cardboard information illegally and use it to form online purchases. For mastercard companies and merchants, it's in-feasible to detect these fraudulent transactions among thousands of normal transactions. If sufficient data is collected and made available, machine learning algorithms are often applied to unravel this problem. during this work, popular supervised and unsupervised machine learning

algorithms are applied to detect mastercard frauds during a highly imbalanced dataset. it had been found that unsupervised machine learning algorithms can handle the skewness and provides best classification results.

In [3], Machine learning and data processing techniques are used extensively so as to detect mastercard frauds. However purchase behaviour and fraudster strategies may change over time. This phenomenon is known as dataset shift [1] or concept drift within the domain of fraud detection [2]. during this paper, they present a way to quantify day-by-day the dataset shift in our face-to-face mastercard transactions dataset (card holder located within the shop) . In practice, they classify the times against one another and measure the efficiency of the classification. The more efficient the classification, the more different the buying behaviour between two days, and the other way around . Therefore, we obtain a distance matrix characterizing the dataset shift. After an agglomerative clustering of the space matrix, we observe that the dataset shift pattern matches the calendar events for this point period (holidays, week-ends, etc). We then incorporate this dataset shift knowledge within the mastercard fraud detection task as a replacement feature. This results in a little improvement of the detection.

III PROBLEM DEFINITION

An effective fraud detection technique have faces these kind of difficulties to achieve best performance.

- **Imbalanced data:** The Dataset of credit card transaction has imbalanced nature.
- **Different misclassification importance:** There are misclassification errors have has an different importance.
- **Overlapping data:** In dataset there are many number of transactions are considered fraudulent means they are normal (false positive) and vice versa also. So it is very difficult to maintaining a low false positive and false negative rate.

IV PROPOSED WORK

In this, we proposed an machine learning classifier based on Random Forest classifier. We performed our experiments analysis on credit card dataset which is mostly used for classification.

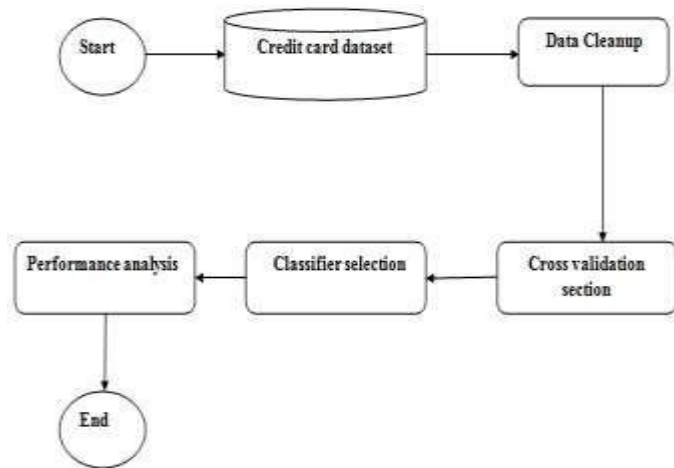


Figure 1. Proposed model

Algorithm

Input:

Data partition, D , which is a set of training tuples and their related class labels;

Attribute_list;

Attribute_selection_method, to determine the splitting criterion that “best” partitions.

Output: A decision tree.

Step 1- CREATING A ROOT NODE

1. Create a root node N
2. If tuples in D are all of the similar class, C then
3. Return N as a leaf node label with the class C ;
4. If attribute list is empty then
5. Return N as a leaf node label with the majority class in D

Step 2- ATTRIBUTE SELECTION

6. Apply attribute_selection_method(D , attribute_list) to discover the “best” splitting_criterion attribute;
7. Label node N with splitting_criterion;

8. Update the attribute_list

Step 3- SPLIT THE TREE

9. for each outcome j of splitting_criterion
 - //partition the tuples and produce subtrees for each partition
10. Based on splitting_criterion attribute
 - Split the tree into two part
12. attach a leaf labeled with the majority class D in node N :
13. else attach the node returned by Generate_decision_tree(D_j :attribute_list) to node N :
 - end for
14. return N ,

Random Forest Algorithm

for $S = 0, \dots, W$ do

$H_i \leftarrow$ Bootstrap sample from H

$T_i \leftarrow$ Construct tree using H_i

for node = 1, ..., No.Nodes do

node _{i} $\leftarrow$ choose random subset m of all features.

end for

end for

$X \leftarrow$ take the majority vote for all trees

Tools & Technology used:

1. R

V EXPERIMENTAL & RESULT ANALYSIS

For experiment we can take credit card data set contains 284807 observations of 31 variables and loaded into Rstudio . We first find the correlation between the

attributes.

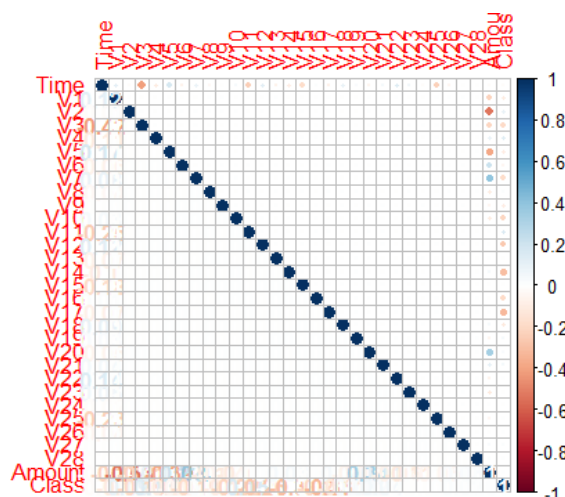


Figure 2. Correlation between the attributes

In these we explore the data and find the important attributes through which we can build the model. figure shows the data exploration through which we can get the important attributes which will help in classification.

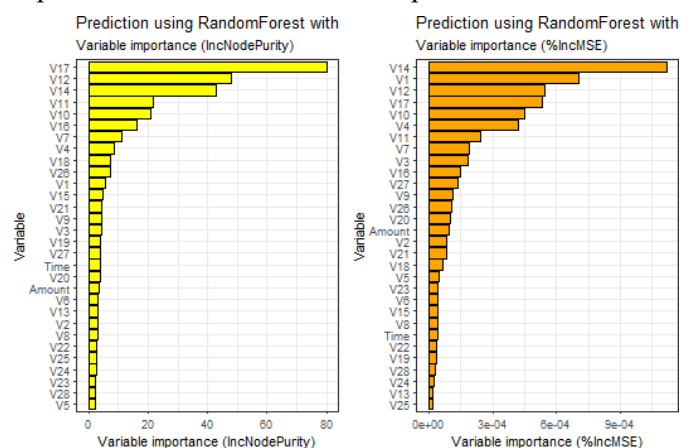


Figure 3. Data Exploration

Build Random Forest Model

Now we build the model based on random forest algorithms and train the model on the training dataset without and data sampling, we just split the data into training and testing dataset and build the model and the predicted performance of the model on test datasets are shown in figure 4.

```
> #plot roc
> threshold <- 0.7 #setting threshold
> threspos <- which.min(abs(roc$threshold-threshold))
> ggplot(roc, aes(roc$FPR, roc$TPR)) =
+   geom_line(color=rgb(0,0,1,alpha=0.3)) +
+   labs(title = sprintf("ROC"), x=lab("FPR"), y=lab("TPR")) +
+   geom_hline(yintercept=roc$threspos, "TPE", alpha=0.5, linetype="dashed") +
+   geom_vline(xintercept=roc$threspos, "FPR", alpha=0.5, linetype="dashed")
> #check the accuracy of this prediction
> accuracy.meas(dtTest$class, dtTest$predicted)

Call:
accuracy.meas(response = dtTest$class, predicted = dtTest$predicted)

examples are labelled as positive when predicted is greater than 0.5

precision: 0.954
recall: 0.817
F: 0.440
>
> #check the final accuracy of this model using ROC curve
> roc.curve(dtTest$class, dtTest$predicted, plotit = T)
Area under the curve (AUC): 0.979
>
```

Figure 4. Performance without data sampling

Apply Sampling techniques

Due to data imbalance problem we use various sampling techniques to resample the datasets and the re-sampling records of these techniques are:

```
> #oversampling
> data_balanced_over <- ovun.sample(class = ., data = dttraining, method = "over", n = 2*nrow(dttraining))$data
> table(data_balanced_over$class)
 0  1
199525 199703
> #undersampling
> data_balanced_under <- ovun.sample(class = ., data = dttraining, method = "under", n = 712, seed = 1)$data
> table(data_balanced_under$class)
 0  1
198 139
>
> #ROC (Random Over-sampling Examples)
> data_RF.rose <- ROSE(class = ., data = dttraining, seed = 1)$data
> table(data_RF.rose$class)
 0  1
99526 99618
>
```

Now we measure the performance of algorithms based on these data sampling techniques and the performance we get are shown in figure 5.

```
> #make predictions on unseen data - using random forest
> pred.RF.rose <- predict(RF.rose, dtTest)
> pred.RF.over <- predict(RF.over, dtTest)
> pred.RF.under <- predict(RF.under, dtTest)
>
> #AUC - Random Forest
>
> roc.curve(dtTest$class, pred.RF.rose, plotit = T)
Area under the curve (AUC): 0.969
> roc.curve(dtTest$class, pred.RF.over, plotit = T)
Area under the curve (AUC): 0.974
> roc.curve(dtTest$class, pred.RF.under, plotit = T)
Area under the curve (AUC): 0.982
>
```

Figure 5. Performance measure of Random Forest on Resampling data

Performance Measure

We used accuracy, which are derived using confusion matrix.

Table-1 Confusion Matrix

Actual Class	Predicted class	
	Class = Yes	Class = No
	Class = Yes	Class = No
Class = Yes	True Positive	False Negative
Class = No	False Positive	True Negative

Where

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

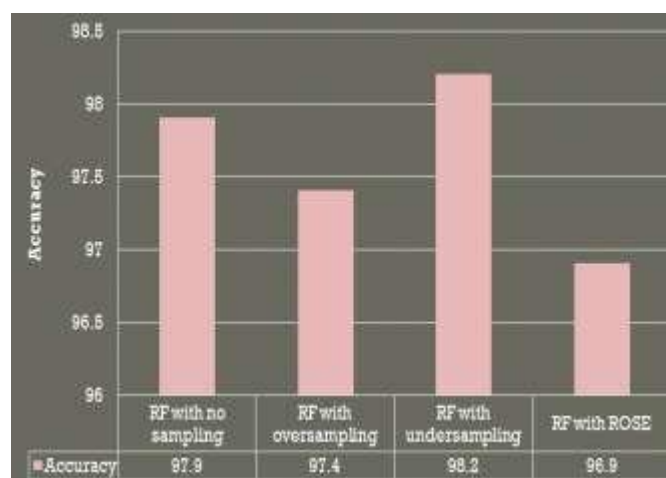


Figure 6. Comparison

VI CONCLUSION

As usage of credit cards become more and more common in every field of the daily life, credit card fraud has become much more rampant. As usage of credit cards become more and more common in every field of the daily life, credit card fraud has become much more rampant. In these we deal with data imbalance

problem which is the total number of a class of data (positive) is far less than the total number of another class of data (negative). In these we proposed various data sampling techniques applied on random forest algorithms and calculate the accuracy of the model based on various sampling techniques such as oversampling, undersampling and ROSE sampling.

REFERENCES

[01] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare, " Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis" in 2017 IEEE .

[02] Bolton, R. J., and Hand, D. J. (2002). Statistical fraud detection: a review. Statistical Science, 17(3), 235-249.

[03] V.Dheepa and Dr. R.Dhanapal, et al. "Analysis of Credit Card Fraud Detection Methods", "IJRTE", Vol 2, No. 3, November 2009

[04] P Chan, W Fan, A Prodromidis & S Stolfo. 1999. Distributed data mining in credit card fraud detection, IEEE Intelligent Systems, 14(6): 67-74.

[05] N.Sivakumar, Dr.R.Balasubramanian "Credit Card Fraud Detection: Incidents, Challenges And Solutions" in IJARCSA.

[06] Srivastava, A., Kundu, A., Sural, S., and Majumdar, A. (2008). Credit card fraud detection using hidden markov model. IEEE Transactions on Dependable and Secure Computing, 5(1), 37-48.

[07] Ghosh, S. and Reilly D. L. 1994. Credit card fraud detection with a neural network. In Proceedings of the 27th Hawaii International Conference on system Science.

[08] Ju, W. H., and Vardi, Y. (2001). A hybrid high-order markov chain model for computer intrusion detection. *Journal of Computational and Graphical Statistics*, 10(2), 277-295.

[09] Chen, R.-C., Luo, S.-T., Liang, X. and Lee, V. C. S.. Personalized approach based on SVM and ANN for detecting credit card fraud. In *Proceedings of the IEEE 2005*.