REPORT OF CREDIT CARD FRAUD DETECTION

Abstract:

Credit cards usage has increased specially after E-Commerce and online shopping has increased, with the usage increasing also the fraud transactions through credit cards have increased. In this project we attempted to model a data-set using machine learning algorithm for credit card fraud detection. We have taken a csv file which has credit-card based transactions from the history, modelled it using Naïve Bayes Algorithm and Random Forest algorithm. Based on the history of transactions the model should be able to predict if the new transaction is normal or fraudulent. The goal of our project is to find a model with accuracy as high as hundred percent, the higher the accuracy it indicates the lesser the chance of misclassifying the data.

Introduction:
" Credit card fraud occurs when an unauthorized person gains access to genuine persons information and uses it to make purchases "

Popular Ways Credit Card Fraud can happen:

- Lost or stolen credit cards
- Calling about fake prizes or wire transfers
- Skimming your credit card, such as at a gas station

In the age of everybody being digitized the necessity to have a good credit card fraud detection system is necessary as every business now involves and encourages online purchase. We will attempt to build a model which will be monitoring or let's say analyzing the behavior of various users and based on their behavior we predict or detect if the other user's behavior/ transaction was undesirable. Before we worked on building the model, our work was to study on all the algorithms and technologies available and the algorithms which will suite credit card fraud detection best. The dataset has transactions which can be classified as normal and fraudulent data, also the other features are analyzed to see if there is a co-relation between them and the type of the transaction. For example, for a particular feature V5, we check if there is a pattern in V5 for it to fall under normal or fraudulent.

About the dataset:

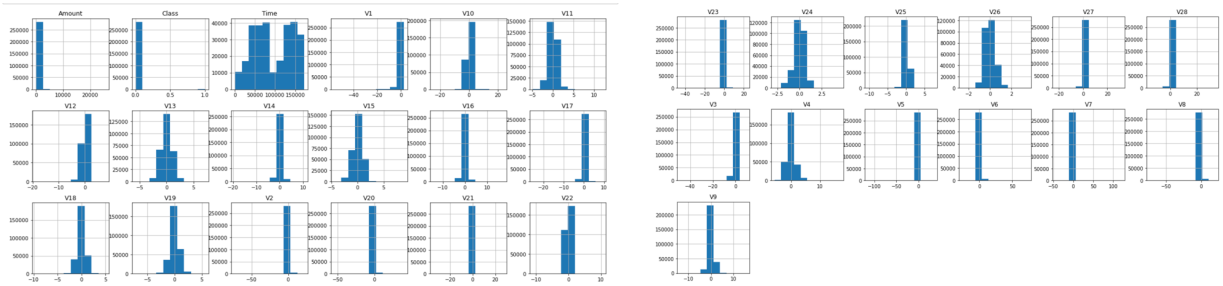
- Source: Kaggle Community
- URL: https://www.kaggle.com/mlg-ulb/creditcardfraud
- The datasets contains transactions made by credit cards in September 2013 by European cardholders.

Dataset has 492 frauds and 284315 normal

- 99.83% transactions are genuine, and 0.17% transactions are fraud.
- The dataset is imbalanced. So, there is requirement of balancing the dataset by taking 500 transactions randomly from the 284,315 transactions to make the overall processed dataset as 50-50 of fraud and genuine transactions for better performance of ML Algorithms.

Our data-set has features labelled as v1-v28 in order to protect sensitive data as this is financial data. The dataset has 31 columns in total, and v1-v28 are 28 of them. Class 0 represents a normal transaction, which means a genuine transaction, Class 1 represents a fraudulent transaction. We have plotted a heatmap in order to get a colored representation of all the data in the dataset to understand the feature co-relation. Next we attempt to fit the data into model by pre-processing it, once the data looks like it can fit into the model we implement the algorithms and check for accuracy. We didn't directly infect the dataset with new transactions to check if the prediction changes all we did was implement different algorithms and find one with higher accuracy. As our goal was to get an algorithm with higher accuracy for credit card transaction fraud dataset.

Plotting the dataset and exploratory data analysis:



From the dataset we can observe that frauds are smaller than regular, also when we observe the Time mark: regular transactions have dropped around the 90,000th second mark and has increased around 110,000th second. Therefore, we can assume that night people make less purchases. Also, we can see that fraud transactions are more during the 100,000 mark indicating that more fraud transactions are likely to happen at night as there will be less surveillance. Next, based on amount we see no significant pattern, both normal and fraudulent transactions were having small amount. As v10-v28 transactions have no label finding a co-relation will help understand them. From the correlation matrix we understand that there is no strong correlation, probably because this dataset is highly unbalanced.

Machine learning for analysis and prediction:

Before we implement an algorithm, we should preprocess features, split the dataset and also handle with unbalanced dataset. In order to have a balanced split of normal and fraud in test and train data we use stratify.
About the Algorithms:

- Naïve Bayes Algorithm: It is a classification technique which is based on Bayes Theorem, in this technique we believe that there is independence among all the predictors, therefore, we assume that one feature is not related to other in this prediction. Therefore, each feature makes an independent and equal contribution to the result that we get. Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

- Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.

- Random Forest Algorithm: This machine learning algorithm can be perceived as a flexible algorithm, it has many decision trees which uses bootstrapping, this algorithm has feature randomness when they build each decision tree, they try to create uncorrelated forest of trees whose prediction will be more accurate than individual tree. It is an ensemble algorithm in nature. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

Random forests are frequently used as "black box" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration

Pseudo Code of two algorithms:

Naïve Bayes Algorithm:

1. Load the dataset
2. Pre-process the dataset
3. Split the dataset into test data and train data
4. Construct the Naïve Bayes classifier for the model by importing from sklearn
5. Get the accuracy of the model

Random Forest Algorithm:

We can say that the Random Forest Algorithm has two stages: one is the creation and the next is the prediction based from the classifier created.

1. Randomly select “k” features from a total “m” features (always make sure k <<m)
2. Once we get the k features, node “d” is calculated using the best split point
3. Further, split the “d” node into daughter nodes using best possible split
4. Continue randomly selecting k features, creating a node based on split and getting daughter nodes till the number of nodes becomes “1”
5. Based on all the steps mentioned above we will get a bunch of decision trees which create a random forest

Advantages: avoids over-fitting problem, also features which impact the result can be found using this algorithm.

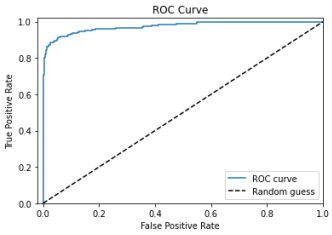
- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

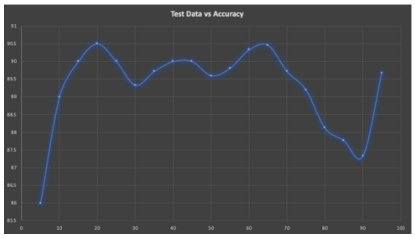
- Random forests are frequently used as "black box" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration

Inference from the code:

Naïve Bayes Algorithm results:

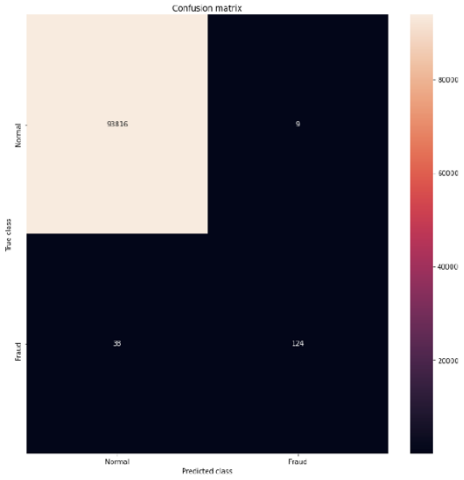
Naïve Bayes being a simple probability calculation without any co-efficient, or optimization by fitting this model learns quickly. We can improve its performance by working around default threshold, most of the fraud cases were predicted right with this model.





Random Forest:

As the features are un-labelled, random forest will rank the variables importance in a regression or classification problem. Random forest is not biased and stable, even if a new datapoint is introduced to the algorithm, it does not get affected. As this dataset has both categorical and numerical features random forest algorithm works well.



Random Forest: 47
0.9994999308414994

	precision	recall	f1-score	support
0	1.00	1.00	1.00	93825
1	0.93	0.77	0.84	162
accuracy			1.00	93987
macro avg	0.97	0.88	0.92	93987
weighted avg	1.00	1.00	1.00	93987

The test split data is for 33% and the confusion matrix obtained is shown on the left.
Here, we can see that are total of 47 data points which are classified wrong from a total transactions of 93,987 (33% of 284,807 transactions).

Conclusion:

As credit card transaction fraud is something which has been increasing over time, having a machine learning prediction is important. Although, our dataset only gives two days data, I think that. a dataset of this kind is a good place to start with as we know the kind of variable we have access to and the model which suits this data best. Future enhancement will be to get higher precision and recall, also getting data with more data will make the data training model train better.

Contribution of each participant in the Project:

Madhav Kaza has described in this paper as to what Naïve Bayes Algorithm is and applied this algorithm on the credit-card fraud detection data-set taken from Kaggle and was able to achieve accuracy as 90%.

Priya Balaji has described in this paper as to what Random Forest Classifier is and applied this algorithm on the credit card detection data-set and was able to achieve accuracy as high as 99%.

Vijaya Nannapaneni in this paper has described about the fraud detection algorithm process, worked on getting the dataset, performed data pre-processing, and worked with comparing the results of both the algorithms and performed visualization step in the code.

REFERENCES

[1] “Credit Card Fraud Detection Based on Transaction Behaviour -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

[5] “Credit Card Fraud Detection through Parenclitic Network Analysis- By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity V olume 2018, Article ID 5764370, 9 pages

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

[7] “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak “Plastic Card Fraud Detection using Peer Group Analysis” Springer, Issue 2008.