Credit Card Fraud Detection – MIS-64060

By S.S. Ananth kumar

Abstract:

Due to a rapid advancement in the electronic commerce technology, the use of credit cards has dramatically increased. Since credit card is the most popular mode of payment, the number of fraud cases associated with it is also rising. Thus, in order to stop these frauds we need a powerful fraud detection system that detects it in an accurate manner. In this paper, I have explained the concept of frauds related to credit cards. I have implemented different machine learning algorithms on an imbalanced dataset such as logistic regression, naïve bayes, random forest with ensemble classifiers using boosting technique. An extensive review is done on the existing and proposed models for credit card fraud detection and has done a comparative study on these techniques. So Different classification models are applied to the data and the model performance is evaluated on the basis of quantitative measurements such as accuracy, precision, recall, f1 score, support, confusion matrix. The conclusion of our study

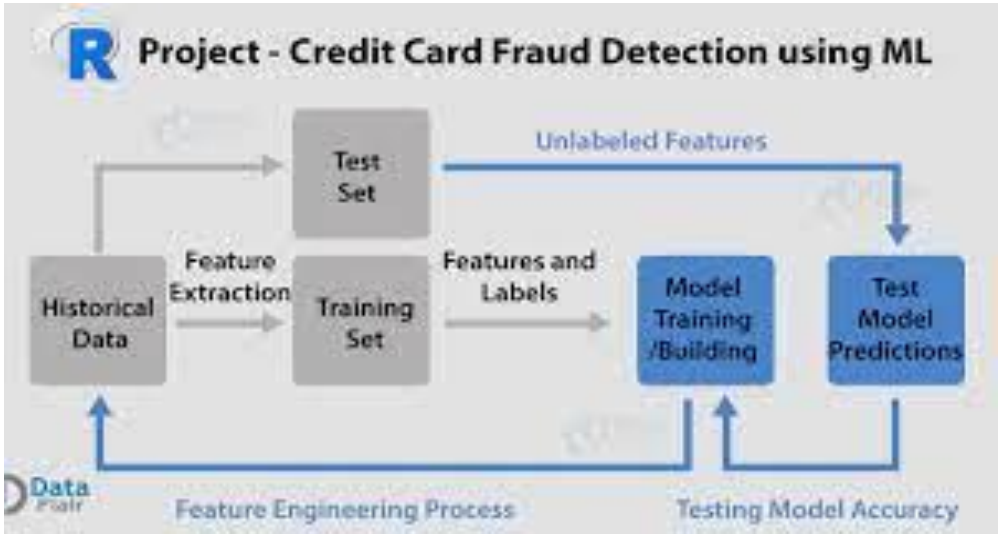
Keywords: Fraud detection, Credit card, Logistic regression, Naïve bayes Algorithms

Introduction:

This project is to propose a credit card fraud detection system using supervised learning algorithm. Supervised algorithms are evolutionary algorithms which aim at obtaining better solutions as time progresses. Credit card is the most popular mode of payment. As the number of credit card users is rising world-wide, the identity theft is increased, and frauds are also increasing. In the virtual card purchase, only the card information is required such as card number, expiration date, secure code, etc. Such purchases are normally done on the Internet or over telephone. To commit fraud in these types of purchases, a person simply needs to know the card details. The mode of payment for online purchase is mostly done by credit card. The details of credit card should be kept private. To secure credit card privacy, the details should not be leaked. Different ways to steal credit card details are phishing websites, steal/lost credit cards, counterfeit credit cards, theft of card details, intercepted cards etc. For security purpose, the above things should be avoided. In online fraud, the transaction is made remotely and only the card's details are needed. A manual signature, a PIN or a card imprint are not required at the purchase time. In most of the cases the genuine cardholder is not aware that someone else has seen or stolen his/her card information. The simple way to detect this type of fraud is to analyze the spending patterns on every card and to figure out any variation to the "usual" spending patterns. Fraud detection by analyzing the existing data purchase of cardholder is the best way to reduce the rate of successful credit card frauds. As the data sets are not available and also the results are not disclosed to the public. The fraud cases should be detected from the available data sets known as the logged data and user

behavior. At present, fraud detection has been implemented by a number of methods such as data mining, statistics, and artificial intelligence

Methodology:



Dataset:

In this paper credit card fraud detection dataset was used, which can be downloaded from Kaggle. This dataset contains transactions done by the customer. The dataset contains 122 features. Since some of the input variables Contains financial information, the PCA transformation of these input variables were performed to keep these data anonymous. Feature "Total_Amount" is the amount of the transactions made by credit card. Feature "Target" represents the label and takes only 2 values: value 1 in case fraud transaction and 0 otherwise. The dataset also contains other information such as cardholder's Annual Income, Total credit debt, family background, employment status, marital status, Children, estimated value of the property owned, credit score and other different information which the credit card company has collected. However, we will not be using all the features as some of the features as no correlation or influence in fraud detection.

Data Exploration:

In this section of the fraud detection, we will explore the data that is contained in the credit card data. Data frame. We will proceed by load the data and displaying the credit card data using the head () function and the tail() function. We will then proceed to explore the other components of this data frame. For the sake of reduced computation time, I have taken 10,000 rows instead of the complete

dataset. The same issue with rebalance in also observed in subsetted data. We also check the structure of the data and numerical data distribution and analyze if the data should be normalized or not. We also need to observe the trends of the data to see if there is pattern in customer income and credit details. When exploring further we have observed that the Target variable is heavily imbalanced, majority of the class in '0' which is not fraud and minority of the class '1'. Due to which when applied algorithm the prediction will be biased or skewed towards the majority class which is not a good prediction. We will resample the data which is covered in the data manipulation section.

Data manipulation:

In this section of the R project, we will scale our data using the scale()function. We will apply this to the amount component of our credit card data amount. Scaling is also known as feature standardization. With the help of scaling, the data is structured according to a specified range. Therefore, there are no extreme values in our dataset that might interfere with the functioning of our model and apply pca for our data. In this dataset there are lot of missing and blank value which were imputed using mean value of the column.

As for categorical variable which are important in prediction, we have applied transformation which will convert the features in the column to levels as 0,1,2,3. As no of levels increases the columns has more input features.

Divide the dataset:

The dataset is divided into trained data set and test data set. 80% of the data set is under training and the remaining 20% is under testing .Here we are using some supervised machine learning algorithms. by using train data train the model and using test data do the predictions and find the accuracy's of each model and select best model by accuracy and displaying graphs.

Logistic Regression accuracy=91.2

Navie bayes accuracy=92.35

Sampling:

When we have a imbalanced data to balance the data we are applying some re sampling methods like SMOTE, ROSE, UP sampling, DOWN sampling. The sampling techniques will balance the target volume equally so that the prediction is not biased. In this project I have considered Up sampling which will increase the minority class to the same level as the majority class i.e if the majority class is 90 and minority class is 10, by applying the up sampling the minority class will have same observation as majority. This technique eliminates the majority and minority class by making them 50 % of either class.

We have applied the sampling only for the target column in training data, as we are trying to predict the target variable and the prediction will not be biased.

Model building:

After the resampling is applied for the training data, a logistic regression is applied to check the model accuracy or to check if the model is built with enough variables or not.

If the initial model is accurate enough, we can go ahead and continue to apply algorithm. For this project I have considered naïve bayes algorithm as it fast and easy classification algorithm. By applying naïve bayes we can identify if there are any miss classification for the fraudulent customers and correctly identify which customer might be a defaulter based on the features selected. After applying the algorithm on the training data, we get model for prediction.

We can predict the customers using the training model and the test data we get the initial model prediction. We must check the performance measure by evaluating the model accuracy, kappa value, sensitivity and specificity. If the kappa value is > 0.75 and the sensitivity and specificity are less than the cut off. If the performance measures are not up to the cut off then we will have to evaluate the feature selection and importance and re-run the who algorithm. If the performance measures have reached the cut off value, we can plot ROC and AUC of the model. The model is set to accurate enough if the AUC value is > 0.65 and the ROC curve is tilted towards the left side of the graph.

On the other hand, we can check the confusion matrix for the miss classification. There is one more measure to check if the prediction is un-biased or not, by checking the no information rate. If the no information rate value is at 0.50 then the model is predicting based on the feature selection and is unbiased. However, in this project few of the performance measures have been met except for Kappa, which did not reach the cut off 0.75, this might be due to the feature selection. Since the data set is huge, there might be a good chance that few of the important features has not been selected. But the model is not biased and the predicting potential defaulters up to certain number of customers. Along with the model performance measure a general hypothetical evaluation of other performance measures.

Modeling Approach

Standard machine learning algorithms struggle with accuracy on imbalanced data for the following reasons:

1. Algorithms struggle with accuracy because of the unequal distribution in dependent variable. This causes the performance of existing classifiers to get biased towards majority class.
2. The algorithms are accuracy driven i.e. they aim to minimize the overall error to which the minority class contributes very little.

3. Algorithms assume that the data set has balanced class distributions.
4. They also assume that errors obtained from different classes have same cost

The methods to deal with this problem are widely known as 'Sampling Methods'. Generally, these methods aim to modify an imbalanced data into balanced distribution using some mechanism. The modification occurs by altering the size of original data set and provide the same proportion of balance.

These methods have acquired higher importance after many research have proved that balanced data results in improved overall classification performance compared to an imbalanced data set. Hence, it's important to learn them.

Below are the methods used here to treat the imbalanced dataset:

- Undersampling
- Oversampling
- Synthetic Data Generation

Undersampling

This method reduces the number of observations from majority class to make the data set balanced. This method is best to use when the data set is huge and reducing the number of training samples helps to improve run time and storage troubles.

Undersampling methods are of 2 types: Random and Informative.

Random undersampling method randomly chooses observations from majority class which are eliminated until the data set gets balanced. Informative undersampling follows a pre-specified selection criterion to remove the observations from majority class.

A possible problem with this method is that removing observations may cause the training data to lose important information pertaining to majority class.

Oversampling

This method works with minority class. It replicates the observations from minority class to balance the data. It is also known as upsampling. Similar to undersampling, this method also can be divided into two types: Random Oversampling and Informative Oversampling.

Random oversampling balances the data by randomly oversampling the minority class. Informative oversampling uses a pre-specified criterion and synthetically generates minority class observations.

An advantage of using this method is that it leads to no information loss. The disadvantage of using this method is that, since oversampling simply adds replicated observations in original data set, it ends up adding multiple observations of several types, thus leading to overfitting.

Synthetic Data Generation (SMOTE and ROSE)

In simple words, instead of replicating and adding the observations from the minority class, it overcome imbalances by generates artificial data. It is also a type of oversampling technique.

In regards to synthetic data generation, synthetic minority oversampling technique (SMOTE) is a powerful and widely used method. SMOTE algorithm draws artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbors in the feature space. ROSE

(random over-sampling examples) uses smoothed bootstrapping to draw artificial samples from the feature space neighbourhood around the minority class.

It is important to note that sampling techniques should only be applied to the training set and not the testing set.

This modeling approach will involve training a single classifier on the train set with TARGET class imbalance suitably altered using each of the techniques above. Depending on which technique yields the best roc-auc score on a holdout test set. we will build subsequent models using that chosen technique.

Evaluation:

There are a variety of measures for various algorithms and these measures have been developed to evaluate very different things .So it should be criteria for evaluation of various proposed method. False Positive(FP),False Negative(FN),True Positive(TP),True Negative(TN) and the relation between them are quantities which usually adopted by credit card fraud detection researchers to compare the accuracy of different approaches.

True Positive(TP): The true positive rate represents the portion of the fraudulent transactions correctly being classified as fraudulent transactions.

True positive=TP/TP+FN

TrueNegative(TN):The true negative rate represents the portion of the normal transactions correctly being classified as normal transactions.

True negative=TN/TN+FP

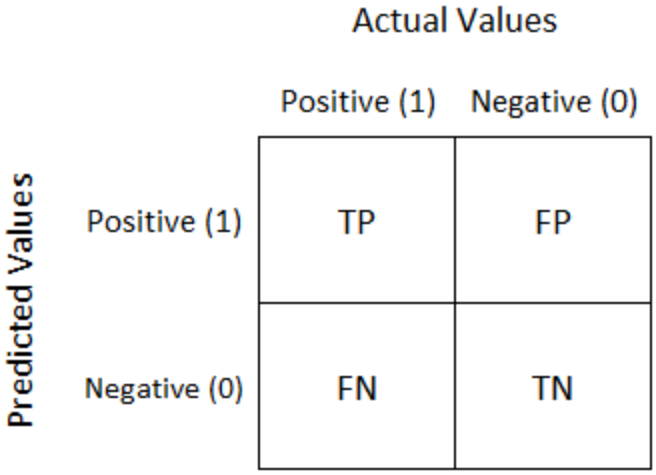
False Positive (FP): The false positive rate indicates the portion of the non-fraudulent transactions wrongly being classified as fraudulent transactions.

False positive=FP/FP+TN

False Negative (FN): The false negative rate indicates the portion of the non-fraudulent transactions wrongly being classified as normal transactions.

False negative=FN/FN+TP

Confusion matrix: The confusion matrix provides more insight into not only the performance of a predictive model, but also which classes are being predicted correctly, which incorrectly, and what type of errors are being made. The simplest confusion matrix is for a two-class classification problem, with negative and positive classes.



Accuracy: Accuracy is the percentage of correctly classified instances. It is one of the most widely used classification performance metrics.

Accuracy=Number of correct predictions/ Total Number of predictions

Or for binary classification models. The accuracy can be defined as:

Accuracy= TP+TN/ TP+TN+FP+FN

Precision : Precision is the number of classified Positive or fraudulent instances that actually are positive instances.

Precision = TP/(TP+FP)

Recall: Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made. Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions.

Recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

Recall = TP / (TP + FN)

F1 score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

CONCLUSION:

I have studied applications of machine learning like Naïve Bayes, Logistic regression with boosting and shows that it proves accurate in deducting fraudulent transaction and minimizing the number of false alerts. Supervised learning algorithms are novel one in this literature in terms of application domain. If these algorithms are 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. Precision, recall f1-score and accuracy are used to evaluate the performance for the proposed system.by comparing both the models naive bayes is the best model.

Git Hub Code Link:

https://github.com/Ananth2021/sseetham_64060.git