"Detection of fake electronic transactions"

Prajakta Patwari[1], MSc., Umadevi Gopalakrishnan[1], MSc, Hemanth Kumar[1], MSc., Nabeel Shahid[1], MSc.

[1]University of South Florida, USA

patwarip@mail.usf.edu, nabeelshahid@mailusf.edu, umadevi@mailusf.edu,hemanth7226@gmail.com

Abstract

Fake electronic transactions is prevalent in modern day society. But it is obvious that the number of fraudulent electronic transaction cases are constantly increasing in spite of the chip cards worldwide integration and existing protection systems. This is why the problem of fraud detection is very important now. In this paper the general description of fake electronic transactions and comparisons between models based on AUC(Area Under Curve) and RMSE (Root Mean Square Error) values are given. Later, we used peer group clustering in which we cluster the data and then apply different modelling techniques to improve our results. After considering various models, we concluded that the Xgboost model gives the best results.

1 Introduction

In Modern day the fraud is one of the major causes of great financial losses, not only for merchants, individual clients are also affected. Many researches have used machine learning algorithms to detect fraudulent transactions. As per the survey from 2014 to 2018 it was found that losses due to counterfeit amounted to three billion U.S. dollars in 2014 and they are projected to decrease to 1.8 billion U.S. dollars by the end of 2018. Therefore, detecting process in optimal way is a time consuming process and mostly is done offline in static operation. The objective of this paper underlying the content is to overcome the above mentioned issue in an efficient way by using R Programming Language. Thus using this project we can analyze and visualize large amount of data.

Our overall statistical modelling approach contains following steps:-

- Data Exploration - explores data for various findings.
- Cleaning Data- removes inconsistencies in the source datasets.
- Data Pre Processing- task-relevant data are retrieved from the source.
- Data Manipulation- data have to be transformed to appropriate task specific form.
- Modeling- appropriate algorithms extract data patterns based on different measures.
- Improvement- Improving the prediction results for better accuracy.
- Knowledge presentation – visualization and knowledge representation to users.

S.no	Country	Cardholders Affected(overall)	Last 5 years
1	United States	42%	37%
2	Mexico	44%	37%
3	United Arab Emirates	36%	33%
4	United Kingdom	34%	31%
5	China	36%	27%
6	India	37%	27%
7	Canada	25%	19%
8	France	20%	18%

Table 1: Cardholders Impacted by Fraud in different Country

The standard data mining methodology is adopted in this research. Table 1 shows the researches have done by researcher based their country; we can see that United State has most part, based on Table 1, we show that the United State suffer for fraud

problem with overall 42% in last years , it means US has good approach to manage this problem.

2 Methods:

In this section we will discuss the different statistical models - logistic regression, random forest, XGBoost, peer group clustering which are used for predicting fake electronic transaction.

2.1 Balancing data using SMOTE

Initially, we considered under-sampling as a technique to balance the dataset of positive and negative instances, it's been taken the dataset with majority label '0' and reduced the dataset in order to match with the minority label '1' This reduction in the number of instances reduced a considerable amount of information that is relevant to improve the fit of the prediction. Similarly, using oversampling, there is a risk of overfitting.

Hence, we used SMOTE (Synthetic Minority Oversampling Technique) hybrid methods that combine under-sampling with the generation of additional data, which will help us to improve the accuracy of predictions. SMOTE(Synthetic Minority Over-sampling Technique) is an approach which is described as the construction of classifiers from imbalanced datasets. A dataset is imbalanced if the classification categories are not approximately equally represented. SMOTE synthesises new minority instances between existing (real) minority instances. The given dataset consists of 90000 observations, out of which 89884 observation have value as '0' and remaining 116 observation have value as '1'.

2.2 Correlation

The correlation between different variables is shown in above figure 1. As it can be observed that there are no highly correlated variables. Moreover, there is no correlation between the response variable 'Fake' and all other variables individually. The variables X1 to X28 are scaled and their names are not provided due to confidential purpose. Hence, none of them can be eliminated based on domain knowledge. However, the time and amount variables are not scaled. The 'Time' variable does not indicate the actual time of the transaction and is more of a list of the data in chronological order. So we assume that the 'Time' feature has little or no significance in classifying a fraud transaction. Therefore, we eliminate this column from further analysis.

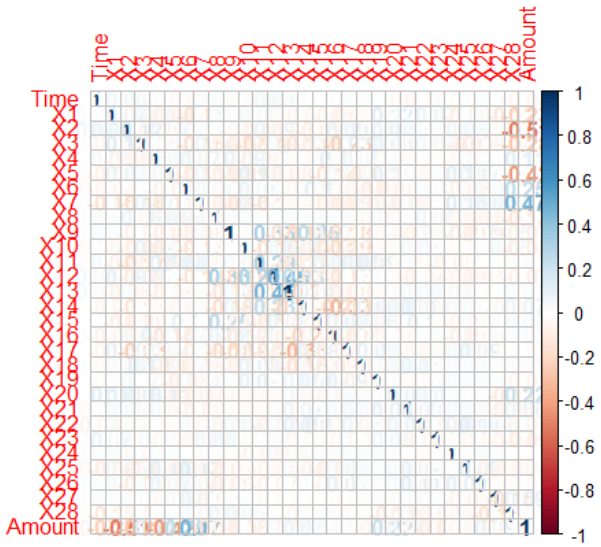

Figure 1:

2.3 Variable Importance Plot:

Variable importance plot provides a list of the most significant variables in descending order by a mean decrease in Gini. The top variables contribute more to the model than the bottom ones and also have high predictive power in classifying default and

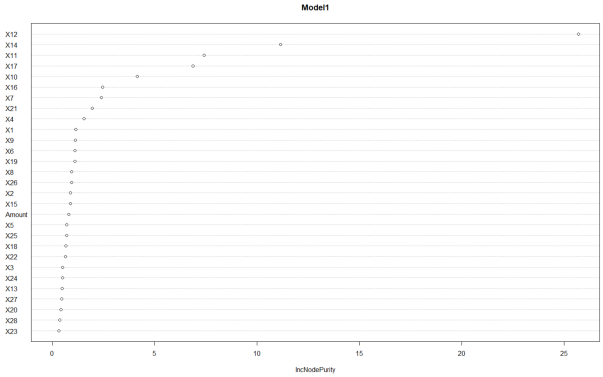

Figure 2: Variable Importance plot

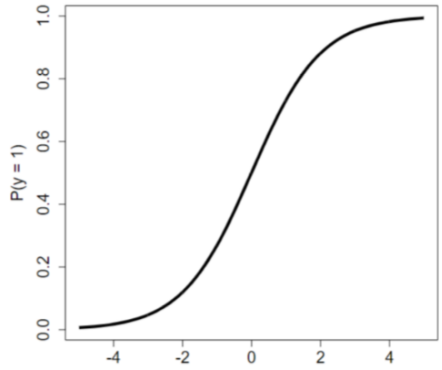

Figure 3: Logistic Regression Graph.

non-default customers.

2.4 Logistic Regression:

Logistic regression is a well-established statistical method for predicting binomial or multinomial outcomes. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic Function:

$$P(y = 1) = 1/(1 + e^{-}\beta_0 + \beta_1 x1 + \beta_2 x2 + ... + \beta_k xk) \quad (1)$$

where, x1, x2..., xk are independent variables.
P(y) = probability prediction.

Although logistic regression is a classification algorithm, we predict probabilities which must be transformed into a binary values (0 or 1) in order to actually make a probability prediction.

2.5 Random Forest:

Random Forest Classifier is supervised classification/regression ensemble algorithm. Ensembled algorithms are those which combines more than one algorithm of same or different kind for classifying objects. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. In a nutshell if we have more trees in the forest robust the forest looks like. In the same way in the random forest algorithm the higher the number of trees in the forest gives the high accuracy results.

There are *four* advantages to illustrate why we use Random Forest algorithm. One is that it can be used for both classification and regression tasks. Over fitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't over fit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

2.6 XGboost

XGBoost is an algorithm that has recently been dominating applied machine learning for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. This algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. the key features of this library rely on model performance and execution speed.

The implementation of XGBoost offers several advanced features for model tuning, computing environments and algorithm enhancement. It is capable of performing the three main forms of gradient boosting (Gradient Boosting (GB), Stochastic GB and Regularized GB) and it is robust enough to support fine tuning and addition of regularization parameters.

2.7 Clustering

Clustering is the process of grouping similar entities together. The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. The goal of cluster analysis is to sort cases (people, things, events) into groups, or clusters, so that the degree of association/relationship is strong between members of the same cluster and weak between members of different clusters. There are many algorithms developed to implement this technique but for this post, let's stick the most popular and widely used algorithms in machine learning.

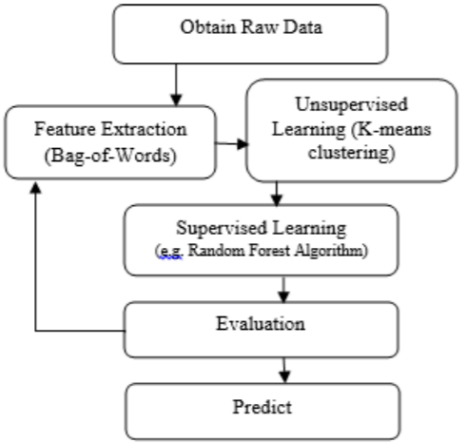

Figure 4: Cluster then predict Model

2.7.1 K- Mean Clustering

The k-means method will produce exactly k different clusters of greatest possible distinction. The algorithm will start with k random clusters, and then move objects between those clusters with the goal to minimize variability within clusters and maximize variability between clusters.

Given k, the k-means algorithm is implemented in four steps:

1. Partition objects into k nonempty subsets
2. Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., mean point, of the cluster)
3. Assign each object to the cluster with the nearest seed point
4. Go back to Step 2, stop when no more new assignment

2.7.2 Hierarchical Clustering

Unlike K-mean clustering Hierarchical clustering starts by assigning all data points as their own cluster. As the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all

the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.
Two important things that you should know about hierarchical clustering are:

- This algorithm has been implemented above using bottom up approach. It is also possible to follow top-down approach starting with all data points assigned in the same cluster and recursively performing splits till each data point is assigned a separate cluster.
- The decision of merging two clusters is taken on the basis of closeness of these clusters. There are multiple metrics for deciding the closeness of two clusters one of them is Euclidean distance. For instance, distance between points i and j is given by Euclidean distance.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (2)$$

3 Important terms:

There are different parameters which can be used to define the model performance:

3.1 Accuracy

The Accuracy metric reports the number of transactions correctly classified, compared to the total number of them. Given a set of transactions ˆT to be classified, it is calculated as shown in Equation, where $|T|$ stands for the total number of transactions, and $|T(+)|$ for the number of those correctly classified.

In this paper the accuracy of the model has been found out by the confusion matrix by the formula given below:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (3)$$

3.2 RMSE

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors).The root-mean-squared error (RMSE) is a measure of how well your model performed. It does this by measuring difference between predicted values and the actual values. The error term is important because we usually want to minimize the error. In other words, our predictions are very close to the actual values.
In general, RMSE is a commonly used metric and serves well as a general purpose error metric.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(d_i - f_i)^2} \quad (4)$$

3.3 AUC

The Area Under the Receiver Operating Characteristic curve (AUC) is a performance measure used to evaluate the effectiveness of a classification model. Its result is in a range [0,1], where 1 indicates the best performance. Given the subset of previous legitimate transactions T+ and the subset of previous fraudulent ones T, the formalization of the AUC metric is reported in the Equation 10, where indicates all possible comparisons between the transactions of the two subsets T+ and T. The result is obtained by averaging over these comparisons.

$$\Theta(t_+, t_-) = \begin{cases} 1, & if\ t_+ > t_- \\ 0.5, & if\ t_+ = t_- \\ 0, & if\ t_+ < t_- \end{cases} \qquad AUC = \frac{1}{|T_+| \cdot |T_-|} \sum_{1}^{|T_+|} \sum_{1}^{|T_-|} \Theta(t_+, t_-)$$

Another way to examine the performance of the model is ROC (Receiver Operator Characteristic) graph. A ROC graph is a plot with the false positive tare on the X-axis and the true positive rate on the Y-axis. The ROC curves for XGBoost is shown in following figure 3:

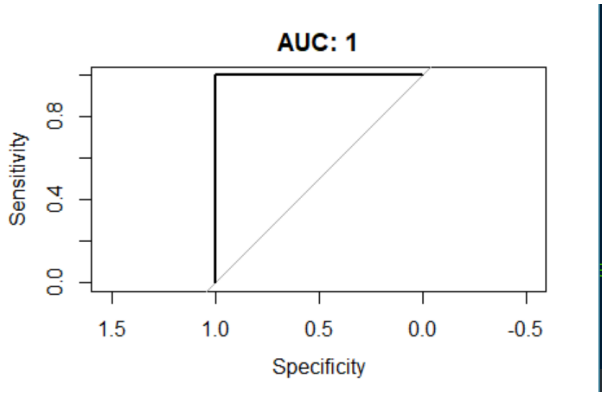

Figure 5: ROC curve for XGBoost

The area beneath an ROC curve can be used as a measure of accuracy.

3.4 Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives you insight not only into the errors being made by your classifier but more importantly the types of errors that are being made. It is this breakdown that overcomes the limitation of using classification accuracy alone.

After selecting and evaluating the proposed 'Cluster-then predict Model', predictions on the test set was done. The predicted and actual results can then be compared using a *confusion matrix* shown in figure 4.

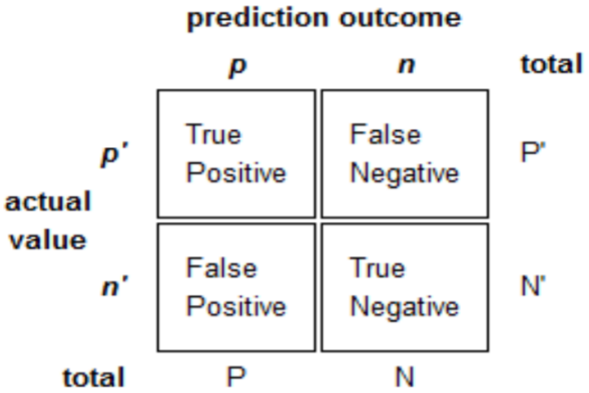

Figure 6: Confusion Matrix

4 Problem description and data:

The data comes from some previously occurred fake electronic transaction. The objective of this report underlying the content is to overcome the above mentioned issue in an efficient way by using R Programming Language. We need to predict the Fake variable (binary) based on various independent variables. The prediction is to be made on new validation dataset.

5 Results:

Table containing all the models and their accuracies:
In the present project we considered three

modelling techniques - Logistic regression, Random forest and XGboost to predict best results. In the process of predicting, initially we applied logistic regression and achieved good results with accuracy of 97.96%. However, this dataset is highly imbalanced with 99.8% of the transactions being not fraudulent. For such a dataset accuracy is not a good measure: the classifier that just says that everything is not fraud already gets 99.8% accuracy for example. Instead we must look at the confusion matrix and see how many of the fraudulent transactions you were actually able to identify. For logistic regression, the AUC is 97.7 which is pretty good but we can still improve the results by using other modelling techniques. In the case of Random Forest we achieved accuracy of 99.18% and AUC value as 99 . At the end when we used XGboost model we got an accuracy of 99.95% and AUC value as 98.6 which is the best of all the predictions. Later we used different approach of cluster wise prediction. In this method we clustered our data into different clusters and then applied different modelling techniques to each clusters for better predictions. However, we got best results when we used logistic regression model for all clusters. We predicted 6 as '1' and rest as '0'. But the rmse slightly increased from 0.020 to 0.03. Hence, among all the models we achieve best results using Xgboost model. The accuracy of this model can be increased by tuning the parameters of xgboost model.

Below is a short review table summarizing the evaluation of different approaches:

S.no	Technique	Accuracy	Area Under Curve(AUC)
1	Logistic regression	99.96%	97.7
2	Random forest	99.18%	99
3	XGboost	9995%	98.6

Table 2: Evaluation of Various Classification Algorithms

6 Conclusion and future work

This paper presents a hybrid mechanism-'Cluster-then predict Model' to improve accuracy of detection of fake electronic transactions. This hybrid model will perform even better when the raw data is very large and diversified. In our project, xgboost model predicted the values more accurately. In future work, author(s) will try to tune the parameters of xgboost model for better accuracy.

7 References

[1]John Kiernan,2013 ,Credit Card Debit Card Fraud Statistics,http://www.cardhub.com/edu/credit-debit-cardfraud-statistics/(2013.)

[2] Linda Delamaire , Hussein Abdou , John Pointon (2009). Credit card fraud and detection techniques, Banks and Bank Systems, Volume 4,57-68

[3] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," The Annals of Statistics, vol. 28, no. 2, pp. 337-407, 2000.

[4] R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in Proc. of the 11th IEEE International Conference on Tools with Artificial Intelligence, Evanston, 1999, pp. 103-106.

[5] S. Kotsiantis, D. Kanellopoulos, P. Pintelas (2006). Handling imbalanced datasets: A review. International Transactions on Computer Science and Engineering.

[6] P.K. Chan, W. Fan, A.L. Prodromidis, S.J. Stolfo (1999). Distributed Data Mining in Credit Card Fraud Detection. IEEE Intelligent Systems, pp 67-74.