Credit Card Fraud detection using Machine Learning Models

Murugavel.P, Akash.S, Adithyan.R, S.Niveditha
CSE Department,
SRM institute of Science and Technology, Vadapalani, Chennai, Tamil Nadu, India.

ABSTRACT

Due to increase of fraud which results in loss of money across the globe, several methodologies and techniques developed for detecting frauds Fraud detection involves analyzing the activities of users in order to understand the malicious behaviour of users. Malicious behaviour is a broad term including delinquency, fraud, intrusion, and account defaulting. This paper presents a survey of current techniques used in credit card fraud detection and evaluates the machine learning approach to identify fraud detection. In the proposed work, we analyze credit card fraud detection using machine learning algorithm namely K-Nearest Neighbor and Ensemble Model of Random Forest. To make the learning process efficient, we used infinite latent feature selection algorithm for feature selection. The performance of the algorithm is evaluated on various measures like Accuracy, Precision and Recall.

KEYWORDS: Machine learning, Ensemble Model of Random Forest, KNN.

1. INTRODUCTION

With the emerging rise of technology today, the dependency on e-commerce and the online payments has grown exponentially.

As the credit card provides convenience to the users but frauds caused due to these activities causes inconvenience. The credit card information is confidential, the bank and the other financial enterprises doesn't want to disclose the information about their customers. Risk management is critical for financial enterprises to survive in such competing industry. The provisional loss arises due to the “bad” accounts bank lends the money to customers who eventually do not have capability to pay back. In the risk management, the chances of false negative

(false “good” accounts) could still be high. However, by leveraging their performance such as credit card utilization, payment information, risks can further be managed to control provisional loss. In this paper, a focus on risk management as well as fraud detection is depicted.

1. It shows an interest in classifying if a booked account as a “bad” account within 12 months since booked. Since an internal

classification model is already available, with a secondary interest to train a better classifier to outperform the benchmark model.

2. Since there are few research initiatives that implements fraud detection.

Concentration on how to optimize fraud detection techniques is brought to light. Since the emergence of many advanced computing and classification systems including the support vector machine and the optimization technique like genetic algorithm shows a greater fluctuation in the implementation of many different technologies due to the accuracy and efficiency it produces. This research uses a hybrid approach of Genetic Algorithm and K Nearest Neighbour to perform fraud detection.

2. RELATED WORK

2.1 Author combined individual classifiers and a multiple classifier system with an increase in classification accuracy is presented. They proposed multiple classifier system based on the Random Forest, Principle Component Analysis and Potential Nearest Neighbor methods As Breiman suggested, the performance of the Random Forest depends on the strength of the weak learners in the forests and diversity among them. The Principle Component Analysis method is applied to transform data at each node to another space when computing the best split at this node. This process increases the diversity of each tree in the forest and thereby improves the overall accuracy. The Random Forest is studied through the perspective of the Adaptive Nearest Neighbor. They introduce the concept of monotone distance measures and potential nearest neighbors and show that the Random Forest can be viewed as an adaptive learning mechanism of k Potential Nearest

Neighbors. Considering the information loss caused by out-of-bag samples, a new voting mechanism based on Potential Nearest Neighbor is also presented to replace the traditional majority vote. The proposed algorithm improves the classification accuracy of the ensemble classifier by improving the difference of the base classifiers.

2.2 The purpose of this study is to construct a valid and rigorous fraudulent financial statement detection model. The research objects are companies which experienced both fraudulent and non-fraudulent financial statements between the years 2002 and 2013. In the first stage, two decision tree algorithms, including the classification and regression trees (CART) and the Chi squared automatic interaction detector (CHAID) are applied in the selection of major variables. The second stage combines CART, CHAID, Bayesian belief network, support vector machine and artificial neural network in order to construct fraudulent financial statement detection models. According to the results, the detection performance of the CHAID–CART model is the most effective,

with an overall accuracy of 87.97 % (the FFS detection accuracy is

92.69 %).

3. PROPOSED WORK

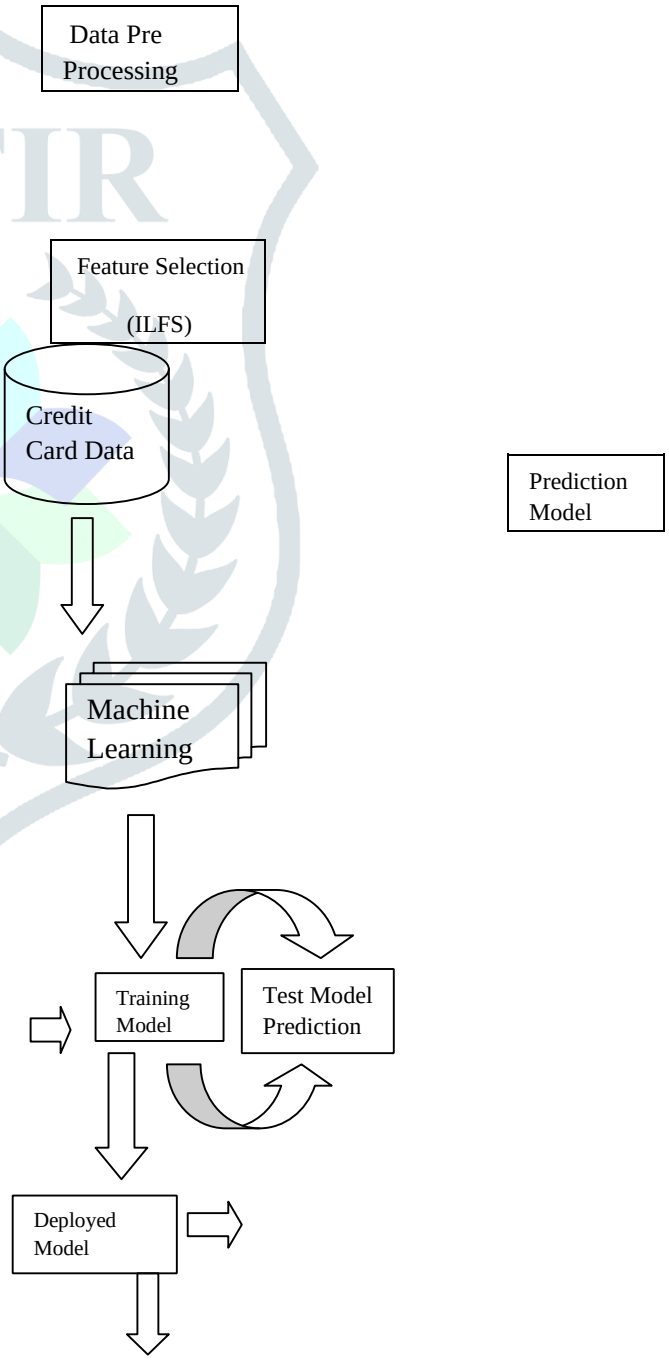
- Credit card logs mining is one of the most significant fields in the area of data mining. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks.
- First load the whole dataset.
- Second we propose feature selection model for predicting the importance variables. Infinite Latent Feature Selection model is used for ranking. Thus we select top 4 attributes from ranked result.
- Third, getting modified dataset from the top 4 attributes in original dataset. After that, whole dataset is splitting into testing and training.
- Finally, training data are trained by using classifier such as k-nearest neighbour and ensemble model of random forest and testing data are evaluated based on trained model. Thus k-nearest neighbour and random forest classifier is used for classify the credit card fraud based on predicted top 4 attributes. At last find out the proposed method accuracy and make comparison with exiting machine learning models.

Advantages

➢ Best accuracy for the study.

➢ ILFS algorithm gets best features.

ARCHITECTURAL DIAGRAM



4.MODULES: In that, our whole dataset is divide into training and testing data; use 80% of data for training and 20% of data for testing.

1. Dataset Collection
2. Feature Selection
3. Data Splitting
4. Prediction

MODULE DESCRIPTION

4.1. Dataset Collection

- The credit card fraud dataset is obtained by Own.
- We use 6 attributes i.e. Time, Amount, Issued Date, Expiry Date, Gender and Age for credit card fraud prediction.

4.2. Feature Selection

- The second module is feature selection which is used to predicting the importance variables.
- In that, Infinite Latent Feature Selection model is used for ranking the variables based on their weight calculation.

 Finally we select top 4 variables from ranked result.

4.3. Data Splitting

 The splitting step is used for creating the training and testing data to analyzing process.

4.4. Prediction

 In classification, splited training and testing data are valuated based on machine learning model.

- First, training data was trained by using machine learning model such as K-Nearest Neighbour and
- After that testing data are validated based on trained data with high classification accuracy rate. Thus we predicted the credit card fraud.

Finally our proposed model compared with exiting machine learning model based on evaluation metric i.e. accuracy, precision and recall.

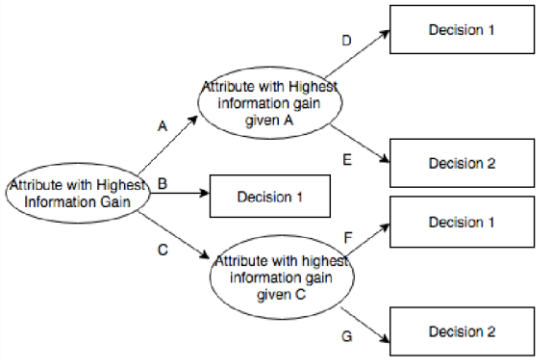
K-NEAREST NEIGHBOUR

Hodges et al. in 1951 presented a pattern based on nonparametric model which is called as K-Nearest Neighbour rule. KNN is the one of the basic and simple but very intelligent classification algorithm. This algorithm did not create any assumptions for data and normally KNN used for classification process when no

knowledge about the data distribution. The working of KNN is finding the k nearest data for the test data in the training set of data. K nearest data finding in training set is based on Euclidean Distance.

DECISION TREE

Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split sample into two or more homogeneous sets (or subpopulations) based on most significant splitter / differentiator in input variables. In decision tree internal node represents a test on the attribute, branch depicts the outcome and leaf represents decision made after computing attribute.



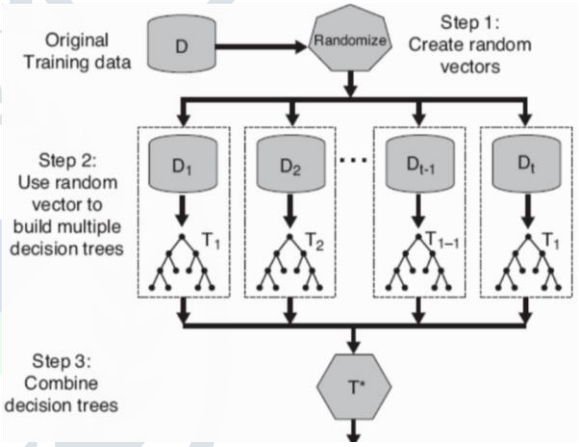
RANDOM FOREST

1. Given there are n cases in the training dataset. From these n cases, sub-samples are chosen at random with replacement. These random sub-samples chosen from the training dataset are used to build individual trees.

2. Assuming there are k variables for input, a number m is chosen such that m < k. m variables are selected randomly out of k variables at each node. The split which is the best of these m variables is chosen to split the node. The value of m is kept unchanged while the forest is grown.

3. Each tree is grown as large as possible without pruning.

4. The class of the new object is predicted based upon the majority of votes received from the combination of all the decision trees.



5.RESULTS

In this part, we show the classified result from two prediction models. We used different parameters for make comparison with different models; the parameters i.e.

Accuracy, Precision and Recall.

Table 1: Confusion Matrix

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Total Dataset- 200; Training Set – 140,

Testing Set - 60

Table 2: Confusion Matrix of KNN

Actual/Predicted	**Real**	**Fake**	**Total**
Real	27	2	29
Fake	2	29	31

Table 3: Confusion Matrix of RF

Actual/Predicted	**Real**	**Fake**	**Total**
Real	27	2	29
Fake	1	30	31

Based on the above prediction, we evaluate the parameters i.e. accuracy, precision and recall.

i) Accuracy: It measures the analysis of TP and TN to the total no. of test images.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

ii) Precision: It is the estimation analysis of true positive to the aggregate value of true positive and false positive rate. It is given in eqn. (2)

$$Precision = \frac{(TP)}{(TP+FP)} \quad (2)$$

iii) Recall: It is the estimation analysis of true positive rate to the aggregate value of the true positive and false negative rate. It is given in eqn. (3).

$$Recall = \frac{(TP)}{(TP+FN)} \quad (3)$$

Table 4: Quantitative Evaluation with two models

Accuracy	0.9333	0.9500
Precision	0.9310	0.9642
Recall	0.9310	0.9310

6.CONCLUSION

In this work we analyzed the two machine learning algorithms i.e. KNN, and Ensemble Model of Random Forest for predicting of fraud credit card. Prediction of credit card fraud followed the step as data preparation in that data was collected from public database; feature selection for selecting important variables; data splitting in that whole dataset split into training and testing; finally

machine learning model based training data are trained and testing data are validated by the trained data in classification step. At last, above two machine learning models are compared interms of finding of quantitative evaluation metrics such as accuracy, precision and recall.

7.REFERENCES

[1] A. O. Adewumi and A. A. Akinyelu, ``A survey of machine-learning and natureinspired based credit card fraud detection techniques," Int. J. Syst. Assurance Eng.

Manage., vol. 8, no. 2, pp. 937953, 2017.

[2] The Nilson Report. (Oct. 2016). [Online].Available:https://www.nilsonreport.com/upload/content_promo/The_Nilson_Report_10-17-2016.pdf

[3] N. Mahmoudi and E. Duman, ``Detecting credit card fraud by modified fisher discriminant analysis," Expert Syst.

Appl., vol. 42, no. 5, pp. 25102516, 2015.

[4] N. S. Halvaiee and M. K. Akbari, ``A novel model for credit card fraud detection using articial immune systems," Appl. Soft Comput., vol. 24, pp. 4049, Nov. 2014.

[5] Y. Sahin, S. Bulkan, and E. Duman, ``A cost-sensitive decision tree approach for fraud detection," Expert Syst. Appl., vol. 40, no. 15, pp. 59165923, 2013.

[6] N. Mahmoudi and E. Duman, ``Detecting credit card fraud by modified fisher discriminant analysis," Expert Syst.

Appl., vol. 42, no. 5, pp. 25102516, 2015.