ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Plastic Money Fraud Detection Using Ensemble Learning Technic

Aditi Gaur[1], Rachna [2]

P.G. Scholar, Department of Computer Science and Engineering, N.G.F College of Engineering and Technology at Palwal, Haryana, India[1]

Assistant Professor, Department of Computer Science and Engineering, N.G.F College of Engineering and Technology at Palwal, Haryana, India[2]

ABSTRACT: Online payments done with credit card is increasing at fast pace these days as online payment system is adapted almost everywhere. Now this fast growth is pushing all the financial systems to make an improvement in their fraud detection system. In the practical scenario, the fraud dataset under observation of credit card is seen heavily imbalanced and we have seen misclassification errors in them. It is very important to control them and resolve them. We have seen in previous research that classification techniques provide solutions to find the fraud and non- fraud transactions. But in some specific scenarios, classification techniques underperform, does not give promising solutions when the situation of huge numbers of difference in minority and majority cases occur. Hence in this study, we present an ensemble machine learning approach as a possible solution to this problem. Our observation is thatis more accurate in detecting normal instances and is for detecting fraud instances. We present an ensemble method based on combination of Naïve Bayes, Linear SVM and Adaboost which is able to predict with high accuracy

KEYWORDS: fraud detection, ensemble learning, naïve bayes, linear SVM.

I. INTRODUCTION

In the present time, the utilization of credit card is expanding day by day. Due to this paper less money transfer trade, this online payment system is rising. In turn which is giving rise to a risk of illegal and unlawful transactions. For fraudsters it is an easy platform to make frauds because it is having low risk and results in good reward. Today's fraud detection is designed to prevent one twelfth of one percent of all transactions processed which still translates into billions of dollars in losses. For example, onMay 15,2016, in a coordinated attack, a group of around 100 individuals used the data of 1600 south African credit cards to steal 12.7 million USD from 1400 convenience stores in Tokyo within three hours. According to a European central report [1] every day billions of euros are lost in Europe due to credit card fraud. Fraudsters persistently keep changing their tactics and masterplan to mislead and cheat the financial systems. To catch these clever and mastermind frauds the old rule-based systems are outdated. Also, they are slow to prevent financial losses. So, this problem motivates to put some machine learning techniques to detect these frauds. Further we discuss some loopholes which are related while using machine learning techniques to find fraudulent use of credit cards.

II. RELATED WORK

In the previous years, many solutions have been proposed to deal with the problem of poor performance of imbalance datasets, both for standard learning algorithms and for ensemble techniques (lopez and Fernandez,2013). They can be categorised into three parts: -

1.Data sampling in this training samples are reorganized in such a way to give more or less balanced class distribution that permit classifiers to perform in identical fashion to standard classifier.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

2. Cost sensitive learning, this type of solutions includes methods which are at data level, or at both levels assimilate.

3. Algorithmic modification, this procedure is inclined towards the adaptation of base learning methods to be more adapted to class imbalance issues.

Sampling methods

Sampling is a popular methodology to counter the problem of class imbalance. One can find some papers about sampling techniques studying effect of changing class distribution in order to deal with imbalanced datasets(ha and bunke, 1997). The goal of sampling methods is to create a dataset that has a relatively balanced class distribution, so that traditional classifiers are better able to capture the decision boundary between the majority and minority classes. These sampling methods have proved that balanced data results in improved overall classification performance compared to an imbalanced data set (analytics vidya content team 2016). The main methods used to treat imbalanced datasets are presented below.

1. Under sampling and oversampling

In under sampling, the majority class instances are discarded at random until a more balanced distribution is reached. For example, a dataset consisting of 10 minority class instances and 100 majority class instances. In under sampling one might attempt to create a balanced class distribution by selecting 90 majority class instances at random to be removed. The resulting dataset will then be consisting of 20 instances: 10 remaining majority class instances and 10 minority class instances. Alternatively, in oversampling, minority class instances are copied and repeated in the dataset until a more balanced distribution is reached. Thus, if there are two minority class instances and 100 majority class instances oversampling would copy the two minority class instances 49 times each. The resulting dataset would then consist of 200 instances: the 100 majority class instances and 100 minority class instances (i.e. 50 copies each of the two minority class instances). Under sampling and oversampling create more balanced distributions, they both suffer from serious drawbacks. In order to overcome these limitations, more sophisticated techniques have been developed i.e. SMOTE.

2. **Synthetic Minority Oversampling Technique (SMOTE)**

In order to overcome the issues mentioned above, (chawla and Bowyer,2002) developed a method of creating synthetic instances instead of merely copying existing instances in the dataset. These techniques are known as the synthetic minority oversampling technique (SMOTE). This approach is inspired by a technique that proved successful in handwritten character recognition (ha and bunke,1997). In SMOTE, the minority class is oversampled by taking each minority class sample and introducing syntheticalong the line segments joining any and introducing synthetic examples along the line segments any of the k minority class nearest neighbours. Depending upon the amount of oversampling required, neighbours from the k nearest neighbours are randomly chosen.

3. Cost-sensitive learning

Cost sensitive learning is another commonly used method to handle classification problems with imbalanced data. In simple words, this approach estimates the cost related with misclassifying consideration. It does not create balanced data distribution. Alternatively, it put light over the imbalanced learning problem by using cost matrices which outline the cost for misclassification. These misclassification cost values can be given by domain experts, it is usually more interesting to recognize the positive instances rather than the negative ones. (lopez Fernandez 2013)

III. PROPOSED SYSTEM

Ensemble learning is a strong way to improve the performance of machine learning model. In this process more than one model such as classifiers are generated and then combine to provide solutions to the problem. The classifiers enhance stability and predictive power of model. Here we have taken the dataset from credit card fraud. This is a case

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

of imbalanced dataset as there are many examples of the non- fraud class. If this was not the case the out of box classifiers such as logistic regression, Support Vector Machines would well on these. But on such datasets, where the ratio of majority: minority is 9:1. The classifier would although show accuracies of 80%-90% but those would be generalized accuracies as it is just a count of how many samples were wrong out of the data the classifier was asked to predict. Our observation is that before applying ensemble of classifiers the accuracy was 98% which was fraud prediction. So, observing this fraud results we tried to predict accurate results and applied new technique using different classifiers which are adaboost, naïve Bayes and linear SVM.

Proposed Algorithm

Step1.Data Preparation: - Data labels are trained and tested and applied classification without any pre-processing.
Train labels: - fit the labels with individual classifiers. The classifiers used are random forest, MLP, K Neighbours.
Test labels: - labels are tested using same classifiers and fetched score through them.

Random Forest	0.999051120863
Neural Network	0.99823403049
K nearest neighbours	0.99832628263

Step2.Applying different classification techniques: -
Under sampling: - Again training is done using random forest and SVM individually and tested then find results.

	Random Forest	SVM
Mean Accuracy in %	99.665418227	99.66541822
Standard Deviation in %	0.5110922652	0.511092265

Oversampling:Again, training is done using Random forest and SVM individually and find results.

	Random Forest	SVM
Mean Accuracy in %	92.21973010	53.750267
Standard Deviation in %	1.784844577	1.467304971

Step3. Hybrid Classification
Three classifiers are used for three different sections of data, training and testing. These are:-
Random forest
MLP
K Neighbour
For final classification all are fit together using a logistic Regression. The data is split into data1, data 2, data 3.
Then accuracy on original data is shown by this ensemble model.
On sampled data, these classifiers are fit for these three sections of data, training and testing and final classification is given by logistic regression.

Using different ensembles: -
Here on original data, three different classifiers are fit for three sections of data, training and testing. Final result is given by the ensemble of " Adaboost, Linear SVM and Naïve bayes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

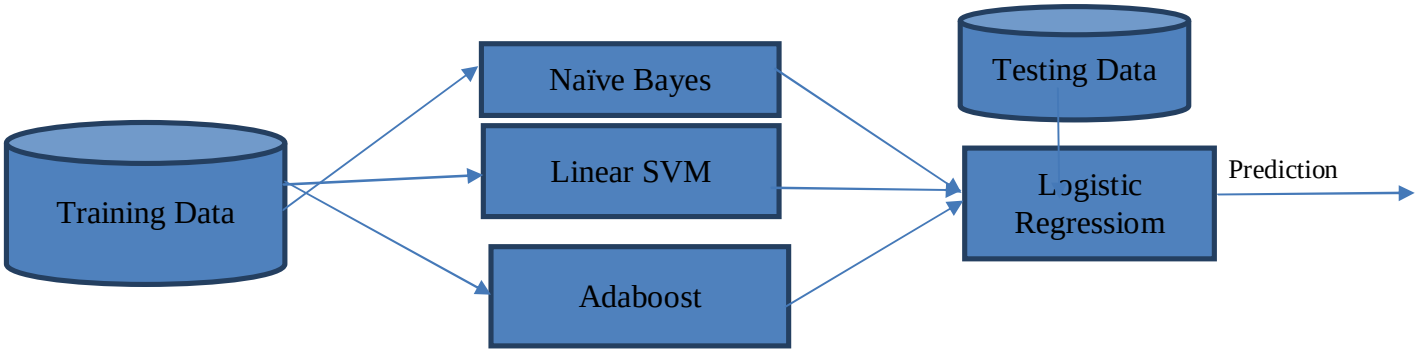
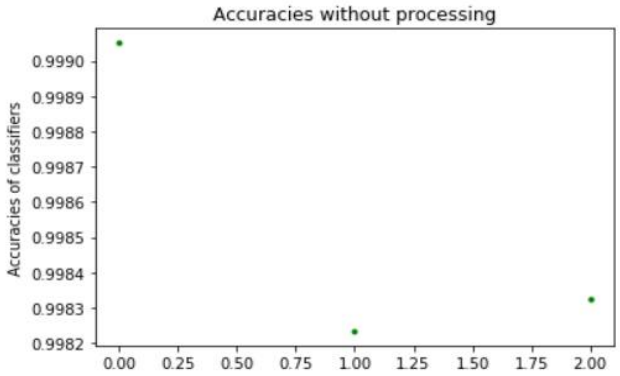


Figure: Proposed Framework

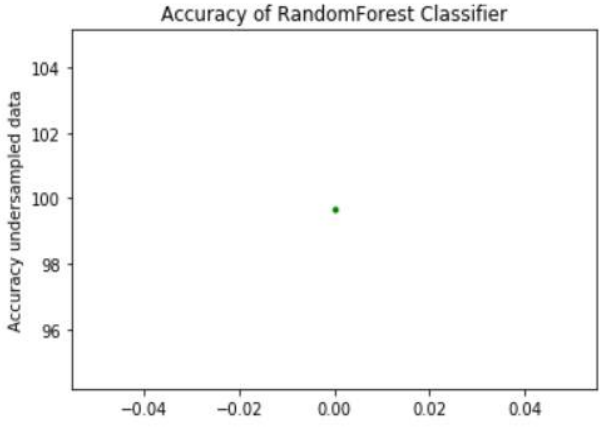
IV. RESULTS

The studies involve the hybrid approach which is formed with the amalgamation of three machine learning models i.e. Naïve Bayes, Linear SVM, Adaboost. The following figures are depicted for proposed scenario and results:

1. Accuracies while applying classification without any pre-processing: -



2. Accuracies while applying different processing techniques for handling imbalanced datasets: - (Undersampling)



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

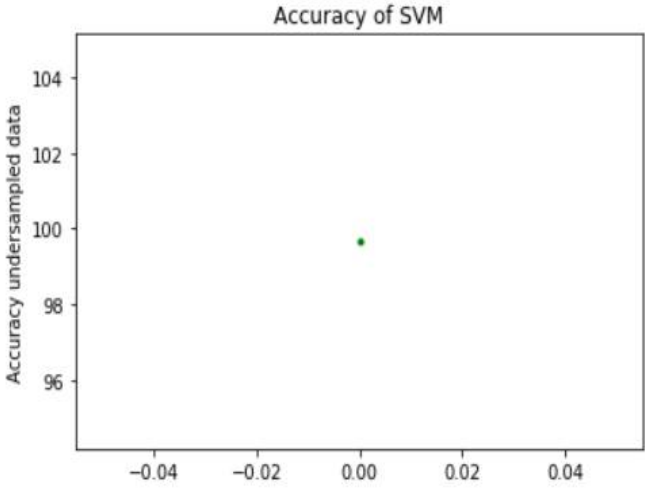
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

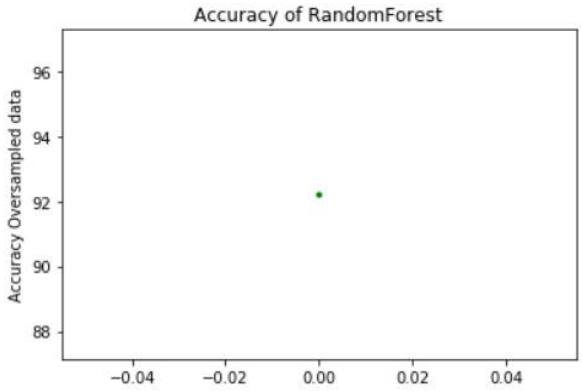
Vol. 7, Issue 5, May 2019

The above-mentioned result is given by taking individual classifier which is Random Forest Classifier using under sampling.



The above-mentioned result is given by taking individual classifier which is Support Vector Machine using same under sampling technique.

Oversampling: -
Here accuracies are fetched while applying different processing techniques for handling imbalanced datasets.



Accuracies are fetched using individual classifiers which is Random Forest, under oversampling technique.

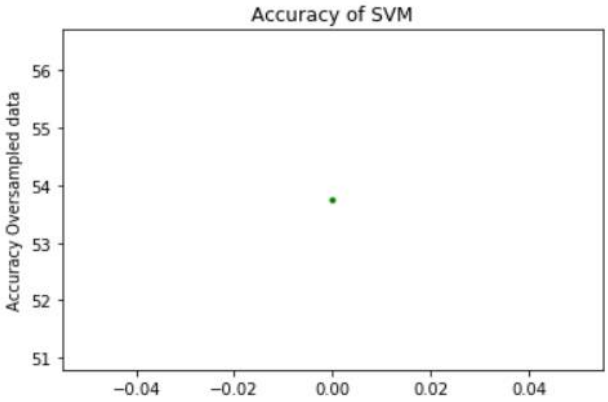
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

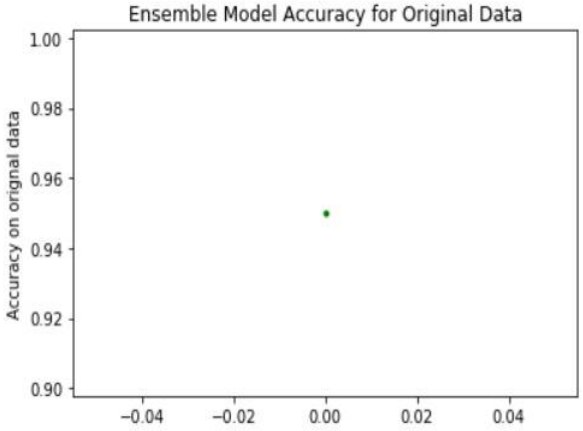
Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

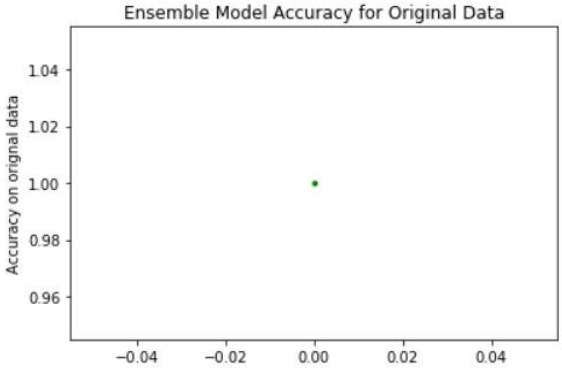


Here also oversampling technique is applied using individual classifier i.e, Support Vector Machine and results are fetched.

Applying hybrid classification (ensemble classifiers):-
. On original data



.Using different ensembles:-(Linear SVM, Adaboost and Naïve bayes)



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

V. CONCLUSION AND FUTURE WORK

In this paper, we have put light over key areas in detecting credit card fraud which is a highly disordered and assymetric problem. We proposed an ensemble model which is a combination of Adaboost, Naïve Bayes and Linear SVM to accurately predict the frauds. Our experiments shown that this is much better approach other than popular approaches. In this research we have worked over improving the results via improving accuracies with the help of robust classifiers. Although the scope of our work is limited to the dataset having numerical values, this work could be extended by including some more robust classifiers like XGBOOST or some more sophisticated techniques like word2vec [12].

REFERENCES

[1] Second report on card fraud. In European Central Bank, 2013.
[2] https://medium.com/razorthink-ai/4-major-challenges-facing-fraud-detection-ways-to-resolve-them-using-machine-learning-cf6ed1b176dd
[3] Sanjeev Jha, Montserrat Guillen, and J Christopher Westland. Employing transactionaggregation strategy to detect credit card fraud. Expert systems with
applications, 39(16):12650–12657, 2012.
[4] Chunhua Ju and Na Wang. Research on credit card fraud detection model basedon similar coefficient sum. In First International Workshop on Database Technologyand Applications, DBTA 2009, Wuhan, Hubei, China, April 25-26, 2009, Proceedings,pages 295–298, 2009.
[5] E. W. T. Ngai, Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification frameworkand an academic review of literature. Decision Support Systems, 50(3):559–569,2011
[6] Piotr Juszczak, Niall M. Adams, David J. Hand, Christopher
[7] https://medium.com/razorthink-ai/4-major-challenges-facing-fraud-detection-ways-to-resolve-them-using-machine-learning-cf6ed1b176dd
[8] Lopéz, VictoriaandAlbertFernández (2013). An insightintoclassificationwithimbalanced data: Empiricalresultsandcurrenttrendsonusingdataintrinsiccharacteristics. Information Sciences 250(2013)113-141(cit. onpp.2,5,8).
[9] H.Han andW.Y.Wang(2005). Borderline-SMOTE: anewover-samplingmethodinimbalanced data setlearning. LectureNotesinComputerScience,3644,878-887(cit.onp.40).
[10] Chawla, NiteshandKevinBowyer (2002). SMOTE: syntheticminorityover-samplingtechnique. Journal ofArtiflcialIntelligenceResearch16(2002)321–357(cit.onpp.7,9).
[11]https://medium.com/razorthink-ai/4-major-challenges-facing-fraud-detection-ways-to-resolve-them-using-machine-learning-cf6ed1b176dd
[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In
Proceedings of the 26th International Conference on Neural Information Processing
Systems, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.