Fraud Detection Using Random Forest Algorithm

Eesha Goel
Computer Science Engineering and Technology, GZSCCET, Bhatinda, India
eesha1992@rediffmail.com

Abhilasha
Computer Science Engineering and Technology, GZSCCET, Bhatinda, India
abd_jain@rediffmail.com

Ankit Agarwal
Manager Analytics, Snapdeal, Gurgaon, India
ankitagar@gmail.com

Abstract—The usage of internet, online shopping and home services are gaining importance among people. But, at the same time customers are also taking undesirable benefits by performing unusual activities. In case of online shopping, a certain limitation in purchasing number of products are specified by the e-commerce websites. When people are getting higher discounts, some people start taking advantage of it and conduct fraudulent activities. Therefore, analysis of fraud is required to detect them and prevent these unusual activities. In this paper, frauds are detected effectively by using Random Forest algorithm in R language.

Keywords-Random Forest, caret, Data Mining, Machine Learning, Supervised Learning, Unsupervised Learning

I. INTRODUCTION

The usage of internet is increasing at a fast rate in the field of e-commerce that means electronic commerce. These websites provide wide variety of products such as home appliances, books, travelling packages, electronic gadgets, software's, clothes, etc. With the introduction of e-commerce websites it becomes possible to buy the products at home itself without going outside and without touching the product physically. This electronic method includes "click and buy" methods using computers and m-commerce by using mobile devices and smart phones. E-commerce websites offer various discounts in order to attract large number of people so that every person is able to buy the product at affordable rates. But, some of the customers start taking advantage of this opportunity by ordering the number of products more than the defined limited offer. Therefore, there is a need to detect these fraudulent activities.

In order to detect frauds, traditional methods are widely used for a longer period of time. These methods are complex and time consuming. The fraudulent activities often consist of same content but usually they are quite different. Detection of frauds is not an easy task. Therefore, proper methods are required to detect frauds efficiently. Statistical techniques and Artificial Intelligence are two primary subject classification where these techniques are utilized.

Statistical techniques includes techniques for pre-processing of data, statistical parameters such as averages, performance metrics, etc. are calculated, user profiles are computed, patterns and association among group of data is determined by clustering and classification method and matching algorithms are performed to determine inconsistent transactions of user.

There are various techniques of Artificial Intelligence that are used for fraud detection such as Data Mining, Expert Systems, Pattern recognition, Machine Learning and Neural Networks

Other techniques like Bayesian Networks, Link Analysis and Decision Tress are also available for fraud detection.

Only statistical techniques are not sufficient to detect fraudulent activities. Therefore, the combination of machine learning and artificial intelligence is required. They are further divided into two methods such as Supervised Learning and Unsupervised Learning.

Supervised Learning is the method that selects sub-sample of data randomly. Then, the selected record will be trained using machine supervised learning algorithm. After training of data model is made which classifies the data into fraudulent and non-fraudulent activities. On the other hand, Unsupervised Learning is similar to the supervised learning but there is only one difference that this method does not utilize labelled records for training, classification and building of model.

In this paper, Random Forest algorithm (a Supervised Machine Learning) is used to detect frauds in R language. This paper is organized as follows: Section II describes about the R language and its features. Section III describes about the Random Forest algorithm and its features. Section IV and V highlights about the implementation and result. Section VI presents conclusion and future work.

II. R LANGUAGE

A. What is R?

R is an open source language like Linux operating system. It does not belongs to a single person rather thousands of persons are responsible for contributing in the development of R. R is a programming language mainly used for statistical analysis and graphics that is supported by the R Foundation of Statistical Computing. It is a dialect of S language [1]. It is a GNU project which is freely available under the GNU General Public License. R was developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand in 1993. It is named as R as the name of both authors begin with the "R" letter. Presently, the R language is developed by the R Development Core Team. It is an interpreted language implemented through command-line interface. With the installation of R a set of packages are inbuilt installed and more than 7,801 packages are available in the Comprehensive R Archive Network (CRAN).

B. Why R is used?

R is not only a statistical computing language rather it is also a programming language. Therefore, user can make his own objects, functions and packages. R programs explicitly make the documentation of the steps that the user executed for analysis. Therefore, it becomes easy to update and correct the issues for better analysis. It is platform independent so, user can use it on any type of operating system [2]. It consist of strong package ecosystem. It is powerful and flexible. It consist of a big library which is composed of several algorithms beneficial for big data analysis. It helps to integrate with many other languages such as python, java, C, C++. It allows to integrate with other databases such as Excel and Access and many statistical packages. It allows to visualize the data by using charts and graphics [3]. It is composed of latest methods of machine learning, predictive modelling and statistics.

III. RANDOM FOREST

A. What is Random Forest?

Random Forest is composed of ensemble of simple tree predictors. It was developed by Brieman in 2001. The response of each tree constituting the random forest depends upon the set of a predictor values that are chosen independently with replacement. The distribution of all trees in the random forest is same and the predictor variables are the subset of the original dataset. It is mainly used for classification and regression problems [4].

For evaluating classification problems, set of simple trees and randomly selected predictor variables are provided. Random forest generates a margin function that determines the extent to which the average number of votes for the correct class exceeds the average vote for any other class present in the dependent variable. But, for the evaluation of regression problems, Random forest is generated with the growth of simple trees. Each tree is capable to provide numerical response value. The predictor variables are selected randomly in the same distribution. The prediction of the Random forest is calculated as the average of predictions calculated by each tree.

If the dataset consist of missing values during the model building, the prediction is evaluated on the basis of the non-terminal node of the respective tree. Therefore, there is no need to delete the missing data.

B. Features of Random Forest

It has excellent accuracy as compared with the current algorithms and runs efficiently on large datasets [5]. It manages several input variables without variable deletion. It estimates which variables are important in the classification. With the processing of forest it produces an internal unbiased estimate of the generalization error. It has the ability to maintain accuracy and estimation of large missing data. It consist of methods for maintaining balance for the unbalanced datasets. It can be saved for the future use. It has the ability to compute the prototypes about the relationship between the variables and the classification and to extend the features for unbalanced data, leading to the unsupervised learning technique. It calculates the proximities between the pair of cases such as clustering, outlier detection or scaling.

IV. IMPLEMENTATION AND RESULTS

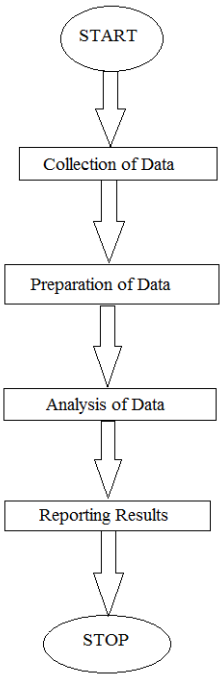

Fig 1 Steps followed for analysis

The steps that are followed for analysis of data is shown in fig 1:

- **Collection of data**: The first step is to collect the data. The data can be collected from various methods such as crawling, Application Program Interface (API) or directly from the company. In this research work, the data is collected directly from the company itself. A dataset of 5100 number of records is obtained having 30 attributes like Name_of_the_Customer, Order_Date, Allowed_Offers, Customer_email_address, Customer_Mobile_No, Payment_method, Result, etc. Six velocity variables are selected such as Customer_email_address, Customer_Mobile_No, IP_address, IP_address and Password, Payment_method and Bank_Account_Number.
- **Preparation of data**: After collecting data, the data is prepared for performing analysis efficiently. The obtained dataset is composed of various attributes that are not necessary for analysis, therefore it is better to prepare the data according to the need so that the algorithm generates accurate results. In this research work, the data is prepared using data.table package and merge function. data.table package provides an upgrade version of data.frame. It allows the user to perform amazing fast manipulation of data and it is widely useful for working with large datasets [6]. Merge function allows to merge two data frames on the basis of common columns or row names by calling the data.frame method. If specified columns are merged then the column names are specified using by.x (column names of first file) and by.y (column names of second file) attribute [7]. First, the original dataset is loaded and the number of occurrences of each velocity variable is calculated by setting Order_Date and Product_ID as a key. Then, the output of all occurrences of each velocity variable are merged with the original dataset by using merge function.
- **Analysis of data**: After preparing the data, the analysis is performed using various algorithms according to the need. In this research work, random forest algorithm is used. It is implemented using caret and randomForest package. Caret package consist of set of functions for training and creating classification and regression predictive models. It is composed of various tools [8] such as splitting of data, Pre-processing of data, Selection of features from the data, Tuning of model using resampling method and Estimation of variable importance. randomForest package implements Random Forest Algorithm that was introduced by Brieman in 2001. This is used for the classification and regression of data. First, with the use of caret package the 90 percent of dataset is divided into training set and rest as test set. Then randomForest package is applied by setting ntree attribute as 1000, mtry attribute as 4 and Result variable as a dependent variable from the training dataset. Random Forest model does not allow missing values (NA). Therefore, to remove them missing values of numeric variables are replaced with 0 and then randomForest package is applied.

- **Reporting results**: After the complete analysis of data, results are obtained. The results are produced by calling the randomForest variable. By observing fig2 it has been observed that the random forest has performed classification and results are displayed in the form of confusion matrix.

  Confusion Matrix is a [9] table that is used to predict the performance of the classification model. The structure of confusion matrix is shown in fig 3.

```
Call:
 randomForest(formula = Result ~ +Order_Date + Paymen
t_method +      Allowed_Offers + N_for_1i + N_for_1ip
 + N_for_1M + N_for_2i +      N_for_2ip + N_for_2M +
N_for_3i + N_for_3ip + N_for_3M +      N_for_4i + N_f
or_4ip + N_for_4M + N_for_5i + N_for_5ip +      N_for
_6ip + N_for_5M + N_for_6i + N_for_6M + N_for_7i + N_
for_7ip +      N_for_7M + N_for_8i + N_for_8M + N_for
_8ip + N_for_9i + N_for_9M +      N_for_9ip + N_for_1
0i + N_for_10ip + N_for_11i + N_for_11ip +      N_for
_11M + N_for_10M + N_for_12i + N_for_12ip + N_for_12M
 +      N_for_13i + N_for_13ip + N_for_13M + N_for_14
i + N_for_14ip +      N_for_14M, data = training, ntr
ee = 1000, mtry = 4)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 87.91%
Confusion matrix:
         FAULT NO FAULT class.error
FAULT      438        3 0.006802721
NO FAULT  4032      117 0.971800434
```

Fig 2: Results obtained using randomForest

	Predicted False	Predicted True	
Actual False	True Negative	False Positive	
Actual True	False Negative	True Positive	

Fig 3 Structure of Confusion matrix

The observation made from confusion matrix is described below:

1. There are only two predicted possibilities such as FAULT and NO FAULT. The faulty data will be in the FAULT category and the non-faulty data is in the NO FAULT category.
2. Total of 4590 predictions are made.
3. Out of 4590 predictions, classifier predicts 4464 times FAULT and 126 times NO FAULT.
4. In reality, 441 records are faulty and 4149 records are non-faulty.

The Out Of Bag (OOB) error estimation is used to get an unbiased test error. It is estimated internally during the processing of random forest.

The accuracy of the model is calculated as (438+117)/4590=0.121.

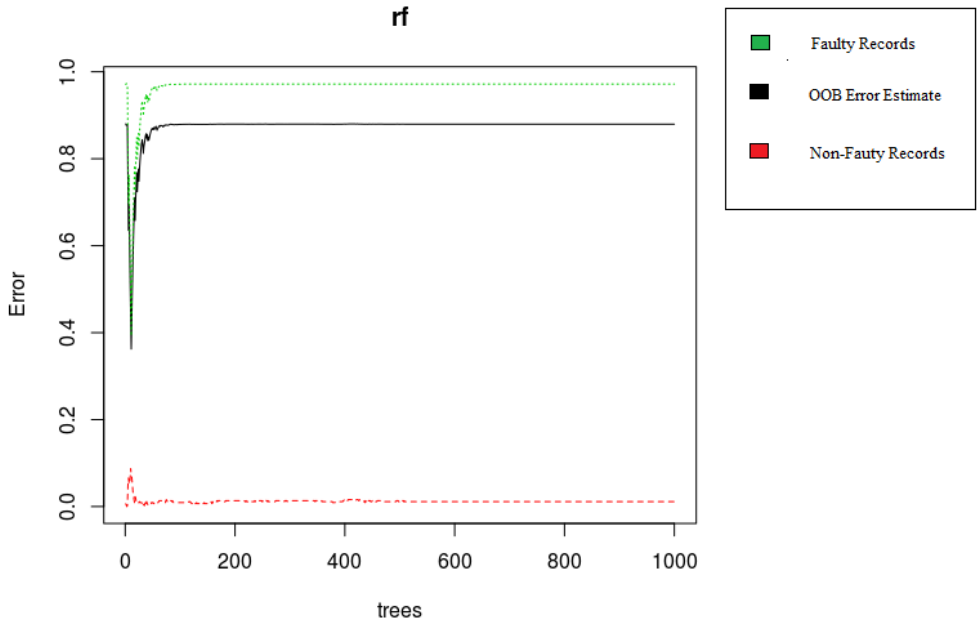

Fig 4: Graphical representation of fault detection using randomForest package

By observing fig4 it is concluded from this plot is that when there are around 10 number of trees the non-faulty records increases but, when the number of tress increases the non-faulty records decreases slightly and then remains constant. On the other hand, when there are less than 7 number of trees both faulty records and OOB error increases and decreases slightly. But, when there are 7 number of trees faulty records and OOB error decreases and with the increase in number of trees they again start fluctuating and lastly remain constant.

V. CONCLUSIONS

With the increase usage of internet, the importance of e-commerce has also increased. Many people are indulging in this activity and make online shopping. But, when huge discounts are provided these activities also produces unusual activities on purchasing products and services. Therefore, fault detection method has solved this problem. In this paper, with the use of Random forest algorithm faults are detected in R language (a powerful package ecosystem).

As the usage of e-commerce website is increasing, the data is becoming large. This makes difficult for R language to analyze the whole data properly. Therefore, another approach is needed to handle this huge amount of data. SparkR tool is efficient for handling large amount of data. This tool will integrate the Spark tool with the R language so that the programmer can take advantage of both powerful analytics tools.

REFERENCES

[1] https://en.wikipedia.org/wiki/R_(programming_language)
[2] http://www.r-bloggers.com/why-use-r/
[3] http://www.inside-r.org/why-use-r
[4] http://www.statsoft.com/Textbook/Random-Forest
[5] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
[6] https://www.datacamp.com/community/tutorials/data-table-cheat-sheet
[7] https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html
[8] (http://topepo.github.io/caret/index.html
[9] http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/