Abstract – The project is mainly focussed on credit card fraud detection in real world. A phenomenal growth in the number of credit card transactions, has recently led to a considerable rise in fraudulent activities. The purpose is to obtain goods without paying, or to obtain unauthorized funds from an account. Implementation of efficient fraud detection systems has become imperative for all credit card issuing banks to minimize their losses. One of the most crucial challenges in making the business is that neither the card nor the cardholder needs to be present when the purchase is being made. This makes it impossible for the merchant to verify whether the customer making a purchase is the authentic cardholder or not. With the proposed scheme, using random forest algorithm the accuracy of detecting the fraud can be improved can be improved. Classification process of random forest algorithm to analyse data set and user current dataset. Finally optimize the accuracy of the result data. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, and precision. Then processing of some of the attributes provided identifies the fraud detection and provides the graphical model visualization. The performance of the techniques is evaluated based on accuracy, sensitivity, and specificity, and precision.

Keywords: Credit Card, Fraud Detection, Random Forest.

1. INTRODUCTION

There are various fraudulent activities detection techniques has implemented in credit card transactions have been kept in researcher minds to methods to develop models based on artificial intelligence, data mining, fuzzy logic and machine learning. Credit card fraud detection is significantly difficult, but also popular problem to solve. In our proposed system we built the credit card fraud detection using Machine learning. With the advancement of machine learning techniques. Machine learning has been identified as a successful measure for fraud detection. A large amount of data is transferred during online transaction processes, resulting in a binary result: genuine or fraudulent. Within the sample fraudulent features, and features are constructed. These are data points namely the age and value of the customer account, as well as the origin of the credit card. There are hundreds of features and each contributes, to varying extents, towards the fraud probability. Note, the level in which each feature contributes to the fraud score is generated by the artificial intelligence of the machine which is driven by the training set, but is not determined by a fraud analyst. So, in regards to the card fraud, if the use of cards to commit fraud is proven to be high, the fraud weighting of a transaction that uses a credit card will be equally so. However, if this were to shrink, the contribution level would parallel. Simply make, these models self-learn without explicit programming such as with manual review. Credit card fraud detection using Machine learning is done by deploying the classification and regression algorithms. We use supervised learning algorithm such as Random forest algorithm to classify the fraud card transaction in online or by offline. Random forest is advanced version of Decision tree. Random forest has better efficiency and accuracy than the other machine learning algorithms. Random forest aims to reduce the previously mentioned correlation issue by picking only a subsample of the feature space at each split. Essentially, it aims to make the trees de-correlated and prune the trees by fixing a stopping criteria for node splits, which I will be cover in more detail later.

1.1 PROBLEM DEFINITION

Billions of dollars of loss are caused every year by the fraudulent credit card transactions. Fraud is old as humanity itself and can take an unlimited variety of different forms. The PwC global economic crime survey of 2017 suggests that approximately 48% of organizations experienced economic crime. Therefore, there is definitely an urge to solve the problem of credit card fraud detection. Moreover, the development of new technologies provides additional ways in which criminals may commit fraud. The use of credit cards is prevalent in modern day society and credit card fraud has been kept on growing in recent years. Hugh Financial losses has been fraudulent affects not only merchants and banks, but also individual person who are using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses. In the short term, may become visible in the long period. For example, if a cardholder is victim of fraud with a certain company, he may no longer trust their business and choose a contender.

1.1 SCOPE OF THE PROJECT

In this proposed project we designed a protocol or a model to detect the fraud activity in credit card transactions. This system is capable of providing most of the essential features required to detect fraudulent and legitimate transactions.
As technology changes, it becomes difficult to track the behaviour and pattern of fraudulent transactions. With the upsurge of machine learning, artificial intelligence and other relevant fields of information technology, it becomes feasible to automate the process and to save some of the effective amount of labor that is put into detecting credit card fraudulent activities.

2. RELATED WORK

[1] The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada. “Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr. Pavol Žavarsky.”

This research paper focuses on the creation of a scorecard from relevant evaluation criteria, features, and capabilities of predictive analytics vendor solutions currently being used to detect credit card fraud. The scorecard provides a side-by-side comparison of five credit card predictive analytics vendor solutions adopted in Canada. From the ensuing research findings, a list of credit card fraud PAT vendor solution challenges, risks, and limitations was outlined.


This paper propose to use two-stage sequence alignment in which a profile Analyser (PA) first determines the similarity of an incoming sequence of transactions on a given credit card with the genuine cardholder’s past spending sequences. The unusual transactions traced by the profile analyser are next passed on to a deviation analyser (DA) for possible alignment with past fraudulent behaviour. The final decision about the nature of a transaction is taken on the basis of the observations by these two analysers. In order to achieve online response time for both PA and DA, we suggest a new approach for combining two sequence alignment algorithms BLAST and SSAHA.

[3] Research on Credit Card Fraud Detection Model Based on Distance Sum.

“Wen-Fang YU, Na Wang”.

Along with increasing credit cards and growing trade volume in China, credit card fraud rises sharply. How to enhance the detection and prevention of credit card fraud becomes the focus of risk control of banks. It proposes a credit card fraud detection model using outlier detection based on distance sum according to the infrequency and unconventionality of fraud in credit card transaction data, applying outlier mining into credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud.


With growing advancement in the electronic commerce field, fraud is spreading all over the world, causing major financial losses. In current scenario, Major cause of financial losses is credit card fraud; it not only affects trades person but also individual clients. Decision tree, Genetic algorithm, Meta learning strategy, neural network, HMM are the presented methods used to detect credit card frauds. In contemplate system for fraudulent detection, artificial intelligence concept of Support Vector Machine (SVM) & decision tree is being used to solve the problem. Thus by implementation of this hybrid approach, financial losses can be reduced to greater extend.


This thesis propose the SVM (Support Vector Machine) based method with multiple kernel involvement which also includes several fields of user profile instead of only spending profile. The simulation result shows improvement in TP (true positive), TN (true negative) rate, & also decreases the FP (false positive) & FN (false negative) rate.


In this study, classification models based on decision trees and support vector machines (SVM) are developed and applied on credit card fraud detection problem. This study is one of the firsts to compare the performance of SVM and decision tree methods in credit card fraud detection with a real data set.

3. SYSTEM ANALYSIS
3.1 EXISTING SYSTEM

In existing System, a research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential. Accuracy of an algorithm is around 50%. Significance of this paper was to find an algorithm and to reduce the cost measure. The result obtained was by 23% and the algorithm they find was Bayes minimum risk.
3.1.1 Disadvantages

1. In this paper a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed.
2. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure.

3.2 PROPOSED SCHEME

In proposed System, we are applying random forest algorithm for classification of the credit card dataset. Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting.

3.3 ADVANTAGES OF PROPOSED SYSTEM

- Random forest ranks the importance of variables in a regression or classification problem in a natural way can be done by Random Forest.
- The ‘amount’ feature is the transaction amount. Feature ‘class’ is the target class for the binary classification and it takes value 1 for positive case (fraud) and 0 for negative case (not fraud).

4. REQUIREMENT SPECIFICATIONS

The requirements specification is a technical specification of requirements for the software products. It is the first step in the requirements analysis process it lists the requirements of a particular software system including functional, performance and security requirements. The purpose of software requirements specification is to provide a detailed overview of the software project, its parameters and goals.

4.1 HARDWARE REQUIREMENTS

- Processor: Intel
- RAM: 4 Gb
- Hard Disk: 260 GB
- Key Board: Standard Windows Keyboard
- Mouse: Two or Three Button Mouse

4.2 SOFTWARE REQUIREMENTS

Python
Anaconda
OS - Windows 7, 8 and 10 (32 and 64 bit)

5. FEASIBILITY STUDY

5.1 TECHNICAL FEASIBILITY

It is evident that necessary hardware and software are available for development and implementation of proposed system

It uses Anaconda.

5.2 ECONOMICAL FEASIBILITY

The cost for the proposed system is comparatively less to other existing software’s.

5.3 OPERATIONAL FEASIBILITY

In this project it requires to configure the necessary software to work on the software.

6. SYSTEM ARCHITECTURE

First the credit card dataset is taken from the source and cleaning and validation is performed on the dataset which includes removal of redundancy, filling empty spaces in columns, converting necessary variable into factors or classes then data is divided into 2 part, one is training dataset and another one is test data set. Now the original sample is randomly partitioned into test and train dataset.

Figure 6.1- ARCHITECTURE OF THE PROPOSED SYSTEM

7. SYSTEM MODULES

7.1 MODULE DESCRIPTION

7.1.1 MODULE 1: DATA COLLECTION

Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know
the target answer. Data for which you already know the
target answer is called labelled data.

7.1.2 MODULE 2: DATA PRE-PROCESSING

Organize your selected data by formatting, cleaning and
sampling from it.

Three common data pre-processing steps are:

Formatting: The data you have selected may not be in a
format that is suitable for you to work with. The data may be
in a relational database and you would like it in a flat file, or
the data may be in a proprietary file format and you would
like it in a relational database or a text file.

Cleaning: Cleaning data is the removal or fixing of missing
data. There may be data instances that are incomplete and
do not carry the data you believe you need to address the
problem. These instances may need to be removed.
Additionally, there may be sensitive information in some of
the attributes and these attributes may need to be removed
from the data entirely.

Sampling: There may be far more selected data available
than you need to work with. More data can result in much
longer running times for algorithms and larger
computational and memory requirements. You can take a
smaller representative sample of the selected data that may
be much faster for exploring and prototyping solutions
before considering the whole dataset.

7.1.3 MODULE 3: FEATURE EXTRATION

Next thing is to do Feature extraction is an attribute
reduction process. Unlike feature selection, which ranks the
existing attributes according to their predictive significance,
feature extraction actually transforms the attributes. The
transformed attributes, or features, are linear combinations
of the original attributes. Finally, our models are trained
using Classifier algorithm. We use classify module on Natural
Language Toolkit library on Python. We use the labelled
dataset gathered. The rest of our labelled data will be used to
evaluate the models. Some machine learning algorithms
were used to classify pre-processed data. The chosen
classifiers were Random forest. These algorithms are very
popular in text classification tasks.

7.1.4 MODULE 4: Evaluation Model

Model Evaluation is an integral part of the model
development process. It helps to find the best model that
represents our data and how well the chosen model will
work in the future. Evaluating model performance with the
data used for training is not acceptable in data science
because it can easily generate overoptimistic and over fitted
models. There are two methods of evaluating models in data
science, Hold-Out and Cross-Validation. To avoid overfitting,
both methods use a test set (not seen by the model) to
evaluate model performance. Performance of each
classification model is estimated base on its averaged. The
result will be in the visualized form. Representation of
classified data in the form of graphs. Accuracy is defined as
the percentage of correct predictions for the test data. It can
be calculated easily by dividing the number of correct
predictions by the number of total predictions.

8. Algorithm Utilized

8.1 Random Forest

Random forest is a type of supervised machine learning
algorithm based on ensemble learning. Ensemble learning is
a type of learning where you join different types of
algorithms or same algorithm multiple times to form a more
powerful prediction model. The random forest algorithm
combines multiple algorithm of the same type i.e. multiple
decision trees, resulting in a forest of trees, hence the name
“Random Forest”. The random forest algorithm can be used
for both regression and classification tasks.

8.2 WORKING OF RANDOM FOREST

The following are the basic steps involved in
performing the random forest algorithm

1. Pick N random records from the dataset.
2. Build a decision tree based on these N records.
3. Choose the number of trees you want in your
algorithm and repeat steps 1 and 2.
4. For classification problem, each tree in the forest
predicts the category to which the new record belongs.
Finally, the new record is assigned to the category that
wins the majority vote.

8.3 ADVANTAGES OF USING RANDOM FOREST

Pros of using random forest for classification and
regression.

1. The random forest algorithm is not biased, since,
there are multiple trees and each tree is trained on a
subset of data. Basically, the random forest algorithm
relies on the power of “the crowd”; therefore, the
overall biasedness of the algorithm is reduced.
2. This algorithm is very stable. Even if a new data
point is introduced in the dataset the overall algorithm
is not affected much since new data may impact one
tree, but it is very hard for it to impact all the trees.
3. The random forest algorithm works well when you
have both categorical and numerical features.

The random forest algorithm also works well when data has
missing values or it has not been scaled well.
9. APPENDICES

9.1 SAMPLE SCREENSHOTS FROM THE PROJECT

![Sample Screenshots](image)

Fig-8.1: Exact figures of fake and original credit card

10. CONCLUSION

The Random forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more pre-processing techniques would also help. The SVM algorithm still suffers from the imbalanced dataset problem and requires more preprocessing to give better results at the results shown by SVM is great but it could have been better if more preprocessing have been done on the data.

11. REFERENCES


