International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April, 2020

Credit Card Fraud Detection and Prevention using Machine Learning

S. Abinayaa, H. Sangeetha, R. A. Karthikeyan, K. Saran Sriram, D. Piyush

Abstract: ***This research focused mainly on detecting credit card fraud in real world. We must collect the credit card data sets initially for qualified data set. Then provide queries on the user's credit card to test the data set. After random forest algorithm classification method using the already evaluated data set and providing current data set[1]. Finally, the accuracy of the results data is optimised. Then the processing of a number of attributes will be implemented, so that affecting fraud detection can be found in viewing the representation of the graphical model. The techniques efficiency is measured based on accuracy, flexibility, and specificity, precision. The results obtained with the use of the Random Forest Algorithm have proved much more effective.***

Keywords: Accuracy, Fraud Detection, Precision, Random Forest Algorithm, Sensitivity.

I. INTRODUCTION

Risk assessment is widely used at banks around the globe. Because credit risk assessment is very important, risk rates are evaluated using a variety of techniques. Banks group clients by profile. During assessment the financial history of clients and subjective consumer considerations are evaluated. Those figures are objective, which reflect the financial statements of the company. Detection of fraud involves monitoring and analysing the behaviour of different users in order to estimate detection unwanted behaviour. To effectively detect credit card fraud, we want to know the diverse technologies, algorithms and types involved in detecting credit card fraud. There are various algorithms to detect the credit card fraud and each have their own advantages and accuracy the algorithms are:-K-nearest neighbour, Linear regression, Ada Boost, Naive Bayes, J48, Logistic Regression, Random Forest algorithm etc. The null hypothesis is the credit card transaction is correct and not fraud. Hence false positive is whether it is a correct and genuine transaction and therefore the system model predicts it as fraud transaction and raises a warning .This means completely normal customers looking to form a sale would deter faraway from making purchases. False negative is a serious issue as the transaction is fraudulent and the system model predicts it as non-fraudulent. In our case, a false negative is far more serious than false positive as our system model would prove costly if it predicts fraudulent transactions as genuine

Revised Manuscript Received on April 25, 2020.
* Correspondence Author
S. Abinayaa*, Assistant Professor, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, Chennai-600087
H. Sangeetha, Assistant Professor, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, Chennai-600087
R. A. Karthikeyan, B. Tech, Information Technology, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, Chennai-600087
K. Saran Sriram, B. Tech, Information Technology, SRM Institute of Science and Technology, Ramapuram Tamilnadu, Chennai-600087
D. Piyush, B. Tech Information Technology, SRM Institute of Science and Technology, Ramapuram Tamilnadu, Chennai-600087

II. ACTUAL SYSTEM

In the existing system, a review of a contextual investigation including the identification of Credit Card misrepresentation where information standardization is applied prior to cluster analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Network on the discovery of extortion has indicated that neuronal data sources may be limited by bundling properties. What's more, encouraging outcomes can be gotten by utilizing standardized information and information ought to be MLP prepared. This examination depended on solo learning. Estimate accuracy is around half. Noteworthiness of this paper was discovering an estimate and reducing the measure of costs. The result was 23% and the calculation they found was the minimum chance of Bayes[3].In this system a collective replacement comparison measure is proposed that represents profits and losses due to fraud detection. Using the existing cost measure, a cost-sensitive method that depends on the Bayes minimum risk is used.

III. PLANNED SYSTEM

We use random forest algorithm in proposed system to classify the credit card data set. Random Forest is a Classification and Regression algorithm. Irregular words have an advantage over the choice tree, as they adjust the propensity to over fit to their set of preparations. A subset of the preparation set is evaluated randomly so that each node at that point parts on an element are chosen from a random subset of the full list of capabilities to prepare each individual tree and then a choice tree is constructed. In any case, it is incredibly fast to prepare for huge information collections with numerous highlights and information events in random forests in events of the fact that each tree is prepared freely of the others[4]. In a natural way Random Forest can do in a problem of regression or classification, ranks the value of variables. Function class is that the conditional classification target class takes value 1 for positive (fraud) cases and value 0 for negative (non-fraud) cases.

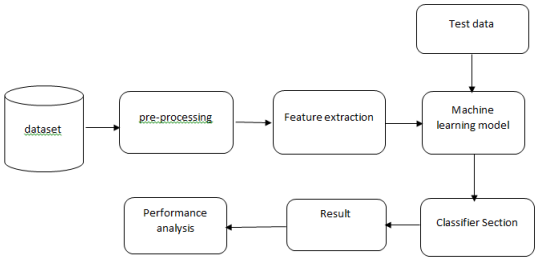

Fig 1:-Architecture Diagram

IV. METHODOLOGY

In credit card transactions, various fraudulent activity detection techniques have been implemented in the minds of researchers to methods for developing models based on artificial intelligence, data mining, fuzzy logic, and machine learning. Apps are installed within fraudulent sample data sets. These data points, include customers name, customers age and value of the customer account, and origin of the credit card. So, with regard to card fraud, if the usage of cards to commit fraud are proven to be more, the fraud of a transaction using a credit card will be equally so, but if this were to decrease, the level of contribution would be equal. Detection and prevention of credit card fraud using Machine Learning is accomplished by using the classification and regression algorithms. We use supervised machine learning algorithms such as Random Forest Algorithms to detect online or offline fraud card transactions. [5].

V.ALGORITHM

A. RANDOM FOREST ALGORITHM

Random Forest algorithm is a machine learning based algorithm that combines multiple decision trees together for obtaining efficient outcome. Decision trees are created by random forest algorithm based on data samples and selects the best solution by means of voting. [6]

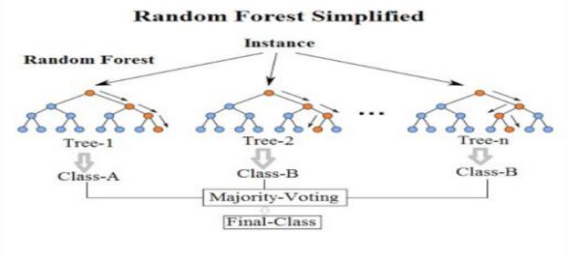

Fig 2:-Random Forest Simplification

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2$$

Where N is the number of data points,
fi is the value returned by the model and
yi is the actual value for data point i.

Fig 3:-Formula for Random Forest Algorithm

VI.MODULES IDENTIFIED

A. DATA COLLECTION

Ideally, ML problems start with data, lots of data (examples or observations) which you already know about the target response[7]. Data that you already know the target answer is named labelled data.

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24
0	51435.0	1.197490	-0.352125	-0.135904	0.222100	0.231128	1.086617	-0.420363	0.391464	0.672499	...	-0.337999	-0.963821	-0.121931	-1.723430
1	70049.0	0.976047	-0.269947	1.465321	1.300002	-1.382887	-0.479586	-0.632572	0.084533	0.710743	...	0.322829	0.790185	-0.101364	0.730461
2	157168.0	-1.395302	0.470266	-0.584911	-1.201527	0.926544	-0.743618	0.755504	-0.141397	-2.118499	...	0.282803	0.903183	-0.444694	0.696438
3	69297.0	1.276014	-0.672705	-0.425494	-0.777398	-0.582088	-0.880396	-0.103505	-0.203036	-1.241653	...	0.256003	0.408211	-0.274815	0.028707
4	144504.0	-0.312745	-1.202565	2.248806	-0.297210	-0.963388	1.207532	-0.837776	-0.057654	1.121421	...	-0.274586	0.662305	0.432717	0.722364

Fig 4:-Data Collected From Bank Server.

B. DATA PREPROCESSING

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

Three common data per-processing steps are:

• **Formatting**: You may not have chosen the details in a format that suits you for working with. The data may also be in in an electronic database and you would like it to be in a spreadsheet, or the information may be in a proprietary file format and you would like it to be in an electronic database or folder.

• **Cleaning:** Cleaning data is the eradication or restoration of unfinished or empty data. There may also be incomplete occurrences of data which do not carry the information that you think you'd like to lever may need to eliminate these occurrences. In addition, there are attributes which carry sensitive information and that the attributes are likely to be omitted.

• **Sampling:** In sampling, there can be more data set selected than required to work with.

Execution of More data can result in greater computational and memory requirements. A smaller ranked sample of the selected data can be taken for consideration, which will be much quicker to discover and model the solutions before considering the entire data set [8].

```
from sklearn.preprocessing import StandardScaler
data['normalizedAmount'] = StandardScaler().fit_transform(data['Amount'].values.reshape(-1,1))
data = data.drop(['Amount'],axis=1)
```

Fig 5:-Code for Data Pre- processing

C. RANDOM FOREST

```
random_forest = RandomForestClassifier(n_estimators=100)
```

```
random_forest.fit(X_train,y_train.values.ravel())
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

```
y_pred = random_forest.predict(X_test)
```

```
random_forest.score(X_test,y_test)
```

```
0.9994002007383512
```

```
cnf_matrix = confusion_matrix(y_test,y_pred)
```

```
labels = [0,1]
sns.heatmap(cnf_matrix, annot=True, cmap="YlGnBu", fmt=".3f", xticklabels=labels, yticklabels=labels)
plt.show()
```

Fig 6:-Code for Random Forest Algorithm.

Retrieval Number: D7327049420/2020©BEIESP
DOI: 10.35940/ijeat.D7327.049420
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. RESULT ANALYSIS

The system points out the amount of false positives it finds and compares them to actual data sets. This is used for measuring the precision and accuracy for the algorithms[10]. The data fraction that we used for faster, more efficient testing is 10% of the entire set of data. The full set of data is also used at the end, and both reports are written out. These results are given as follows in the output, along with the classification report for each algorithm, where class 0 means the transaction has been determined to be valid and 1 means it has been determined to be fraudulent. This result matched the category values to show for false positives.

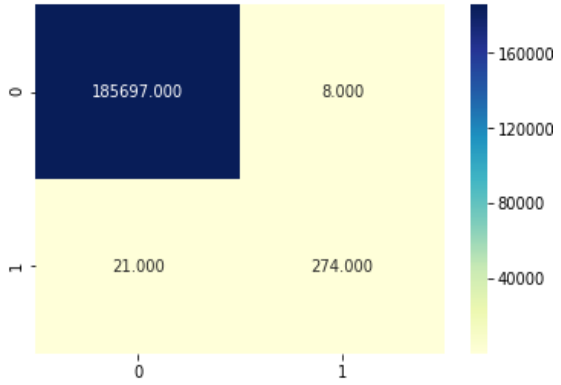

Fig 7:-Result and Comparison between Data sets

VIII. CONCLUSION

Credit card fraud is a criminal act of dishonesty. This article points the most common methods of fraud alongside their detection methods and algorithms, and examined recent findings in this area. This paper also explained in detail how machine learning can be implemented in combination with the random forest algorithm, pseudo code, in order to obtain better results in the detection of fraud. While the algorithm reaches over 60% accuracy, its precision only remains at 28% when considering a tenth of the info set[9]. When the entire data set is fed into the algorithm, however, the precision increases to 33%.

FUTURE ENHANCEMENTS

Although we did not meet the target of 100 percent precision in identifying fraud, we ended up building a program with enough time and resources that could come really similar to that goal.The very design of this project allows the incorporation of multiple algorithms as modules, and the combination of their results can increase the accuracy of the end result. [12].

REFERENCES

1. https://www.ijitee.org/wpcontent/uploads/papers/v8i12S/L102810812S19.pdf.
2. https://www.researchgate.net/publication/336800562_Credit_Card_Fraud_Detection_using_Machine_Learning_and_Data_Science
3. https://ieeexplore.ieee.org/document/8717766
4. https://pdfs.semanticscholar.org/6f4a/a57eb9335f6e2658c78a7a2264e779a09307.pdf
5. http://www.ijesrt.com/issues%20pdf%20file/Archive-2019/March-2019/26.pdf
6. https://www.ijrte.org/wpcontent/uploads/papers/v7i6s4/F10440476S419.pdf
7. https://www.ijert.org/credit-card-fraud-detection-using-machine-learning-and-data-science
8. https://www.ijeat.org/wpcontent/uploads/papers/v8i6S/F11640886S19.pdf
9. https://www.ijitee.org/wpcontent/uploads/papers/v8i12S/L102810812S19.pdf
10. http://www.ijirset.com/upload/2019/april/107_Credit.pdf
11. https://www.irjet.net/archives/V6/i3/IRJET-V6I3710.pdf
12. http://www.iosrjournals.org/iosrjce/papers/Vol21-issue3/Series-5/H2103054552.pdf

AUTHORS PROFILE

S. Abinayaa
Assistant Professor, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, Chennai-600087

H. Sangeetha
Assistant Professor, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, Chennai-600087

R. A. Karthikeyan
B. Tech Information Technology, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, Chennai-600087

K. Saran Sriram
B. Tech Information Technology, SRM Institute of Science and Technology, Ramapuram Tamilnadu, Chennai-600087

D. Piyush
B. Tech Information Technology, SRM Institute of Science and Technology, Ramapuram Tamilnadu, Chennai-600087