CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Dr. Swati A. Bhavsar[1], Shital V. Avhad[2] , Nilam B. Avhad[3], Pooja K. Palve[4] , Komal D. Sanap[5]

Department of Computer Engineering, Matoshri College of Engineering and Research Centre, Nashik-422105

---***---

Abstract - Electronic trade or web based business is a plan of action that lets organizations and people over the web trade anything. As of late, in the age of the Internet and sending to E-business, parts of information are put away and moved starting with one area then onto the next. Information that moved can be presented to risk by fraudsters. There is an enormous expansion in misrepresentation which is prompting the deficiency of a huge number of dollars worldwide consistently. There are different current methods of distinguishing extortion that is consistently proposed and applied to a few business fields. The primary undertaking of Fraud identification is to notice the activities of huge loads of clients to recognize undesirable conduct. To recognize these different sorts, information mining strategies and AI to have been proposed and carried out to decrease down the assaults. A quite some time ago, numerous strategies are used for misrepresentation discovery framework like Support Vector Machine (SVM), K-closest Neighbor (KNN), neural organizations (NN), Fuzzy Logic, Decision Trees, and numerous more. This large number of methods have yielded respectable outcomes yet expecting to further develop the precision even further, by fostering the actual strategies or by utilizing a crossover learning approach for distinguishing cheats..

***Key Words*:** Monitoring, Credit Card, Authentication, security

INTRODUCTION

Presently in the innovation of days because of quick turn of events web utilization is all over the place. In today‟s development electronic world, numerous little and enormous organizations have put their organizations on to the WWW to offer types of assistance to client. Internet business draws on innovations like electronic asset move, online exchange handling, web banking, and mechanized information assortment frameworks, etc. Internet shopping will be well known step by step. Online business installment frameworks have become well known because of far and wide utilization of the web based shopping and banking. Quick augmentation of this time billions of dollars are lost consistently because of Visa misrepresentation. Misrepresentation is a demonstration of treachery planned for individual utilization or on the other hand to hurt a misfortune to somebody. Fraudster just needs to know the individual data identified with (card number, card expiry date and so on) It tends to be conceivable genuinely or for all intents and purposes. It is regularly comprehended as deceptive nature to acquire some benefit which is frequently monetary, over someone else. It very well may be seen in generally normal, securing or exchanging of property, including genuine property, Personal Property, and immaterial property, for example,stocks, bonds, and copyrights.[2].

LITRATURE SURVEY

1. Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison, Samidha Khatri [1], In today's economic scenario, credit card use has become extremely commonplace. These cards allow the user to make payments of large sums of money without the need to carry large sums of cash. They have revolutionized the way of making cashless payments and made making any sort of payments convenient for the buyer. This electronic form of payment is extremely useful but comes with its own set of risks. With the increasing number of users, credit card frauds are also increasing at a similar pace. The credit card information of a particular individual can be collected illegally and can be used for fraudulent transactions. Some Machine

Learning Algorithms can be applied to collect data to tackle this problem. This paper presents a comparison of some established supervised learning algorithms to differentiate between genuine and fraudulent transactions

2. Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection, Vinod Jain [2], Credit cards are very commonly used in making online payments. In recent years' frauds are reported which are accomplished using credit cards. It is very difficult to detect and prevent the fraud which is accomplished using credit card. Machine Learning(ML) is an Artificial Intelligence (AI) technique which is used to solve many problems in science and engineering. In this paper, machine learning algorithms are applied on a data set of credit cards frauds and the power of three machine learning algorithms is compared to detect the frauds accomplished using credit cards. The accuracy of Random Forest machine learning algorithm is best as compared to Decision Tree and XGBOOST algorithms.

3. A Survey On Fraud Detection Techniques in E-Commerce, Suha Mohcen Najem [3], Electronic commerce or e-commerce is a business model that lets companies and persons over the internet buy and sell anything. Recently, in the age of the Internet and forwarding to E-commerce, lots of data are stored and transferred from one location to another. Data that transferred can be exposed to danger by fraudsters. There is a massive increase in fraud which is leading to the loss of many billions of dollars worldwide every year. There are various modern ways of detecting fraud that is regularly proposed and applied to several business fields. The main task of Fraud detection is to observe the actions of tons of users to detect unwanted behavior. To detect these various kinds, data mining methods & machine learning to have been proposed and implemented to lessen down the attacks A long time ago, many methods are utilized for fraud detection system such as Support Vector Machine (SVM), K-nearest Neighbor (KNN), neural networks (NN), Fuzzy Logic, Decision Trees, and many more. All these techniques have yielded decent results but still needing to improve the accuracy even further, by developing the techniques themselves or by using a hybrid learning approach for detecting frauds.

4. Tharindu Madushan Bandara; Wanninayaka Mudiyanselage; Mansoor Raza [4] , In this paper, a review to describe the latest studies on fraud detection in e-commerce between (2018-2020), and a general analysis of the results-achieved and upcoming challenges for further researches. This will be useful for giving us complete visualization about how can we present the most suitable, most accurate methods for fraud detection in e-commerce transactions.

PROBLEM STATEMENT:

To design and develop syestem for credit card fraud detection that can accurately identify fraudulent transactions from a large dataset of credit card transaction. The model should be able to analyze various features of the transaction, such as the transaction amount, location, time, and other realted information to detect patterns that are indicative of fraud.

The goal is to create highly accurate fraud detection system that can help financial institutions prevent losses due to fradulent transactions while minimizing false positive the can impact legitimate transactions. The model should be able to adapt to new patterns of fradulent behavior and maintain high accuracy levels over time.

MOTIVATION

The main Motive of implementing this system is current methods of distinguishing extortion that is consistently proposed and applied to a few business fields. The primary undertaking of Fraud identification is to notice the activities of huge loads of clients to recognize undesirable conduct.

OBJECTIVES

The objectives of the system are

- To implement the discussion module for better understanding of problem.
- To implement fraud detection system for normal user.
- To test and validate the results

ADVANTAGES

1. Innovative.
3. Centralised Database.
4. Easy to use.
5. Efficient cost.

APPLICATION:

1. Ecommerce
2. Personals
3. Organizations

PROPOSED SYSTEM :

Card payments are always different when compared to former payments made by the client. This creates a problem called conception drift. Concept drift can be said as a variable which changes over time and in unlooked-for ways. These variables create a high imbalance in data. The main agenda of our exploration is to overcome the problem of Concept drift to apply on real- world script. In our proposed system we will be using machine learning algorithm like Random Forest Classifier and calculate their accuracy scores. We will also calculate the confusion matrix for this algorithm and take that into consideration along with the accuracy score to the best algorithm. Also, we need to consider the fact that our data set that we are about to look at is very much imbalanced.

DATASET:

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

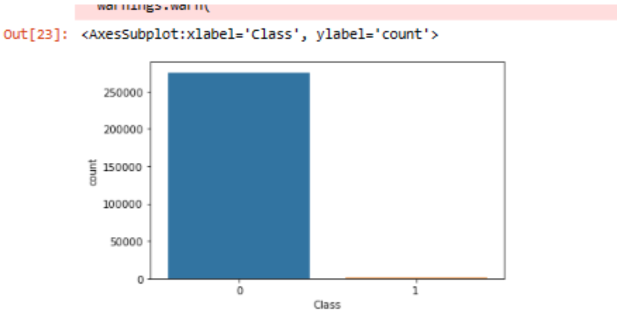
STEPS AND IMPLEMENTATION

Steps to develop the Classifier in Machine Learning

- Complete the Exploratory Data Analysis on the dataset
- Apply ML algorithm on our dataset
- Train and evaluate the model and the best one

Step 1. Complete the Exploratory Data Analysis on the dataset

First, we will import the required modules, load the dataset, and perform EDA on it. Then we will make sure there are no null values in our dataset. The feature that we will be focusing is “Amount”.
Now, if we traverse the existence of each class tag and plot the data using matplotlib the plot will be as follows

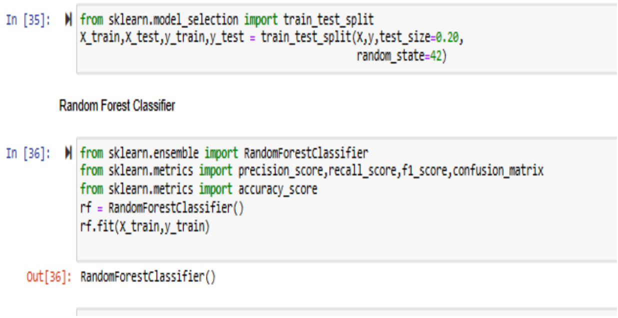


```
    warnings.warn(
Out[23]: <AxesSubplot:xlabel='Class', ylabel='count'>
```

We can observe from the above bar graph that the normal transactions are over 99%. So, to avoid this problem we can apply the scaling techniques on the "Amount" feature to transform them to the range of values. We will remove the "Amount" column and add a new column with the scaled values in its place. We will also remove the "Time" column as it is not required.

Step 2: Use ML Algorithm to the Dataset

Let's use the Random Forest Classifiers which are present in the sklearn package as RandomForestClassifier() respectively.

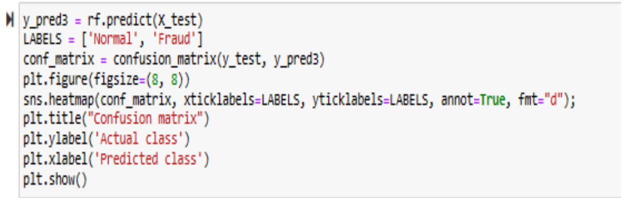


```
In [35]:  from sklearn.model_selection import train_test_split
          X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.20,
                                                           random_state=42)

Random Forest Classifier

In [36]:  from sklearn.ensemble import RandomForestClassifier
          from sklearn.metrics import precision_score,recall_score,f1_score,confusion_matrix
          from sklearn.metrics import accuracy_score
          rf = RandomForestClassifier()
          rf.fit(X_train,y_train)

Out[36]: RandomForestClassifier()
```

Step 3: Train and Evaluate the Model

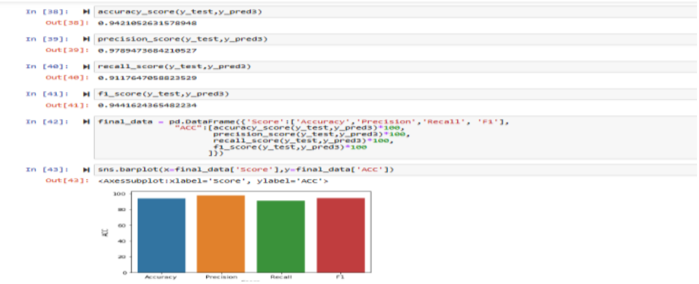
Now, Let's train and evaluate the recently created model and the best one. Train the random forest model using the fit() function.

The Random Forest classifier has somewhat an advantage over the Decision Tree classifier. Now we will calculate the accuracy, precision, recall, and f1-score for both of the classifiers by creating a function commonly used to calculate these values



```
y_pred3 = rf.predict(X_test)
LABELS = ['Normal', 'Fraud']
conf_matrix = confusion_matrix(y_test, y_pred3)
plt.figure(figsize=(8, 8))
sns.heatmap(conf_matrix, xticklabels=LABELS, yticklabels=LABELS, annot=True, fmt="d");
plt.title("Confusion matrix")
plt.ylabel('Actual class')
plt.xlabel('Predicted class')
plt.show()
```

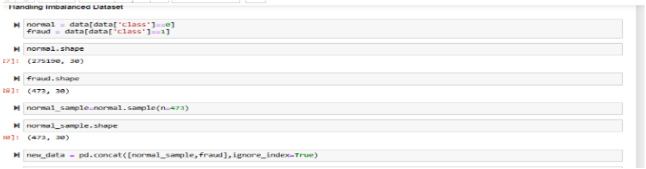
The evaluation metrics of the ***Random Forest model*** will be as follows



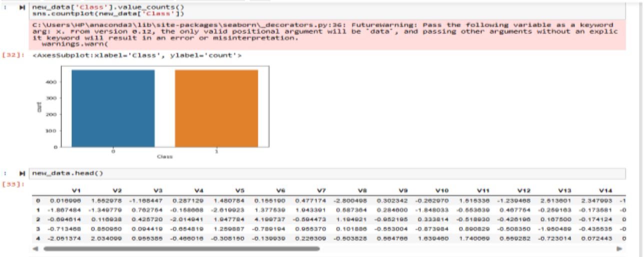
```
In [38]:  accuracy_score(y_test,y_pred3)
Out[38]:  0.9421052631578948

In [39]:  precision_score(y_test,y_pred3)
Out[39]:  0.9789473684210527

In [40]:  recall_score(y_test,y_pred3)
Out[40]:  0.9117647058823529

In [41]:  f1_score(y_test,y_pred3)
Out[41]:  0.9441624365482234

In [42]:  final_data = pd.DataFrame({'Score':['Accuracy','Precision','Recall', 'F1'],
                         "ACC":[accuracy_score(y_test,y_pred3)*100,
                               precision_score(y_test,y_pred3)*100,
                               recall_score(y_test,y_pred3)*100,
                               f1_score(y_test,y_pred3)*100
                              ]})

In [43]:  sns.barplot(x=final_data['Score'],y=final_data['ACC'])
Out[43]:  <AxesSubplot:xlabel='Score', ylabel='ACC'>
```

Address the Class-Imbalance issue

The Random Forest does better than the Decision Trees. But our dataset has a serious problem of class imbalance. The normal transactions are more than 99% and the fraud transactions instituting 0.17%. With such a diffusion, if we train our model without taking care of the imbalance issues, it predicts the data as normal transactions as there is more data about them and hence gets more accuracy even though there are some fraud transactions and these are ignored as there is less data about them. The class imbalance problem can be resolved by many methods.



```
Handling Imbalanced Dataset

     normal = data[data['class']==0]
     fraud = data[data['class']==1]

     normal.shape
27]: (275190, 30)

     fraud.shape
28]: (473, 30)

     normal_sample=normal.sample(n=473)

     normal_sample.shape
30]: (473, 30)

     new_data = pd.concat([normal_sample,fraud],ignore_index=True)
```



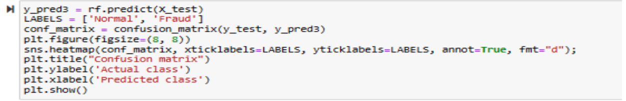
```
:    new_data['Class'].value_counts()
     sns.countplot(new_data['Class'])

     C:\Users\HP\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword
     arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explic
     it keyword will result in an error or misinterpretation.
       warnings.warn(
[32]: <AxesSubplot:xlabel='Class', ylabel='count'>
```

```
:    new_data.head()
[33]:
          V1         V2        V3        V4        V5        V6        V7        V8        V9       V10       V11       V12       V13       V14
 0   0.016996   1.552978 -1.166447  0.287129  1.480784  0.156190  0.477174 -2.800498  0.302342 -0.262970  1.515336 -1.239468  2.513601  2.347993 -1
 1  -1.867484  -1.349779  0.762764 -0.158668 -2.619923  1.377539  1.943391  0.587364  0.284600 -1.848033 -0.553639  0.467764 -0.259163 -0.173581 -0
 2  -0.694614   0.116938  0.425720 -2.014941  1.947784  4.199737 -0.594473  1.194921 -0.952195  0.333814 -0.518930 -0.426196  0.167500 -0.174124  0
 3  -0.713468   0.850950  0.094419 -0.654819  1.259887 -0.789194  0.955370  0.101886 -0.553004 -0.873984  0.890829 -0.508350 -1.950489 -0.436535 -0
 4  -2.051374   2.034099  0.955385 -0.466016 -0.308150 -0.139939  0.228309 -0.503528  0.564786  1.639460  1.740069  0.559282 -0.723014  0.072443  0
```

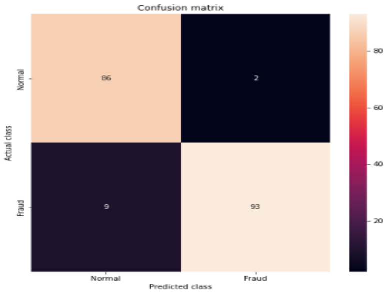
RESULTS AND DISCUSSION

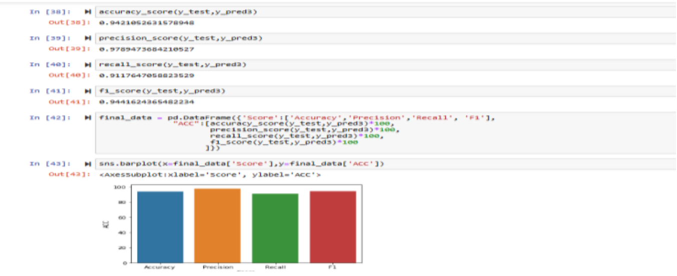
Easily, Random Forest model works better than Decision Trees. But if we observe our dataset suffers a serious problem of class imbalance. The normal (not fraud) deals are further than 99 with the fraud deals constituting of 0.17. With similar kind of distribution, if we train our model without taking care of the imbalance issues, it predicts the label with more significance given to normal deals (as there are more data about them) and hence obtains further fragility. The class imbalance problem can be resolved by reasonable number of ways. Finally, after Handling imbalanced dataset the confusion matrix and the accuracy scores are calculated.

The evaluation metrics for the ***Random Forest model*** are as follows:



```
y_pred3 = rf.predict(X_test)
LABELS = ['Normal', 'Fraud']
conf_matrix = confusion_matrix(y_test, y_pred3)
plt.figure(figsize=(8, 8))
sns.heatmap(conf_matrix, xticklabels=LABELS, yticklabels=LABELS, annot=True, fmt="d");
plt.title("Confusion matrix")
plt.ylabel('Actual class')
plt.xlabel('Predicted class')
plt.show()
```





```
In [38]:  accuracy_score(y_test,y_pred3)
Out[38]:  0.9421052631578948

In [39]:  precision_score(y_test,y_pred3)
Out[39]:  0.9789473684210527

In [40]:  recall_score(y_test,y_pred3)
Out[40]:  0.9117647058823529

In [41]:  f1_score(y_test,y_pred3)
Out[41]:  0.9441624365482234

In [42]:  final_data = pd.DataFrame({'Score':['Accuracy','Precision','Recall', 'F1'],
                         "ACC":[accuracy_score(y_test,y_pred3)*100,
                                precision_score(y_test,y_pred3)*100,
                                recall_score(y_test,y_pred3)*100,
                                f1_score(y_test,y_pred3)*100
                               ]})

In [43]:  sns.barplot(x=final_data['Score'],y=final_data['ACC'])
Out[43]:  <AxesSubplot:xlabel='Score', ylabel='ACC'>
```

As we can see the accuracy scores of the Random Forest model after the handling imbalance dataset which is done to avoid the class imbalance issue, is quite good and better than the different algorithm approaches. So we can say that the Random Forest algorithm does a good job of predicting the anomalies in a huge imbalanced dataset.

CONCLUSION

In conclusion, the credit card fraud detection system using machine learning algorithms is a valuable and effective solution for preventing financial losses due to fraudulent transactions. The project involves the development of a system that can analyze large volumes of transactional data, identify patterns and anomalies that may indicate fraud, and classify transactions as either fraudulent or legitimate. The project typically involves data preprocessing, feature selection, algorithm selection,
and hyperparameter tuning, as well as the development of a user-friendly interface for real-time fraud detection.

REFERENCES

1. Aishwarya Arora; Arun Prakash Agrawal " Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison Samidha Khatri", 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020. https://ieeexplore.ieee.org/document/9057851.
2. T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences", *J. Molecular Biology*, vol. 147, no. 1, pp. 195-197, 2015.
3. C. Chiu and C. Tsai, "A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection", *Proc. IEEE Int'l Conf. e-Technology e-Commerce and e-Service*, pp. 177-181, 2004..
4. V. Vatsa, S. Sural and A.K. Majumdar, "A Game-Theoretic Approach to Credit Card Fraud Detection", *Proc. Int'l Conf. Information Systems Security*, pp. 263-276, 2005
5. J.W. Slocum and H. Lee, "Mathews Social Class and Income as Indicators of Consumer Credit Behavior", *J. Marketing*, vol. 34, no. 2, pp. 69-74, 1970.
6. R. Wheat and D.G. Morrison, "Estimating Purchase Regularity with Two Interpurchase Times", *J. Marketing Research*, vol. 27, no. 1, pp. 87-93, 1990
7. B.E. Kahn and D.C. Schmittlein, "Shopping Trip Behavior: An Empirical.

Investigation", *Marketing Letters*, vol. 1, no. 1, pp. 55-69, 1989.

8. K. Liano and J.T. Lindley, "An Analysis of the Weekend Effect within the Monthly Effect", *Rev. of Quantitative Finance and Accounting*, vol. 5, no. 4, pp. 419-426, 1995.
9. C.H. Joha, H.J.P. Timmermansa and P.T.L. Popkowski-Leszczyc, "Identifying Purchase-History Sensitive Shopper Segments Using Scanner Panel Data and Sequence Alignment Methods", *J. of Retailing and Consumer Services*, vol. 10, no. 3, pp. 135-144, 2003
10. K. Takeda, "The Application of Bioinformatics to Network Intrusion Detection", *Proc. Int'l Carnahan Conf. Security Technology (CCST)*, pp. 130-132, 2005.