e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

STUDY OF MACHINE LEARNING ALGORITHMS FOR CREDIT CARD FRAUD DETECTION

Tanuj Palaspagar[*1], Bhushan Patil[*2], Mitali Mane[*3], Preethi Ambati[*4], Dr. Shaveta Malik[*5]

[*1,2,3,4,5]Department Of Computer Engineering, Terna Engineering College, Nerul University Of Mumbai, India.

ABSTRACT

The biggest reason behind the fast growth of the modern industry is trading through online commerce. Exchanging goods through online services proves to be convenient for the customer and the provider has facilities to reach a wider audience. High dependence on internet technology has posed a threat to financial institutions and their customers. Advancements in recent technology methods are by far the answer to aid in defending against such bank frauds. The manual investigation by industry experts can be used to focus on more complex threats while efficient systems can be put in place to detect unusual transactions in bank account behaviors. Detecting Credit Card fraud is a Data Mining problem, which is hampered by the fact that normal and fraudulent profiles change constantly and the great dependence on the quality of the dataset, sampling approach on the dataset, and detection technique. In this research, we have explored the implementation of different machine learning techniques that can be used for classifying transactions as fraudulent or genuine. The simulation results show the accuracy and give a comparative study of the methods to detect anomalies.

Keywords: Fraud Detection, Machine Learning, Logistic Regression, Random Forest, Support Vector Machine.

I. INTRODUCTION

With the growth of e-commerce websites, people and financial companies rely on online services to carry out their transactions that have led to an exponential increase in credit card frauds. Fraudulent credit card transactions lead to a loss of a huge amount of money. An effective fraud detection system is necessary to recognize and reduce the losses incurred by the customers and financial companies. A good practically implemented fraud detection system should be able to identify the fraud transaction accurately and should be able to make the detection possible in real-time transactions. Fraud detection can be divided into two groups: anomaly detection and misuse detection. Anomaly detection systems bring normal transactions to be trained and use techniques to determine novel frauds. A misuse fraud detection system uses the labeled transaction as normal or fraud transaction from the database history to categorize the fraud according to the existing fraud pattern. Hence, this misuse detection system entails a system of supervised learning and an anomaly detection system a system of unsupervised learning. Customers' typical behavior is used by fraudsters to deceive them. Because transaction trends change over time, the fraud detection system must constantly learn and update the fraud patterns. Traditional card-related frauds (application, stolen, account takeover, fake and counterfeit), merchant-related frauds (merchant collusion and triangulation), and Internet frauds (site cloning, credit card generators, and false merchant sites) are the three types of credit card frauds.

The criteria used to assess the authenticity of detection algorithms become crucial in the development of a model that accurately scores fraudulent transactions while taking into consideration case imbalance and the cost of misclassifying a case as genuine when it is not.

Financial fraud can be defined as mendacious actions that lead to the financial gain of the fraudster. The prevalent form of financial fraud is Credit Card Fraud [5]. The main objective of the project is to identify frauds during credit card transactions.

II. RELATED WORKS

Numerous literature pertaining to Credit fraud detection have been published already and are available for public usage and we found these resources notable to our project.

A Survey on Credit Card Fraud Detection on Machine Learning (2019) [8] focused on Regression, classification, Support vector machines, Neural networks, Fuzzy logic-based system, etc. explained existing techniques based on statistics and computation they have concluded that all the present machine learning techniques mentioned

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

provide high accuracy for the detection rate.

A Survey on Credit Card Fraud Detection using Machine Learning (2018) [6] proposed an unsupervised fraud detection method using autoencoder-based clustering. The autoencoder is an auto-associative neural network they have used to lower the dimensionality, extract the useful features, and increase the efficiency of learning in a neural network.

Random forest for credit card fraud detection [18] provides a comparison between Random-tree-based random forest and CART-based random forest. It compares different random forest algorithms to train the behavior features of normal and abnormal transactions; both of the algorithms are different in their base classifications and their performance.

Literature Review of Different Machine Learning Algorithms for Credit Card Fraud Detection (2021) [4] provides an in-detail comparative analysis for SVM, Logistic Regression, Gaussian Naïve Bayes, K-Neighbor classifier, and Random Forest algorithms using previous references.

III. METHODOLOGY

The dataset is obtained in comma-separated values (CSV) format, which contains the values of 'Time' and 'Amount' as unchanged metrics, along with 28 features that may be a result of transformation. The 'Class' attribute is binary-it contains values of 0 or 1, which describe the nature of the transaction, where 1 denotes fraud and 0 is used for genuine transactions [7]. The values of this class hence, are used as labels in the model. To reduce the biasing of the unbalanced data, the data is sampled and pre-processed to fit the necessary parameters for the model. The resulting dataset is split into testing and training datasets, where 80% is used for training the model and the model is tested on the remaining 20% of data. The data is randomized to get a nearly homogenous dataset. We pass all the attributes and target class in two different variables. After these, we form a model of our chosen method in which training data is fitted to train the model. Once the model is ready it predicts the values for our testing data. The model is run in multiple iterations and observations are noted after each one. After analyzing the observations, the outliers are discarded and the median of the observations is considered as the final result. This is to make sure we do not consider any edge cases or special cases. Along with model accuracy and confusion matrix, model attributes of precision, recall, f1-score, and support are taken into consideration while evaluating the algorithm.

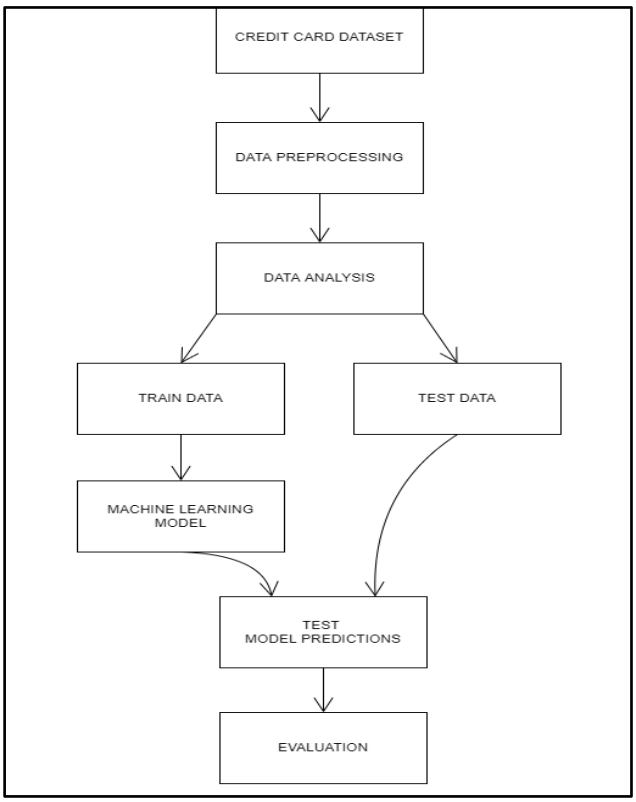

Fig 1: Process Diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

3.1 DATASET

Machine Learning relies heavily on data acquisition. Data collection is the process of obtaining and analyzing data from a variety of sources. To be able to identify some input parameters more correctly, Machine Learning requires a large quantity of data with numerous qualities. The crucial feature of data collecting is that it allows the algorithm to be trained. It has been discovered that having a larger number of qualities leads to a better outcome [17].

Public dataset provided by Kaggle was used for the implementation of this study [7]. This dataset covers transactions done by European cardholders in September 2013 over two days [7].

The dataset contains 31 numerical features. The PCA transformation of these input variables was performed in order to keep the data anonymous since some of the input variables contain financial information. Three of the given features weren't transformed. Feature "Time" shows the time between the first transaction and every other transaction in the dataset. Feature "Amount" is the amount of the transactions made by credit card. Feature "Class" represents the label, and takes only 2 values: value 1 in case of fraud transaction and 0 otherwise [7][2].

```
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64
dtypes: float64(30), int64(1)
```

Fig 2: Dataset columns, entries and data type

3.2 PREPROCESSING

The Dataset contains 284,807 transactions where 492 transactions were frauds and the rest were genuine[7]. Considering the numbers, it is observed that this dataset is highly imbalanced, where only 0.172% of transactions are labeled as frauds.

The use of this imbalanced data causes the model to be heavily biased. To reduce this bias, sampling is a necessity; we have chosen to undersample our data to avoid data duplication. A comparable number of non-fraudulent cases were chosen at random to be trained and tested, along with all of the known fraud transactions.

Since the distribution ratio of classes plays an important role in model accuracy and precision, preprocessing of the data is crucial. The PCA method prohibits fraudulent users from exploiting the credit card information

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

present in this dataset, but it poses difficulty in being vague in feature labeling, which is a result of data masking and dimensionality reduction.

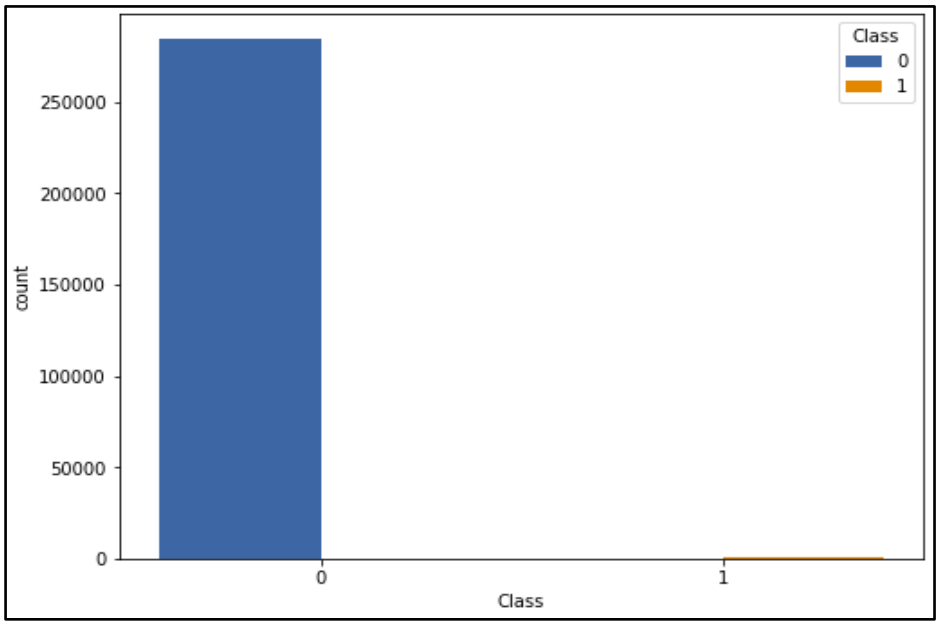

Fig 3: Data Imbalance

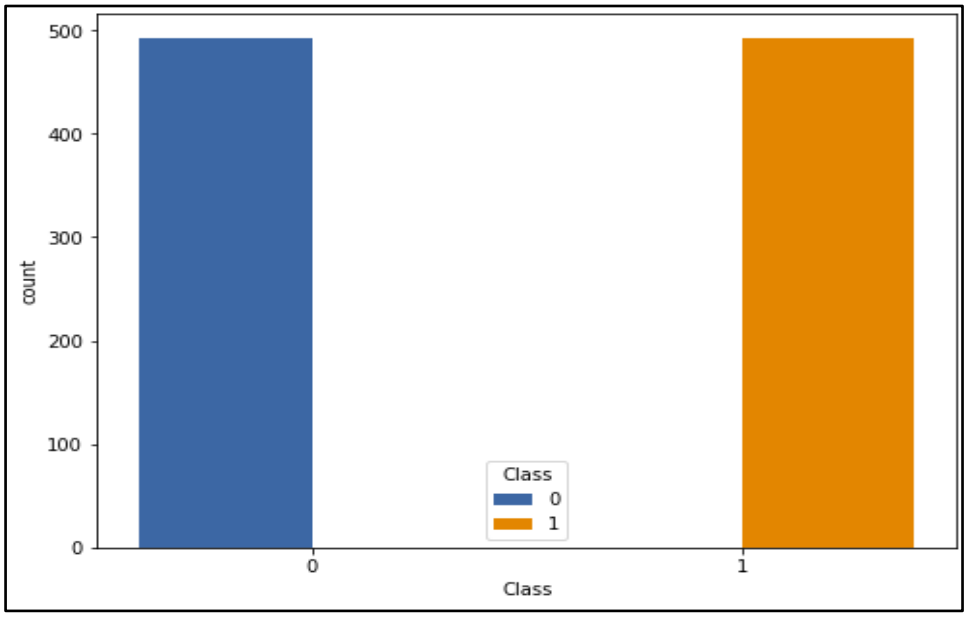

Fig 4: Preprocessed and Balanced Dataset

IV. RESEARCH AND FINDINGS

Evaluation of the algorithms on the dataset is done based on the following performance metrics:

1. Confusion matrix: The confusion matrix provides more knowledge about the performance of our model by providing the information of correctly, incorrectly classified classes through which we can identify errors.
2. Accuracy: Accuracy is the percentage of correctly predicted outputs.
3. Precision: ratio of true positives to the sum of true positives and false positives.
4. Recall: ratio of true positives to the sum of true positives and false negatives.
5. F1 score: Harmonic Mean of Precision and Recall. The closer this number is to 1, the better the performance.
6. Support: Number of actual occurrences of class in the dataset. It does not vary between models.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

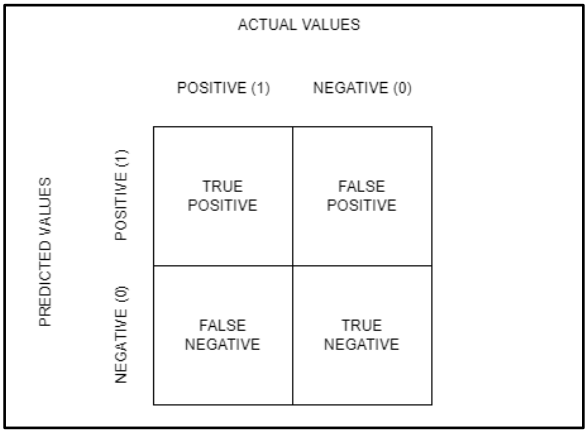

Fig 5: Formation of a generic Confusion Matrix

4.1 LOGISTIC REGRESSION

Logistic regression is one of the most popular classification algorithms in machine learning. The logistic regression model describes relationships between predictors that can be continuous, binary, and categorical.

```
Accuracy on Training data :  0.9990168755074722
Confusion Matrix :
 [[56828    36]
 [   26    72]]
Accuracy of Logistic Regression model on test dataset :
 0.9989115550718023
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.67      0.73      0.70        98

    accuracy                           1.00     56962
   macro avg       0.83      0.87      0.85     56962
weighted avg       1.00      1.00      1.00     56962
```

Fig 6: Logistic Regression implemented on unsampled dataset

```
Accuracy on Training data :  0.9428208386277002
Confusion Matrix :
 [[95  4]
 [10 88]]
Accuracy of Logistic Regression model on test dataset :
 0.9289340101522843
Classification Report :
              precision    recall  f1-score   support

           0       0.90      0.96      0.93        99
           1       0.96      0.90      0.93        98

    accuracy                           0.93       197
   macro avg       0.93      0.93      0.93       197
weighted avg       0.93      0.93      0.93       197
```

Fig 7: Logistic Regression implemented after sampling the dataset

The confusion matrix shows us that for the unsampled data the true positives are 56828 and false positives are 36, the true negatives are 26 and false negatives are 72. For the sampled data the true positives are 95 and false positives are 4, the true negatives are 10 and false negatives are 88.

In our implementation, the dependent variables are binary. Based on some predictors we predict the likelihood of an event occurring. For a given collection of predictors, we calculate the likelihood of belonging to each category [2]. It is most effective when used on data with associated attributes[4].

Logistic Regression uses a functional approach to estimate the probability of a binary response based on one or more variables (features). It finds the best-fit parameters to a nonlinear function called the sigmoid [10].

4.2 RANDOM FOREST

Random forest is an algorithm that can be used in both classification and regression problems. It consists of many decision trees. This algorithm gives better results when there is a higher number of trees in the forest.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

Since it uses bagging and ensemble learning, it prevents model overfitting and improves the accuracy of the model. Each decision tree in the forest gives some results. These results are merged in order to get more accurate and stable predictions [2].

```
Confusion Matrix :
 [[56859     5]
 [   21    77]]
Accuracy of Random Forest model on test dataset :
 0.9995435553526912
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.94      0.79      0.86        98

    accuracy                           1.00     56962
   macro avg       0.97      0.89      0.93     56962
weighted avg       1.00      1.00      1.00     56962
```

Fig 8: Random Forest implemented on unsampled dataset

```
Confusion Matrix :
 [[98  1]
 [11 87]]
Accuracy of Random Forest model on test dataset :
 0.9390862944162437
Classification Report :
              precision    recall  f1-score   support

           0       0.90      0.99      0.94        99
           1       0.99      0.89      0.94        98

    accuracy                           0.94       197
   macro avg       0.94      0.94      0.94       197
weighted avg       0.94      0.94      0.94       197
```

Fig 9: Random Forest implemented after sampling the dataset

The confusion matrix shows us that for the unsampled data the true positives are 56859 and false positives are 5, the true negatives are 21 and false negatives are 77. For the sampled data the true positives are 98 and false positives are 1, the true negatives are 11 and false negatives are 87.

While using Random Forest algorithm, data is not required to be rescaled or transformed. It can be applied to Classification and Regression problems. The algorithm divides the data based on their features and each tree has high variance and low bias that leads to a good result. It trains the model with high speed and is also easy to implement and can handle a good amount of feature loss and errors in the data set [4]. These results are merged to get more accurate and stable predictions.

4.3 K-NEAREST NEIGHBORS

The KNN algorithm handles noisy data remarkably. It's a memory-based technique that allows us to employ both categorization kinds (Binary and Multi) with no additional work. We can also utilize both classification and regression. The initial parameter makes parameter selection difficult but later on, it aligns with the first parameter.

```
Confusion Matrix :
 [[56864     0]
 [   86    12]]
Accuracy of KNN model on test dataset :
 0.9984902215512096
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       1.00      0.12      0.22        98

    accuracy                           1.00     56962
   macro avg       1.00      0.56      0.61     56962
weighted avg       1.00      1.00      1.00     56962
```

Fig 10: KNN implemented on unsampled dataset

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

```
Confusion Matrix :
 [[71 28]
 [40 58]]
Accuracy of KNN model on test dataset :
 0.6548223350253807
Classification Report :
              precision    recall  f1-score   support

           0       0.64      0.72      0.68        99
           1       0.67      0.59      0.63        98

    accuracy                           0.65       197
   macro avg       0.66      0.65      0.65       197
weighted avg       0.66      0.65      0.65       197
```

Fig 11: KNN implemented after sampling the dataset

The confusion matrix shows us that for the unsampled data the true positives are 56864 and false positives are 0, the true negatives are 86 and false negatives are 12. For the sampled data the true positives are 71 and false positives are 28, the true negatives are 40 and false negatives are 58.

The Euclidean distance measure is used in this study for the kNN classifier. For every data point in the dataset, the Euclidean distance between an input data point and the current point is calculated [10].

These distances are sorted in increasing order and k items with the lowest distances to the input data point are selected. The majority class among these items is found and the classifier returns the majority class as the classification for the input point [10].

Since the model does not have to be re-trained for each new data point, the system can be made dynamic; where new data can be added easily. The accuracy score is found to be the least among all the methods. This is due to the fact that the algorithm needs scaled data, and hence it cannot work with high dimensionality, and the algorithm cannot scale to effectively handle all variables.

4.4 SUPPORT VECTOR MACHINE

Support Vector Machine or SVM algorithm used for classification and pattern analysis. It is a classification technique to classify or predict patterns into two classes; fraud or legitimate. This Technique is used for binary classifications. SVM is used in classification as well as in pattern recognition systems. Risk Minimization theory is developed and supported by SVM.

```
Confusion Matrix :
 [[56864     0]
 [   98     0]]
Accuracy of SVM model on test dataset :
 0.9982795547909132
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56864
           1       0.00      0.00      0.00        98

    accuracy                           1.00     56962
   macro avg       0.50      0.50      0.50     56962
weighted avg       1.00      1.00      1.00     56962
```

Fig 12: SVM implemented on unsampled dataset

```
Confusion Matrix :
 [[45 54]
 [29 69]]
Accuracy of SVM model on test dataset :
 0.5786802030456852
Classification Report :
              precision    recall  f1-score   support

           0       0.61      0.45      0.52        99
           1       0.56      0.70      0.62        98

    accuracy                           0.58       197
   macro avg       0.58      0.58      0.57       197
weighted avg       0.58      0.58      0.57       197
```

Fig 13: SVM implemented after sampling the dataset

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

The confusion matrix shows us that for the unsampled data the true positives are 56864 and false positives are 0, the true negatives are 98 and false negatives are 0. For the sampled data the true positives are 45 and false positives are 54, the true negatives are 29 and false negatives are 69.

It is found that when the function dimension is high and the size of data is huge, the performance and expandability are still unsatisfactory. A major reason is that they have to check all the instances of data for each feature to assess the collection of information from all potential splitting positions, which takes a long time. SVM was proposed to address this issue. SVM is a gradient boosting application that uses tree-based learning algorithms. SVM works primarily on the Histogram-based, and at the same time retains relatively accurate results. SVM is usually faster than other gradient boosting algorithms [3].

V. FUTURE ENHANCEMENT

As viewed through our observations, the precision of the algorithms improves as the dataset size grows. As a result, more data will undoubtedly improve the model's accuracy in detecting frauds while lowering the number of false positives.

When it comes to credit card fraud detection, the current system detects the fraud after it has occurred. The existing system stores a vast quantity of data. When a customer notices an inconsistency in a transaction, he or she files a complaint, and the fraud detection system kicks in. It tries to detect whether or not fraud has occurred before moving on to tracking the location of the scam and so on. In the event of the existing system, there is no guarantee of fraud recovery or client satisfaction [6].

VI. PROBLEMS AND LIMITATIONS

This proposal is difficult to execute in real life since it requires collaboration from banks, which aren't eager to exchange information due to their market competitiveness, and also owing to legal concerns and protection of data of their consumers. As a consequence, we searched for some reference publications that used comparable methods and gathered data[16].

It is hard to have some figures on the impact of fraud since companies and banks do not like to disclose the number of losses due to fraud. At the same time, public data is scant due to confidentiality concerns, leaving many questions concerning the ideal technique unanswered.

Another issue with estimating credit-card fraud loss is that we can only evaluate the loss of frauds that have been found; it is impossible to estimate the extent of scams that have gone unreported or unnoticed. Fraud patterns are changing rapidly and fraud detection needs to be re-evaluated from a reactive to a proactive approach.

VII. CONCLUSION

On successful implementation of Logistic Regression, Random Forest, k-Nearest Neighbors, and Support Vector Machine algorithms, we have found Random Forest Classifier gives the best results, although it was the most complex to implement. Every other algorithm has its merits in other metrics, so there is no single best algorithm for this use, although they can be used collectively for predicting different values based on their merits. The contribution of the paper is summarized in the following:

1. Three classifiers based on different machine learning techniques are trained on real-life credit card transaction data and their performances on credit card fraud detection are evaluated and compared based on several relevant metrics.
2. The highly imbalanced dataset is sampled in a hybrid approach where the positive class is oversampled and the negative class under-sampled, achieving two sets of data distributions.
3. The performances of the three classifiers are examined on the two sets of data distributions using accuracy, precision, recall, f1-score, and support metrics.

This study shows the effect of hybrid sampling on the performance of binary classification of imbalanced data. The results from the algorithms vary across different evaluation metrics, but we believe real-world data with no pre-transformation will avail us to consider more known parameters, resulting in a better-trained model, thereby increasing the efficiency in predicting outcomes [1].

VIII. REFERENCES

[1] **Credit Card Fraud Detection using Machine Learning-** by D. Tanouz, R Raja Subramanian, G V Parameswara Reddy, A. Ranjith Kumar of Dept of Computer Science and Engineering Kalasalingam Academy of Research and Education Virudhunagar, TamilNadu, India.

[2] **Credit Card Fraud Detection- Machine Learning methods- by** Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andra Anderla. 2019.

[3] **Implementation of Credit Card Fraud Detection using Support Vector Machine- by** M.Amarender Reddy, Dr. Pravin R Kshirsagar, D. Akshitha, G. Alekya, K. Divya Rosy JES 2021.

[4] **Literature Review of Different Machine Learning Algorithms for Credit Card Fraud Detection-** by Nayan Uchhana, Ravi Ranjan, Shashank Sharma, Deepak Agrawal, Anurag Punde IJITEE 2021.

[5] **Credit Card Fraud Detection using Machine Learning and Deep Learning Techniques-** by Mohammed Azhan and Shazli Meraj IEEE Xplore 2020

[6] **A Survey on Credit Card Fraud Detection using Machine Learning-** by Mohamad Zamini & Gholam Ali Montazer IEEE Xplore 2018.

[7] **Machine Learning Group** - ULB, Kaggle.com (2021). Credit Card Fraud Detection. [online] Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud

[8] **A Survey on Credit Card Fraud Detection using Machine Learning** by Rimpal R. Popat and Jayesh Chaudhary IEEE Xplore 2019.

[9] **Credit Card Fraud Detection Using Machine Learning-** by Ruttala Sailusha, V. Gnaneswar, R. Ramesh G. Ramakoteswara Rao IEEE Xplore 2020.

[10] **Credit Card Fraud Detection Using Machine Learning Techniques-** by John O. Awoyemi, Adebayi O. Adetunmbi, Samuel A. Oluwadare IEEE 2017.

[11] **Credit Card Fraud Detection Techniques: A Review-** by Sonal Mehndiratta and Mr. Kamal Gupta International Journal of Computer Science and Mobile Computing 2019.

[12] **Credit card fraud detection using Machine learning algorithms-** by Andhavarapu Bhanusri K.Ratna Sree Valli, P.Jyothi, G.Varun Sai, R.Rohith Sai Subash Quest Journal 2020.

[13] **Credit Card Fraud Detection System-** by V. Filippov, L. Mukhanov, B.Shchukin 2008

[14] **Credit Card Fraud Detection using Machine Learning Algorithms-** by Varun Kumar KS, Vijaya Kumar V G, Vijay Shankar A, Pratibha K IJERT 2020.

[15] **Credit Card Fraud Detection System-** Pragya Mittal, Rajat Kr. Sharma, Rishabh Rastogi, Varun Kumar 2020.

[16] **Credit Card Fraud Detection using Machine Learning and Data Science-** by S P Maniraj, Aditya Saini, Swarna Deep Sarkar Shadab Ahmed published by IJERT 2019

[17] **Credit Card Fraud Detection Techniques-** by Nikita Shirodkar, Pratikesh Mandrekar, Rohit Shet Mandrekar, Rahul Sakhalkar, K.M. Chaman Kumar, Shailendra Aswale Zeichen Journal 2020

[18] **Random forest for credit card fraud detection-** by Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, Changjun Jiang IEEE Xplore 2018