Fraudulent Activities Detection in E-commerce Websites

Venkat Dinesh Seetha[1], Srikanth Narabathoju[2], Gnaneshwar Bollam[3], Charan Pulipalupula[4], Y. Lakshmi Prasanna[5]

[1,2,3,4]*Student, Dept. of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India*
[5]*Assistant professor, Dept. of Computer Science and Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana, India*

---***---

Abstract - *In e-commerce money is transferred through websites in the form of transaction. As the number of users in e-commerce increases the number of transactions made by the users also increases as well. The chances of the online transaction being fraud also increases. Through using machine learning, detection of fraud in e-commerce can be developed. There are various number of machine learning algorithms such as Decision Trees, Random Forest. Analysis is done on these machine learning algorithms to find a suitable machine learning algorithm. The amount of money processed through transaction by users in e-commerce can be large or small. The chances for the user engaging in fraudulent activities are very high. The fraudulent activities that user can engage are such as Use of stolen credit cards, money laundering, etc. Due to wide spread of e-commerce in last years, there is a rapid increase in the online transactions by many numbers of users. There has also been a huge growth in the percentage of fraudulent transactions. Hence it necessary to develop and apply different techniques that can help in detecting fraud transactions.*

Key Words: **Fraud Detection, K-Nearest Neighbors, Decision Tree, Random Forest, Extreme Gradient Boosting.**

1. INTRODUCTION

The electronic buying and sale of goods via the Internet utilising online services is known as e-commerce. Electronic funds transfers, Mobile commerce, Internet marketing, supply chain management, electronic data interchange (EDI), online transaction processing, automated data gathering systems, and inventory management systems are all examples of electronic commerce technologies. Technological improvements in the semiconductor industry help electronic commerce, which is the dominant industry of the electronics sector.

Even though other services, such as e-mail, are sometimes utilised, e-commerce frequently employs the internet for at least part of the transaction's life cycle. Purchases of products or services are a common e-commerce transaction. E-commerce is made up of three types: online selling, online auctions, and electronic markets. Electronic commerce makes e-commerce possible.

eCommerce fraud is when a criminal or fraudster uses stolen payment information to conduct online transactions without the account owner's knowledge on an eCommerce platform. Purchase fraud is another name for it. It may be accomplished through the use of a fraudulent identity, a stolen credit card, forged cards and information, and false personal and card information, among other methods.

It goes without saying that the growth of the eCommerce industry, as well as the proliferation of payment methods like cards and online payment solutions, is linked to an increase in fraud.

According to the poll, eCommerce fraud has expanded rapidly in recent years, exceeding eCommerce sales by a factor of two. The chargeback rate is increasing at a rate of more than 20% per year. Since FY17, the number of online shopping scams reported to the National Consumer Helpline has nearly doubled, from 977 to 5,620 cases in FY20, bringing the total number of cases to 13,993. There are several grounds for fraud in eCommerce stores, to say the least. As everything gets digital and AI is employed, fraudsters are becoming more intelligent, creating new tactics, and becoming more sophisticated with each passing year. With today's advanced technologies, stealing data and purchasing information is simple. The use of internet aliases makes identifying and apprehending the criminal harder. In comparison, acquiring evidence and prosecuting cases are subject to less time and resource constraints. You must use a high-quality fraud detection and management system and include creative approaches into your firm to combat fraud.

2. LITERATURE REVIEW

Fraud on credit cards is a type of fraud detection that has evolved quickly. The fraud approach is discussed in many studies. The auto-encoder and constrained Boltzmann machine is one of the deep learning studies [1].

Deep learning is being utilised to create a fraud detection model that works similarly to a human neural network, with data being created in various layers that are tied together for the process, starting with the encoder at layer 1 and ending with the hinge decoder at layer 4. The Deep Learning approach is compared to other algorithms such as the Hidden Markov Model by the researcher (HMM) [2].

Machine learning was also used in the identification of credit card fraud [3] Decision Tree algorithms, Neural Networks, Naive Bayes, and Random Forests are all examples of machine learning algorithms. Because it is simple to use, decision trees are commonly employed in fraud detection. A decision tree is a hierarchical or tree-structured prediction model.

Because Naive Bayes is a classification based on statistical and probability approaches, it is utilised in fraud detection credit cards. In real-world situations, Naive Bayes is incredibly quick and accurate. Genetic algorithms are used to decide the number of hidden layer topologies on neural networks for fraud detection on credit cards [4] The genetic algorithm generates the most ideal number of hidden layers with the genetic algorithm [5].

Random forest is also used in credit card fraud detection [6].

Each excellent tree is integrated into one model via Random Forest. A random value of vector is used in Random Forest with the equivalent distributions across all trees, with a maximum depth of each decision tree. [7].

3. METHODOLOGY

3.1 Proposed Method

When it comes to identifying fraud, machine learning is quite effective. A risk team is responsible for avoiding machine learning fraud on any website where you enter your credit card information. The influence of applying multiple techniques to conduct online transaction fraud detection on the online shopping website, as well as the merchant and consumer, is discussed in this study. The top accuracy results from the transaction dataset in e-commerce will be compared using this machine learning approach.

The steps of the system architecture are as follows:

1. Import of Required Packages.
2. Take a look at the data set.
3. The features in the dataset should be normalized.
4. Train/Test Split divides the entire dataset into train and test data sets.
5. Build the model, i.e., train it.
6. Model prediction is used to test the model.
7. System evaluation (accuracy score, F1-score, etc.)

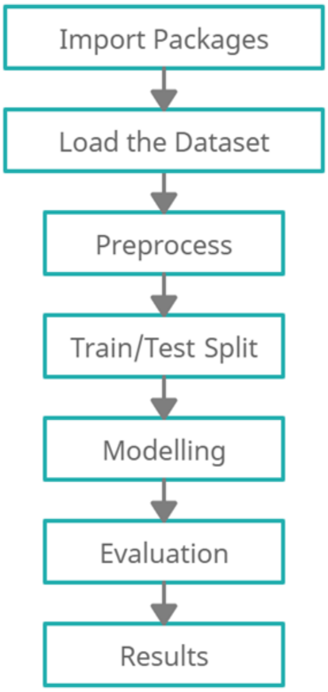

Fig - 1: System Architecture

3.2 K-Nearest Neighbors

1. Neighbors-based categorization is a sort of instance-based learning, also known as non-generalizing learning, because it doesn't try to build a general internal model, instead storing instances of the training data.
2. Classification is determined by a simple majority vote of each point's nearest neighbors: a query point is assigned to the data class with the most representatives among the point's nearest neighbors.
3. A ball tree is a data structure that can be extremely efficient on highly organized data, even when the dimensions are extremely large.
4. A ball tree divides data recursively into nodes defined by a centroid C and radius r, with each node's point falling within the hyper-sphere described by r and C.

3.2.1 Ball Tree

The Ball Tree Algorithm can be described using a metric tree. Metric trees use the metric space in which the points are placed to organise and arrange data points. Points don't have to be finite-dimensional or in vectors when using metrics.

Ball tree algorithm gets its name from the cluster's sphere shape. A node of the tree is represented by each cluster. Let's take a peek at the algorithm in action.

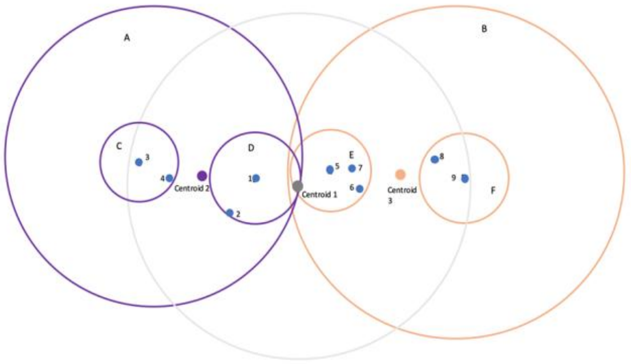

Fig – 2: Ball Tree

At each level of the tree, the offspring are picked to have the greatest possible spacing between them, often using the following architecture.

The centroid of the whole cloud of data points is first determined. The centre of the first cluster and child node is chosen as the site with the greatest distance to the centroid. The second cluster's centre is picked as the point furthest away from the previous cluster's centre. The node is then allocated to all other data points, and the cluster is assigned to the closest centre, either cluster 1 or cluster 2. Each point can only belong to one cluster. Although the sphere lines might cross, the points must be unambiguously assigned to one cluster. A point must be allocated to one cluster if it is exactly in the middle of both centres and has the same distance to both sides. Unbalanced clusters are possible. The Ball Tree Algorithm is based on this fundamental principle. Within each cluster, the procedure of separating the data points into two clusters/spheres is continued until a predetermined depth is attained. As a result, a layered cluster with more and more circles emerges.

3.3 Decision Tree

1. Decision trees are a non-parametric supervised learning approach for classification and regression.

2. The goal is to construct a model using machine learning that can predict the result which is similar to the target variable value by learning basic decision rules from data properties.

3. With the help of decision tress, objects can be classified in binary ([-1, 1]) and multiclass ([0, . . ., K-1]) modes.

4. When unusual behaviours in a transaction from an authorised user need to be classified, decision tree algorithms are used in fraud detection.

5. These algorithms use restrictions that are taught on the dataset to categorise fraud transactions.

Consider a scenario in which a user performs transactions. Based on the transaction, we'll create a decision tree to forecast the likelihood of fraud.

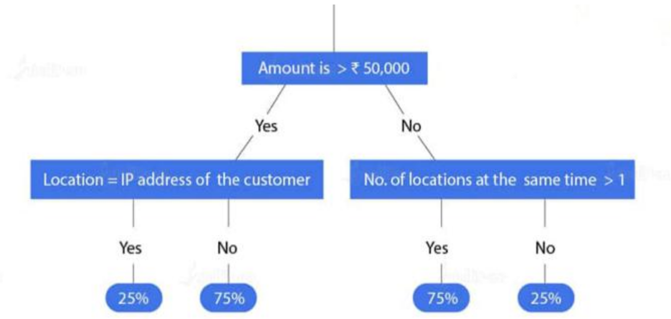
Amount is > ₹ 50,000
Yes
No
Location = IP address of the customer
No. of locations at the same time > 1
Yes
No
Yes
No
25%
75%
75%
25%

Fig – 3: Decision Tree

First, we'll use the decision tree to see if the transaction is higher than $50,000. If the answer is yes, we will investigate the transaction's location. If the answer is no, we'll look into the transaction's frequency. The transaction will then be classified as 'fraud' or 'non-fraud' based on the probabilities obtained for these conditions.

There is only a 25% chance of "fraud" and a 75% chance of "non-fraud" when the amount is greater than 50,000 and the location is equal to the customer's IP address. There is a 75 percent possibility of fraud and a 25% chance of non-fraud if the amount is greater than 50,000 and the number of sites is greater than one. In Machine Learning, a decision tree can help in the creation of fraud detection systems.

3.4 Random Forest

1. A random forest is made up of a collection of basic tree predictors.

2. In random forest, each tree in the ensemble is constructed using a sample selected from the training set with replacement.

3. By fitting numerous decision tree classifiers to distinct sub-samples of the dataset, a Random Forest uses averaging to increase projected accuracy and control over-fitting. Each decision tree evaluates a different set of conditions.

4. To improve the results, Random Forest employs a variety of decision trees. Different conditions are checked by each decision tree.

5. They're trained on random datasets, and depending on the decision trees' training, each tree shows the likelihood of a transaction being 'fraud' or 'non-fraud.' In this way, the model forecasts the outcome.

Consider the following scenario: a transaction is made. Now we'll look at how Machine Learning's random forest is employed in fraud detection methods.

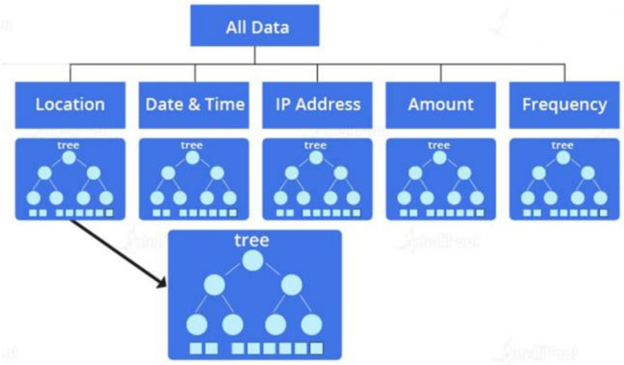

Fig – 4: Random Forest

Whenever the model gets a transaction request, it looks for information such the transaction's location, date, time, IP address, amount, and frequency. As an input, the whole dataset is delivered to the fraud detection algorithm. The fraud detection algorithm then chooses factors from the given dataset to aid in the splitting up of the dataset. The dataset is divided into multiple decision trees in the illustration.

As a result, the sub-trees are made up of variables and the criteria that must be met in order for the variables to be checked for an approved transaction. All the sub-trees will indicate the probability for a transaction being 'fraud' or 'non-fraud' once all the requirements have been checked. The model classifies the transaction as 'fraud' or 'real' based on the combined results. This is how a random forest is used in Machine Learning for fraud detection algorithms.

3.5 Extreme Gradient Boosting

Extreme Gradient Boosting, or XGBoost for short, is a fast implementation of the stochastic gradient boosting machine learning technique. The stochastic gradient boosting algorithm, also known as gradient boosting machines or tree boosting, is a powerful machine learning technique that excels at a variety of difficult machine learning situations. For a wide range of regression and classification predictive modelling applications, the XGBoost method is useful.

It's a fast implementation of the stochastic gradient boosting algorithm with a variety of hyperparameters for fine-grained control of the model training process. Although the technique performs well in general, even on unbalanced classification datasets, it does provide a way to modify the training process to pay greater attention to minority class misclassification in datasets with a skewed class distribution.

This is a supervised learning approach that uses a decision tree-based machine learning algorithm. It's an ensemble approach that aims to build a strong classifier out of poor ones. When we have a large number of observations, we employ this method.

3.6 DATASET

The dataset was split into two files, each containing the following information:

Fraud Data: Data concerning each user's initial transaction.

1. **user id:** The user's identifier. User-specific
2. **signup time:** the time the user signed up for an account (GMT time)
3. **purchase time:** the time when the item was purchased by the user (GMT time)
4. **purchase value:** the price of the item you bought (USD)
5. **device id:** device id is the identifier for the device. It's safe to presume that it's a one-of-a-kind device. Transactions having the same device ID, for example, indicate that the same physical device was used to make the purchase.
6. **user marketing channel:** advertisements, SEO, and direct marketing (i.e., came to the site by directly typing the site address on the browser).
7. browser: The user's browser is referred to as "browser."
8. **sex:** male/female
9. **age:** age of the user
10. **Ip address:** numeric Ip address of the user
11. **class:** we're trying to figure out whether or not the conduct was fraudulent (1) or genuine (0).

IpAddress to Country: assigns a country to each numeric IP address. It shows a range for each country. If the numeric Ip address fits inside the range, it is associated with the appropriate country.

1. **lower bound Ip address:** the numeric Ip address of that country's lower bound.
2. **upper bound Ip address:** the numeric Ip address of that country's upper bound.

3. **country:** the related country's numeric Ip address's numeric Ip address's numeric Ip address's numeric Ip address's numeric Ip address If a user's Ip address falls within the top and lower bounds, she is a resident of this nation.

The dataset used in this paper has a total of 151,112 records. The dataset classified as fraud is 14,151 records. The percentage of fraud data is 0.093 percent.

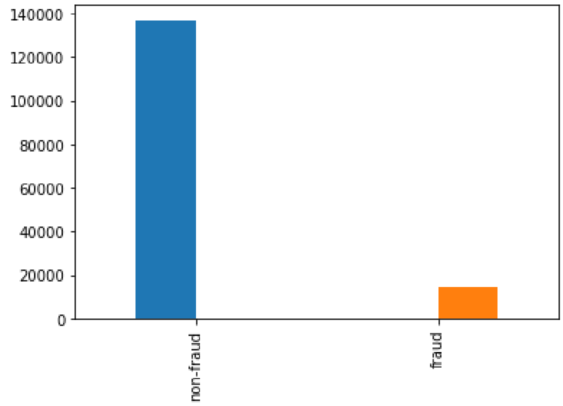

Fig – 5: Fraud and Non-Fraud transactions

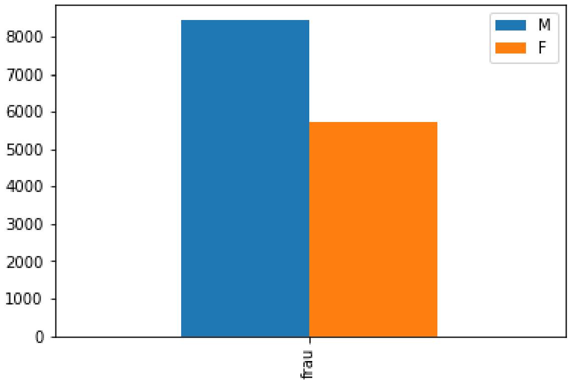

Fig – 6: Gender

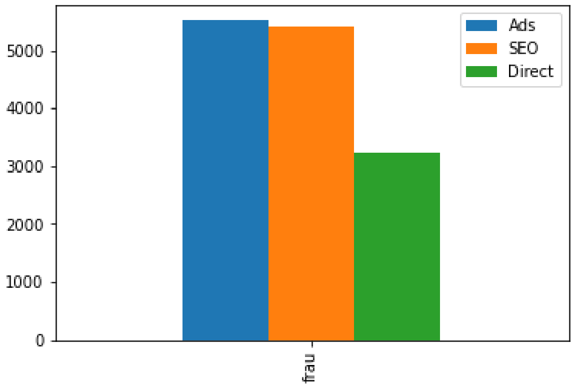

Fig – 7: Marketing Channel

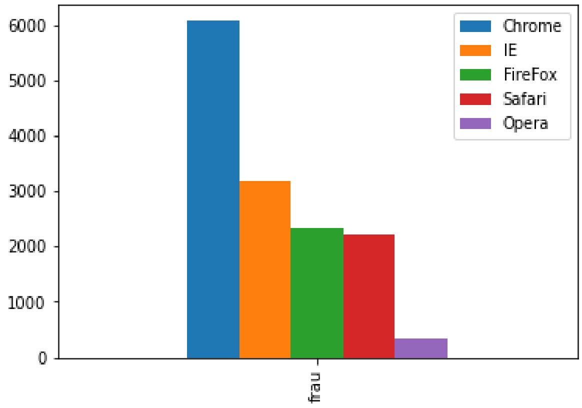

Fig – 8: Browser

Understanding the dataset

To comprehend the dataset, a correlation matrix is employed. If there is little or no association between specific attributes and the desired column, the correlation matrix will show it. It offers a notion of how features are related to one another and can assist in determining which features are more important for our forecast.

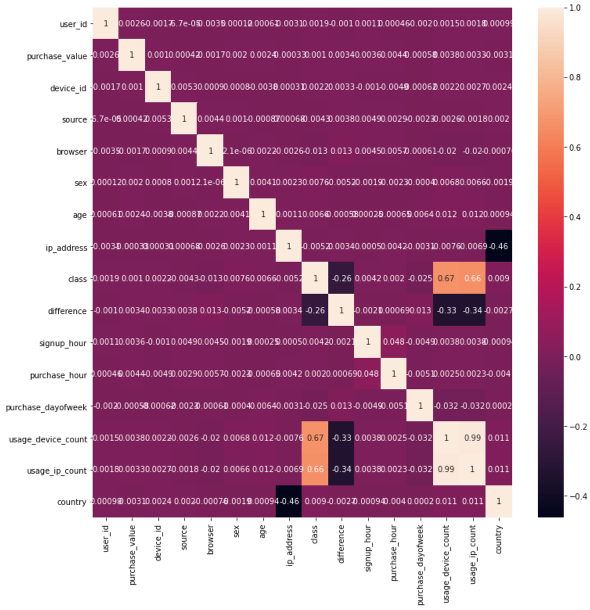

Fig – 9: Correlation

The correlation between the variables on each axis is shown in each square. Correlation might be anywhere between -1 and +1. Closer to 0 suggests that the two variables do not have a linear connection. The closer the correlation is to one, the more positively associated they are; that is, when one rises, so does the other, and the closer to one they are, the stronger the association. A correlation around -1 is similar, except instead of both variables rising, one will drop as the other grows. Because the squares are connecting each variable to itself (thus it's a perfect correlation), the diagonals are all 1/white. For the remainder, the stronger the correlation between the two variables, the larger the number and the lighter the colour.

3.7 PRE-PROCESSING

Pre-processing is a data mining approach for transforming raw data into a format that is both usable and efficient. Pre-processing is the process of extracting, transforming, normalising, and scaling new features for use in the machine learning algorithm process. Pre-processing is the process of converting raw data into a usable format. We map the Ip addresses of the transactions with their respective countries. We also find the difference in the signup time and the purchase time to make a purchase and convert everything in the dataset to numeric values. Save the pre-processed data in a new .csv file.

4. EXPERIMENTAL RESULTS

4.1 Accuracy

Accuracy is calculated by dividing the correctly predicted results with the total number of observations. It refers to the percentage of test samples that are correctly classified. This statistic assesses how close the model's prediction is to the actual data.

The accuracy of each classifier is found by using cross validation score. It is found that Extreme Gradient Boosting shows better accuracy than the other classifiers i.e., KNN, Decision Trees and Random Forest.

Table -1: Accuracy

Model	Accuracy
KNN	76%
Decision Trees	77%
Random Forest	83%
Extreme Gradient Boosting	84%

The below figure shows the graphical representation of the accuracy of each classifier i.e., KNN, Decision Tree, Random Forest, Extreme Gradient Boosting which clearly shows that Extreme Gradient Boosting beats the Random Forest by a close margin. Extreme Gradient Boosting is the best classifier for the fraud detection in ecommerce.

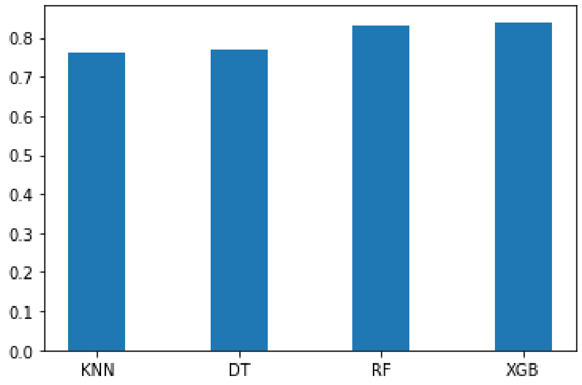

Fig – 10: Accuracy

4.2 ROC-AUC Curve

The ROC-AUC Curve is a performance statistic for identifying problems at different thresholds. The AUC stands for the degree of separation, and it's a probability curve. It expresses the model's ability to differentiate across classes. The AUC measures how effectively the model correctly predicts 0s as 0s and 1s as 1. With TPR on the y-axis and FPR on the x-axis, TPR is plotted versus FPR.

TPR=TP/TP+FN

FPR=1-Precision=FP/TN+FP

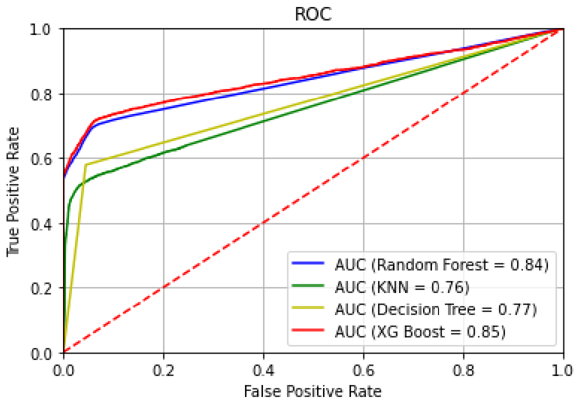

Fig – 11: Performance of models

5. CONCLUSION

In this paper, Fraudulent Activities Detection in E-commerce Websites is studied using transaction dataset. We have trained the dataset on four different classifiers for fraud detection. We found a best approach among those classifiers for detecting fraud activities. We found that Extreme Gradient Boosting gives better results than KNN, Decision Tree and Random Forest. From that, we deduced that Extreme Gradient Boosting is best suited for fraud detection in ecommerce. This paper's main goal was to examine a range of machine learning techniques for detecting fraudulent transactions. The comparison revealed that the XGBoost algorithm delivers the best results, i.e., correctly detects whether transactions are fraudulent or not. Accuracy and the AUC-roc curve, were used to determine this.

REFERENCES

[1] Pumsirirat, Apapan, and Liu Yan. "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine." International Journal of advanced computer science and applications 9.1 (2018): 18-25.

[2] Srivastava, Abhinav, et al. "Credit card fraud detection using hidden Markov model." IEEE Transactions on dependable and secure computing 5.1 (2008): 37-48.

[3] Lakshmi, S. V. S. S., and S. D. Kavilla. "Machine Learning for Credit Card Fraud Detection System." International Journal of Applied Engineering Research 13.24 (2018): 16819-16824.

[4] Aljarah, Ibrahim, Hossam Faris, and Seyedali Mirjalili. "Optimizing connection weights in neural networks using the whale optimization algorithm." Soft Computing 22.1 (2018): 1-15.

[5] Bouktif, Salah, et al. "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches." Energies 11.7 (2018): 1636.

[6] Xuan, Shiyang, Guanjun Liu, and Zhenchuan Li. "Refined weighted random forest and its application to credit card fraud detection." International Conference on Computational Social Networks. Springer, Cham, 2018.

[7] Hong, Haoyuan, et al. "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)." Catena 163 (2018): 399-413.