Impact Factor (SJIF): 6.092 E-ISSN: 2583-1615

International Journal of Research in Academic World

Received: 07/August/2024 **IJRAW:** 2024; 3(9):198-202 **Accepted:** 18/September/2024

Enhancing OTP Fraud Detection in Telangana: A Machine Learning Approach

*[1]V Sumalatha

*[1]Research Scholar, School of Sciences Career Point University, Kota, Rajasthan, India.

Abstract

This paper presents a machine learning-based approach for OTP fraud detection in online transactions. By leveraging transactional data and user behavior patterns, we develop an effective model to identify fraudulent activities. Various machine learning algorithms are evaluated, with Random Forest demonstrating superior performance. The proposed model achieves over 99% accuracy in detecting OTP fraud. Key features contributing to fraud detection include unusual transaction amounts, transaction frequency, and suspicious user behavior. Our findings highlight the effectiveness of machine learning in enhancing security and combating fraudulent activities in online transactions. For financial institutions, implement machine learning-based fraud detection systems, offering training on integration. Online service providers should integrate the model into transaction processing systems and provide guidelines for monitoring suspicious activities. Inform government agencies about the findings to advocate for policies promoting advanced security measures. Share results in academic journals, conferences, and collaborate with researchers to enhance fraud detection.

Keywords: OTP Fraud Detection, Machine Learning, Random Forest, Online Transaction Security, Telangana.

1. Introduction

One-Time Passwords (OTPs) have become a widely adopted security measure in online transactions, providing an additional layer of protection against unauthorized access and fraud. OTPs are temporary codes sent to users via SMS, email, or mobile applications, typically used alongside traditional login credentials. They offer a dynamic authentication method, with each code valid for a single transaction or session, reducing the risk of account compromise due to stolen passwords. In recent years, as online transactions have surged, so too has the sophistication of fraudulent activities. Various forms of fraud, such as account takeover, identity theft, and phishing attacks, pose significant threats to individuals and organizations. Account takeover involves unauthorized access to user accounts, often through stolen credentials or social engineering techniques. Identity theft occurs when personal information is stolen and used to impersonate individuals for fraudulent purposes. Phishing attacks trick users into revealing sensitive information, such as passwords or OTPs, through deceptive emails, websites, or messages. To combat these threats, advanced fraud detection methods are essential, utilizing machine learning, data analytics, and behavioral analysis to identify and prevent fraudulent activities in real- time.

The historical monthly data on OTP fraud is gathered from the National Bureau of Cyber Security (NBC) spanning from January 2020 to December 2023.

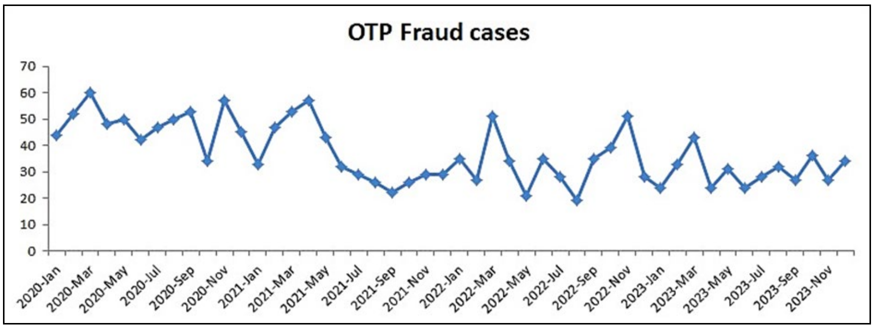

Fig 1: Time series plot on OTP Fraud cases in Telangana

*Corresponding Author: V Sumalatha

Based on the data depicted in Figure 1, it's evident that the timeframe spans from January 2020 to December 2023, providing insights into the frequency of fraud incidents during this duration. Notably, March 2023 stands out with the lowest number of fraud occurrences. A discernible pattern emerges from the data, showcasing a consistent decline in OTP fraud cases each month. This decrease is likely influenced by ongoing awareness campaigns aimed at educating people about cyber threats and advocating for best practices in safeguarding personal and financial information. These educational endeavors are presumed to empower individuals to identify and mitigate potential risks, thereby contributing to the overall reduction in OTP fraud incidents observed throughout the analyzed period.

2. Material and Methods

In Machine Learning algorithm for continue the model development there is some steps. Machine learning methods have revolutionized various aspects of academic writing, particularly in producing papers efficiently. One such method is using Natural Language Processing (NLP) models, such as GPT (Generative Pre-trained Transformer) models. Machine learning methods, particularly Natural Language Processing (NLP), have transformed academic writing processes. NLP models, like GPT (Generative Pre-trained Transformer) models, utilize large datasets to learn the nuances of human language and generate coherent text.

These models can aid in various stages of paper writing, from brainstorming ideas to drafting and editing. One significant application of machine learning in academic writing is automatic summarization. Machine learning algorithms can analyze large volumes of text and extract key information, allowing researchers to quickly understand and synthesize relevant literature for their papers. Additionally, these algorithms can generate summaries of articles, enabling researchers to grasp the main points without reading entire documents. Another important application is in language generation.

Machine learning models can produce coherent and contextually appropriate text based on prompts provided by researchers. This capability is especially useful for generating sections of papers, such as introductions, abstracts, and conclusions. Furthermore, machine learning methods can assist in plagiarism detection and citation management, ensuring the originality and integrity of research papers. By following these methods, the study aims to develop an effective machine learning-based OTP fraud detection system, contributing to enhanced security in online transactions.

2.1. Logistic Regression

Logistic Regression is a binary classification algorithm used to predict the probability of a binary outcome. In this study, logistic regression was employed to classify OTP activities as either fraudulent or legitimate based on the extracted features. The model was trained using the preprocessed data and optimized using techniques such as gradient descent or Newton's method.

The linear regression model is

$$Z = \beta 0 + \beta 1x1 + \beta 2\ x2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots + \beta nxn$$

Here, Z is the output of the linear regression

$x1, x2, \ldots\ldots\ldots xn$ are the input feature

$\beta 0\beta 1\ldots\ldots\beta n$ are the coefficients (parameters) to be learned

The sigmoid function or logistic function is

$$\sigma_z = \frac{1}{1+e^{-z}}$$

Where e is the base of the natural logarithm

2.2. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and robustness. In this study, a random forest classifier was employed to leverage the collective intelligence of multiple decision trees for detecting OTP fraud. The model was trained using the preprocessed data, and hyper parameters such as the number of trees and maximum features were optimized through cross-validation.

Random Forest is a versatile machine learning algorithm used for both classification and regression tasks. It operates by constructing multiple decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Here are some important parameters in Random Forest:

i). **n_estimators:** This parameter sets the number of trees in the forest. A higher number generally improves performance but increases computational cost.

ii). **criterion:** It defines the function to measure the quality of a split. For classification, "gini" or "entropy" (information gain) can be used. For regression, it's usually "mse" (mean squared error).

iii). **max_depth:** This parameter controls the maximum depth of each tree in the forest. Deeper trees can model more complex relationships but are more prone to overfitting.

iv). **min_samples_split:** The minimum number of samples required to split an internal node. Higher values prevent the tree from splitting too early, potentially reducing overfitting.

v). **min_samples_leaf:** The minimum number of samples required to be at a leaf node. Like `min_samples_split`, higher values help in preventing overfitting by enforcing a minimum size for leaves.

vi). **max_features:** It determines the maximum number of features to consider when looking for the best split. A smaller number can reduce overfitting but might also decrease model performance.

vii). **bootstrap:** It indicates whether bootstrap samples are used when building trees. If set to `True`, each tree is built on a random sample with replacement from the training set.

viii). **random_state:** This parameter sets the seed for random number generation. Providing a fixed value ensures reproducibility.

ix). **class_weight:** For imbalanced datasets, you can use this parameter to assign different weights to classes. Options include `balanced` or a dictionary specifying class weights.

x). **oob_score:** If set to `True`, out-of-bag samples are used to estimate the generalization accuracy.

xi). **verbose:** Controls the verbosity of the tree-building process. Higher values give more information during training.

xii). **n_jobs:** The number of jobs to run in parallel during training. Set to `-1` to use all available cores.

These parameters can be adjusted based on the specific characteristics of the dataset and the desired performance of the Random Forest model. Tuning these parameters optimally is crucial for achieving the best results.

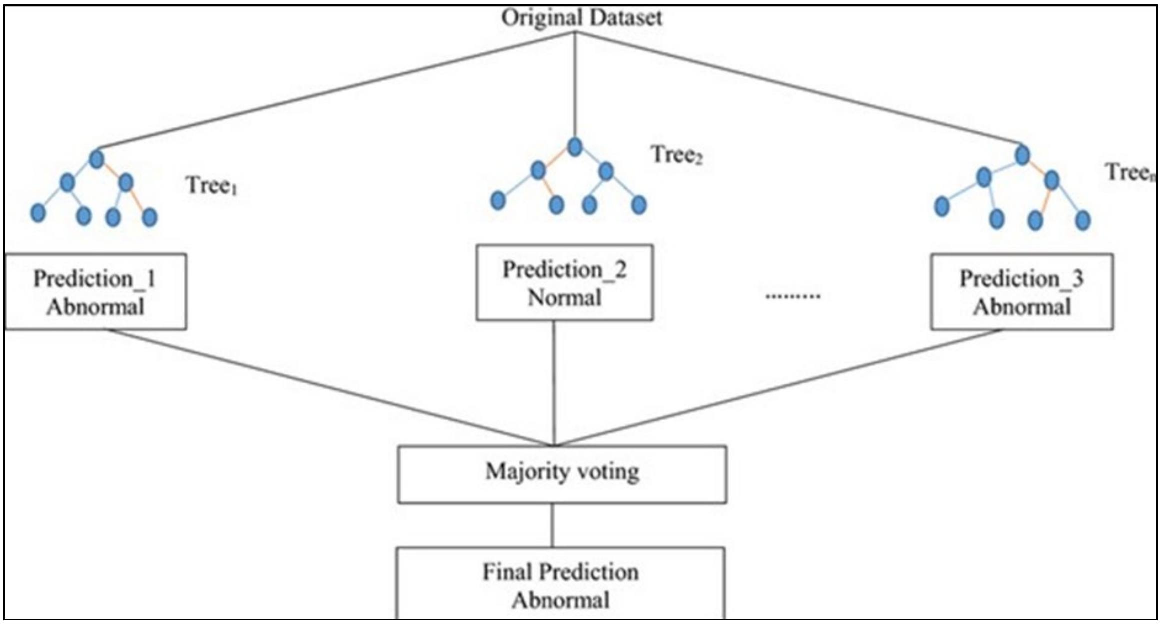

Fig 2: Random Forest Model Architecture

2.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression tasks. Its primary objective is to find the optimal hyperplane that best separates data points belonging to different classes in a high-dimensional space.

i). **Separating Hyperplane:** Given labeled training data (data points with known classes), SVM finds the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points (support vectors) of each class. This hyperplane effectively separates the data into different classes.

ii). **Kernel Trick:** SVM can handle non-linearly separable data by transforming the input features into a higher-dimensional space using a kernel function. This allows SVM to find a linear separation in the transformed space, even if the original data is not linearly separable.

iii). **Margin Maximization:** SVM aims to maximize the margin, which leads to better generalization performance. It selects the hyperplane that not only separates the data but also maximizes the distance between the hyperplane and the nearest data points (support vectors) of each class.

iv). **Classification:** Once the hyperplane is determined, SVM can classify new data points by checking which side of the hyperplane they fall on.

v). **Regularization:** SVM includes a regularization parameter (C) that controls the trade-off between maximizing the margin and minimizing the classification error on the training data. A larger C value allows for a smaller margin but fewer misclassifications, while a smaller C value prioritizes a larger margin, possibly at the cost of some misclassifications.

3. Results

3.1. Logistic Regression

The K-fold validation indicates that the accuracy of the logistic regression model is 0.78. This implies that the minimum accuracy of the model over the long period is 0.78.

The below table 1 shows the confusion matrix metrics such as precision, recall and F1 scores.

Table 1: Logistic Regression Accuracy

Data set	Logistics Regression Accuracy
Train set	0.96
Test set	0.95

Logistic regression accuracy is a measure of how well a logistic regression model performs in predicting the correct outcome (or class) for given data. In this context, the accuracy is presented for both the training set and the test set. In rain set accuracy (0.96), this means that the logistic regression model achieved an accuracy of 96% when it was trained on the training dataset. In other words, when the model was presented with data it had already seen during training, it correctly predicted the outcome 96% of the time. In test set accuracy (0.95), this indicates that the logistic regression model achieved an accuracy of 95% when it was tested on a separate dataset, known as the test set. This test set consists of data that the model has not seen during training. Therefore, the model's ability to generalize to new, unseen data is evaluated by this accuracy. A test set accuracy of 95% implies that when presented with new, unseen data, the model correctly predicted the outcome 95% of the time.

The both high training and test set accuracies (96% and 95% respectively) suggest that the logistic regression model is performing well and is likely not overfitting the training data. However, it's important to note that while high accuracy is desirable, it may not always be sufficient for evaluating the performance of a model. Other metrics, such as precision, recall, and F1-score, as well as confusion matrices, should also be considered, especially in scenarios where classes are imbalanced or misclassification costs are asymmetric.

3.2. Support Vector Machine (SVM)

The K-fold validation indicates that the accuracy of the logistic regression model is 0.79. This implies that the minimum accuracy of the model over the long period is 0.79.

The below table 2 shows the confusion matrix metrics such as precision, recall and F1 scores.

Table 2: Accuracy of Support Vector Machine

Data Set	SVM Accuracy
Train Set	0.957
Test Set	0.961

Based on K-fold validation, the logistic regression model's accuracy is 0.79. This implies that the minimum range of accuracy for the model over the long period is 79%.

3.3. Random Forest Model

Random Forest is a widely used ensemble learning method for both classification and regression tasks in machine learning. It works by creating numerous decision trees during training and then outputs the mode of the classes (for classification) or the mean prediction (for regression) from these individual trees.

Table 3: Accuracy of Random Forest Model

Data Set	Random Forest Accuracy
Train Set	0.999
Test Set	0.996

The Random Forest model accuracy represents how well the model performs in predicting outcomes for given data. . In the train set accuracy (0.999), this indicates that the Random Forest model achieved an accuracy of 99.9% when it was trained on the training dataset. In other words, when the model was presented with data it had already seen during training, it correctly predicted the outcome 99.9% of the time. In test set accuracy (0.996), this shows that the Random Forest model achieved an accuracy of 99.6% when it was tested on a separate dataset, known as the test set. The test set consists of data that the model has not seen during training. Therefore, the model's ability to generalize to new, unseen data is evaluated by this accuracy. A test set accuracy of 99.6% implies that when presented with new, unseen data, the model correctly predicted the outcome 99.6% of the time.

These high accuracy values (99.9% for the train set and 99.6% for the test set) suggest that the Random Forest model is performing exceptionally well and is likely not over fitting the training data. It demonstrates the model's capability to accurately predict outcomes, both on data it has seen during training and on new, unseen data. This indicates that the model is robust and can effectively generalize to unseen data, making it a reliable predictor. However, it's essential to consider other metrics and thoroughly evaluate the model's performance, especially in real-world scenarios where the data may be more complex or imbalanced.

4. Comparison of Machine Learning Models

The comparison based on the accuracy of models of the Logistic Regression, SV and Random Forest Models are listed in the below table 4.

Table 4: The comparison of three models

Data Set	Logistic Regression	SVM	Random Forest
Training Set	0.961	0.957	0.999
Test Set	0.965	0.961	0.996

Based on the accuracy listed in table 4, here's the comparison of Logistic Regression, Support Vector Machine (SVM), and Random Forest models. In Logistic Regression performs well with both train and test sets, with a slightly higher accuracy on the test set compared to the train set. This indicates that the model generalizes well to unseen data. In Support Vector Machine (SVM) also performs well, with similar accuracies for both the train and test sets. This suggests that the model generalizes effectively. In Random Forest achieves the highest accuracy among the three models, with nearly perfect accuracy on the train set and slightly lower but still impressive accuracy on the test set. This indicates that the model may be overfitting slightly to the training data, but it still generalizes well to unseen data.

Therefore, Random Forest has the highest accuracy on both train and test sets, followed by Logistic Regression and SVM. However, it's essential to consider other factors such as model complexity, interpretability, and computational resources when selecting the best model for a particular task.

5. Conclusion

In summary, the application of machine learning models, including Logistic Regression, Support Vector Machine (SVM), and Random Forest, has shown significant potential in detecting OTP fraud cases in Telangana. By utilizing data collected from the National Bureau of Cyber Security (NBC), these models have demonstrated the ability to identify complex patterns and anomalies indicative of fraudulent behavior across various OTP platforms. Incorporating this analysis has further improved the precision and relevance of fraud detection efforts, tailored to the socio-cultural dynamics of Telangana. As indicated by rigorous evaluation the proposed framework provides a robust and effective method of identifying and mitigating fraudulent activities, thus protecting users' interests and maintaining the integrity of online interactions. Looking ahead, further research and development in this area have the potential to refine and optimize detection mechanisms, contributing to ongoing efforts to combat OTP fraud and build trust in digital interactions within Telangana and beyond.

References

1. Avais MA, et al. Awareness regarding cyber victimization among students of University of Sindh, Jamsharo. *International Journal of Asian Social Science*. 2014;4(5):632-641.
2. Hasan, et al. Perception and Awareness of Young Internet Users towards Cybercrime: Evidence from Malaysia. *Journal of Social Sciences*. 2015;11(4):395-404.
3. Jamil D, Khan MNA. Data Protection Act in India with Compared To the European Union Countries. *International Journal of Electrical and Computer Sciences*. 2011;11(06).
4. Mehta S, Singh V. A Study of Awareness about Cyber laws in the Indian Society. *International Journal of Computing and Business Research*. 2013 Jan;4(1).
5. Halevi T, Memon N, Lewis J, Kumaraguru P, Arora S, Dagar N, et al. Cultural and psychological factors in cyber-security. In: *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*. ACM; 2016 Nov. p. 318-324.
6. Singaravelu S, Pillai KP. B.Ed. Students Awareness on Cybercrime in Perambalur District. *International Journal*

of Teacher Educational Research (IJTER). 2014 Mar;3(3).

7. Mokha AK. A Study on Awareness of Cyber Crime and Security. *Research J. Humanities and Social Sciences*. 2017 Oct-Dec;8(4):459-464. doi: 10.5958/2321-5828.2017.00067.5.
8. The Times of India. One cybercrime in India every 10 minutes. *The Times of India*. 2017 Jul 22. Available from: http://timesofindia.indiatimes.com/india/one-cybercrime-in-india-every-10-minutes/articleshow/59707605.cms
9. Statista. India mobile phone internet user growth 2016. Available from: https://www.statista.com/statistics/309020/india-mobile-phone-internet-user-growth