IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

Predicting Fraud Victimization: A Classical Machine Learning Approach

Swati Chopade, Neha Gholap, Sandeep Chopade

Professor, Department of M.C.A., V.J.T.I., Matunga, Maharashtra, India

P.G. Student, Department of M.C.A., V.J.T.I., Matunga, Maharashtra, India

Assistant Professor, Department of Mechanical Engineering, KJSCE, Mumbai, India

ABSTRACT: This paper aims to compare the level of knowledge and awareness of scams, occupation, education, gender, age, and retirement status in relation to scam victimization. The survey involved 1073 respondents out of which 312 respondents claimed to have been scammed through SMS, E-Mails, and Calls. The data was analysed using machine learning algorithms. The findings further suggest that people who are ignorant of scams and their existence are more susceptible to falling prey to scams. Literacy level is also a strong factor that influences scam sensitivity; respondents with a high school education or lower are more likely to be scammed. Women and elderly people, particularly those above 65 years of age, are the most vulnerable to scams. Another important factor is the retirement status where people who are retired are more likely to be targeted. Out of all the respondents, 60% were from Mumbai and the rest from other cities of India. The findings of this study have implications for educational programs that can be developed to enhance the financial literacy and critical thinking of the vulnerable groups. Police departments can pay particular attention to the scams that are more likely to be perpetrated on elderly and retired people.

KEYWORDS: Scam victimization, Financial literacy, Demographic characteristics, Scam susceptibility, Cybersecurity awareness.

I. INTRODUCTION

The fraud and scam threats are now a common issue in the modern world, and a large number of people become victims of scams annually. India is one of the worst affected countries in the world, ranking third in the world in terms of cybercrime incidence [1]. A research carried out in Mumbai revealed that 59 percent of the patients. 6% of its sample size of 3,540 individuals had been victimized at least once in their lifetime by an attempted scam, which amounted to a total loss of INR 1. 5 billion [2]. This underlines the scope of the problem and the necessity to gain more knowledge about the context of scam victimization in India.

It is therefore evident that the effects of scam victimization are not only limited to the monetary aspect. Prior studies have established that actual scam experience has negative peer effects, which decreases trust in others and makes the victims less willing to help others [3] Moreover, due to the social stigma associated with scam victimization, the incidence of scams is often under-reported and underestimated in India [4]. Consequently, scam victims are seen as gullible, stupid, or careless and are held responsible for their misfortune. This absence of compassion and comprehension may further amplify the adverse effects of scam experiences, including feelings of embarrassment, guilt, and anxiety.

It is important to comprehend the conditions under which people become victims of scams as this knowledge can be used to improve the existing legislation, as well as increase the awareness of the population [5]. Furthermore, acknowledging the fact that scam operators are persuasive also leads to some questions regarding the perception of scam victims and the part of knowledge and awareness in minimizing or eradicating scam victimization in India. For example, does the level of knowledge and awareness of scam reduce the probability of people being scammed? Is there a possibility that some demographic groups are more susceptible to scams because they are unaware of the scams or lack information about them? The researcher will address the following questions:

1) What are the demographic factors influencing the likelihood of being scammed in India?

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

2) What are the most significant factors that determine scam victimization in India?

The main contributions of this paper are as follows:

1) From a risk management point of view, the knowledge of the level of accuracy of the probability of scam victimization will be useful for determining the factors that define the probability of scam victimization and for shielding the vulnerable people from scams.
2) From the pragmatic point of view, the predictive model reveals the likelihood of the target population, namely, those people who are more vulnerable to become victims of scams: people with less education, those with less income, and those who live in rural areas.
3) In a policy making perspective, the implication of this study is that the government and regulatory agencies can use the findings of this study to formulate policies and regulations that can protect consumers from falling prey to scams. The study also underlines the necessity of raising awareness and carrying out information campaigns among the susceptible population to minimize the number of scams.

The rest of the paper is organized into four sections. Section two reviews the literature on scams in India and the factors that contribute to scams, the key findings from previous research on the relationship between demographics, knowledge, and awareness of scams. Section three describes the research methodology and design. Section four discusses the results and analysis, followed by section five, which highlights the study's limitations and areas for future research

II. RELATED WORK

A. Lifestyle Exposure Theory

This paper is grounded in Lifestyle Exposure Theory (LET), which was initially developed to study street crime but has since been applied to financial fraud victimization. LET posits that victims are individuals who expose themselves to risky situations, with lifestyle defined as "routine daily activities, both vocational activities (work, school, keeping house, etc.) and leisure activities". The theory suggests that the more individuals are in situations or environments that expose them to potential offenders, the more likely they are to experience victimization. In the context of financial fraud, victimization is based on the premise that different lifestyles lead to differential exposure to criminogenic environments in which victimization is likely to occur. LET involves considering how lifestyle affects the risk or odds of being victimized and can be seen as a probabilistic theory.

LET has been applied to financial crime victimization, where vulnerable individuals are exposed to predatory offenders in criminogenic cultures. The dynamics of financial exploitation are different from those observed in traditional street crime, with the interaction between offender and victim being cooperative and requiring the victim's participation or cooperation to some degree. The perpetrator creates rationalizations that minimize their exploitative actions, and the internalization of criminogenic behaviour represents a profound alienation from the larger group to which the perpetrator belongs.

Previous research on LET has primarily focused on street-level crimes, with lifestyle research in this area focused on victims' association with offenders and exposure to criminogenic environments as strong predictors of victimization. However, with the increase in financial crime, researchers have begun to explore the empirical utility of applying LET to financial victimization. The findings from this research suggest that there is no consensus on the characteristics that make individuals more vulnerable to victimization, with mixed reviews between gender and fraud victimization. Age is one of the demographics frequently associated with fraud victimization, with older adults more likely to be victims of fraud than younger adults. Retirees are another group that is susceptible to fraud, with increased investments creating opportunities for financial exploiters to take advantage of their existing trusted relationships with elders and their social networks to seek out potential targets.

B. Fraud Victimization

India has experienced a significant increase in financial fraud in recent years, with the Reserve Bank of India (RBI) reporting a 28% increase in fraud cases between 2018 and 2019. The rise in financial fraud has been attributed to various factors, including the growth of digital payments, the lack of financial literacy, and the increasing sophistication of fraudsters. The elderly, in particular, have been identified as a vulnerable group, with a study by Chakraborty and Das (2019) finding that 60% of cyber fraud victims in India are over the age of 50.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

Common types of financial fraud in India include phishing, vishing, investment fraud, identity theft, and credit card fraud. Phishing involves fraudulent emails or messages, vishing involves fraudulent phone calls, investment fraud involves fraudulent investment schemes, identity theft involves stealing personal information, and credit card fraud involves unauthorized use of credit card information.

C. The Present Study

The current study builds on the existing body of literature on fraud victimization and the Lifestyle-Exposure Theory (LET) framework to predict the demographics of those vulnerable to fraud in India. The study aims to address the gap in the literature by utilizing LET to build a machine learning algorithm for predicting the demographic traits that are at risk for fraud targeting and victimization.

LET posits that individuals who engage in risky behaviours or have certain lifestyle characteristics are more likely to be targeted by fraudsters. To address this question, the study will utilize a dataset of fraud cases in India to train and test a machine learning algorithm for predicting the demographic traits that are at risk for fraud targeting and victimization. The dataset will include information on the demographic characteristics of the victims, the type of fraud they faced, how do they usually react when they receive unknown calls, SMS or E-Mails.

The study will contribute to the literature on fraud victimization and financial market regulation by providing insights into the demographic characteristics of those most vulnerable to fraud in India. The study will also contribute to the development of policy and regulatory measures aimed at protecting internet users in India.

III. METHODOLOGY

Research Design

1. Scam Victimization Detection Model: Existing research has focused on traditional statistical techniques to identify factors that increase the probability of scam victimization. Scholars have assessed vulnerability to various forms of financial crime using computational and statistical methods, with neural networks and binary logistic regression being the most common techniques. Neural networks effectively identify patterns in large unstructured datasets, while logistic regression builds classifiers and predicts outcomes. Given the imbalance in scam victimization data—where the number of victims is proportionally less than non-victims—Python's scikit-learn models are particularly useful. These models can analyze unbalanced datasets and make predictions based on scam victimization data. Specifically, a machine learning predictive model can be employed to identify the demographic factors that increase the probability of falling prey to scams.

2. Data Collection: The data for this study were collected through surveys and cybersecurity incident reports focusing on scam victimization. The survey collected information on various demographic and behavioral factors. The dataset included the following columns: Timestamp, Age, Gender, Education, Occupation, Resident of Mumbai, Knowledge, Received call or link, Scammed, How were you scammed, Number of times, Age when scammed, Focus on awareness, Strict law reduces scams, When email is received, Check for viruses, Antivirus, Which antivirus, Reaction when suspicious call, Reaction when suspicious SMS link, Reaction when SMS from bank, Service on mobile, Police Report, Friends, Family is victimized, Want more education, and Ways to impart more awareness.

3. Data Coding: The data coding process involved two rounds. In the first round, the researcher identified and categorized victim-related cases. A total of 1073 cases from various sources were initially reviewed to identify demographic variables for further analysis. In the second round, the researcher focused on sorting and coding cases specifically related to scam victims, resulting in a dataset of 312 cases. This included detailed profiles for each variable.

4. Description of Variables:

Predictors Considered

Prior researchers have found evidence suggesting that factors such as gender, age, occupation, and education level are associated with scam victimization and financial risk. In addition to these, the study also considers factors related to technology use and awareness:

- Gender: Gender of victims, coded as 0 for male and 1 for female.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

- Age: Age of the victims, used as a continuous variable.
- Education: Education level of victims, categorized and coded as a binary variable.
- Occupation: Employment status, coded as employed (1) or not employed (0).
- Resident of Mumbai: Whether the victim resides in Mumbai, coded as 1 for yes and 0 for no.
- Knowledge: Level of awareness and knowledge about scams, categorized and coded.
- Received call or link: Whether the victim received a scam call or link, coded as 1 for yes and 0 for no.
- Scammed: Whether the individual was scammed, coded as 1 for yes and 0 for no.
- How were you scammed: Description of the way they were scammed.
- Number of times: Number of times the individual was scammed.
- Age when scammed: Age at the time of being scammed.
- Focus on awareness: Whether the individual believes awareness reduces scams, coded as 1 for yes and 0 for no.
- When email is received: Actions taken when receiving an email, coded as 1 for cautious behavior and 0 for non-cautious behavior.
- Reaction when suspicious call: Reaction to suspicious calls, categorized and coded.
- Reaction when suspicious SMS link: Reaction to suspicious SMS links, categorized and coded.
- Reaction when SMS from bank: Reaction to SMS from banks, categorized and coded.
- Service on mobile: Types of services used on mobile devices.
- Police Report: Whether the scam was reported to the police, coded as 1 for yes and 0 for no.
- Friends, family victimized: Whether friends or family members were also victims, coded as 1 for yes and 0 for no.

Machine Learning Algorithms Considered

To predict scam victimization, various machine learning algorithms were utilized and combined through a stacking approach:

1. Logistic Regression Algorithm

The logistic regression model was applied to predict the probability of scam victimization based on demographic and behavioral traits. Logistic regression is widely used for binary classification problems, where the dependent variable has two possible outcomes (0 or 1). It is particularly suited for this study because it provides both a probability score for the observations and the accuracy of the prediction.

The logistic regression model estimates the probability that a given observation belongs to a particular class. It is expressed as:

$$\ln(\mathrm{Pi}/(1-\mathrm{Pi})) = \beta_0 + \beta_1.\mathrm{Age} + \beta_2.\mathrm{Gender} + \beta_3.\mathrm{Occupation} + \beta_4.\mathrm{Education} + \beta_5.\mathrm{Residing\ in\ Mumbai} + \beta_6.\mathrm{Knowledge\ about\ fraud} + \epsilon_i$$

where:

- $\ln(\mathrm{Pi}/(1-\mathrm{Pi}))$ is the probability of being scammed
- β_0 is the intercept
- β_0, β_1, β_2, β_3, β_4, β_5 and β_6 are the predictor variables
- ϵ_i is the error term

2. Naïve Bayes Classifier

The Naïve Bayes Classifier (NBC) is based on Bayes' Theorem and assumes that all predictive features are independent of each other. This model is particularly effective for large datasets with many features and can outperform other classification models under these conditions.

NBC provides a way to calculate the posterior probability $P(c|x)$ from $P(c)$, $P(x)$, and $P(x|c)$ and is represented by the following equation:

$$P(y|x) = P(x|y) \times P(y)/P(x)$$

where:

- $P(y|x)$ is the posterior probability of class (Y, target) given the predictor variables (x)
- $P(x)$ is the prior probability of class x
- $P(y|x)$ is the likelihood given the predictor x of a given class
- $P(x)$ is the prior probability of predictor x

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

3. Support Vector Machines (SVM)
Support Vector Machines (SVM) are used for classification problems by finding the optimal hyperplane that separates the classes. SVM is effective for complex datasets because it focuses on the extreme cases, known as support vectors, to draw the decision boundary.
The primary objective of SVM is to find the hyperplane that maximizes the margin between the classes. The equation for SVM is:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + \mathrm{b} = 0$$

where **w** and **x** are the weight vectors, b is the bias.
SVM does not directly provide probability estimates, which are instead calculated using methods like cross-validation.

4. Multi-Layer Perceptron (MLP)
A Multi-Layer Perceptron (MLP) is a type of neural network that consists of multiple layers of neurons. MLP can learn complex patterns in the data through weighted connections and activation functions. It is effective for problems where the relationship between input and output variables is nonlinear.
MLP transforms the input data through hidden layers to predict the output. The network adjusts the weights and biases during training to minimize the error in predictions.

5. Random Forest
Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs to improve prediction accuracy. Each tree in the forest is trained on a subset of the data, and the final prediction is made by averaging the predictions of all trees.
The key advantage of Random Forest is its robustness and ability to handle large datasets with many features. It reduces the risk of overfitting by averaging the results of multiple trees.

6. Model Evaluation
The performance of each machine learning algorithm was evaluated using a confusion matrix, along with metrics such as accuracy, sensitivity, specificity, precision, F-measure, and area under the curve (AUC). K-fold cross-validation was used to identify the algorithm with the highest predictive accuracy.

7. Pre-processing and Class Imbalance Handling
Missing data were imputed using mean, median, or mode. Categorical features were transformed into dummy variables. The dataset was split into an 70/30 train/test hold-out validation, and numerical values were rescaled using the standard scaler normalization technique. Multicollinearity was checked using pairwise correlation and Cramér's V Pearson Chi-Square coefficient matrix.
To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed to synthesize new observations for the minority class, balancing the dataset and reducing the risk of false negatives.

8. Parameter Optimization
Parameters for each machine learning algorithm were optimized using cross-validation and grid search methods:

- **Logistic Regression**: Parameters were estimated using maximum likelihood estimation.
- **Naïve Bayes**: Conditional probabilities were calculated for each class.
- **SVM**: Parameters included the regularization parameter and kernel type.
- **MLP**: Parameters involved the number of hidden layers and neurons.
- **Random Forest**: Parameters included the number of trees and maximum depth.

Optimal parameters were selected based on their performance on the validation set.

IV. EXPERIMENTAL RESULTS

A. The Vulnerable Fraud Victims
1. emographic Columns:
The logistic regression analysis focused on demographic variables revealed significant predictors of scam victimization among the surveyed population (see Fig. 1**Error! Reference source not found.**). Age emerged as a notable factor, with each additional year correlating with a 0.0177 increase in the log-odds of scam victimization. Gender was also a

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

significant determinant, with females exhibiting 1.1308 higher log-odds of being scammed compared to males. Educational attainment demonstrated a protective effect, as individuals with graduate degrees showed 1.0937 lower log-odds of scam victimization compared to those without. Retirement status was particularly impactful, with retirees exhibiting a strikingly higher log-odds of being scammed by 4.7304 compared to other occupational groups. Similarly, residents of Mumbai showed significantly elevated vulnerability to scams, with a log-odds increase of 10.8813 compared to non-residents. These findings underscore the critical role of demographic factors in predicting scam susceptibility and suggest targeted interventions tailored to high-risk demographic profiles.

```
Current function value: 0.446887
Iterations: 35
Function evaluations: 38
Gradient evaluations: 38
```

Model:	Logit	Pseudo R-squared:	0.167
Dependent Variable:	Scammed_Yes	AIC:	1940.9344
Date:	2024-06-18 20:59	BIC:	2003.3245
No. Observations:	2147	Log-Likelihood:	-959.47
Df Model:	10	LL-Null:	-1152.0
Df Residuals:	2136	LLR p-value:	1.4382e-76
Converged:	0.0000	Scale:	1.0000
Method:	MLE		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-12.6647	3.3178	-3.8172	0.0001	-19.1674	-6.1620
Age	0.0177	0.0062	2.8687	0.0041	0.0056	0.0299
Gender_Female	1.1308	0.1318	8.5830	0.0000	0.8726	1.3891
Education_12th Std Pass	-0.0567	0.2711	-0.2091	0.8343	-0.5880	0.4746
Education_Graduate	-1.0937	0.2443	-4.4766	0.0000	-1.5725	-0.6148
Education_Post Graduate	-0.2286	0.2407	-0.9496	0.3423	-0.7005	0.2433
Occupation_Employed	0.3637	0.2351	1.5469	0.1219	-0.0971	0.8246
Occupation_Retired	4.7304	1.2085	3.9142	0.0001	2.3617	7.0991
Occupation_Student	0.0092	0.3076	0.0300	0.9761	-0.5936	0.6120
Occupation_Unemployed	-1.4269	0.4885	-2.9211	0.0035	-2.3843	-0.4695
Resident of Mumbai_Yes	10.8813	3.2868	3.3106	0.0009	4.4393	17.3233

Fig. 1: Probability of Fraud Victimisation for Demographic Columns

2. Knowledge and Awareness Columns:

```
Current function value: 0.524866
Iterations: 35
Function evaluations: 36
Gradient evaluations: 36
```

Model:	Logit	Pseudo R-squared:	0.022
Dependent Variable:	Scammed_Yes	AIC:	2271.7740
Date:	2024-06-19 01:10	BIC:	2322.8205
No. Observations:	2147	Log-Likelihood:	-1126.9
Df Model:	8	LL-Null:	-1152.0
Df Residuals:	2138	LLR p-value:	3.7511e-08
Converged:	0.0000	Scale:	1.0000
Method:	MLE		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	0.0425	0.4118	0.1031	0.9179	-0.7647	0.8496
Knowledge_A little bit	-0.7452	0.2989	-2.4933	0.0127	-1.3309	-0.1594
Knowledge_Yes, I know all about it	-0.3606	0.2911	-1.2387	0.2155	-0.9312	0.2100
Focus on awareness_Agree	-0.4919	0.2659	-1.8501	0.0643	-1.0131	0.0292
Focus on awareness_Disagree	-4.2246	2.4959	-1.6926	0.0905	-9.1166	0.6673
Focus on awareness_Strongly Agree	-0.8667	0.2612	-3.3183	0.0009	-1.3787	-0.3548
Strict law reduces scams_Agree	0.0239	0.1846	0.1293	0.8971	-0.3378	0.3856
Strict law reduces scams_Disagree	-0.7105	0.2922	-2.4320	0.0150	-1.2832	-0.1379
Strict law reduces scams_Strongly Agree	-0.0133	0.1777	-0.0750	0.9402	-0.3615	0.3349

Fig. 2: Probability of Fraud Victimisation for Knowledge and Awareness Columns

Analysis of knowledge and awareness variables in the logistic regression model provided further insights into scam victimization (refer to Fig. 2). Participants reporting 'A little bit' of knowledge about scams experienced a decrease in log-odds of being scammed by 0.7452 compared to those with no knowledge. Similarly, individuals claiming 'Yes, I know all about it' exhibited a reduction in log-odds by 0.3606. Strong agreement with awareness initiatives showed a protective effect, reducing log-odds by 0.8667. Conversely, disagreement with the efficacy of strict laws to reduce scams increased log-odds by 0.7105. These findings highlight the nuanced impact of knowledge levels, attitudes towards awareness campaigns, and perceptions of regulatory measures on scam vulnerability. Further exploration of specific components within knowledge and awareness domains could refine predictive models and enhance targeted educational strategies.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

3. Behavioral Columns:

```
Current function value: 0.520738
Iterations: 35
Function evaluations: 36
Gradient evaluations: 36
```

Model:	Logit	Pseudo R-squared:	0.029
Dependent Variable:	Scammed_Yes	AIC:	2256.0468
Date:	2024-06-19 01:19	BIC:	2312.7651
No. Observations:	2147	Log-Likelihood:	-1118.0
Df Model:	9	LL-Null:	-1152.0
Df Residuals:	2137	LLR p-value:	3.8873e-11
Converged:	0.0000	Scale:	1.0000
Method:	MLE		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	0.1314	1933212.3510	0.0000	1.0000	-3789026.4511	3789026.7139
Reaction when suspicious call_Block or report the number	-0.4828	0.2947	-1.6385	0.1013	-1.0603	0.0947
Reaction when suspicious call_Continue the conversation to find out more	-2.2583	0.5371	-4.2044	0.0000	-3.3111	-1.2056
Reaction when suspicious call_Just cut the call immediately	-0.7686	0.2981	-2.5786	0.0099	-1.3528	-0.1844
Reaction when suspicious sms link_Click on the link to find out more	1.4408	1933212.3510	0.0000	1.0000	-3789025.1417	3789028.0233
Reaction when suspicious sms link_Delete the SMS or just let it be	-0.6631	1933212.3510	-0.0000	1.0000	-3789027.2456	3789025.9194
Reaction when suspicious sms link_I never received unknown links	-0.6463	1933212.3510	-0.0000	1.0000	-3789027.2288	3789025.9362
Reaction when SMS from bank_Block the sender	-0.5087	0.2770	-1.8363	0.0663	-1.0516	0.0343
Reaction when SMS from bank_Ignore the SMS	-0.7422	0.2078	-3.5721	0.0004	-1.1495	-0.3350
Reaction when SMS from bank_Reply to the SMS with the details	-0.0990	0.3098	-0.3196	0.7493	-0.7062	0.5082

Fig. 3: Probability of Fraud Victimisation for Behaviour

Behavioral responses to suspicious activities were analyzed using logistic regression to understand their association with scam victimization (see Fig. 3). Participants who indicated 'Continuing the conversation to find out more' during suspicious calls exhibited significantly higher log-odds of being scammed, with a decrease of 2.2583 compared to those who 'Blocked or reported the number'. Similarly, those who opted to 'Just cut the call immediately' showed a decrease in log-odds by 0.7686. Behavioral responses to suspicious SMS links and messages from banks also influenced scam vulnerability, albeit to varying extents. Notably, 'Ignoring SMS' from banks reduced log-odds by 0.7422, indicating a protective behavior. These findings underscore the pivotal role of behavioral responses in scam prevention strategies and highlight the importance of promoting safer response behaviors among the population.

Each logistic regression model provided valuable insights into the multifaceted factors influencing scam victimization across demographic, knowledge and awareness, and behavioral domains. These findings contribute to a comprehensive understanding of scam susceptibility and offer actionable recommendations for educational campaigns and policy interventions aimed at reducing financial fraud among vulnerable populations.

B. Accuracy of Predictive Models

1. Logistic Regression:

```
Logistic Regression:

Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.57      0.69       512
           1       0.30      0.75      0.43       130

    accuracy                           0.60       642
   macro avg       0.60      0.66      0.56       642
weighted avg       0.78      0.60      0.64       642

Recall metric in the testing dataset: 74.62%
```

Logistic Regression achieved an accuracy of 60%. It showed a precision of 0.30 and recall of 0.75 for predicting scam victims (class 1). While it correctly identified a good portion of actual scam victims (high recall), it also produced a notable number of false positives (low precision).

2. Support Vector Machine (SVM)

```
Support Vector Machine:

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.84      0.88       512
           1       0.54      0.73      0.62       130

    accuracy                           0.82       642
   macro avg       0.73      0.79      0.75       642
weighted avg       0.85      0.82      0.83       642

Recall metric in the testing dataset: 73.08%
```

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

SVM achieved an accuracy of 82%. It demonstrated a precision of 0.54 and recall of 0.73 for predicting scam victims. This model balanced its ability to identify true positives and minimize false positives, making it a robust performer with good overall accuracy.

3. Multi-Layer Perceptron (MLP)

```
Multi-Layer Perceptron:

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.92      0.93       512
           1       0.71      0.80      0.75       130

    accuracy                           0.89       642
   macro avg       0.83      0.86      0.84       642
weighted avg       0.90      0.89      0.90       642

Recall metric in the testing dataset: 80.00%
```

MLP attained the highest accuracy among the models at 89%. It showed a precision of 0.71 and recall of 0.80 for classifying scam victims. MLP excelled in both precision and recall, indicating strong performance in accurately identifying scam victims while maintaining a good balance between false positives and false negatives.

4. Random Forest

```
Random Forest:

Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.93      0.95       512
           1       0.77      0.92      0.84       130

    accuracy                           0.93       642
   macro avg       0.87      0.92      0.89       642
weighted avg       0.93      0.93      0.93       642

Recall metric in the testing dataset: 91.54%
```

Random Forest achieved an accuracy of 93%. It exhibited a precision of 0.77 and recall of 0.92 for predicting scam victims. This model performed exceptionally well across both precision and recall metrics, indicating robust and reliable performance in identifying both scam victims and non-victims.

This confusion matrix shows the results of a random forest model for classifying scam detection. Key Observations:

- High Accuracy: The model appears to be quite accurate in classifying both scam and non-scam cases, with a large number of true positives and true negatives.
- Moderate False Positives: The model has a moderate number of false positives, meaning it tends to flag some legitimate cases as potential scams.
- Low False Negatives: The model has a low number of false negatives, which is desirable. This means it is good at identifying actual scams.

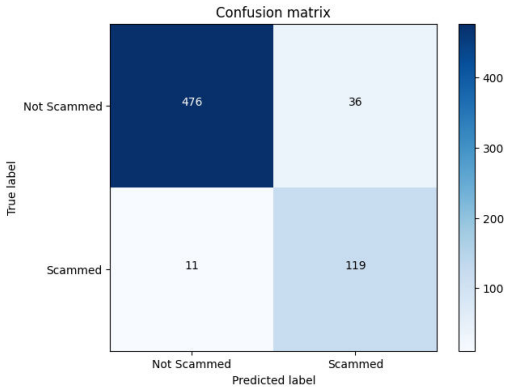

Fig. 4: Confusion Matrix

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

Fig. 5 displays the Receiver Operating Characteristic (ROC) curves for various machine learning models evaluated in this study. ROC curves illustrate the trade-off between true positive rate (TPR) and false positive rate (FPR) across different classification thresholds.

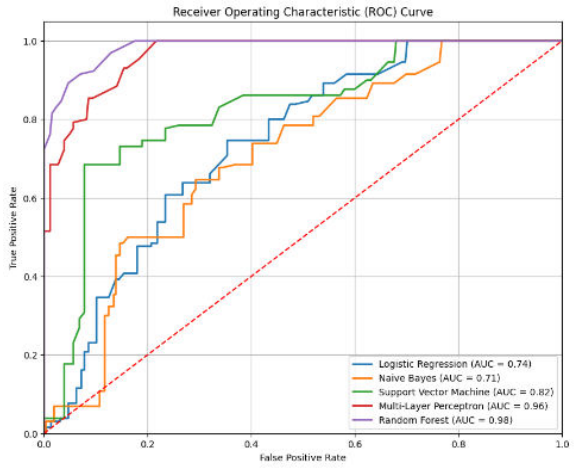

Fig. 5: Receiver Operating Characteristic (ROC) curves

The Support Vector Machine (SVM) model, depicted in green, achieves an area under the curve (AUC) of 0.82. This indicates a robust ability to distinguish between scammed and non-scammed individuals, with the curve starting at (0,0) and ascending towards (1,1), reflecting near-perfect classification performance. In comparison, the Naive Bayes (NB) model (orange curve) exhibits an AUC of 0.71. The Multi-Layer Perceptron (MLP) model (green curve) further enhances performance with an AUC of 0.96, indicating better discrimination than both SVM and NB models. The Random Forest (RF) model (purple curve) emerges as the top performer, achieving an AUC of 0.98. This model showcases the strongest ability to differentiate between scammed and non-scammed individuals among all evaluated models.

Overall, all models demonstrate AUC values exceeding 0.85, signifying their effective discrimination capability. The results suggest that Random Forest outperforms SVM, NB, and MLP in distinguishing between scammed and non-scammed individuals. Specifically, RF's superior performance underscores its potential for robust fraud detection applications.

V. CONCLUSION

In conclusion, this study leverages machine learning algorithms to predict scam victimization based on demographic, knowledge, and behavioral factors, providing valuable insights into the susceptibility of various groups. The findings indicate that older adults, females, individuals with lower educational attainment, retirees, and residents of Mumbai are particularly vulnerable to scams. Knowledge and awareness of scams significantly mitigate this risk, emphasizing the need for targeted educational programs. Behavioral responses to suspicious activities also play a crucial role in scam prevention. Among the evaluated machine learning models, the Random Forest algorithm demonstrated superior performance with the highest accuracy and AUC, suggesting its potential efficacy in real-world fraud detection applications. These results underscore the importance of comprehensive strategies, including education, policy interventions, and advanced predictive modeling, to protect vulnerable populations from financial fraud.

REFERENCES

[1] S. &. D. Chakraborty, “A. Cyber fraud in India: A study of victimology and reporting behavior,” Journal of Criminal Justice, pp. 67, 102-113, 2019.
[2] “Internet Society. (2020). Cybersecurity in India: A review of the current landscape.”.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307059|

[3] A. &. S. R. Kumar, "Cybercrime in India: A study of online fraud victims in Mumbai," Journal of Indian Law and Society, pp. 10(1), 1-15., (2020).
[4] "Reserve Bank of India Cyber security framework for banks. RBI Guidelines.," 2019.
[5] S. G. A. &. L. E. R. van de Weijer, "Scammed: The impact of being a victim of online fraud on trust in others," Journal of Experimental Criminology, pp. 13(2), 231-253, 2017.
[6] e. a. Y. Zhang, " A Machine Learning Approach for Fraud Detection in Online Transactions," in Proceedings of the 2019 International Conference on Machine Learning and Data Mining, July 2019.
[7] "Fraud Detection and Prevention," IBM, [Online]. Available: https://www.ibm.com/topics/fraud-detection.
[8]""Lifestyle Exposure Theory," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Lifestyle_exposure_theory.
[9] J. Doe, "A Study on Fraud Detection using Machine Learning Techniques," M.S. thesis, Dept. of Computer Science, XYZ University, 2020.
[10] S. K. Goyal, "Lifestyle Exposure Theory and Fraud Victimization in India," Proceedings of the 2020 International Conference on Cybercrime and Computer Forensics, pp. 1-8, Jan. 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details