Capstone Project:

Understanding Cyber Security and Safety Perceptions

for

Digital Payments and E-Services

विद्याविनियोगाद्विकासः

Presented By

EPABABL05 Group – eShield.

Chintan Joshi	EPABABL0523009
Hardik Raval	EPABABL0523011
Rajiv Trivedi	EPABABL0523029

Project Mentor

Prof. Kavitha Ranganathan, Prof. Arnab Laha

1 Contents

List of Figures

List of Tables

List of Code Snippet

Source Code Location

(Github Project Source, 2023)

1 Executive Summary

The Capstone Project for EPABA Batch 05 focuses on investigating the factors that impact cybersecurity in online transactions, particularly in the context of digital payments and e-Services. The study aims to understand consumer perceptions of safety in these transactions and analyse how various factors influence the adoption of digital payments and e-Services. Additionally, the project aims to identify associations with demographic variables and address the challenges hindering the widespread adoption of digital payments and related e-Services. The insights gained from this analysis will contribute valuable information to enhance cybersecurity measures and facilitate a smoother transition to digital payment methods and e-Services.

2 Introduction

In today's digital world, using things like online payments and services on the internet has become common. But, as we do more things online, like paying for stuff, there's a worry about how safe it is. This project, part of EPABA Batch 05, is all about figuring out what makes online transactions safe or risky, especially when it comes to digital payments.

A lot of people are concerned about using digital payments. Some don't know much about the risks of cyber fraud, and others worry they might lose money when they pay for things online. We want to understand how different groups of people, like those of different ages, incomes, and jobs, feel about using digital payments. This way, we can come up with good ideas to encourage more people to use digital payments and make sure it's safe.

To get all this info, we're talking to people and asking them questions through surveys. We want to know what they think about digital payments, if they know about cyber fraud, and if their age, income, or job influences how they use these services. After collecting all this info, we'll study it closely to find out what patterns we can see and come up with practical suggestions. Our goal is to make digital payments widely accepted and safe for everyone. The process involves talking to different people and using surveys to get a complete picture, and then carefully studying the data to find useful insights.

3 Objective

The objective of this project is to investigate and understand the factors influencing cybersecurity in online transactions, with a specific focus on digital payments and e-Services. Leveraging the analytical tools and methods acquired during the EPABA program, our study aims to answer key questions through exploratory analysis, predictive analysis and shedding light on various dimensions of consumer behaviour and perceptions.

Key Questions to be Addressed.

3.1 Exploratory Analysis

- How do individuals of different age groups, income levels, and occupations engage with digital payments and e-Services?
- What patterns emerge when we explore the data based on demographic factors such as age, income, and occupation?

3.2 Cybersecurity Concerns

- To what extent are people concerned about the cybersecurity of online transactions, particularly in the context of digital payments?
- How do these concerns vary across different demographic segments?

3.3 Awareness of Fraud Possibilities

- How aware are individuals of the possibilities of cyber fraud in the realm of digital payments and e-Services?
- Are there demographic factors that influence awareness levels?

3.4 Likelihood of Financial Loss in Fraud

- Who is more likely to experience financial losses in cases of cyber fraud during online transactions?
- Are there identifiable patterns related to age, income, or occupation that correlate with a higher likelihood of losing money?

3.5 Usage of Online Learning Platforms

- How are people utilizing online learning platforms, and is there a correlation between their engagement with digital payments and participation in online learning?
- Can demographic factors provide insights into the adoption and usage patterns of online learning platforms in conjunction with digital transactions?

By addressing these questions through a comprehensive analysis of the gathered data, we aim to contribute valuable insights that can inform strategies to promote the safe adoption of digital payments, mitigate cybersecurity risks, and understand the intersection between digital transactions and online learning behaviours.

4 Methodology

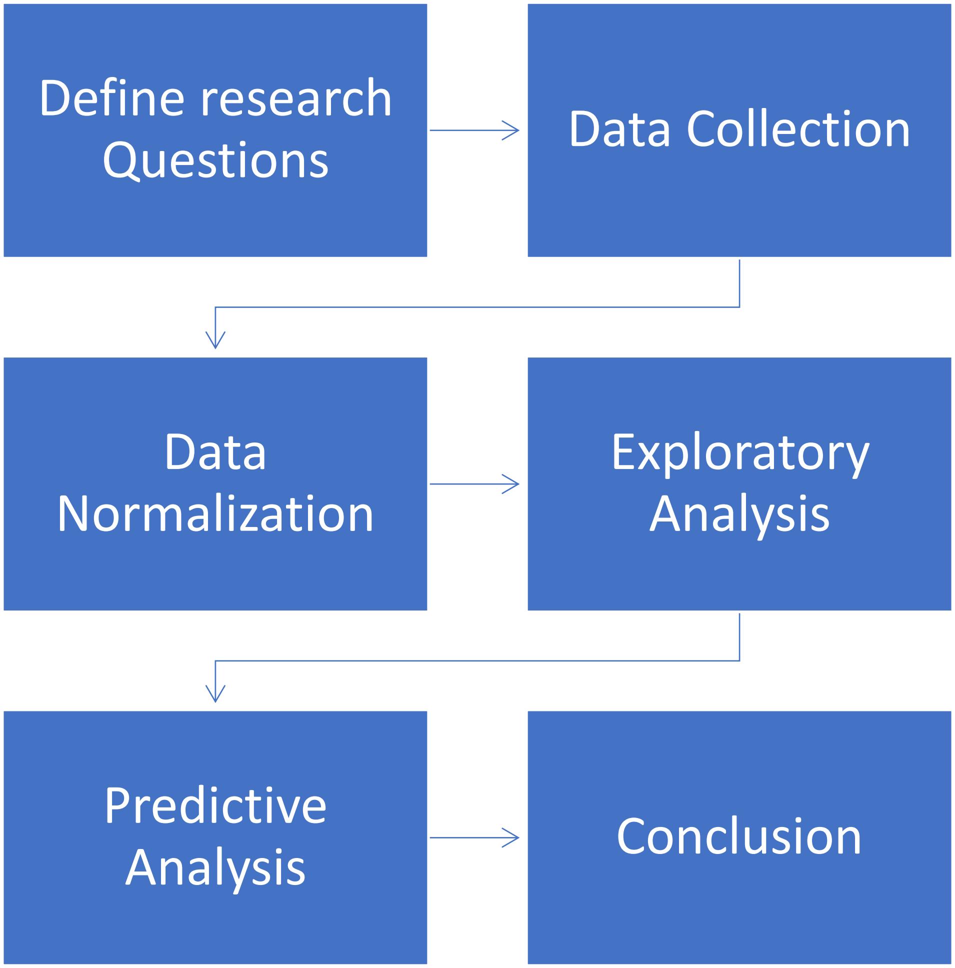


Figure 1 Project Methodology

5 Data Collection

5.1 Target Population

To capture user perception over wide age group, income group and service group, we targeted population which can cover college students, professionals, home maker as well as those who are in their matured stage of service.

College going students provide unique insights on their comfort and concern over usage of various online platforms for payment, purchase and learning. Home maker provides very distinct view being different challenges faced by them which may include type of joint or nuclear family. Professionals from various fields bring different perspectives for their experience and challenges faced by them. Businessman by economic trends market dynamics may give different insight of online platform usage which may have different perceptive after covid19 and their comfort of online transaction over cash transactions.

To ensure respondents comfort for answering relevant questions we limited questions which are not specific any individual for example income data, business data, family type, etc. This will ensure data privacy as well.

Geography for the sample is planned in Ahmedabad and Target sample size was at least 100+

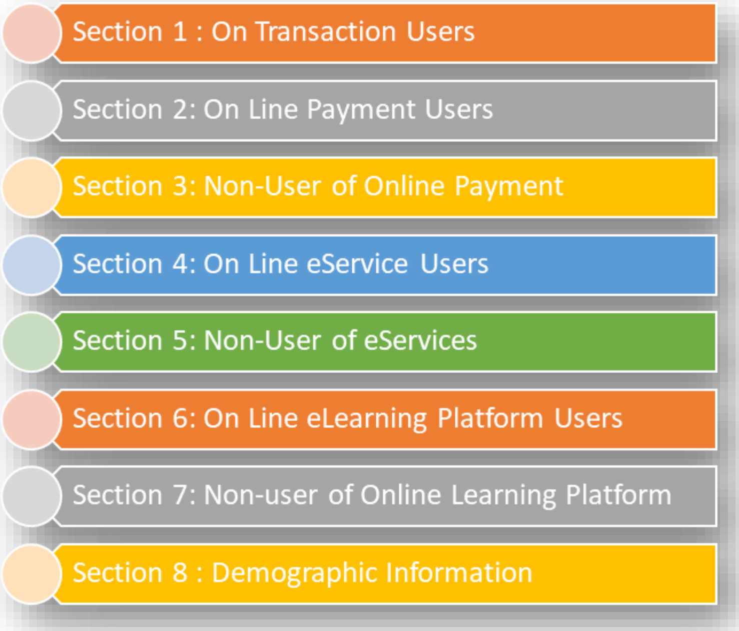
5.2 Survey Type

Our intention was to have primary data for our project hence out of various options of face-to-face interview, paper survey questionnaire, we decided to go for google survey and defined our actions accordingly.

5.3 Survey Questionnaire Design

While designing questions we have targeted following information and accordingly divided various sections in google form.

- Reponses by a range of factors: Age, Income, Occupation
- How people are concerned about Cyber Security
- How much people are aware of the possibilities of fraud?
- Who is likely to lose the money in fraud?



Section 1 : On Transaction Users

Section 2: On Line Payment Users

Section 3: Non-User of Online Payment

Section 4: On Line eService Users

Section 5: Non-User of eServices

Section 6: On Line eLearning Platform Users

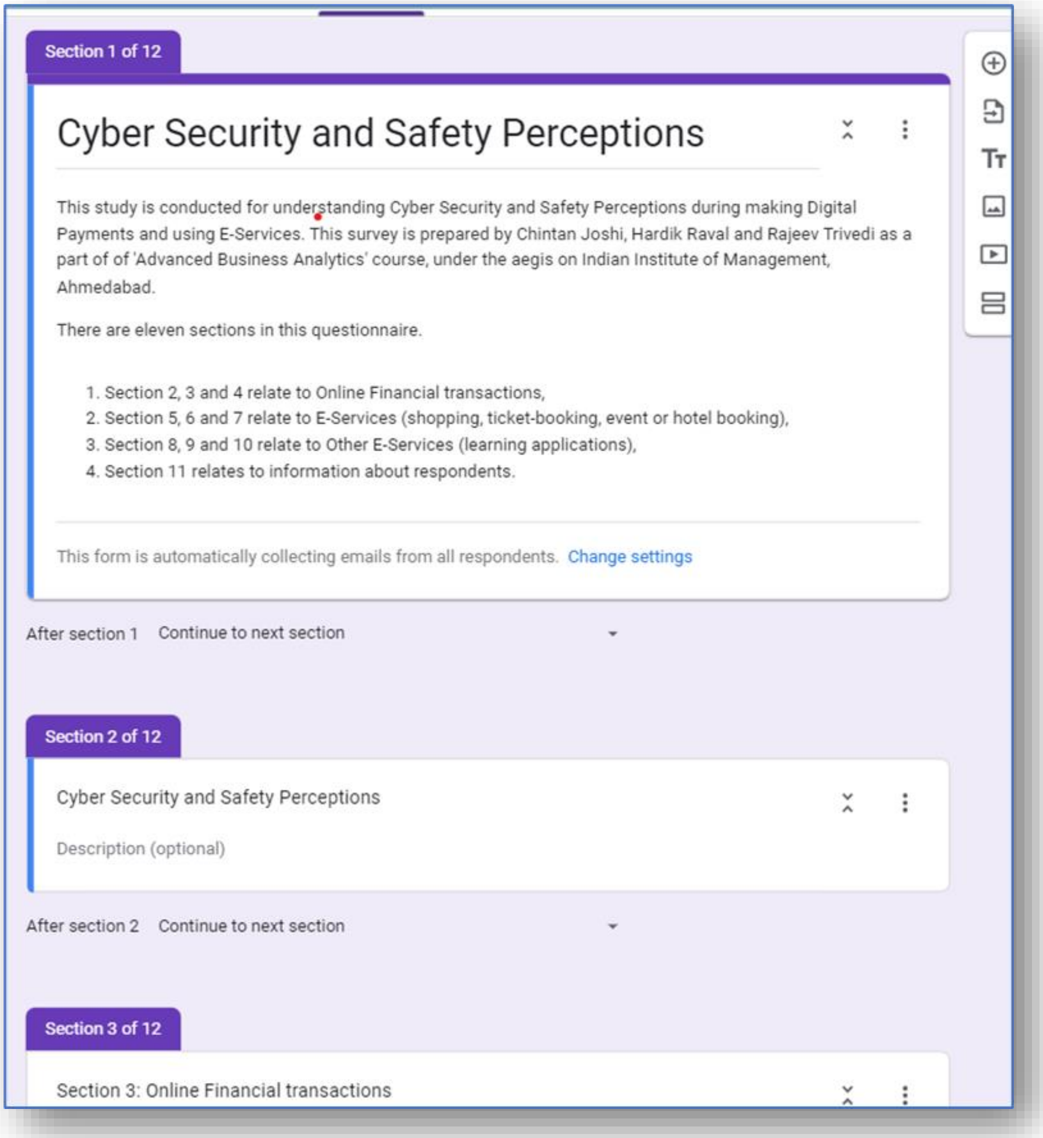
Section 7: Non-user of Online Learning Platform

Section 8 : Demographic Information

Figure 2 Survey Questionnaire Sections

Questionnaire has major eight sections. Wherein we tried to collect information regarding respondents view related to Online transactions and if they do so we asked other questions.

If respondents are not doing online transactions, then we did not ask any other questions other than understanding demography. We also created sections for various online usage which includes Online Payment, Online Services, Online Learning. For every section we asked respondents if they are using online platforms and if the answer is “No” then we route questions to the next section. The questionnaire covers 54 questions combining all sections.



Section 1 of 12

Cyber Security and Safety Perceptions

This study is conducted for understanding Cyber Security and Safety Perceptions during making Digital Payments and using E-Services. This survey is prepared by Chintan Joshi, Hardik Raval and Rajeev Trivedi as a part of of 'Advanced Business Analytics' course, under the aegis on Indian Institute of Management, Ahmedabad.

There are eleven sections in this questionnaire.

1. Section 2, 3 and 4 relate to Online Financial transactions,
2. Section 5, 6 and 7 relate to E-Services (shopping, ticket-booking, event or hotel booking),
3. Section 8, 9 and 10 relate to Other E-Services (learning applications),
4. Section 11 relates to information about respondents.

This form is automatically collecting emails from all respondents. Change settings

After section 1 Continue to next section

Section 2 of 12

Cyber Security and Safety Perceptions

Description (optional)

After section 2 Continue to next section

Section 3 of 12

Section 3: Online Financial transactions

Figure 3 Google survey form sample page

5.4 Survey Method

Due to time constraints, we decided to go for online survey instead of personal survey to various area. We agreed that google survey form sharing via email, WhatsApp or other social media can have wider reach and data can be collected in shorter duration. So, we emailed our google survey form survey as well as share link on WhatsApp.

Prior to sharing google survey we have taken dummy trials on how user feels and how much time user need to spend on filling survey.

Few close respondents were interviewed as for first pilot run of survey form and noted their suggestion to make it more convenient for respondents to answer questions.

Data Collection period was planned for maximum two weeks; however, we could collect most of the data in one week of time.

5.5 Data Collection, Cleaning and Validation

Google form (Google Survey Form, 2023) data have been collected in Excel file which have been further reviewed for cleanliness, some of responses which were descriptive were segregated and added extra columns for categorising them.

As our data collection was categorical most of answers we got were “Yes” or “No” and in some questions we also tried to measure depth of particular variable hence options were mentioned which were ordinal in nature. We kept it ordinal with an intention to find out user’s strong associations for particular variable, however in our analysis we did not focus much on same and converted original data to “Yes” and “No” only. For example, for a question’s response we asked about frequency of usage in which respondents were having options to answer either “Regularly “, “Occasionally”, “Rarely”, “Never” so for the said question and its importance as critical variable we only considered value as “Yes” for responses “Regularly” and for reaming responses like “Occasionally” “Rarely” “Never” we gave value “No”.

For deriving Exploratory Analysis, we need to clean data further, however for Predictive Analysis we converted data to Binary. So “Yes” responses have been given value as “1” and “No” responses have been given value as “0”.

We have used clean data to run in “R Program” to validate if this is working fine, for which our results were good, however we could notice that as data variables are too large it might be good break data in certain sections for in depth analysis.

5.6 Data Primary Analysis and Improvements

On duly validation of data various predictive models run which were working fine in terms of results however while inferring these variables we have noted it does not give us true meaning of our objectives which we were looking, at this joint we took help of mentor and noted that as data is skewed towards more “ Yes” and little information about “ No” all results are also skewed towards “ Yes” hence it’s not giving true inferences. In other words, our data was unbalanced and need to be balanced. So, we use SMOTE function and balanced data for further analysis.

5.6.1 SMOTE

Synthetic Minority Oversampling Technique is called as **SMOTE** in short form. SMOTE uses nearest neighbouring approach to generate minority class samples. There are two methods to use SMOTE, under sampling and Over Sampling, when we have large data set for example 1000 responses in which data is skewed towards one specific response for example 800 responses are “Yes” and 200 responses are “No” then we go for under sampling method in which majority class data is reduced to lower size for example 800 becomes 400 and so we avoid any further skewness of data. We may have to go for few trials before we reach to particular number of data set, which is usually done on test data.

Other method is Oversampling method in which we increase minority class data by synthetically duplicating responses. Method is used is of k-nearest neighbours. Algorithm identifies feature vector and its nearest neighbour the difference between two is calculated and multiplied with random number 0 and 1 and new data point is created, this process is repeated until desired balance is achieved. There are some limitations also of SMOTE which need to be noted while using this tool. The limitations are due to data which is not new but from the sample data so it does not provide any new information, sometimes it can overfit also. If data has issue on separation of various category, then this may not beneficial, this can also lead to different inferences when data is too small or too large.

Considering all above points, we have validated our model results from SMOTE Test data and noted that it provides us logical and practical inferences of our study hence we accepted SMOTE data for further analysis. We have referred books/materials from library (IIM Library, 2023)

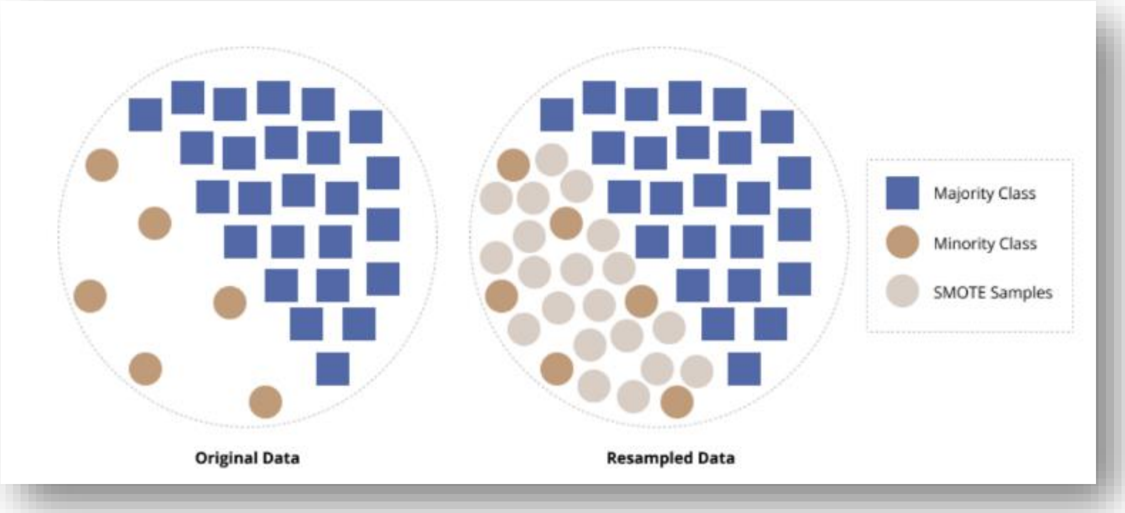


Figure 4 SMOTE data method

(Kegelmeyer, 2002)

(Managing imbalance Data Set with SMOTE in Pythone, 2020)

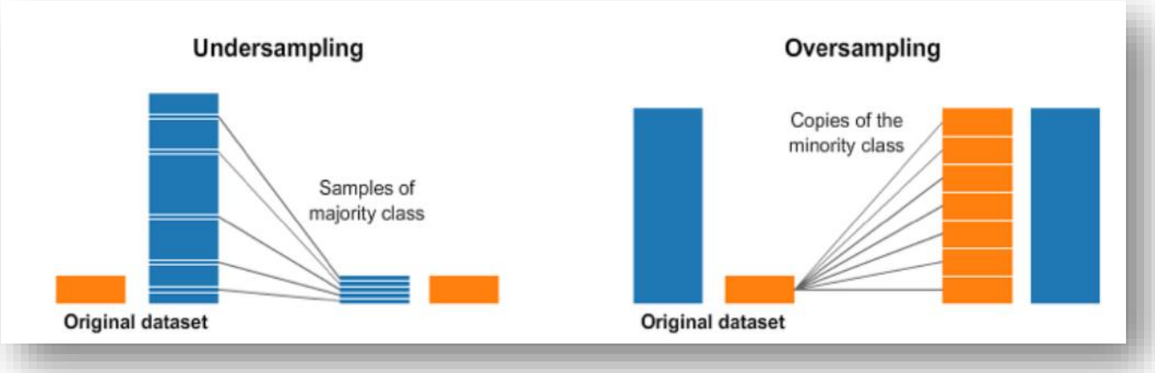


Figure 5 SMOTE data types

6 Data Analysis

6.1 Exploratory Data Analysis

For the Exploratory Data analysis, we used Tableau software, where in main validated file has been uploaded without conversion to binary. Various worksheets created for all the target variables. Used Tableau inbuilt features of calculating % of specific variable out of total variable. Some data were having multiple responses (Ordinal Type or Different Category Type).

Tableau inbuilt feature of splitting of data and pivot table helped for understanding each variable separately. All respective worksheets have been combined in Dashboard for overall inference about specific variables and related responses.

6.1.1 Online Application Users

Figure 6 Digital Platforms

We studied how many users are using online platform for various applications like payment, purchase, or learning.

FIGURE 7 illustrates about.

- Online Users for Payment Applications is 96%
- Online Users for Purchase Applications is 93%
- Online Users for Learning Applications is 74%

So, from Figure we can infer that Online payment applications is most significant among other applications.

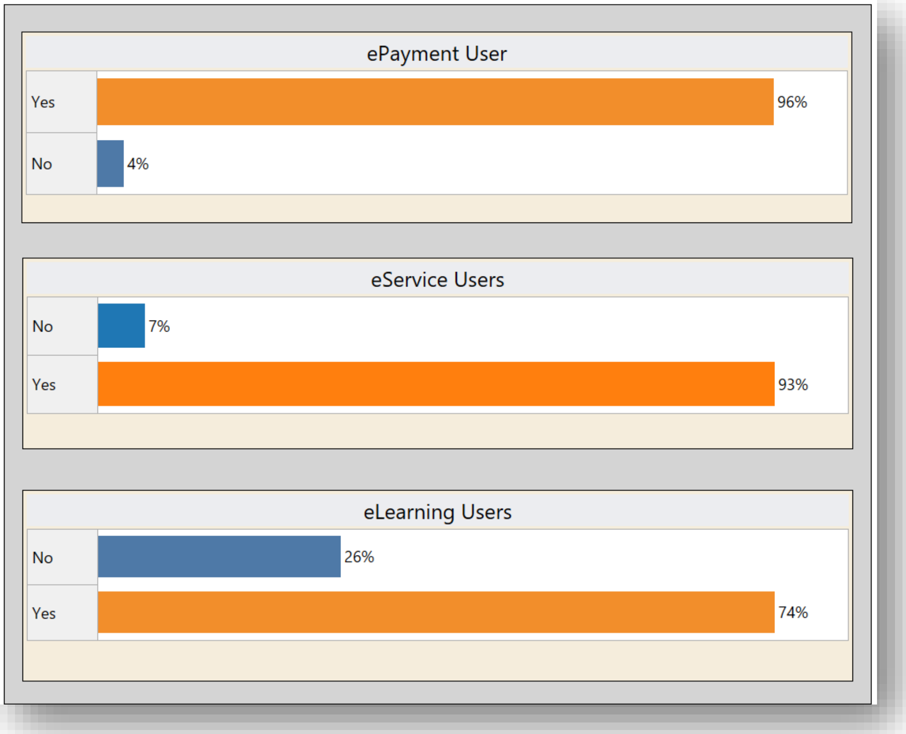


Figure 7 Online Application Users

6.1.1.1 Reasons for not using Online Applications

6.1.1.1.1 NON-USERS OF ONLINE PAYMENT APPLICATIONS

From **Figure 8** we can infer that those who are not using Online Payment is because they feel risk of Fraud during transactions and some of them don't use because of transaction charges.

Reasons for Not Using

- Risk of Fraud 82%
- Due to Transaction Charges 18%

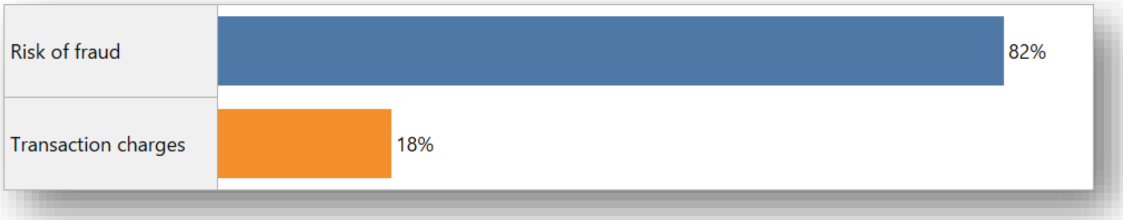


Figure 8 Reasons for not using online payment.

6.1.1.1.2 NON-USERS OF ONLINE PROCUREMENT APPLICATIONS

From **FIGURE 9** we can understand about reasons of respondents who are not using Online eservices i.e. online procurement and following are main reasons.

Reasons for Not Using

- Privacy and Security Concern 63%
- Lack of In-Store Engagement 15%
- Vulnerabilities to Frad 11%
- Lack of Knowledge 7%
- Shipping Problems 4%

We can summarise that privacy and vulnerability are major factors of user not using Online Procurement platforms.

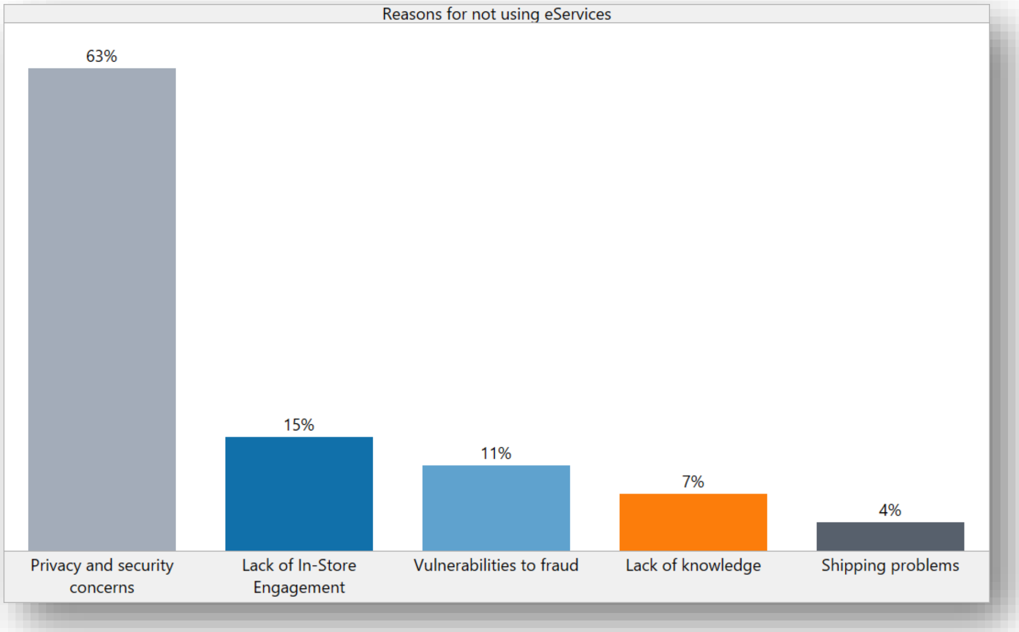


Figure 9 Reasons for not using eServices.

6.1.1.1.3 NON-USERS OF ONLINE LEARNING APPLICATIONS

FIGURE 10 Illustrates about top two reasons for not using online learning applications. We can see that users do not feel risk of cyber fraud for the online learning; however, they don't use because of following two main reasons,

- We may infer that as learning applications use may not require commercial transactions as compulsory, user may not see risk of cyber fraud.
- Some of respondents have express about their fear of unknown, this bring interesting insights into understanding how Online learning applications can be made more popular to reduce fear, and that's an opportunity for further research in this area, however as our objective is limited, we consider this point for further research.

Reasons for Not Using Learning Applications

- Lack of motivation 80%
- Fear of Unknown 20%

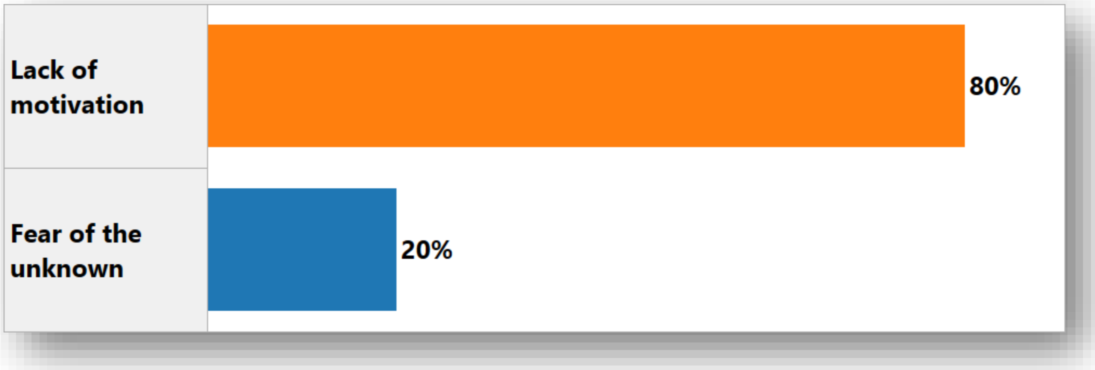


Figure 10 Reasons for not using Online Learning Applications

6.1.2 Demography

Respondents Demographic profile is well illustrated in **FIGURE 11** , this provides clear representation of diversity regarding.

- Gender
- Age
- Education
- Occupation and
- Income.

The Majority of respondents are male and representing 75% of total respondents. Among age category we can see that nearly

- half of total respondents are Young having age less than 25 years, followed by
- 26% between age of 25 to 40 and 26% between age of 40 to 61.

This means our respondents covers well spread diversity among different age groups which will help us for understanding different perspectives of their experiences.

Similarly, we can see that almost.

- half of respondents are students and remaining.
- most are working professionals.
- very little who are home maker.

With these two distinctive classes we will have interesting perspective of risk aversion capability between students and professionals as they will be at different stages of their experience.

We can also see that most respondents are well educated and 21% represents even higher education, this also brings good information about their awareness related to online applications as well as various tools for performing transactions and that may not limit their opportunities if they wish to use online services, this is very important factor for our study.

Income category analysis suggests that.

- most respondents are of middle-income categories, this can be interesting study during analysis about their online platform usage frequency,
    - as it might be possible that higher income category respondents may like to use traditional way of transaction instead of online transaction due to high risk of cyber fraud

We will be having interesting insight of various demographic factors while we do statistical analysis with help of various predictive analysis tools.

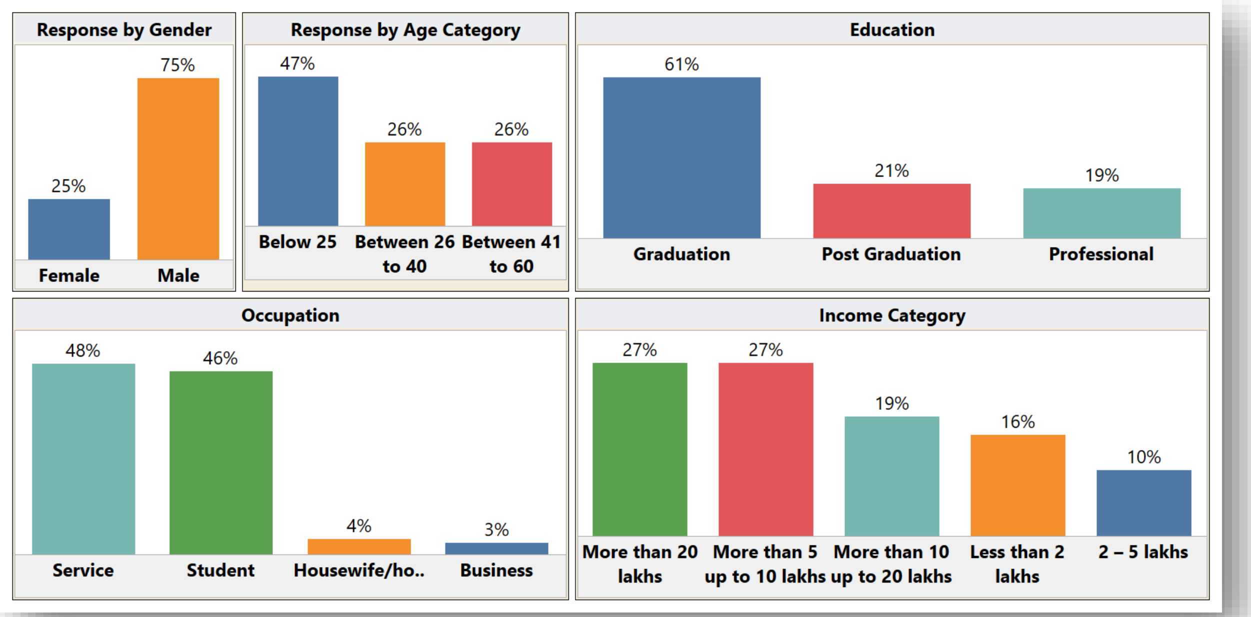


Figure 11 Demographic information of the data

6.1.3 Transactions Methods Used for Online Payment

While analysing further about which online payment method is most popular among users, we found that.

- UPI accounted for maximum % of user which is 63% among total respondents which is very close to current trend.

FIGURE 12 illustrates about UPI payment growth year over year which is published by website “money control “in their article “Economic Survey 2023:

- UPI accounted for 52% of India's total digital transactions in FY22” where in data source is from “National Payments Corporation of India (NPCI)”

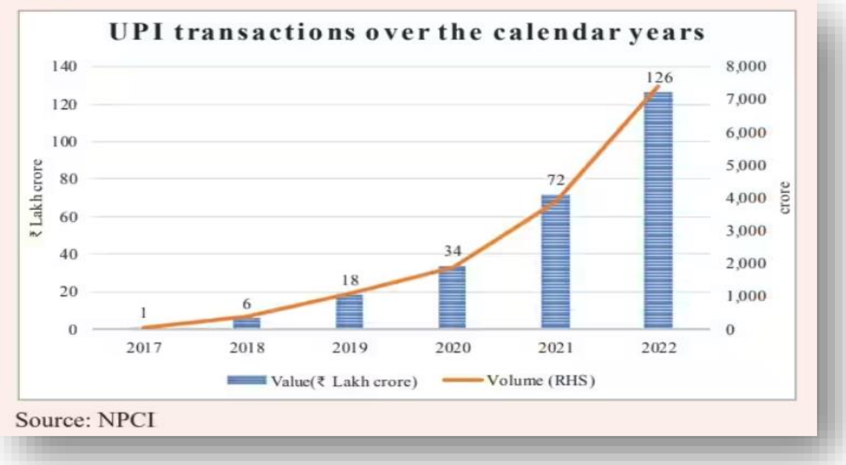


Figure 12 UPI Transactions in India

(Economic Survey 2023, 2023)

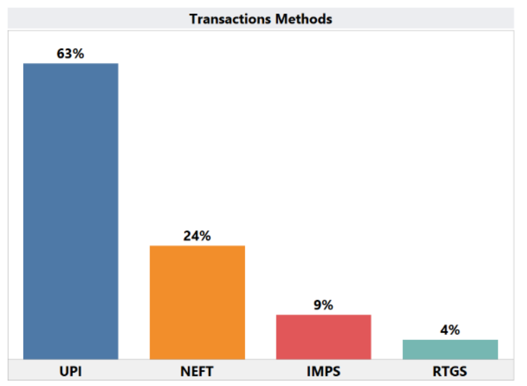


Figure 13 Online Payment Methods

Other popular payment methods are.

- NEFT 24% followed by
- IMPS 9% and
- RTGS 4%

this says about convenience of UPI payment methods and credit goes to ease of doing online transactions which also increases risk of cyber fraud.

6.1.4 User Experience about Cyber Fraud

From **FIGURE 7** we have noted that 96% of respondents are using online payment platforms, so we also studied while doing so what is their experience in response to cyber fraud.

FIGURE 14 illustrates that.

- 13% of respondents have experienced cyber bullying while performing online payment transactions and
    - 16% have lost money who has experienced cyber bullying.
- Percentage of respondents about cyber bullying concern is 65% as they are concerned about cyber bullying.
- Around 24% of respondents has little concern about cyber bullying and
- 11% are risk takers as they know they should be concerned about cyber bullying, but they are not concerned, primary reason for the could be because.
    - nearly 50% of respondents are young generation.

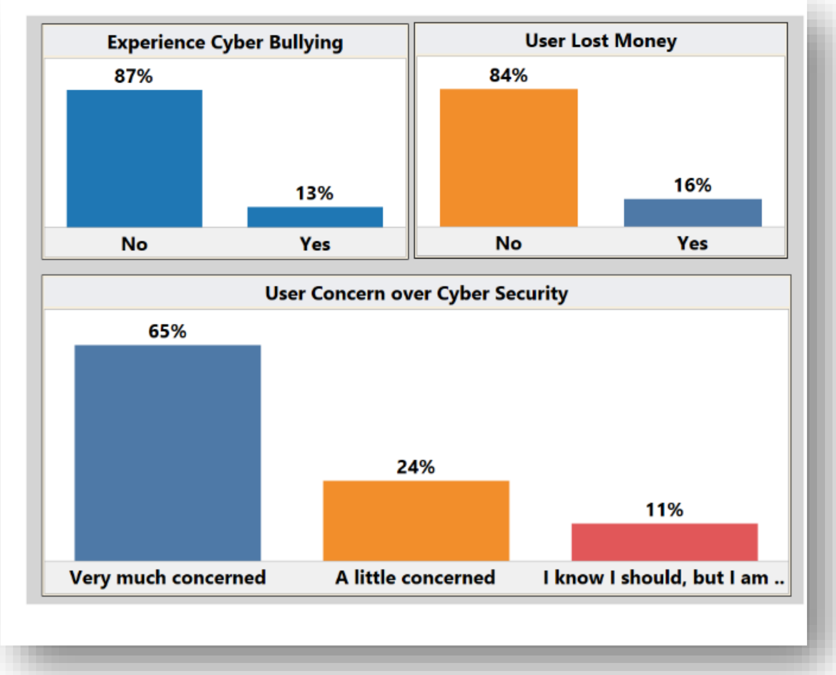


Figure 14 User Experience and Perception Towards Cyber Fraud

6.1.5 User Experience about Artificial Intelligence Tool (Chatbot)

While performing Online Payment transactions, there are several incidents wherein users have queries for which they need help for the resolution, through questionnaire we tried to find out whether they are concerned about using Artificial Intelligence Tools like Chatbot? as during interaction with Chatbot they might need to provide confidential information and if so, how are they perceiving their risk about cyber fraud during same interaction.

From **FIGURE 15** we can note that

- 26% respondents are frequently using chatbot as their tools for resolving queries.
- Around 35% of respondents uses chatbot sometimes and
- 39% of respondents never used chatbot.

With this we can infer that chatbot is taking popularity as one important tool for resolving queries quickly.

We can also infer from figure that,

- 43% felt that its very useful tool for them and
- 41% believes that it may compromise their privacy while using such Artificial Intelligence tool, there are.
- 32% of respondents says about their concern over privacy,

however remaining respondents seems more comfortable, so we can infer that there is good opportunity for establishing more AI tools and more awareness among users for making it more effective. More research in this area will be very interesting and can bring different insights of digital payment and artificial intelligence collaboration.

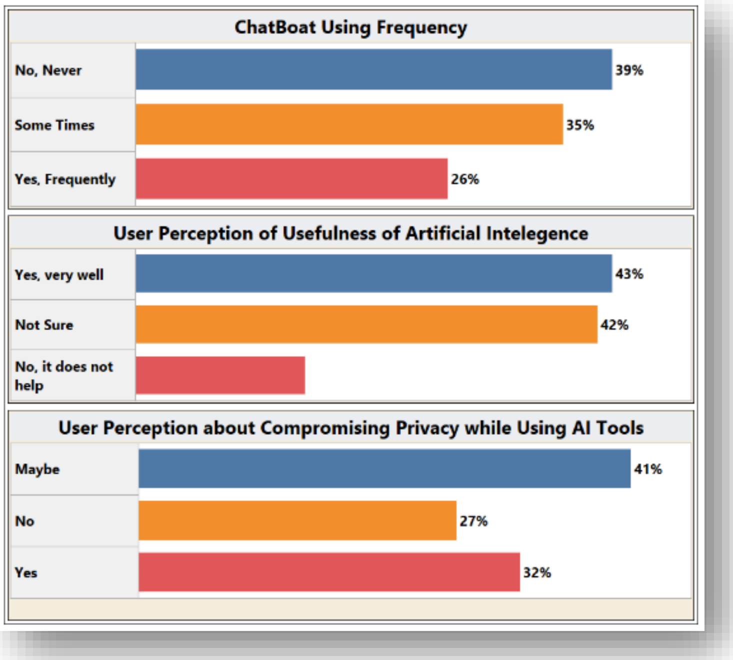


Figure 15 Usage of Chatbot (artificial intelligence tools)

6.1.6 Authentication Method and Secured Webservice Usage

While performing online payment transactions we tried to find out how much user is aware about risk involved when using public network or when they try to use Web address which are not secured i.e. https://

FIGURE 16 provides very important information about user's risk exposure as

- 54% of respondents uses public network sometimes if we add 5% who uses regularly this total is almost 60%
  - which shows very high exposure about Risk while using public network.
- Around 19% respondents do not ensure whether website is secured or not that further added their risk of cyber fraud. However, we can note that.
- 42% ensures that web address is secured while they perform online transactions, followed by
- 38% who uses secured web address sometimes, we can infer from this about well awareness among users for using secured web address while performing online transactions.
- Fingerprint authentication is most popular authentication method which tells about most users using mobile as their instruments for online transactions, followed by security pin, password and FaceID.

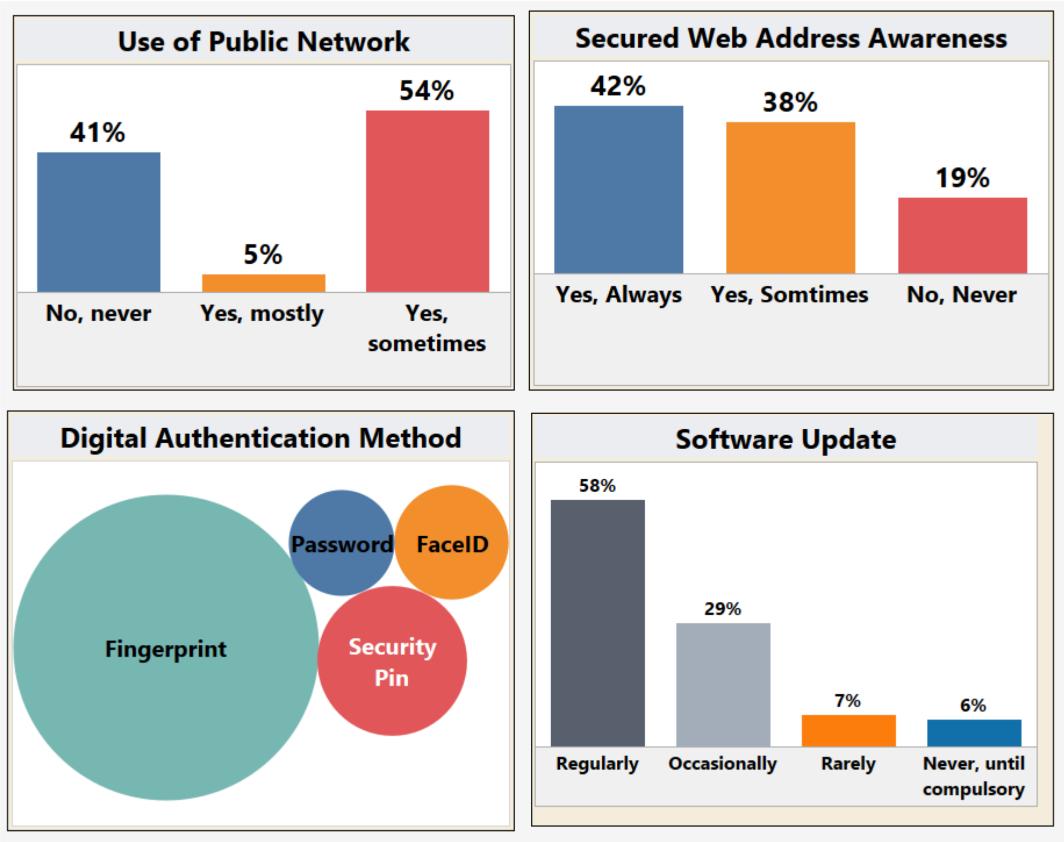


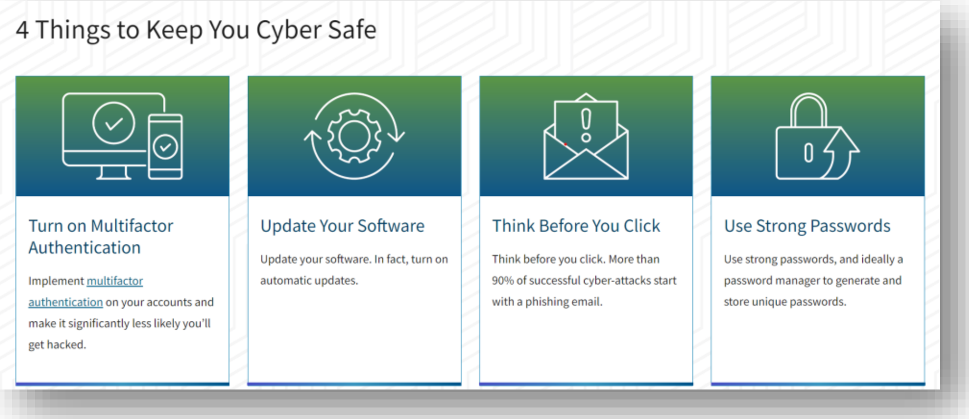
Figure 16 Authentication Method and Secured Webservice Usage

6.1.6.1 Cyber Security and Infrastructure Security Agency (CISA)

CISA is short name of “Cyber Security and Infrastructure Security Agency “of America’s Cyber Défense Agency which is the operational lead for federal cybersecurity and the national coordinator for critical infrastructure security and resilience.

Based on CISA’s study they recommended four things to protect against cyber fraud. **Figure 17** illustrates these four things which helps to combat cyber fraud.

- Turn on Multifactor Authentication
- Update Software Regularly
- Think Before you Click.
- Use Strong Password



4 Things to Keep You Cyber Safe

Turn on Multifactor Authentication

Implement multifactor authentication on your accounts and make it significantly less likely you'll get hacked.

Update Your Software

Update your software. In fact, turn on automatic updates.

Think Before You Click

Think before you click. More than 90% of successful cyber-attacks start with a phishing email.

Use Strong Passwords

Use strong passwords, and ideally a password manager to generate and store unique passwords.

FIGURE 17 CISA Recommendation for Cyber Fraud Prevention

(CISA, 2022)

When we consider CISA’s guidelines and analysis from our responses, we can infer that.

- User has good amount of awareness in this area as
  - people are using multifactor authentication,
  - using secured website,
  - avoiding public network as well as
  - they are updating software regularly.

So, if there is continuous awareness drive there is high possibility for reducing extent of cyber fraud.

6.1.7 Online Buying

Online buying or ecommerce is referred as eServices for our project wherein we focused on understanding users' perspective towards cyber fraud related concern and if so, what are the factors effecting it.

From **FIGURE 18** we can understand that

- 92% online buyers are aware about cyber fraud as they believe cyber fraud can happen during performing online transactions as it correspondence to commercial transactions.
- Most users have started online buying post 2015 and their preferred platform is.
- Amazon as 93% respondents voted Amazon compared to other options.

Their inclination towards online buying is primarily due to

- Lower prices,
- Convenience and
- Wide product availability.

However, though it looks on prima facie about time savings, but user has considered it as least factor as it felt by 7% as time is saved during performing online buying transactions.

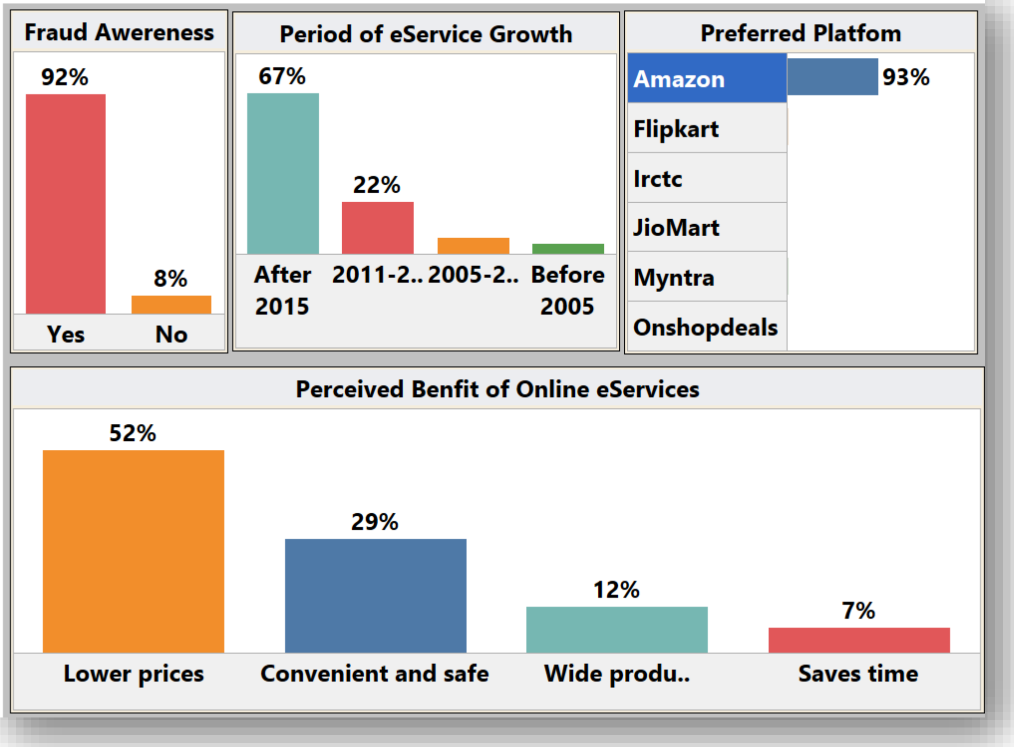


Figure 18 Online Buying (eServices) Experience

6.1.8 Online Learning

Along with user perception about cyber fraud in Online Learning, we also tried to find out which platform is most popular among users.

FIGURE 19 Illustrates that out of total respondents 74% of respondents are using online learning platforms and them.

- Preferred learning platforms are.
  - YouTube (46%),
  - Udemy (17%),
  - Coursera (14%).

It's interesting to know that approximately.

- 1/3 of respondents were using online learning platforms even before Covid19 and approximately.
- 2/3 of respondents started online learning post covid19.

Users' perception about fraud in Online learning is about 13% and most of them feels it has not much risk compared to Online payment and Online Buying.

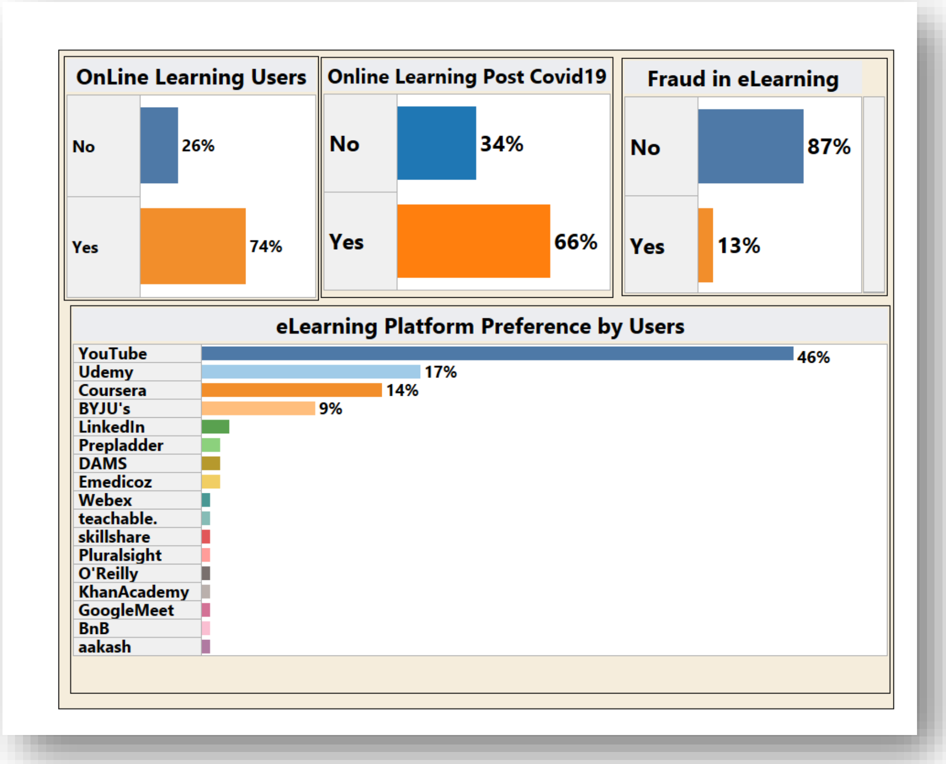


Figure 19 Online Learning Platforms

6.1.9 Summary of Exploratory Analysis

Entire exploratory analysis is based on Tableau software and following is key summary of entire exploratory data analysis.

We focused following areas of our interest,

- Digital platform usage,
- Reasons for not using specific online applications,
- User demographics.

Key findings include high UPI transaction percentages, concerns about cyber fraud, and the popularity of artificial intelligence tools such as chatbots. The analysis revealed insights into authentication methods, online buying behaviour, and online learning preferences.

Following key insights can be derived from our exploratory analysis,

- Digital Platform Usage: UPI emerged as the most popular online payment method (63%), exceeding national trends.
- Cyber Fraud Concerns: A significant portion of respondents (65%) expressed concerns about cyber fraud, particularly in online payment transactions.
- Artificial Intelligence Tools: Chatbots were used by 26% of respondents, with 43% finding them very useful but 41% expressing privacy concerns.
- Authentication Methods: Fingerprints were the most popular authentication method, with 60% using public networks, highlighting a potential risk.
- Online Buying: 92% of online buyers were aware of cyber fraud risks, and Amazon was the preferred platform for 93% of respondents.
    - Online Learning: YouTube was the most popular online learning platform (46%), and 74% of respondents engaged in online learning, with 13% expressing fraud concerns.

6.2 Predictive Analysis

6.2.1 Logistic Regression Model to find how people are concerned.

6.2.1.1 Selection of Response Variable

The selection of the response variable as a binary indicator of whether people are concerned or not is based on several considerations aligned with the goals of the analysis.

- Clear Focus: A binary choice simplifies the analysis, making it clear whether individuals express concern about security on the internet.
- Alignment with Research Question: It directly relates to our research goal of understanding factors influencing people's concerns.
- Compatibility with Logistic Regression: Binary responses work well with logistic regression.
- Easy Interpretation: Coefficients in logistic regression directly link to the likelihood of expressing concern, making results easy to understand.
- Actionable Insights: The binary variable guides targeted strategies to address concerns and promote positive behaviours.
- Statistical Suitability: Binary outcomes simplify statistical analyses, facilitating model fit assessments and hypothesis testing.
- Practical Relevance: Decision-makers often seek insights on the prevalence of concerns, aligning with real-world needs.

6.2.1.2 Explanatory Variable Selection:

Explanatory variables were selected to find pattern of how demographic information affects people’s concern about security on internet.

Response Variable	Values
Concern_Binary	1 for Concerned, 0 for less concerned
Explanatory Variables	**Values**
is_male	1 for male, 0 for not male
is_female	1 for female, 0 for not female
age_group	1 for below 25, 2 for 25-40,3 for above 40
is_graduate	1 for graduate, 0 for not graduate
is_postgraduate	1 for postgraduate, 0 for not postgraduate
is_schooling	1 for schooling, 0 for not schooling
is_other_education	1 for other education, 0 for not other education
is_business	1 for business, 0 for not business
is_service	1 for service, 0 for not service
is_student	1 for student, 0 for not student
is_homemaker	1 for homemaker, 0 for not homemaker
income_group	2 – 5 lakhs: 1, Less than 2 lakhs: 2, More than 10 up to 20 lakhs:3, More than 20 lakhs:4, More than 5 up to 10 lakhs:5

Table 1 Variable Selection for Logistic Regression Model to find how people are concerned.

6.2.1.3 Data Preprocessing

- Converted columns like age_group and income group from text response like

Income Group	Variable
2 – 5 lakhs	1
Less than 2 lakhs	2
More than 10 up to 20 lakhs	3
More than 20 lakhs	4
More than 5 up to 10 lakhs	5

Table 2 Income group encoding in logistic model.

Age Group	Variable
Below 25	1
Between 26 to 40	2
Between 41 to 60	3

Table 3 Age group encoding in logistic model.

6.2.1.4 Data Scaling

Since the response variable concerned was more skewed toward more people concerned, we use upscaling techniques in the R code to upscale data using library caret.

6.2.1.5 One-Hot Encoding

- Introduced dummy columns representing different categories to capture categorical information. e.g. is_male, is_female, is_student.

6.2.1.6 Collinearity Assessment

- Use Variance Inflation Factor (VIF) analysis to identify and address multicollinearity among explanatory variables.
- Removed variables with high VIF values to improve model stability and interpretability.

6.2.1.7 Variable Selection

- Selected is_student and age_group as these were the statistically significant data with the sample data.

6.2.1.8 Model Building

Built logistic model using R language.

```
summary(final_model)
# Call:
#   glm(formula = concern_binary ~ age_group + is_student, family =
binomial,
#       data = data)
#
# Coefficients:
#   Estimate Std. Error z value Pr(>|z|)
# (Intercept)   5.0151     1.6287   3.079 0.002076 **
#   age_group    -1.7268     0.6137  -2.814 0.004897 **
#   is_student   -3.5396     1.0728  -3.299 0.000969 ***
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
# Null deviance: 142.79  on 103  degrees of freedom
# Residual deviance: 128.11  on 101  degrees of freedom
# AIC: 134.11
#
# Number of Fisher Scoring iterations: 4
# Conclusion ---------------------------------------------------------
# Overall, this model suggests that both 'age_group' and 'is_student'
# are important predictors of the binary response variable
'concern_binary',
# and the model provides a good fit to the data
```

Code Snippet 1 Logistic Model for Predicting Concern About Security on Internet

6.2.1.9 Model Evaluation

Model checked for accuracy by following metrices.

Metric	Training Data	Validation Data
TP	25	37
TN	52	27
FP	33	21
FN	6	19
Accuracy	0.614	0.595
Precision	0.431	0.638
Recall	0.806	0.661
Specificity	0.612	0.563
F1 Score	0.562	0.649

Table 4 Accuracy matrix for the cyber security concern analysis

6.2.1.10 Conclusion on Model:

- The model's intercept and both explanatory variables (age_group and is_student) are statistically significant, indicating that these variables collectively contribute to explaining the variation in the likelihood of expressing concern for security on internet.
- The negative coefficients for age_group and is_student suggest that, holding other variables constant, older age groups and being a student are associated with lower log-odds of expressing concern.
- The p-values for all coefficients are below conventional significance levels (e.g., 0.05), suggesting robust statistical evidence for their significance.
- In conclusion, this logistic regression model provides insights into the influence of age group and student status on the likelihood of expressing concern, offering a statistically significant understanding of the factors associated with the binary response variable.

6.2.2 Random Forest Analysis to find fraud awareness.

6.2.2.1 Purpose

We wanted to predict whether people are aware of fraud in digital payments using the response in our questioner.

6.2.2.2 Variable Selection

Response Variable	**Values**
fraud_awareness	1 for Yes, 0 for No
Explanatory Variables	**Values**
doing_transaction	1 for yes 0 for no
transaction_method	Credit cards, ECS, IMPS, NEFT, Net banking, RTGS, UPI
concerned	A little concerned, I know I should, but I am not concerned, very much concerned
os_update	Never, until compulsory, Occasionally, Rarely, regularly
public_network	No-never, Yes- mostly, Yes- sometimes
information_security_knowledge	Yes, No
data_backup	Yes, No
data_backup_scheule	Occasionally, Rarely, Regularly
network_security	Yes, No
lost_money	Yes, No
two_factor	Yes, No
gender	Female, Male, Prefer not to say
Age	Below 25, Between 26 to 40, Between 41 to 60
Education	Graduation, Others, Post Graduation, Schooling
Occupation	Business, Housewife/househusband, Service, Student
income	2 – 5 lakhs, Less than 2 lakhs, More than 10 up to 20 lakhs, More than 20 lakhs, More than 5 up to 10 lakhs

Table 5 Variable Selection for The Random Forest Method for Fraud Awareness

6.2.2.3 Upscale data

- Used SMOTE for upscale data as our data has skewness toward awareness of fraud.

6.2.2.4 Training testing division

- Split data into train and testing sets

6.2.2.5 Model

Fitted the random forest method like below.

```
explanatory_variables <- c("doing_transaction","transaction_method
                           ","concerned","os_update","

public_network","information_security_knowledge
                           ","data_backup","data_backup_scheule",
                           "network_security",
                           "lost_money",
                           "two_factor",
                           "gender",
                           "Age","Education","Occupation","income")

formula <- as.formula(paste(response_variable, "~",
paste(explanatory_variables, collapse = "+")))




# Train the random forest model
rf_model <- randomForest(formula, data = train_no_missing_rows,
                         mtry=4,
                         ntree=500)
```

Code Snippet 2 Random Forest Model for Predicting Fraud Awareness.

6.2.2.6 Validation of model

Model gives us accuracy of 100% on training and testing data, it seems to be overfit however since the data sample were limited, we conclude this could happen with small dataset and having categorical values.

6.2.2.7 Important Variables

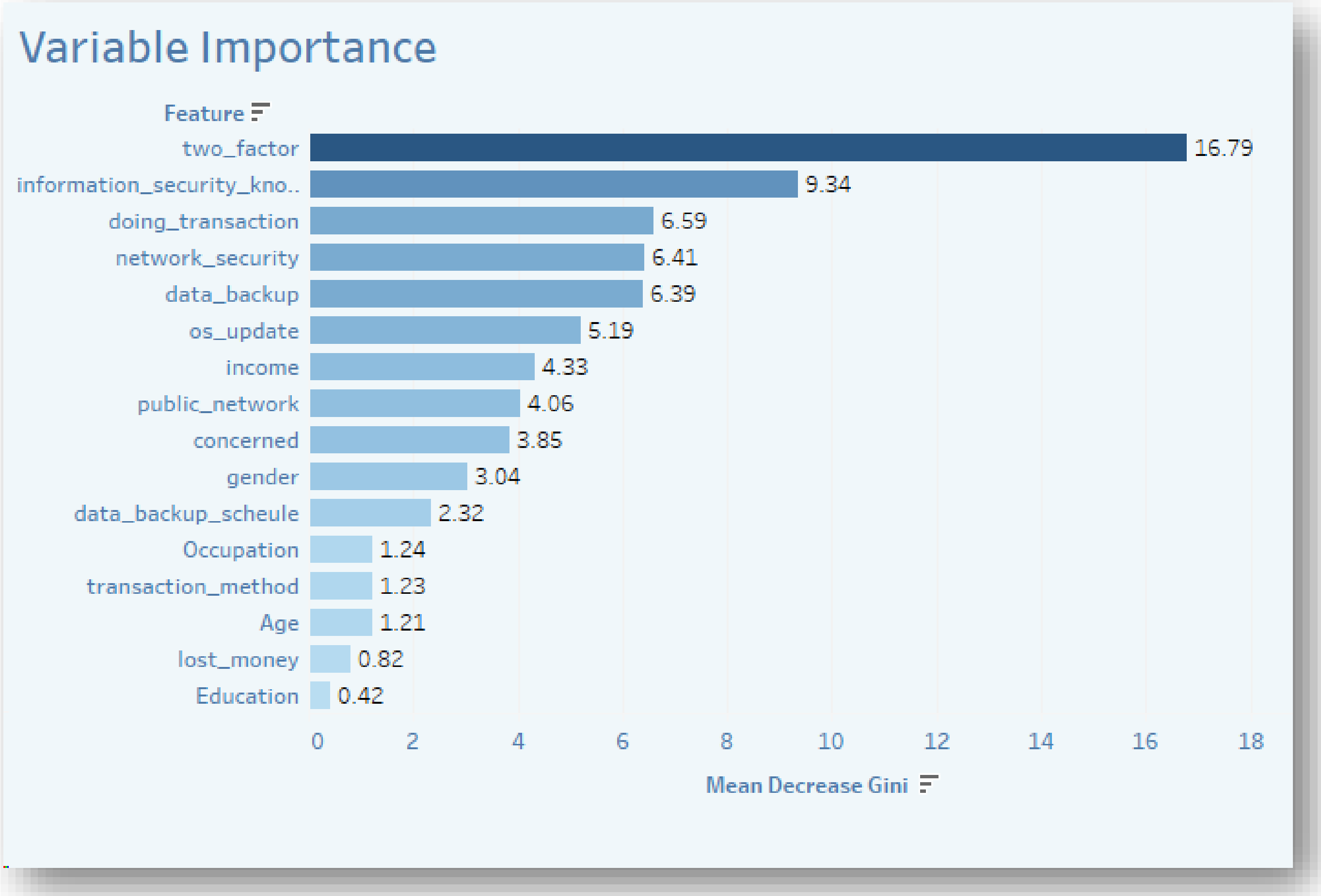


Figure 20 Important Variables for The Random Forest analysis on Fraud Awareness

6.2.2.8 Conclusion

From the analysis above, the random forest model can predict fraud awareness of the new observations and Two factor authentication, information security knowledge, network security, income are amongst the most significant variables influencing the awareness.

6.2.3 Market Basket Analysis to find the rules for the people concerned.

6.2.3.1 Purpose

To find the rules or combination of other factors that impact peoples' concern about security on internet. To answer this question, we applied Apriori algorithm to find set of rules among the data and filing rules where RHS is.

Concern = Very much concerned

With this we can find other influencing factors that attributes to people's concern on the security on internet.

6.2.3.2 Method

- Upscaled the data using caret library in R.
- Used minimum support as 20%
- Used Threshold for confidence as 70%

6.2.3.3 Derived Rules

We got many rules that were suggesting pattern of features in combinations, we filtered those rules where RHS was {Concern=Very much concerned}

LHS	RHS	Support	Confidence	Lift
{Age=Between 26 to 40}	=> {Concern=Very much concerned}	0.2211538	0.8214286	1.472906
{Occupation=Service}	=> {Concern=Very much concerned}	0.3365385	0.7142857	1.280788
{Age=Between 26 to 40, Occupation=Service}	=> {Concern=Very much concerned}	0.2211538	0.8214286	1.472906
{Gender: =Male, Occupation=Service}	=> {Concern=Very much concerned}	0.2884615	0.7692308	1.379310

Table 6 Market Basket Rules for people very much concerned about security on internet.

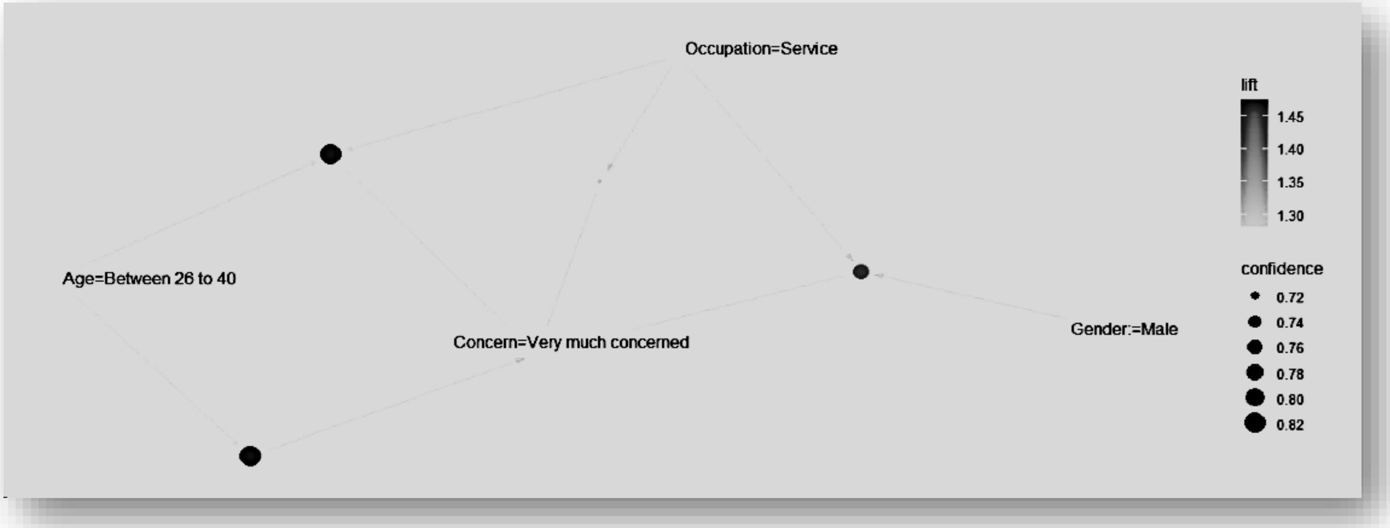


Figure 21 Rules for People Very Much Concerned About Security on Internet

6.2.3.4 Conclusion

People with following attributes are likely to have very much concerned about Cyber Security

- Age group 26 to 40
- Occupation: Service
- Gender: Male

6.2.4 Market Basket Analysis to find factors behind lack of fraud awareness.

6.2.4.1 Purpose

We wanted to predict who are the people that lacks fraud awareness, those are the people who can be knowing victim of fraud as they have less awareness that how prevalent is fraud in digital payments and fraud can happen to them.

For this we applied Apriori algorithm to find the rules where fraud awareness was no. With this we wanted to find other influencing feature that contributes to lack of fraud awareness.

6.2.4.2 Method for finding features for lack of fraud awareness.

- Upscaled the data using caret library in R.
- Used minimum support as 20%
- Used Threshold for confidence as 80%

6.2.4.3 Derived Rules

We got many rules that were suggesting pattern of features in combinations, we filtered those rules where RHS was {fraud_awareness=No}

LHS	RHS	Support	Confidence	lift
{information_security_knowledge=No,	{fraud_awareness=No}	0.02884615	1	26
data_backup=No,				
two_factor=No,				
Gender: =Female}				
{public_network=Yes, sometimes,	{fraud_awareness=No}	0.02884615	1	26
data_backup=No,				
two_factor=No,				
Gender: =Female}				
{transaction_method=UPI,	{fraud_awareness=No}	0.02884615	1	26
data_backup=No,				
two_factor=No,				
Gender: =Female}				
{transaction_method=UPI,	{fraud_awareness=No}	0.02884615	1	26
information_security_knowledge=No,				
data_backup=No,				
two_factor=No,				
Gender: =Female}				

Table 7 Market Basket Rules for Lack of Fraud Awareness

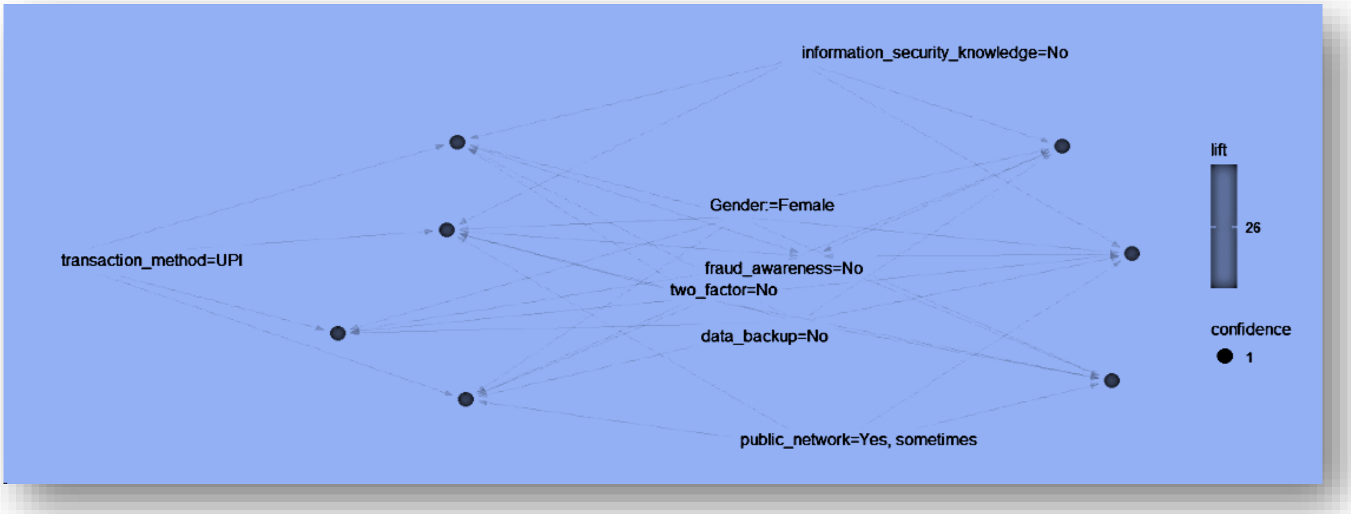


Figure 22 Rules for Lack of Fraud Awareness.

6.2.4.4 Conclusion

With above associate mining rules we could infer that data suggests people with following features lacks

Awareness about frauds

- lack of information security knowledge
- not doing data backup
- not using two-factor authentication
- and are female.

6.2.5 Market Basket Analysis to find online learning usage patterns.

6.2.5.1 Purpose

To find the patten of usage of online learning platforms.

6.2.5.2 Method for finding patterns in online learning.

- Data Corrections

Formatted data for the consistency as our data was open ended.

Manually updated values for consistency

e.g. YouTube->YouTube

- Run the Apriori algorithm using excel workbook to find first level support.
- Used minimum support as 5%
- Used Threshold for confidence as 75%

- Used **Excel** pivot for first level support.

Platform	Count of Platform
aakash	1
BnB	1
BYJU's	13
Coursera	18
DAMS	4
Emedicoz	2
Google Meet	1
Khan Academy	1
LinkedIn	3
Marrow	15
O'Reilly	1
Pluralsight	1
Prepladder	2
skillshare	1
teachable.	1
Udemy	22
Uworld	1
Webex	1
YouTube	61
(blank)	

Table 8 Market basket Analysis for Online Learning Platform Usage - First Level Support

- Filtered the values not matching minimum support.

Platform	Count	Support Percentage	is greater than Min Support
aakash	1	1%	FALSE
BnB	1	1%	FALSE
BYJU's	13	13%	TRUE
Coursera	18	17%	TRUE
DAMS	4	4%	FALSE
Emedicoz	2	2%	FALSE
Google Meet	1	1%	FALSE
Khan Academy	1	1%	FALSE
LinkedIn	3	3%	FALSE
Marrow	15	14%	TRUE
O'Reilly	1	1%	FALSE
Pluralsight	1	1%	FALSE
Prepladder	2	2%	FALSE
skillshare	1	1%	FALSE
teachable.	1	1%	FALSE
Udemy	22	21%	TRUE
Uworld	1	1%	FALSE
Webex	1	1%	FALSE
YouTube	61	59%	TRUE

Table 9 Market Basket Analysis for Online Learning Platform Usage – First Level Support Filters

- Crated pairs from values matching minimum support in first level support and checked support of pairs.

Pairs		support	Support Percentage	is Support Percentage Greater Than Min Support
BYJU's	Coursera	3	3%	FALSE
BYJU's	Marrow	2	2%	FALSE
BYJU's	Udemy	3	3%	FALSE
BYJU's	YouTube	9	9%	TRUE
Coursera	Marrow	1	1%	FALSE
Coursera	Udemy	8	8%	TRUE
Coursera	YouTube	16	15%	TRUE
Marrow	Udemy	2	2%	FALSE
Marrow	YouTube	10	10%	TRUE
Udemy	YouTube	19	18%	TRUE

Table 10 Market Basket Analysis for online learning platform usage - Support for pairs

- Crated triplets from values matching minimum support in pairs support and checked support of pairs.

Triplets			Support	Support Percentage	is Support > Min Support
BYJU's	Coursera	Marrow	0	0%	FALSE
BYJU's	Marrow	Udemy	0	0%	FALSE
BYJU's	Udemy	YouTube	2	2%	FALSE
Coursera	Marrow	Udemy	0	0%	FALSE
Coursera	Udemy	YouTube	7	7%	TRUE
Marrow	Udemy	YouTube	2	2%	FALSE

Table 11 Market Basket Analysis for online learning platform usage - Support for triplets

- Created association mining rules from the triplets.

A			B			Support Of A U B	Support of A	Confidence Support Of A U B/Support Of A	Confidence Greater than Min Confidence?
LHS			RHS						
Coursera	YouTube	->	Udemy			7	16	44%	FALSE
Coursera	Udemy	->	YouTube			7	8	88%	TRUE
YouTube	Udemy	->	Coursera			7	19	37%	FALSE
	Coursera	->	Udemy	YouTube		7	18	39%	FALSE
	Udemy	->	Coursera	YouTube		7	22	32%	FALSE
	YouTube	->	Coursera	Udemy		7	61	11%	FALSE

Table 12 Market Basket Analysis for online learning platform usage - Rules

6.2.5.3 Conclusion

- With above associate mining rules we could infer that data suggests people using Coursera, YouTube also uses Udemy.
- With more such rules on different features can give valuable insight into the usage patterns of digital platforms, e-services.
- For small number of categories, market basket analysis can be performed with traditional tools like Microsoft Excel

6.2.6 Random forest analysis to find important factors contributing to probability of money loss in fraud.

6.2.6.1 Purpose

To find important features that can be used for further analysis to predict money loss behaviour.

6.2.6.2 Data Upscaling - SMOTE

Using Random Forest machine learning technique, we want to find out the most important features for the response variable money lost during online transactions. Before that we need to check the raw data structure.

Raw data structure
Total Variables: 43
Response variable: money lost during online transaction (count of people who lost money during online transaction)
Total no of observations: 97.

As **FIGURE 23** reveals that data is imbalanced. The majority of responses fall in "Yes" category where Yes indicates a person has lost money during online transactions while "No" indicates not to lose money during online transactions. Using SMOTE technique, data is made balanced.

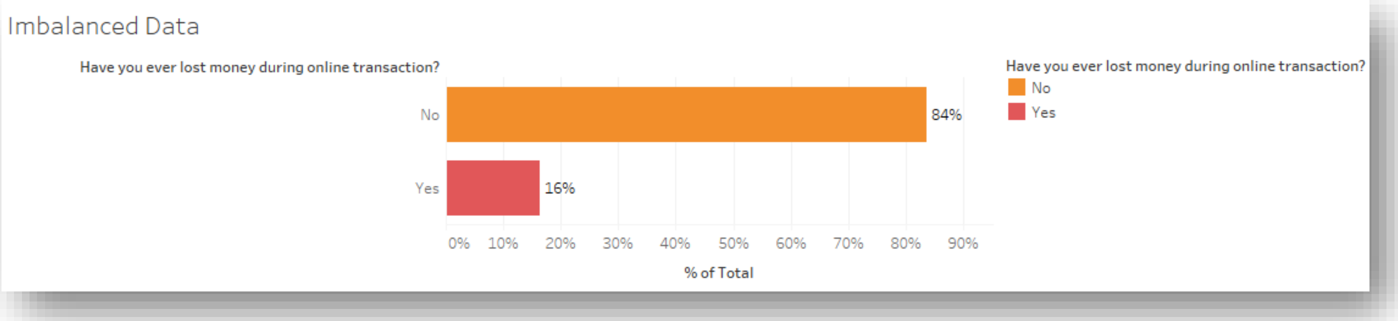


Figure 23 Imbalanced Data of Response Variable Money Lost During Online Transaction

6.2.6.3 Data Partitioning

Now, we divide the data set into two categories.

Particular	% of total data
Training Data	80%
Testing Data	20%

Table 13 Data Partitioning for Random Forest Model

6.2.6.4 RF Model building

To build an effective Random Forest model, we have used Param grid function to fine tune hyper parameters. Below are opted values of hyper parameters.

No of estimators= 100
Minimum sample split=2
Minimum sample leaf=2
Maximum features=6(square root of total features)
Maximum depth=30

6.2.6.5 RF model Performance

Using the above hyper parameters, the model is now applied to test data to check the various performance metrics. Below table represents the performance matrix.

Particular	Training Data	Testing Data
Accuracy	100%	91%
F1 Score	90	91
Out-of-bag error rate	NA	15.5%

Table 14 Random Forest Model Performance Metrics

6.2.6.6 Important variable features selection

RF model is not an interpretative model, but it gives prominent features related to response variables. Using this unique feature of the model we will extract important variables from the plot of the Variable importance plot. **FIGURE 24** is a variable importance plot where in there are total 41 features are plotted in descending order of importance.

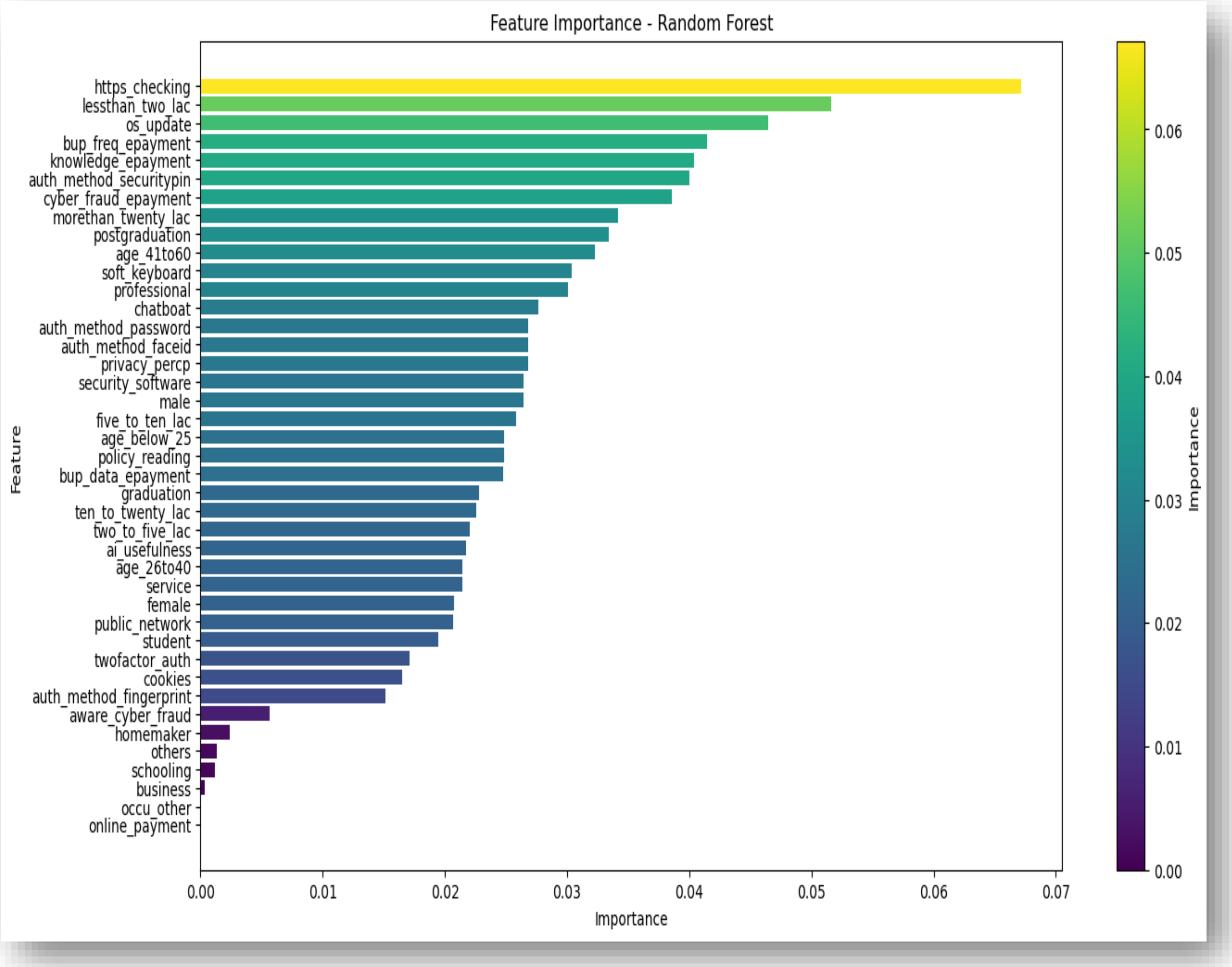
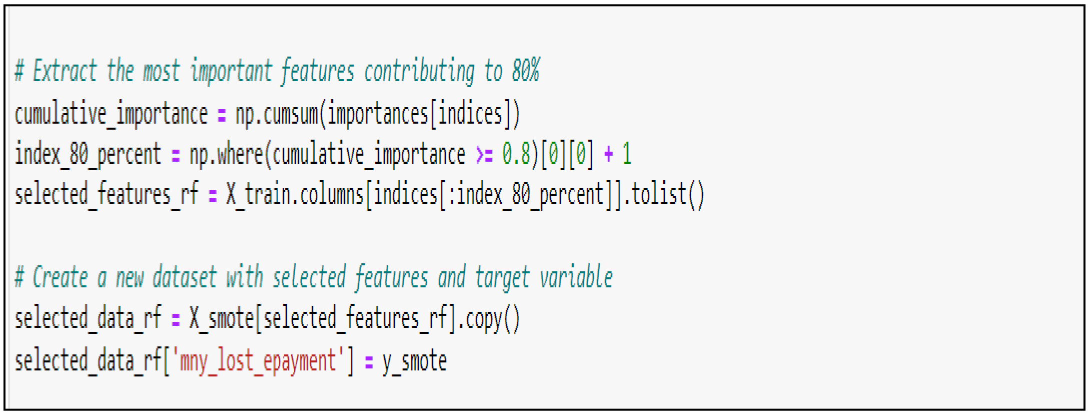


Figure 24 Variable Importance Plot for Response Money Lost

6.2.6.7 Extracting the most important features- Pareto Rule

As we can see in **Figure 24**, there are 41 features in plot, hence, to reduce the number of features to work further we now, use Pareto rule to extract the features whose total contribution in importance is 80%

After application we are left with a total 25 features from 41 features. The code snip shot of Pareto rule is shown in **Code Snippet 3**



```
# Extract the most important features contributing to 80%
cumulative_importance = np.cumsum(importances[indices])
index_80_percent = np.where(cumulative_importance >= 0.8)[0][0] + 1
selected_features_rf = X_train.columns[indices[:index_80_percent]].tolist()

# Create a new dataset with selected features and target variable
selected_data_rf = X_smote[selected_features_rf].copy()
selected_data_rf['mny_lost_epayment'] = y_smote
```

Code Snippet 3 Pareto rule application to get 80% contributed features from variable importance plot.

Total features (Variable importance plot)	Extracted features after Pareto rule application
41	26

Table 15 Extraction of Features Using Pareto Rule

https checking	Security method: Pin	Use of soft keyboard	Privacy perception	Age< 25 Years	Annual Income: 2-5 lakhs
Annual income < 2 lakhs	Face any cyber fraud	Working Professional	Usage of security software	Reading Policy?	-
Regular OS update	Annual income > 20 lakhs	Usage of chatbot	Male	Back up data	-
Back up frequency of data	Postgraduate	Security method: Password	Female	Graduate	-
Knowledge of e payment	Age 41 - 60	Security method: Face id	Annual income: 5-10 lakhs	Annual income: 10-20 lakhs	-

Table 16 Extracted Features for Further Analysis.

6.2.6.8 Extracted features - Raw data for next analysis.

Selected 26 features after the Pareto rule application, we have used these data set along with money lost as our response for the next Logistic regression model building.

6.2.7 Logistic regression to find probability of money loss in fraud.

6.2.7.1 Purpose

To find the variable that could help to reveal the probability of money loss in fraud and prevent cyber frauds. This model not only helps in classifying the categories of people who lost money but inherently it also reveals the behavioural pattern of those who are prone to lose money or vice versa.

We, now, use selected features data from the rf model and Pareto application to work further.

6.2.7.2 Data Partitioning

Particular	% of total data
Training Data	70%
Validating Data	15%
Testing Data	15%

Table 17 Data Splitting for Logistic Regression Model

6.2.7.3 Building Logistic Regression Model

Using Training data, we get the final model of logistic regression which is further tested on test data. On train data, we build an initial logistic regression model by applying a Backward elimination method. We will get the most significant variables which are the prime driving variables to decide/predict the money lost (fraud).

6.2.7.4 Backward Elimination method- Logistic Regression

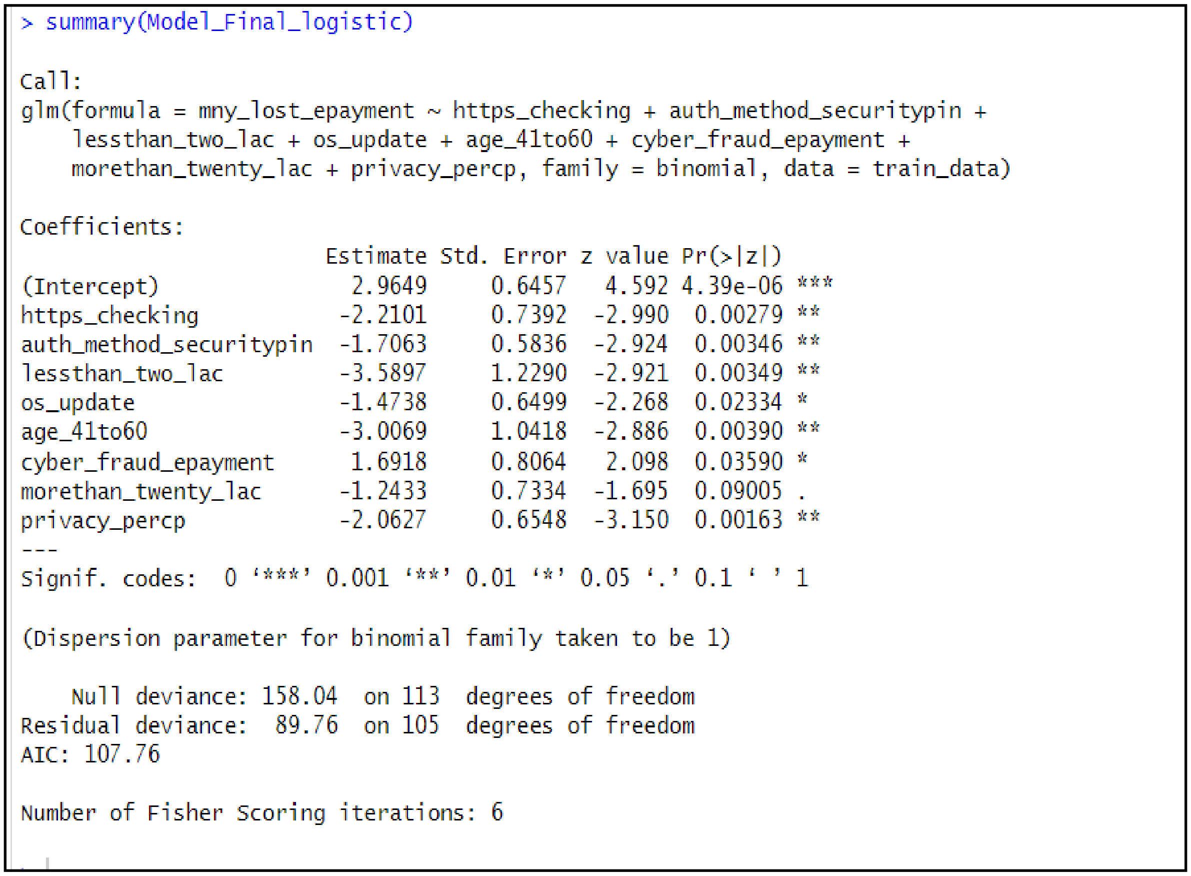
On training data, now, we use backward elimination method and threshold value for cut off are as mentioned below in table.

Particular	Cut-off value
VIF	< 3 (less than 3)
P value	< 0.1 (less than 0.1)

Table 18 Backward Elimination Cut-off Value Threshold

6.2.7.5 Most Significant variables- Logistic Regression

Figure below is a summary of LR model on validation data, which gives important insights on money fraud during online transaction.



```
> summary(Model_Final_logistic)

Call:
glm(formula = mny_lost_epayment ~ https_checking + auth_method_securitypin +
    lessthan_two_lac + os_update + age_41to60 + cyber_fraud_epayment +
    morethan_twenty_lac + privacy_percp, family = binomial, data = train_data)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.9649     0.6457   4.592 4.39e-06 ***
https_checking           -2.2101     0.7392  -2.990  0.00279 **
auth_method_securitypin  -1.7063     0.5836  -2.924  0.00346 **
lessthan_two_lac         -3.5897     1.2290  -2.921  0.00349 **
os_update                -1.4738     0.6499  -2.268  0.02334 *
age_41to60               -3.0069     1.0418  -2.886  0.00390 **
cyber_fraud_epayment      1.6918     0.8064   2.098  0.03590 *
morethan_twenty_lac      -1.2433     0.7334  -1.695  0.09005 .
privacy_percp            -2.0627     0.6548  -3.150  0.00163 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 158.04  on 113  degrees of freedom
Residual deviance:  89.76  on 105  degrees of freedom
AIC: 107.76

Number of Fisher Scoring iterations: 6
```

Code Snippet 4 LR model summary (test data)

Now, we have clear insights for that class of people who have not lost money during online transaction. Out of 8 significant variables, 6 variables are in negative association with response money lost. Now, let us discuss some insights we get from the above summary.

6.2.7.6 Important key notes from- Logistic Regression model.

Below table tells the probability of each class of people who are having very less likelihood of losing money during online transactions as are having negative association with response.

Class of Person	Probability (%) of losing Money during online transaction
Age bracket of 41 to 60	4.7%
Annual Income < 2 Lakhs	2.69%

Table 19 Least Likelihood of Losing Money During Online Transaction

(keeping other variable constant while calculating each class probability)

It is very interesting to note that those who are having an annual income of less than 2 lakh are very least prone to money fraud during online transactions.

If we talk about the age bracket, then surprisingly, people of age 41 to 60 during online transactions have very less probability to lose money.

Table 20 shares, the rest of the class of people's probability of not losing money.

Class of Person	Probability (%) of losing Money during online transaction
https checking (authentic web address)	9.8%
Usage of Security method: Pin	15.36%
Regular OS update	18.64%
Annual income > 20 lakhs	22.39%

Table 20 Lower Probability of Losing Money During Online Transaction

(keeping other variable constant while calculating each class probability)

There are two more variables which are positively associated with response money lost and probability of losing money are shown in below **Table 21**.

Class of Person	Probability (%) of losing Money during online transaction
Victim of cyber fraud (other than monetary fraud)	84.45%
Privacy perception (who believes sharing data to ChatGPT, and chat-boat hamper their privacy)	11.28%

Table 21 Higher Probability of Losing Money During Online Transaction

6.2.7.7 Validating Predictive Model

Now, we use validation data to get an optimal threshold to make the model more accurate and effective.

We have chosen F1 score as priority to get an optimal threshold value. We have also considered Area under curve value of ROC plot.

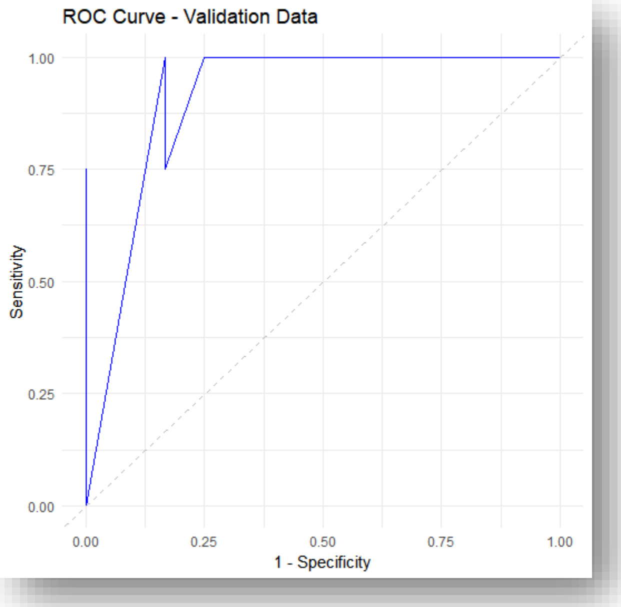


Figure 25 ROC Curve

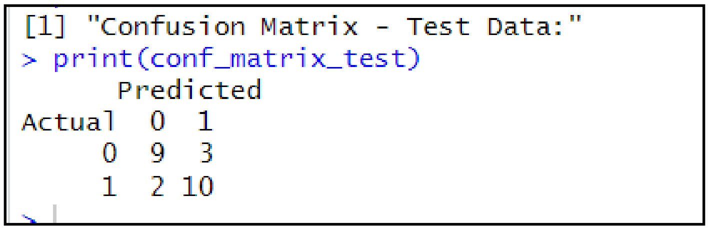
FIGURE 25 is the ROC curve and the value of AUC is 0.95, which is reasonably high to continue further with this model. While keeping F1 score high as priority, the threshold value obtained is mentioned below in table.

Threshold Value obtained	0.23
Area Under Curve	0.96

Table 22 Optimal Threshold Value

6.2.7.8 Testing Predictive ability of Model

It is time to evaluate the model on testing data, here we will mainly have testing performance to check the predictive ability of the model. Various performance measures are mentioned in the table below. For better understanding of the model performance, we need to compare it with train data performance metrics as well. Related to measures, confusion matrix is also shown in **CODE SNIPPET 5**.



```
[1] "Confusion Matrix - Test Data:"
> print(conf_matrix_test)
      Predicted
Actual  0  1
     0  9  3
     1  2 10
>
```

Code Snippet 5 Confusion matrix (test data)

Performance Measures	Training Data	Testing Data
Accuracy	87.7%	79.16%
Precision	84.12%	76.92%
Recall	92.98%	83.33%
F1 Score	88.3	80.0

Table 23 Performance Measures of Training and Testing Data

Our model exhibits effective predictive capabilities regarding individuals' susceptibility to financial loss during online transactions. Performance measures, including accuracy, precision, recall, and F1 score, have been assessed for both training and testing data, showcasing the model's reliability in capturing relevant patterns and behaviours. Refer to the table for specific metric values.

Note: All the analysis codes are checked into **(Github Project Source, 2023) for** easy reference, version control and collaboration amongst the team

7 Results

Cyber Security Concerns:

- Logistic Regression: Age group and student status influence concerns.
- Random Forest: Successfully predicts cyber fraud awareness.
- Apriori Rules: Higher concern in age 26 to 40, service occupation, and male gender.

Fraud Awareness:

- Apriori Rules: Lack of awareness linked to no information security knowledge, no data backup, no two-factor authentication, and female gender.

Fraud Loss Prediction:

- Random Forest: Non-loss characteristics include HTTPS checking, security pin authorization, income less than two lac, regular OS updates, age 41 to 60, income more than twenty lac, and privacy concern.
- Logistic Regression: Effective in predicting money loss.

8 Conclusion

These findings collectively emphasize the intricate nature of cybersecurity concerns, fraud awareness, and the prediction of financial losses in online transactions. The influence of demographic factors, information security practices, and individual characteristics contributes to a nuanced understanding of the challenges and opportunities in promoting secure digital transactions. Strategies to enhance cybersecurity should consider targeted measures based on age, occupation, gender, and awareness levels to effectively address concerns and mitigate the risk of financial loss.

9 Recommendations

9.1 Concentrated Ad Campaigns

- Leverage the insights gained from the project study to design and implement concentrated advertising campaigns. Tailor these campaigns to address specific concerns identified in different demographic segments, such as age groups, occupations, and gender.
- Highlight the security features and benefits of digital payment systems, emphasizing the measures taken to address the concerns raised by various user groups.

9.2 Develop and Market Cybersecurity Products

- Use the findings from the study to inform the development of cybersecurity products that specifically target the identified risk factors. Consider creating user-friendly interfaces that align with the preferences and concerns of different demographic groups.
- Market these cybersecurity products as essential tools for securing online transactions, with a focus on addressing the vulnerabilities highlighted in the study.

9.3 Establish Targeted Educational Programs

- Develop targeted educational programs based on the insights and inferences obtained from the project study. Tailor these programs to address the specific awareness gaps identified in different demographic categories, such as age, gender, and knowledge levels.
- Collaborate with educational institutions, businesses, and community organizations to implement these programs, ensuring widespread access to cybersecurity education.

9.4 Customized Communication Strategies:

- Implement customized communication strategies for different audience segments, considering the preferences and concerns revealed in the study. This could include designing communication materials that resonate with specific age groups, occupations, and genders.
- Utilize multiple channels, including social media, community events, and targeted online platforms, to disseminate information about cybersecurity measures and promote safe digital practices.

By implementing these recommendations, organizations and policymakers can proactively address cybersecurity concerns, enhance awareness, and contribute to the promotion of secure digital transactions. The customization of strategies based on demographic factors ensures a more targeted and effective approach in reaching and influencing diverse user groups.

9.5 Future Scope of this Study

This study lays the foundation for potential future research endeavours. Designing studies that target more granular audience groups could yield richer data for analysis, thereby enhancing the robustness of predictions and insights. Expanding the scope to encompass a more detailed examination of specific audience segments holds promise for advancing our understanding in this field.

10 References

CISA. (2022, December 18). *4 Things You Can Do To Keep Yourself Cyber Safe*. Retrieved from CISA.com: https://www.cisa.gov/news-events/news/4-things-you-can-do-keep-yourself-cyber-safe

Economic Survey 2023. (2023, January 31). Retrieved from moneycontrol.com: https://www.moneycontrol.com/news/business/economic-survey-2023-upi-accounted-for-52-of-indias-total-digital-transactions-in-fy22-9970741.html

Github Project Source. (2023, 11 11). *Github*. Retrieved from Github: https://github.com/rajivtrivedi/epababl05

Google Survey Form. (2023, 11 15). *Google Form*. Retrieved from Google Form: https://forms.gle/eidJ1KNXuTGur3NB9

IIM Library. (2023, 11 01). Retrieved from IIM Library: https://library.iima.ac.in/

Kegelmeyer, N. C. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intellegence Research*, 328-347.

Managing imbalance Data Set with SMOTE in Pythone. (2020, March 11). Retrieved from oralytics.com: https://oralytics.com/2019/07/01/managing-imbalanced-data-sets-with-smote-in-python/