**Capstone Project:  
Understanding Cyber Security and Safety Perceptions for Digital Payments and E-Services**



**Project Mentor**

**Prof. Kavitha Ranganathan, Prof. Arnab Laha**

**Presented By  
EPABABL05 Group – eShield**

**Chintan Joshi EPABABL0523029**

**Rajiv Trivedi EPABABL0523029**

**Hardik Raval EPABABL0523029**

Contents

[1. Executive Summary 6](#_Toc153108118)

[2. Introduction 6](#_Toc153108119)

[3. Objective 7](#_Toc153108120)

[3.1. Exploratory Analysis 7](#_Toc153108121)

[3.2. Cybersecurity Concerns 7](#_Toc153108122)

[3.3. Awareness of Fraud Possibilities 7](#_Toc153108123)

[3.4. Likelihood of Financial Loss in Fraud 7](#_Toc153108124)

[3.5. Usage of Online Learning Platforms 8](#_Toc153108125)

[4. Methodology 9](#_Toc153108126)

[5. Data Analysis 10](#_Toc153108127)

[5.1. Exploratory Data Analysis 10](#_Toc153108128)

[5.2. Predictive Analysis 13](#_Toc153108129)

[5.2.1. Logistic Regression Model to find how people are concerned. 13](#_Toc153108130)

[5.2.1.1. Selection of Response Variable 13](#_Toc153108131)

[5.2.1.2. Explanatory Variable Selection: 13](#_Toc153108132)

[5.2.1.3. Data Preprocessing: 14](#_Toc153108133)

[5.2.1.4. Data Scaling: 14](#_Toc153108134)

[5.2.1.5. One-Hot Encoding: 14](#_Toc153108135)

[5.2.1.6. Collinearity Assessment 15](#_Toc153108136)

[5.2.1.7. Variable Selection 15](#_Toc153108137)

[5.2.1.8. Model Building 15](#_Toc153108138)

[5.2.1.9. Model Evaluation 16](#_Toc153108139)

[5.2.1.10. Conclusion on Model: 16](#_Toc153108140)

[5.2.2. Random Forest Analysis to find fraud awareness. 17](#_Toc153108141)

[5.2.2.1. Purpose 17](#_Toc153108142)

[5.2.2.2. Variable Selection 17](#_Toc153108143)

[5.2.2.3. Upscale data 17](#_Toc153108144)

[5.2.2.4. Training testing division 17](#_Toc153108145)

[5.2.2.5. Model 17](#_Toc153108146)

[5.2.2.6. Validation of model 18](#_Toc153108147)

[5.2.2.7. Important Variables 19](#_Toc153108148)

[5.2.2.8. Conclusion 19](#_Toc153108149)

[5.2.3. Market Basket Analysis to find the rules for the people concerned. 20](#_Toc153108150)

[5.2.3.1. Purpose 20](#_Toc153108151)

[5.2.3.2. Method 20](#_Toc153108152)

[5.2.3.3. Derived Rules 20](#_Toc153108153)

[5.2.3.4. Conclusion 21](#_Toc153108154)

[5.2.4. Market Basket Analysis to find factors behind lack of fraud awareness. 22](#_Toc153108155)

[5.2.4.1. Purpose 22](#_Toc153108156)

[5.2.4.2. Method for finding features for lack of fraud awareness. 22](#_Toc153108157)

[5.2.4.3. Derived Rules 23](#_Toc153108158)

[5.2.4.4. Conclusion 24](#_Toc153108159)

[5.2.5. Market Basket Analysis to find online learning usage patterns. 25](#_Toc153108160)

[5.2.5.1. Purpose 25](#_Toc153108161)

[5.2.5.2. Method for finding patterns in online learning. 25](#_Toc153108162)

[5.2.5.3. Conclusion 29](#_Toc153108163)

[5.2.6. Random forest analysis to find important factors contributing to probability of money loss in fraud. 30](#_Toc153108164)

[5.2.6.1. Purpose 30](#_Toc153108165)

[5.2.7. Logistic regression to find probability of money loss in fraud. 30](#_Toc153108166)

[5.2.7.1. Purpose 30](#_Toc153108167)

[6. Results 30](#_Toc153108168)

[7. Conclusion 31](#_Toc153108169)

[8. Recommendations 31](#_Toc153108170)

[8.1. Concentrated Ad Campaigns 31](#_Toc153108171)

[8.2. Develop and Market Cybersecurity Products 31](#_Toc153108172)

[8.3. Establish Targeted Educational Programs 31](#_Toc153108173)

[8.4. Customized Communication Strategies: 32](#_Toc153108174)

[9. References 32](#_Toc153108175)

[Figure 1 Demographic information of the data 10](#_Toc153108176)

[Figure 2 Usage of AI tools 11](#_Toc153108177)

[Figure 3 usage of online learning platforms 12](#_Toc153108178)

[Figure 4 important variables for the random forest analysis on fraud awareness 19](#_Toc153108179)

[Figure 5 rules for people very much concerned about security on internet 21](#_Toc153108180)

[Figure 6 Rules for Lack of Fraud awareness. 24](#_Toc153108181)

[Table 1 Variable Selection for Logistic Regression Model to find how people are concerned. 14](#_Toc153108182)

[Table 2 income group encoding in logistic model 14](#_Toc153108183)

[Table 3 age group encoding in logistic model 14](#_Toc153108184)

[Table 4 Accuracy matrix for the cyber security concern analysis 16](#_Toc153108185)

[Table 5 variable selection for the random forest method for fraud awareness 17](#_Toc153108186)

[Table 6 Market Basket Rules for people very much concerned about security on internet 20](#_Toc153108187)

[Table 8 Market Basket Rules for lack of fraud awareness 23](#_Toc153108188)

[Table 9 Market basket analysis for online learning platform usage - first level support 26](#_Toc153108189)

[Table 10 Market Basket Analysis for online learning platform usage – First level support Filters 27](#_Toc153108190)

[Table 11 Market Basket Analysis for online learning platform usage - Support for pairs 28](#_Toc153108191)

[Table 12 Market Basket Analysis for online learning platform usage - Support for triplets 28](#_Toc153108192)

[Table 13 Market Basket Analysis for online learning platform usage - Rules 29](#_Toc153108193)

[Code Snippet 1 Logistic model 16](#_Toc153108194)

[Code Snippet 2 random forest model for predicting fraud awareness. 18](#_Toc153108195)

# 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 analyze 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.

# Introduction

In today's digital world, using things like online payments and services on the internet has become really 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.

# 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, shedding light on various dimensions of consumer behavior and perceptions.

Key Questions to be Addressed

## 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?

## 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?

## 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?

## 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?

## 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.

# Methodology

# Data Analysis

## Exploratory Data Analysis

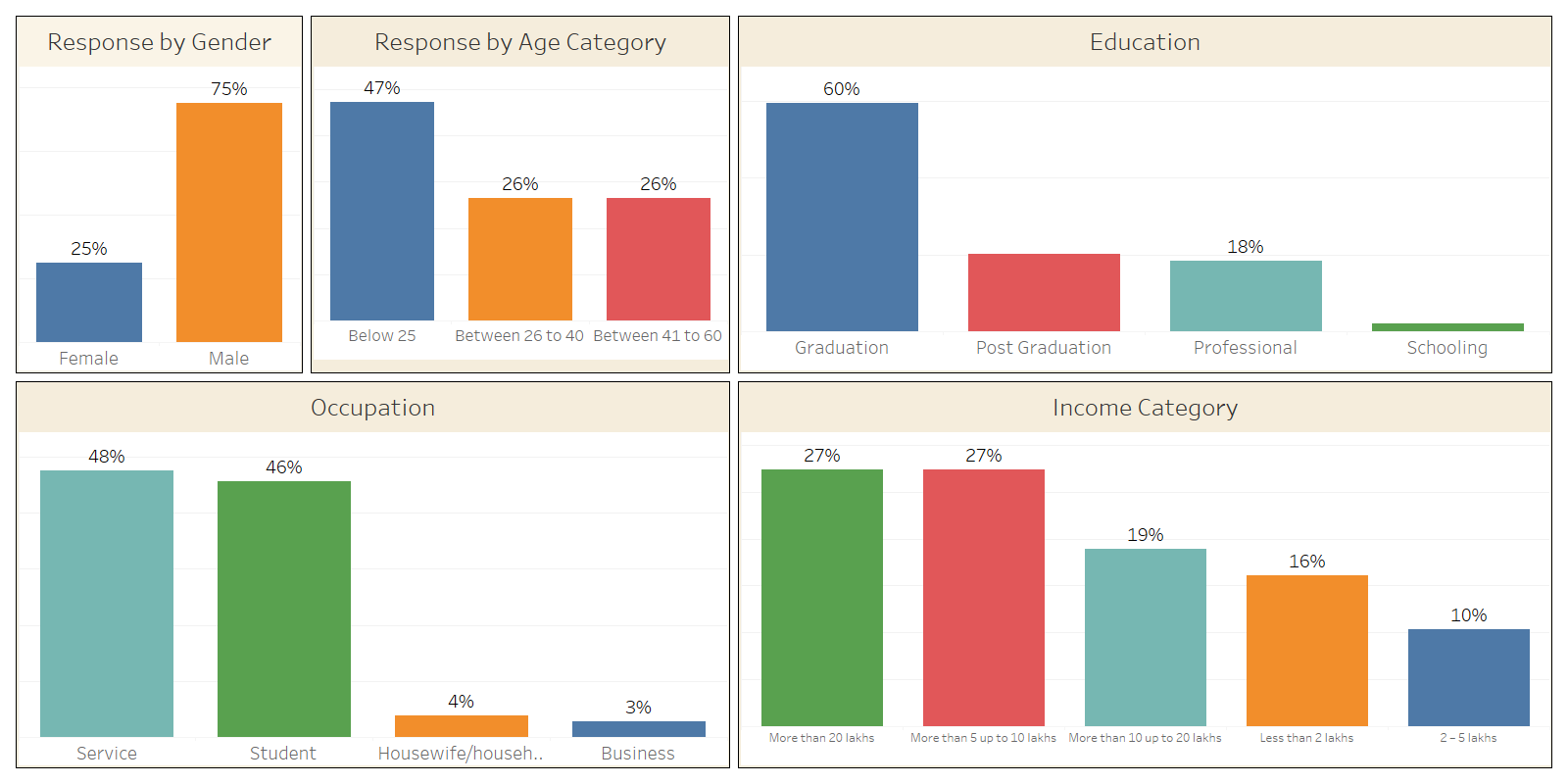


Figure 1 Demographic information of the data

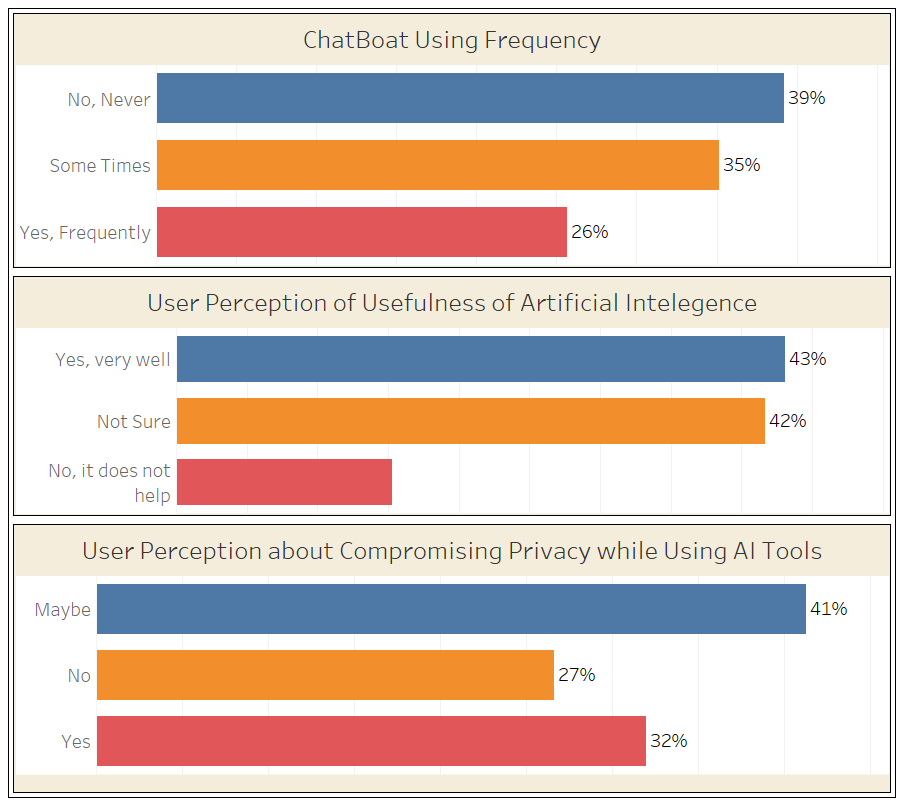


Figure 2 Usage of AI tools

A screenshot of a computer

Description automatically generated

Figure 3 usage of online learning platforms

## Predictive Analysis

## Logistic Regression Model to find how people are concerned.

## 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, a tool suited for predicting outcomes with two categories.
* 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 behaviors.
* 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.

## 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.

## 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

## 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

## One-Hot Encoding:

* Introduced dummy columns representing different categories to capture categorical information

e.g. is\_male, is\_female,is\_studen

## 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.

## Variable Selection

* Selected is\_student and age\_group as these were the statistical significant data with the sample data

## Model Building

Built logistic model using

|  |
| --- |
| 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

## 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

## 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.

## Random Forest Analysis to find fraud awareness.

## Purpose

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

## 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

## Upscale data

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

## Training testing division

* Splitted data into train and testing sets

## 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.

## 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

## Important Variables

A graph with numbers and text

Description automatically generated

Figure 4 important variables for the random forest analysis on fraud awareness

## 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.

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

## 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.

## Method

Upscaled the data using caret library in R

Used minimum support as 20%

Used Threshold for confidence as 70%

## 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

A white background with black text and red dots

Description automatically generated

Figure 5 rules for people very much concerned about security on internet

## Conclusion

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

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

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

## 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.

## 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%

## 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 8 Market Basket Rules for lack of fraud awareness

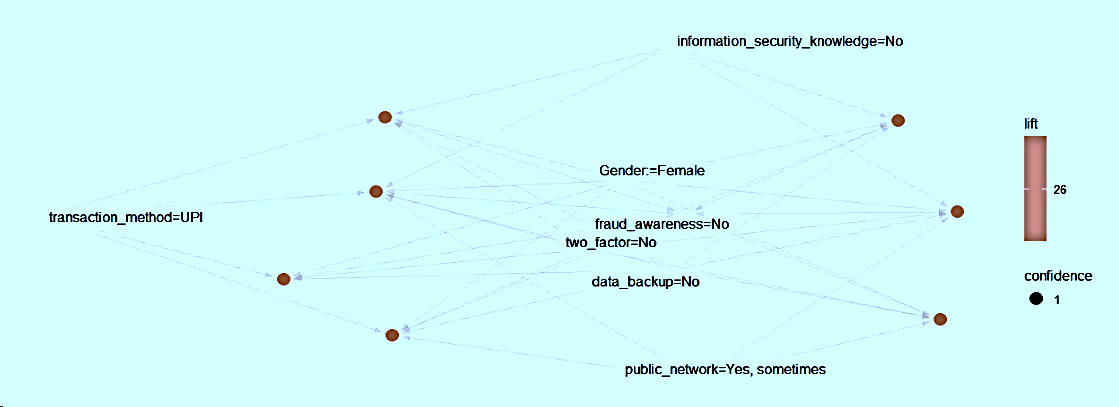


Figure 6 Rules for Lack of Fraud awareness.

## 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

## Market Basket Analysis to find online learning usage patterns.

## Purpose

To find the patten of usage of online learning platforms.

## 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 9 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 10 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 11 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 12 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 13 Market Basket Analysis for online learning platform usage - Rules

## 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

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

## Purpose

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

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

## Purpose

To find the variable that could impact probability of money loss in fraud and prevent cyber frauds.

# 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.

# 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.

# Recommendations

## 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.

## 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.

## 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.

## Customized Communication Strategies:

* Implement customized communication strategies for different audience segments, taking into account 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.

# References

1. <https://forms.gle/eidJ1KNXuTGur3NB9>
2. <https://library.iima.ac.in/>
3. https://github.com/rajivtrivedi/epababl05
4. Galit Shmueli, Peter C. Bruce, Peter Gedeck, Nitin R. Patel - Data Mining for Business Analytics\_ Concepts, Techniques and Applications in Python-Wiley (2019)