CASHLESS INDIA

PGDASS PROJECT
2016-17

TABLE OF CONTENTS

UNIVERSITY OF MUMBAI
DEPARTMENT OF STATISTICS
VIDYANAGARI MUMBAI – 400098

CERTIFICATE

This is to certify that Mr./Ms._____ of PGDASS has successfully completed the project entitled "Cashless India" during the academic year 2016-2017.

The Team Comprises of: -

Abhijit Ghosh
Amruta Panhalkar
Austin D'sa
Heet Shah
Kiran Patil
Pallavi Pawar
Rushali Salvi
Siddhi Newalkar
Vipul Mishra

This work is to the best of our knowledge and belief is original.

Dr. Santosh Gite
(Head of Department & Project Guide)

ACKNOWLEDGEMENT

From the conception of topic to the final presentation, we are extremely thankful to our guide **Dr. Santosh Gite** (*Head of Department of Statistics at University of Mumbai*) for his immense support and guidance throughout the process of making this entire project. It was only with the help of his suggestions & corrections, at every step that we were able to complete the project on time & achieve our objectives to our satisfaction.

In addition, we would like to express our gratitude to all the professors and non-teaching staff of the Department of Statistics for their co-operation as well as providing us with requisite amenities over the duration of the project work.

Last but not the least, we would like to give our special thanks to our respondents who diligently and honestly filled our survey purely for academic purposes and taking the time out of their schedule to answer our questions genuinely.

INTRODUCTION

There is tremendous interest worldwide among policy makers, academicians and commercial enterprises to explore the possibility of moving towards a cashless economy. There are several reasons not to like cash, but ours remains a cash-based world.

The burden of cash usage on national economies is substantial. Heavy cash usage may also be an indicator of other economic problems.

Branch Banking is a traditional way of banking where customer visits a bank branch and operates his account. However, cash still continues to remain the predominant form of transaction. The initial hesitation has subsided now and the people are beginning to realize the safety and convenience offered by digital payment systems.

However, the benefits of this move are trickling in with more and more people switching to digital modes of receiving and making payment. Digital transactions are traceable, therefore easily taxable, leaving no room for the circulation of black money.

The whole country is undergoing the process of modernization in money transaction. “Faceless, paperless, cashless” is one of the professed role of Digital India.

As a part of promoting cashless transaction and converting India into less-cash society, various modes of cashless payments are available.
These modes are:

- Banking Cards
- Cheque/DD
- Internet Banking
- Mobile Banking
- ECS
- E-Wallets

Banking Cards

- A **bank card** is typically a plastic card issued by a bank to its clients that can perform one or more of number of services that relate to giving the client access to funds, either from the client's own bank account, or through a credit account. It can also be a smart card.
- Physically, a bank card will usually have the client's name, the issuer's name, and a unique card number printed on it. It will have a magnetic strip on the back enabling various machines to read and access information.
- Depending on the issuing bank and the preferences of the client, this may allow the card to be used as an ATM card, enabling transactions at automatic teller machines; or as debit card, linked to the client's bank account and able to be used for making purchases at the point of sale; or as a credit card attached to a revolving credit line supplied by the bank.
- **Advantages** of a Banking Card: -
  - Easy to obtain
  - Convenience
  - Safety
  - Readily accepted

Cheque/DD

- A **cheque** is a document that orders a bank to pay a specific amount of money from a person's account to the person in whose name the cheque has been issued. The person writing the cheque, the *drawer*, has a transaction banking account (often called a current, cheque, chequing or checking account) where their money is held. The drawer writes the various details including the monetary amount, date, and a payee on the cheque, and signs it, ordering their bank, known as the *drawee*, to pay that person or company the amount of money stated.
- Cheques are a type of bill of exchange and were developed to make payments without the need to carry large amounts of money. Paper money evolved from promissory notes, another form of negotiable instrument like cheques in that they were originally a written order to pay the given amount to whomever had it in their possession (the "bearer").

- A **demand draft** is a negotiable instrument like a bill of exchange. A bank issues a demand draft to a client (drawer), directing another bank (drawee) or one of its own branches to pay a certain sum to the specified party (payee).
- A demand draft can also be compared to a cheque. However, demand drafts are difficult to countermand. Demand drafts can only be made payable to a specified party, also known as pay to order. But, cheques can also be made payable to the bearer. Demand drafts are orders of payment by a bank to another bank, whereas cheques are orders of payment from an account holder to the bank.

Internet Banking

- **Online banking**, also known as **internet banking**, **e-banking** or **virtual banking**, is an electronic payment system that enables customers of a bank or other financial institution to conduct a range of financial transactions through the financial institution's website. The online banking system will typically connect to or be part of the core banking system operated by a bank and contrasts with branch banking which was the traditional way customers accessed banking services.
- To access a financial institution's online banking facility, a customer with internet access would need to register with the institution for the service, and set up a password and other credentials or customer verification. The credentials for online banking is normally not the same as for telephone or mobile banking. Financial institutions now routinely allocate customers numbers, whether customers have indicated an intention to access their online banking facility. Customer numbers are normally not the same as account numbers, because several customer accounts can be linked to the one customer number. Technically, the customer number can be linked to any account with the financial institution that the customer controls, though the financial institution may limit the range of accounts that may be accessed to, say, cheque, savings, loan, credit card and similar accounts.
- The customer visits the financial institution's secure website, and enters the online banking facility using the customer number and credentials previously set up. The types of financial transactions which a customer may transact through online banking are determined by the financial institution, but usually includes obtaining account balances, a list of the recent transactions, electronic bill payments and funds transfers between a customer's or another's accounts. Most banks also enable a customer to

download copies of bank statements, which can be printed at the customer's premises (some banks charge a fee for mailing hard copies of bank statements). Some banks also enable customers to download transactions directly into the customer's accounting software. The facility may also enable the customer to order a cheque book, statements, report loss of credit cards, stop payment on a cheque, advise change of address and other routine actions.

- Today, many banks are internet-only institutions. These "virtual banks" have lower overhead costs than their brick-and-mortar counterparts. In the United States, many online banks are insured by the Federal Deposit Insurance Corporation (FDIC) and can offer the same level of protection for the customers' funds as traditional banks.

Mobile Banking

- **Mobile banking** is a service provided by a bank or other financial institution that allows its customers to conduct financial transactions remotely using a mobile device such as a mobile phone or tablet. It uses software, usually called an app, provided by the financial institution for the purpose. Mobile banking is usually available on a 24-hour basis. Some financial institutions have restrictions on which accounts may be accessed through mobile banking, as well as a limit on the amount that can be transacted.
- Transactions through mobile banking may include obtaining account balances and lists of latest transactions, electronic bill payments, and funds transfers between a customer's or another's accounts. Some apps also enable copies of statements to be downloaded and sometimes printed at the customer's premises; and some banks charge a fee for mailing hardcopies of bank statements.

ECS

- **ECS** is an electronic mode of funds transfer from one bank account to another. It can be used by institutions for making payments such as distribution of dividend interest, salary, pension, among others.
- It can also be used to pay bills and other charges such as telephone, electricity, water or for making equated monthly installments payments on loans as well as SIP investments. ECS can be used for both credit and debit purposes.

E-wallet

- E-wallet is a type of electronic card which is used for transactions made online through a computer or a smartphone. Its utility is same as a credit or debit card. An E-wallet needs to be linked with the individual's bank account to make payments.
- It is a type of pre-paid account in which a user can store his/her money for any future online transaction. An E-wallet is protected with a password. With the help of an E-wallet, one can make payments for groceries, online purchases, and flight tickets, among others.
- Some of the prominent e-wallets are PayTM, Freecharge, BHIM, etc.

Pre-requisites for Going Cashless:

- Access to financial services
  - Measures of the availability and affordability of financial services and whether people use bank accounts and electronic payment products.
- Macro-economic and cultural factors
  - Measures the factors impacting preference for cash, such as ease of doing business and size of informal economy.
- Merchant scale and competition
  - Measures indicating the potential for uptake of new payment solutions by large scale merchants. Measures the intensity of local competition.
- Technology and infrastructure
  - Measures of access to and uptake of new technology as well as innovation. Also measures the quality of infrastructure.

CASH – Boon or Bane?

For businesses, paper money has to be managed, it must be stored, guarded, and accounted for. It can be difficult to transport and is inherently insecure. The cash-dependent small businesses cannot afford sophisticated security and cash transportation services.

Cash – by which I mean paper currency and coins — has many benefits. It's safe from hackers. It doesn't require any special hardware or software. There is no fee charged to retailers who use it and no exorbitant interest rates lying in wait for consumers. It's accepted almost everywhere and it offers anonymity.

While it has been steadily displaced by a variety of competitors, such as credit and debit cards, mobile payments, and crypto currencies, there are many good reasons paper money has stuck around. There's an assumption that cash is best when money is tight – best for the poor, and best for small businesses running on tight margins.

OBJECTIVES

As the world rapidly advances towards a cashless-based economy, there are many challenges and benefits at the same time during this process. In order to make this transition with minimum bottlenecks, we need to identify people that might need assistance in moving towards this new era.

Our primary aim of this project is to distinct individuals that support cashless payment systems from individuals that don't. In addition, we also would like to know the factors that play a significant role in cashless decision making but are usually hidden from plain sight.

Therefore, our objectives are as follows: -

- To identify and analyze the socio-demographic factors that affect the people's decision on whether to go cashless.
- To identify and analyze other latent (qualitative) factors that influence the people's preference to choose cash over cashless.
- To find the deciding factors leading up-to a person's preference for using cashless in daily transactions.
- To find out the most preferred cashless mode of payment/ online wallet and the factors behind its popularity.
- To find out the major concerns with current cashless payment systems among the people who use cashless and those who don't.

RESEARCH DESIGN

To meet aim and objectives of the study, it is important that the researcher select the most appropriate research design. The research design identifies the procedures by which the study population will be selected. The research design refers to the overall strategy that you choose to integrate the different components of the study in a coherent and logical way, thereby, ensuring you will effectively address the research problem; it constitutes the blueprint for the collection, measurement, and analysis of data.

1. Research Methodology:

- **Defining the objectives of the project**
  - Defining the problem clearly. (Ex. – Our primary aim was to identify people who use cashless and significant factors that affect cashless payment systems)
  - The scope of the project was chalked out to determine the project plan.
  - The plan of action was developed which included the start and the end dates of various steps of the project.
- **Questionnaire Design**
  - As per our objectives, a questionnaire was prepared. It was not the final questionnaire as changes would be made (if necessary) after the pilot survey.
- **Pilot Survey**
  - We collected primary data by means of an online survey.
  - Pilot survey was done on the questionnaire with a sample size of 50 on the basis of which changes were made in the questionnaire. The questionnaire contains 31 questions which gives factors affecting final conclusion/decision.
  - The pilot survey was conducted for the purposes of framing our objectives better and to get an idea about the data we can expect from the full-fledged survey.
- **Preparing final questionnaire**
  - A final questionnaire was prepared using all the input from various sources.

- **Collection of data**
  - The data was collected using Google Forms.
  - The response of people was recorded by open-ended questions; multiple choice questions, grade scale.
  - The filled up information was later obtained to give the required interpretation and findings of most preferred cashless mode and factors behind its popularity.
- **Performing data analysis**
  - Different statistical tests (Ex. – Binary logistic regression, Factor Analysis, etc.) were conducted to test the different hypothesis using various software.
- **Arriving at the conclusions**
  - From the results obtained after performing the tests, conclusions were given accordingly.

2. Research approach:

There are two main approaches to research:

Qualitative approach is concerned with assessment of attitudes, opinions and behaviour. We employed qualitative research to fulfil our objectives 2 and 5.

Quantitative approach emphasizes objective measurements and the statistical, mathematical or numerical analysis of data. We employed quantitative research to fulfil our objectives 1, 3 and 4.

3. Types of research design:

- **Exploratory research:**

  - Exploratory research is the type of research conducted for a problem that has not been clearly defined. Exploratory research helps to determine the best research design, data collection method and selection of subjects. It also helps in framing of our objectives.
  - The results of exploratory research are not usually useful for decision-making by themselves, but they can provide significant insight into a given situation.

- We went forward with the exploratory research by conducting a pilot survey to check whether there were any issues in the questionnaire, to reframe our objectives better and to get an idea of the kind of responses we would receive especially, in the open-ended questions.

- **Descriptive research:**

  - Descriptive research refers to research that provides an accurate description of characteristics of a particular individual, situation, or group. Descriptive research is also known as statistical research.
  - In short, descriptive research deals with something that can be counted and studied, which has an impact in the lives of the people it deals with. After our pilot survey our actual main research was descriptive research based on the objectives we framed after the exploratory research.

4. Target population:

The respondents will be the users who at least have knowledge about cashless modes of payment. Our respondents were primarily belonging to an urban setting.

- **Gender:** Male and Female
- **Age:** 15 – 70

5. Sample size:

A large sample size helps to negate the biasness if any that arise due to sampling. Our aim was to collect as many responses as we could within a month's time from as many sources as we could.

Size required: At least 1000 responses.

6. Statistical software used:

- Microsoft Excel
- SPSS
- SAS
- R

QUESTIONNAIRE DESIGN

1. Study protocol

This involves getting detailed knowledge of our topic cashless payment systems, decide on objectives, formulate a hypothesis, and define the main information needed to test the hypothesis.

2. Draw a list of the information needed

From the plan of analysis, we have drawn a list of the information that needs to be collected from participants. In this step, we have determined the type and format of the variables needed.

3. Design different parts of the questionnaire

In this step, we started designing different parts of the questionnaire considering the required information and objectives.

4. Framing questions

We have framed questions knowing the education, occupation, ethnicity/background, language, knowledge and special sensitivity level of our study population. We have also kept in mind that the questionnaire needs to be adapted to our study population.

5. Order of the questions asked

We started from easy/general, moderate to difficult questions. We had to make sure the most sensitive questions are correctly placed. They should be rather placed in the middle or towards the end of the questionnaire. And also we needed to be careful that we do not to put the most important item last, since some people might not complete the interview/ survey.

E.g. If respondent does not use cashless systems then it is very important that they skip all the related questions based on use of cashless method and straight away answer the reason for not using cashless system and end the questionnaire.

6. Complete the questionnaire

We added instructions and definitions of key words for participants ensuring a smooth flow throughout the survey. We inserted jumps between questions if some

questions are not meant to be answered by respondents. E.g. Use of E-Wallets (Paytm, Freecharge, Mobile banking apps, etc.)

7. Verification of the content and style of the questions

We verified each question's answer for each objective. We have deleted questions that are not directly related to objectives and made sure that each question is clear, unambiguous, simple and short. We checked the logical order and flow of the questions and made sure that the questionnaire is easy to read and has a clear layout.

8. Conduct a pilot study

We started to conduct a pilot study among the intended population before starting the live survey.

9. Refine your questionnaire

Depending on the results of the pilot study, we have made some amendment to the questionnaire before the main survey starts.

QUESTIONNAIRE

1. What is your age? *

   Enter the appropriate number

   _____

2. What is your gender? *

   Mark only one oval

   - Male
   - Female

3. What is the highest level of education you have received? *

   Mark only one oval

   - Not studied
   - Up to 12th Standard
   - Graduate Degree / Diploma
   - Masters / Graduate Degree
   - Research Studies (M.Phil / PhD)
   - Prefer not to answer
   - Other:

4. What is your occupation? *

   Mark only one oval

   - Private Sector
   - Government Service
   - Business
   - Student
   - Unemployed
   - Retired
   - Homemaker
   - Freelance
   - Other:

5. What is your religion? *

Mark only one oval

- Hinduism
- Islam
- Christianity
- Sikhism
- Judaism
- Buddhism
- Atheist
- Agnostic
- Other:

6. What option best describes your house/dwelling? *

Mark only one oval

- Residential Apartment
- Bungalow/ Villa
- Chawl
- Slums

7. Please select the range below that best describes your total annual household income. *

Mark only one oval

- Less than ₹50,000
- ₹50,000 - ₹1,00,000
- ₹1,00,000 - ₹2,50,000
- ₹2,50,000 - ₹5,00,000
- ₹5,00,000 - ₹10,00,000
- Above ₹10,00,000

8. Which option best describes your family? *

Mark only one oval

- o Joint Family
- o Nuclear Family

9. State the number of dependents in your family? *

Enter the appropriate number

_____

10. Which of the following statements best describes your current marital status? *

Mark only one oval

- o Single (not living with a significant other or partner)
- o Married
- o Not married but living with my partner/significant other
- o Widowed
- o Divorced

11. How many members are there in your family including you? *

Enter the appropriate number

_____

12. Do you own a smartphone? *

Mark only one oval

- o Yes
- o No

13. How many BANK ACCOUNTS do you have? *

Mark only one oval

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7

14. What is the type of bank you have your account in? *

Check all that apply

- ▪ Public Sector bank
- ▪ Private bank
- ▪ Co-Operative bank
- ▪ Other:

15. Do you use cashless payment methods? *

Mark only one oval

- ○ Yes
- ○ No

16. What are the mode of payment you use? *

Check all that apply

- • By cash
- • Cheque/DD
- • ECS
- • Credit/Debit card
- • e-wallets (E.g. PayTM, Freecharge, Mobile banking apps, etc.)
- • Net Banking
- • Other:

17. How likely are you to use cashless payment methods in the following situations? *

Mark only one oval per row

	Very Likely (100%)	Likely (75%)	Neither likely nor unlikely (50%)	Unlikely (25%)	Very unlikely (0%)
Food in a restaurant					
Drinks at a bar/ coffee shop, etc.					
Physical goods in a retail store					
Trades work in the home					
Tickets for events					
Donation to charity					
Paying a friend					
Taxi/ Cab					
Motor Fuel					
Public Transport					
Goods over the phone					
Goods on web					
Goods in the market					
Utility bills					
Govt. Service (TV License, Passport, etc.)					

18. How comfortable are you with using mobile/computer for payments? *

Mark only one oval

	1	2	3	4	5	
Not at all comfortable						Very Comfortable

19. Are you using cashless payment methods/system because of demonetization? *

Mark only one oval

- ○ Yes
- ○ No

20. How did you come to know about cashless payment methods/system? *

Check all that apply

- ▪ TV/ Radio
- ▪ Reference (Family, friends, colleagues, etc.)
- ▪ Social Media
- ▪ Newspaper/ Magazines
- ▪ Other:

21. Give rating to the following factors for preferring cashless methods/ system? *

Mark only one oval per row

		1 – Strongly Disagree	2	3	4	5 – Strongly agree
1.	User – friendly					
2.	Time saving					
3.	Convenience					
4.	Trend					
5.	Secure					
6.	Discount / cashback/ offers					

22. Will you refer/ recommend someone to use cashless payment system? *

Mark only one oval

- Yes
- No
- Maybe

23. Do you do/ prefer online shopping? *

Mark only one oval

- Yes
- No

24. How frequently do use mobile banking/ online payment? *

Mark only one oval

- ○ Once a week
- ○ More than once a week
- ○ Less than 4 transactions in a month
- ○ Not using

25. Average amount spent in a month using e-payments (Mobile wallet)? *

Mark only one oval

- ○ Less than 1,000
- ○ Between 1,000 and 5,000
- ○ Between 5,000 and 10,000
- ○ More than 10,000

26. Do you think that CASHLESS transactions are more convenient than cash transactions for daily transactions? *

Mark only one oval

- ○ Yes
- ○ No

27. Which are the mobile wallets you use? *

Check all that apply

- ▪ Paytm
- ▪ Freecharge
- ▪ MobiKwik
- ▪ BHIM/ UPI
- ▪ Ola Money
- ▪ Airtel Money
- ▪ Banking Apps
- ▪ Other:

28. Which of the following systems do you prefer? *

Mark only one oval

- o Banking Apps
- o Mobile wallet apps

29. What are your reasons for preferring one over the other system above? *

_____

_____

_____

30. What are your major concerns with cashless payment systems? *

_____

_____

_____

31. Any reason as why you do not prefer cashless payment systems? *

_____

_____

_____

DATA COLLECTION

Our objectives demanded primary data and hence a questionnaire was prepared with the help of Google Forms. The preliminary step of data collection was to carry out a pilot survey of suitable size. This helped us to formulate our objectives and better understand the structure of data.

The next step was to send the questionnaire to as many people as possible. Because when it comes to data, the higher the number of responses the better. The data was collected through various mediums like e-mails, offline surveys, and lastly, with the help of social networking sites like WhatsApp and Facebook.

We decided to close the survey a month after it went live. We finally closed it once we received 1150 responses approximately a month after we started it. We then entered the entire data in Microsoft Excel. Among the 1150 responses about 36 responses were discarded due to false and incomplete information, fake/ spam responses. The final data we were left with (1114 responses) was then coded for further analysis.

DATA CLEANING

Preparing the data for analysis is very essential process as data needs to be cleaned before bringing it into an actionable form where all the final analysis can be conducted on the sample. Typically, in a survey data there are ample of issues which can occur if the data is not cleaned properly as responses can be duplicate, can contain some bias or the respondent might not have taken the survey seriously. The aim of data cleaning is to remove these overt biases and invalid responses to get the good quality of data for performing accurate analysis.

Steps followed during data cleaning:

1. **Importing the data into Excel**

   Excel is a very efficient tool for data cleaning as it has many functions in it, which makes the process less taxing. Our survey was designed and conducted on Google forms. After reaching the deadline of the survey, we closed the survey and downloaded the data in an Excel file and went forward with data cleaning. We had collected 1150 responses by the deadline.

2. **Removing duplicate entries**

   Removal of duplicate entries are necessary since it allows the analysis to be conducted more accurately and the output obtained is more logical. After importing our sample data in Excel, we went through the data to check if survey was completed for more than once from the same IP address. Our aim was to eliminate such respondent from our data set.

3. **Removal of unnecessary responses**

   This was the major issue faced by us during cleaning as our survey had many open ended questions. So, many non-serious respondents took unnecessary advantage of that and gave erroneous responses. We had to go through the open ended questions very thoroughly and we came across very absurd responses which were then eliminated. The responses which are not related to the survey needs to be removed for improving the quality of data on which analysis will be done since better the quality of data, the more accurate will be the analysis.

4. Removal of speeders and laggards

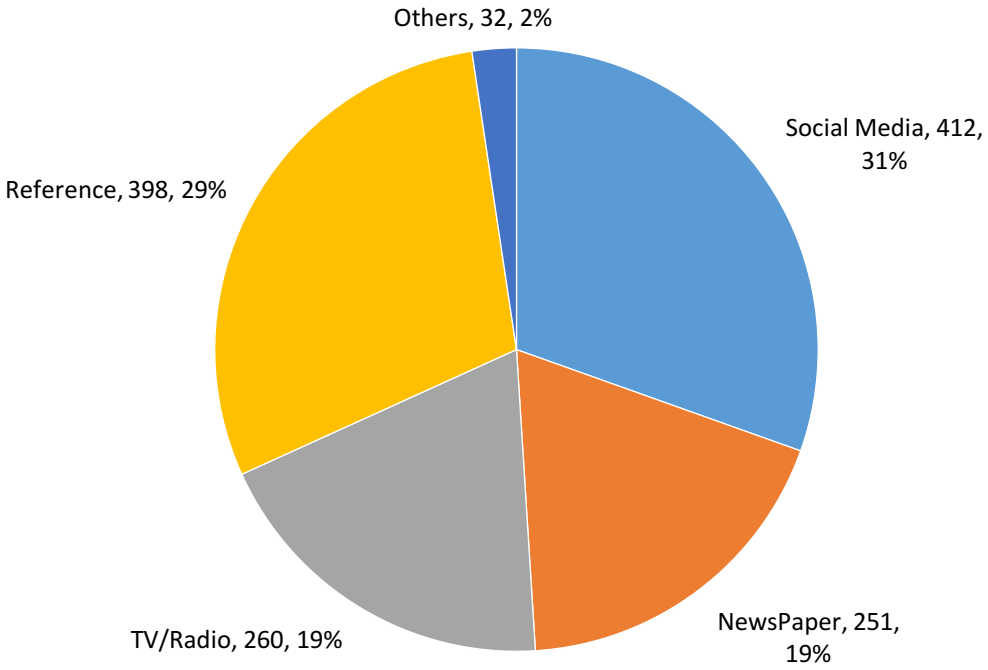
In this step, we identified the respondents who have completed the survey in very less time or took a lot of time than estimated time. We have estimated our survey to be completed in 7-10 minutes and removed all the respondents who completed the survey in less than 2 minutes (people choosing "yes" in Q.15) or have taken more than 30 minutes.

5. Code the responses

After all the above process was done, we came to the last step, which was making the data compatible with the tools on which we would run our analysis. The first step was to label a variable name into the required format. Now, the second step was to code the open end data into categories. Then, thirdly, we coded the data using whole numbers both in open and close ended questions. For example, in question where the respondent had to choose between yes and no we coded "yes" as 1 and "No" as 2. This made our data compatible with tools which we had used for further analysis.

AWARENESS OF CASHLESS

We did graphical representation to find out the sources from where people came to know about cashless payment systems and this is what we found:



As we can see from the above pie chart, majority of the population came to know about Cashless payment methods/systems by Social Media i.e. around 31% of the sample. Secondly, 29% of the sample population came to know about cashless method by their references i.e. from their family, friends, colleagues, etc. Lastly, 38% of the sample population came to know from Newspaper and from TV/Radio.

This tells us the increasing usage and importance of social media networks in our life in 2017 and how much of a major impact it has in the adoption of new systems. This can be leveraged further by the government and corporations to increase adoption and make it easier for people to use cashless systems especially, when shopping online.

OBJECTIVE 1

To identify and analyze the socio-demographic factors that affect the people's decision on whether to go cashless.

BINARY LOGISTIC REGRESSION

Logistic regression is part of a category of statistical models called generalized linear models.

Logistic regression is a predictive analysis, like linear regression. Logistic regression allows one to predict a discrete outcome from a set of variables that may be continuous, discrete and dichotomous or a mix of any of these. Here the dependent variable is dichotomous such as presence/ absence.

Binary logistic regression is a form of regression which is used when the dependent is a binary and the independents are of any type. Continuous variables are not used as dependents in logistic regression. Unlike logit regression, there can be only one dependent variable.

The goal of an analysis using logistic regression method is find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables and to determine the percent of variance in the dependent variable explained by the independents; to rank the relative importance of independents; to assess interaction effects; and to understand the impact of covariate control variables.

We used binary logistic regression since our dependent variable is categorical taking 2 values viz. using cashless payment systems and not using cashless payment systems.

The binary Regression model is,

$$P = \frac{e^{(\beta_0 + \sum_{i=1}^{k} \beta_i X_i)}}{1 + e^{(\beta_0 + \sum_{i=1}^{k} \beta_i X_i)}}$$

Where,

P = Pr (Y=1 | X) = Conditional probability that the outcome is present

Y : Response Variable

X : Vector of Independent Variables

It models the logit-transformed probability as a linear relationship with the predictor variables. More formally, let **y** be the binary outcome variable indicating failure/success with 0/1 and **p** be the probability of **y** to be 1,

p=prob(**y**=1). Let $x_1,\ldots,x_k$ be a set of predictor variables. Then the logistic regression of y on $x_1,\ldots,x_k$ estimates parameter values for $\beta_1,\ldots,\beta_k$ via maximum likelihood method of the following equation.

$$Logit(p) = \log\frac{p}{p-1} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

P: Probability that Y=1
Y: Dependent Variable

Note that LHS of the model can lie between $-\infty \; to \; \infty$.

We have 16 independent variables, out of which 4 are continuous and 12 are categorical. The independent variables are *Age, Gender, Education, Occupation, Religion, Type of House, Income, Marital Status, Family Type, Phone, Account in public sector bank, Account in Private sector bank, Account in a Co-operative bank, No. of Dependents, No. of members in the family, No. of Accounts by the person*.

ASSUMPTIONS

Logistic regression does not make many of the key assumptions of linear regression and general linear models that are based on ordinary least squares algorithms – particularly regarding linearity, normality, homoscedasticity, & measurement.

Firstly, it does not need a linear relationship between the dependent and independent variables. Logistic regression can handle all sorts of relationships, because it applies a non-linear log transformation to the predicted odds ratio. Also, the independent variables do not need to be multivariate normal – although multivariate normality yields a more stable solution. Also, the error terms (the residuals) do not need to be multivariate normally distributed. In addition to that, homoscedasticity is not needed. Logistic regression does not need variances to be heteroscedastic for each level of the independent variables. Lastly, it can handle ordinal and nominal data as independent variables. The independent variables do not need to be metric (interval or ratio scaled).
However, some other assumptions still apply.

Secondly, since logistic regression assumes that P(Y=1) is the probability of the event occurring, it is necessary that the dependent variable is coded

accordingly. That is, for a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.

Thirdly, the model should be fitted correctly. Neither over fitting nor under fitting should occur. That is only the meaningful variables should be included. A good approach to ensure this is to use a stepwise method to estimate the logistic regression.

Fourthly, the error terms need to be independent. Logistic regression requires each observation to be independent. That is that the data-points should not be from any dependent samples design, e.g., before-after measurements, or matched pairings.

Also the model should have little or no multicollinearity. That is that the independent variables should be independent from each other. However, there is the option to include interaction effects of categorical variables in the analysis and the model. If multicollinearity is present, centering the variables might resolve the issue, i.e. deducting the mean of each variable. If this does not lower the multicollinearity, a factor analysis with orthogonally rotated factors should be done before the logistic regression is estimated.

Fifthly, logistic regression assumes linearity of independent variables and log odds. Whilst it does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. Otherwise the test underestimates the strength of the relationship and rejects the relationship too easily, that is being not significant (not rejecting the null hypothesis) where it should be significant. A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the model. Another approach would be to use discriminant analysis, if the assumptions of homoscedasticity, multivariate normality, and absence of multicollinearity are met.

Lastly, it requires quite large sample sizes. Because maximum likelihood estimates are less powerful than ordinary least squares (e.g., simple linear regression, multiple linear regression); whilst OLS needs 5 cases per independent variable in the analysis, ML needs at least 10 cases per independent variable, some statisticians recommend at least 30 cases for each parameter to be estimated.

TESTING FOR MULTICOLLINEARITY

To test for multi-collinearity before we run the logistic regression model, we decided to run correlation of estimates on our chosen variables to make sure we only include those variables are actually relevant to our model analysis and ensure that bias remains minimum.

Correlation Matrix

		Constant	Age	Gender	Education	Occupation	Religion	House	Income	Family_Type	Dependents	Marital	No_Members	Phone	No_Accounts	Public	Private	Co_op
Step 1	Constant	1.000	-.174	-.072	-.216	.062	-.109	-.167	-.128	-.348	-.072	-.056	-.344	-.486	-.379	-.682	-.686	-.627
	Age	-.174	1.000	.199	.187	.016	.147	.112	-.115	-.082	.015	-.521	.098	-.019	-.214	.063	-.005	-.012
	Gender	-.072	.199	1.000	-.130	-.064	.081	-.018	-.025	-.158	.102	-.064	-.062	-.052	.018	-.066	-.020	-.088
	Education	-.216	.187	-.130	1.000	.131	.047	.121	-.082	-.004	-.039	-.129	.019	-.110	-.077	.118	.063	-.037
	Occupation	.062	.016	-.064	.131	1.000	-.085	.015	.018	-.190	-.067	.076	.015	-.226	.028	-.041	-.221	-.022
	Religion	-.109	.147	.081	.047	-.085	1.000	-.077	-.010	-.022	-.012	.007	-.035	.057	-.018	-.005	.001	-.035
	House	-.167	.112	-.018	.121	.015	-.077	1.000	.177	.088	.050	.004	-.125	-.062	-.032	.000	-.035	.115
	Income	-.128	-.115	-.025	-.082	.018	-.010	.177	1.000	-.096	-.079	-.036	-.037	.097	-.142	-.003	.032	.066
	Family_Type	-.348	-.082	-.158	-.004	-.190	-.022	.088	-.096	1.000	.089	.108	.363	.166	.072	.024	.081	.073
	Dependents	-.072	.015	.102	-.039	-.067	-.012	.050	-.079	.089	1.000	-.060	-.364	.000	.077	.030	.061	.098
	Marital	-.056	-.521	-.064	-.129	.076	.007	.004	-.036	.108	-.060	1.000	-.040	.003	.029	.008	.044	.008
	No_Members	-.344	.098	-.062	.019	.015	-.035	-.125	-.037	.363	-.364	-.040	1.000	.141	.016	.048	.102	.018
	Phone	-.486	-.019	-.052	-.110	-.226	.057	-.062	.097	.166	.000	.003	.141	1.000	.071	.185	.197	.175
	No_Accounts	-.379	-.214	.018	-.077	.028	-.018	-.032	-.142	.072	.077	.029	.016	.071	1.000	.384	.374	.376
	Public	-.682	.063	-.066	.118	-.041	-.005	.000	-.003	.024	.030	.008	.048	.185	.384	1.000	.661	.521
	Private	-.686	-.005	-.020	.063	-.221	.001	-.035	.032	.081	.061	.044	.102	.197	.374	.661	1.000	.473
	Co_op	-.627	-.012	-.088	-.037	-.022	-.035	.115	.066	.073	.098	.008	.018	.175	.376	.521	.473	1.000

Our cut-off for significant correlation was < -0.8 and > 0.8. So, as we can see from the above table, none of the correlations came close to this value. So, there is very little or no multi-collinearity in our selected variables.

MODEL BUILDING

Frequencies of dependent variables are given in following tables:

The LOGISTIC Procedure

Model Information		
Data Set	WORK.CASHLOGIT	
Response Variable	Cashless	Cashless
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	1114
Number of Observations Used	1114

Response Profile		
Ordered Value	Cashless	Total Frequency
1	1	912
2	0	202

Probability modeled is Cashless='1'.

DESIGN VARIABLES

The independent variables are converted to design variables. Twelve design variables are defined corresponding to the variable.

Class Level Information									
CLASS	VALUE	STRING VALUE	DESIGN VARIABLES						
Gender	1	Male	1						
	2	Female	0						
Education	1	Not studied	1	0	0	0	0		
	2	Upto_12th	0	1	0	0	0		
	3	Graduate	0	0	1	0	0		
	4	Masters	0	0	0	1	0		
	5	Research	0	0	0	0	1		
	6	No_answer	0	0	0	0	0		
Occupation	1	Private_sector	1	0	0	0	0	0	0
	2	Govt_sector	0	1	0	0	0	0	0
	3	Business	0	0	1	0	0	0	0
	4	Student	0	0	0	1	0	0	0
	5	Unemployed	0	0	0	0	1	0	0
	6	Retired	0	0	0	0	0	1	0
	7	Homemaker	0	0	0	0	0	0	1
	8	Freelance	0	0	0	0	0	0	0
Religion	1	Hindu	1	0	0	0	0	0	0

Class Level Information									
CLASS	**VALUE**	**STRING VALUE**	**DESIGN VARIABLES**						
	2	Islam	0	1	0	0	0	0	0
	3	Christian	0	0	1	0	0	0	0
	4	Sikh	0	0	0	1	0	0	0
	5	Buddhism	0	0	0	0	1	0	0
	6	Atheist	0	0	0	0	0	1	0
	7	Agnostic	0	0	0	0	0	0	1
	8	Other	0	0	0	0	0	0	0
Marital	1	Single	1	0	0				
	2	Married	0	1	0				
	3	Live_in	0	0	1				
	4	Divorced	0	0	0				
House	1	Apartment	1	0	0				
	2	Villa	0	1	0				
	3	Chawl	0	0	1				
	4	Slums	0	0	0				
Income	1	Less than 50,000	1	0	0	0	0		
	2	Between 50,000 and 1,00,000	0	1	0	0	0		
	3	Between 1,00,000 and 2,50,000	0	0	1	0	0		
	4	Between 2,50,000 and 5,00,000	0	0	0	1	0		
	5	Between 5,00,000 and 10,00,000	0	0	0	0	1		
	6	Above 10,00,000	0	0	0	0	0		
Family_Type	1	Joint	1						
	2	Nuclear	0						
Phone	1	Yes	1						
	2	No	0						
Public	1	Yes	1						
	2	No	0						
Private	1	Yes	1						

Class Level Information										
CLASS	**VALUE**	**STRING VALUE**	**DESIGN VARIABLES**							
	2	No	0							
Co_op	1	Yes	1							
	2	No	0							

MODEL

Here our dependent variable has categories:

Y = 1 Using cashless payment systems
= 0 Not using cashless payment systems

The logit function is:

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \sum_{j=1}^{5} \beta_{3j} X_{3j} + \sum_{j=1}^{7} \beta_{4j} X_{4j} + \sum_{j=1}^{7} \beta_{5j} X_{5j} + \sum_{j=1}^{3} \beta_{6j} X_{6j} + \sum_{j=1}^{5} \beta_{7j} X_{7j} + \sum_{j=1}^{3} \beta_{8j} X_{8j} + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{14} X_{14} + \beta_{15} X_{15} + \beta_{16} X_{16}$$

where,

- β_0 = intercept
- β_1 = coefficient of the Age variable
- β_2 = coefficient of the Gender variable
- β_{3j}= coefficient of the jth design variable of the 3rd categorical variable (Education) where j = 1 to 5.
- β_{4j}= coefficient of the jth design variable of the 4th categorical variable (Occupation) where j = 1 to 7.
- β_{5j} = coefficient of the jth design variable of the 5th categorical variable (Religion) where j = 1 to 7.
- β_{6j} = coefficient of the jth design variable of the 6th categorical variable (House) where j = 1 to 3.

- β_{7j} = coefficient of the jth design variable of the 7th categorical variable (Income) where j = 1 to 5.
- β_{8j} = coefficient of the jth design variable of the 8th categorical variable (Marital Status) where j = 1 to 3.
- β_9 = coefficient of the Family Type variable.
- β_{10}= coefficient of the Smartphone variable
- β_{11} = coefficient of the Public Sector bank variable.
- β_{12} = coefficient of the Private Sector bank variable
- β_{13} = coefficient of the Co-operative Sector bank variable
- β_{14} = coefficient of the No. of dependents in a family variable.
- β_{15} = coefficient of the No. of members in a family variable.
- β_{16} = coefficient of the No. of accounts variable.
- Errors follow binomial distribution
- X is the vector of design and independent variables.

The conditional probabilities of each Outcome Category, given the covariate vectors are as follows:

$$P_1 = P(Y = 1|X) = e^{(\frac{g(x)}{1+e^{g(x)}})}$$
$$P_2 = P(Y = 0|X) = e^{(\frac{g(x)}{1+e^{g(x)}})}$$

LIKELIHOOD FUNCTION

The likelihood function expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of these parameters are chosen to be those values which maximise this function.

The likelihood function: -

$$I(\beta) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

Where,

If Y = 1 then $Y_1 = 1, Y_2 = 0$
If Y = 0 then $Y_1 = 0, Y_2 = 1$

The likelihood equation are found by taking the first partial derivative of $L(\beta) = \ln l(\beta)$ with respect to each of the unknown parameters. The maximum likelihood estimator is obtained by setting these equations equal to zero and solving for β.

STEPWISE SELECTION PROCEDURE

Stepwise method is a process of building a model by successively adding or removing variables.

Any stepwise procedure for selection or deletion of variables from a model is based on statistical algorithm which check for the importance of variables, and either include or exclude them on the basis of fixed decision rule. The importance of the variable is defined in terms of a measure of the statistical significance is assumed via likelihood ratio chi-squared test.

Thus, at any step in the procedure the most important variable, in statistical terms, will be the one that produces the greatest change in the log-likelihood relative to a model not containing the variable (i.e., the one that would result in the largest likelihood ratio statistic).

For the stepwise selection procedure, we kept SLS and SLE at **5%**.

The summary of stepwise selection procedure given by SAS is as follows:

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	House		3	1	38.2205		<.0001
2	No_Accounts		1	2	24.7078		<.0001
3	Age		1	3	11.9170		0.0006
4	Occupation		7	4	29.2177		0.0001
5	Gender		1	5	20.0351		<.0001
6	Private		1	6	12.9773		0.0003
7	Phone		1	7	14.0380		0.0002
8	Income		5	8	19.7080		0.0014
9	Public		1	9	4.8561		0.0275

Note: No (additional) effects met the 0.05 significance level for entry into the model.

As we can see from the summary of stepwise selection procedure, there were 9 variables extracted from the 16 we originally entered into the model. We can see that these 9 variables have p-value < 0.05 and so have significant impact on our dependent variable.

GLOBAL TESTING

Global testing is used to test whether or not, at least one of the independent variables influence the dependent variables i.e. at least one of the independent variables insignificant.

H_0 = The Design variables entered into the model by stepwise procedure are insignificant

$$\beta_1 = \beta_2 = \ldots = \beta_k = 0$$

V/s

H_1 = The Design variables entered into the model by stepwise procedure are significant OR At least one coefficient is not zero.

Test Statistic:

$\chi^2 = L_1 - L_2$ which follows chi-square distribution with k df.

$L_1 = -2$ Log L with only constant term and no independent variables.

$L_2 = -2$ Log L with k independent variables and a constant term.
Where, L is the likelihood function.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	176.3446	21	<.0001
Score	160.8875	21	<.0001
Wald	127.7253	21	<.0001

Conclusion:

Since p-value < 0.05, we reject H_0.Hence, at least one of the independent variable is significant.

ANALYSIS OF EFFECTS

The hypothesis is:

H_0 = Individual independent variable is insignificant

H_1 = Individual independent variable is significant

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Age	1	28.9109	<.0001
Gender	1	19.6550	<.0001
Occupation	7	28.4680	0.0002
House	3	13.9932	0.0029
Income	5	20.2287	0.0011
Phone	1	13.1420	0.0003

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Public	1	4.8237	0.0281
Private	1	18.8495	<.0001
No_Accounts	1	11.3951	0.0007

Analysis of Effects Eligible for Entry			
Effect	**DF**	**Score Chi-Square**	**Pr > ChiSq**
Education	5	3.1842	0.6716
Religion	7	13.0189	0.0716
Marital	3	1.8503	0.6040
Family_Type	1	1.5677	0.2105
Co_op	1	0.2319	0.6301
Dependents	1	1.4457	0.2292
No_Members	1	0.3265	0.5677

Conclusion:

Since 9 variables have p-values less than 0.05, we reject H_0 for each of the 9 variables, i.e., each of the 9 variables have overall significance in the model.

As we can see in the above table of variables eligible of entry, there are 7 variables that have the p-value greater than 0.05, we fail to reject H_0 for each of these 6 variables, i.e., each of these 7 variables fail to have any significant effects in the model.

PARAMETER ESTIMATION AND INDIVIDUAL TESTING

The parameter (logit) estimates are nothing but the M.L.E. estimates obtained by partial differentiation of the natural log of likelihood function with respect to each of the unknown parameters and equating the resultant equations to zero. Iterative method is used for computing the estimates. The standard interpretation of the multinomial logit is that for a unit change in the predictor change, the logit of the outcome relative to the reference group is expected to change by its respective parameter (which is in log-odds unit) given the other variables in the models are held constant.

Intercept: This is the multinomial logit estimates when the predictor variables in the model are evaluated at zero.

To test the hypotheses:

H_0 = Individual coefficients of independent variables are zero OR $\beta_i = 0$

H_1 = Individual coefficients of independent variables are not zero OR $\beta_i \neq 0$

To test the above hypotheses, Wald's Statistic is used. It is defined as the ratio of estimated coefficient to its estimated standard error.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	1.2976	1.3944	0.8660	0.3521
Age		1	-0.0545	0.0101	28.9109	<.0001
Gender	**1**	1	0.8331	0.1879	19.6550	<.0001
Occupation	**1**	1	-3.4134	1.1832	8.3220	0.0039
Occupation	**2**	1	-2.9177	1.2454	5.4884	0.0191
Occupation	**3**	1	-4.1107	1.2085	11.5707	0.0007
Occupation	**4**	1	-3.8645	1.1909	10.5300	0.0012
Occupation	**5**	1	-3.3596	1.2561	7.1531	0.0075
Occupation	**6**	1	10.5377	506.0	0.0004	0.9834
Occupation	**7**	1	-0.6233	1.5322	0.1655	0.6842
House	**1**	1	1.8721	0.7022	7.1067	0.0077
House	**2**	1	1.6169	0.7438	4.7260	0.0297
House	**3**	1	1.2079	0.7204	2.8111	0.0936
Income	**1**	1	-0.3165	0.4096	0.5974	0.4396
Income	**2**	1	-0.4730	0.3529	1.7963	0.1802
Income	**3**	1	-0.3533	0.3059	1.3336	0.2482
Income	**4**	1	0.0882	0.2995	0.0868	0.7682
Income	**5**	1	0.8592	0.3371	6.4956	0.0108
Phone	**1**	1	1.9463	0.5369	13.1420	0.0003
Public	**1**	1	0.4701	0.2140	4.8237	0.0281

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Private	**1**	1	0.9705	0.2235	18.8495	<.0001
No_Accounts		1	0.3420	0.1013	11.3951	0.0007

The variables – Occupation (6,7) = Retired, Homemaker as well as House (Chawl) are insignificant. Almost all income levels except the one between 5,00,000 and 10,00,000 are insignificant.

GOODNESS OF FIT

With logistic regression, instead of R^2 as the statistics for overall fit of the linear regression model, deviance between observed values from the expected values is used. In linear regression, residuals can be defined as $y_i - \hat{y}_i$.

Where y_i is the observed dependent variable for the i^{th} subject, and $\hat{y}_i$ the corresponding prediction from the model. The same concept applied to logistic regression, where y_i is equal to either 1 or 0.

We can use 2 tests for testing the goodness-of-fit.

HOSMER & LEMESHOW TEST:

The **Hosmer–Lemeshow test** is a statistical test for goodness of fit for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar are called well calibrated.

The Hosmer-Lemeshow statistic evaluates the goodness of fit by creating ordered groups of subjects and then comparing the number actually in each group (observed) to the number predicted by the logistic regression model (predicted).

The statistic used is a chi-square statistic with desirable outcome of non-significance, indicating the model prediction does not significantly differ from the observed.

Hypothesis:

H_0 = Model is a good fit for the data

H_1 = Model is not a good fit for the data

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
4.3204	8	0.8271

Since p-value > 0.05, we fail to reject H_0 and conclude that our model is a good fit.

RESIDUAL CHI-SQUARE TEST:

The residual chi-square test is carried out to test the significance of the remaining independent variables which have not been entered into the model.

Hypothesis:

H_0 = The reduced model is as good as the full model.

H_1 = The reduced model is not as good as the full model.

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
23.8676	19	0.2013

Since the p-value is > 0.05, we fail to reject H_0.

Hence, the left out variables are insignificant and we proceed with the 9 variables given by the stepwise procedure.

DETECTING INFLUENTIAL OBSERVATIONS

Cook's distance measures a level of influence of an observations by taking into account both the size of the residuals and the amount of leverage for that observations.

For our data,

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Analog of Cook's influence statistics	1114	.00003	.25953	.0153237	.02766839
Valid N (listwise)	1114				

All values of cook's distance Di < 1, so there is no influential observations.

FITTED MODEL

$$g(x) = 1.2976 - 0.0545(X_{11}) + 0.8331(X_{21}) - 3.4134(X_{41}) - 2.9177(X_{42}) - 4.1107(X_{43}) - 3.8645(X_{44}) - 3.3596(X_{45}) + 1.8721(X_{61}) + 1.6169(X_{62}) + 0.8592(X_{75}) + 1.9463(X_{101}) + 0.4701(X_{111}) + 0.9705(X_{121}) + 0.3420(X_{161})$$

Where,

X_{11} = Age
X_{21} = Gender
X_{41} = Occupation 1 (Private Sector)
X_{42} = Occupation 2 (Govt. Sector)
X_{43} = Occupation 3 (Business)
X_{44} = Occupation 4 (Student)
X_{45} = Occupation 5 (Unemployed)
X_{61} = House 1 (Apartment)
X_{62} = House 2 (Villa / Bungalow)
X_{75} = Income 5 (Between 5,00,000 and 10,00,000)
X_{101} = Smartphone
X_{111} = Account in Public Sector bank
X_{121} = Account in Private Sector bank
X_{161} = No. of accounts

ODDS RATIO

Odds ratio is a measure of association. It approximates how much more likely it is for outcome to be present among the different levels of independent variables.

In general, for binary logistic regression model, if Y = 0 is the reference outcome then, odds ratio of outcome j versus outcome 0 for X = a versus X = b is given as follows:

$$\psi j(a,b) = \frac{[\frac{P\left(\frac{Y=j}{x=a}\right)}{P\left(\frac{Y=0}{x=a}\right)}]}{[\frac{P\left(\frac{Y=j}{x=b}\right)}{P\left(\frac{Y=0}{x=b}\right)}]}$$

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
Age	0.947	0.928	0.966
Gender 1 vs 2	2.300	1.592	3.325
Occupation 1 vs 8	0.033	0.003	0.335
Occupation 2 vs 8	0.054	0.005	0.621
Occupation 3 vs 8	0.016	0.002	0.175
Occupation 4 vs 8	0.021	0.002	0.216
Occupation 5 vs 8	0.035	0.003	0.408
Occupation 6 vs 8	>999.999	<0.001	>999.999
Occupation 7 vs 8	0.536	0.027	10.803
House 1 vs 4	6.502	1.642	25.750
House 2 vs 4	5.037	1.173	21.642
House 3 vs 4	3.346	0.815	13.734
Income 1 vs 6	0.729	0.327	1.626
Income 2 vs 6	0.623	0.312	1.244
Income 3 vs 6	0.702	0.386	1.279
Income 4 vs 6	1.092	0.607	1.964
Income 5 vs 6	2.361	1.220	4.572
Phone 1 vs 2	7.002	2.445	20.056
Public 1 vs 2	1.600	1.052	2.434
Private 1 vs 2	2.639	1.703	4.090
No_Accounts	1.408	1.154	1.717

Interpretation:

1. Gender:
    a. Males are 2.3 times more likely to use cashless payment systems compared to females.
2. Occupation:
    a. Retired individuals are almost 1000 times more likely to be using cashless payment systems rather than people in freelance occupations.
3. House:
    a. People living in apartments are 6.5 times more likely to be using cashless payment systems compared to people living in slums.
    b. People living in villas/ bungalows are 5.307 times more likely to be using cashless payment systems compared to people living in slums.
    c. People living in chawls are 3.346 times more likely to be using cashless payment systems compared to people living in slums.
4. Income:
    a. People earning between 2,50,000 and 5,00,000 are 1.092 times more likely to be using cashless payment systems compared to those earning above 10,00,000.
    b. People earning between 5,00,000 and 10,00,000 are 2.361 times more likely to be using cashless payment systems compared to those earning above 10,00,000.
5. Phone
    a. People who use smartphones are 7.002 times likely to be using cashless payment systems compared to people who don't.
6. Public
    a. People who have accounts in public banks are 1.6 times likely to be using cashless payment systems compared to those who don't.
7. Private
    a. People who have accounts in private banks are 2.639 times more likely to be using cashless payment systems compared to those who don't.

CLASSIFICATION TABLES

The classification table is a cross-tabulation of observed and predicted frequencies for the dependent values of Y.
These observed and predicted values y values are cross tabulated to get the classification table as follows:

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	887	24	178	25	81.8	97.3	11.9	16.7	51.0

The table shows that 911 out of 1114 (i.e. 81.8%) matches are correct. Hence, we can conclude that the model predicts the values correctly 81.8% of the times.

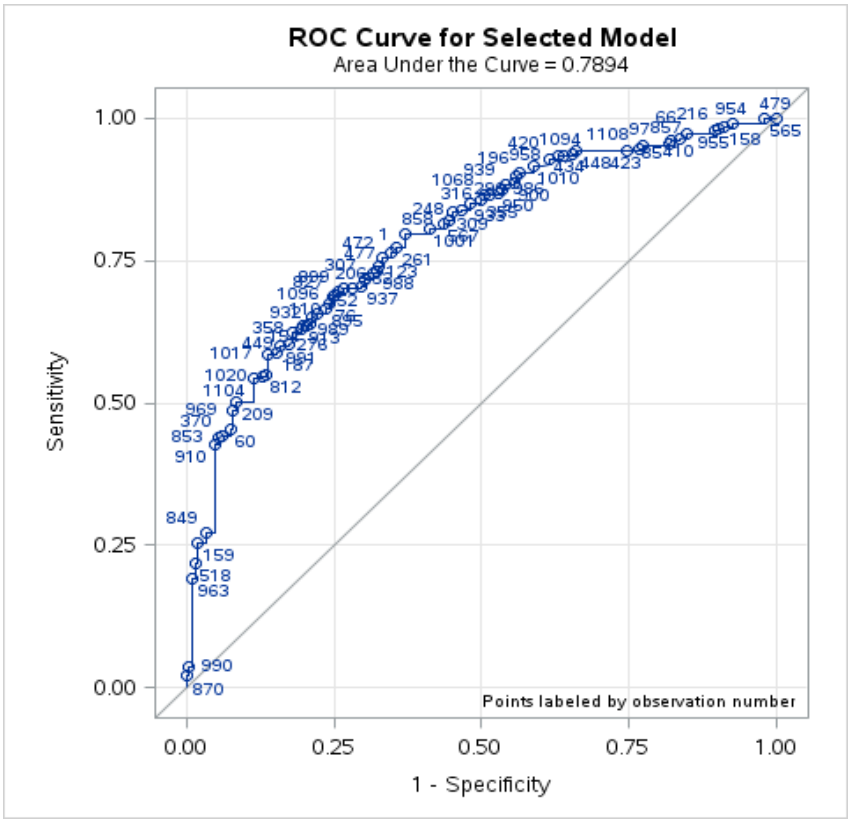
ROC CURVE

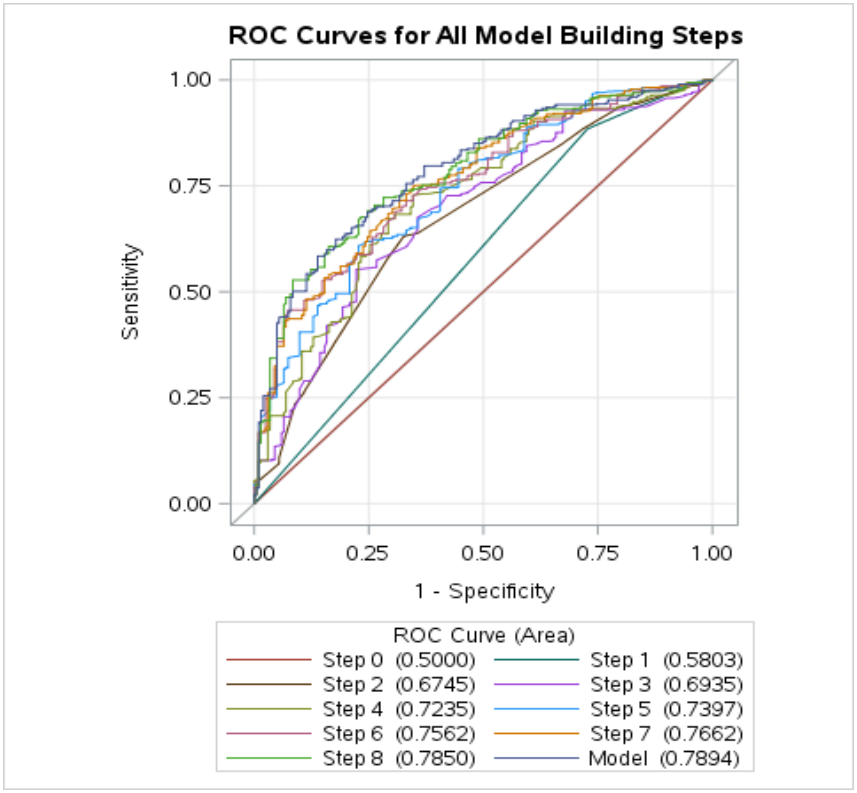
Receiver Operating Characteristic i.e. ROC curve are used to evaluate and compare the performance of diagnostic tests. They can also be used to evaluate model fit. An ROC curve is just a plot of proportion of true positives (events predicted to be events i.e. Sensitivity) versus the proportion of false positives (non-events predicted to be events i.e. Specificity).

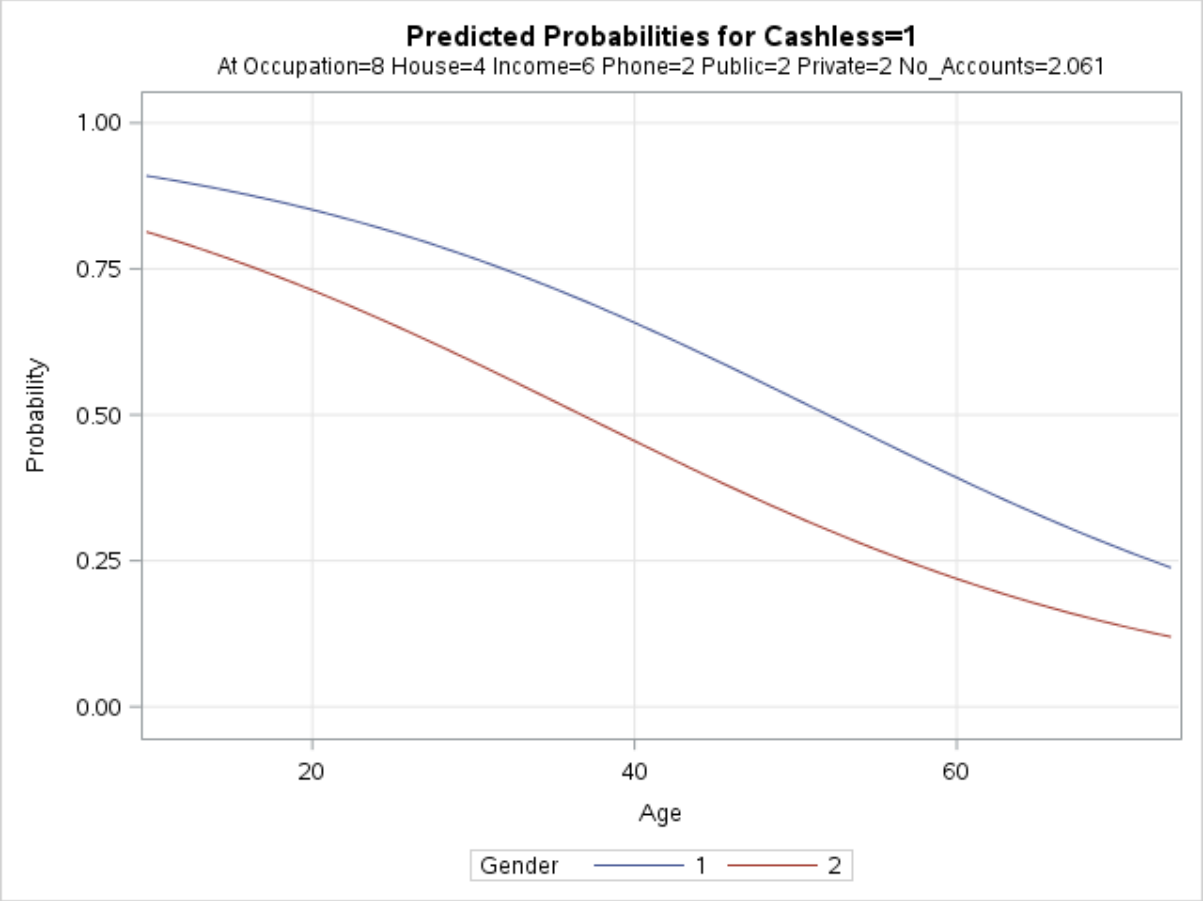
The accuracy of test is measured by the area under the ROC curve. An area of 1 represent a perfect test, while an area of 0.5 represents a worthless test. The closer the curve follows the left hand border and then the top border of the ROC space, the more accurate the test, the true positive rate (sensitivity) is high and the false positive rate (1-specificity) is low. Statistically, more area under the curve means that it is identifying more true positives while minimising the number/ percent of false positives.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.9	**Somers' D**	0.579
Percent Discordant	21.1	**Gamma**	0.579
Percent Tied	0.0	**Tau-a**	0.172
Pairs	184224	**c**	0.789

Area under the ROC curve is estimated by the statistic c in the "Association of predicted Probabilities and Observed Responses" table. Hence, the areas under the ROC curve is 0.789.







Since our ROC curve rises quickly, the sensitivity is high while keeping low 1-specificity. Hence, we have high predictive accuracy.

OBJECTIVE 2

To identify and analyse other latent (qualitative) factors that influence the people's preference to choose cash over cashless.

FACTOR ANALYSIS

INTRODUCTION

Analyze the structure of the inter-relationship (correction) among a large set of decision variables to determine whether information can be summarized into smaller set of factors.

Factor analysis is a useful method of reducing data complexity by reducing the number of variable under study. In general, factor analysis is a set of techniques which analyzes correlation between variables, and reduces their number into fewer uncorrelated and unobservable "Factors" which explain much of the original data, more economically.

Thus, factor analysis is a general name denoting a class of procedures primarily used for data reduction and summarization.

OBJECTIVE

To summarize the information contained in the number of decision variables into smaller set of factors subjected to its minimum loss of information.

It provides as a tool to better interpret the results of observations when a large number of decision variables is grouped into smaller set of factors.

ASSUMPTION

Mean

1. The specific factors or errors all have mean zero:
   $E(\varepsilon_i) = 0; i = 1,2, \dots, p$

   We assume that these errors are random. So, they will have a mean and the mean we will assume will be zero here.
2. The common factors, the f's, have mean zero:
   $E(f_i) = 0; i = 1,2, \dots, m$

   All of the unobserved explanatory variables will have a mean zero.

A consequence of these assumptions is that the mean response of the j[th] trait is μ_i. That is,

$E(X_i) = \mu_i$

VARIANCE

1. The common factors have variance one: $Var\ (f_i) = 0; i = 1,2, \dots, m$ These unobserved common factors are all assumed to have variance of one.
2. The variance of specific factors i is $?_i : Var\ (\varepsilon_i) = 0; i = 1,2, \dots, m$ or the errors or specific factors are assumed to have variances, s $?_i$, for the ith specific factor. Here, $?_i$ is called the specific variance.

CORRELATION

1. The common factors are uncorrelated with one another:
$$Cov\ (f_i, f_j) = 0\ for\ i \neq j = 1,2, \dots, m$$
2. The specific factors are uncorrelated with one another:
$$Cov\ (\varepsilon_i, \varepsilon_j) = 0\ for\ i \neq j = 1,2, \dots, p$$

3. The specific factors are uncorrelated with the common factors:
$$Cov\ (\varepsilon_i, f_j) = 0\ for\ i = 1,2, \dots, p; j = 1,2, \dots, m$$

These assumptions are necessary because in this case if we do not add these constraints to our model it is not possible to uniquely estimate any of the parameters. You could get an infinite number of equally well fitting models with different values for the parameters unless we add these constraints if these assumptions are not made.

MODEL REPRESENTATION

The basic model representation of exploratory factor analysis in matrix terms is:

$$\boldsymbol{x = \mu + LF + \varepsilon}$$

Where,

μ = Means
L = Factor Loadings
F = Factors
ε = Error Matrix

EXTRACTION METHODS

- Principal Component Method - the goal is to extract as much variance with the least number of factors.
- Maximum Likelihood - Computational intensive method for estimating loadings that maximize the likelihood (probability) of the correlation matrix.

ROTATION METHODS

Orthogonal Rotation Methods

Varimax

- It has a simple structure by maximizing variance loadings within factors across variables.
- It makes large loadings larger and small loadings smaller.
- It spreads the variances from first (largest) factor to other smaller factors.

Quartimax

- It is the opposite of varimax.
- It simplifies variables by maximizing variance with variables across factors.
- Varimax works on the columns of the loading matrix; Quartimax works on the rows.
- It is not used often.
- It is usually not a goal to simplify variables.

Equamax

- It is a hybrid of the earlier two that tries to simultaneously simplify factors and variables.
- It is not that popular either.

FINAL ANALYSIS

To check the factors responsible for how likely people use cashless payment method:

We have considered the following 15 variables for the analysis:

F1 : Food in Restaurant
F2 : Drinks at a bar/coffee shop etc.
F3 : Physical goods in retail shop
F4 : Trade works in the home
F5 : Tickets for events
F6 : Donation to charity
F7 : Paying a Friend
F8 : Taxi/cab
F9 : Motor fuel
F10 : Public Transport
F11 : Goods over the phone
F12 : Goods on web
F13 : Goods in market
F14 : Utility bills
F15 : Gov. services (TV License, passport, etc.)

CORRELATION MATRIX

Correlations

		F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
F1	F1	1.00000	0.63673	0.40080	0.38916	0.52187	0.28404	0.27124	0.39868	0.47833	0.25512	0.31477	0.29865	0.33842	0.22162	0.39178
F2	F2	0.63673	1.00000	0.37863	0.37271	0.40134	0.36949	0.34979	0.34255	0.41895	0.27452	0.30129	0.36278	0.36679	0.30288	0.36675
F3	F3	0.40080	0.37863	1.00000	0.55862	0.43285	0.43295	0.43807	0.43275	0.39718	0.48567	0.34177	0.32965	0.61363	0.38966	0.48277
F4	F4	0.38916	0.37271	0.55862	1.00000	0.40743	0.43587	0.43505	0.43863	0.42374	0.51980	0.42372	0.24403	0.49711	0.29338	0.41198
F5	F5	0.52187	0.40134	0.43285	0.40743	1.00000	0.31249	0.34513	0.43913	0.43477	0.29717	0.34311	0.39954	0.30683	0.32998	0.37544
F6	F6	0.28404	0.36949	0.43295	0.43587	0.31249	1.00000	0.49084	0.40345	0.38301	0.48891	0.43353	0.33772	0.34600	0.37948	0.37387
F7	F7	0.27124	0.34979	0.43807	0.43505	0.34513	0.49084	1.00000	0.53199	0.40960	0.53763	0.47728	0.36525	0.50045	0.36581	0.37912
F8	F8	0.39868	0.34255	0.43275	0.43863	0.43913	0.40345	0.53199	1.00000	0.48009	0.53379	0.47256	0.45672	0.47408	0.33389	0.47807
F9	F9	0.47833	0.41895	0.39718	0.42374	0.43477	0.38301	0.40960	0.48009	1.00000	0.43125	0.45464	0.38216	0.46694	0.37803	0.48767
F10	F10	0.25512	0.27452	0.48567	0.51980	0.29717	0.48891	0.53763	0.53379	0.43125	1.00000	0.48647	0.24441	0.55775	0.32546	0.40890
F11	F11	0.31477	0.30129	0.34177	0.42372	0.34311	0.43353	0.47728	0.47256	0.45464	0.48647	1.00000	0.52558	0.45094	0.34737	0.38004
F12	F12	0.29865	0.36278	0.32965	0.24403	0.39954	0.33772	0.36525	0.45672	0.38216	0.24441	0.52558	1.00000	0.37764	0.41543	0.37150
F13	F13	0.33842	0.36679	0.61363	0.49711	0.30683	0.34600	0.50045	0.47408	0.46694	0.55775	0.45094	0.37764	1.00000	0.43685	0.45275
F14	F14	0.22162	0.30288	0.38966	0.29338	0.32998	0.37948	0.36581	0.33389	0.37803	0.32546	0.34737	0.41543	0.43685	1.00000	0.45883
F15	F15	0.39178	0.36675	0.48277	0.41198	0.37544	0.37387	0.37912	0.47807	0.48767	0.40890	0.38004	0.37150	0.45275	0.45883	1.00000

A correlation matrix (or covariance) matrix is a symmetric matrix showing the simple correlation (or covariance), between all possible pairs of variables included in the analysis.

This table gives the correlations between the original variables (which are specified on the **/variables** subcommand). Before conducting a principal components analysis, you want to check the correlations between the variables. If any of the correlations are too high (say above .9), you may need to remove one of the variables from the analysis, as the two variables seem to be measuring the same thing. Another alternative would be to combine the variables in some way (perhaps by taking the average). If the correlations are too low, say below .1, then one or more of the variables might load only onto one principal component (in other words, make its own principal component). This is not helpful, as the whole point of the analysis is to reduce the number of items (variables).

So, as we can see from the above table, none of the correlations came close to these values.

KMO

- The KMO measure is used for sample adequacy.
- A small value of the KMO statistic indicates that the correlations between pairs of variables cannot be explained by the other variables and that the factor analysis may not be appropriate.
- Generally, a value greater than 0.5 is desirable.

BARLETT'S TEST OF SPHERICITY

- Barlett's test of sphericity test the hypothesis that the correlation matrix is an identity matrix.
- This test which is often done prior to factor analysis, tests whether the data comes from multivariate normal distribution with zero covariances.
- We proceed with factor analysis only if the above null hypothesis is rejected.

Below is the output of the data we analysed:

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.914
Bartlett's Test of Sphericity	Approx. Chi-Square	6109.306
	Df	105
	Sig.	.000

The value of KMO statistics (0.914) is large (>0.5).
Thus we proceed for FACTOR ANALYSIS as an appropriate technique of data reduction.

Hypothesis for Barlett's test:

H_0 = Population correlation matrix is an identity matrix

H_1 = Population correlation matrix is not an identity matrix

Since, the p-value < 0.05, we reject null hypothesis. Therefore, the population correlation matrix is not an identity matrix.

COMMUNALITIES

The portion of the variance of the i^{th} variable contributed by the m common factors is called the i^{th} communality. This is the proportion of each variable's variance that can be explained by the factors (e.g., the underlying latent continua). It is also noted as h^2 and can be defined as the sum of squared factor loadings for the variables.

- **Initial** – By definition, the initial value of the communality in a principal components analysis is 1.
- **Extraction** – The values in this column indicate the proportion of each variable's variance that can be explained by the principal components. Variables with high values are well represented in the common factor space, while variables with low values are not well represented. (In this example, we don't have any particularly low values.) They are the reproduced variances from the number of components that you have saved. You can find these values on the diagonal of the reproduced correlation matrix.

Communalities

	Initial	Extraction
Food_restaurant	1.000	.800
Drinks	1.000	.645
Physicalgoods_store	1.000	.611
Tradesworkathome	1.000	.648
Tickets_events	1.000	.550
Donation	1.000	.456
Paying_friend	1.000	.569
Taxi	1.000	.542
Motorfuel	1.000	.520
Publictransport	1.000	.710
Goods_phone	1.000	.584
Goods_web	1.000	.755
Goods_market	1.000	.594
Utility_bills	1.000	.482
Govt_services	1.000	.470

Extraction Method: Principal Component Analysis.

As we can see from the above table, it shows the proportion of variation explained by each of the original variable, before and after extraction.

TOTAL VARIANCE EXPLAINED

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.698	44.655	44.655	6.698	44.655	44.655	3.706	24.710	24.710
2	1.245	8.303	52.958	1.245	8.303	52.958	2.656	17.709	42.418
3	.992	6.612	59.570	.992	6.612	59.570	2.573	17.152	59.570
4	.825	5.499	65.069						
5	.712	4.747	69.816						
6	.639	4.263	74.079						
7	.629	4.196	78.275						
8	.567	3.779	82.054						
9	.517	3.449	85.503						
10	.444	2.961	88.464						
11	.423	2.823	91.287						
12	.403	2.688	93.975						
13	.343	2.288	96.263						
14	.289	1.924	98.187						
15	.272	1.813	100.000						

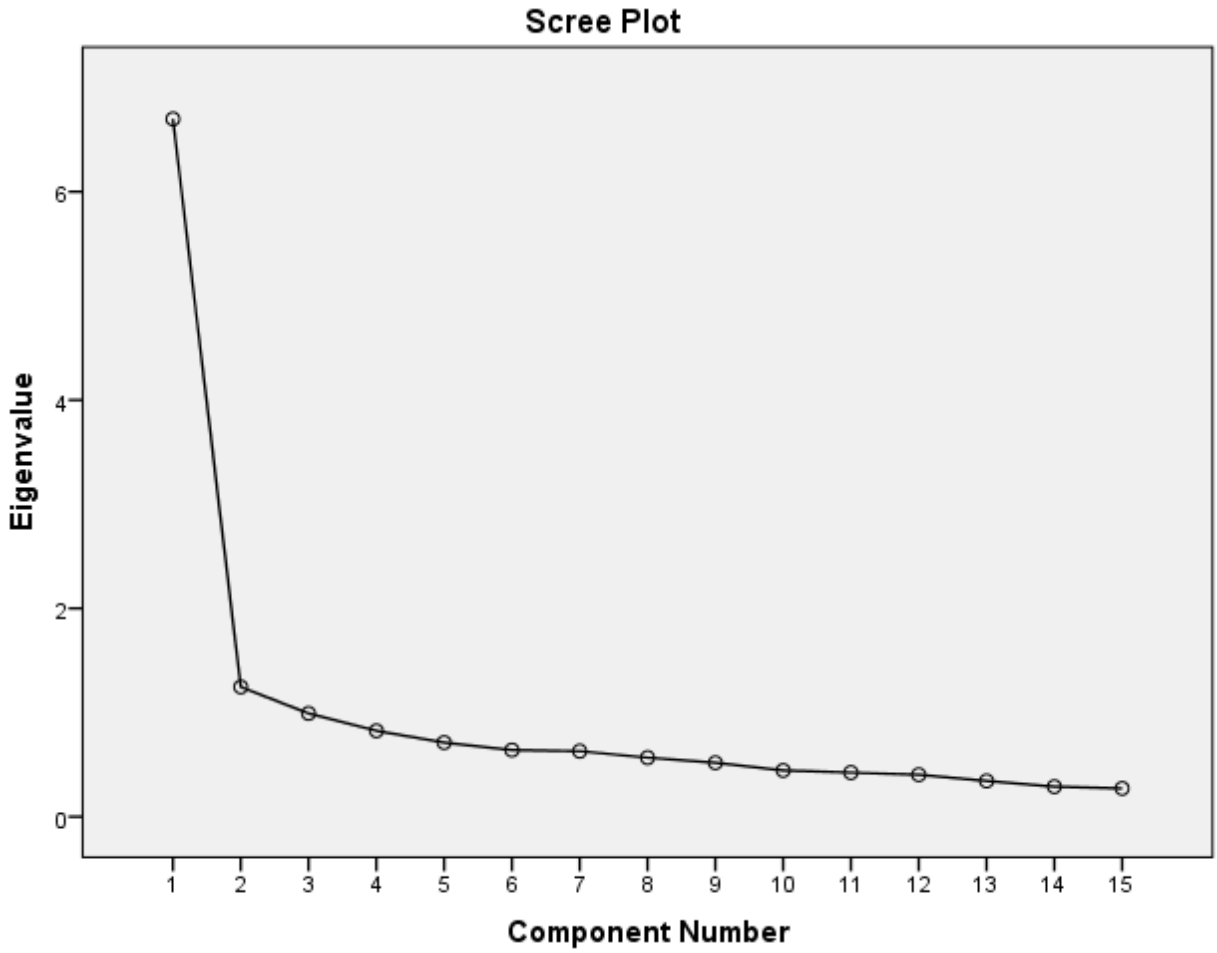
Extraction Method: Principal Component Analysis.

- **Component** – There are as many components extracted during a principal components analysis as there are variables that are put into it. In our study, we used 15 variables, so we have 15 components.
- **Initial Eigenvalues** – Eigen values are the variances of the principal components. Because we conducted our principal components analysis on the correlation matrix, the variables are standardized, which means that each variable has a variance of 1, and the total variance is equal to the number of variables used in the analysis, in this case, 15.
- **Total** – This column contains the eigenvalues. The first component will always account for the most variance (and hence have the highest eigenvalue), and the next component will account for as much of the left over variance as it can, and so on. Hence, each successive component will account for less and less variance.

- **% of Variance** – This column contains the percent of variance accounted for by each principal component.
- **Cumulative %** – This column contains the cumulative percentage of variance accounted for by the current and all preceding principal components. For example, the third row shows a value of 59.570. This means that the first three components together account for 59.570% of the total variance. (Remember that because this is principal components analysis, all variance is considered to be true and common variance. In other words, the variables are assumed to be measured without error, so there is no error variance.)
- **Extraction Sums of Squared Loadings** – The three columns of this half of the table exactly reproduce the values given on the same row on the left side of the table. The number of rows reproduced on the right side of the table is determined by the number of principal components whose eigenvalues are 1 or greater or in our case, 3 factors extracted.

Since, the first 2 Eigen values are greater than 1 and 3rd one is close to 1 with 0.992 value, we considered three factors for further analysis. The three factors together explain 59.57% of the total variation in the sample. Hence, three factors will be retained using MINIMUM EIGEN criteria.

SCREE PLOT



The scree plot graphs the eigenvalue against the component number. You can see these values in the first two columns of the table. From the third component on, you can see that the line is almost flat, meaning that each successive component is accounting for smaller and smaller amounts of the total variance. In general, we are interested in keeping only those principal components whose eigenvalues are greater than or closer to 1. Components with an eigenvalue of less than 1 account for less variance than did the original variable (which had a variance of 1), and so are of little use. Hence, you can see that the point of principal components analysis is to redistribute the variance in the correlation matrix (using the method of eigenvalue decomposition) to redistribute the variance to first components extracted.

From our data, The Scree plot shows that the graph starts to decrease after first 3 factors. This also supports the analysis regarding number of factors.

COMPONENT MATRIX

	Component		
	1	2	3
Food_restaurant	.612	.625	
Drinks	.615	.508	
Physicalgoods_store	.717		
Tradesworkathome	.690		
Tickets_events	.628		
Donation	.645		
Paying_friend	.694		
Taxi	.727		
Motorfuel	.704		
Publictransport	.693		
Goods_phone	.676		
Goods_web	.602		.619
Goods_market	.726		
Utility_bills	.589		
Govt_services	.683		
Extraction Method: Principal Component Analysis. Only three components extracted.			

This table contains component loadings, which are the correlations between the variable and the component. Because these are correlations, possible values range from -1 to +1. On the **/format** subcommand, we used the option **blank(.45)**, which tells SPSS not to print any of the correlations that are .45 or less. This makes the output easier to read by removing the clutter of low correlations that are probably not meaningful anyway.

Component – The columns under this heading are the principal components that have been extracted. As we can see, three components were extracted. We usually do not try to interpret the components the way that we would factors that have been extracted from a factor analysis. Rather, we are interested in the component scores, which are used for data reduction (as opposed to factor analysis where we are looking for underlying latent continua).

As we can from the above table, Food in a restaurant and drinks at a bar/ coffee shop as well as Goods on web is explained by more than one component. So, we will do a Varimax rotation to clarify these extractions.

ROTATED COMPONENT MATRIX

	Component		
	1	2	3
Food_restaurant		.873	
Drinks		.758	
Physicalgoods_store	.674		
Tradesworkathome	.719		
Tickets_events		.643	
Donation	.551		
Paying_friend	.626		
Taxi	.499		.459
Motorfuel		.474	
Publictransport	.814		
Goods_phone			.635
Goods_web			.828
Goods_market	.680		
Utility_bills			.620
Govt_services			
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.			
a. Rotation converged in 6 iterations.			

This table contains the rotated component loadings, which represent both how the variables are weighted for each component but also the correlation between the variables and the component. Because these are correlations, possible values range from -1 to +1. On the **/format** subcommand, we used the option **blank(.45)**, which tells SPSS not to print any of the correlations that are .45 or less. This makes the output easier to read by removing the clutter of low correlations that are probably not meaningful anyway.

Component – The columns under this heading are the rotated components that have been extracted. As we can see, three components were extracted (the three components that we requested). For example, the first component might be called "Official purposes" because items like "Trades work at home" and "Public transport" load highly on it. The second factor might be called "relating hangout" because items like "Tickets for events" and "Drinks at a bar" load highly on it. The third factor has to do with e-services.

COMPONENT TRANSFORMATION MATRIX

Component	1	2	3
1	.680	.511	.526
2	-.542	.834	-.109
3	-.494	-.211	.844
Extraction Method: Principal Component Analysis. **Rotation Method: Varimax with Kaiser Normalization.**			

This matrix displays the correlation matrix among the components prior to and after rotations. The rotated component matrix helps to clarify segregate the factors among three different components. As shown in the matrix, following three components can be formed.

Official purposes:

- Physical goods in a retail shop
- Trade works in the home
- Donation to charity
- Paying a friend
- Taxi/cab
- Public transport
- Goods in market

Leisure:

- Food in restaurant
- Drinks at a bar/coffee shop etc.
- Tickets for events
- Motor fuel

e-Services:

- Taxi/cab
- Goods over the phone
- Goods on web
- Utility bills (TV License, passport, etc.)

Here we define the factors as follows:

Factor 1: Official Purposes

The component Official purposes includes scenarios where a person might be using cashless payment systems for commercial/ business purposes. One of the main reasons for naming this component as such are the high loadings from "Trades work at home", "Physical goods at a retail shop" and "Goods in market". Other factors like "paying a friend" and "donation to charity" can be viewed as an official purpose or a personal endeavor but considering the context we take it as official purposes here. Lastly, we have factors like "Taxi/ cab" and "Public transport" which are commuting expenses to and from office or travelling expenses to client site.

Factor 2: Leisure

The component Leisure includes scenarios where a person might be using cashless payment systems for personal/ social outings or hangouts. It includes "Food in a restaurant" and "Drinks at a bar/ coffee shop" which are clearly luxuries reserved for outing with family/ friends during the free time. This component also includes "Tickets for events" like concerts, plays, entertainment shows, movies, etc. which are again social outings during leisure time. Lastly, the component includes "Motor Fuel" and it comes under leisure as the fuel is burned for personal purposes and not official purposes. The fuel consumed can be used to go for outings, leisure drives, etc.

Factor 3: e-Services

We extracted this 3rd component even though its Eigen value was less than 1 because of the distinction in scenarios with compared to the other 2 components above. As we named this component "e-Services", the reason for that is the scenarios under this component are pertaining to when we purchase/ spend money over electronic mediums like phone, web, mobile, etc. to buy goods or pay bills. We included "Taxi/ cab" in this component as well because in 2017, we have majority of urban respondents booking cabs via Uber, Ola and other such real-time applications which rely primarily on cashless modes of payment for their services.

CONCLUSION:

1. A principal component analysis was conducted on 15 scenarios with orthogonal rotation (Varimax).
2. The KMO value is 0.914 which is above the acceptable level of 0.5 hence we proceed factor analysis.
3. After analysis, 3 factors were extracted which explained combined of 59.57% total variance.
4. With this output we would be able to group all the variables under consideration in 3 factors as follows:
    a. Official Purposes
    b. Leisure
    c. E-Services

OBJECTIVE 3

To find the deciding factors leading up-to a person's preference for using cashless in daily transactions.

CART MODEL

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning.

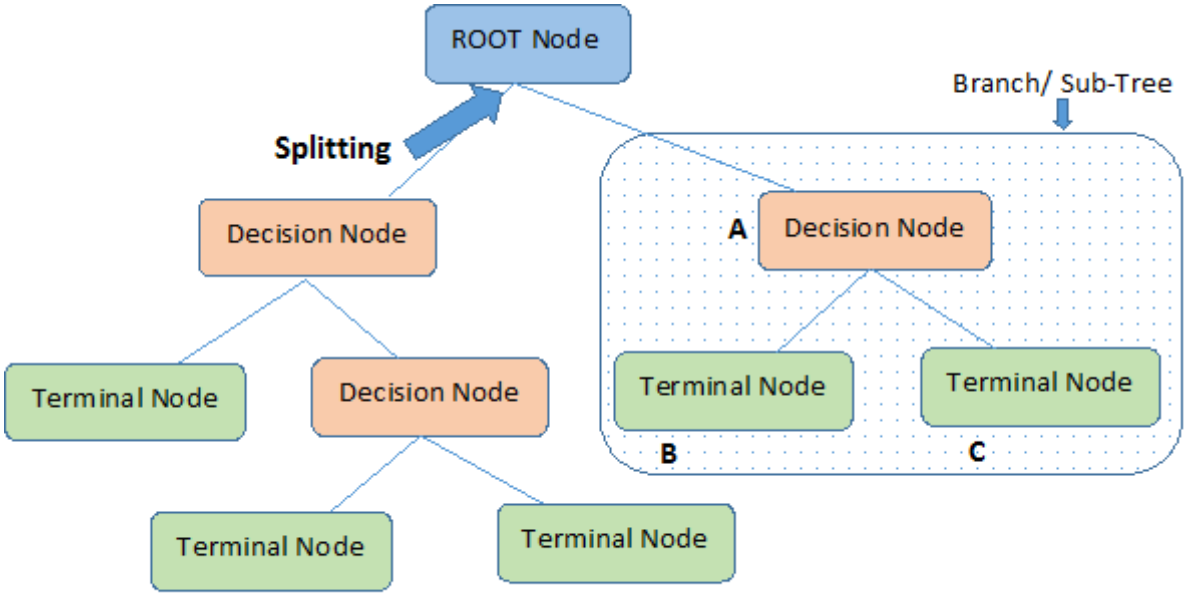
Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).

Types of decision tree is based on the type of target variable we have. It can be of two types:

1. **Categorical Variable Decision Tree:** Decision Tree which has categorical target variable then it called as categorical variable decision tree.
2. **Continuous Variable Decision Tree:** Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Let's look at the basic terminology used with Decision trees:

1. **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
6. **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.



Note:- A is parent node of B and C.

These are the terms commonly used for decision trees. As we know that every algorithm has advantages and disadvantages, below are the important factors which one should know.

- **Simple to understand and interpret.** People are able to understand decision tree models after a brief explanation. Trees can also be displayed graphically in a way that is easy for non-experts to interpret.
- **Able to handle both numerical and categorical data.** Other techniques are usually specialised in analysing datasets that have only one type of variable. (For example, relation rules can be used only with nominal variables while neural networks can be used only with numerical variables or categorical converted to 0-1 values.)
- **Requires little data preparation.** Other techniques often require data normalization. Since trees can handle qualitative predictors, there is no need to create dummy variables.
- **Uses a white box model.** If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model, the explanation for the results is typically difficult to understand, for example with an artificial neural network.
- **Possible to validate a model using statistical tests.** That makes it possible to account for the reliability of the model.
- **Non-statistical approach that makes no assumptions of the training data or prediction residuals**; e.g., no distributional, independence, or constant variance assumptions
- **Performs well with large datasets.** Large amounts of data can be analysed using standard computing resources in reasonable time.

- **Mirrors human decision making more closely than other approaches.** This could be useful when modelling human decisions/behaviour.
- **Robust against co-linearity, particularly boosting**.

Disadvantages:

- **Over fitting:** Over fitting is one of the most practical difficulty for decision tree models. This problem gets solved by setting constraints on model parameters and pruning (discussed in detailed below).
- **Not fit for continuous variables**: While working with continuous numerical variables, decision tree loses information when it categorizes variables in different categories.

We all know that the terminal nodes (or leaves) lies at the bottom of the decision tree. This means that decision trees are typically drawn upside down such that leaves are the bottom & roots are the tops (shown below).

Both the trees work almost similar to each other, let's look at the primary differences & similarity between classification and regression trees:

1. Regression trees are used when dependent variable is continuous. Classification trees are used when dependent variable is categorical.
2. In case of regression tree, the value obtained by terminal nodes in the training data is the mean response of observation falling in that region.

Thus, if an unseen data observation falls in that region, we'll make its prediction with mean value.

3. In case of classification tree, the value (class) obtained by terminal node in the training data is the mode of observations falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mode value.
4. Both the trees divide the predictor space (independent variables) into distinct and non-overlapping regions. For the sake of simplicity, you can think of these regions as high dimensional boxes or boxes.
5. Both the trees follow a top-down greedy approach known as recursive binary splitting. We call it as 'top-down' because it begins from the top of tree when all the observations are available in a single region and successively splits the predictor space into two new branches down the tree. It is known as 'greedy' because, the algorithm cares (looks for best variable available) about only the current split, and not about future splits which will lead to a better tree.
6. This splitting process is continued until a user defined stopping criteria is reached. For example: we can tell the algorithm to stop once the number of observations per node becomes less than 50.
7. In both the cases, the splitting process results in fully grown trees until the stopping criteria is reached. But, the fully grown tree is likely to over-fit data, leading to poor accuracy on unseen data. This bring 'pruning'. Pruning is one of the technique used tackle overfitting. We'll learn more about it in following section.

The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria are different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

The algorithm selection is also based on type of target variables. Let's look at the four most commonly used algorithms in decision tree:

Gini Index

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

1. It works with categorical target variable "Success" or "Failure".
2. It performs only Binary splits
3. Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses Gini method to create binary splits.

Steps to Calculate Gini for a split

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split

ANALYSIS

After performing binary logistic regression on our socio-demographic variables for the purpose of identifying the significant variables that affect people's decision to use cashless payment systems, we extracted 9 prominent variables with significant effect:

1. Age
2. Gender
3. Occupation
4. House
5. Income
6. Phone
7. No. of Account
8. Private
9. Public

We will use these variables further with few more new variables to identify the strongest variables that affect the decision of people to use cashless payment systems for their daily transactions.

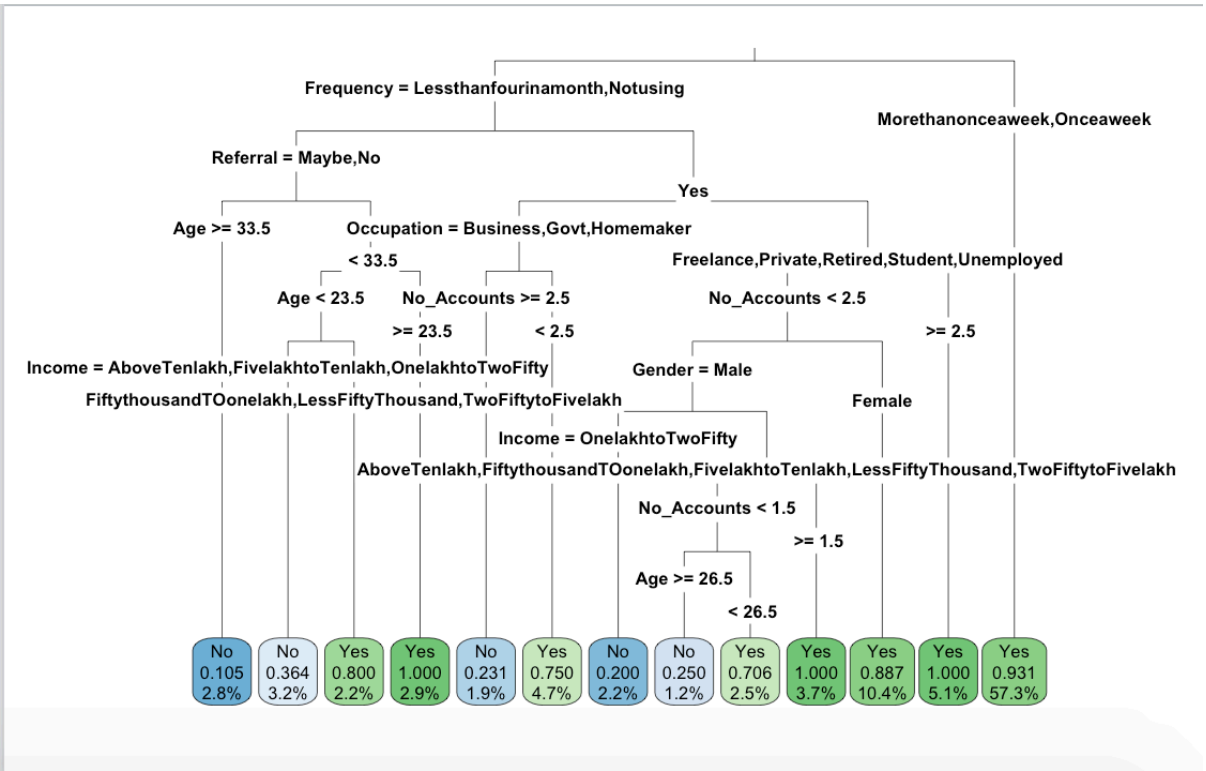
The new variables that we will take into account are:

10. Demonetization
11. Referral
12. Online Shopping
13. Frequency of using Cashless payments
14. Average amount spent using cashless payments

We will use CART Model – specifically Classification trees to come up with the best possible model to explain people's decision. We will R Software to come up with this model.

```
#Install the necessary packages to perform the
analysis
install.packages("rpart")
install.packages("rpart.plot")
```

```r
#Load the required packages
library(rpart)
library(rpart.plot)

#Load the dataset
data(DailyCash)

#The structure of the dataset
str(DailyCash)

#Set a common seed to verify the results later
set.seed(12345)

#Split the dataset into training and testing data by
the dependent variable using 75:25 ratio to test the
accuracy of the model later. Training data set
contains 684 observations and testing data set
contains 228 observations.
spl = sample.split(DailyCash$DailyCashless,
SplitRatio = 0.75)
Train = subset(DailyCash, spl==TRUE)
Test = subset(DailyCash, spl==FALSE)

#Create the classification tree on our training
dataset and plot the tree.
m1 <- rpart(DailyCashless ~ ., data=Train)
rpart.plot(m1, type = 3, digits = 3, fallen.leaves =
TRUE)
prp(m1)
```



As we can see from the above tree, there are 7 variables removed from the Full-Model tree: Public, Phone, House, Demonetization, Private, Average amount spent, Online Shopping in cashless payments.

This tree takes into account all variables and observations with no conditions applied to its splitting and branches.

So, let's test its accuracy on the testing dataset.

```
#Predicting the observations in testing data set
using the tree created by the training data set.
p1 <- predict(m1, DailyCash)
PredictCASH = predict(m1, newdata = Test, type =
"class")

#Create a classification table to test the accuracy.
table(Test$DailyCashless, PredictCASH)
```

Actual Values	**Predicted Values**		
		No	**Yes**
	No	17	19
	Yes	8	184

The accuracy of the model is:

$$\frac{184 + 17}{228} = 88.16\%$$

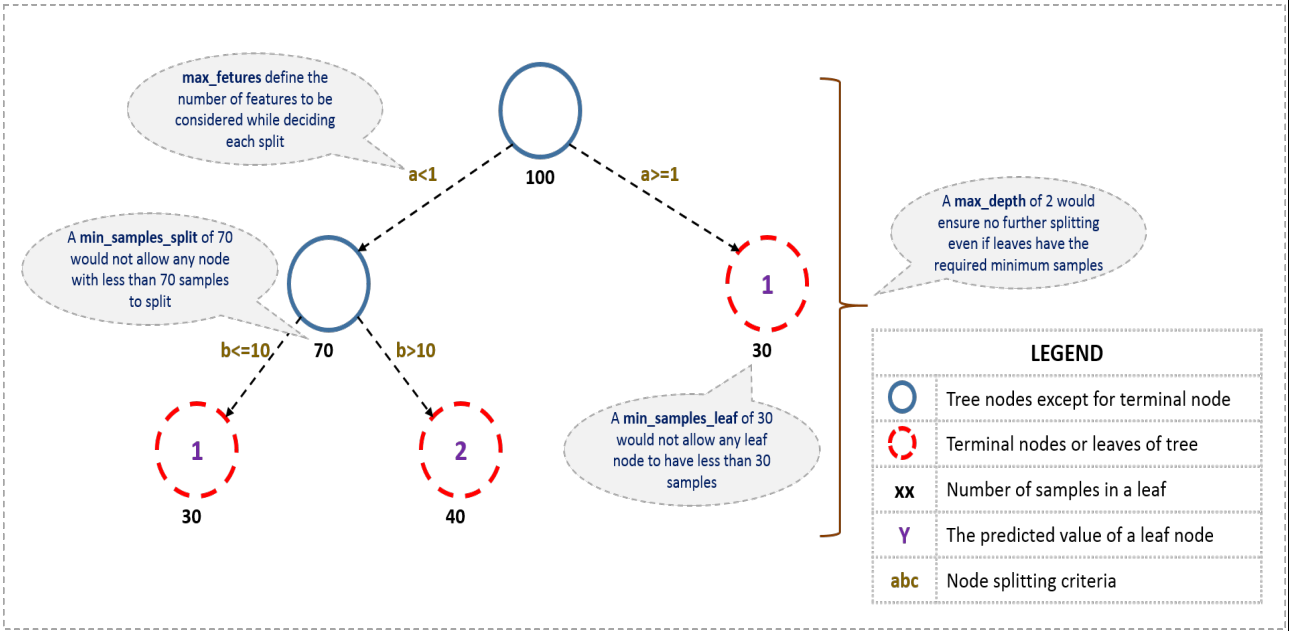
Since, this is the full-model, the tree suffers from over-fitting making the tree seem complicated with one of the terminal nodes accounting for only 1.9% or about 13 of the total 684 observations.

Overfitting is one of the key challenges faced while modelling decision trees. If there is no limit set for a decision tree, it will give you 100% accuracy on training set because in the worst case, it will end up making 1 leaf for each observation. Thus, preventing overfitting is pivotal while modelling a decision tree and it can be done in 2 ways:

1. Setting constraints on tree size
2. Tree pruning

SETTING CONSTRAINTS ON TREE SIZE

This can be done by using various parameters which are used to define a tree. First, let's look at the general structure of a decision tree:



The parameters used for defining a tree are further explained below. The parameters described below are irrespective of tool. It is important to understand the role of parameters used in tree modeling.

1. **Minimum samples for a node split**
    - Defines the minimum number of samples (or observations) which are required in a node to be considered for splitting.

- Used to control over-fitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample selected for a tree.
- Too high values can lead to under-fitting hence, it should be tuned using CV.

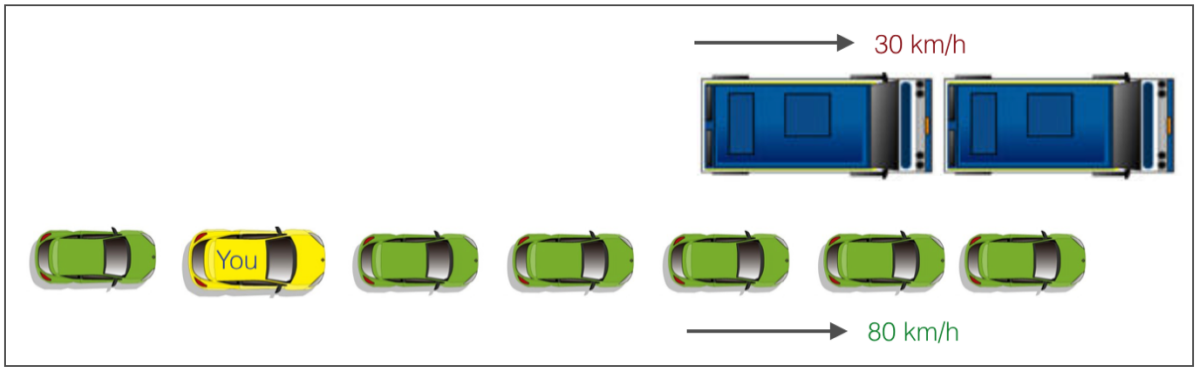
2. **Minimum samples for a terminal node (leaf)**
   - Defines the minimum samples (or observations) required in a terminal node or leaf.
   - Used to control over-fitting similar to min_samples_split.
   - Generally lower values should be chosen for imbalanced class problems because the regions in which the minority class will be in majority will be very small.
3. **Maximum depth of tree (vertical depth)**
   - The maximum depth of a tree.
   - Used to control over-fitting as higher depth will allow model to learn relations very specific to a particular sample.
   - Should be tuned using CV.
4. **Maximum number of terminal nodes**
   - The maximum number of terminal nodes or leaves in a tree.
   - Can be defined in place of max_depth. Since binary trees are created, a depth of 'n' would produce a maximum of 2^n leaves.
5. **Maximum features to consider for split**
   - The number of features to consider while searching for a best split. These will be randomly selected.
   - As a thumb-rule, square root of the total number of features works great but we should check upto 30-40% of the total number of features.
   - Higher values can lead to over-fitting but depends on case to case.

TREE PRUNING

As discussed earlier, the technique of setting constraint is a greedy-approach. In other words, it will check for the best split instantaneously and move forward until one of the specified stopping condition is reached. Let's consider the following case when you're driving:

There are 2 lanes:

1. A lane with cars moving at 80km/h
2. A lane with trucks moving at 30km/h



At this instant, you are the yellow car and you have 2 choices:

1. Take a left and overtake the other 2 cars quickly
2. Keep moving in the present lane

Let's analyze these choices. In the former choice, you'll immediately overtake the car ahead and reach behind the truck and start moving at 30 km/h, looking for an opportunity to move back right. All cars originally behind you move ahead in the meanwhile. This would be the optimum choice if your objective is to maximize the distance covered in next say 10 seconds. In the later choice, you sail through at the same speed, cross trucks and then overtake maybe depending on situation ahead. Greedy you!

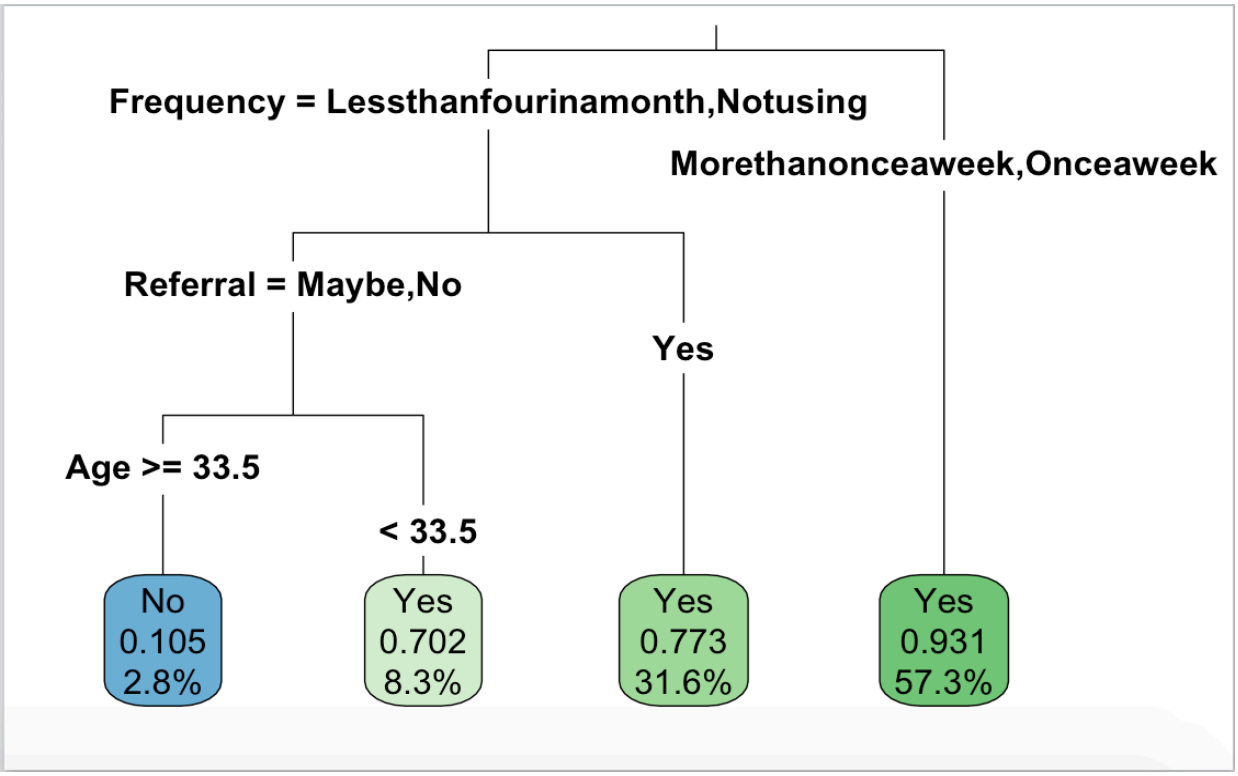
This is exactly the difference between normal decision tree & pruning. A decision tree with constraints won't see the truck ahead and adopt a greedy approach by taking a left. On the other hand, if we use pruning, we in effect look at a few steps ahead and make a choice.
So we know pruning is better. But how to implement it in decision tree? The idea is simple.

1. We first make the decision tree to a large depth.
2. Then we start at the bottom and start removing leaves which are giving us negative returns when compared from the top.
3. Suppose a split is giving us a gain of say -10 (loss of 10) and then the next split on that gives us a gain of 20. A simple decision tree will stop at step 1 but in pruning, we will see that the overall gain is +10 and keep both leaves.

MINSPLIT = 50

So, in order prevent overfitting and prune the tree, we tried a minimum split of 50 to see the newer decision tree and check its accuracy.

```
set.seed(12345)
m1 <- rpart(DailyCashless ~ ., minsplit=50,
data=Train)
rpart.plot(m1, type = 3, digits = 3, fallen.leaves =
TRUE)
```



As we can see from the tree above, it only takes into account 3 variables – Frequency, Age and Referral which is much lesser than 7 variables from the full-model.

So, let's test its accuracy on the testing dataset.

```
#Predicting the observations in testing data set
using the tree created by the training data set.
p1 <- predict(m1, DailyCash)
PredictCASH = predict(m1, newdata = Test, type =
"class")

#Create a classification table to test the accuracy.
```

```
table(Test$DailyCashless, PredictCASH)
```

Actual Values	Predicted Values		
		No	**Yes**
	No	7	29
	Yes	4	188

The accuracy of the model is:

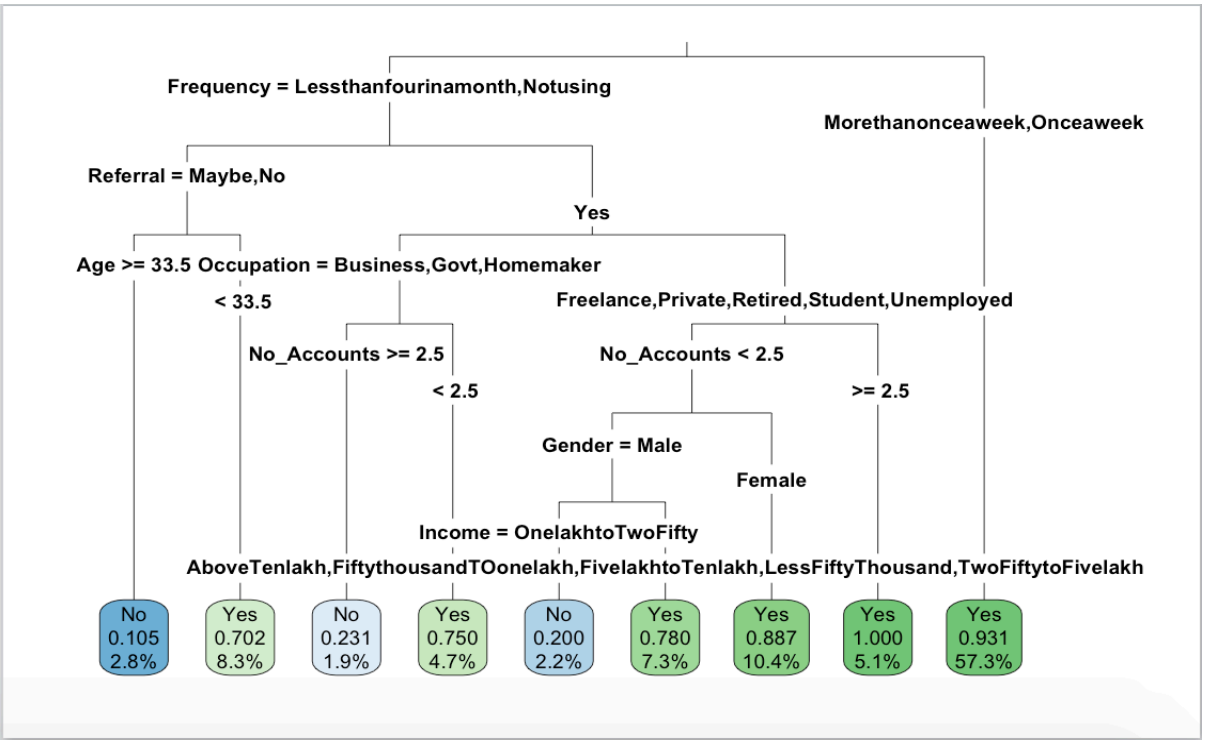
$$\frac{188 + 7}{228} = 85.52\%$$

So, as we can see this actually reduces our accuracy but makes the tree more simpler and efficient.

MINSPLIT = 40

Next, we tried a minimum split of 40 to see the newer decision tree and check its accuracy.

```
set.seed(12345)
m1 <- rpart(DailyCashless ~ ., minsplit=40,
data=Train)
rpart.plot(m1, type = 3, digits = 3, fallen.leaves =
TRUE)
```



As we can see from the tree above, it takes into account 7 variables – Frequency, Income, Occupation, Gender, Referral, Age and No. of accounts which are the same 7 variables from the full-model and more than 3 variables from the Minsplit=50 model but instead of 13 terminal nodes in the full model, it has reduced to 9 terminal nodes.

So, let's test its accuracy on the testing dataset.

```
#Predicting the observations in testing data set
using the tree created by the training data set.
p1 <- predict(m1, DailyCash)
PredictCASH = predict(m1, newdata = Test, type =
"class")

#Create a classification table to test the accuracy.
table(Test$DailyCashless, PredictCASH)
```

Actual Values	**Predicted Values**		
		No	**Yes**
	No	12	24
	Yes	5	187

The accuracy of the model is:

$$\frac{187 + 12}{228} = 87.28\%$$

So, we can see that the accuracy has reduced a smidge by 0.88% from the full model but we have reduced considerable no. of nodes in our decision tree. The lowest number of observations in a terminal node are 1.9% ~ 13 observations with the highest still being the same 57.3% ~ 391 as it has been in the earlier to variants of the classification tree.

CONCLUSION

We can conclude from the 3 variants of decision tree that minsplit=40 would be optimal for our analysis purpose that prevents overfitting as well as prevents the tree from being too simplistic in nature.

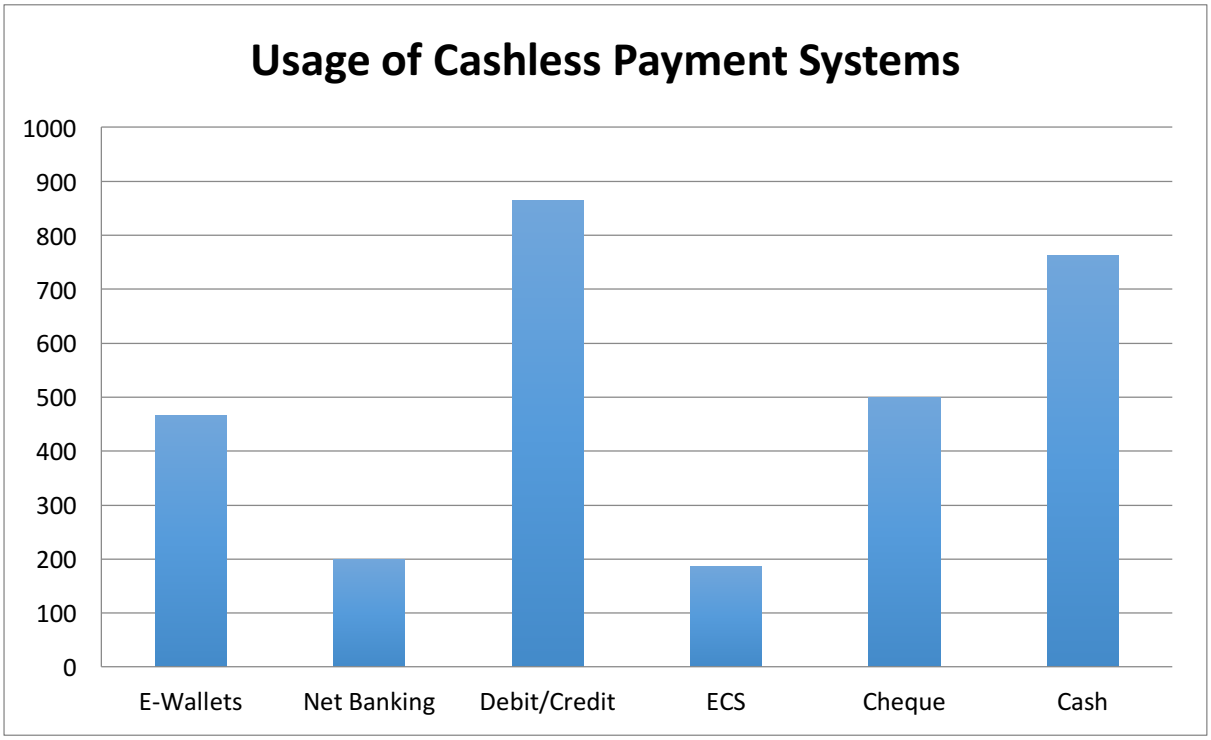
It has been repeatedly observed that Frequency, Referral and Age are the most significant variables that affect people's decision on whether they prefer cashless payment systems for daily transactions.

Later we see that new variables are added like Occupation, No. of Accounts, Gender and Income. This is to be expected as this were some of the most significant variables in the logistic regression model and now they have similar effect here as well.

OBJECTIVE 4

To find out the most preferred cashless mode of payment/ online wallet and the factors behind its popularity.

BAR CHART



We asked the respondents of the number of different payment systems they use and the above bar chart represents the number of votes each payment method received. As we can clearly notice, Cash and debit/ credit card payments are top 2 most preferred payment mechanisms in India.

Since the survey was taken in the post-demonetization period, it is to be expected that people are relying more on cards over cash as cash was in short supply and there was general fear among the public to keep cash reserves at the minimum[12].

RBI data at End-March shows 24.5 million credit cards and 661.8 million debit cards in the country, compared to 1.3 million PoS terminals[3].

Additionally, the growth in the acceptance infrastructure has not been uniform across all locations in the country with higher concentration of such infrastructure noted in urban areas and larger towns and with larger merchants. Thus, the usage

[1] http://economictimes.indiatimes.com/news/economy/finance/demonetisation-from-cash-to-less-cash-to-cashless/articleshow/55674538.cms
[2] http://indianexpress.com/article/india/india-others/black-money-bill-in-lok-sabha10-years-jail-for-concealing-foreign-funds/
[3] http://www.business-standard.com/article/economy-policy/new-committee-swipes-towards-cashless-society-116052601108_1.html

of cards has been constrained by lack of accessible acceptance infrastructure, especially in rural areas where the growth in card issuance has been very high in recent times[4].

Since, our survey respondents predominantly hail from urban areas, it is quite likely they have higher usage of card payments compared to people from rural areas with less access to necessary payment infrastructure like POS machines and ATMs.

On the other hand, 68% of transactions across India are cash-based[5]. The number of electronic card/automated clearing house (ACH) transactions also increased from 2.6% in FY07 to 6.8% in FY12. But transactions through other modes — cheques, demand drafts, etc. — decreased from 4.1% in FY07 to 2.5% in FY12[6].

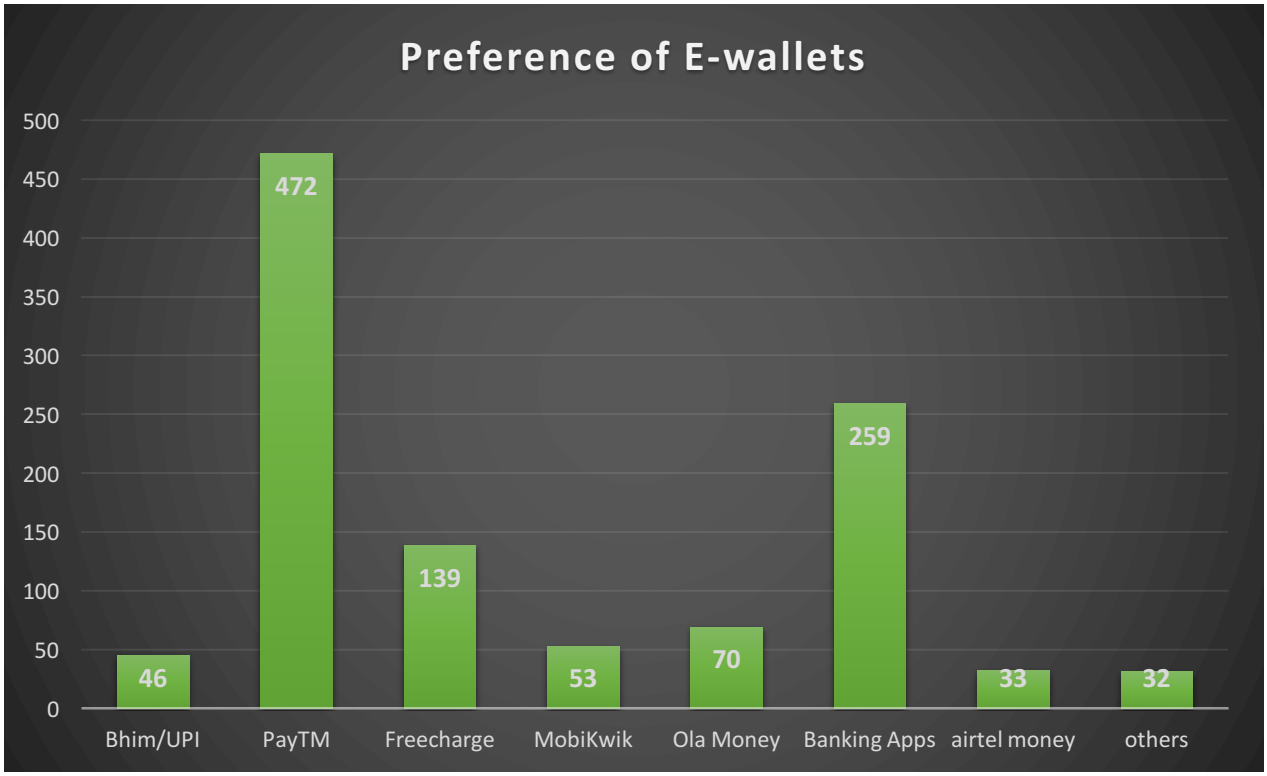
Since demonetization, there has been a significant push toward digital transactions by Prime Minister Mr. Narendra Modi and the NDA Central Govt. Administration[7].

[4] https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/MDRDBEDA36AB77C4C81A3951C4679DAE68F.PDF - Concept Paper on Card Acceptance Infrastructure

[5] http://www.business-standard.com/article/economy-policy/infographic-68-of-transactions-in-india-are-cash-based-116111400495_1.html

[6] http://www.business-standard.com/article/economy-policy/demonetisaton-cash-is-still-king-in-india-banking-penetration-improves-116112400575_1.html

[7] http://economictimes.indiatimes.com/news/politics-and-nation/narendra-modi-calls-for-making-india-a-cashless-society-in-mann-ki-baat/articleshow/52386384.cms



Mobile wallets are the next-gen payment systems and their usage has been increasing steadily and very much exponentially in the past few months since demonetization with the biggest beneficiaries being Paytm[8]. Paytm's traffic increased by 435%, app downloads grew 200%, and there was 250% rise in overall transactions and transaction value[9].

India has launched a new payments system called Unified Payment Interface, or UPI, which is designed to make person-to-person and e-commerce transactions swifter and easier. Doing transactions, the government says, will be as easier and as faster as sending a text message.

Unveiled earlier this year by the National Payments Corporation of India (NPCI), the primary body that governs all retail payments in the country, UPI aims to propel the economy towards more cashless transactions. This would, among other things, help the government curb unreported money exchanges that aren't subjected to tax. The number of non-cash transactions per person in India is only six per year.

The government has opted for smartphone as the computing platform for UPI.

[8] http://www.hindustantimes.com/business-news/mobile-wallets-see-a-soaring-growth-post-demonetisation/story-zwdBi3UGqG1qZD92AEF9GK.html

[9] http://www.businessinsider.in/Thanks-to-demonetization-Paytm-is-making-Rs-120-crore-per-day-achieves-target-before-deadline-crosses-5-billion-GMV/articleshow/55541691.cms

India is the world's fastest growing smartphone market. The country has over 350 million smartphone users, and the number is projected to grow past 700 million by 2020. By making digital payments available in full-fledged capacity on smartphones, the country is diminishing the barrier point and making it easier for more users to join the bandwagon.

The other integrals of UPI are inter-operability — allowing transactions across banks — and need for a single-identifier to make transactions. It is built on top of IMPS protocol, a 24x7 and real-time transaction service. But unlike IMPS, which requires the concerned parties to provide a fair amount of bank details such as account number and IFSC code, UPI-enabled apps will only require the knowledge of a virtual address. For this, it leverages on India's biometrics-enabled national ID system, *Aadhaar*.

This virtual address is a 12-digit number which is unique to a person based on their fingerprint, iris, and facial features. Not only does utilizing *Aadhaar* makes the whole process more secure, it also gives the government a way to target a wider consumer base. Over a billion people now have an *Aadhaar* card[10].

Lots of banks have their own banking apps linked with UPI that have the 2nd highest preference after Paytm and this is due to various reasons like security, convenience and user-friendly which we will see in more detail in further analysis.

[10] http://mashable.com/2016/08/30/india-upi-payments-system/#oIMXIBh59SqZ

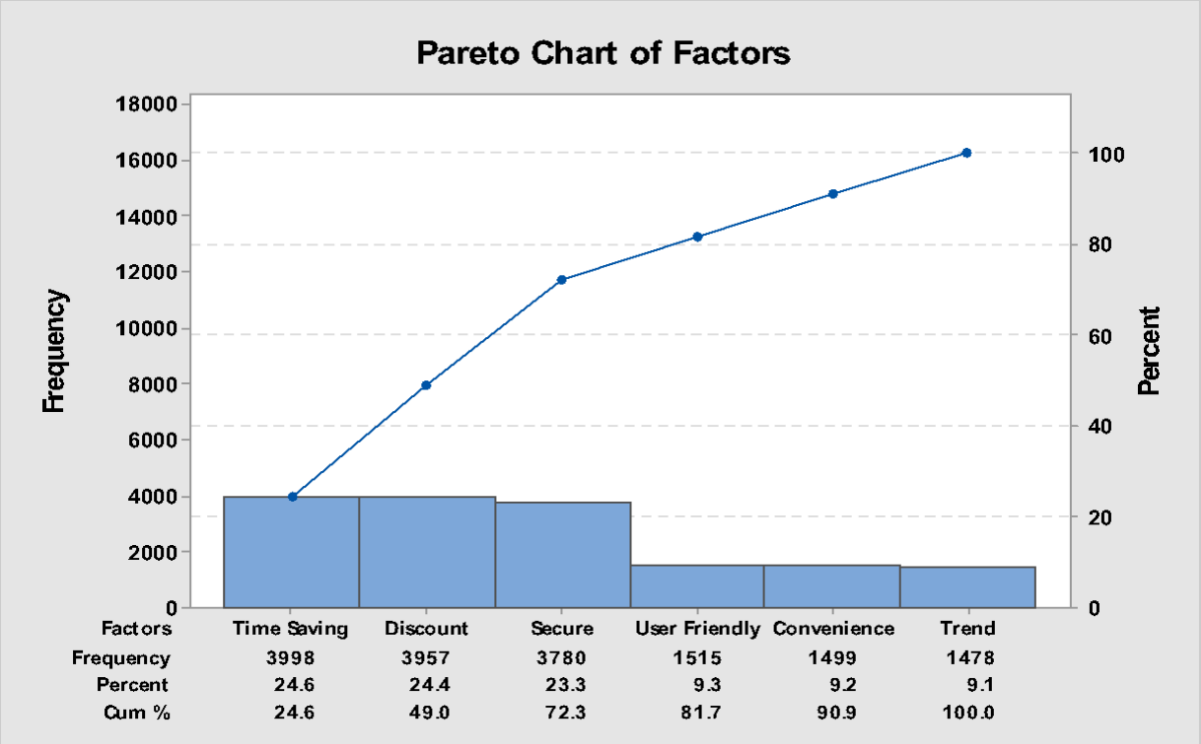
PARETO ANALYSIS

A Pareto chart is used to graphically summarize and display the relative importance of the differences between groups of the data. The Pareto Chart is a very simple but effective tool for prioritizing problem causes.

Pareto analysis aims at highlighting those elements which demand attention and should be examined first. It conveys that, tackle the 'Vital Few' and ignore the 'Trivial Many'.

It is a combination of bar graph and line graph that puts items in order (from the highest to the lowest) relative to some measurable effect of interest such as frequency, cost or time. The Pareto principle describes a phenomenon in which 80% of variation observed can be explained by a mere 20% of the causes of that variation. The line graph indicates the cumulative percentage of occurrence at each bar of the bar graph. This line graph, referred to as a 'cumulative percentage line', is used to determine which of the bars belong to the 'vital few' and which ones are relegated to the 'trivial many'.

The Pareto curve makes it clear as to where effort must be concentrated so as to give maximum effect.



Factors	Time Saving	Discount	Secure	User Friendly	Convenience	Trend
Frequency	3998	3957	3780	1515	1499	1478
Percent	24.6	24.4	23.3	9.3	9.2	9.1
Cum %	24.6	49.0	72.3	81.7	90.9	100.0

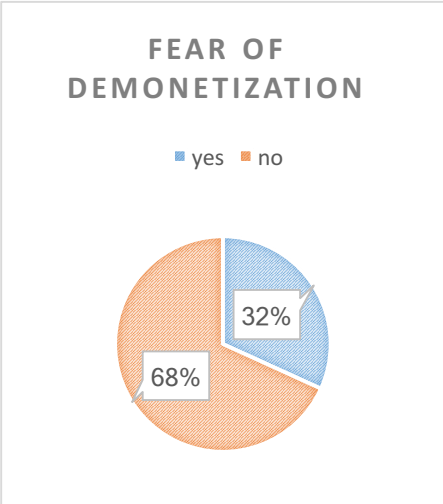
We did a Pareto chart on the factors that make people prefer cashless payment systems. We had six factor options given to the respondents: time saving, discount, secure, user friendly, convenience and trend[11].

As we can from the Pareto chart of the responses we have received, close to 81.7% of responses concentrate on just 4 factors out of the total six. The three most important factors are time saving, discount, secure and User-friendly.

It is well known that cashless payments help save time both for the user as well as the government in transporting the money. You will no longer need to carry wads of cash, plastic cards, or even queue up for ATM withdrawals. It's also a safer and easier spending option when you are travelling[12].

The recent waiver of service tax on card transactions up to Rs 2,000 is one of the incentives provided by the government to promote digital transactions. This has been followed by a series of cuts and freebies. There are always cashback offers and discounts offered by mobile wallets like Paytm, as well as the reward points and loyalty benefits on existing credit and store cards, and this is why it's one of the prominent reasons in people's preference in using cashless payment systems.

Cashless payment systems are more secure than ever - If stolen, it is easy to block a credit card or mobile wallet remotely, but it's impossible to get your cash back. This is especially true while travelling, especially abroad, where loss of cash can cause great inconvenience. Besides, if the futuristic cards evolve to use biometric ID (finger prints, eye scan, etc.), it can be extremely difficult to copy, making it a very safe option.



We also asked our respondents whether demonetization played a huge role in their decision to go cashless and 32% (almost 292) respondents said no. So, we can say that digital payments are the future and are being adopted increasingly with more acceptance and not fear.

[11] http://economictimes.indiatimes.com/wealth/spend/ready-to-go-cashless/articleshow/56269830.cms

[12] http://economictimes.indiatimes.com/articleshow/55908649.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

CART MODEL

We again applied CART model to find out the preference of our respondents between banking apps and private e-wallets and the reasons behind such a preference.

As we noticed in the chart above, Paytm and banking apps were the 2 highest used digital payment platforms used by the respondents.

Nearly two dozen Indian banks including ICICI Bank, Canara Bank, Andhra Bank have announced they will be releasing a UPI-enabled app on Google Play Store. ICICI Bank announced that the feature is live on its 'iMobile' and 'Pockets' mobile apps. The latter allows peer-to-peer and e-commerce payments by users including those who don't have an ICICI Bank account. It is the first time in the world that a project of such a scale is being introduced to customers.

Several other major banks in the country such as SBI, HDFC, and Kotak have announced that they will be releasing a UPI-enabled app soon. Any bank with over 1,000 pilot customers, 5,000 transactions, and an 80 percent higher success rate is eligible to adopt UPI on their app and offer it to the general public through the Google Play.

Once they become available, you're required to download a UPI-enabled app, set a PIN code, create a virtual address and link it to any bank account. Once you have set it up, you only need to know the receiver's unique ID. When you have it, you open the app, select the amount, enter the unique ID, and select "send." The app will ask you for an authentication ID, which you will receive on your phone. After which, regardless of the bank in which the receiver has her account with, the money will go through instantly.

UPI offers a considerably wide payment range — between Rs. 50 to Rs. 100,000 in one transaction. The payments system is designed to serve as a replacement for all the apps that you needed to make money transactions on online shopping websites, pay electricity bills, barcode-based payments, and deposit college tuition[13].

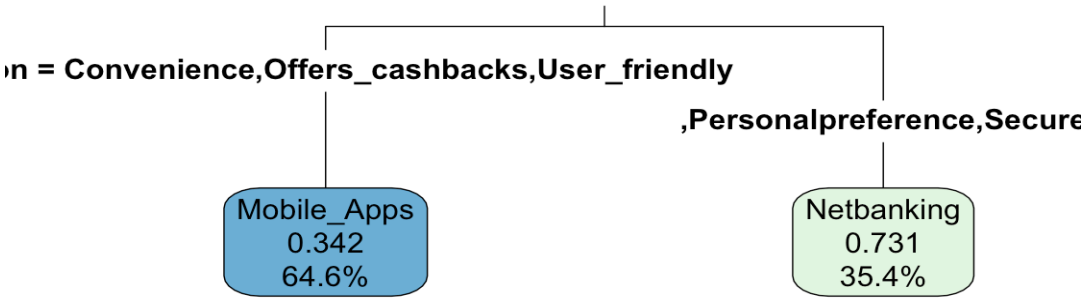
So, in order to find out the reasons for such a choice, we gave them an open-ended question and coded the data in five categories: Security, Convenience, Offers/ Cashbacks, User-friendly and lastly, personal preference[14].

[13] http://mashable.com/2016/08/30/india-upi-payments-system/#oIMXIBh59SqZ

[14] http://economictimes.indiatimes.com/articleshow/55908649.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

Here is the code for how we went about with our analysis:

```
#Load the required packages
library(caTools)
library(rpart)
library(rpart.plot)

#Set a common seed to verify the results later
set.seed(3000)

#Split the dataset into training and testing data by
the dependent variable using 75:25 ratio to test the
accuracy of the model later. Training data set contains
684 observations and testing data set contains 228
observations.
spl = sample.split(Prefer$Preference, SplitRatio =
0.75)
Train = subset(Prefer, spl==TRUE)
Test = subset(Prefer, spl==FALSE)

#Create the classification tree on our training dataset
and plot the tree.
PreferTree = rpart(Preference ~ ., data = Train,
method="class", minbucket=25)
prp(PreferTree)
rpart.plot(PreferTree, type = 3, digits = 3,
fallen.leaves = TRUE)
```



```
#Predicting the observations in testing data set using
the tree created by the training data set.
PredictCART = predict(PreferTree, newdata = Test, type
= "class")

#Create a classification table to test the accuracy.
table(Test$Preference, PredictCART)
```

Actual Values	Predicted Values		
		Mobile Apps	**Banking Apps**
	Mobile Apps	94	25
	Banking Apps	43	66

The accuracy of the model is:

$$\frac{94 + 66}{228} = 70.18\%$$

So, as we can the model is pretty accurate at 70.18% where Convenience, Offers/ Cashbacks, User-friendly lead to Mobiles Apps/E-wallets while reasons like Personal preference and security lead them to banking apps.

To further try to improve the accuracy of our model, we tried the random forest method.

RANDOM FOREST MODEL

Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

In Random Forest, we grow multiple trees unlike the CART model, where we use a single tree. To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

It works in the following manner. Each tree is planted & grown as follows:

1. Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.
3. Each tree is grown to the largest extent possible and there is no pruning.
4. Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

Advantages of Random Forest:

- This algorithm can solve both type of problems i.e. classification and regression and does a decent estimation at both fronts.
- One of benefits of Random forest which excites me most is, the power of handle large data set with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods. Further, the model outputs **Importance of variable,** which can be a very handy feature (on some random data set).
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.
- It has methods for balancing errors in data sets where classes are imbalanced.
- The capabilities of the above can be extended to unlabelled data, leading to unsupervised clustering, data views and outlier detection.

- Random Forest involves sampling of the input data with replacement called as bootstrap sampling. Here one third of the data is not used for training and can be used to testing. These are called the **out of bag** samples. Error estimated on these out of bag samples is known as *out of bag error*. Study of error estimates by Out of bag, gives evidence to show that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. Therefore, using the out-of-bag error estimate removes the need for a set aside test set.

Disadvantages of Random Forest:

- It surely does a good job at classification but not as good as for regression problem as it does not give precise continuous nature predictions. In case of regression, it doesn't predict beyond the range in the training data, and that they may over-fit data sets that are particularly noisy.
- Random Forest can feel like a black box approach for statistical modellers – you have very little control on what the model does. You can at best – try different parameters and random seeds!

Here is the code for how we went about with our analysis:

```
# Install randomForest package
install.packages("randomForest")
library(randomForest)

# Build random forest model
PreferForest = randomForest(Preference ~ ., data =
Train, ntree=200, nodesize=25 )

# Convert outcome to factor
Train$Preference = as.factor(Train$Preference)
Test$Preference = as.factor(Test$Preference)
```

The Confusion matrix we got at this point is as follows:

Actual Values	**Predicted Values**	
	Mobile Apps	**Banking Apps**
Mobile Apps	291	65
Banking Apps	151	177

The accuracy of the model is:

$$\frac{291 + 177}{684} = 68.42\%$$

This is lesser than the accuracy we got from the CART model, but let's validate the model on our testing dataset.

```
# Install the required packages
install.packages("party")
library("party")

# Make predictions
PredictForest = predict(PreferForest, newdata = Test)
table(Test$Preference, PredictForest)
```

The Confusion matrix we got at this point is as follows:

Actual Values	**Predicted Values**	
	Mobile Apps	**Banking Apps**
Mobile Apps	94	25
Banking Apps	43	66

The accuracy of the model is:

$$\frac{94 + 66}{228} = 70.18\%$$

This is exactly the same matrix we got in our CART model. So, we can conclude that the CART model we initially made is the best fit model for us and it has been further confirmed by our random forest model analysis.

CONCLUSION

To analyze the cause for the use of Banking Apps and E-wallets, we ran cart model. After running the cart model, the results clearly showcased that out of our 912 people who took our survey there were around 35% users of banking applications and 65% were users of private e-wallets applications. The major reasons which were identified by us through our model for the use banking apps were security and personal preference while it was user-friendliness, convenience and offers/ cashbacks for e-wallet apps.

As cashless transactions are not so popular in India compared to other countries, people are always skeptical regarding the use of applications on digital and electronic mediums but due to rise in awareness and government initiatives there have been people especially in urban India, who are now using these applications increasingly.

But our survey clearly showed that users of e-wallets are much higher than the users of net banking apps. As the users of cashless payment systems increase so the number of e-wallets companies and to attract more users they go on a large scale advertisement campaigns and they also have attractive offers for the users which automatically attract more number of customers. They also tend to be more convenient and user friendly as compared too many of the banking applications as stated by our surveyors.

OBJECTIVE 5

To find out the major concerns with current cashless payment systems among the people who use cashless and those who don't.

WORD CLOUD

A word cloud is a graphical representation of frequently used words in a collection of text files. The height of each word in a word cloud is an indication of frequency of occurrence of the word in the entire text. Such diagrams are very useful when doing text analytics.

Text mining methods allow us to highlight the most frequently used keywords in a paragraph of texts. One can create a **word cloud**, also referred as *text cloud* or *tag cloud*, which is a visual representation of text data.

We used word cloud technique on our two open-ended questions we had asked our respondents. The questions were similar in nature but were asked to 2 different sets of respondents: The ones who use cashless payments systems and the ones who don't. So, we asked them the first set of respondents what were their major concerns with cashless payment systems despite using cashless payment systems so we received 912 responses for this particular question.

Secondly, we asked the people who don't use cashless payment systems as to why they don't and we received 202 responses for this particular questions since majority of our respondents were familiar and users of cashless payment systems so the data size is small it gives an idea whether both of the respondents share similar concerns and whether they can be worked upon to bring satisfaction to both users and non-users to adopt and keep using cashless payment systems in the future.

USERS

Here is how we went about with our analysis:

```
#Create a text file and load the text file
View(Opendata)
Opendata.text <- sapply(Opendata, function(x)
x$getText())

#DATA CLEANING
#convert all text to lower case
Opendata <- tolower(Opendata)

# Remove punctuation
Opendata <- gsub("[[:punct:]]", "", Opendata)
```

```
# Remove links
Opendata <- gsub("http\\w+", "", Opendata)

# Remove tabs
Opendata <- gsub("[ |\t]{2,}", "", Opendata)

# Remove blank spaces at the beginning
Opendata <- gsub("^ ", "", Opendata)

# Remove blank spaces at the end
Opendata <- gsub(" $", "", Opendata)
Opendata <- sapply(Opendata,function(row) iconv(row,
"latin1", "ASCII", sub=""))

#Install and load required packages
install.packages("tm")
library("tm")

#create corpus
Opendata.corpus <- Corpus(VectorSource(Opendata))

#clean up by removing stop words
Opendata.corpus <- tm_map(Opendata.corpus,
function(x)removeWords(x,stopwords()))

#install wordcloud if not already installed
install.packages("wordcloud")
library("wordcloud")

#generate wordcloud
wordcloud(Opendata.corpus, min.freq = 2,
scale=c(2.5,0.5),colors=brewer.pal(8, "Dark2"),
random.color= FALSE, random.order = FALSE, max.words
= 50)
```

This is result we received:

As we can see from the above word cloud, lots of words have similar meanings and are ever repeated with singular and plural verbs.

Looking through all the words, we came up with following coding for our open-ended data:

1 = Security (Fraud, Hacking, Risk, Safety, Secure)
2 = Time Consuming (Time)
3 = Distribution problem (availability, acceptance, vendors)
4 = Technical Issues (Internet, issues)
5 = MDR Charges (Charges)
6 = No Concerns (No, Nothing)

NON-USERS

Here is how we went about with our analysis:

```
#Create a text file and load the text file
View(LastQ)

#convert all text to lower case
LastQ <- tolower(LastQ)

# Remove punctuation
LastQ <- gsub("[[:punct:]]", "", LastQ)
```

```
# Remove links
LastQ <- gsub("http\\w+", "", LastQ)

# Remove tabs
LastQ <- gsub("[ |\t]{2,}", "", LastQ)

# Remove blank spaces at the beginning
LastQ <- gsub("^ ", "", LastQ)

# Remove blank spaces at the end
LastQ <- gsub(" $", "", LastQ)
lastQ <- sapply(LastQ,function(row) iconv(row,
"latin1", "ASCII", sub=""))

#Install and load required packages
install.packages("tm")
library("tm")

#create corpus
LastQ.corpus <- Corpus(VectorSource(LastQ))

#clean up by removing stop words
LastQ.corpus <- tm_map(LastQ.corpus,
function(x)removeWords(x,stopwords()))

#install wordcloud if not already installed
install.packages("wordcloud")
library("wordcloud")

#generate wordcloud
wordcloud(LastQ.corpus, min.freq = 2,
scale=c(2.5,0.5),colors=brewer.pal(8, "Dark2"),
random.color= FALSE, random.order = FALSE, max.words
= 50)
```

This is result we received:

As we can see from the above word cloud, lots of words have similar meanings and are ever repeated with singular and plural verbs.

Looking through all the words, we came up with following coding for our open-ended data:

1 = User-Unfriendliness / technical issues (Complicated)
2 = Not Comfortable (Comfortable, Aware, tried)
3 = Distribution/ Accessibility (available, heard)
4 = Security (Risky, secure)
5 = No reasons

CONCLUSION

After applying the word cloud technique on our users and non-users responses for text mining purposes, we can conclude that there are 3 reasons which are common concerns among both these set of people: Security, Distribution/ Accessibility and lastly, Technical issues.

Since it's the advent of era of cashless payment systems, the systems and the corporations manning these are in still their infancy and so it is to be expected that people are still doubtful about their credibility and ability to securely conduct transactions[15]. We believe that as time goes by, the adoption of these systems will gradually increase and given the boost with the current demonetization, these concerns will reduce with the backing from the central government.

[15] http://economictimes.indiatimes.com/wealth/spend/ready-to-go-cashless/articleshow/56269830.cms

FINAL CONCLUSION

The focus of the project was not only confined to one or two aspects of cashless payment systems but also on multiple factors which influence its use and popularity. Since our aim was to find out the challenges faced by people trying to use cashless payments, we ran statistical analysis to find out the major factors and concerns surrounding it.

Thus, our project was divided into five objectives assessing the factors influencing the use of cashless and how we can strategize them effectively to help make this transition easier for everyone.

- Objective one was focused on identifying and analyzing socio-demographic factors influencing the preference of cash v/s cashless. We used logistic regression to assess the significance of different factors in the survey responses. We extracted 9 significant factors that we will use in further analysis and gain more information.
- Objective two was to assess the scenarios and group them by their similarity to understand cashless habits of people better. For this, we went forward with Principal Component Analysis method. After our analysis, we categorized our scenarios into three broad categories which are Official purposes, Leisure and E-services. This gave an idea of the situational based factors that affect people's decision making on using cashless.
- After extracting quantitative and qualitative factors in objectives 1 and 2, we wanted to know the factors which affect a person's preference to go for cashless over cash in daily transactions. So, to find the decision making process, we ran a CART (Classification and Regression tree) model on our data using the factors extracted in objective 1 and additional few factors. We came up with a full model and finally a reduced model with similar accuracy and a pruned tree. This gave us an idea of the factors that we will need to focus on to make people use cashless more.
- In objective four, we wanted to analyze the popularity of different cashless payment systems and reasons behind their popularity. We used simple bar charts for representing the respondent's data as well as Pareto analysis for the factors behind them. After analyzing the factors and the different payment systems, we found out banking apps and e-wallets had lot of preference among the cashless users so we did CART model again to find out the factors that make a person prefer one over the other. We further did random forest model for checking the validity of our CART model and we found out that our CART model was the best fit model.

- Our final objective was to see the major concerns among users and non-users of our current cashless system. Since our data for the same was from open-ended questions, we did a text mining technique called Word cloud. We found out that 3 concerns were common among users and non-users that were security, distribution and technical issues. So, if we come up with solutions to tackle these issues first, we can satisfy a lot of people and increase the usage of cashless payment systems.

Going from a cash based to a cashless economy is a very slow and time consuming process as people have their concerns. But there are countries, where use of cashless methods is very common form of payment. Since our government has taken initiative towards the transition and is promoting cashless extensively, we need to make this process happen in a short period of time and minimum hiccups. Cashless is the way to go whether some people like it or not since it is safer, faster, more convenient and more transparent. It is just like the dot com boom of the 2000s where internet as a disruptive force changed the world forever. Additionally, many experts also believe that going cashless is better for the economy as economy could get rid of considerable amount of illicit money.

So, after reviewing the analysis and objectives of our project, we feel we have found the requisite factors and the type of people who prefer cashless distinctly from people who don't.

RECOMMENDATIONS

- In the times of cyber-hacks and user data leaks, e-wallet companies and government needs to strengthen the encryption behind user's data and their money. The government needs to formulate stringent laws against such security breaches and come up with a contingency plan to compensate or bring the systems back in order should such a scenario arise. People need assurance of the safety of their money just like how they used to stash cash away rather than deposit in banks for the fear of bankruptcy and loss.
- Since this is the advent of cashless systems, it's understandable that it is not acceptable and available everywhere but it should be. With government plans like Digital India and disruption of telecom industry with 4G, the distribution of merchant acceptance has to spread widely, especially in the rural areas where it will reap most benefits. One major concern noted among the non-users of cashless payments is the MDR charged by various banks and that is a huge deterrent for people trying to make the transition and accept cashless as primary mode of payment.

SAS CODES

FOR BINARY LOGISTIC REGRESSION

```
#Create a library and reference the data files to
that library
%let path=/folders/myfolders/;

libname orion "/folders/myfolders/";

#Load the required Excel file
libname orion xlsx "&path\Surveydata.xlsx";

#See the contents of the library
proc contents data=orion._all_;
run;

#Create a SAS data set from the Excel file
data work.CashLogit;
   set orion.Sheet1;
run;

#Turn graphics mode on
ods graphics on;

#Run the logistic regression model using PROC
LOGISTIC step
proc logistic data=CashLogit plots(only)=(roc(id=obs)
effect);
class Gender Education Occupation Religion Marital
      House Income Family_Type Phone Public
      Private Co_op / param=ref;
model cashless = Age Gender Education Occupation
Religion
      House Income Marital Family_Type Phone Public
      Private Co_op Dependents No_members
      No_Accounts / selection=stepwise details
sle=0.05 sls=0.05 ctable pprob=0.5;
output out=prob p=pr reschi=resi1;
run;
```

BIBLIOGRAPHY

BOOKS

Krzanowski, W. J., and F. H. C. Marriott. *Multivariate analysis*. London: Edward Arnold, 1994. Print.

Cody, Ronald P., and Jeffrey K. Smith. *Applied statistics and the SAS: programming language*. Upper Saddle Ricver NJ: Prentice Hall, 1997. Print.

Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Hoboken, NJ: Wiley, 2013. Print.

Anderson, T. W., and Jermy D. Finn. *SPSS guide to new Statistical Analysis of Data*. N.p.: Wiley, n.d. Print.

Malhotra, Naresh K. *Essentials of marketing research: an applied orientation*. Frenchs Forest, N.S.W.: Pearson Education Australia, 2008. Print.

WEBSITES

- edX – Data analytics course
- Wikipedia
- StackOverFlow
- AnalyticsVidhya
- Hindustan Times
- Times of India
- SAS Support
- Stats.Idre.UCLA.Edu
- Stack Exchange
- Business Standard
- Mashable
- Slide Share
- You Tube
- Quora
- Statistics Solutions
- R-Bloggers