PROJECT REPORT

on

Data mining using pivot table to define logic for classifying digital payment user

CSE408 - DATA WAREHOUSING AND DATA MINING

Project By:-

Vishal Sharma

(14BCE0298)

Submitted to:-

Prof. Hari Seetha

Abstract

In this project the emphasis will be on trying to identify the necessary attributes of citizens that can be used to analyze the need for promotional and educational programs in region to create awareness regarding various digital payment options by creating a sample data set, containing combinations of some attributes and then analyzing the data in multiple dimensions using pivot table and defining logical statements based on these attributes that will help in identifying the required regions based on results obtained for test cases based on the acquired rule based classifier.

1.1 **Introduction**

The problem posed in current economic and social spheres of society is to keep up with Government's decisions that encourage use of Digital Modes of payment for various services offered by government as well as for daily activities. This project aims to achieve logical relationship between attributes possessed by residents of a region that dictates the use of digital modes of payment and hence identifying the regions and social-economic groups that lack in upgrading their economic practices so that attempts can be made to spread awareness regarding these services through the medium of advertising, promotions or providing required infrastructure. The increase in digital mode of payment will eventually lead to transparency in economic transaction which will be followed by reduction in corruption and black money.

1.2 **Literature survey**

1. Policies for digitalization promotion

The article by yourstory.com clearly describes the methods the Government went forward with to promote digital payments which can be found at

(https://yourstory.com/2016/12/narendra-modi-package-digital-cashless-economy).

2. Creating a PivotTable to analyze worksheet data

The material available at Microsoft website supports in the generation and usage of pivot tables for data analysis and can be found at:-

(https://support.office.com/en-us/article/Create-a-PivotTable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576)

3. Demonetisation: Impact on the Economy

This paper published No. 182 on 14-Nov-2016 by Tax Research Team elucidates the impact of De-monetization on the availability of credit, spending, level of activity and government finances

1.3 Proposed work/system

The formation of an estimated test data set will be the first task to be completed containing attributes that are possessed by citizens and are possible to collect. This will be used to demonstrate the approach proposed using Pivot tables that leads to the formation of logical relations on which he rule based classifier works:-

1. A data set with 1000 tupels is used for training, which will be analysed to form the logical relations.

2. The data set will have entities identified by their UID and have six attributes:-

1. Literacy (Educated, Uneducated)

2. Smartphone User (Yes, No)

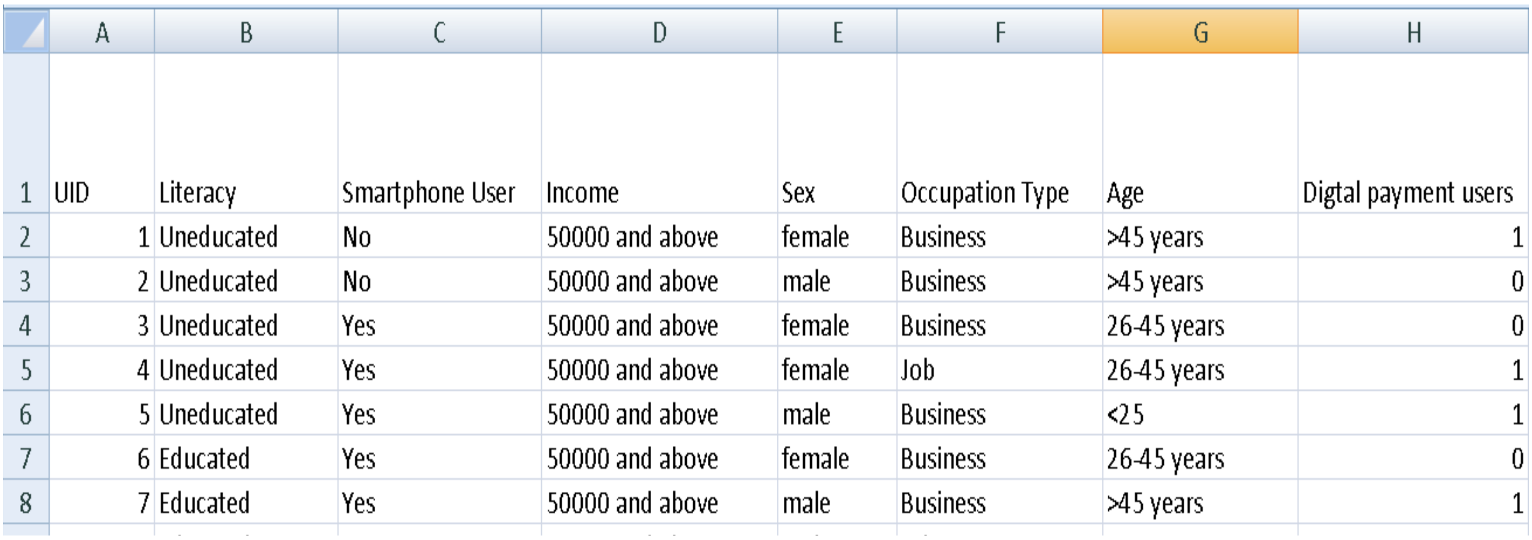
3. Income (50000 and above, 20000-50000, 20000 and below)

4. Sex (Male, Female)

5. Occupation type (Job, Business)

6. Age (<25 years, 26-45 years, >45 years)

Which will act as characteristics of an entity to be analyzed to find relation with weather a person is digital payment user or not which will also be an attribute.

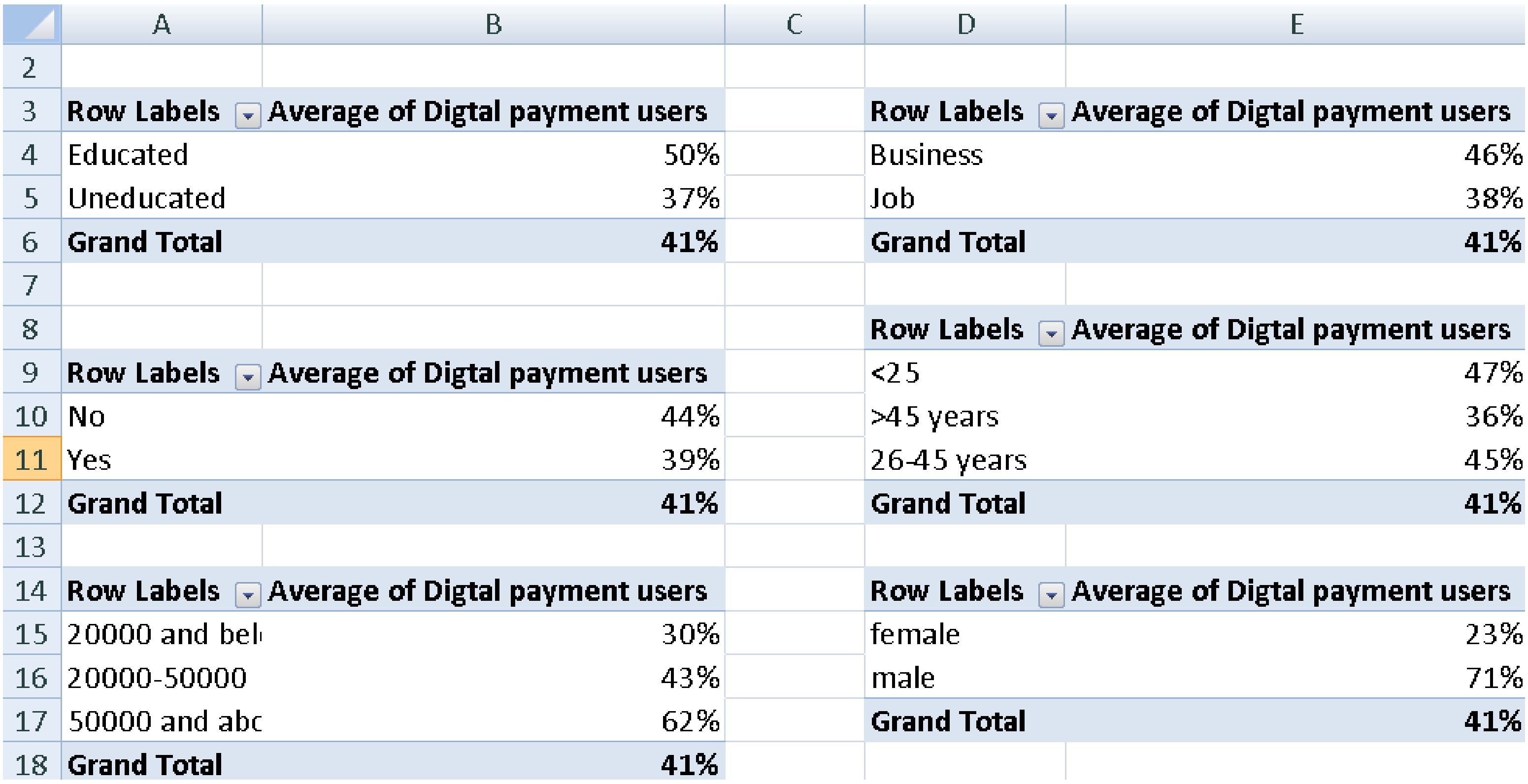


	A	B	C	D	E	F	G	H
1	UID	Literacy	Smartphone User	Income	Sex	Occupation Type	Age	Digtal payment users
2	1	Uneducated	No	50000 and above	female	Business	>45 years	1
3	2	Uneducated	No	50000 and above	male	Business	>45 years	0
4	3	Uneducated	Yes	50000 and above	female	Business	26-45 years	0
5	4	Uneducated	Yes	50000 and above	female	Job	26-45 years	1
6	5	Uneducated	Yes	50000 and above	male	Business	<25	1
7	6	Educated	Yes	50000 and above	female	Business	26-45 years	0
8	7	Educated	Yes	50000 and above	male	Business	>45 years	1

Fig 1. The dataset attributes

3. Once the dataset is created we will start the analysis based on the above mentioned attribute through pivot table tool present in Excel to understand the patterns that occur and interdependence of attributes that determine digital payment user.

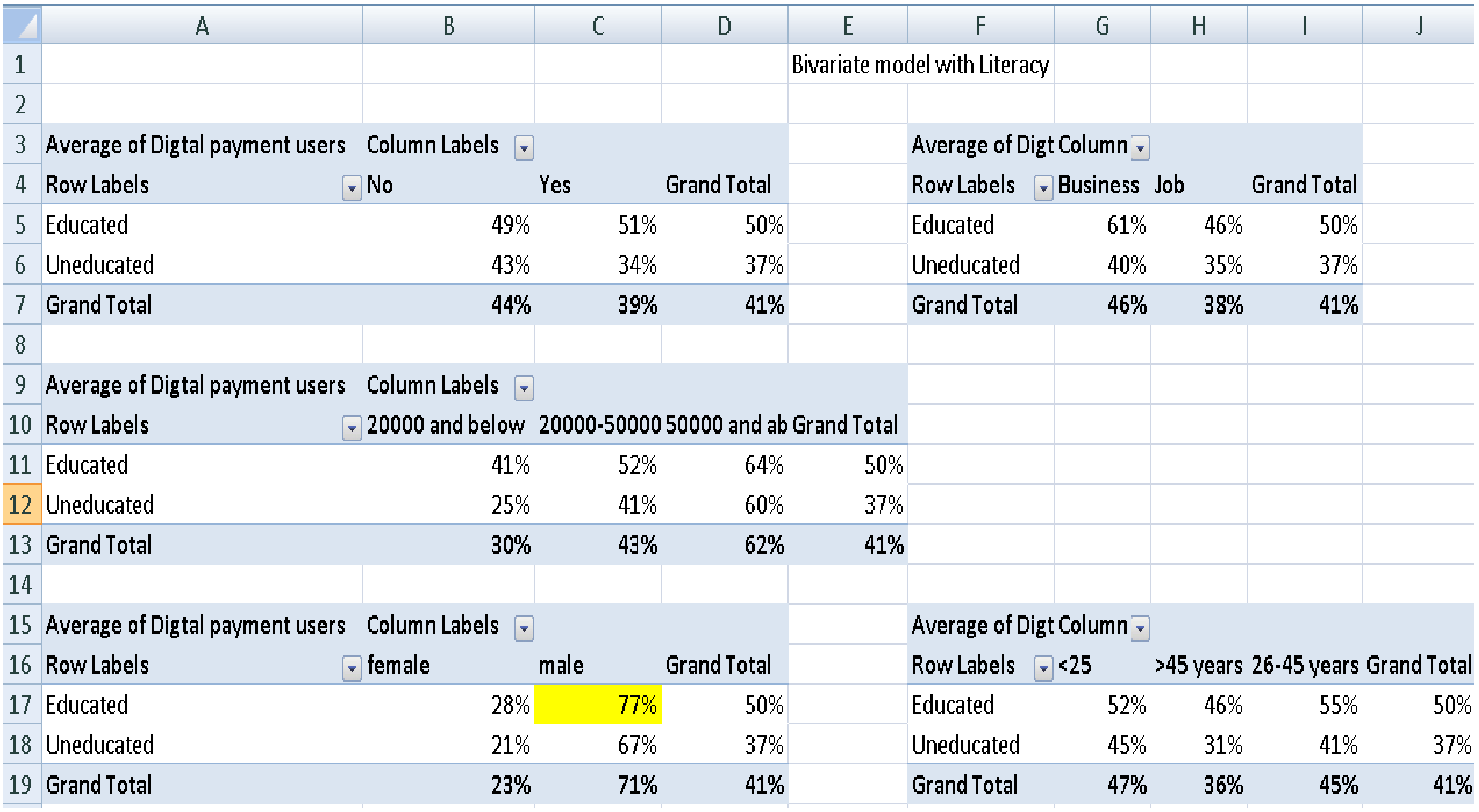
4. Analysis begins with an Univariate model to check independent dependencies between attributes and of tendency to use digital payments The values used for analysis are average value of digital payment users in that category converted to percentage.



	A	B	C	D	E
2					
3	**Row Labels**	**Average of Digtal payment users**		**Row Labels**	**Average of Digtal payment users**
4	Educated	50%		Business	46%
5	Uneducated	37%		Job	38%
6	**Grand Total**	**41%**		**Grand Total**	**41%**
7					
8				**Row Labels**	**Average of Digtal payment users**
9	**Row Labels**	**Average of Digtal payment users**		<25	47%
10	No	44%		>45 years	36%
11	Yes	39%		26-45 years	45%
12	**Grand Total**	**41%**		**Grand Total**	**41%**
13					
14	**Row Labels**	**Average of Digtal payment users**		**Row Labels**	**Average of Digtal payment users**
15	20000 and bel	30%		female	23%
16	20000-50000	43%		male	71%
17	50000 and abc	62%		**Grand Total**	**41%**
18	**Grand Total**	**41%**			

Fig 2. Univariate Model

5. From Univariate Model we can conclude that all the attributes play role in determining the usage of digital payment. Further clarity is obtained by analyzing the dependencies obtained from Univariate model and designing possible bivariate models. The attributes that depend on only one other attribute are identified to form rules. Where there is dependence above 70% those attributes are selected to form rules. The highlighted conditions will be used as rules.



Bivariate model with Literacy

Average of Digtal payment users	Column Labels		
Row Labels	No	Yes	Grand Total
Educated	49%	51%	50%
Uneducated	43%	34%	37%
Grand Total	44%	39%	41%

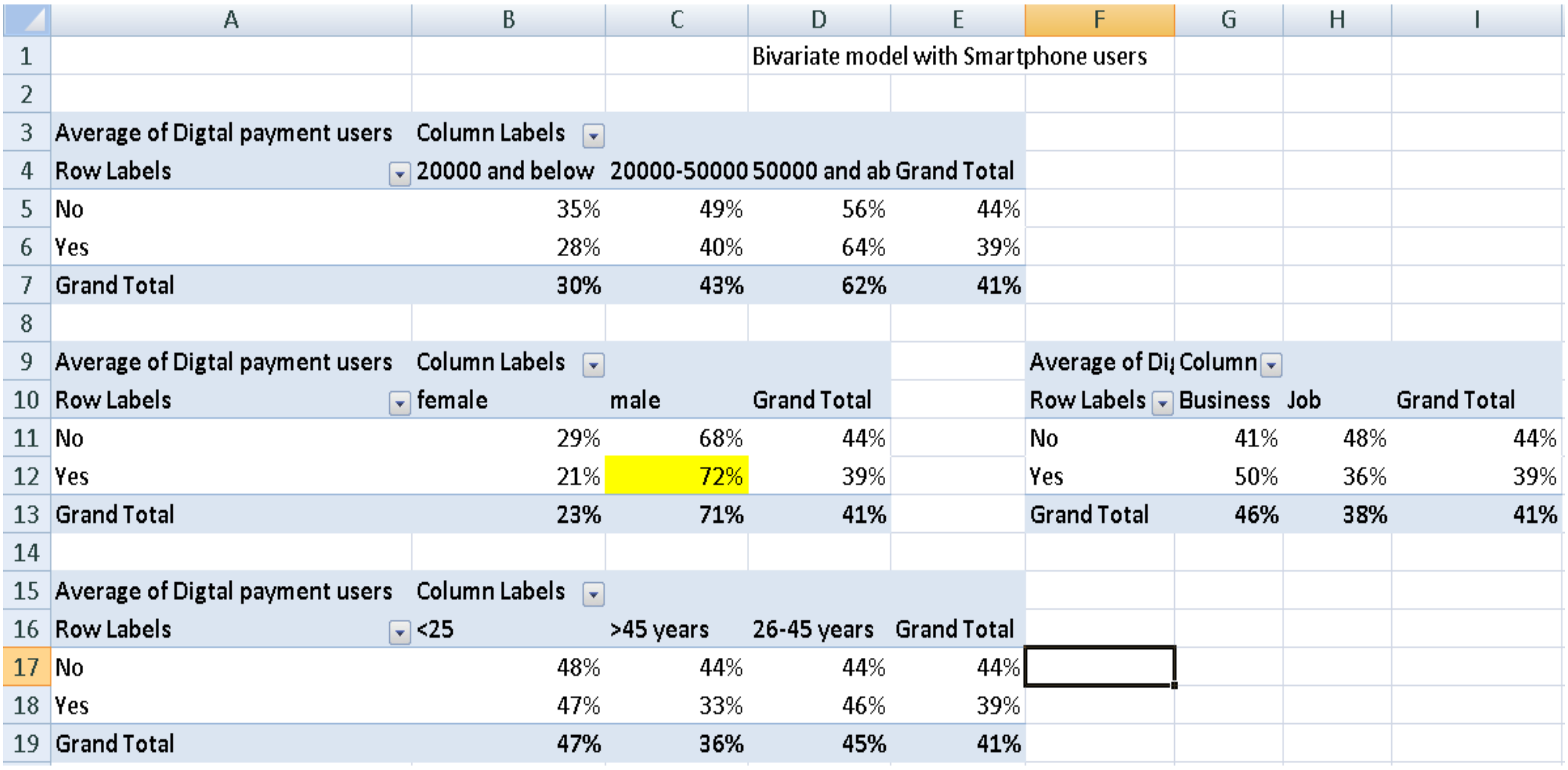
Average of Digt	Column		
Row Labels	Business	Job	Grand Total
Educated	61%	46%	50%
Uneducated	40%	35%	37%
Grand Total	46%	38%	41%

Average of Digtal payment users	Column Labels			
Row Labels	20000 and below	20000-50000	50000 and ab	Grand Total
Educated	41%	52%	64%	50%
Uneducated	25%	41%	60%	37%
Grand Total	30%	43%	62%	41%

Average of Digtal payment users	Column Labels		
Row Labels	female	male	Grand Total
Educated	28%	77%	50%
Uneducated	21%	67%	37%
Grand Total	23%	71%	41%

Average of Digt	Column			
Row Labels	<25	>45 years	26-45 years	Grand Total
Educated	52%	46%	55%	50%
Uneducated	45%	31%	41%	37%
Grand Total	47%	36%	45%	41%

Fig 3. Identifing rules from Bivariate Model with Literacy



Bivariate model with Smartphone users

Average of Digtal payment users	Column Labels			
Row Labels	20000 and below	20000-50000	50000 and ab	Grand Total
No	35%	49%	56%	44%
Yes	28%	40%	64%	39%
Grand Total	30%	43%	62%	41%

Average of Digtal payment users	Column Labels		
Row Labels	female	male	Grand Total
No	29%	68%	44%
Yes	21%	72%	39%
Grand Total	23%	71%	41%

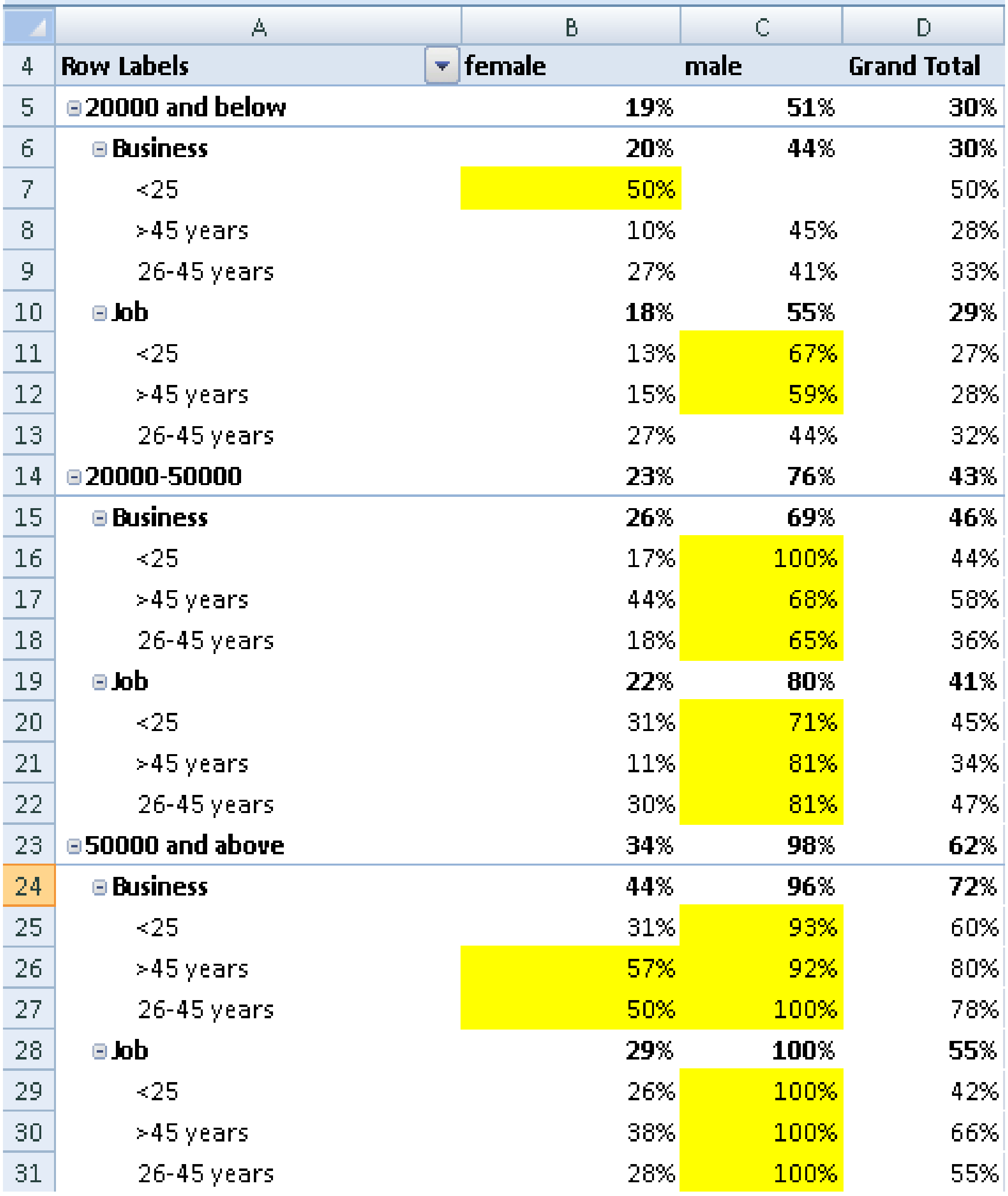
Average of Di;	Column		
Row Labels	Business	Job	Grand Total
No	41%	48%	44%
Yes	50%	36%	39%
Grand Total	46%	38%	41%

Average of Digtal payment users	Column Labels			
Row Labels	<25	>45 years	26-45 years	Grand Total
No	48%	44%	44%	44%
Yes	47%	33%	46%	39%
Grand Total	47%	36%	45%	41%

Fig 4. Identifing rules from Bivariate Model

with Smartphone Users

6. Once we identify attribute’s dependence on one another we continue with the Multivariate Model on the dataset and identify possible combinations that will lead to identifying person using digital payments, Since our multivariate model has 4 attributes simultaneously we take the selection percentage to be 50% as less tupels are coveres per case as compared to Bivariate model. The highlighted condition will be used as rules.



	A	B	C	D
4	Row Labels	female	male	Grand Total
5	⊟20000 and below	19%	51%	30%
6	⊟Business	20%	44%	30%
7	<25	50%		50%
8	>45 years	10%	45%	28%
9	26-45 years	27%	41%	33%
10	⊟Job	18%	55%	29%
11	<25	13%	67%	27%
12	>45 years	15%	59%	28%
13	26-45 years	27%	44%	32%
14	⊟20000-50000	23%	76%	43%
15	⊟Business	26%	69%	46%
16	<25	17%	100%	44%
17	>45 years	44%	68%	58%
18	26-45 years	18%	65%	36%
19	⊟Job	22%	80%	41%
20	<25	31%	71%	45%
21	>45 years	11%	81%	34%
22	26-45 years	30%	81%	47%
23	⊟50000 and above	34%	98%	62%
24	⊟Business	44%	96%	72%
25	<25	31%	93%	60%
26	>45 years	57%	92%	80%
27	26-45 years	50%	100%	78%
28	⊟Job	29%	100%	55%
29	<25	26%	100%	42%
30	>45 years	38%	100%	66%
31	26-45 years	28%	100%	55%

Fig 5. Multivariate model with sex, income

Occupation type and age

7. On completion of identification of required attribute values convert the obtained into logical relations.

8. Implement the rules on a test dataset and find final result, followed by accuracy of model. If required accuracy is obtained we confirm the rules for implementation in the process of identifying individuals using digital payments.

1.4 Results and Discussions

The result obtained after analysis of dataset using pivot tables can be summarized into following 10 rules for identifying individuals using digital payments:-

1. IF sex="male", income="20000-50000" and occupation type="Business"
2. IF sex="male", income="20000-50000" and occupation type ="Job"
3. IF sex="male" and income="50000 and above"
4. IF E2="male", income="20000 and below", occupation type ="Job" and age="<25 years"
5. IF sex="female", income="50000 and above", occupation type ="Business" and age=">45 years"
6. IF sex ="male", income="20000 and below", occupation type ="Job" and age=">45 years"
7. IF sex="female", income="20000 and below", occupation type ="Business" and age="<25"
8. IF sex="female", income="50000 and above", occupation type ="Business" and age="26-45 years"
9. IF sex="male" ,literacy="Educated"
10. IF sex="male", smartphone user="Yes"

The result is verified on Test data and obtained results are tested against known values to check accuracy, the obtained accuracy is calculated as:-

Cases in coverage = 372

Cases correctly predicted = 282

% accuracy = 282/372*100 = 76%

Hence we can classify digital payment users based on their attributes with accuracy of 76% which will help us in identifying regions which need promotion programs and social groups that lack the in the usage of digital payments.

1.5 Conclusions and Scope for future work

In conclusion of the project we define logical relations among attributes to test the requirement for promotion or infrastructure in a region to support Digitization. The use of pivot tables enabled us to classify people on distinct attribute values they posses and how it affected their outlook toward use of digital payments method.

This type of analysis following the approach of gradual identification of pattern for discrete data set to define rules for classification can be used to analyze any number of phenomenon and how they effect or modify the people effected by it for any region and even in an organization provided with required data set.

1.6 References

- Policies for digitalization promotion
- Creating a PivotTable to analyze worksheet data by Microsoft
- Demonetisation: Impact on the Economy
- Data Mining Concepts and Techniques by Jiawei Han, Micheline Kamber, Jian Pei