greatlearning
Learning for Life

Project report on Predicting the success of 'pre-paid scheme' in share brokerage

Submitted towards partial fulfilment of the criteria
for award of PGPBABI by Great Lakes Institute of Management
Submitted By

Group No. 1[Batch: August 2019, Chennai]

Submitted by

Mohamed Yusuf Silarsha

Research Supervisor

Mr. P. V. Subramanian

Great Lakes Institute of Management

ABSTRACT

Abstract

This project is to understand investor preferences and predicting the success towards the Pre-paid Model of stock broking through detailed market survey was undertaken among investors in Hyderabad city in order to know the reasons for preference. I would be considering various aspects which are key factors for the company growth i.e. Product, price, place, promotion and market share, project would also include the strategies adopted by LKP security in all the outlets relating to all the 4p‘s of marketing. The report would mainly be used by the managers to identify market opportunities and develop targeted promotion plans for various products sold under different categories.

This project includes all the details of the company in research and various product and services of the LKP shares and report the major strengths and weaknesses of the company so that it would help them to identify and determine the scope of the competitors which can help to determine better strategies to overcome the weak areas. In brokerage industry service is very important. So, I try to compare LKP service process with the competitor’s service process. We try to identify the number of competitors present in the market. Here, we try to analyze the competitor’s strategy.

This project would also contain an extensive study of various factors influencing the investment decision of investors and to determine the investment pattern by collecting the responses from the investors with the help of questionnaire. This project documents the market research, which was undertaken from 12th March 2021 till 23rd June 2021, which has gathered valuable insights on the various thought processes that go into the minds of the investors when they look to make an investment in the pre-paid scheme. By tapping into these insights from the investors, LKP shares could mainly focus on those aspects of the product or services.

- Techniques: Predictive modelling and data mining
- Tools: R and SPSS
- Domain: Factor Analysis and Predictive Model

ACKNOWLEDGEMENT

"I certify that the work done by me for conceptualizing and completing this project is original and authentic."

Date: 12.Dec.2021

Place: Chennai

greatlearning
Learning for Life

CERTIFICATE OF COMPLETION

I hereby certify that the project titled **'Predicting the success of pre-paid scheme'** for case resolution was undertaken and completed under my supervision

by Mr. S. Mohamed Yusuf

of **Post Graduate Program in Business Analytics and Business Intelligence (PGPBABI).**

Mr. P. V. Subramanian

Date:
Place: Chennai

CHAPTER 1

LIST OF TABLES AND GRAPHS

CHAPTER 1

1. Introduction

1.1 Project Background

This project has been undertaken at a juncture where an LKP share has increased its presence in the Indian market since its foray into the equity segment in the year 2009-10. By offering higher margins to investors and approximately 10% more return to investors, LKP shares brand Life has been able to increase its presence throughout India in states such as Punjab, Maharashtra, Gujarat, Haryana, and Uttar Pradesh. Reliance is now looking to increase its presence in the North Indian region, Maharashtra being the focus of the brand for scaling up its operations. The reason for this is that Maharashtra is one of the front-line states of India for more investors. It occupies 1st place in people income in the country.

1.2 Project Objective

1. The project aims to identify the various reasons for Pre-paid model preference and non-preference among the investors in Hyderabad city. And also understand the penetration of the Pre-paid model in the brokerage firms.
2. To understand and analyse the competition that the company faces in terms of brokerage fees from the competitors. This would give us a brief idea of the competitors in the field and the marketing process followed by them.
3. To identify the Pre-paid scheme advantages and disadvantages and also identify brand wise market share. In addition to this, the project also looks to identify various insights that would help a newly established brand to foray deeper into the market on a large scale.

1.3 Scope of the report

The identification of the factors which act as drivers in the choice of a Pre-paid and subsequently in the decision making process of a Investors, proved to be a challenge since every Investors had to be studied keenly. A lot of discussion with the investors and the brokerage company ensured that the insights kept flowing in. Investors too were approached in order to identify the factors which motivate them in making a purchase. General trade and the modern trade were covered during the entire market research. This project focuses on providing insights which would help

LKP Shares form measures to take on their more established counterparts. In that capacity, this project looks to suggest key result areas which the company could look at focusing in order to successfully build its brand in the Pre-paid model.

1.4 Limitations of the Project

1. The research was conducted in the entire Hyderabad city which has lasted for 6 weeks. Six weeks is not enough for the researcher to observe the investors and advisers for real understanding of their behaviour. It would be better if it was done in a longer time.
2. A lot of responses are collected from investors, but it is difficult to collect the primary data. The reasons are most of the time investors are busy, brokerage firms are not allowing me to get response from the investors. So, that it will take long time to get the response from the investors.
3. The data collection was confined to only Hyderabad city of India since constraints were faced during data collection. The replication of the study at different regions of India would enable better generalization ability of the findings of the study.

1.5 Methodology Used

1. Sample Size	-	200 investors
2. Sample Unit	-	Investors in Stocks.
3. Sampling Technique	-	Stratified random sampling Technique.
4. Sampling Frame	-	Different parts of Hyderabad.
5. Collection of Data	-	Self-designed Questionnaire.
6. Analysis of Data .	-	Pie-charts, Bar graph, Factor analysis, ANOVA and Correlation.

The project employs the observation and the survey method for research. Both questionnaires and personal interviews were administered in order to collect the required data for analysis purposes. The research was done across various locales of the Hyderabad city.

2 Industry Analysis

2.1 Origin of Indian Capital market

The history of the Indian capital markets and the stock market can be traced back to 1850's when stockbrokers would gather under banyan trees in front of Mumbai's Town hall. The group eventually moved to Dalal Street in 1874 and in 1875 became an official organization known as 'The Native Share & Stock Brokers Association'. In 1956, the BSE became the first stock exchange to be recognized by the Indian government under the Securities Contract Regulation Act. The imposition of wealth and expenditure tax in 1957 by Mr. T.T. Krishanmachari, the then finance minister led huge fall in the markets.Mr. Manmohan Singh as Finance Minister came with a reform agenda in 1991 and this led to a resurgence of interest in the capital market, only to be punctured by the Harshad Mehta scam in 1992.

3 Company Analysis

3.1 About the Company

LKP Securities is basically a Brokerage house and is a pioneer in the Retail Broking Industry. LKP provided Securities Broking and Advisory services. It is a corporate member of the Capital market, Wholesale Debt market and the Derivative segment of NSE and BSE. LKP has been assigned the rating BQ-1 by Crisil, the highest rating attained by a brokerage house.
LKP Securities provides product and services related to the purchase and sale of Securities listed in NSE and BSE through various types of brokerage accounts.

3.2 Market share of Company

LKP occupies 0.72 market share in the share brokerage industry whereas ICICI direct is the market leader in the brokerage industry.

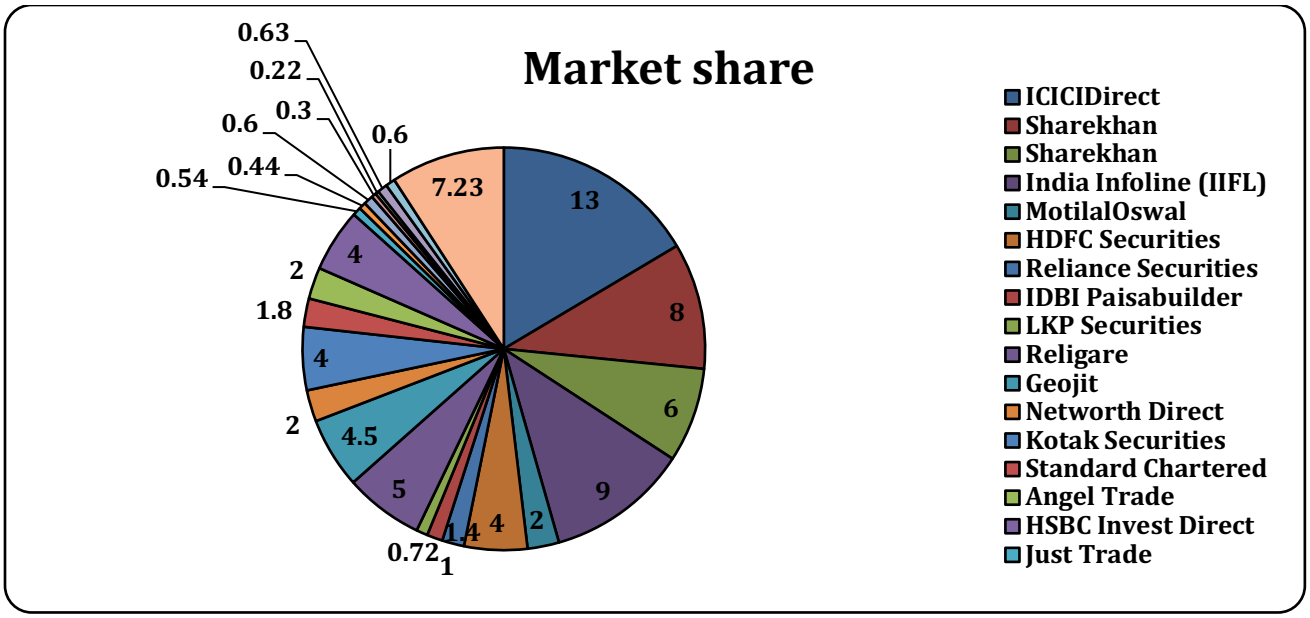


Figure 1.1

4 About LKP Pre-paid scheme

What is it?

Instead of paying a brokerage every time you execute a deal, you can make a one-time deposit with the broker in advance and let him deduct charges as and when the transactions take place. You can replenish the deposit whenever you exhaust it.

Subscription amount limit: There are five types of value trade schemes are there. They are

1. Value trade scheme 14,995
2. Value trade scheme 19,995
3. Value trade scheme 24,995
4. Value trade scheme 49,995
5. Value trade scheme 99,995

5 Interpretation and Analysis

5.1 Exploratory Data Analysis

5.1.1 Data Summary

The response data file contains following attributes

S.No	Description	Variable name	Type
1	Investors look most when they invested in the shares	Invest_pref	Categorical
2	At what Extent brokerage fees influences investors	brok_fee_influence	Categorical
3	What Investor consider most while investing in scheme	Invest_criteria	Categorical
4	**Interested to invest in Pre-paid scheme (**TARGET variable**)**	**Invest_on_scheme**	Categorical
5	Trading Type	Trade_type	Categorical
6	Willing to switch company if scheme offered is attractive	Switch_co	Categorical
7	Value Trade Scheme	Scheme_value	Categorical
8	Investor income level	Income_level	Categorical
9	Invest in this scheme if scheme price is less	lessprice_same_ co	Categorical
10	Preference to change scheme	Change_scheme	Categorical
11	Advertisement is essential for pre-paid scheme	Advertisement	Categorical
12	Pre-paid scheme fetch more return	Prepaid_high_profit	Categorical
13	Pre-paid scheme beneficial for large volume investors	benefit_big_investor	Categorical
14	Brand is more important than brokerage fees	Brand	Categorical
15	Pre-paid scheme is beneficial for frequent investors	benefit_freq_investor	Categorical
16	Brokerage & value trade Charge in scheme are reasonable	Reasonable_fees	Categorical
17	Value Trade charge is high	High_trade_fee	Categorical
18	Respondent Experience	Experience	Continous
19	Respondent Age	Age	Continous
20	Response collected Location	Location	Nominal
21	Respondent Preferred Branded company name	Pref_brand	Nominal
22	Respondent Name	Name	Nominal

Tab 1.1

5.1.2 Data Overview:

There are 22 columns with 4400 records and the data also contains header name for better understanding of the data. In the given dataset, variables such as Invest_on_scheme, Switch_co, lessprice_same_co, are dichotomous in nature

5.1.3 Structure of Data

```
> str(org_sdata) # Structure
'data.frame':   200 obs. of  22 variables:
 $ Name                 : chr  "HARI.B" "P.S.KUMAR" "PRADEEP SIGH" "KISHORE BABU" ...
 $ Invest_pref          : int  5 5 3 5 3 5 5 3 5 1 ...
 $ brok_fee_influence   : int  2 2 3 5 4 4 4 4 4 3 ...
 $ Invest_criteria      : int  1 1 4 1 1 2 1 4 1 3 ...
 $ Invest_on_Scheme     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Trade_type           : int  1 1 1 1 2 4 1 1 1 3 ...
 $ Switch_co            : int  2 2 2 2 2 1 2 2 2 2 ...
 $ Scheme_value         : int  NA 1 1 2 1 1 1 1 2 1 ...
 $ Income_level         : int  2 1 2 1 2 1 4 1 4 4 ...
 $ lessprice_same_.co   : int  2 2 1 1 1 2 1 2 2 1 ...
 $ Change_scheme        : int  2 1 2 2 2 2 2 2 2 2 ...
 $ Advetisement         : int  4 5 5 5 5 5 5 5 5 5 ...
 $ Prepaid_high_profit  : int  4 4 5 4 4 4 4 5 4 5 ...
 $ benefit_big_investor : int  3 3 3 2 1 3 2 4 2 1 ...
 $ Brand                : int  4 4 5 4 5 4 5 5 4 5 ...
 $ benefit_freq_investor: int  4 3 3 2 3 2 4 3 3 4 ...
 $ Reasonable_fees      : int  5 5 4 5 4 5 4 5 4 5 ...
 $ High_trade_fee       : int  4 4 3 4 2 4 3 5 4 3 ...
 $ Exp                  : int  15 8 3 4 9 7 3 12 4 5 ...
 $ Age                  : int  44 29 26 28 32 32 28 38 32 30 ...
 $ Loc                  : chr  "Ammerpet" "Ammerpet" "Ammerpet" "Ammerpet" ...
 $ Alt_co               : chr  "HSBC" "SHAREKHAN" "INDIA INFOLINE" "SHAREKHAN, STANDARD CHATTERED" ...
```

Fig 1.2

5.2 Data Preprocessing:

5.2.1 Detection of Missing Values:

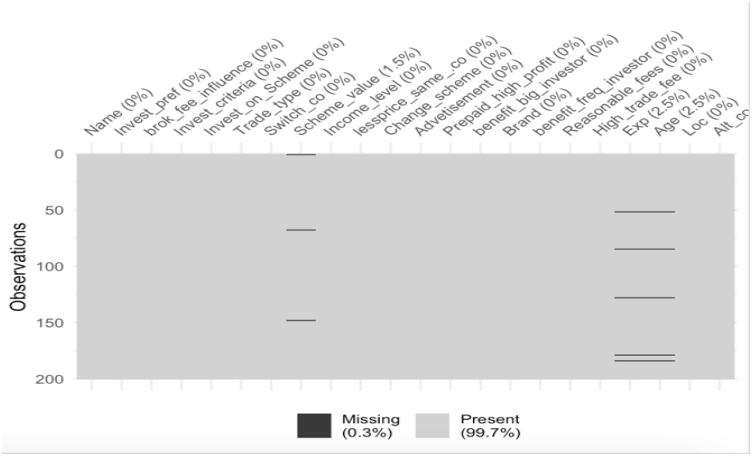


Fig 1.3

Checked for missing values and handled them. Total of 14 null values found in data and imputed accordingly.

5.2.2 Outliers Detection

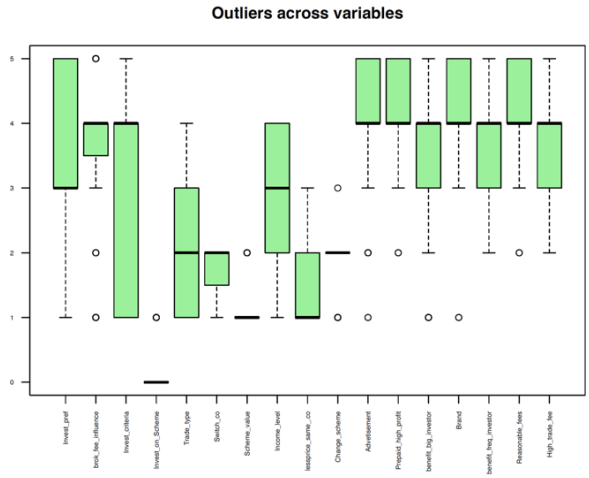


Fig 1.4

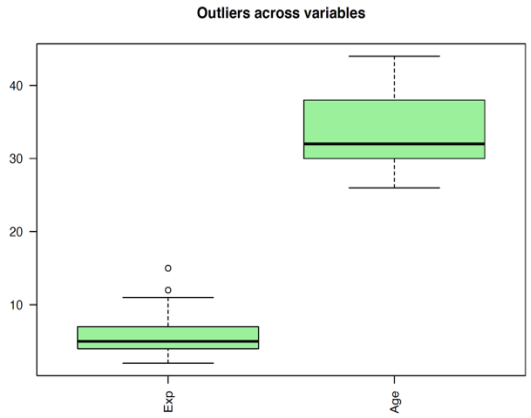


Fig 1.5

5.3 Univariate and Bi-variate Analysis

5.3.1 Univariate Analysis

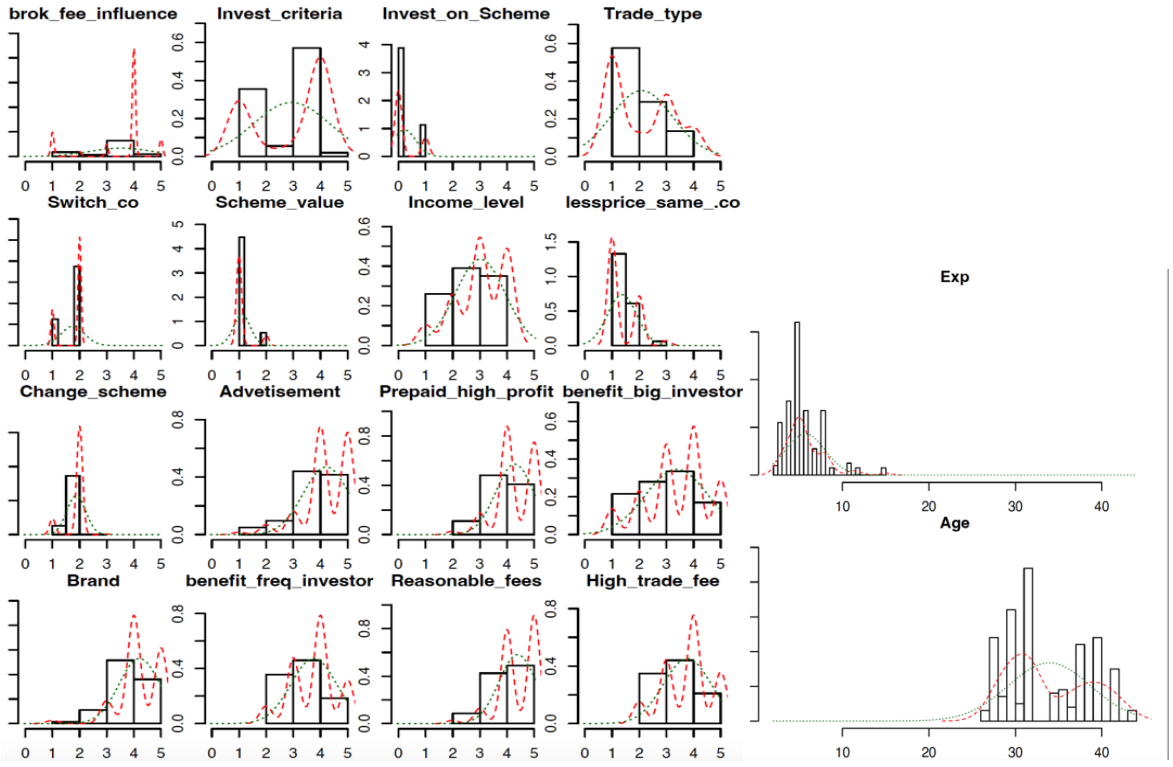


Fig 1.6

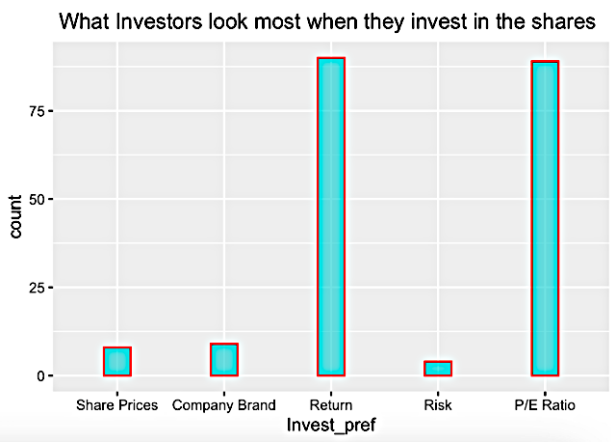


Fig 1.7

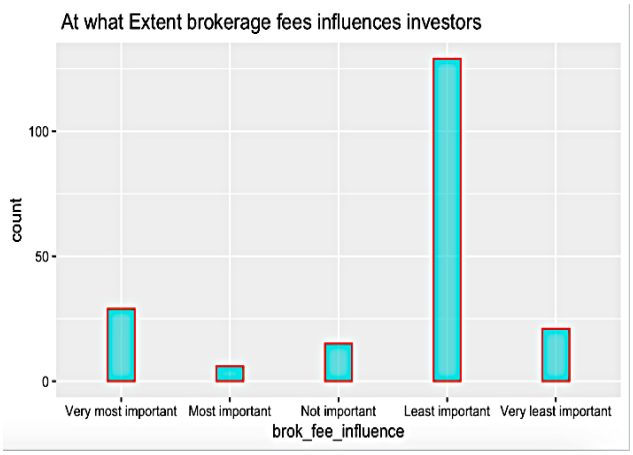


Fig 1.8

The **price-to-earnings ratio** is the ratio of total market capital value over earnings. The P/E ratio of a company is a major focus for many managers.

Brokerage Fees: A fee charged by a brokerage company to facilitate transactions between buyers and sellers.

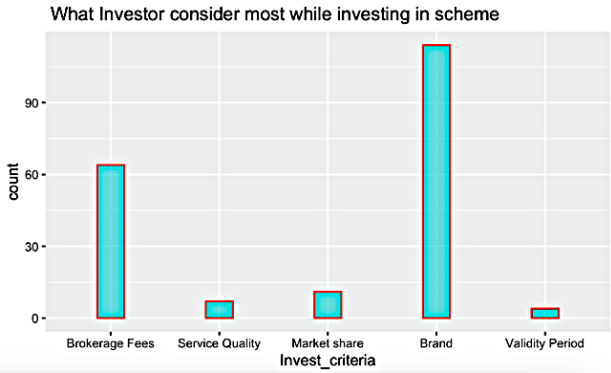


Fig 1.9

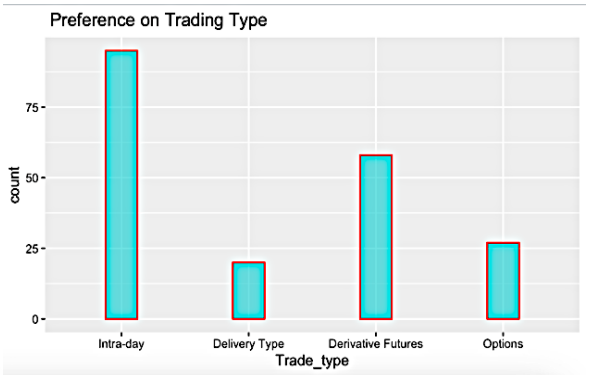


Fig 1.10

Brand: A brand is often the most valuable asset of a corporation. Brand owners manage their brands carefully to create shareholder value, and brand.

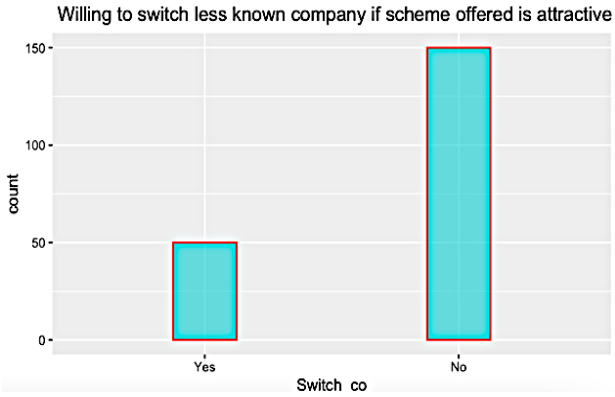


Fig 1.11

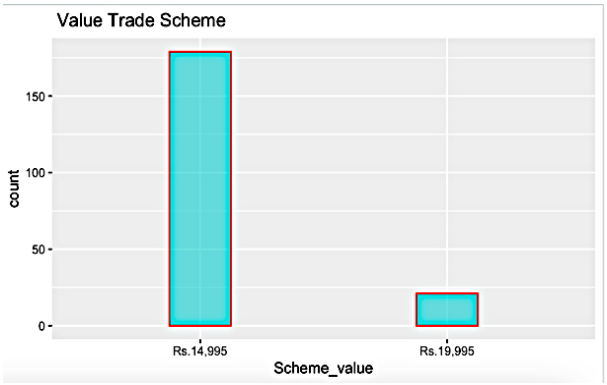


Fig 1.12

Investors prefer to stick with well-known companies even if they are offered attractive offers. Also, the Scheme Value of Rs.14,995 is preferred most among along scheme which is budget friendly to investors.

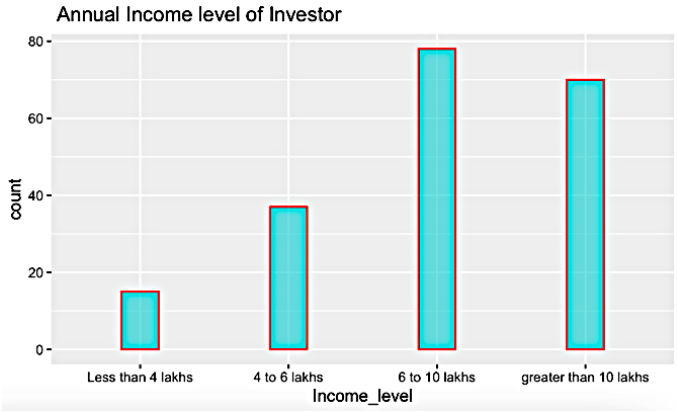


Fig 1.13

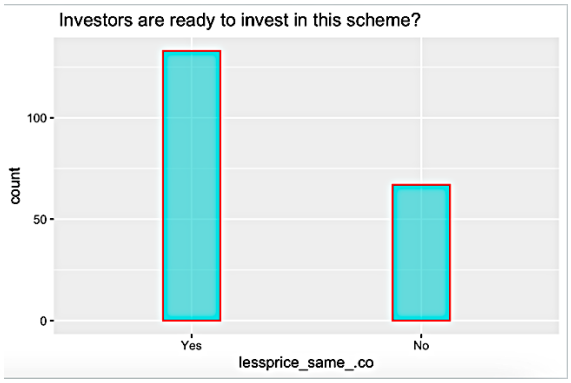


Fig 1.14

Annual Income of investors are in higher slab where the income group of 6 to 10 lakhs seems highest. Investor is ready to invest in this scheme if lower price offered to them.

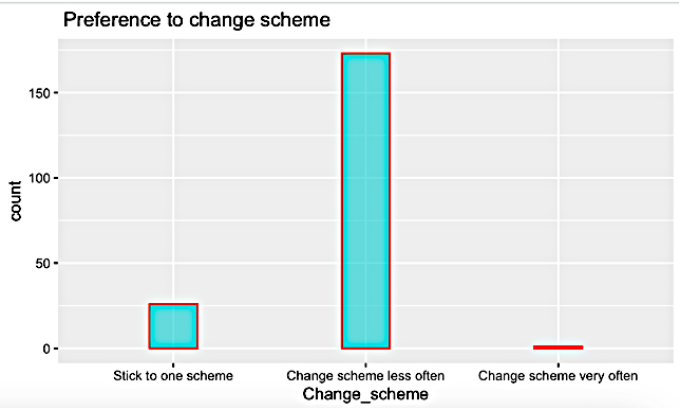


Fig 1.15

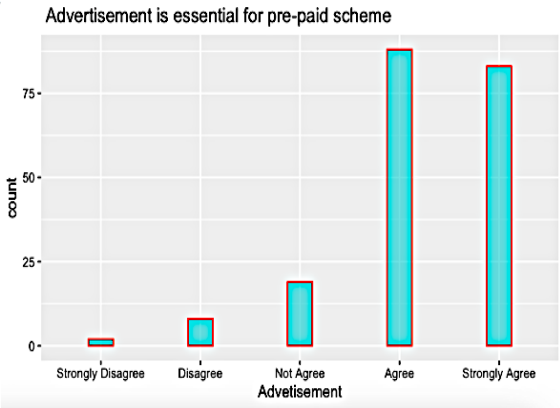


Fig 1.16

Investors prefers to change scheme less often and Advertisement play vital role in promoting the scheme as preferred by investors.

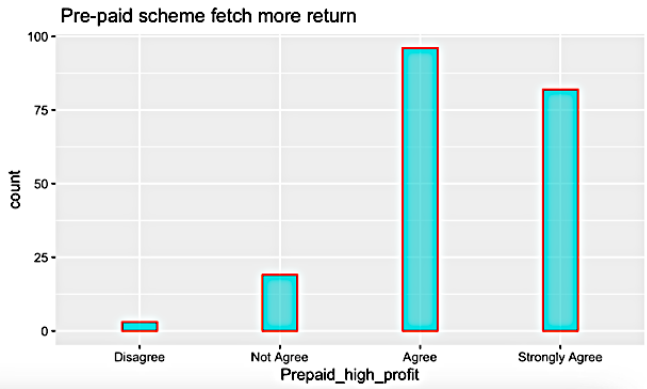


Fig 1.17

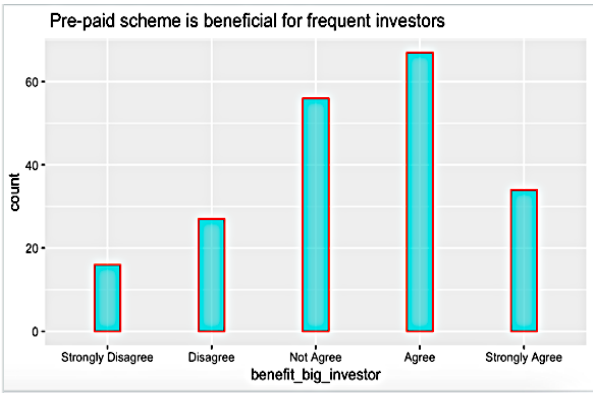


Fig 1.18

Many agree that pre-paid scheme would fetch more returns to them and also its more beneficial to investors.

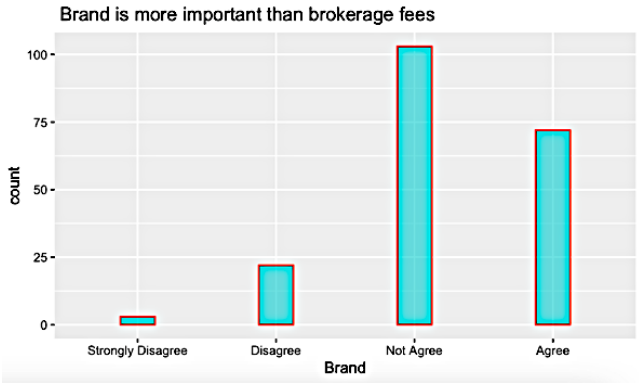


Fig 1.19

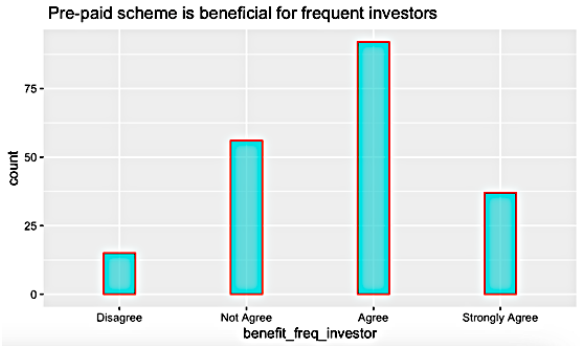


Fig 1.20

We could understand that brokerage fees seem to be more important than brand. Also, pre-paid scheme is beneficial to investors who frequently invest.

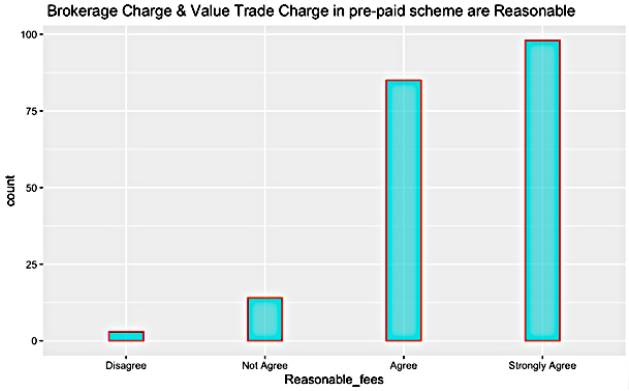


Fig 1.21

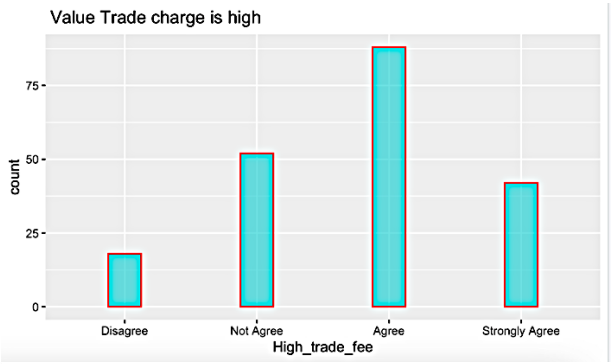
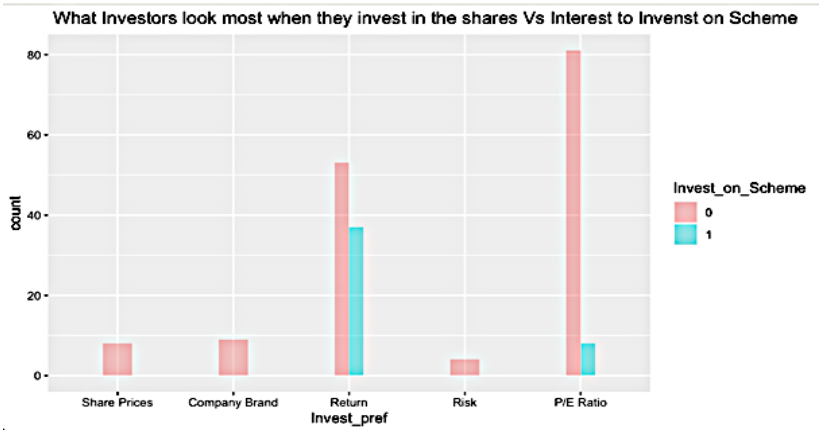


Fig 1.22

Most of investors strongly agree that charges for brokerage and Value trade are reasonable. Also investors agreed that value trade charges are high when it is considered alone.

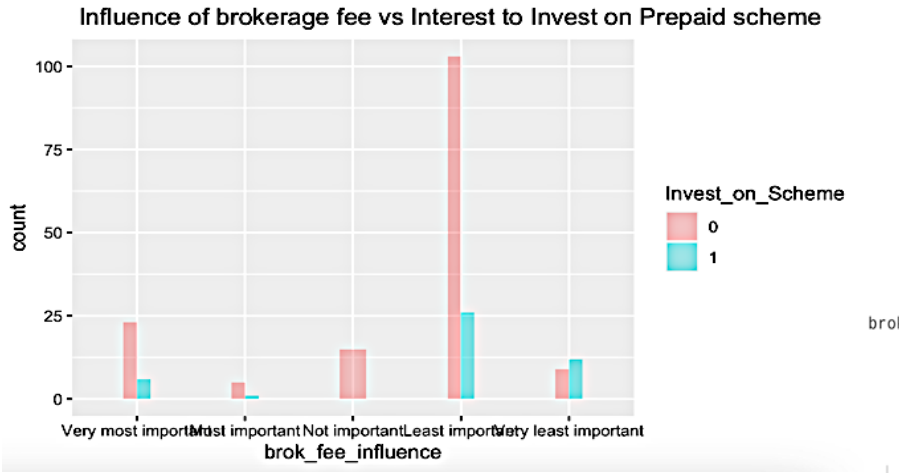
5.3.2 Bi Variate Analysis



Invest_pref	Invest_on_Scheme 0	1
1	100.000000	0.000000
2	100.000000	0.000000
3	58.888889	41.111111
4	100.000000	0.000000
5	91.011236	8.988764

Fig 1.23

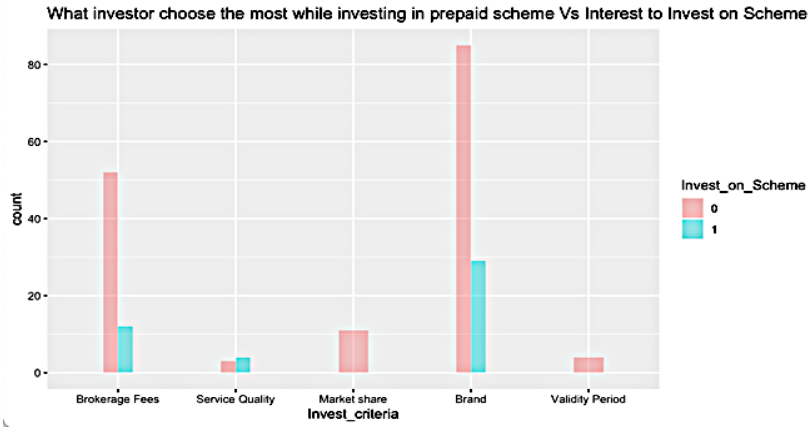
P/E Ratio seems to have higher influence in affecting the target variable with higher difference.



brok_fee_influence	Invest_on_Scheme 0	1
1	79.31034	20.68966
2	83.33333	16.66667
3	100.00000	0.00000
4	79.84496	20.15504
5	42.85714	57.14286

Fig 1.24

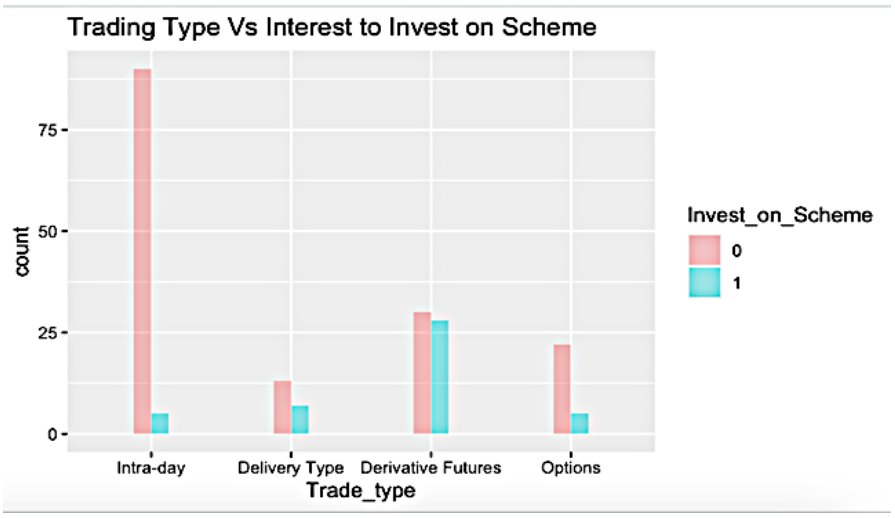
Brokerage fees with Least important have higher significance over Invest on Scheme.



Invest_criteria	Invest_on_Scheme 0	1
1	81.25000	18.75000
2	42.85714	57.14286
3	100.00000	0.00000
4	74.56140	25.43860
5	100.00000	0.00000

Fig 1.25

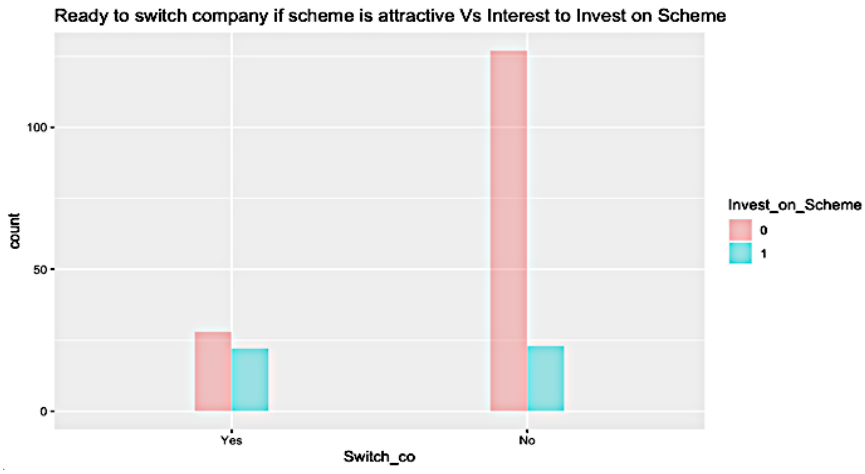
Both Brokerage fees and Brand have higher impact on decision to invest on Scheme.



Trade_type	Invest_on_Scheme 0	1
1	94.736842	5.263158
2	65.000000	35.000000
3	51.724138	48.275862
4	81.481481	18.518519

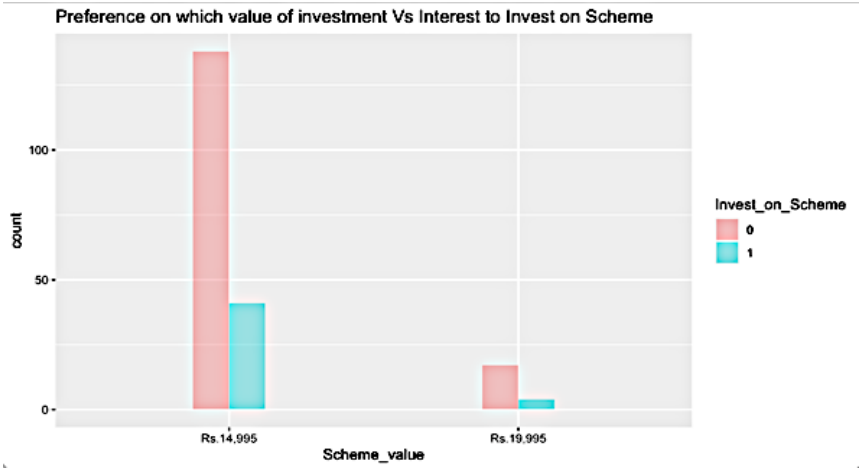
Fig 1.26

Intra-day shows higher impact on target variable



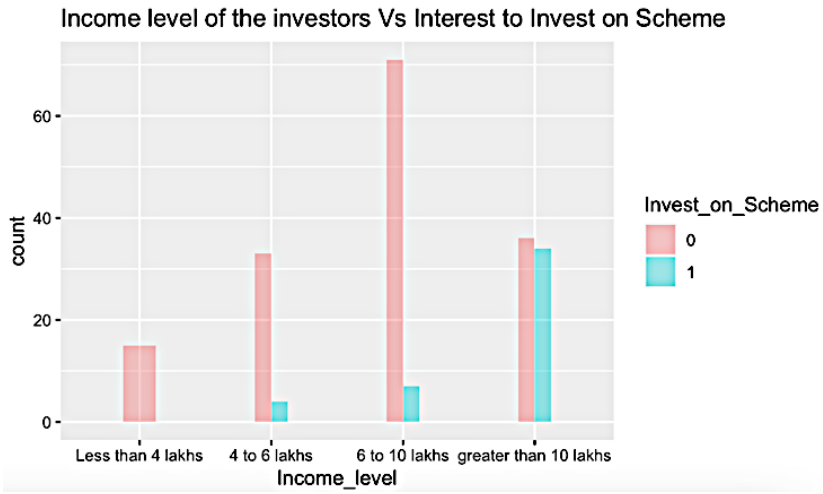
Switch_co	Invest_on_Scheme 0	1
1	56.00000	44.00000
2	84.66667	15.33333

Fig 1.27



Scheme_value	Invest_on_Scheme 0	1
1	77.09497	22.90503
2	80.95238	19.04762

Fig 1.28

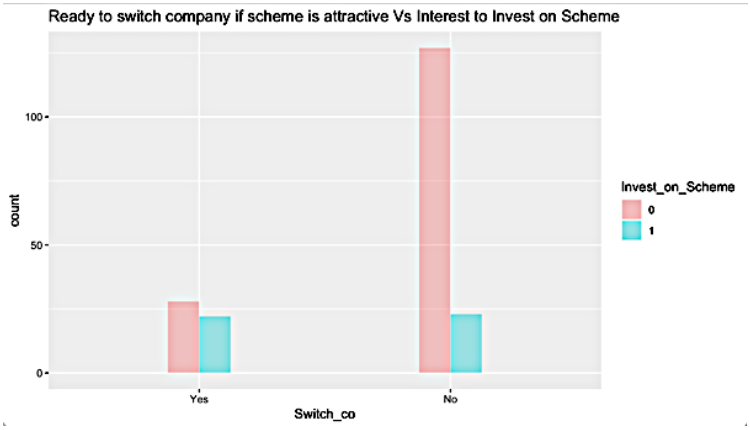


Income_level	Invest_on_Scheme 0	1
1	100.000000	0.000000
2	89.189189	10.810811
3	91.025641	8.974359
4	51.428571	48.571429

Fig 1.29

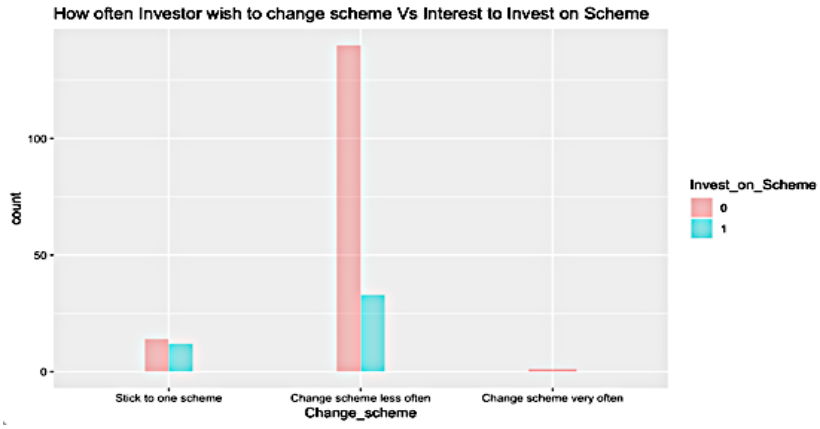
lessprice_same_.co	Invest_on_Scheme 0	1
1	82.70677	17.29323
2	67.16418	32.83582

Fig 1.30



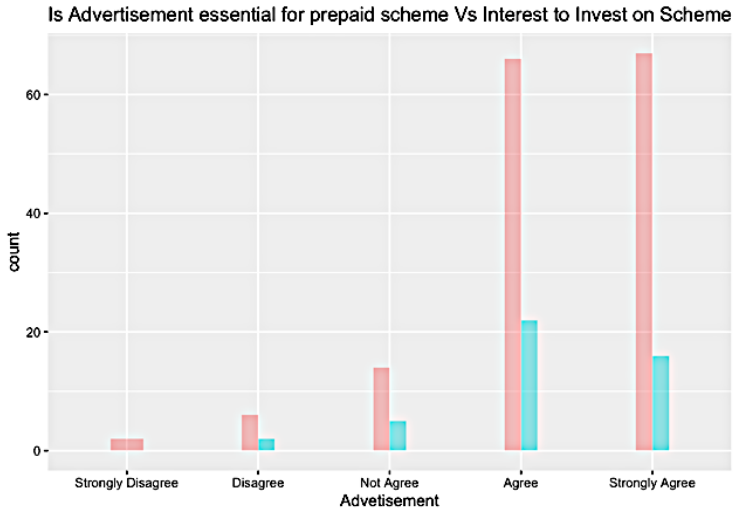
Change_scheme	Invest_on_Scheme 0	1
1	53.84615	46.15385
2	80.92486	19.07514
3	100.00000	0.00000

Fig 1.31



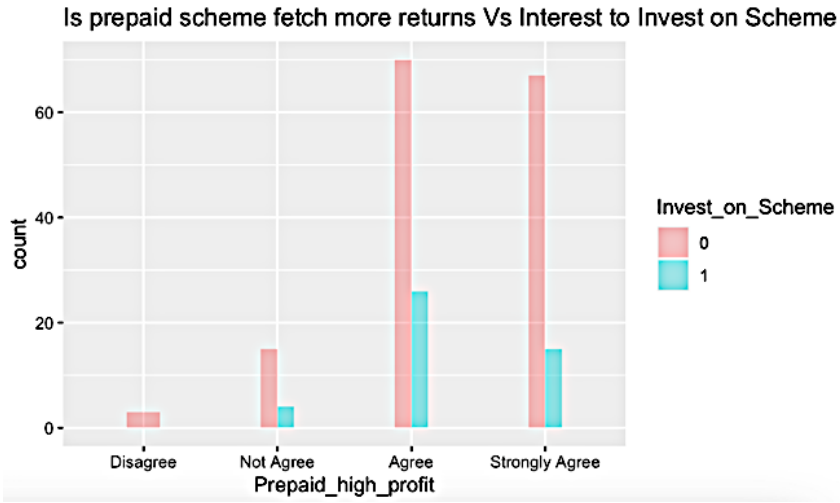
Change_scheme	Invest_on_Scheme 0	1
1	53.84615	46.15385
2	80.92486	19.07514
3	100.00000	0.00000

Fig 1.32



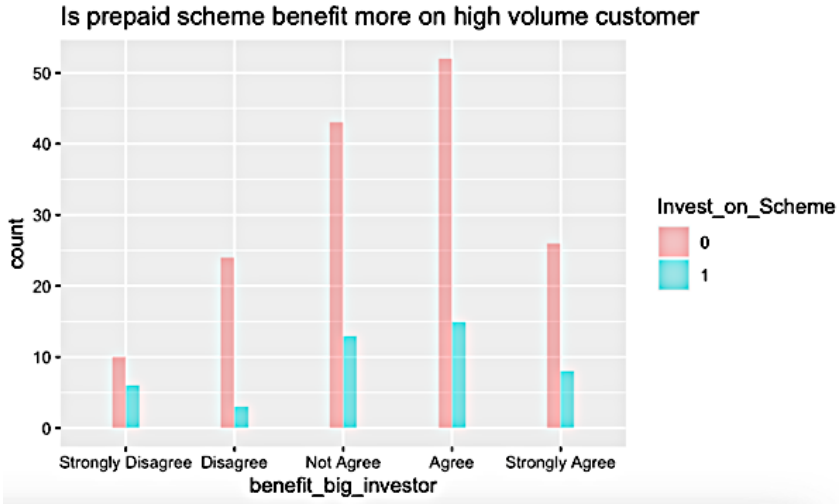
Advetisement	Invest_on_Scheme 0	1
1	100.00000	0.00000
2	75.00000	25.00000
3	73.68421	26.31579
4	75.00000	25.00000
5	80.72289	19.27711

Fig 1.33



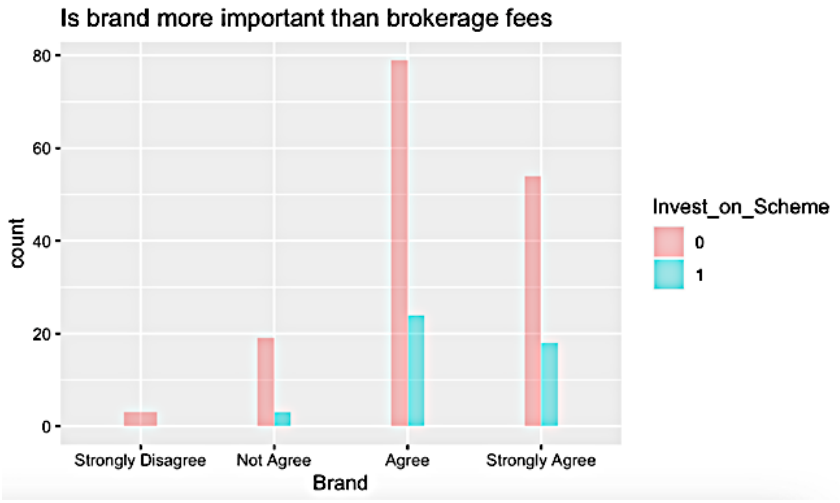
Prepaid_high_profit	Invest_on_Scheme 0	1
2	100.00000	0.00000
3	78.94737	21.05263
4	72.91667	27.08333
5	81.70732	18.29268

Fig 1.34



benefit_big_investor	Invest_on_Scheme 0	1
1	62.50000	37.50000
2	88.88889	11.11111
3	76.78571	23.21429
4	77.61194	22.38806
5	76.47059	23.52941

Fig 1.35

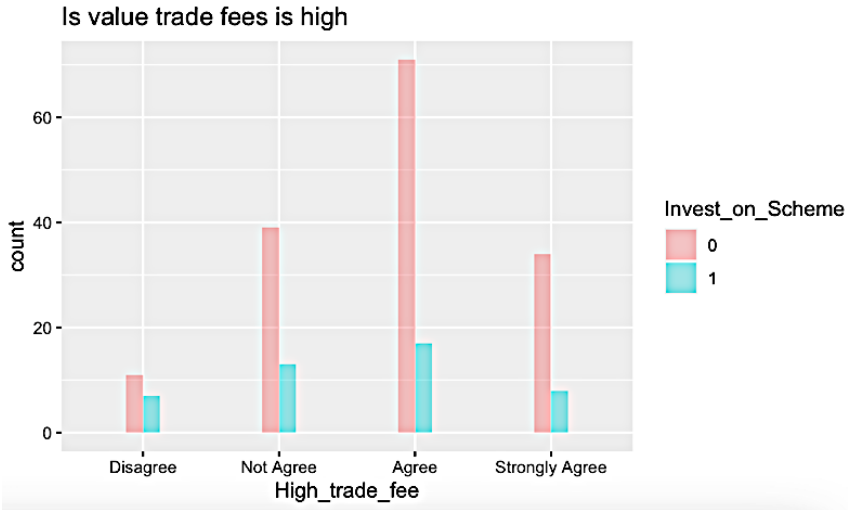


Brand	Invest_on_Scheme 0	1
1	100.00000	0.00000
3	86.36364	13.63636
4	76.69903	23.30097
5	75.00000	25.00000

Fig 1.36

Reasonable_fees	Invest_on_Scheme 0	1
2	33.33333	66.66667
3	71.42857	28.57143
4	75.29412	24.70588
5	81.63265	18.36735

Fig 1.3



High_trade_fee	Invest_on_Scheme 0	1
2	61.11111	38.88889
3	75.00000	25.00000
4	80.68182	19.31818
5	80.95238	19.04762

Fig 1.38

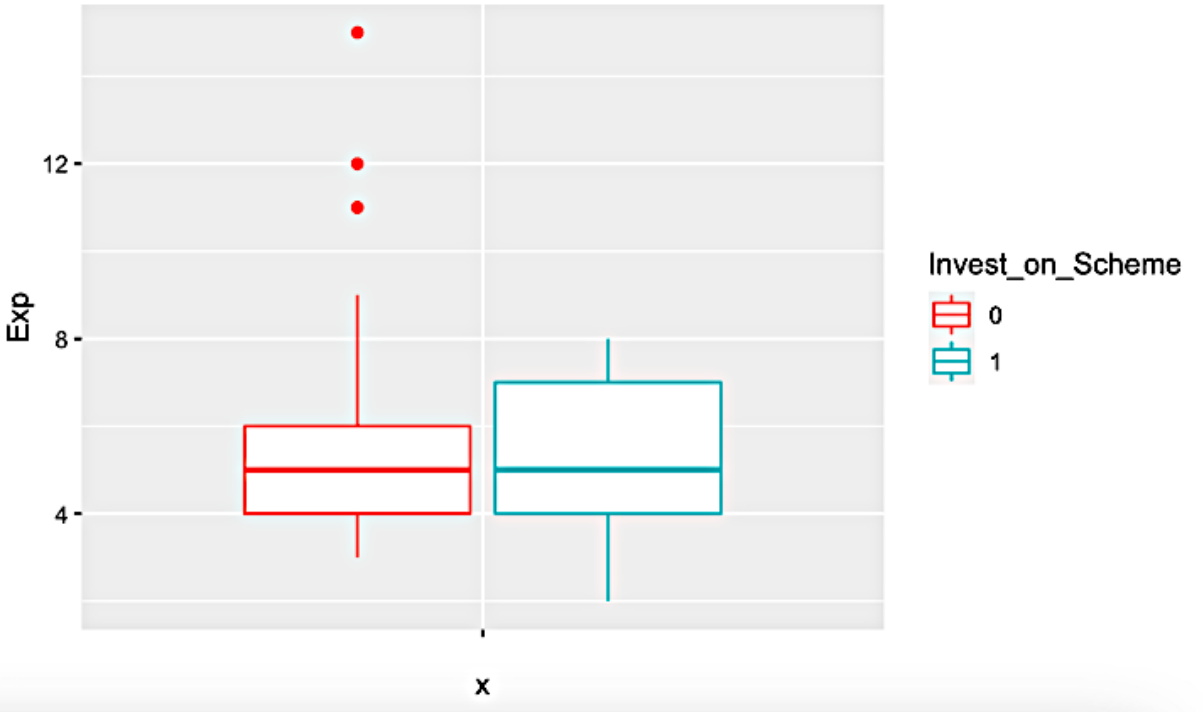


Fig 1.39

Experience is an important variable that influences the target variable i.e. whether to invest or not

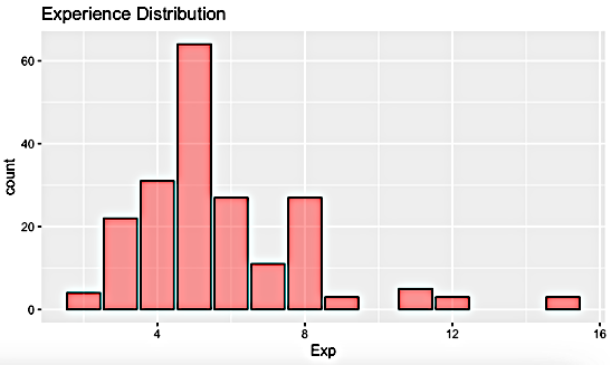


Fig 1.40

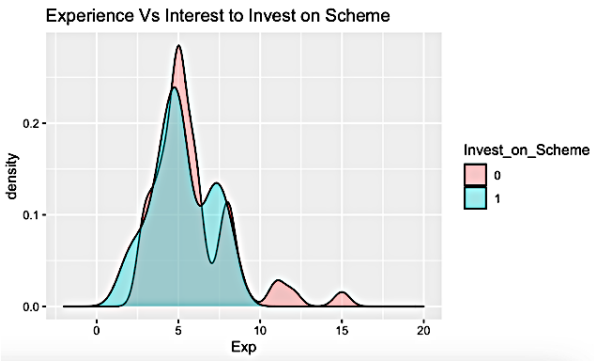


Fig 1.41

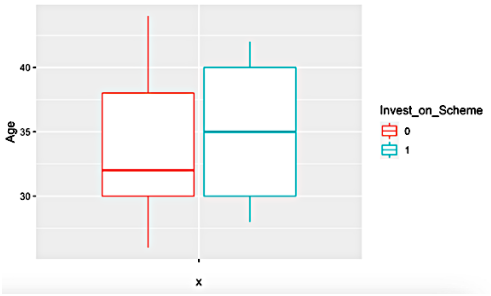


Fig 1.42

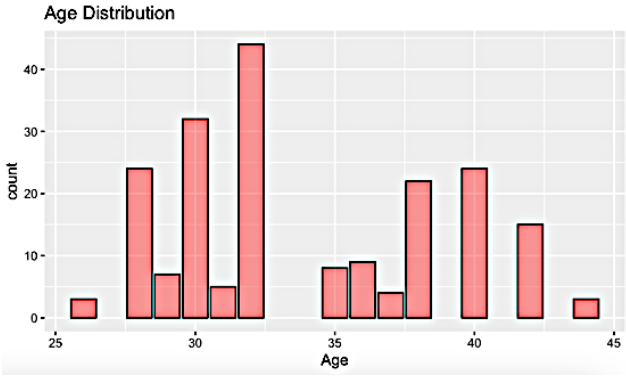


Fig 1.43

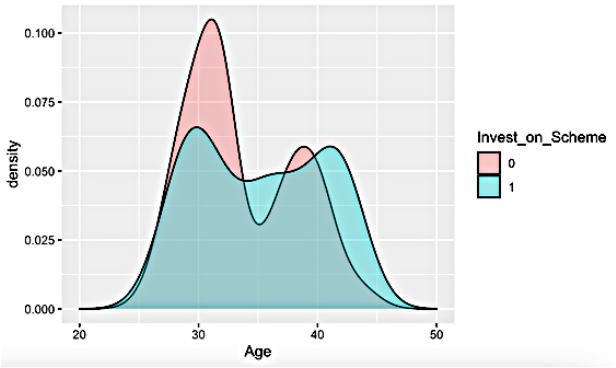


Fig 1.44

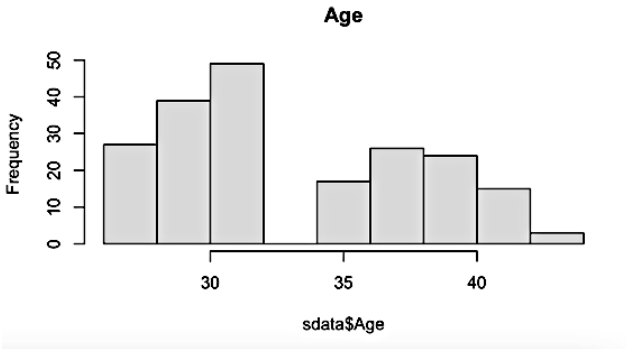


Fig 1.45

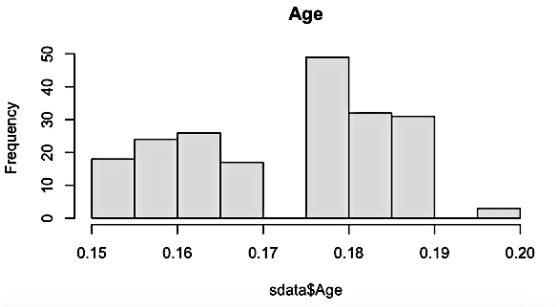


Fig 1.46

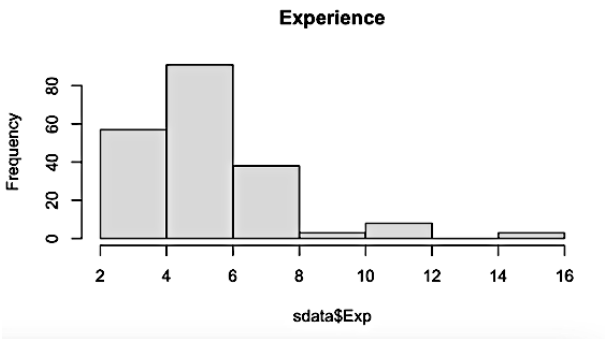


Fig 1.47

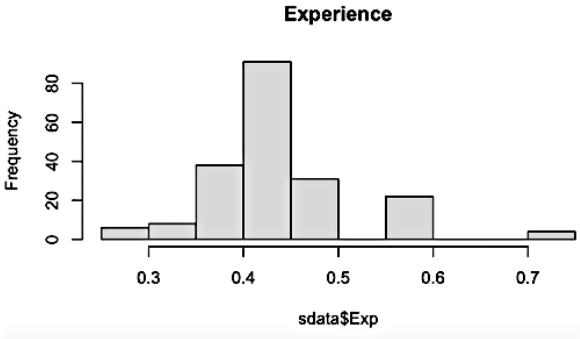


Fig 1.48

5.4 Analytical Approach

5.4.1 KMO value

KMO value is a measure of adequacy. It is a measure that tells whether the number of samples taken for analysis is sufficient or not. If the KMO value is greater than 0.5, the number of samples is sufficient. Else the analysis must be repeated by increasing the number of samples. % Variation explained by factors in the variables selected. Higher the value of KMO better it supports the idea of factor analysis.

```
> KMO(sdata_num)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = sdata_num)
Overall MSA =  0.55
MSA for each item =
          Invest_pref    brok_fee_influence       Invest_criteria
                 0.40                  0.61                  0.41
     Invest_on_Scheme            Trade_type             Switch_co
                 0.56                  0.68                  0.53
         Scheme_value          Income_level    lessprice_same_.co
                 0.50                  0.64                  0.53
        Change_scheme          Advetisement   Prepaid_high_profit
                 0.46                  0.57                  0.65
 benefit_big_investor                 Brand benefit_freq_investor
                 0.47                  0.53                  0.60
      Reasonable_fees        High_trade_fee                   Exp
                 0.59                  0.60                  0.57
                  Age
                 0.46
```

Fig 1.49

The above-mentioned table indicates the KMO factor obtained after the analysis with 200 samples. Since the KMO value obtained (**0.55**) is greater than 0.5, the number of samples taken for the analysis is sufficient.

5.4.2 Factor Analysis:

Factor Analysis is done to basically identify the important factors or variables that influence the measuring variable. In this project, Factor Analysis is done to identify the important factors that influence the investor preference towards pre-paid products.

A set of techniques for finding the number and characteristics of variables underlying many measurements made on individuals or objects. The variables that have been chosen for analysis are as follows:

a. **Brand**
b. **Advertisements**
c. **Fetch more return**
d. **Benefits for large volume investors**
e. **Frequent investors**
f. **Brokerage fees**
g. **Value Trade fees**

Output on Factor Analysis

```
Factor Analysis using method =  alpha
Call: fa(r = sdata_num, nfactors = 3, rotate = "none", fm = "alpha")
Standardized loadings (pattern matrix) based upon correlation matrix
                     alpha1 alpha2 alpha3    h2     u2 com
Invest_pref           -0.17   0.00  -0.11 0.039  0.961 1.7
brok_fee_influence    -0.17  -0.16   0.08 0.059  0.941 2.4
Invest_criteria       -0.08   0.15   0.35 0.152  0.848 1.4
Invest_on_Scheme      -0.38   0.17   0.33 0.282  0.718 2.4
Trade_type            -0.39  -0.06   0.04 0.155  0.845 1.1
Switch_co              0.64  -0.15  -0.61 0.805  0.195 2.1
Scheme_value          -0.03  -0.40   0.21 0.206  0.794 1.5
Income_level          -0.17  -0.11   0.37 0.182  0.818 1.6
lessprice_same_.co    -0.73   0.21  -0.11 0.595  0.405 1.2
Change_scheme          1.07  -0.11   0.36 1.275 -0.275 1.2
Advetisement           0.03   0.00  -0.13 0.018  0.982 1.1
Prepaid_high_profit    0.00   0.13   0.02 0.016  0.984 1.0
benefit_big_investor  -0.03   0.09   0.03 0.010  0.990 1.5
Brand                  0.11   0.30   0.07 0.109  0.891 1.4
benefit_freq_investor -0.06   0.16  -0.06 0.034  0.966 1.5
Reasonable_fees        0.04   0.23  -0.31 0.147  0.853 1.9
High_trade_fee         0.04   0.31  -0.16 0.122  0.878 1.5
Exp                    0.27   0.94  -0.34 1.076 -0.076 1.4
Age                    0.33   0.82   0.25 0.843  0.157 1.5

                     alpha1 alpha2 alpha3
SS loadings            2.68   2.18   1.27
Proportion Var         0.14   0.11   0.07
Cumulative Var         0.14   0.26   0.32
Proportion Explained   0.44   0.36   0.21
Cumulative Proportion  0.44   0.79   1.00
```

Fig 1.50

```
Loadings:
                    alpha1 alpha2 alpha3
Invest_pref         -0.166         -0.106
brok_fee_influence  -0.168 -0.155
Invest_criteria             0.145  0.353
Invest_on_Scheme    -0.379  0.173  0.329
Trade_type          -0.386
Switch_co            0.644 -0.151 -0.606
Scheme_value                -0.401  0.210
Income_level        -0.171 -0.113  0.374
lessprice_same_.co  -0.734  0.210 -0.110
Change_scheme        1.065 -0.110  0.359
Advetisement                       -0.130
Prepaid_high_profit          0.127
benefit_big_investor
Brand                0.105  0.304
benefit_freq_investor        0.164
Reasonable_fees              0.226 -0.308
High_trade_fee               0.309 -0.157
Exp                  0.275  0.942 -0.336
Age                  0.334  0.816  0.255

               alpha1 alpha2 alpha3
SS loadings     2.680  2.176  1.269
Proportion Var  0.141  0.115  0.067
Cumulative Var  0.141  0.256  0.322
```

Fig 1.51

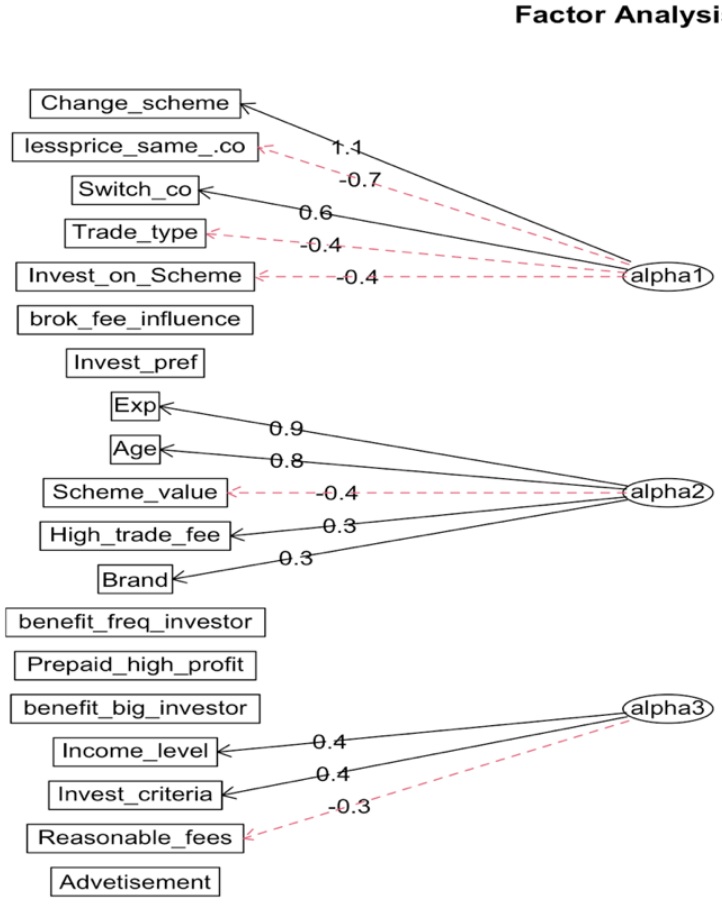


Fig 1.52

5.4.3 Factor Pattern:

Factor Pattern is a matrix showing the factor loadings i.e., the variances between the variables and the factors. The factor loadings, also called component loadings in PCA, are the correlation coefficients between the variables (rows) and factors (columns).

In **factor analysis**, the original variables are defined as **linear combinations of the factors**. In principal components analysis, the goal is to explain as much of the **total variance** in the variables as possible. The goal in factor analysis is to explain the **covariances or correlations** between the **variables.**

5.4.4 Principal Component Analysis

PCA is a dimensionality reduction that **identifies important relationships** in our data, **transforms the existing data** based on these relationships, and then **quantifies the importance** of these relationships so we can keep the most important relationships and drop the others. To remember this definition, we can break it down into four steps:

1. We identify the relationship among features through a Covariance Matrix.
2. Through the linear transformation or eigen-decomposition of the Covariance Matrix, we get eigen vectors and eigen values.
3. Then we transform our data using Eigenvectors into principal components.
4. Lastly, we quantify the importance of these relationships using Eigenvalues and keep the important principal components**.**

Principal Component Extraction

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     5.0104097 1.91658313 1.39134315 1.28593361 1.22369564
Proportion of Variance 0.6178919 0.09041101 0.04764688 0.04070081 0.03685639
Cumulative Proportion  0.6178919 0.70830290 0.75594978 0.79665060 0.83350699
                          Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.1889396 1.03222182 0.95220163 0.80267068 0.74213239
Proportion of Variance 0.0347925 0.02622479 0.02231639 0.01585772 0.01355591
Cumulative Proportion  0.8682995 0.89452428 0.91684066 0.93269838 0.94625430
                         Comp.11    Comp.12     Comp.13     Comp.14
Standard deviation     0.72251808 0.64893918 0.567792819 0.542043775
Proportion of Variance 0.01284883 0.01036511 0.007934979 0.007231606
Cumulative Proportion  0.95910312 0.96946824 0.977403215 0.984634820
                          Comp.15     Comp.16     Comp.17     Comp.18
Standard deviation     0.474331373 0.394815630 0.335439448 0.272323744
Proportion of Variance 0.005537704 0.003836672 0.002769455 0.001825312
Cumulative Proportion  0.990172524 0.994009196 0.996778651 0.998603962
                          Comp.19
Standard deviation     0.238158229
Proportion of Variance 0.001396038
Cumulative Proportion  1.000000000
```

Fig 1.53

```
> round(PCA$loadings[,1:3],3)
                        Comp.1 Comp.2 Comp.3
Invest_pref              0.003  0.094  0.169
brok_fee_influence      -0.045 -0.120 -0.502
Invest_criteria          0.027 -0.368  0.499
Invest_on_Scheme         0.009 -0.078 -0.076
Trade_type              -0.039 -0.170 -0.366
Switch_co               -0.006  0.048  0.007
Scheme_value            -0.015 -0.003  0.026
Income_level            -0.026 -0.300 -0.103
lessprice_same_.co       0.003  0.033 -0.008
Change_scheme            0.015  0.003  0.001
Advetisement             0.008  0.075 -0.134
Prepaid_high_profit      0.006  0.009  0.045
benefit_big_investor     0.018  0.017  0.462
Brand                    0.019  0.021  0.052
benefit_freq_investor -0.002  0.030  0.164
Reasonable_fees          0.011  0.092  0.010
High_trade_fee           0.022  0.102  0.227
Exp                      0.327  0.778 -0.045
Age                      0.942 -0.285 -0.054
```

Fig 1.54

5.4.5 Co-efficient of PCA

```
> round(PCA$loadings[,1:3],3)
                      Comp.1 Comp.2 Comp.3
Invest_pref            0.003  0.094  0.169
brok_fee_influence    -0.045 -0.120 -0.502
Invest_criteria        0.027 -0.368  0.499
Invest_on_Scheme       0.009 -0.078 -0.076
Trade_type            -0.039 -0.170 -0.366
Switch_co             -0.006  0.048  0.007
Scheme_value          -0.015 -0.003  0.026
Income_level          -0.026 -0.300 -0.103
lessprice_same_.co     0.003  0.033 -0.008
Change_scheme          0.015  0.003  0.001
Advetisement           0.008  0.075 -0.134
Prepaid_high_profit    0.006  0.009  0.045
benefit_big_investor   0.018  0.017  0.462
Brand                  0.019  0.021  0.052
benefit_freq_investor -0.002  0.030  0.164
Reasonable_fees        0.011  0.092  0.010
High_trade_fee         0.022  0.102  0.227
Exp                    0.327  0.778 -0.045
Age                    0.942 -0.285 -0.054
```

Fig 1.55

The above-mentioned table shows the factor loadings between all the variables and the three factors.

5.4.6 PCA without Rotation

```
Principal Components Analysis
Call: principal(r = sdata_num, nfactors = 3, rotate = "none")
Standardized loadings (pattern matrix) based upon correlation matrix
                        PC1    PC2    PC3     h2    u2  com
Invest_pref           0.061 -0.168  0.061 0.0355 0.964 1.54
brok_fee_influence   -0.435  0.100 -0.030 0.1998 0.800 1.12
Invest_criteria      -0.150  0.162  0.169 0.0771 0.923 2.97
Invest_on_Scheme     -0.467  0.388  0.446 0.5679 0.432 2.93
Trade_type           -0.523  0.398  0.056 0.4355 0.564 1.89
Switch_co             0.301 -0.198 -0.389 0.2809 0.719 2.42
Scheme_value         -0.162 -0.274 -0.250 0.1637 0.836 2.62
Income_level         -0.606  0.313 -0.008 0.4653 0.535 1.50
lessprice_same_.co   -0.028  0.385  0.213 0.1941 0.806 1.57
Change_scheme         0.290 -0.415  0.017 0.2562 0.744 1.79
Advetisement          0.140 -0.188  0.096 0.0639 0.936 2.39
Prepaid_high_profit   0.225  0.381 -0.169 0.2250 0.775 2.06
benefit_big_investor  0.146  0.203 -0.003 0.0627 0.937 1.82
Brand                 0.208  0.277  0.165 0.1473 0.853 2.55
benefit_freq_investor 0.320  0.620 -0.367 0.6210 0.379 2.19
Reasonable_fees       0.501  0.351 -0.249 0.4368 0.563 2.32
High_trade_fee        0.558  0.477 -0.248 0.6003 0.400 2.36
Exp                   0.671 -0.025  0.557 0.7607 0.239 1.94
Age                   0.424  0.076  0.748 0.7446 0.255 1.61

                       PC1   PC2   PC3
SS loadings          2.705 1.946 1.688
Proportion Var       0.142 0.102 0.089
Cumulative Var       0.142 0.245 0.334
Proportion Explained 0.427 0.307 0.266
Cumulative Proportion 0.427 0.734 1.000

Mean item complexity =  2.1
Test of the hypothesis that 3 components are sufficient.
```

Fig 1.56

```
> summary(UnrotatedProfilePCA)

Call:
lm(formula = UnrotatedProfile$Invest_on_Scheme ~ ., data = UnrotatedProfile)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64811 -0.19254 -0.03159  0.17105  0.82637

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22500    0.01961  11.476  < 2e-16 ***
PC1         -0.19544    0.01966  -9.943  < 2e-16 ***
PC2          0.16246    0.01966   8.265 2.07e-14 ***
PC3          0.18692    0.01966   9.510  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2773 on 196 degrees of freedom
Multiple R-squared:  0.5679,    Adjusted R-squared:  0.5613
F-statistic: 85.87 on 3 and 196 DF,  p-value: < 2.2e-16
```

Fig 1.57

5.4.7 Rotated Factor Pattern:

The rotated factor pattern is obtained by **rotating** the factor pattern along the **90-degree** axis. The rotated factor pattern can be used to assign the variables to the suitable factors. The sum of Eigen values is not affected by rotation, but rotation will alter the Eigen values (and percent of variance explained) of particular factors and will change the factor loadings. factor analysis.

```
> summary(RotatedProfilePCA)

Call:
lm(formula = RotatedProfile$Invest_on_Scheme ~ ., data = RotatedProfile)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64811 -0.19254 -0.03159  0.17105  0.82637

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.22500    0.01961  11.476  < 2e-16 ***
RC1          0.27826    0.01966  14.157  < 2e-16 ***
RC2         -0.06749    0.01966  -3.434 0.000726 ***
RC3          0.13247    0.01966   6.739 1.73e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2773 on 196 degrees of freedom
Multiple R-squared:  0.5679,    Adjusted R-squared:  0.5613
F-statistic: 85.87 on 3 and 196 DF,  p-value: < 2.2e-16
```

Fig 1.58

5.4.8 Varimax rotation is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. A varimax solution yields results which make it as easy as possible to identify each variable with a single factor. This is the most common rotation option.

Varimax Rotation

```
Principal Components Analysis
Call: principal(r = sdata_num, nfactors = 3, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                          RC1     RC2     RC3      h2     u2   com
Invest_pref           -0.1438 -0.1174  0.0326 0.03552 0.9645 2.039
brok_fee_influence     0.3894 -0.1407 -0.1683 0.19976 0.8002 1.650
Invest_criteria        0.2406 -0.0361  0.1339 0.07712 0.9229 1.618
Invest_on_Scheme       0.6647 -0.1612  0.3164 0.56792 0.4321 1.567
Trade_type             0.6582 -0.0090 -0.0473 0.43551 0.5645 1.011
Switch_co             -0.4097  0.1871 -0.2794 0.28094 0.7191 2.223
Scheme_value          -0.0864 -0.1724 -0.3557 0.16373 0.8363 1.582
Income_level           0.6579 -0.0855 -0.1586 0.46528 0.5347 1.151
lessprice_same_.co     0.2954  0.1698  0.2793 0.19413 0.8059 2.593
Change_scheme         -0.4806 -0.1576  0.0178 0.25616 0.7438 1.216
Advetisement          -0.2115 -0.1062  0.0889 0.06391 0.9361 1.864
Prepaid_high_profit    0.0465  0.4705  0.0385 0.22498 0.7750 1.033
benefit_big_investor   0.0176  0.2253  0.1078 0.06269 0.9373 1.450
Brand                  0.0413  0.2360  0.2999 0.14730 0.8527 1.939
benefit_freq_investor  0.0973  0.7812 -0.0365 0.62103 0.3790 1.035
Reasonable_fees       -0.1938  0.6282  0.0672 0.43675 0.5632 1.213
High_trade_fee        -0.1567  0.7487  0.1231 0.60025 0.3997 1.144
Exp                   -0.4460  0.0836  0.7449 0.76068 0.2393 1.666
Age                   -0.1669 -0.0571  0.8447 0.74461 0.2554 1.087

                       RC1    RC2    RC3
SS loadings          2.3789 2.1014 1.8580
Proportion Var       0.1252 0.1106 0.0978
Cumulative Var       0.1252 0.2358 0.3336
Proportion Explained 0.3753 0.3315 0.2931
Cumulative Proportion 0.3753 0.7069 1.0000

Mean item complexity =  1.5
Test of the hypothesis that 3 components are sufficient.
```

Fig 1.59

5.4.9 Naming the Factor

S.NO	Factor	Eigen Value
1	**Factor 1: Frequency of trading related variables.**	
i	Invest in this scheme if scheme price is less	**-0.7**
ii	Willing to switch company if scheme offered is attractive	**0.6**
iii	Trading Type	**-0.4**
2	**Factor 2: Experience and Brand Awareness related variables.**	
i.	Investor Experience	**0.9**
ii.	Investor Age	**0.8**
iii.	Brand	**0.3**
3	**Factor 3: Brokerage fees Preference related variables**.	
i.	Income Level	**0.4**
ii.	Reasonable fess	**-0.3**

Tab 1.2

The above-mentioned factors can be named based on the characteristics of the variables lying underneath.

Factor 1 can be named as **Frequency of trading related variables.**
Factor 2 can be named as **Experience and Brand Awareness related variables.**
Factor 2 can be named as **Brokerage fees Preference related variables**.

5.4.10 Factor Scoring Coefficients:

This is a measure of the importance of each variable i.e., how much does a variable influence the measuring factor.

5.4.11 Eigen Values:

Eigen value is a measure of sum of variances of the variables present in a factor. If the Eigen value for a factor is greater than 1, it means that the factor is significant else it can be ignored. It is a measure of the variance of each factor, and if divided by the number of variables (i.e., the total variance), it is the percent of variance summarized by the factor. The Eigen value for a given factor measures the variance in all the variables which is accounted for by that factor. Eigen values measure the amount of variation in the total sample accounted for by each factor.

Calculate Eigenvalues

```
> EigenValue
 [1] 2.7046606 1.9460960 1.6875184 1.5865512 1.4667844 1.2970305 1.1059494
 [8] 1.0142835 0.9238618 0.8623163 0.7589986 0.7062433 0.5839303 0.5653929
[15] 0.4453034 0.4192580 0.3890639 0.3170042 0.2197532
```

Fig 1.60

5.4.12 Scree plot:

The scree criterion may result in fewer or more factors than the Kaiser criterion.

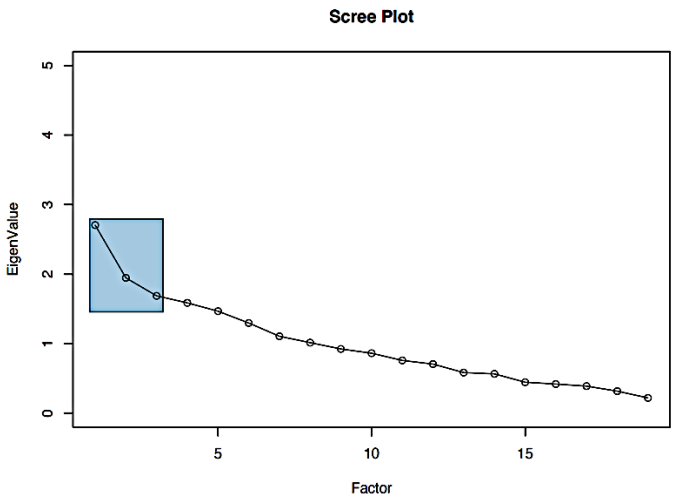


Fig 1.61

5.4.13 Correlation

Correlation is a measure of association between two variables. The variables are not designated as dependent or independent. The value of a correlation coefficient can vary from minus one to plus one. A minus one indicates a perfect negative correlation, while a plus one indicates a perfect positive correlation. A correlation of zero means there is no relationship between the two variables.

Check multicollinearity in independent variables using VIF

VIF Matrix

```
> vifmatrix
          Invest_pref    brok_fee_influence       Invest_criteria
             1.160198              1.254654              1.137558
           Trade_type             Switch_co          Scheme_value
             1.395562              1.307129              1.315867
         Income_level    lessprice_same_.co         Change_scheme
             1.579443              1.301483              1.453016
        Advetisement   Prepaid_high_profit  benefit_big_investor
             1.232146              1.157345              1.375040
                Brand benefit_freq_investor       Reasonable_fees
             1.103396              1.561013              1.720253
       High_trade_fee                   Exp                   Age
             1.795856              2.463827              2.162732
```

Fig 1.62

VIF starts at 1 and has no upper limit, if VIF = 1, no correlation between the independent variable and the other variables VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others

Correlation matrix:

The correlation matrix computes the correlation coefficients of the columns of a matrix. The diagonal elements of the correlation matrix will be 1 since they are the correlation of a column with itself.

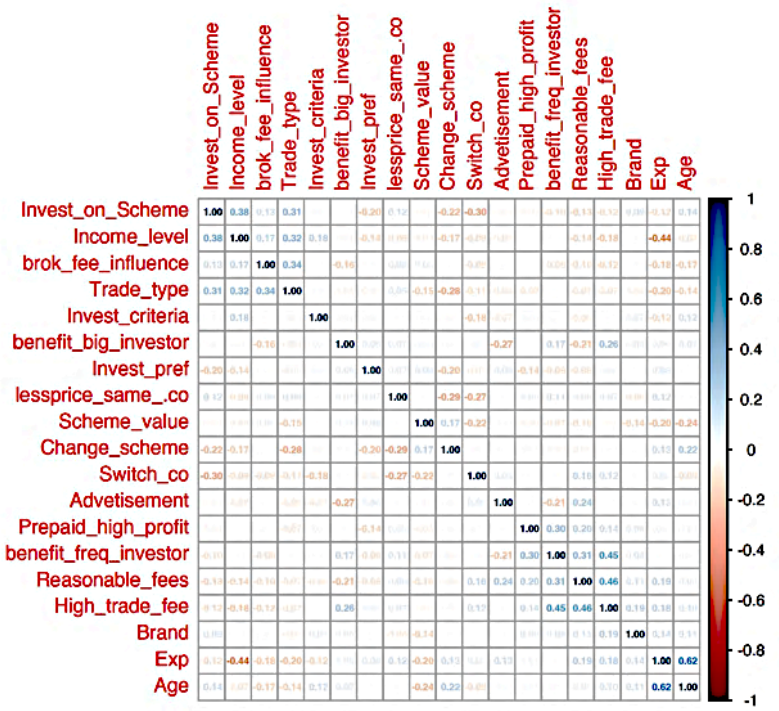


Fig 1.63

```
lm(formula = Invest_on_Scheme ~ ., data = sdata_num)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64279 -0.25629 -0.03175  0.19687  0.76044

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.2361106  0.4835636   0.488 0.625948
Invest_pref          -0.0813555  0.0220872  -3.683 0.000304 ***
brok_fee_influence    0.0063275  0.0229687   0.275 0.783260
Invest_criteria      -0.0365329  0.0184062  -1.985 0.048676 *
Trade_type            0.0571618  0.0250591   2.281 0.023709 *
Switch_co            -0.2101931  0.0638043  -3.294 0.001187 **
Scheme_value          0.0910900  0.0904255   1.007 0.315112
Income_level          0.1088567  0.0332398   3.275 0.001267 **
lessprice_same_.co    0.0366694  0.0510371   0.718 0.473385
Change_scheme        -0.2910332  0.0843082  -3.452 0.000693 ***
Advetisement          0.0015070  0.0317027   0.048 0.962139
Prepaid_high_profit  -0.0074070  0.0373767  -0.198 0.843133
benefit_big_investor -0.0007843  0.0246118  -0.032 0.974613
Brand                 0.0660785  0.0338308   1.953 0.052337 .
benefit_freq_investor -0.0672284  0.0359590  -1.870 0.063157 .
Reasonable_fees      -0.0401794  0.0463352  -0.867 0.387009
High_trade_fee        0.0195228  0.0367357   0.531 0.595765
Exp                  -0.0237652  0.0165426  -1.437 0.152554
Age                   0.0276307  0.0074763   3.696 0.000290 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3417 on 181 degrees of freedom
Multiple R-squared:  0.3939,    Adjusted R-squared:  0.3336
F-statistic: 6.534 on 18 and 181 DF,  p-value: 2.304e-12
```

Fig 1.64

5.4.14 Hypothesis Testing

Hypothesis testing of correlation between Brokerage fees and Trading Type:

The statistical test for the significance of a correlation coefficient is conducted using a t-statistic. The hypothesis to be tested is mentioned below.

Ho: ρ = 0 = No correlation between Brokerage fees and Trading Type.

Ha: ρ =! 0 = Correlation between Brokerage fees and Trading Type.

Level of Significance = **0.05**

n-2 = The degrees of freedom.

Correlations SPSS Output			
		Brokerage fees	**Trading Type**
Brokerage fees	Pearson Correlation	1	**.339****
	Sig. (2-tailed)		.000
	N	200	200
Trading Type	Pearson Correlation	**.339****	1
	Sig. (2-tailed)	.000	
	N	200	200

Tab 1.3

Correlation coefficient = **0.3385651**, t-test for the significance of the coefficient = 4.10

Result:

The table value of t at a **5 %** level of significance is given by 3.39. The corresponding sample value is 4.10. Since the computed values are greater that the corresponding Table values, the null hypothesis is rejected in these cases. Therefore, it can be concluded that there is Positive correlation between Brokerage fees and Trading Type. When there is a positive correlation between two variables, as the value of one variable increases, the value of the other variable also increases. The variables move together.

Hypothesis testing of correlation between investor income level &preference to change scheme:

The statistical test for the significance of a correlation coefficient is conducted using a t-statistic. The hypothesis to be tested is mentioned below.

Ho: ρ = 0 = No correlation between investor income level and their preference to change scheme.
Ha: ρ =! 0 = correlation between investor income level and their preference to change scheme
Level of Significance = **0.05**
n-2 = The degrees of freedom.

Correlations SPSS Output			
		Investor income level	**Preference to change scheme**
Investor income level	Pearson Correlation	1	-.100
	Sig. (2-tailed)		.160
	N	200	200
Preference to change scheme	Pearson Correlation	-.100	1
	Sig. (2-tailed)	.160	
	N	200	200

Tab 1.4

Correlation coefficient = **-0.0996666**, T-test for the significance of the coefficient = 4.45.

Result: The table value of t at a **5 %** level of significance is given by 4.12. Therefore, it can be concluded that there is Negative correlation between investor income level and their preference to change scheme. When there is a negative correlation between two variables, as the value of one variable increases, the value of the other variable decreases, and vice-versa. In other words, for a negative correlation, the variables work opposite each other.

5.4.15 ANOVA:

In a randomized block design, there is only one primary factor under consideration in the experiment. Similar test subjects are grouped into blocks. Each block is tested against all treatment levels of the primary factor at random order. This is intended to eliminate possible influence by other extraneous factors. In the statistical theory of the design of experiments, blocking is the arranging of experimental units in groups (blocks) that are similar to one another. Typically, a blocking factor is a source of variability that is not of primary interest to the experimenter.

1. **Investor's Income Level Verses Trading Type Table:**

Investors Income level	TRADING TYPE				
	Intra-day	Delivery Type	Derivative Futures	Options	Total
Very Higher income investors (>10 Lakhs)	8	8	54	3	73
Higher income investors (6to10Lakhs)	62	12	4	0	78
Middle income investors (4to6Lakhs)	28	3	4	0	35
Lower income investors (<4 Lakhs)	14	0	0	0	14
Total	112	23	62	3	200

Fig 1.5

2. **Hypothesis Testing:**

   A. **Income Level:**

   Ho: Average numbers of investors in all the trading type are same.

   Ha: At least two mean are different.

   B. **Trading Type:**

   Ho: Average numbers of investors in all the income level are same.

   Ha: At least two mean are different.

   Level of Significance = 0.05

The F-values for both testing the null hypotheses of equality of means $H_0 : \mu_1 = \mu_2 = \mu_3$, and equality of block effects, $H_0 : \beta_1 = \beta_2 = \beta_3$ should be rejected at $\alpha = .05$ significance level.

ANOVA TABLE: **Level of Sign: 0.05%**

Sources of Variation	Degree of freedom	Sum of Squares	Mean Square	F
Treatments	3	2808.5	936.16667	4.1261019
Blocks	3	2631.5	887.16667	3.9101371
Error	9	3122	226.88889	
Total	15	5562		

Fig 1.6

3. **Result** : The table value of F at a **5 %** level of significance is given by 3.86. The corresponding sample values for both are 4.126 and 3.91. Since the computed values are greater that the corresponding Table values, the null hypothesis is rejected in both the cases. Therefore, it can be concluded that there is difference in the numbers of investors in the trading type as well as the income level.

6 Findings and Recommendations

1. 50% of the Investors were found to be in the age group of 28-38 years while 25% were in 18-28 years group.
2. The largest segment of Investors with Annual income was found to be in the bracket of 2-5 lakh who were mostly salaried.
3. 54% were Brand conscious Investors since they looked for brands before investing in a Pre-paid model.
4. The company should provide training to its Relationship Managers in order to improve the services also because 4 of the 150 respondents were not satisfied with the service and were considering switching to other firms
5. The company should provide training to its Relationship Managers in order to improve the services also because 4 of the 150 respondents were not satisfied with the service and were considering switching to other firms.
6. LKP Shares can push Investors to trade daily which will result in profit to the Investors and brokerage for the company because 50% of the Investors were investors who would not trade daily.

7 Predictive Model Building

Let factor the required variables for the purpose of model building

7.1 Feature Engineering

```
> str(sdata_num)
'data.frame':   200 obs. of  19 variables:
 $ Invest_pref          : Factor w/ 5 levels "1","2","3","4",..: 5 5 3 5 3 5 5 3 5 1 ...
 $ brok_fee_influence   : Factor w/ 5 levels "1","2","3","4",..: 2 2 3 5 4 4 4 4 4 3 ...
 $ Invest_criteria      : Factor w/ 5 levels "1","2","3","4",..: 1 1 4 1 1 2 1 4 1 3 ...
 $ Invest_on_Scheme     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Trade_type           : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 2 4 1 1 1 3 ...
 $ Switch_co            : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 2 2 2 2 ...
 $ Scheme_value         : Factor w/ 2 levels "1","2": 1 1 1 2 1 1 1 1 2 1 ...
 $ Income_level         : Factor w/ 4 levels "1","2","3","4": 2 1 2 1 2 1 4 1 4 4 ...
 $ lessprice_same_.co   : Factor w/ 3 levels "1","2","3": 2 2 1 1 1 2 1 2 2 1 ...
 $ Change_scheme        : Factor w/ 3 levels "1","2","3": 2 1 2 2 2 2 2 2 2 2 ...
 $ Advetisement         : Factor w/ 5 levels "1","2","3","4",..: 4 5 5 5 5 5 5 5 5 5 ...
 $ Prepaid_high_profit  : Factor w/ 4 levels "2","3","4","5": 3 3 4 3 3 3 3 4 3 4 ...
 $ benefit_big_investor : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 2 1 3 2 4 2 1 ...
 $ Brand                : Factor w/ 4 levels "1","3","4","5": 3 3 4 3 4 3 4 4 3 4 ...
 $ benefit_freq_investor: Factor w/ 4 levels "2","3","4","5": 3 2 2 1 2 1 3 2 2 3 ...
 $ Reasonable_fees      : Factor w/ 4 levels "2","3","4","5": 4 4 3 4 3 4 3 4 3 4 ...
 $ High_trade_fee       : Factor w/ 4 levels "2","3","4","5": 3 3 2 3 1 3 2 4 3 2 ...
 $ Exp                  : num  15 8 3 4 9 7 3 12 4 5 ...
 $ Age                  : num  44 29 26 28 32 32 28 38 32 30 ...
```

Fig 1.65

Check ratio of target value

```
> summary(sdata$Invest_on_Scheme)
  0   1
155  45
```

As we see from above data distribution that most of the investor are not preferring to invest on the Prepaid scheme. Let split the dataset into train and test dataset in an proportion of 80:20.

Split set 1:

```
> #Compare response variable levels in train & test
> round(prop.table(table(train1$Invest_on_Scheme))*100, digits = 1)

   0    1
77.1 22.9
> round(prop.table(table(test1$Invest_on_Scheme))*100, digits = 1)

   0    1
78.3 21.7
```

Data balance of Train Split 1:

```
Classes: 2
  0   1
108  32
Positive class: 1
```

This shows imbalance in data which will create a bias so let balance the data using SMOTE technique

After SMOTE technique on Train Split 1:

```
Classes: 2
  0   1
108  96
Positive class: 1
```

7.2 Building KNN model 1

Model 1 in KNN using Spilt set 1

```
> knn.cm1 #Accuracy = 88%, sensitivity = tpr = 77%, specificity = 91%
    predicted
true 0         1
   0 43        4         tpr: 0.77 fnr: 0.23
   1 3         10        fpr: 0.09 tnr: 0.91
     ppv: 0.71 for: 0.07 lrp: 9.04 acc: 0.88
     fdr: 0.29 npv: 0.93 lrm: 0.25 dor: 35.83
```

Model 2 in KNN using Split set 2

Data Balance on Spilt set 2

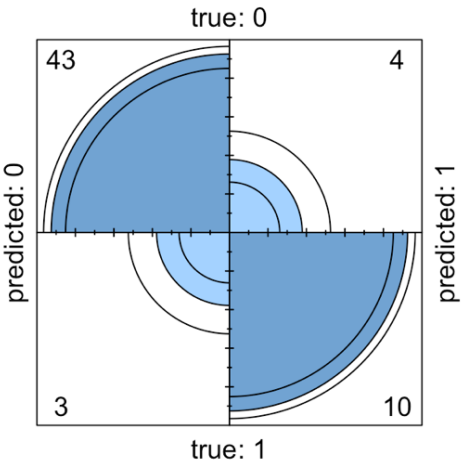
```
> round(prop.table(table(train2$Invest_on_Scheme))*100, digits = 1)

   0    1
77.1 22.9
> round(prop.table(table(test2$Invest_on_Scheme))*100, digits = 1)

   0    1
78.3 21.7
```

7.2.1 Interpretation of KNN Model 1

In our business case we need to consider both accuracy and Sensitivity. Since we need to identify the customers who may invest to avoid unnecessary marketing spends, we shall give relative importance to Sensitivity.

KNN model 1

```
> knn.cm1 #Accuracy = 88%, sensitivity = tpr = 77%, specificity = 91%
    predicted
true 0         1
   0 43        4         tpr: 0.77 fnr: 0.23
   1 3         10        fpr: 0.09 tnr: 0.91
     ppv: 0.71 for: 0.07 lrp: 9.04 acc: 0.88
     fdr: 0.29 npv: 0.93 lrm: 0.25 dor: 35.83
```




```
> performance(knn.predict1,auc)
      auc
0.893617
```

Fig 1.66

This model has a good accuracy and sensitivity of 77% which is good performance for a model

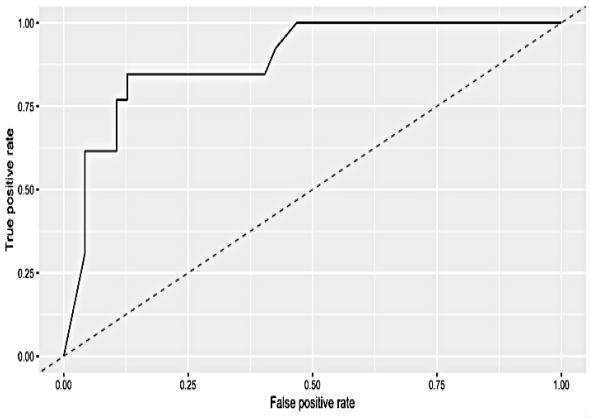


Fig 1.67

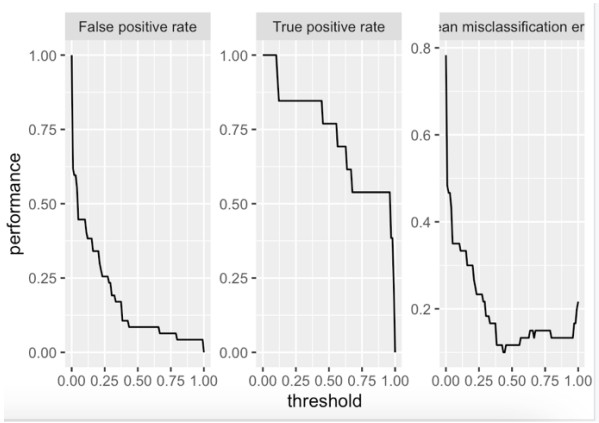
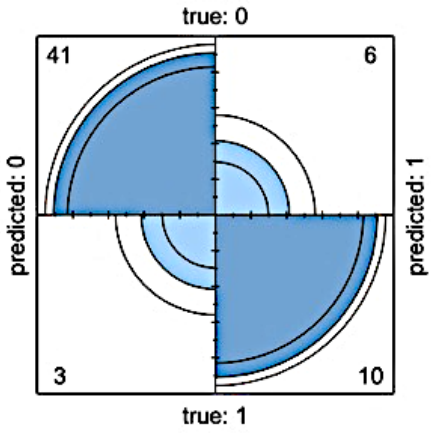


Fig 1.68

7.2.2 Interpretation of KNN Model 2

```
> knn.cm2 #Accuracy = 85%, sensitivity = tpr = 77%, specificity = 85%
    predicted
true 0         1
   0 41        6          tpr: 0.77 fnr: 0.23
   1 3         10         fpr: 0.13 tnr: 0.87
     ppv: 0.62 for: 0.07 lrp: 6.03 acc: 0.85
     fdr: 0.38 npv: 0.93 lrm: 0.26 dor: 22.78
```




```
> performance(knn.predict2,auc)
      auc
0.9638219
```

Fig 1.69

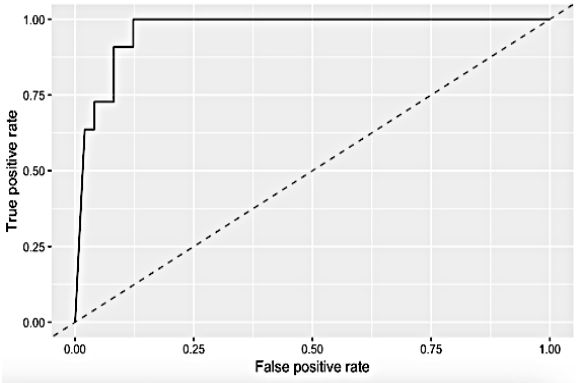


Fig 1.70

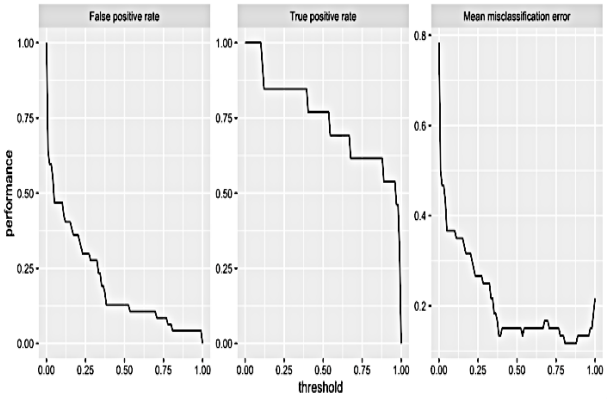


Fig 1.71

The AUC of KNN model 1 has better performance than KNN Model2. Also, model predicted well with Accuracy of 88%, sensitivity of 77% and specificity of 91% which is good.

7.3 Interpretation of Naïve Bayes Model

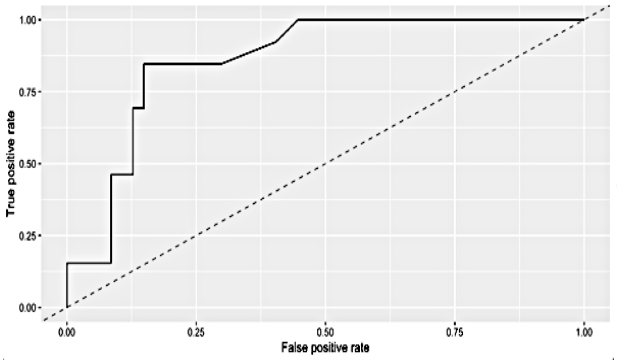


Fig 1.72

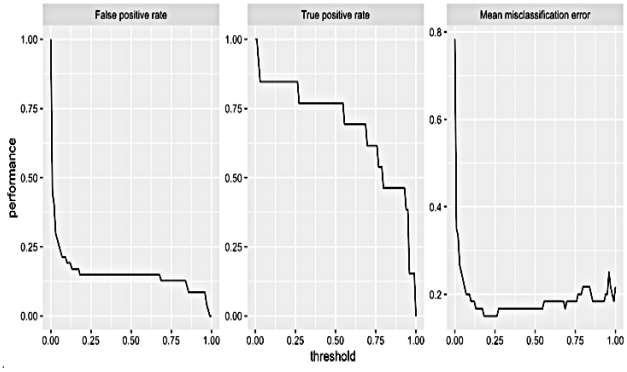
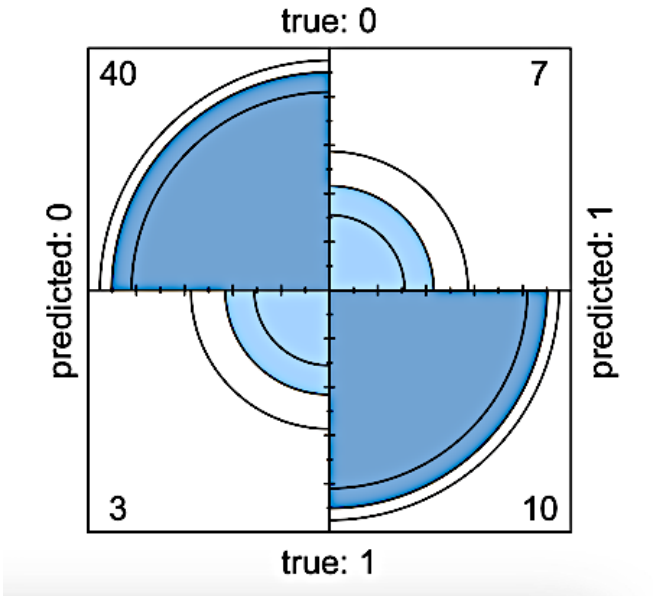


Fig 1.73

Accuracy of Naïve Bayes model is less compared to the KNN models but it has good sensitivity similar to other models.

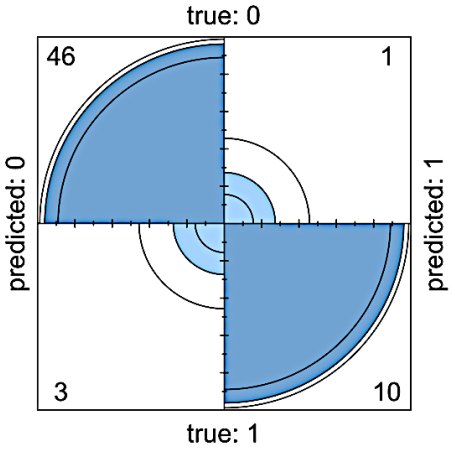



```
> nb.cm1
    predicted
true 0         1
   0 40        7          tpr: 0.77 fnr: 0.23
   1 3         10         fpr: 0.15 tnr: 0.85
     ppv: 0.59 for: 0.07 lrp: 5.16 acc: 0.83
     fdr: 0.41 npv: 0.93 lrm: 0.27 dor: 19.05
```

Fig 1.74

7.4 Interpretation of **CART Model**

Classification and Regression Trees (CART) models can be implemented by using the rpart package in R. The recursive structure of CART models is ideal for uncovering complex dependencies among predictor variables. If a response variable depends strongly on a predictor variable in a nonlinear fashion, then a CART model will be better at detecting this relationship




```
> cart.cm
    predicted
true 0         1
   0 46        1         tpr: 0.77  fnr: 0.23
   1 3         10        fpr: 0.02  tnr: 0.98
     ppv: 0.91 for: 0.06 lrp: 36.15 acc: 0.93
     fdr: 0.09 npv: 0.94 lrm: 0.24  dor: 153.33
```

```
> performance(cart.predict,auc)
      auc
0.9844517
```

Fig 1.75

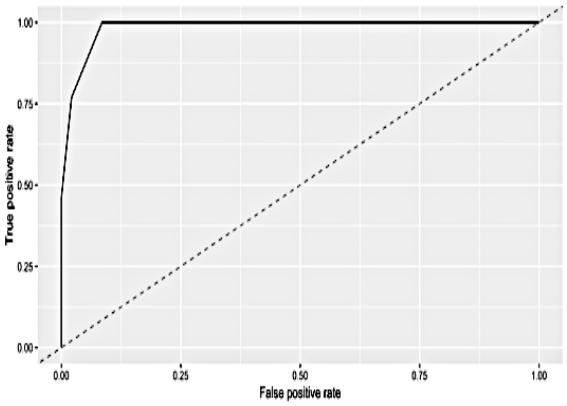


Fig 1.76

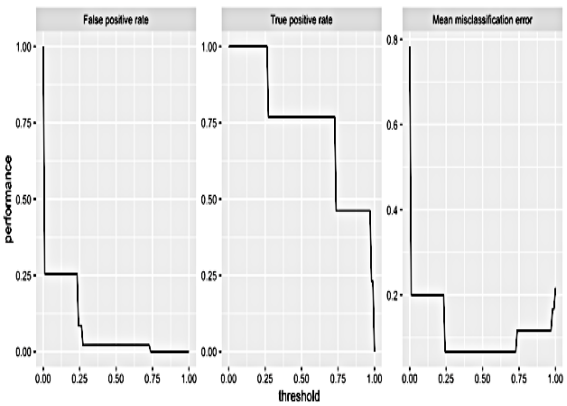
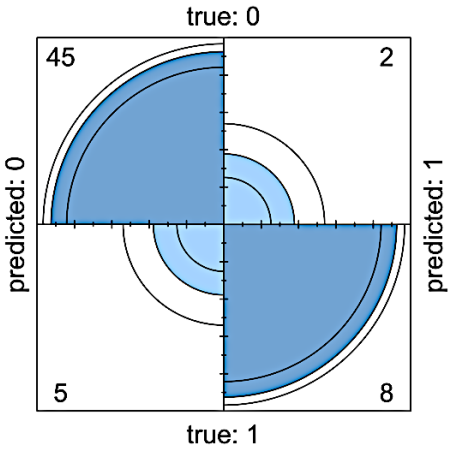


Fig 1.77

CART model gives better accuracy compared to the above two models which is 93%. And also, the sensitivity is goof with 77%.

7.5 Interpretation of **Random Forest Model**

Random forest approach is used over decision trees approach as decision trees lack accuracy and decision trees also show low accuracy during the testing phase due to the process called over-fitting.




```
> rf.cm
     predicted
true 0        1
   0 45       2        tpr: 0.62  fnr: 0.38
   1 5        8        fpr: 0.04  tnr: 0.96
     ppv: 0.8 for: 0.1 lrp: 14.46 acc: 0.88
     fdr: 0.2 npv: 0.9 lrm: 0.4   dor: 36
```

```
> performance(rf.pred,auc)
      auc
0.9639935
```

Fig 1.78

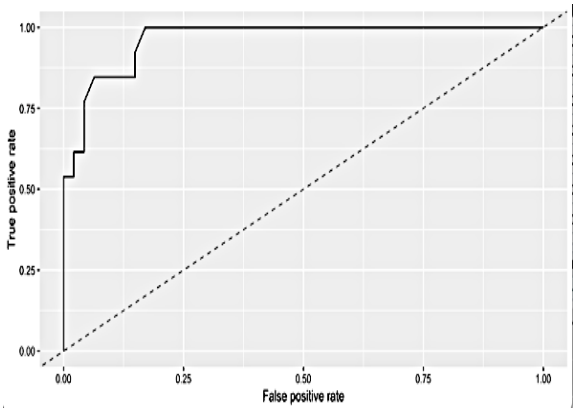


Fig 1.79

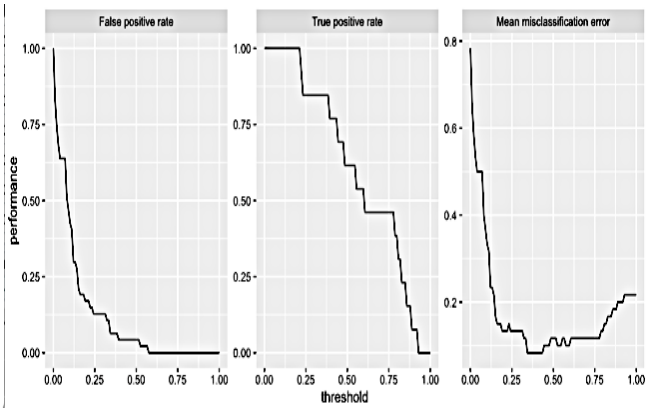
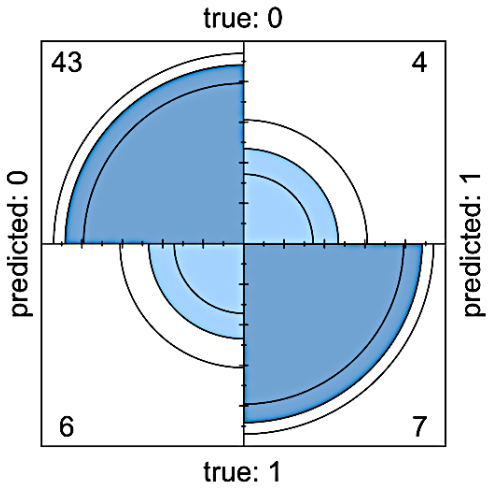


Fig 1.80

Random forest model generates better accuracy than naïve bayes model but not higher than the CART model. Performance of Sensitivity seems to be good.

7.6 Interpretation of **GBM Model**

This method is to improve (boost) the week learners sequentially and increase the model accuracy with a combined model.




```
> gbm.cm
    predicted
true 0         1
   0 43        4         tpr: 0.54 fnr: 0.46
   1 6         7         fpr: 0.09 tnr: 0.91
     ppv: 0.64 for: 0.12 lrp: 6.33 acc: 0.83
     fdr: 0.36 npv: 0.88 lrm: 0.5  dor: 12.54
```

```
> performance(gbm.pred,auc)
      auc
0.8903437
```

Fig 1.81

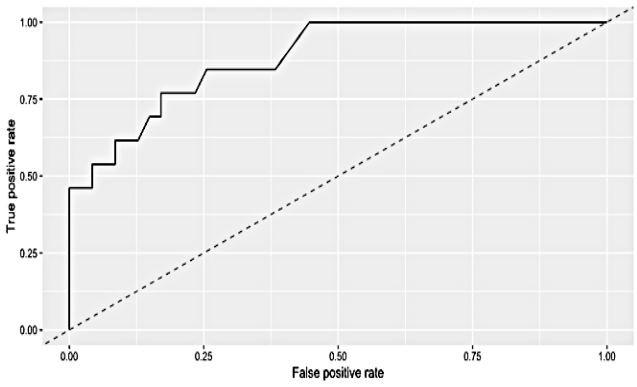


Fig 1.82

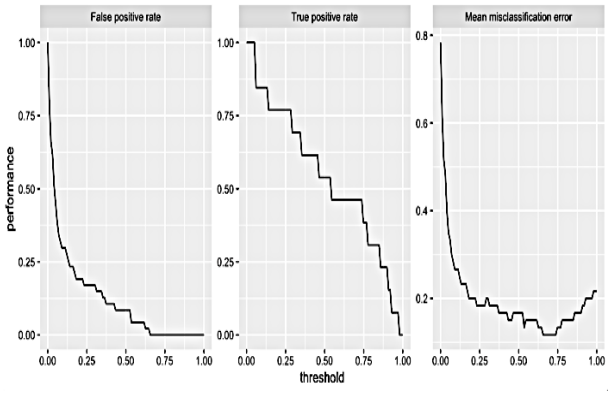
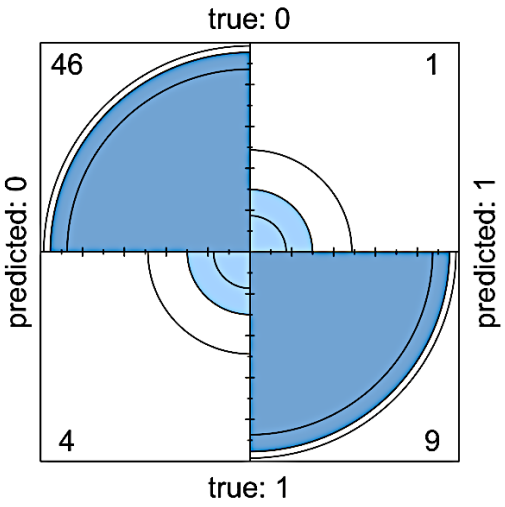


Fig 1.83

The performance of GBM is not good as compared to any other models we previously generated. Sensitivity is poor el which is only 54%

7.7 Interpretation of **SVM Model**

SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces. classification is performed by finding the hyper-plane that best differentiates the two classes.




```
> svm.cm
    predicted
true 0        1
   0 46       1         tpr: 0.69  fnr: 0.31
   1 4        9         fpr: 0.02  tnr: 0.98
     ppv: 0.9 for: 0.08 lrp: 32.54 acc: 0.92
     fdr: 0.1 npv: 0.92 lrm: 0.31  dor: 103.5
```

```
> performance(svm.pred,auc)
      auc
0.9345336
```

Fig 1.84

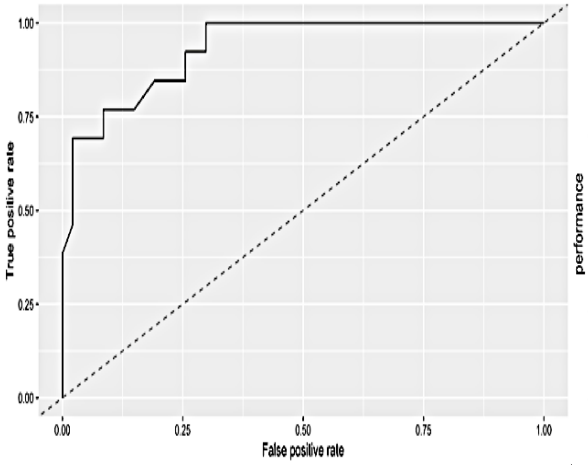


Fig 1.85

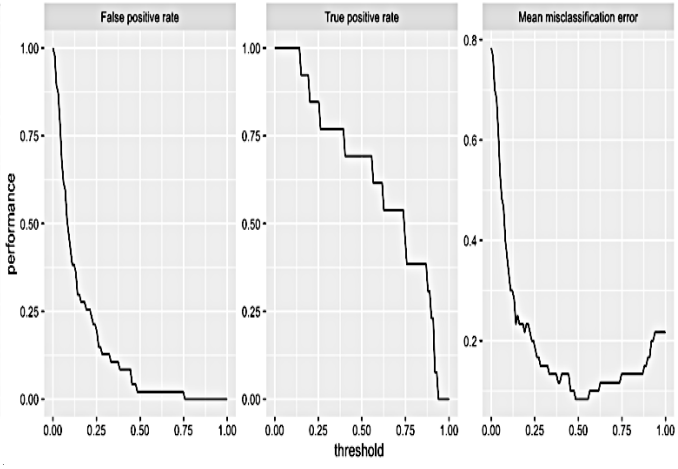
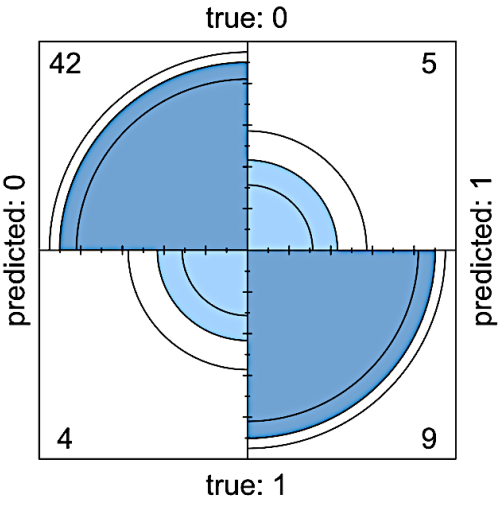


Fig 1.86

SVM has a good accuracy of 92% which looks to be a good model but with lower sensitivity of 69%.

7.8 Interpretation of **Lasso Model**

"LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is good for models showing high levels of multicollinearity or when you want to automate certain parts of model selection i.e. variable selection or parameter elimination. Lasso regression is a classification algorithm that uses shrinkage in simple and sparse models(i.e model with fewer parameters).




```
> glmnet.cm
    predicted
true 0         1
   0 42        5         tpr: 0.69 fnr: 0.31
   1 4         9         fpr: 0.11 tnr: 0.89
     ppv: 0.64 for: 0.09 lrp: 6.51 acc: 0.85
     fdr: 0.36 npv: 0.91 lrm: 0.34 dor: 18.9
```

```
> performance(glmnet.pred,auc)
      auc
0.9050736
```

Fig 1.87

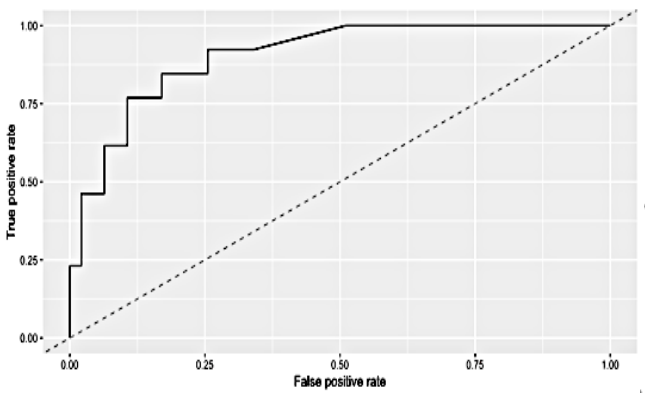


Fig 1.88

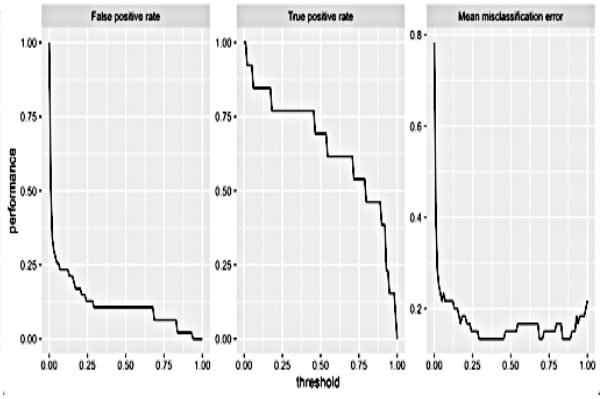


Fig 1.89

Lasso model gives the accuracy of 85% which is less compared the other models but less sensity of 69%

7.9 Model Comparison

Comparision within Model Performances			
Model Name	Accuracy	sensitivity	specificity
KNN	88%,	77%,	91%
KNN	85%,	77%,	87%
Naïve Bayes	83%,	77%,	85%
CART	93%,	77%,	98%
Random Forest	88%,	62%,	96%
GBM	83%,	54%,	91%
SVM	92%,	69%,	98%
Lasso	85%,	69%,	89%

Tab 1.7

From all the above model CART has better performance in terms of Sensitivity, Accuracy and Specificity.

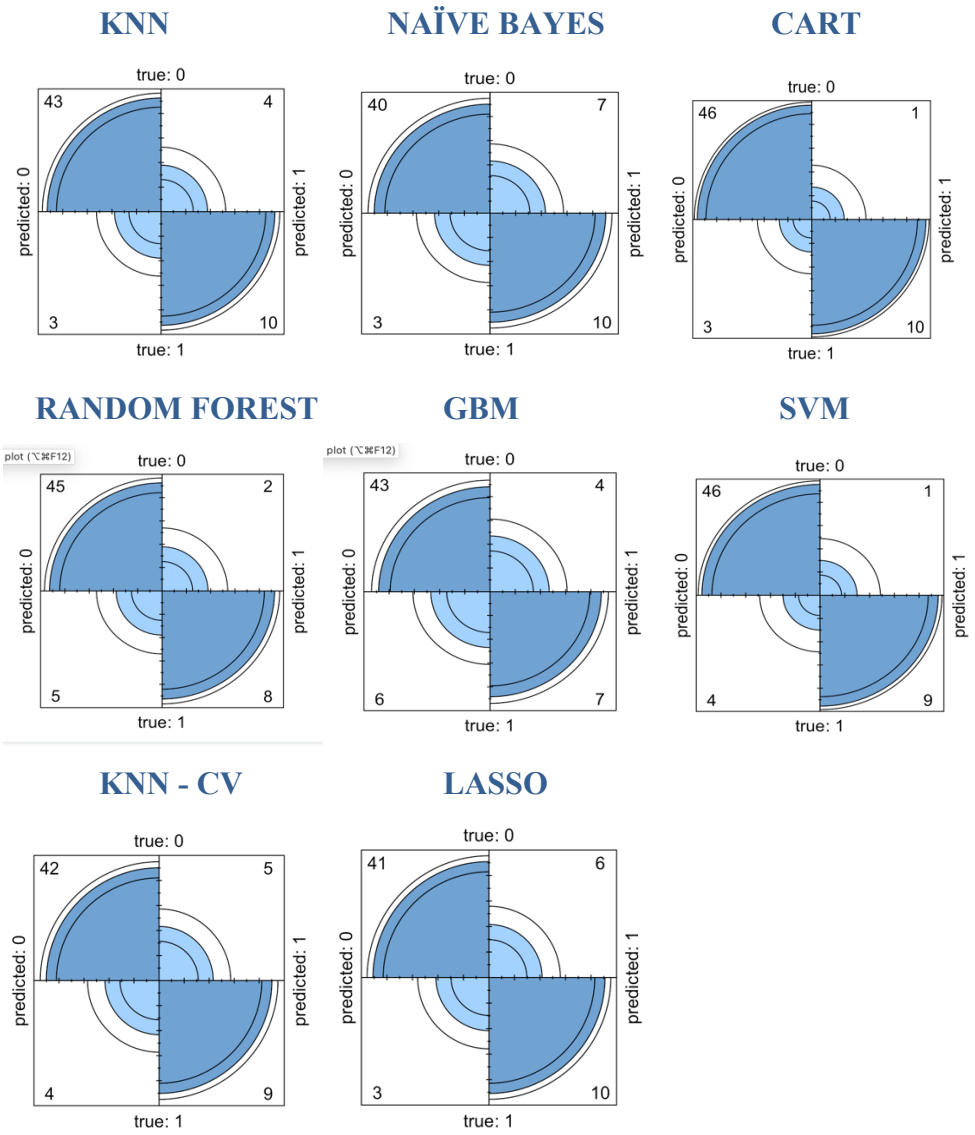


Fig 1.90

8 References and Bibliography

Books:

1. Regression Modelling Strategies With applications to Linear Models, Logistics and Ordinal. Regression and Survival Analysis. Frank E Harrell Jr. 2015 publication

2. Mastering Predictive Analytics using R. Second Edition- James D. Miller, Rui Miguel Forte

3. Weiss Bach, S., 1997. Using the Internet for quantitative survey research, Quirk's Marketing Research Review. 43 – 49.

Website:

1. Lucas, Samuel R., 2012. LKP merger with Kingfishers shares. [Online]. Available from: http://economictimes.indiatimes.com/topic/LKP-Securities-Ltd [Accessed 12 June 2012]

2. Chambers R L., 2005. LKP Shares in Nifty. [Online]. Available from: www.niftydirect.com/nsebse/market-gyan/Learning%20Session%205th.pdf [Accessed 21 March 2005]

3. Graubard, B.I.. 2009. Factor Analysis. [Online]. Available from: www.statisticshell.com/docs/factor.pdf [Accessed 04 July 2009]

4. Tyrrell, Sidney., 2002.SPSS for Analysis part. [Online]. Available from: https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php [Accessed 11 April 2002]

5. https://www.lkpsec.com/research/market-research.aspx

9 Appendix

Below are the libraries which we used in the R

```
library(ggplot2)
library(mlr)
library(dplyr)
library(explore)
library(parameters)
library(PerformanceAnalytics)
library("ggpubr")
```

Raw codes attached in separate file. Output projected in respective figures.