

Customer Behavioral Segmentation and
Recharge Prediction of Telecom Company of
Pakistan Using Machine Learning
Algorithms

By

Muhammad Fahad Arjad
Moazam Ali
Zain Imran



Bachelors in Information & Technology

Department of Computing
School of Electrical Engineering & Computer Science
National University of Sciences & Technology
Islamabad, Pakistan
2013



SEECSP01910

“Customer Behavioral Segmentation and Recharge Prediction of Telecom Company of Pakistan Using Machine Learning Algorithms”

By

Muhammad Fahad Amjad (2009-NUST- BIT-262)

Moazam Ali (2009-NUST- BIT-258)

Zain Imran (2010-NUST- BIT-414)



SEEGS LIBRARY

Project documentation submitted in partial fulfillment of the requirements for the degree of

Bachelors in Information & Technology (BIT)

Department of Computing

School of Electrical Engineering & Computer Science

National University of Sciences & Technology

Islamabad, Pakistan 2013

Certificate

It is certified that the contents and form of this Report entitled "Customer Behavioral Segmentation and Recharge Prediction of Telecom Company of Pakistan using Machine Learning Algorithm" submitted by "Moazam Ali (2009-NUST-SEECS-BIT-258)", "M Fahad Amjad (2009-NUST-SEECS-BIT-262)" and "Zain Imran (2010-NUST-SEECS-BIT-414)" have been found satisfactory for the requirements of the degree.

Advisor: Ali Mustafa 12/6/14

(Dr. Ali Mustafa Qamar)

Co. Advisor: Muneeb Ullah 17/10/14

(Dr. Muneeb Ullah)

DEDICATION

To Allah the Almighty

&

To our Parents and Faculty

ACKNOWLEDGEMENTS

We are deeply thankful to our Advisor and Co-Advisors, Dr. Ali Mustafa Qamar and Dr. Muneeb Ullah for helping us throughout the course in accomplishing our final project. Their guidance, support and motivation enabled us in achieving the objectives of the project.

We are also thankful to our mentors from IBM, Mr. Salar Masood and Mr. Ahsan Rehman for their help and valuable feedback at the crucial stages of the project.

Table of Contents

1.0 Introduction:	7
1.1 Abstract:	7
1.2 Introduction to Data analytics in Telecommunication Industry:	7
1.3 Problem Statement:	8
1.4 What is Data mining?:	8
1.5 Data Analytics:	8
1.6 Importance of Data Analytics:	9
1.7 Applications in Telecom Industry of Data Analytics:	9
1.8 Techniques of Data Analytics:	9
1.9 Project Objectives:	9
1.10 Requirements Analysis:	10
2.0 Literature Review:	11
3.0 Methodology:	12
3.1 Data Collection:	12
3.2 Data Attributes:	13
4.0 Exploratory Data Analysis (EDA):	14
4.1 Definition:	14
4.2 EDA of Attributes We Used:	14
4.2 Overall EDA of our Selected Attributes:	20
4.3 Steps Involved in EDA:	20
4.4 Data Details:	20
5.0 UnSupervised Learning:	21
5.1 Pre-Processing:	21
6.0 Clustering & Customer Segmentation:	23
6.1 Clustering Algorithms:	23
6.2 Definition of Clusters:	28
6.3 Graphical Representation of Clusters for all 6 months:	29
7.0 Inter Cluster Migrations:	29
7.1 Migration Matix:	29
7.2 Migrations Report:	30
7.3 Graphical Representation of Migrations:	33
7.4 Analyzing Migration Trends to Month 6 & Marketing Strategies:	33
8.0 Supervised Learning:	34

8.1 Types of Supervised Learning:	34
9.0 Data Preprocessing:	36
9.1 Data Cleansing:	36
9.2 Feature Selection:.....	37
9.3 Selecting Attributes:.....	37
9.4 Partitioning of data:.....	39
10.0 Modeling:.....	40
10.1 Steps involved:.....	40
11.0 Algorithms Used for classification:	42
11.1 C 5.0 Decision tree algorithm:	42
11.2 Neural Networks Algorithm:	46
11.3 Discriminant algorithm:.....	48
11.4 Comparison:.....	49
12.0 Formation of BCG matrix:.....	50
13.0 Future Work:.....	51
14.0 Difficulties Faced:.....	52
15.0 References:.....	52

1.0 Introduction:

1.1 Abstract:

Pakistan hosts the world's largest and most experienced telecom companies. The number of mobile subscribers has reached 123 million, with more than 90% of country having cellular services, and with a tele density of over 62% as of January 2013.

In our project we first perform customer segmentation using different algorithms and analyzing customer behavior over a period of 6 months, followed by recharge prediction using IBM SPSS and finally creating reports.

Customer segmentation is to classify customers into different groups according to one or more attributes. Due to a fierce competition between different telecom companies, to generate maximum revenue and to retain their customers requires the need for a company to better analyze their customers for producing optimal price plans and reducing churn.

Moreover, we created a recharge prediction model to understand how the business will perform in the future.

Finally, BI Reports were generated to analyze Key performance indicators and help companies evaluate previously-made decisions and answer questions through layout-led and data-led discovery and data mining.

1.2 Introduction to Data analytics in Telecommunication Industry:

As Pakistan holds competitive and fluid telecom industry, creating profit and loyal customers is one of the major targets of the telecom companies. For this purpose companies must understand their customers. In order to fulfill this need data mining technique clustering is used.

Clustering classifies customers into different groups (clusters) on the basis of selected attributes which is helpful in separating customers and dividing them into similar groups. So different and selected pricing plans can be offered to customers of different groups (clusters) to reduce churn, retain a good customer base and increase profits.

In this research we will be using K-means clustering algorithm to classify real time data using attributes ; customers total revenue, voice call revenue, voice call duration, voice call on net duration, voice call off net duration, voice call off net revenue, sms revenue, sms frequency and

recharge amount. We selected K-means algorithm because it produced better results compared to other algorithms such as 2-step clustering which we also tested on the data.

Furthermore we will also be predicting the recharge of the customer using a prediction model.

Lastly useful insights will be presented using BI reports to better understand the trends and outcomes of the model on the basis of which decisions will be made to increase profits, reduce churn and retain the customers of the company.

1.3 Problem Statement:

Customer Behavioral Segmentation and Recharge Prediction of Telecom Company of Pakistan Using Machine Learning Algorithms. We were provided with real time data of a telecommunication company on which we had to apply machine learning algorithms to extract useful information and patterns and generate reports.

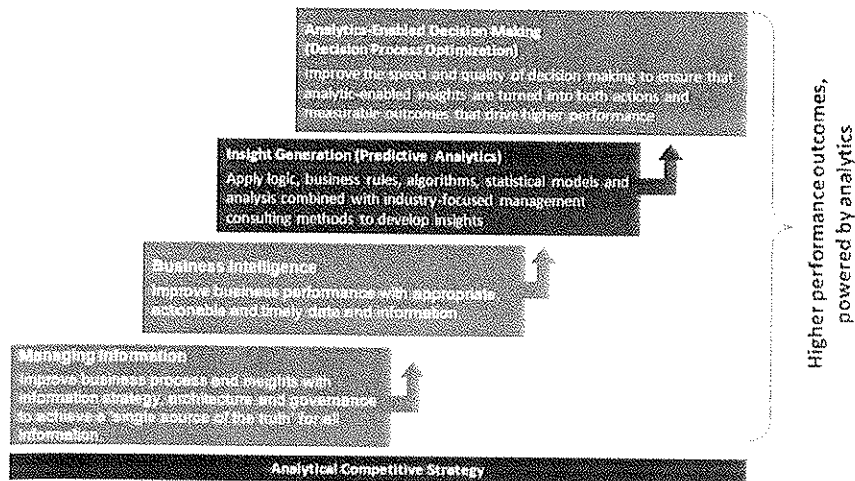
1.4 What is Data mining?

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. (Wikipedia)

1.5 Data Analytics:

Analytics is the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, data mining techniques and operations research to quantify performance. Analytics often favors data visualization to communicate insight. (Wikipedia)

1.6 Importance of Data Analytics:



1.7 Applications in Telecom Industry of Data Analytics:

- Customer Profiling and Segmentation.
- Churn Prediction.
- Identifying fraudulent activity.
- Prediction Models.
- Network analysis.

1.8 Techniques of Data Analytics:

Following are some techniques used in our project:

- Clustering
- Prediction Model

We used following tools in our project:

- WEKA
- IBM SPSS Statistics
- IBM SPSS Modular
- ORACLE 10g
- Google Refine

1.9 Project Objectives:

Following are the objectives of our project.

- Literature review.
- Requirements analysis.
- Data Representation.

- Data Preprocessing.
- Clustering.
- Revenue Prediction Model.

1.10 Requirements Analysis:

Following are the modules of the project:

1. Data preparation/data pre-processing
 - Aggregate data (e.g. daily sales into weekly etc)
 - Sampling of data
 - Dimensionality reduction
 - Feature selection
 - Discretizing (switching from real to integer values)
 - Attribute transformation (from old attributes to new attributes)
2. Defining the study (what is to be mined)
 - Understanding limits(set of problems faced by a user)
 - Choosing an appropriate study
 - Types of studies
 - Selection of elements for analysis
 - Issue of sampling
 - Reading the data and building the model
3. Model accuracy
 - Power of model to provide correct and reliable information.

4. Model intelligibility

- Refers to the characteristic of being easily understood by different people with different degrees/types of training.

5. Performance of Model

- The performance of a data mining model defined by both the time needed to be built and its speed of processing data in order to provide a prediction.

6. Noise handling

- Each model has a threshold of tolerance to noise and this is one of the reasons for an initial data pre-processing stage.

7. Clustering/Segmentation

- Grouping customers into distinct groups based on some attributes.

8. Recharge Prediction Model

- Predicting customer recharge based on ground realities.

9. Financial and Operational Reporting

- Generating Financial and Operational reports for the organization so that the organization can analyze its performance and can plan what to do.

2.0 Literature Review:

1. Customer Segmentation and Analysis of a Mobile Telecommunication Company of Pakistan using Two Phase Clustering algorithm

- Binning
- Two-Phase algorithm for clustering

- Mapping Clusters on Bins
- Marketing Strategies

2. Case study on cluster analysis of the telecom customers based on consumers' behavior.

- Kohonen clustering algorithm to generate customer types
- The clustering results were arranged to the eight quadrants

3. Prepaid Telecom Customers Segmentation Using The K-Mean Algorithm

- Used K-Means Clustering
- Created 7 Clusters, depending on the recharge amount
- BI Report

3.0 Methodology:

3.1 Data Collection:

- We gathered our Dataset from 3rd Party company
- It's a real time Data of Telecom company of Pakistan
- 40000 Subscribers Data for 6 Months
- 3 General attributes for whole 6 Months
- 17 Attributes for each Month
- Usage, Revenue & Recharge Data

3.2 Data Attributes:

General Attributes:

- SUBSCRIBER_ID,
- TENURE,
- ARPU,

Month 1 Attributes:

- M1_TOT_REVENUE,
- M1_CALL_DUR,
- M1_CALL_FREQ,
- M1_CALL_REV,
- M1_CALL_ONNET_DUR,
- M1_CALL_ONNET_FREQ,
- M1_CALL_ONNET_REV,
- M1_CALL_OFFNET_DUR,
- M1_CALL_OFFNET_FREQ,
- M1_CALL_OFFNET_REV,
- M1_SMS_FREQ,
- M1_SMS_REV,
- M1_CALL_FIXLINE_DUR,
- M1_CALL_FIXLINE_FREQ,
- M1_CALL_FIXLINE_REV,
- M1_NO_OF_RECHARGES,
- M1_RECHARGES_AMT

Note: All 6 months have attributes similar to the attributes mentioned in **Month 1 Attributes** section.

4.0 Exploratory Data Analysis (EDA):

4.1 Definition:

In statistics, **exploratory data analysis (EDA)** is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

4.2 EDA of Attributes We Used:

Month 1 Total Revenue:

Selected attribute

Name: M1_TOT_REVENUE

Missing: 0 (0%)

Distinct: 8987

Type: Numeric

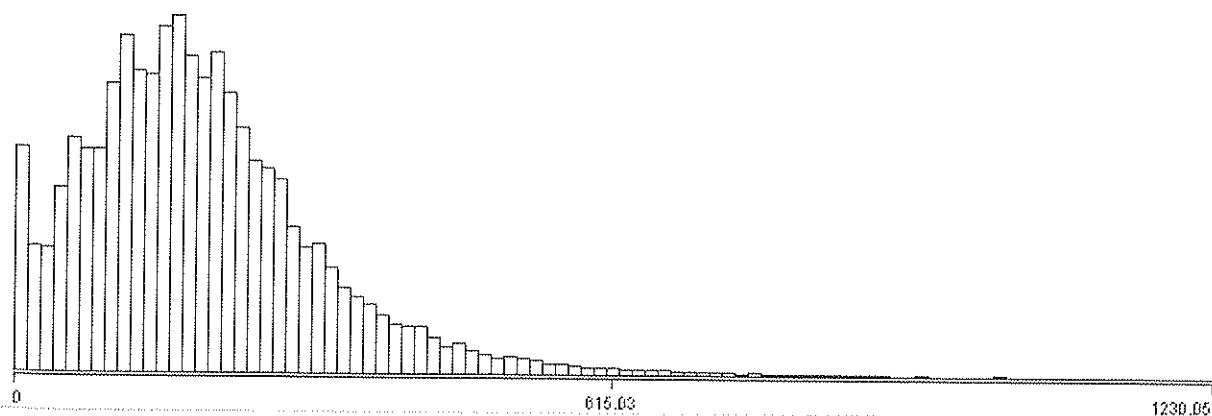
Unique: 2996 (7%)

Statistic	Value
Minimum	0
Maximum	1230.05
Mean	193.278
StdDev	129.691

Class: M1_RECHARGES_AMT (Num)



Visualize All



Outgoing Voice Call Revenue:

Selected attribute

Name: M1_OG_VOICE_CALL_REV

Missing: 0 (0%)

Distinct: 8406

Type: Numeric

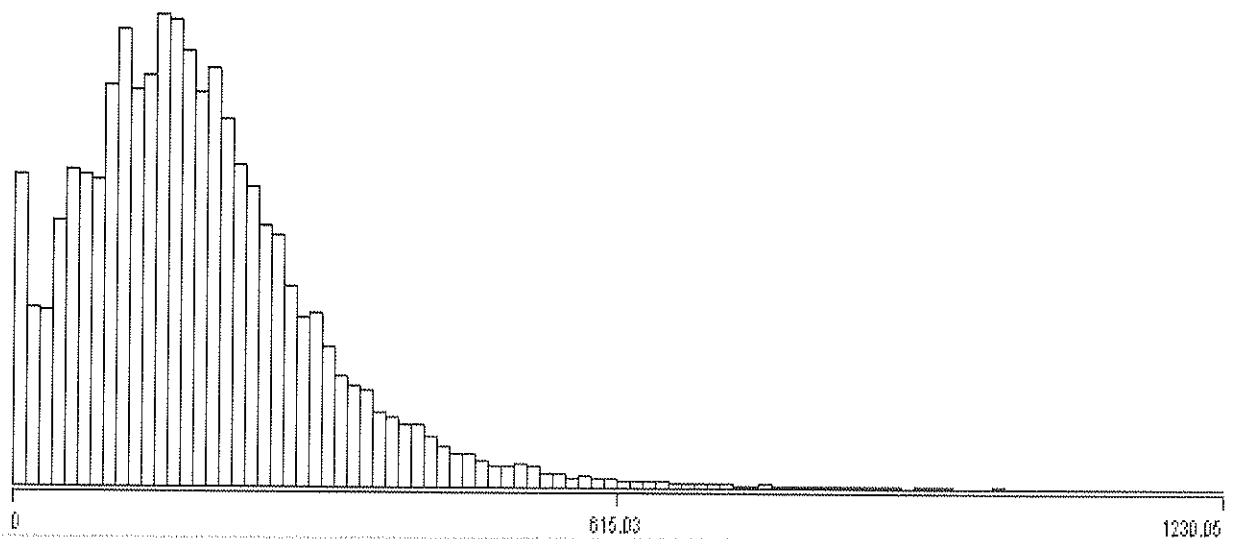
Unique: 3072 (8%)

Statistic	Value
Minimum	0
Maximum	1230.05
Mean	189.434
StdDev	128.148

Class: M1_RECHARGES_AMT (Num)



Visualize All



Outgoing On-net Voice Calls Revenue:

Selected attribute

Name: M1_OG_VOICE_CALL_ONNET_REV

Missing: 0 (0%)

Distinct: 4665

Type: Numeric

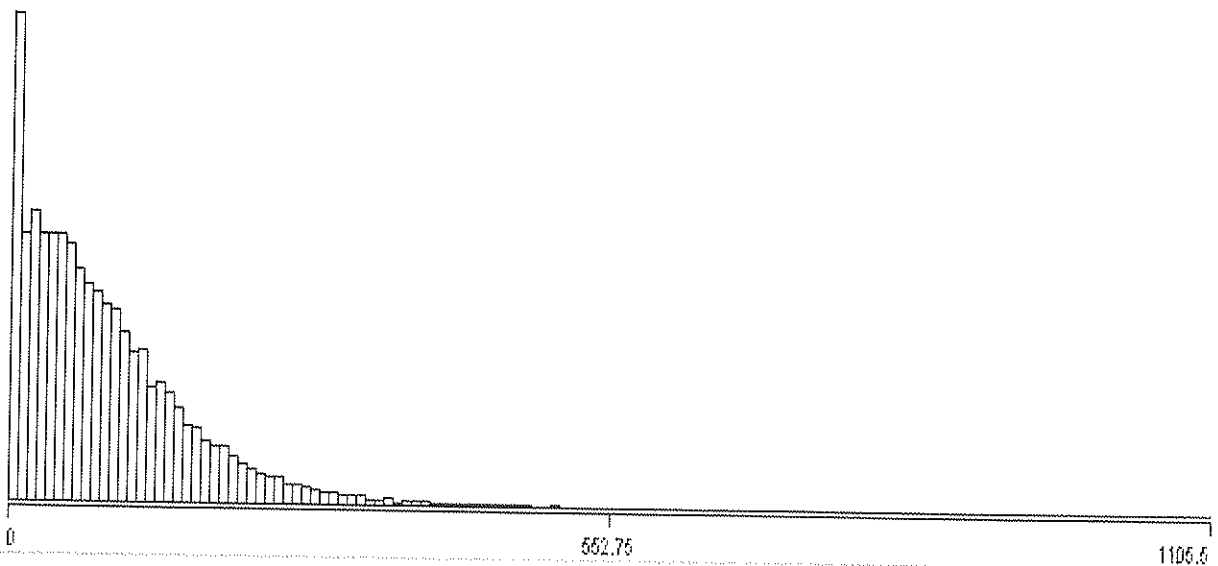
Unique: 1936 (5%)

Statistic	Value
Minimum	0
Maximum	1105.5
Mean	89.689
StdDev	82.997

Class: M1_RECHARGES_AMT (Num)



Visualize All



Outgoing Voice Call Off-net Revenue:

Selected attribute

Name: M1_OG_VOICE_CALL_OFFNET_REV

Type: Numeric

Missing: 0 (0%)

Distinct: 5911

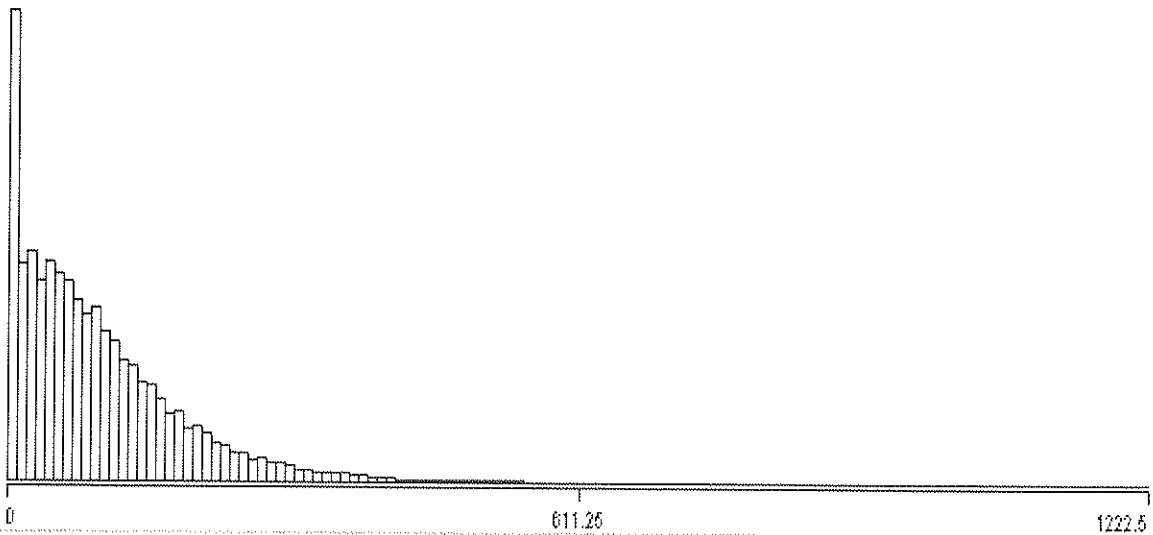
Unique: 2389 (6%)

Statistic	Value
Minimum	0
Maximum	1222.5
Mean	99.429
StdDev	96.767

Class: M1_RECHARGES_AMT (Num)



Visualize All



Outgoing SMS Revenue:

Selected attribute

Name: M1_OG_SMS_CALL_REV

Type: Numeric

Missing: 0 (0%)

Distinct: 260

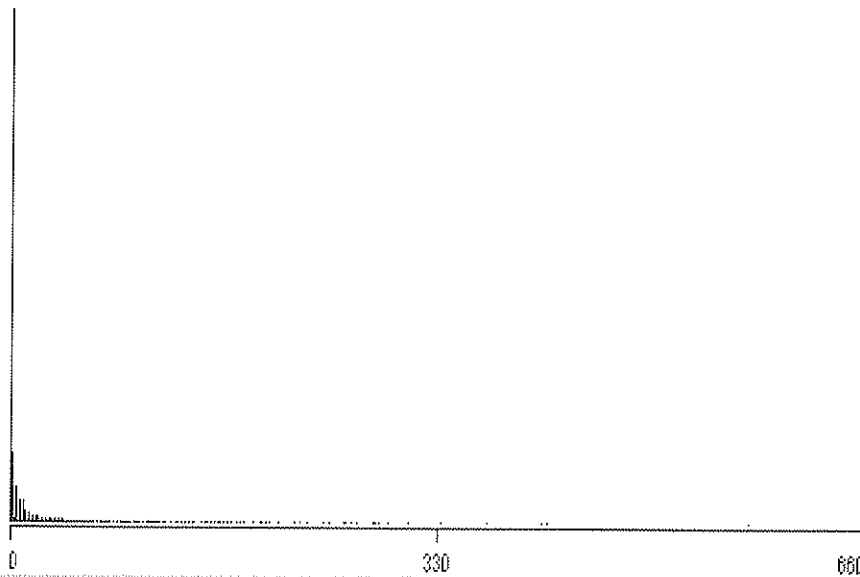
Unique: 127 (0%)

Statistic	Value
Minimum	0
Maximum	660
Mean	3.844
StdDev	13.856

Class: M1_RECHARGES_AMT (Num)



Visualize All



Recharge Amount:

Selected attribute

Name: M1_RECHARGES_AMT

Missing: 0 (0%)

Distinct: 62

Type: Numeric

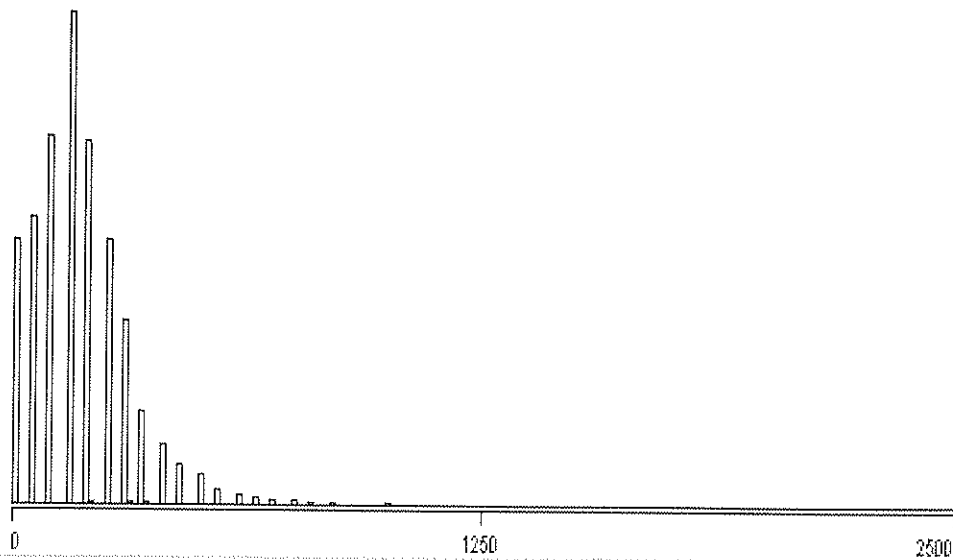
Unique: 16 (0%)

Statistic	Value
Minimum	0
Maximum	2500
Mean	180.11
StdDev	143.538

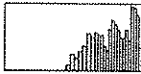

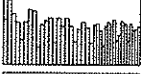
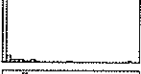
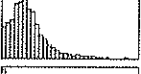
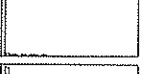
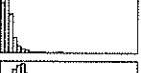
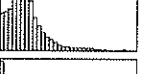


Class: M1_RECHARGES_AMT (Num)



Visualize All



4.2 Overall EDA of our Selected Attributes:

Audit		Quality		Annotations							
Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Unique	Valid			
SUBSCRIBER_ID		Continuous	86.000	9992538.000	7410038.740	1636677.827 ...	--	40000			
TENURE		Continuous	182.000	2189.000	766.668	516.802 ...	--	40000			
ARPU		Continuous	141.000	210.000	173.450	20.177 ...	--	40000			
GAP_NO_OUTGOING_VOICE		Continuous	1.000	143.000	3.139	4.706 ...	--	40000			
M1_TOT_REVENUE		Continuous	0.000	1230.050	193.278	129.691 ...	--	40000			
M1_OG_VOICE_CALL_DUR		Continuous	0.000	251887.000	2433.863	4913.875 ...	--	40000			
M1_OG_VOICE_CALL_FREQ		Continuous	0.000	1005.000	47.140	34.442 ...	--	40000			
M1_OG_VOICE_CALL_REV		Continuous	0.000	1230.050	189.434	128.148 ...	--	40000			
M1_OG_VOICE_CALL_ONNET...		Continuous	0.000	249889.000	1412.604	4761.019 ...	--	40000			
M1_OG_VOICE_CALL_ONNET...		Continuous	0.000	949.000	23.577	21.943 ...	--	40000			

4.3 Steps Involved in EDA:

- Initial Data Analysis is carried out
- All attributes Means, Ranges & Standard Deviations are analyzed
- It was analyzed that Fixline durations, frequencies & revenues for all 6 months have a very low mean values
- Fixline attributes are removed from the dataset (For Segmentation Model)

4.4 Data Details:

- Total number of Instances = 40000
- Total number of Attributes = 105
- Total number of Attributes for Month 1 = 17
- Missing Instances in Month 1 = 6947 (in total 17 attributes)

5.0 UnSupervised Learning:

In machine learning, the problem of unsupervised learning is that of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution.

5.1 Pre-Processing:

Null Values Handling:

- There were almost 6947 in Month 1
- Null Values are handled by replacing them with Zero
- As Null values shows no activity so replacing them with Zero is justified through Business Logic

1. Anomaly Detection:

- Anomaly Detection Node in IBM SPSS Modeler is used
- It not only detects outliers in complete dataset
- It uses Two Step Clustering algorithm & detects anomalies within the clusters as well
- It determines ratio of the group deviation index to its average over the cluster that the case belongs to
- Anomaly index value is set at 2 %

2. Sparseness of Data:

- Sparseness of data is checked
- Sparseness of data increases by replacing Zero's as Null values
- Rows with all null values are removed to handle sparseness

3. Attribute Selection:

- Correlation for all attributes were calculated with Total Month's Revenue
- To get Discrete Clusters for all Six Models, different combinations were tested
- Finally, 9 Best attributes Selected on the basis of Human Intelligence with Silhouette Value 0.5 & Above for Clusters.

1. Correlation:

- Correlation for first month's all attributes is carried out
- Keeping the Business logic in view
- We took only 9 most important attributes in our Model
- Their correlation with Month's revenue is shown below

2. Selected attributes for 6 Segmentation Models:

The Attributes are as follow:

- Total Month's Revenue
- Call Duration
- Call Revenue
- On-net Call Duration
- Off-net Call Duration
- Off-net Call Revenue
- SMS Frequency
- SMS Revenue
- Month's Recharge Amount

Correlations

		M1_TOT_REV ENUE	M1_CALL_DU R	M1_CALL_RE V	M1_CALL_ON NET_DUR	M1_CALL_OF FNET_DUR	M1_CALL_OF FNET_REV	M1_OG_SMS _FREQ	M1_OG_SMS _REV	M1_REC GES_AM
REVENUE	Pearson Correlation	1	.327	.994	.214	.579	.755	.139	.164	
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	
	Sum of Squares and Cross-products	672771285.1	8336831125	660976462.0	5296046116	2846654893	378844574.1	4301781.288	11794823.08	6224813
	Covariance	16819.703	208425.989	16524.825	132404.463	71168.152	9471.351	107.547	294.876	15562.
	N	40000	40000	40000	40000	40000	40000	40000	40000	40

Figure 5.4.1

Figure 5.4.1 shows the correlation for the first month's attribute with first month's total revenue attribute. Similarly all attributes were checked against all others for the correlation values. These steps were repeated for the rest of 5 months as well to determine and select attributes for Segmentation Model of these months.

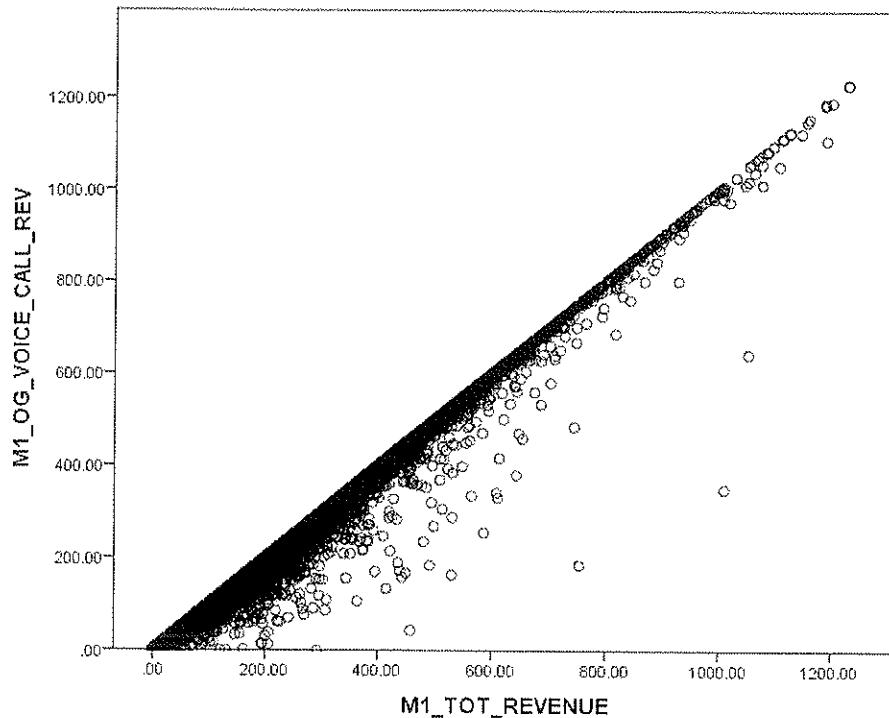


Figure 5.4.2

Figure 5.4.2 shows the correlation graph of first month's total revenue with the first month's voice call revenue.

6.0 Clustering & Customer Segmentation:

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. We used two different type of clustering algorithm, Two step clustering algorithm is discussed but only K-Means algorithms results & further findings are shown below.

6.1 Clustering Algorithms:

Two Step Clustering Algorithm:

TwoStep Clustering Component is a scalable cluster analysis algorithm designed to handle very large datasets. Capable of handling both continuous and categorical variables or attributes, it requires only one data pass in the procedure. In the first step of the Procedure, you pre-cluster the records into many small sub-clusters. Then, cluster the sub-clusters from the pre-cluster step into the desired number of clusters. If the desired number of clusters is unknown, the Two Step Cluster Component will find the proper number of clusters. Two step cluster method is an algorithm which can handle up to a large amount of data, this algorithm has several features which distinguish it from conventional clustering techniques:

- **Auto cluster number selection:**

It compares the values of a model choice standard across different clustering solutions. This method can automatically determine the ideal number of clusters required.

- **Scalability:**

It creates a cluster feature tree which summarizes all the records, and the two step algorithm provides you with the ability to evaluate very large datasets.

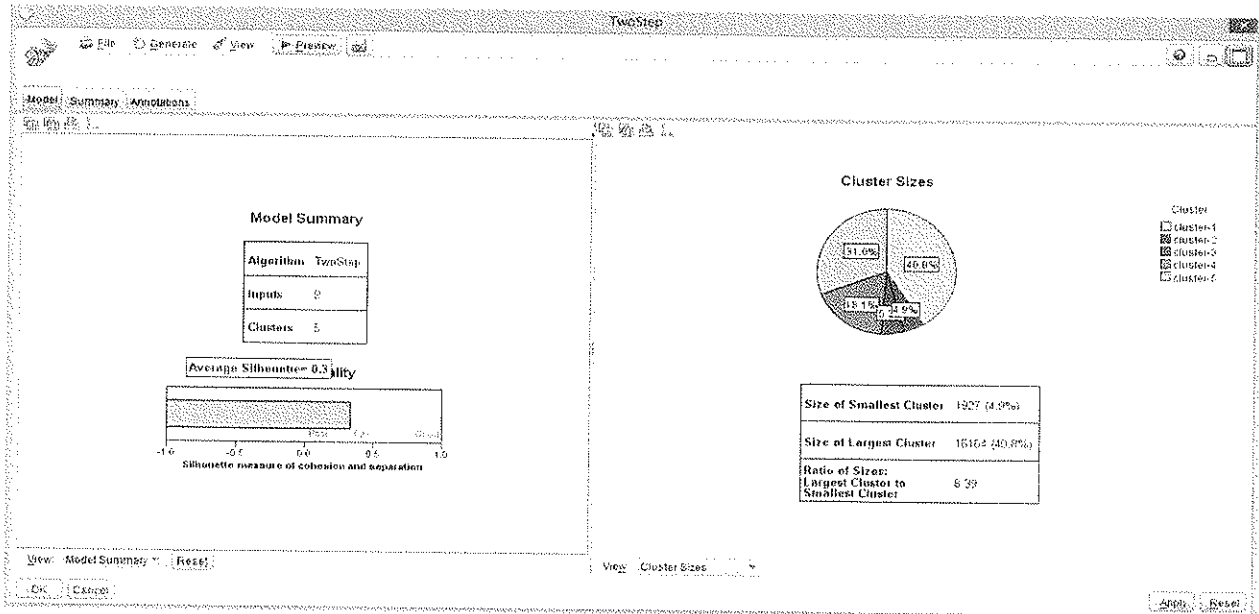
- **Categorical and continuous variable handling:**

It assumes that the variables are not dependent on each other; it handles both the continuous and categorical data.

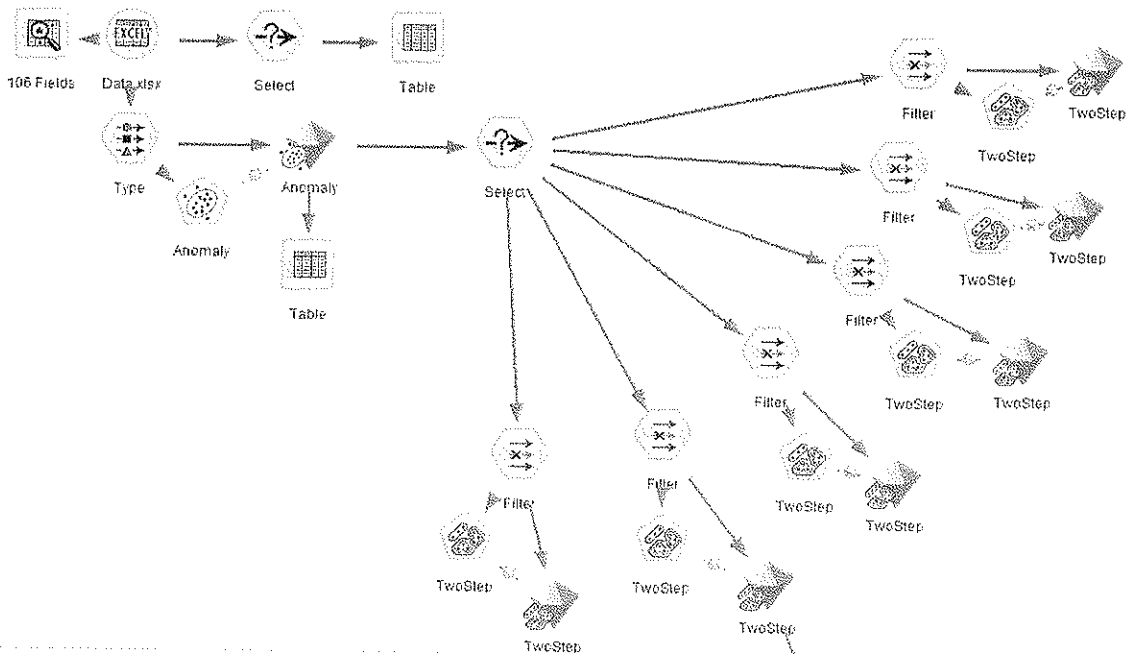
Approach for the Algorithm:

- The first step makes a single pass through the data to compress the raw input data into a manageable set of subclusters.
- The second step uses a hierarchical clustering method to progressively merge the subclusters into larger and larger clusters.
- TwoStep has the advantage it can handle mixed field types and large datasets efficiently.
- 9 Attributes included in our approach after analyzing the correlation of all data attributes month wise & selecting best possible attribute through human intelligence.
- Number of clusters 5 was selected by depending on our project domain & scope.
- This Algorithm does not suite our required scope of the project. We needed 6 segmentation models for 6 months using same attributes & to fine tune the models with same attributes for entire 6 models we needed more flexible clustering algorithm. Two Step Algorithm does not provide us with good cluster quality having low silhouette values for almost all the 6 models.

Results of Using Two Step Clustering:



Stream:



K-Means Clustering Algorithm:

Following are main features of K-Means algorithm:

1. It is an unsupervised clustering algorithm.
2. K stands for number of clusters to be formed.
3. Means mean that it uses cluster mean to group data.
2. It runs in multiple iterations.
3. It stops when means of clusters stop changing their positions.
4. If distance between clusters is more and within cluster distance is less then this means that algorithm has properly clustered the data.

Working:

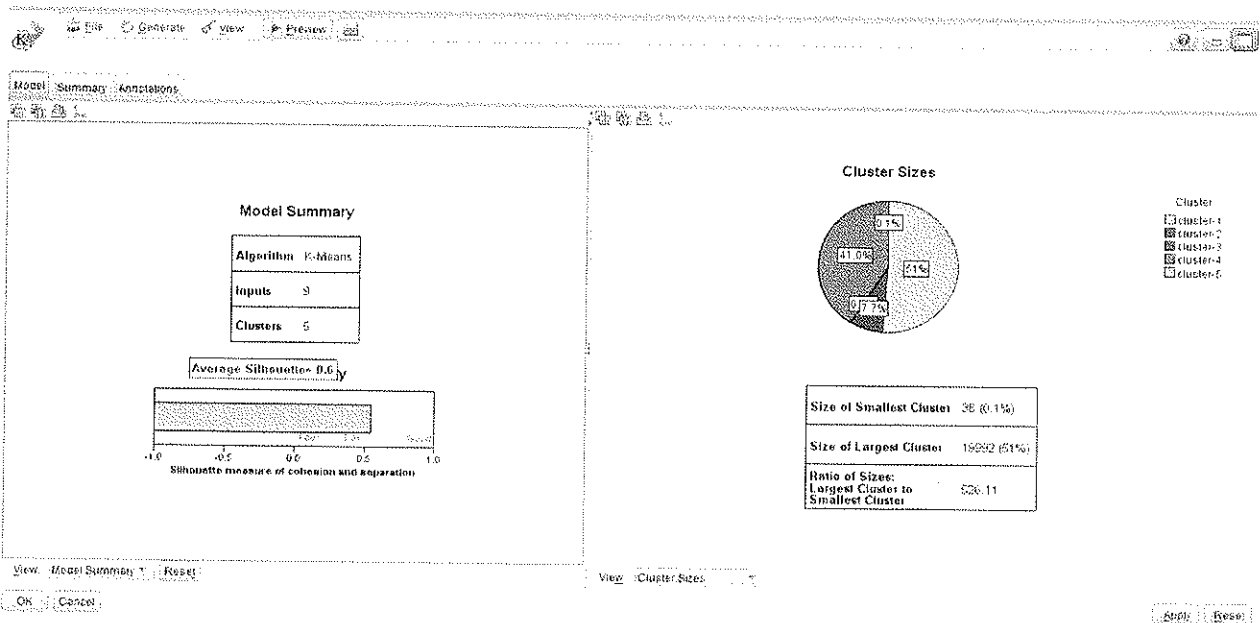
1. Initially the algorithm randomly places k points into the space represented by the objects that are being clustered. These points represent initial cluster centers.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the k centroids, hence changing positions.
4. Repeat step 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups.
5. K-means is an iterative algorithm; an initial set of clusters is defined, and the clusters are repeatedly updated until no more improvement is possible (or the number of iterations exceeds a specified limit)

Approach to the Algorithm

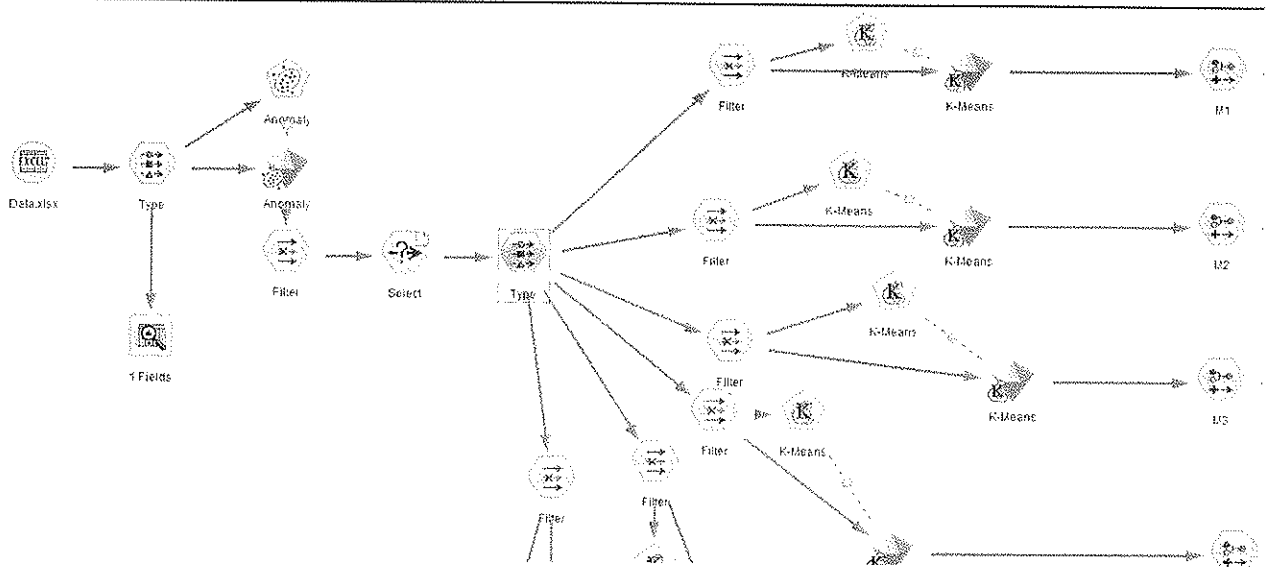
- The k-means method is a clustering method, used to group records based on similarity of values for a set of input fields. We have almost 17 attributes per month with data of 6 months. Our project scope is to track migrations within the clusters of similar quantities in 6 different segmentation models.
- As this algorithm regulates around the user defined number of clusters, iteratively assigns records to clusters, and adjusts the cluster centers until the defined number is reached & refinement can no longer improve the model.

- We used variable number of Iterations for this model from 25 to 30, to adjust the cluster quality using same 9 attributes (which gave best possible correlation while pre-processing), we maintained the cluster quality for these 6 segmentation models by adjusting this iteration quantity.
- Selected attributes to perform K-means Clustering Algorithm are mentioned in Pre-Processing.
- We determined 5 clusters in all 6 models using this algorithm & categorized them into 5 categories by analyzing 'Revenue', 'Call Revenue' & 'Recharge Amount' for every cluster.
- Categories for the respective 5 clusters are as given below in the table:

Results of using K-means Algorithm:



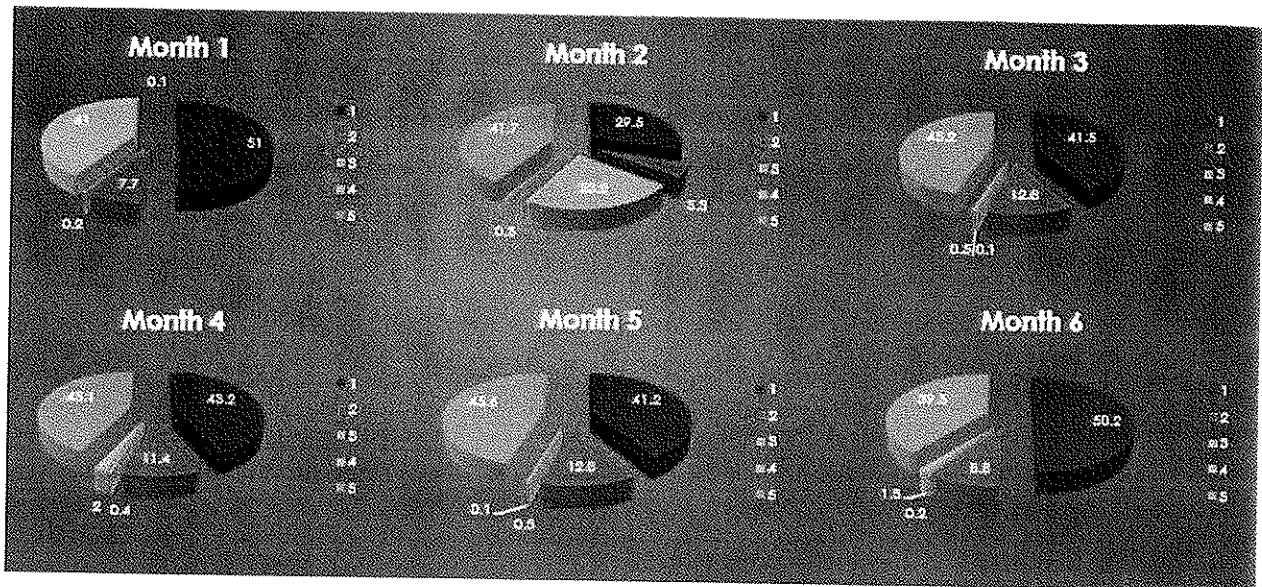
Stream:



6.2 Definition of Clusters:

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Low Revenue	High Revenue	High Revenue	Average Revenue	Low Revenue
Low Call Revenue	High Call Revenue	High Call Revenue	Average Call Revenue	Low Call Revenue
Low Recharge Amount	High Recharge Amount	Average Recharge Amount	Average Recharge Amount	Average Recharge Amount

6.3 Graphical Representation of Clusters for all 6 months:



7.0 Inter Cluster Migrations:

- Inter-Cluster Migrations were analyzed
- Migration of customers from one cluster in a particular month to another cluster in next month, were recorded
- Migrations from Month 1 to Month 6 were recorded

7.1 Migration Matix:

		Month 2				
		C1	C2	C3	C4	C5
Month 1	C1	31.15%	0.41%	0.00%	40.69%	19.46%
	C2	30.50%	10.39%	0.17%	32.92%	25.91%
	C3	28.75%	0.00%	48.75%	18.75%	3.75%
	C4	23.83%	5.30%	0.17%	44.81%	27.43%
	C5	31.57%	23.68%	0.00%	28.94%	15.79%

		Month 3				
		C1	C2	C3	C4	C5
Month 2	C1	37.64%	0.08%	11.32%	50.56%	0.41%
	C2	36.78%	0.00%	14.98%	48.18%	0.05%
	C3	10.00%	12.50%	0.00%	29.17%	48.33%
	C4	49.81%	0.01%	12.31%	37.54%	0.32%
	C5	48.87%	0.03%	15.20%	35.68%	0.21%

		Month 4				
		C1	C2	C3	C4	C5
Month 3	C1	39.28%	10.75%	0.06%	1.79%	48.11%
	C2	21.43%	0.00%	71.43%	3.57%	3.57%
	C3	45.48%	12.00%	0.08%	1.44%	40.99%
	C4	46.85%	46.85%	0.24%	2.27%	38.54%
	C5	32.58%	4.49%	41.01%	5.06%	16.85%

		Month 5				
		C1	C2	C3	C4	C5
Month 4	C1	44.89%	0.17%	13.61%	0.07%	41.28%
	C2	41.63%	0.07%	13.57%	0.09%	44.65%
	C3	34.25%	46.58%	4.11%	0.00%	15.07%
	C4	43.69%	0.39%	11.70%	0.13%	44.08%
	C5	36.91%	0.04%	11.91%	0.05%	50.55%

		Month 6				
		C1	C2	C3	C4	C5
Month 5	C1	59.47%	7.68%	0.16%	1.21%	31.46%
	C2	40.74%	0.00%	47.22%	1.85%	10.19%
	C3	41.93%	13.51%	0.04%	1.24%	43.28%
	C4	37.50%	29.17%	0.00%	4.17%	29.17%
	C5	44.18%	8.49%	0.56%	1.34%	45.93%

7.2 Migrations Report:

Cluster	Name	Month1		Month 2	
		% Subs	Subs	% Subs	Subs
1	Low Total Revenue, Low Call Revenue, Low Recharge Amount	51	19992	29.5	11
2	High Total Revenue, High Call Revenue, High Recharge Amount	7.7	3010	5.3	20
3	High Total Revenue, High Call Revenue, Avg. Recharge Amount	0.2	80	23.2	90
4	Avg. Total Revenue, Avg. call Revenue, Avg. Recharge Amount	41	16080	0.3	1
5	Low Total Revenue, Low Call Revenue, Avg. Recharge Amount	0.1	38	41.7	16

Cluster	Name	Month 2		Month 3	
		% Subs	Subs	% Subs	Subs
1	Low Total Revenue, Low Call Revenue, Low Recharge Amount	29.5	11563	41.5	16
2	High Total Revenue, High Call Revenue, High Recharge Amount	5.3	2069	12.8	50
3	High Total Revenue, High Call Revenue, Avg. Recharge Amount	23.2	9091	0.1	1
4	Avg. Total Revenue, Avg. call Revenue, Avg. Recharge Amount	0.3	120	0.5	1
5	Low Total Revenue, Low Call Revenue, Avg. Recharge Amount	41.7	16357	45.2	17

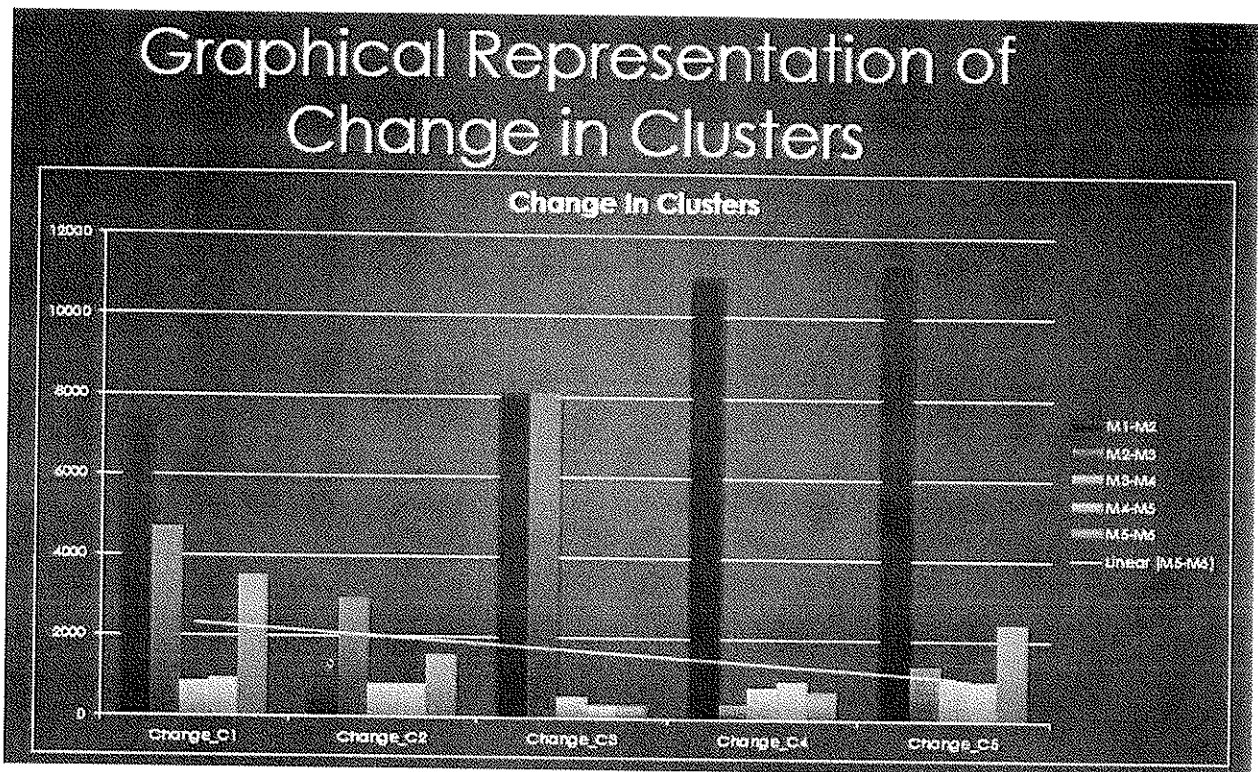
Cluster	Name	Month 3		Month 4	
		% Subs	Subs	% Subs	Subs
1	Low Total Revenue, Low Call Revenue, Low Recharge Amount	41.5	16263	43.2	16
2	High Total Revenue, High Call Revenue, High Recharge Amount	12.8	5015	11.4	40
3	High Total Revenue, High Call Revenue, Avg. Recharge Amount	0.1	28	0.4	1
4	Avg. Total Revenue, Avg. call Revenue, Avg. Recharge Amount	0.5	178	2	7
5	Low Total Revenue, Low Call Revenue, Avg. Recharge Amount	45.2	17716	43.1	16

Cluster	Name	Month 4		Month 5	
		% Subs	Subs	% Subs	Subs
1	Low Total Revenue, Low Call Revenue, Low Recharge Amount	43.2	16924	41.2	16
2	High Total Revenue, High Call Revenue, High Recharge Amount	11.4	4482	12.8	50

3	High Total Revenue, High Call Revenue, Avg. Recharge Amount	0.4	146	0.3	1
4	Avg. Total Revenue, Avg. call Revenue, Avg. Recharge Amount	2	769	0.1	
5	LOW Total Revenue, Low Call Revenue, Avg. Recharge Amount	43.1	16879	45.6	17

Cluster	Name	Month 5		Month 6	
		% Subs	Subs	% Subs	Subs
1	Low Total Revenue, Low Call Revenue, Low Recharge Amount	41.2	16170	50.2	19
2	High Total Revenue, High Call Revenue, High Recharge Amount	12.8	5018	8.8	34
3	High Total Revenue, High Call Revenue, Avg. Recharge Amount	0.3	108	0.2	8
4	Avg. Total Revenue, Avg. call Revenue, Avg. Recharge Amount	0.1	24	1.3	5
5	LOW Total Revenue, Low Call Revenue, Avg. Recharge Amount	45.6	17880	39.5	15

7.3 Graphical Representation of Migrations:



7.4 Analyzing Migration Trends to Month 6 & Marketing Strategies:

- While Analyzing Migration Matrix for Month 5 to Month 6
- Month 5 Cluster 2 is High Revenue Cluster its major portion of 40.47% is moved to Month 6 Cluster 1 (LOW Revenue Cluster)
- It is alarming situation for the company
- Company should immediately throw marketing campaigns for this portion of subscriber to increase retain their revenue
- Next target should be Month 5 Cluster 3 which is second High Revenue Cluster of Month 5, its major portion of 41.93% has moved to Month 6 Cluster 1 (LOW Revenue Cluster)

8.0 Supervised Learning:

Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset. (Matlab Works).

A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

In order to solve a given problem of supervised learning, one has to perform the following steps:

- Determine the type of training examples. Before doing anything else, the user should decide what kind of data is to be used as a training set. In the case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.
- Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.
- Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.
- Determine the structure of the learned function and corresponding learning algorithm. For example, the engineer may choose to use support vector machines or decision trees.
- Complete the design. Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.
- Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

8.1 Types of Supervised Learning:

Looking on the whole there are two types of supervised learning.

1. Classification
2. Regression

Following are the details of both types of machine learning.

Classification:

In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.).

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

Often, the individual observations are analyzed into a set of quantifiable properties, known variously explanatory variables, features, etc. These properties may variously be categorical (e.g. "A", "B", "AB" or "O", for blood type), ordinal (e.g. "large", "medium" or "small"), integer-valued (e.g. the number of occurrences of a part word in an email) or real-valued (e.g. a measurement of blood pressure). Other classifiers work by comparing observations to previous observations by means of a similarity or distance function.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category.

Regression:

Regression analysis is a statistical process for estimating the relationships among variables. With reference to our project we have used classification to solve the problem. Some details about the regression process are as follow.

Regression includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of

these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable.

9.0 Data Preprocessing:

Data preprocessing includes different stages such as:

- Data Cleansing.
- Attribute selection.
- Data Transformation.
- Class Imbalance problem.
- Sparseness of data.

9.1 Data Cleansing:

- **Noise Handling:**

It includes removing outliers from your data because they effect your results badly. Outlier is that data which lies on the boundary of your graph.

- **Duplicate removal:**

Removing duplicate records from the data as they can affect the result.

- **Anomaly Removal:**

Removing anomalous records from the data as they can also effect the results. As for prediction model we removed records using anomaly index of 1% as we did not want to lose more records with class values False. Records with class value false were already less so in order to save those records we reduced the anomaly index from 2% (as used in segmentation model) to 1%.

- **Class Imbalance Problem:**

Occurs when the total number of a class of data (positive) is far less than the total number of another class of data (negative).

Most machine learning algorithms works best when the number of instances of each classes are roughly equal. If there is class imbalance problem, this might result in biasness towards the class which greater.

Initially in our case we had only 15% of the records with recharge amount=0, i.e the records labeled as False were only 15% instead of those labeled as True were 85%. This would result in biasness towards the True class.

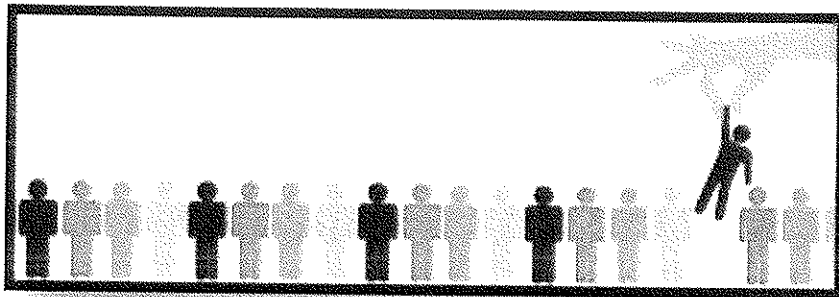
Handling the class imbalance problem:

We handled the class imbalance problem by using the balance node in the modeler. Balance node is used to adjust the ratios of the class records in order to balance them in data and get better results.

After balancing we were able to obtain the proportion of around 45% for false records and 55% for the true labeled records.

Value	Proportion	%
F		45.42
T		54.58

9.2 Feature Selection:



In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.

The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context.

Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).

Feature selection techniques provide three main benefits when constructing predictive models:

- Improved model interpretability.
- Shorter training times.
- Enhanced generalization by reducing over fitting.

9.3 Selecting Attributes:

In accordance with our data we selected features on the basis of following factors.

- **Maximum percentage of missing values:**

Screens fields with too many missing values. Fields with a large percentage of missing values provide little predictive information.

- **Maximum number of categories as a percentage of records:**

Screens fields with too many categories relative to the total number of records. If a high percentage of the categories contains only a single case, the field may be of limited use.

- **Minimum coefficient of variation:**

Screens fields with a coefficient of variance less than or equal to the specified minimum. This measure is the ratio of the input field standard deviation to the mean of the input field.

- **Minimum standard deviation:**

Screens fields with standard deviation less than or equal to the specified minimum.

On the basis of all these factors we selected **92** features. Some of them are as shown in the figure. Screen fields with:

<input checked="" type="checkbox"/> Maximum percentage of missing values	70.0
<input checked="" type="checkbox"/> Maximum percentage of records in a single category	90.0
<input checked="" type="checkbox"/> Maximum number of categories as a percentage of records	95.0
<input checked="" type="checkbox"/> Minimum coefficient of variation	0.1
<input checked="" type="checkbox"/> Minimum standard deviation	0.0

Rank	Field	Measurement	Importance	Value
74	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
75	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
76	M6_OG_VOICE_CAL...	Continuous	Import...	1.0
77	M6_OG_VOICE_CAL...	Continuous	Import...	1.0
78	M3_OG_VOICE_CAL...	Continuous	Import...	1.0
79	TENURE	Continuous	Import...	1.0
80	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
81	M3_NO_OF_RECHA...	Continuous	Import...	1.0
82	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
83	M3_OG_VOICE_CAL...	Continuous	Import...	1.0
84	M5_OG_VOICE_CAL...	Continuous	Import...	1.0
85	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
86	M3_OG_VOICE_CAL...	Continuous	Import...	0.998
87	M5_OG_SMS_CALL...	Continuous	Import...	0.997
88	GAP_NO_OUTGOIN...	Continuous	Import...	0.993
89	M5_OG_SMS_CALL...	Continuous	Import...	0.985
90	M4_OG_VOICE_CAL...	Continuous	Import...	0.983
91	M1_OG_VOICE_CAL...	Continuous	Import...	0.981
92	M1_OG_VOICE_CAL...	Continuous	Import...	0.979
93	M4_OG_VOICE_CAL...	Continuous	Marginal	0.931
94	M2_OG_VOICE_CAL...	Continuous	Marginal	0.927

9.4 Partitioning of data:

We divided our data into three parts using the partition node.

- **Training :**

First partition was of 40% which was used to train the data for the model.

- **Testing:**

Second partition was 40% as well which was used for testing and further refining the model.

- **Validation:**

Third partition was 20% which was used to validate the results of the model.

Partitions: Train and test Train, test and validation

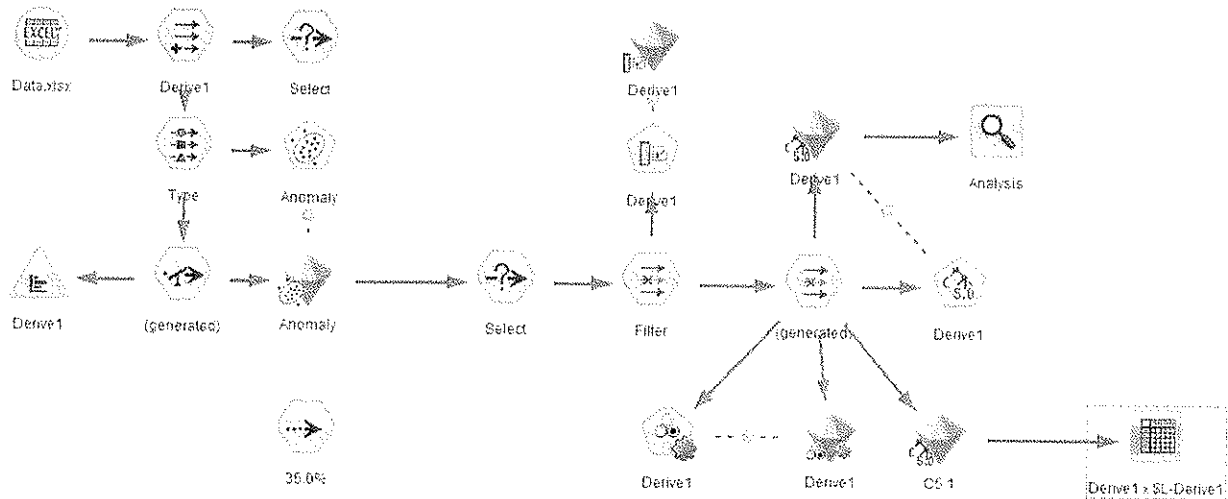
Training partition size: Label:

Testing partition size: Label:

Validation partition size: Label:

10.0 Modeling:

We started with a baseline model for classification purpose and used different algorithms for this purpose and further refined the model in order to get better results, Following is our baseline model.



10.1 Steps involved:

Step by step explanation of the model is as follow starting from the data node.

- **Input Data:**

Inserting data node in the mode and selecting the data file to be used for analysis. Data can be in several formats like .csv or .xlsx. In our case the data was in .xlsx format. In this node we give the path of data file to the node so that it can be made ready for analysis.

- **Deriving class:**

Data node is then connected to derive node where we derive the flag class that is our target class as well. Here we labeled the records having recharge_amount > 0 as T (true) and the records having recharge_amount = 0 as F (false). The new derived class is formed and records are labeled accordingly.

- **Selecting target class:**

Derive node is further connected to type node where we define our target class. We select the newly derived flag class as our target class.

- **Class imbalance:**

Derive node is further connected to balance where we solve the class imbalance problem. In balance node we set the ratios of the records in order to balance them according to our need. In our case we balanced the class ratio as 1 : 0.45 , so we got 45% F labeled records against 55% T labeled records.

- **Anomaly Detection:**

Balance node is connected to anomaly node where anomalous data is removed. The anomaly index is set to 1% in order to prevent loss of minimum possible records.

- **Feature selection:**

Anomaly node is then passed to select node where feature selection is done. Useful attributes are selected on the basis of earlier described criterion. In our case we selected 92 features which were used for the classification purpose.

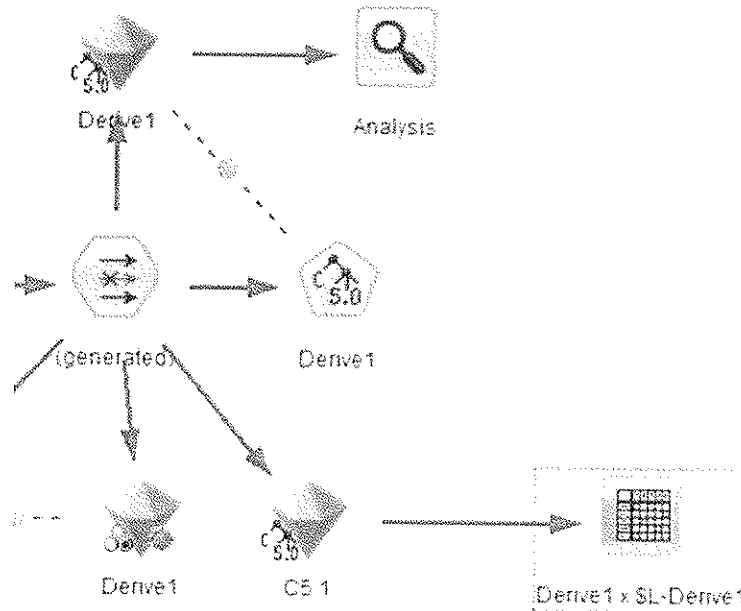
Rank	Field	Measurement	Importance	Value
75	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
76	M6_OG_VOICE_CAL...	Continuous	Import...	1.0
77	M6_OG_VOICE_CAL...	Continuous	Import...	1.0
78	M3_OG_VOICE_CAL...	Continuous	Import...	1.0
79	TENURE	Continuous	Import...	1.0
80	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
81	M3_NO_OF_RECHA...	Continuous	Import...	1.0
82	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
83	M3_OG_VOICE_CAL...	Continuous	Import...	1.0
84	M5_OG_VOICE_CAL...	Continuous	Import...	1.0
85	M4_OG_VOICE_CAL...	Continuous	Import...	1.0
86	M3_OG_VOICE_CAL...	Continuous	Import...	0.998
87	M5_OG_SMS_CALL...	Continuous	Import...	0.997
88	GAP_NO_OUTGOIN...	Continuous	Import...	0.993
89	M5_OG_SMS_CALL...	Continuous	Import...	0.985
90	M4_OG_VOICE_CAL...	Continuous	Import...	0.983
91	M1_OG_VOICE_CAL...	Continuous	Import...	0.981
92	M1_OG_VOICE_CAL...	Continuous	Import...	0.979
93	M4_OG_VOICE_CAL...	Continuous	Marginal	0.931
94	M2_OG_VOICE_CAL...	Continuous	Marginal	0.927

- **Filter Node:**

Selected node Is then connected to filter node from where only the selected attributes are passed on to be used for classification.

- **Classification using different algorithms:**

Finally the classification algorithms are applied and model is refined in accordance with the algorithms.



Analysis:

Finally the resulting nugget is connected to the analysis node in order to analyze the results.

11.0 Algorithms Used for classification:

We used three algorithms for our model and selected the best algorithm in accordance with our model.

Following are the classification algorithms that we used in our model:

- C 5.0 decision tree algorithm
- Neural Networks
- Discriminant

11.1 C 5.0 Decision tree algorithm:

C 5.0 is an algorithm used to generate a decision tree developed by Ross Quinlan. C 5.0 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C 5.0 can be used for classification, and for this reason, C 5.0 is often referred to as a statistical classifier.

C 5.0 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set $S = \{s_1, s_2, \dots\}$ of already classified samples. Each sample s_i consists of a p -dimensional vector $(x_{\{1,i\}}, x_{\{2,i\}}, \dots, x_{\{p,i\}})$, where the x_j represent attributes or features of the sample, as well as the class in which s_i falls.

At each node of the tree, C 5.0 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized

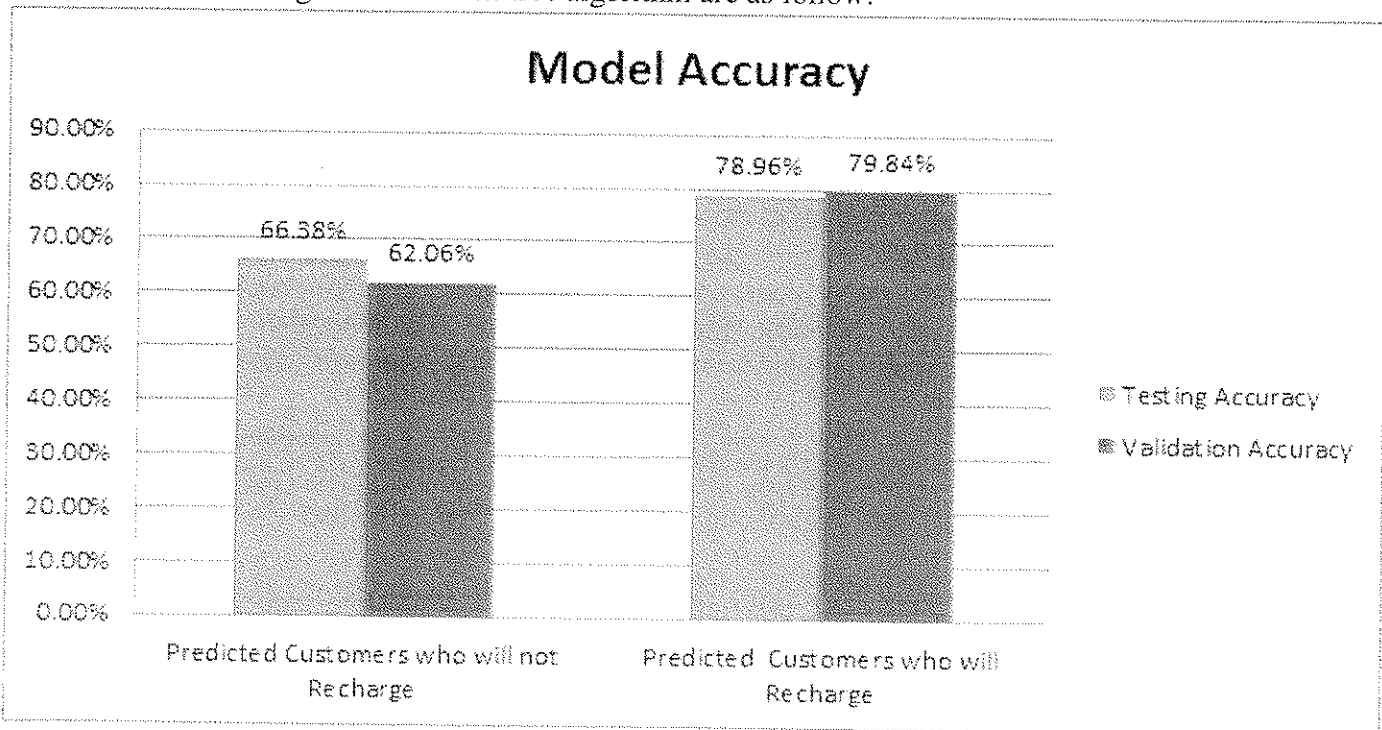
information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C 5.0 algorithm then recurs on the smaller sub lists.

This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C 5.0 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C 5.0 creates a decision node higher up the tree using the expected value.

Initial results:

Our initial results using C 5.0 decision tree algorithm are as follow:



Initial Insights:

- Our initial results were not good as per industrial requirements. We had high results for the true records. Validation accuracy was around 80% for the true records meaning that 80% of the records from data were predicted right.

- Our main concern was the records which were labeled false and were not likely to recharge in future. For this scenario our validation accuracy was only 62% meaning that only 62% of the total false records were predicted right.

Refining of the model:

We further refined our model using the following techniques in C 5.0 decision tree algorithm.

- **Boosting:**

- It works by building multiple models in a sequence. The first model is built in the usual way. Then, a second model is built in such a way that it focuses on the records that were misclassified by the first model. Then a third model is built to focus on the second model's errors, and so on.
- Increase in boosting increase accuracy.
- Massive increase in boosting may result in over fitting.
- Number of boosting trials were 25.

- **Over Fitting:**

Over fitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Over fitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model which has been over fit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data.

- **Pruning Severity:**

- Determines the extent to which the decision tree or rule set will be pruned.
- Increase in this value results in a smaller, more concise tree.
- Decrease in this value results in more accurate tree.
- Pruning severity was set o 50.

- **Minimum records per child branch:**

- Used to limit the number of splits in any branch of the tree.
- A branch of the tree will be split only if two or more of the resulting sub branches would contain at least this many records from the training set.
- Increase in this value to help prevent overtraining with noisy data.
- Minimum records per child branch were set to 3.

Use boosting Number of trials:

Cross-validate Number of folds:

Mode: Simple Expert

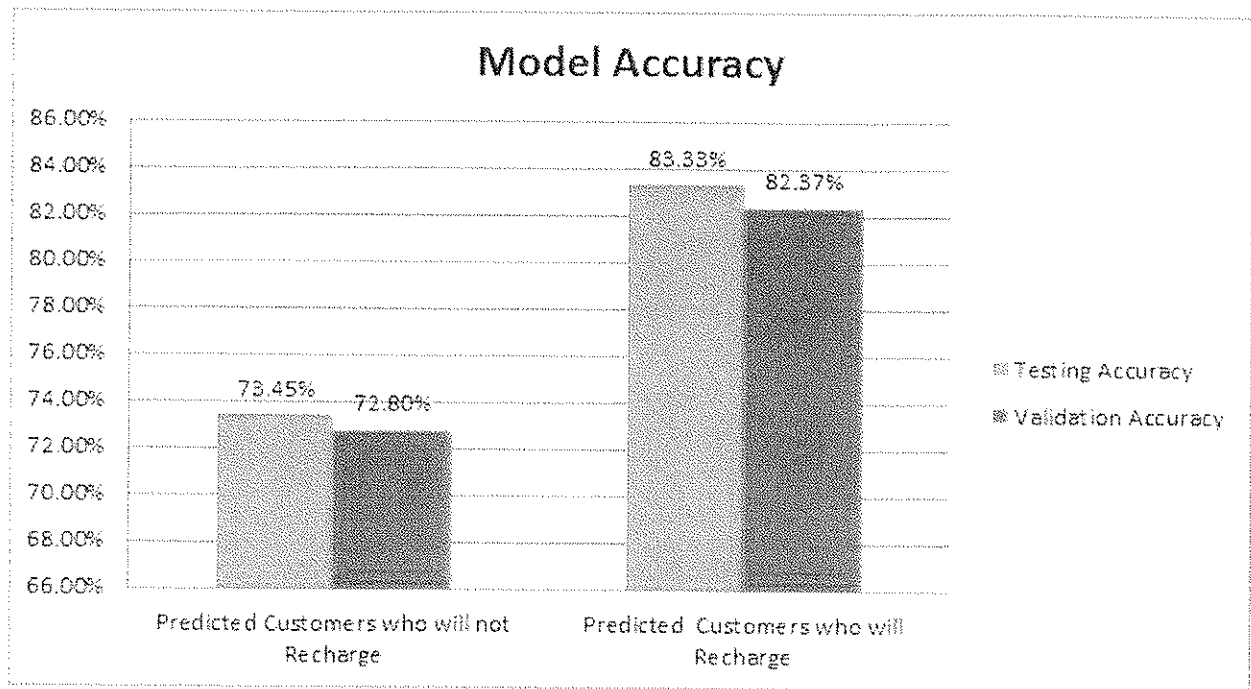
Pruning severity:

Minimum records per child branch:

Final Results:

With all these techniques and refining of the model the results were significantly improved. The model accuracy for the false records rose up to around 73% which is considered as one of the best according to industrial point of view.

Following are the results.



11.2 Neural Networks Algorithm:

In machine learning and related fields, artificial neural networks (ANNs) are computational models inspired by an animal's central nervous systems (in particular the brain) which is capable of machine learning as well as pattern recognition. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs.

For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

Like other machine learning methods - systems that learn from data - neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

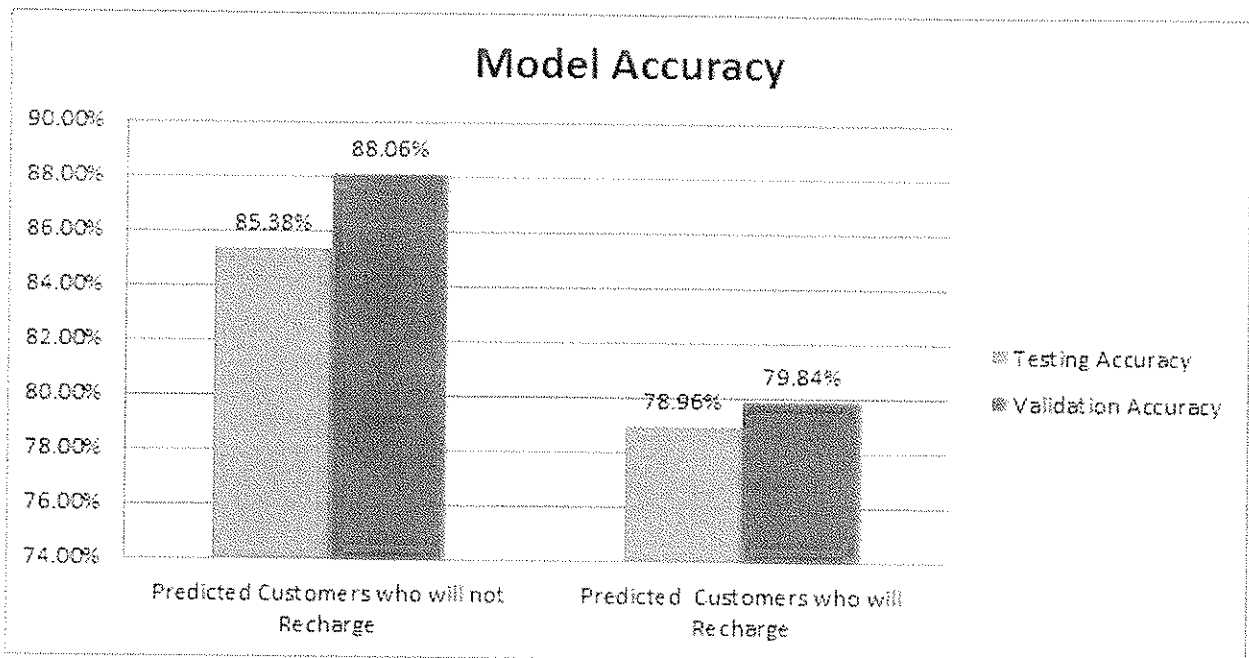
There are many types of neural networks, many of which fall into one of two categories:

- Feed-forward Networks where input is provided on one side of the network and the signals are propagated forward (in one direction) through the network structure to the other side where output signals are read. These networks may be comprised of one cell, one layer or multiple layers of neurons. Some examples include the Perceptron, Radial Basis Function Networks, and the multi-layer perceptron networks.

- Recurrent Networks where cycles in the network are permitted and the structure may be fully interconnected. Examples include the Hopfield Network and Bidirectional Associative Memory.

Results:

With respect to our model following were the results of neural networks algorithm.



Insights:

- The validation accuracy of the false records i.e the customers that will not recharge in the sixth month was 88% which was more than testing accuracy which was 85%.
- The results shows the case of over fitting.
- The accuracy was very low on the totally unknown data .
- The accuracy for the True records i.e the customers who will recharge in the sixth month was 80%, although we had more records for the true case hence making the model over fit.

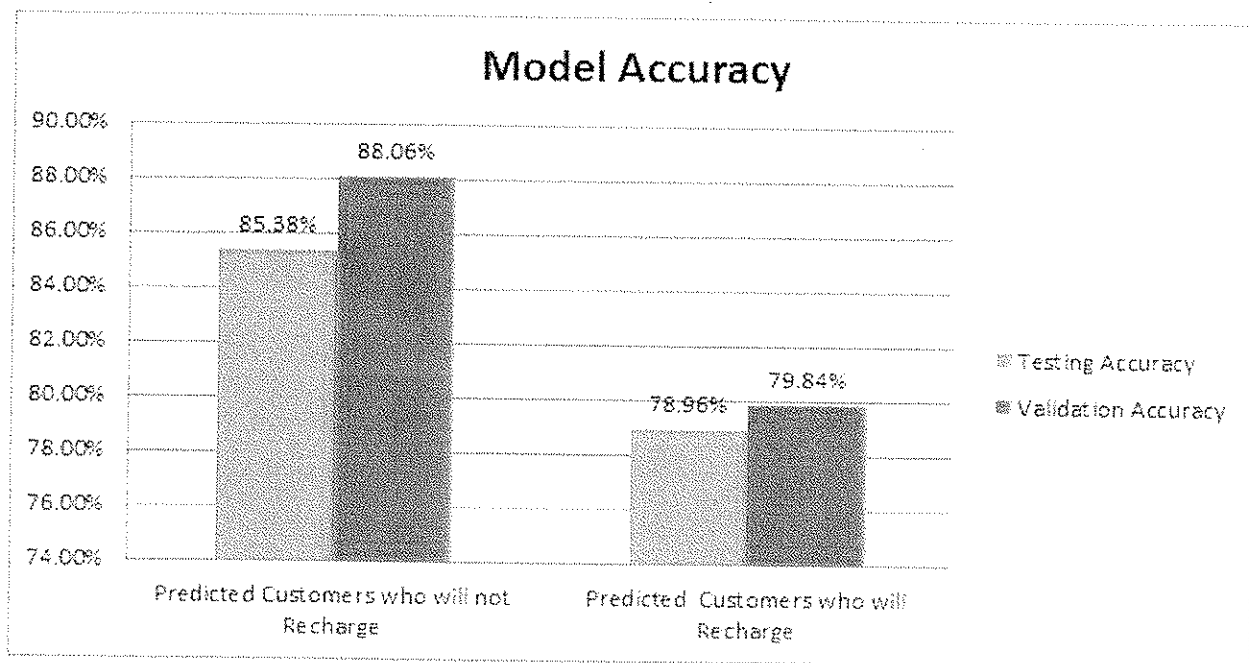
11.3 Discriminant algorithm:

Linear discriminant analysis (LDA) and the related Fisher's linear discriminant are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements.

Results:

Following are the results of the discriminant analysis algorithm

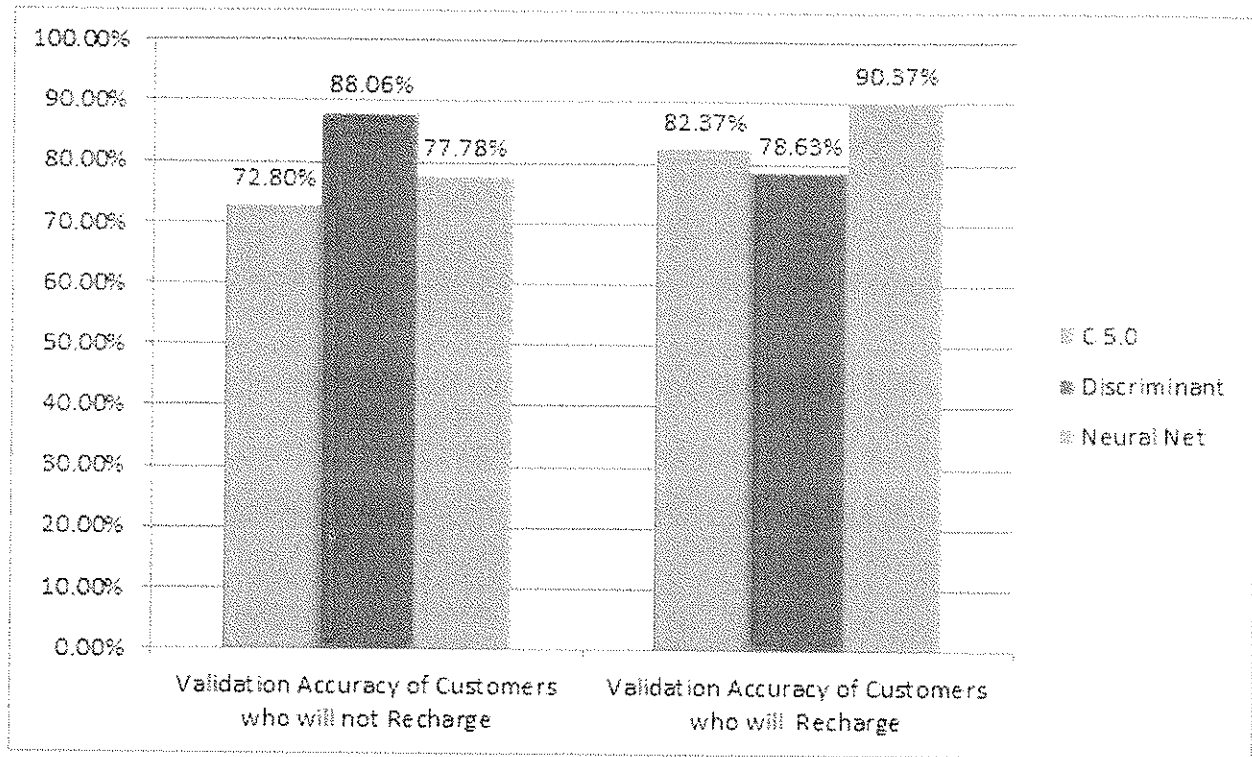


Insights:

- The validation accuracy of the false records i.e the customers that will not recharge in the sixth month was 88% which was more than testing accuracy which was 85%.
- The results shows the case of over fitting.
- The accuracy was very low on the totally unknown data .
- The accuracy for the True records i.e the customers who will recharge in the sixth month was 80%, although we had more records for the true case hence making the model over fit.

11.4 Comparison:

Following are the results comparison of all the algorithms that were applied in our model.



Insights:

- **C 5.0 is the best suited algorithm**
- **Flexible**
C 5.0 is the most flexible algorithm as it provides several options to fine tune the model like boosting, pruning and child records per branch.
- **Best results in case of unknown data**
- **No over fitting**
Over fitting occurs when the model or the algorithm fits the data too well. Specifically, over fitting occurs if the model or algorithm shows low bias but high variance.
- **No under fitting**

Under fitting occurs when the model or the algorithm does not fit the data well enough. Specifically, under fitting occurs if the model or algorithm shows low variance but high bias

12.0 Formation of BCG matrix:

We shaped our results in the form of BCG matrix in order to see that who are the immediate targets of company, who are the threats and similarly who are the potential revenue generators.

For this purpose we calculated the probabilities of the predicted records that with what probability the model have predicted any record as true or false.

Following are the results of the matrix and insights based on this matrix:

Loyal Customers: Recharge=T and Score ≥ 0.7 34.71%	Next Targets: Recharge=T and Score < 0.6
Potential Revenue Targets: Recharge=F and Score ≤ 0.6 13.82%	Immediate Targets: Recharge=F and Score ≥ 0.7 21.69%

Insights:

Based on the above matrix following are some insights

- 21.69% of the customers in the fourth quadrant are not likely to recharge with the score > 0.7 .

Throw marketing campaigns immediately: this segment is needed to be targeted immediately because they are predicted by a very high probability that these customers will not recharge.

- **13.03%** customers lying in the second quadrant of the matrix are our next targets because
 - Recharge = $T < 0.5$

There are two assumptions regarding this segment which are

- **Model fault**

Firstly that this might be a model error and records were wrongly predicted in this segment as they belong to next or previous segment.

- **Might not recharge**

Second and the most likely possibility is that these are the customers who will not recharge so these must be on the target and second in the priority list of the company.

- **13.82%** customers in the third quadrant are the potential revenue targets because
 - Recharge = $F \leq 0.6$
 - Throw campaigns to retain them so that they can progress towards the first quadrant.
- **34.71%** customers in the first quadrant are the customers who are most likely to recharge with the **score ≥ 0.7 .**

13.0 Future Work:

1. Add ons to this Project:

- Target Recharge Prediction only for those subscribers who have moved from High Revenue Clusters to Low Revenue Clusters
- Targeted Marketing campaigns can be achieved by analyzing user's behavior throughout 6 months & analyzing their recharge prediction for Month 5 & month 6
- Marketing campaigns will be more effective

2. Moving Forward:

- Using IBM COGNOS BI Reporting
- Dynamic Dashboards

14.0 Difficulties Faced:

1. Data Acquisition
 - Acquiring real time telecom data was the most difficult phase of our project
2. Selecting Project Domain
 - While analyzing the data, the hard part was to decide what to do with this data at start
 - As we moved forward into the project, we had to change our target domain several times
3. Availability of Tools
 - Started with Weka, then moved to IBM SPSS Statistics. Later on we were introduced to IBM SPSS Modeler towards the mid of project.

15.0 References:

[1] Salar Masood, Moaz Ali, Faryal Arshad, Ali Mustafa Qamar, Ahsan Rehman and Aatif Kamal.

[2] J. F. Tang T. J. Zhang, X. H. Huang and X. G. Luo. In Industrial Engineering and Engineering Management (IE EM), 2011 IEEE 18th International Conference on, volume Part 2, pages 1358–1362, 2011

[3] Băcilă Mihai-Florin, Rădulescu Adrian, Mărar Ioan Liviu
IBM SPSS:

- -Introduction to Statistical Analysis Using IBM SPSS Statistics (Student Guide)
- -Introduction to IBM SPSS Modeler and Data Mining (Student Guide)
- IBM SPSS Modeler Help

