UCD Michael Smurfit Graduate Business School

MSc in Business Analytics

MIS41270

Data Management and Mining

Assignment: Telco Churn – Team 16

By

Chaitanya Dave - 19201077

Kaushal Kumar - 19200110

Siddhesh Bhure - 19200063

Bhagyashree Janyani - 19200337

Assessment Submission Form

Student Name	Chaitanya Dave, Bhagyashree Janyani, Kaushal Kumar, Siddhesh Bhure
Student Number	19201077, 19200337, 19200110, 19200063
Assessment Title	Telco Churn
Module Title	Data Management and Mining
Module Co-ordinator	Ms Aoife D'Arcy
Tutor (if applicable)	
Date Submitted	24.04.2020
OFFICE USE ONLY **Date Received**	
OFFICE USE ONLY **Grade/Mark**	

A SIGNED COPY OF THIS FORM MUST ACCOMPANY ALL SUBMISSIONS FOR

ASSESSMENT. STUDENTS SHOULD KEEP A COPY OF ALL WORK SUBMITTED.

Procedures for Submission and Late Submission
Ensure that you have checked the School's procedures for the submission of assessments.
Note: There are penalties for the late submission of assessments. For further information please see the University's ***Policy on Late Submission of Coursework***, (http://www.ucd.ie/registrar/)

Plagiarism: the unacknowledged inclusion of another person's writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism designed to impose sanctions that reflect the seriousness of University's commitment to academic integrity. Ensure that you have read the University's ***Briefing for Students on Academic Integrity and Plagiarism*** and the UCD

Declaration of Authorship

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

Signed: Bhagyashree, Siddhesh, Date: 24/04/2020

Kaushal, Chaitanya Date: 24/04/2020

Plagiarism Statement, Plagiarism Policy and Procedures, (http://www.ucd.ie/registrar/)

Executive Summary

Acme Telecom company's Customer retention management team has asked the analytics division to address the current customer base of the company and build strategies to retain them, or better, to build a bigger customer base and be prepared for the new incoming customers. This spectrum of propositions is backed by the available data.

The analytics team will be required to breakdown the ongoing transaction and deduce actionable items. This was done by making use of Machine learning concepts and the tools that operate on them.

The customer breakdown led us to find 4 prominent clusters, for which distinct divisions could be made. With Cluster 3 having the greatest number of people, this cluster would be a great opportunity to maximise the revenue of the company and spread the good word about the company. The dissatisfaction is proportional to the usage of the services that company provides as seen from cluster 4. Cluster 3 has not used the network to its potential yet. However, if that's to happen, company will need to fix underlying networking issues before approaching bigger underwhelming clusters such as Cluster 3. The report is followed by a series of recommendations which was quite evident by seeing the trends in clusters, the bundles being insufficient is an awakening call for the company and need to act swiftly along with network problem in suburbs.

Table of Contents

Chapter 1: All About Business

Section 1: Business problem statement

The company has been suffering from rising attrition over the years and becoming a concern for the company. *Figure 1* gives information of the inclination as compared to past few years.

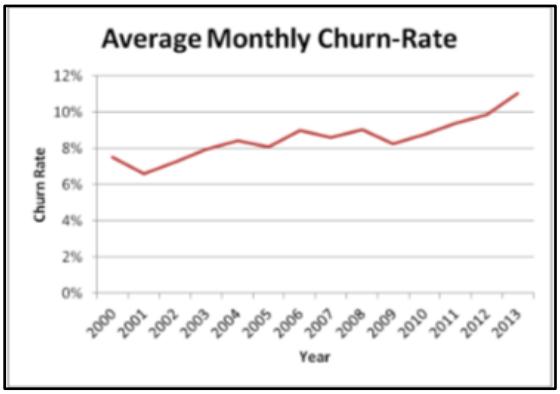

Figure 1: Churn Rate for Acme Telecom

"Your most unhappy customers are your greatest source of learning." -- Bill Gates

Understanding customers by addressing data-backed different opportunities to help retain them and even adding new customers. The strategies will allow the company to stop the current attrition and gain traction to more and more consumers by ways of personalised messages and marketing schemes.

Ergo, the report is defined to address two problems: -

a) Identifying and predicting the people who have churned and learning from them.
b) Keeping different strategies by targeting customers with specific strategies.

The report also recommends few investigative steps backed by data to improve the service and provide better customer experience. These steps can be further analysed by appropriate teams.

Section 2: What we know so far!

Analyses that are required from the team will be derived from the below columns in *figure 2*.

Variable	Type	Description
customerID	Numeric	Customer ID
children	Categorical	There are children present in the customer's household {true, false}
credit	Categorical	The customers credit rating {a, aa, b, c, de, gy, z}
creditCard	Categorical	The customer owns a credit card {true, false}
custcare	Numeric	average number of calls to customer calls in the last 6 months
custcareTotal	Numeric	total calls to customer calls in the last 6 months
custcareLast	Numeric	calls to customer calls in the last month
directas	Numeric	The number of directory assisted calls made in the last 6 months
directasLast	Numeric	The number of directory assisted calls made last month
dropvce	Numeric	The number of calls dopped in the last 6 months
dropvceLast	Numeric	The number of calls dopped the last month
income	Numeric	The cutomer's income {0 - 9}
marry	Categorical	The customer's marital status {yes, no, unknown}
mou	Numeric	Number if minutes last month
mouTotal	Numeric	The total number of minutes used in the last 6 months
mouChange	Numeric	% change in minutes
occupation	Categorical	The occupation of the cutomer { clerical, crafts, homemaker, professional, retired, self-employed, student}
outcalls	Numeric	The number of calls made
overage	Numeric	The number of minutes over the customer's bundle used this month
overageMax	Numeric	Max overage
overageMin	Numeric	Min overage
peakOffPeak	Numeric	The total number of peak calls made the last 6 months
peakOffPeakLast	Numeric	The total number of peak calls made last month
recchrge	Numeric	The recurring bundle charge this month
regionType	Categorical	The type of region in which the customer lives {rural, suburban, town}
revenue	Numeric	Reveue from customer last month
revenueTotal	Numeric	total revenue in the last 6 months
revenueChange	Numeric	% change in revenue
roam	Numeric	The number of roaming events in the time period
churn	Categorical	Flag indicating if the customer has churned

Figure 2: Data Dictionary

- The information above is stored for each customer(~10k).
- The data tell about customer's demography such as marital status, children, region type.
- The data also provides financial aspects of the customer such as Credit rating, Credit card holder, income bands.
- The volume of usage of our services are identified by columns Minutes of usage (MoU), overage, peakoffPeak.
- There are distinct services provided by the company such as Customer Care, roaming, directory assisted and obvious outcall service.
- The dynamic usage of these services spanning over different time intervals results into revenue for the company.
- However, the churn of the customer is a consolidation of several reasons such as drop calls which perhaps could be caused by overage, usage of network during peak times or the even the location.

Section 3: How the report is expected to help?

- Factors affecting churn and prediction of new customer churning in future.
- Building a definitive predictive model with a confidence to identify the churn of the customer.
- Co-relation between several factors affecting the business.

- Report that classifies our customers into different baskets and how each basket responds to our services.

Section 4: What the team has used?

The team used a series of computational techniques with the help of some academic references. Python was used for initial assessment of data because of quick prototyping ability to provide multiple functions to operate on data set. Concepts of Machine learning and application of the team's understanding of business. Team used several combinations of decision trees with different settings and decided that since multiple categorical columns are there, to go ahead with Gradient Boosting in SAS Viya.

Eventually the team has used clustering technique "K-means clustering" to divide customers targetable segments, where different features of the company showed prominence in each segment. This helped the team to recommend some fixable issues with the help of other teams. There was a strong preference for some of the company's features in some clusters.

Section 5: Analytical Approaches

1. Use cutting-edge analytical techniques. Cutting-edge analytics let operators apply advanced algorithms to vast troves of data without needing to program specific transformations. These algorithms can identify previously hidden variables and combinations of variables that predict customer behaviors such as churn. Companies can then analyze the reasons behind those behaviors to come up with solutions.

   For example, a leading operator used an analytical technique called "feature discovery" to identify over 50 variables that contributed to customer churn, as well as their relative importance. These variables included specific thresholds, such as combinations of phone type, data usage, and call-center history that, once reached, reliably predicted customer attrition. With a list of these thresholds in hand, the company was able to hold a series of cross-functional workshops to identify root causes of customer discontent, such as issues at specific call centers and dropped calls. (Jain, P. and Surana, k., 2017)

2. Break the customer base into scores of microsegments. The full value of data analytics can only be realized when companies can personalize the treatment of a precisely targeted group of customers with the highest propensity to leave. Such a tailored approach requires a granular micro-segmentation of the customer base, which is then matched to a broad, well-classified library of offers. One leading operator, for example, developed a library of over 50 offers and then set up a mechanism for rapidly launching and measuring the related campaigns. As a result, the company was able to reduce churn by 10–15% over the following 18 months. (Julien Boudet, Brian Gregg, Jason Heller, and Caroline Tufft, 2017)

Section 6: Business Understanding

To follow an unbiased approach and get academic insights we used ResearchGate by *Mahajan R. 2017.et al International Journal of Data Analysis Techniques and Strategies* to identify what factors affects a telecom industry. The study helped us to identify the factors based on which we could better understand the best conventions followed in the industry.

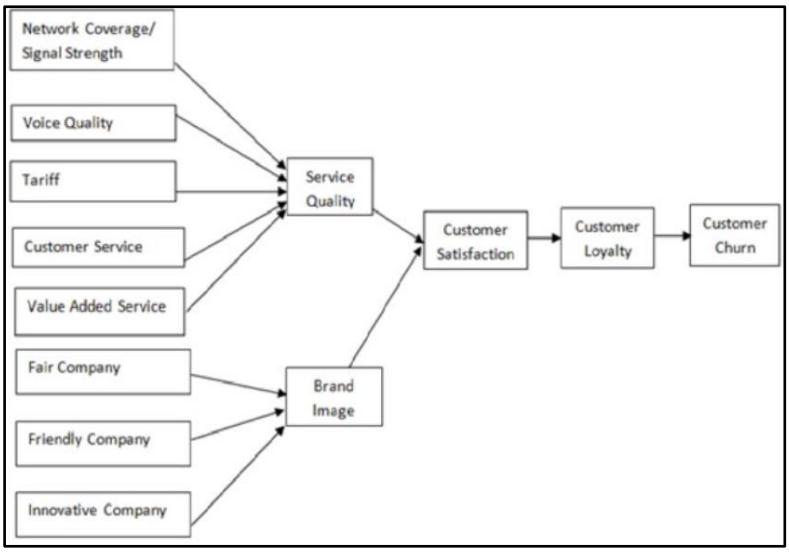

Figure 3: Mahajan R. 2017.et al International Journal of Data Analysis Techniques and Strategies

Figure 3 is an excerpt from Research gate, which we used to get second opinion about the industry. The rules that govern this industry are universally the same, to say the least, the socio-economic factors of our customers play an equally important role as compared to the services provided by the company.

Therefore, an equal emphasis will be made by the report on understanding our customer's innate nature and how the company can adapt to that. The Data dictionary provided to us, aligns well with this study mentioned in *Figure 3* and will give a good reflection of the steps to be taken to address the long-standing issues of the business.

-

Chapter 2: All About Data

Section 1: Data Cleaning:

1. Replace values:
   Region type: Wrote code to replace: 4 rows with value 'r' to 'rural', 36 rows with value 's' to 'suburban', 6 rows with value 't' to 'town'.
2. Removed outliers with values above 3 Standard deviation from mean.
3. Also, below columns were filtered based on the metadata limits mentioned in table below:

Column	Lower Limit	Upper Limit
Directas	0	8
Dropvce	0	33
Outcalls	0	730
Peakoffpeak	0	11
Recharge	0	119
Revenue	0	194
Revenuechange	-2	1.8
RevenueTotal	0	1009
Roam	0	19

Table 1: Metadata Limits

This step excluded 384 observations from 4810 selected by SAS as training data.

4. Wrote the code to combine marry column with children column to get a combine result in single column named "Family".
   (Please refer Appendix 1 figure 9 for source code)

Section 2: Data Characterization Report

Data at a Glance

- Thirty variable characterises the Original data set, including seven categorical, and Twenty-two numerical variable and one target variable.
- Data replacement resulted into creation of on new variable Family using values from variable children and marry
- Missing values and outliers are removed from variable occupation and region type in analytics base table.
- Cases that were found too skewed were also removed from analytics-based table.

(See figure Appendix 2 figure 11 for statistical and graphical summary of data)

Family
This variable is important from strategic point of view as different consumer plans can be derived from this data. This variable is further divided in two different categories family consumers & Non-family consumers. Preliminary analysis on data indicates that majority of customer are Non-Family Consumers as compared to Family consumers. This inferential statistic is derived from two column children and marry. Rows for which marry value is 'NO' and children value is 0 are categorized as Non-Family consumers.

Credit Card
Most consumers are having credit cards (68%) while sizable minority do not. This variable is important from marketing perspective as company can give different schemes on recharges and bill payments. Offers may vary depends on other parameters.

Credit Score
This variable is indicating the dominance of different credit score bands. As most of the consumers are credit card holders (from above point) it is not surprising to see that majority of the consumers' credit score lie between top two bands 'AA' and 'B'. According to data there also a significant amount of population having good credit score but do not have credit card.

Income bands
Among all the data significant chunk is under income band 0 and 6. Income bands 5, 4, 3 are having almost equal number of populations.

Family Consumers and Churn
It has been observed from cluster analysis that consumer from cluster 4 which are non-family consumers are churning more as compared to cluster 1 and 3. Cluster 2 is next on this pattern after cluster 4.

Customer care calls and drop calls
These variables are important from service point of view the mean value of total customer care calls in last 6 months is 8 ranging from 0 to 213 calls in last 6 months. Average number of drop calls in last month is 5.9 ranging from 0 to 93.

Minutes of Usage
Usage statistics shows mean of Number of minutes used in last month is 521.88 with standard deviation of 542.25. This variable can be used to decide talk time/recharge plans according to different segments.

Last month Revenue
This is an important feature from marketing point of view. Average value for last month revenue is around 60.5 with deviation of 41.71 maximum consumers are lying between 0-120 revenue range. Consumers from Revenue range from greater than 150 are significantly less.

Roaming service
Overall data suggests maximum no of consumers used roaming service at least 15 times in given time period, whereas heavy roaming service users are not that much. Variable importance for this is from different recharge plans according to different clusters.

Worth of each column – "Feature Importance" is part of almost every ML modelling techniques. This attribute basically ranks the columns based on its relevance in the entire dataset. For the analysis of our dataset, we have found the relevant columns in 3 different iterations:

1. Considering all columns: Seeing below, the most important columns were Revenue of last 6 months, Revenue of last month. Profit being the KPI of any organisation, it became very crystal clear to take into consideration the columns related to revenue.

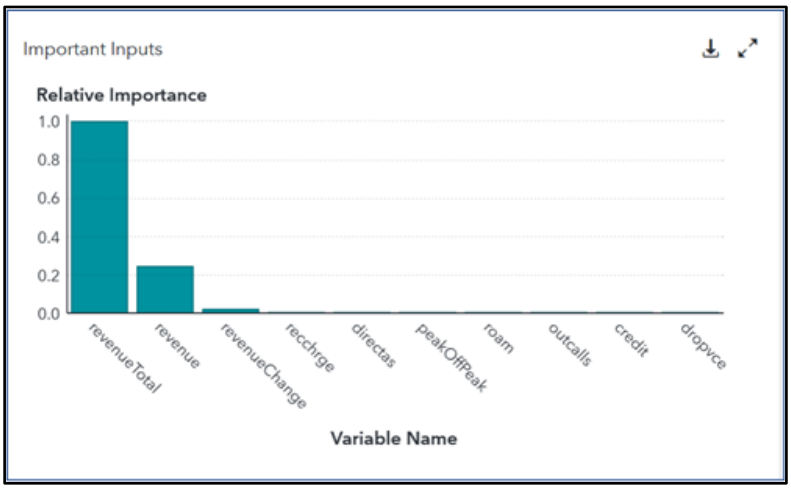

Figure 4: Worth of columns

2. Next set of relevant columns were found by rejecting the columns observed in first set of this analysis. The columns observed were "Minutes of usage in last 6 month & last month". For a company, these parameters help to find the spectrum utilisation.

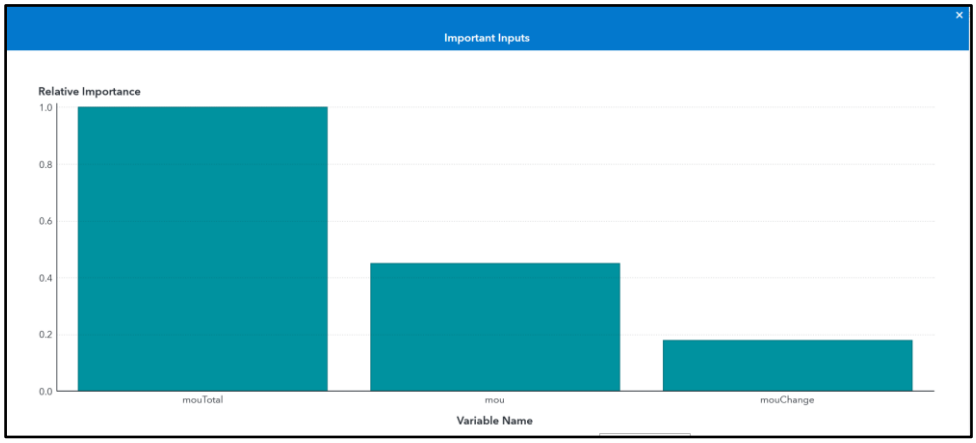

Figure 5: Worth of Columns

3. Third set of relevant columns were found by rejecting the columns of set 1 & 2, so as to find the ranking of columns independent of the most relevant columns “Revenue” & “Minutes of Usage”.
   This showcased that columns related to “Customer care calls” make good analysis data and thus are ranked in relevant list of columns.

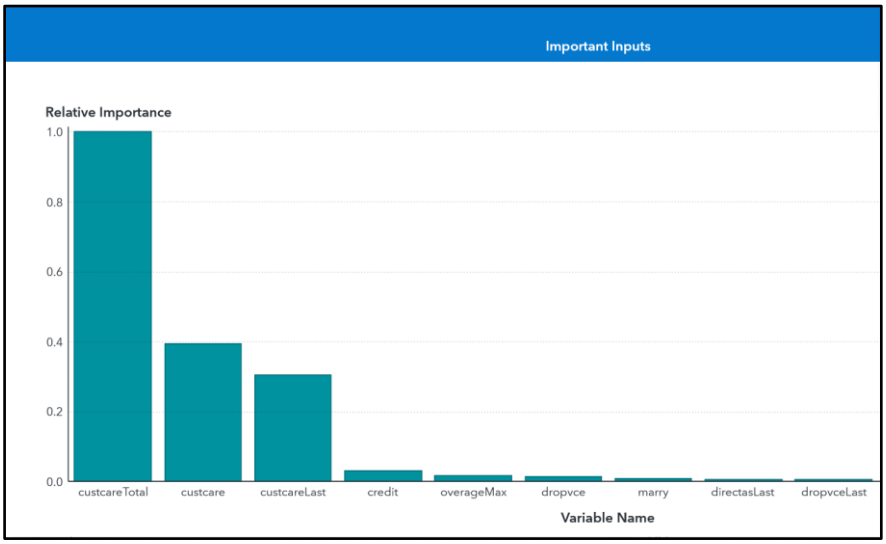

Figure 6: Worth of Columns

Chapter 3: Prediction Model

Several predictive models we used after cleaning the data. Variables were selected and passed to this step (*see Appendix 3 Figure 13*). To see how different models, respond to the dataset, multiple pipelines were developed where different predictive had different configurations see Appendix 3 figure 14. In order to choose the best performing model a comparison node was added in the pipeline (*see Appendix 3 figures 15-18)*.

Multiple decision trees were setup with different configuration (see *Appendix 3 figure 14,23*)

An ensemble (multiple models collaborated under one hood) was also devised to see if (*Appendix figure 21*) if a cumulative effect would be better off than individual decision trees.

Ultimately Gradient boosting was created to compete against the tradition predictive models and there was difference in its favour. (*See Appendix 3 figure 19*)

All the models were compared (figure) and drawn with their corresponding ROC values (*Appendix Figure 24*). The model with highest **A**rea **U**nder **C**urve value was chosen as the final model. Another metric Accuracy score (See *Appendix Figure 23*) was also used to find give better understanding to the business.

Gradient Boosting:

An incremental technique where weak decision trees are converted into strong decision trees. This process is particularly helpful when we do not want to do any pre-processing of data and when working with categorical variables. In our, dataset we are working with Family, income, credit card, credit rating, region type. Multiple choice of hyperparameters will allow to customise the learning rate so that overfitting can be avoided, which are not available with most of machine learning technique and are dependent on dataset.

Chapter 4: Customer Segmentation

K-means Clustering – Is a data science technique where we don't already have a label for the customer cluster but we classify a customer based on similarities that it shares with other customers. Since we don't have a separate column which summarises each customer inclination to a group, we want the system to do that for us based on other customers cohesiveness.

The closer a customer is to any cluster more is the probability for it to behave in the same way. (See figure 8).

The business perspective or the business' biggest problem that is taken to decide on these clusters plays an important role in defining how clusters are made.

Since this problem definition is related to churn and the dissatisfaction of our customers, we will classify our target clusters mainly on two reasons which define this problem. That is, Revenue change and drop calls. Any customer would have a valid reason to leave the network if the service(calls) provided do not meet a benchmark or standard. As a reaction to this dissatisfaction, the revenue would decrease over time or the usage of the network may decrease over time.

It is imperative that the report highlights that this clustering analysis is independent of the predictive analysis where prediction of churn data was being checked in the previous section. The clustering assigns each customer in one cluster. A significant difference drop calls and revenue change would help distinguish the data between two clusters and assign them into their respective baskets. Please refer *Appendix 4 Figure 28- Figure 35* for this section.

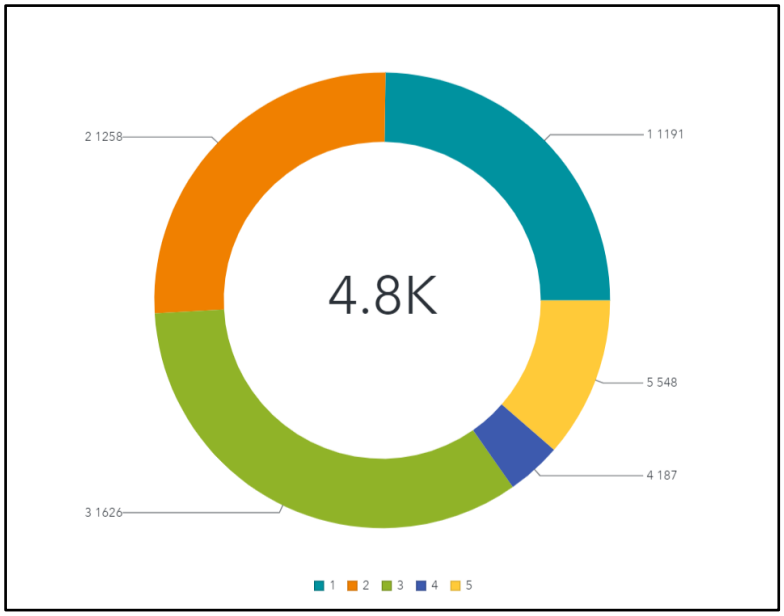

Figure 7: Clustering

Cluster Centroids

Cluster ID	dropvce	revenueChange	Standard: dropvce	Standard: revenueChange
1	6.3034	-0.0213	0.0460	-0.1102
2	2.3317	0.0373	-0.4097	-0.0106
3	0.3948	-0.0463	-0.6319	-0.1526
4	35.0313	0.1917	3.3420	0.2517
5	14.0947	0.3508	0.9399	0.5219

Figure 8: Centroids

Cluster 3 (The Elephants) –

Cluster 3 contains equal no of people churning and not churning. However, this is reflective of population and contains max population from data. Any pilot program should be kept far from this cluster. Maximum number of people in cluster 3 having credit cards (*Appendix 4 figure 31*).

Customer belonging to cluster 3, go for smaller recurring plans and in turn use less network and face lesser drop calls (*Appendix 4 figure 35*). Cumulative earning of cluster 3 is highest as compared to others given the fact that they have most population in the cluster (*Appendix 4 figure 32*). Almost equally poised family demography in this cluster i.e. 50% of non-family and about the same for non-family members (*Appendix 4 figure 33*)

Dominating factors	Credit Rating of population	Credit Card owner?	Family or Individuals	Popular Income bands	Churn & Non churn
Cluster 3	a, aa, b	**yes**	**Equal**	0,6,9	**Equal**

Highest number of credit card owners are in this cluster. Along with that, most of these credit card owners have good credit ratings. This is a gold mine if pilot program runs successfully in other smaller clusters.

Cluster 4 (The Tigers) –

Population in this cluster, has made the maximum utilization of all the services (Minutes of Usage in last month, recharge, overage in last month, customer care calls in last month, directory assisted calls) see *Appendix 4 figure 35* and have in turn contributed most to the revenue. As a result, they have also been affected by the greatest number of drop calls.

Dominating factors	Credit Rating of population	Credit Card owner?	Family or Individuals	Popular Income bands	Churn & Non churn
Cluster 4	aa, b, de	**no**	**Individuals**	0,6	**Equal**

As compared to family members, there are more individuals more in this cluster (*Appendix 4 figure 33*). Since they contribute most to company's revenue, these customers should have a higher priority in customer servicing.

Cluster 1 & Cluster 2 (The Wanderers) –

Consists of most number customer roaming which is perhaps why the minutes of usage is less for them (*Appendix 4 figure 35*). The calls made in this cluster often are at peak timings.

Dominating factors	Credit Rating of population	Credit Card owner?	Family or Individuals	Popular Income bands	Churn & Non churn
Cluster 1	aa, b, de	**yes**	**Individuals**	0,6,9	**Equal**
Cluster 2	a, aa, b	**yes**	**Equal**	0,6,9	**Equal**

Introduce roaming friendly bundles and tie-ups with foreign telecom networks. Since the cumulative population in cluster 1 and 2 surpasses that of cluster 3 they can also be attracted to make calls in non-peak hours.

Cluster 5 (The Penultimate Group) –

Cluster 5 has always been at 2nd position irrespective the element of business. The cluster has more population than cluster 4 and quite symbolic when representing revenue of the company.

It will serve an ideal group before targeting clusters with higher population.

Dominating factors	Credit Rating of population	Credit Card owner?	Family or Individuals	Popular Income bands	Churn & Non churn
Cluster 1	aa, b, de	**yes**	**Individuals**	0,6,9	**Equal**
Cluster 2	a, aa, b	**yes**	**Equal**	0,6,9	**Equal**

See Appendix 4

Chapter 5: Recommendations:

Please refer to A*ppendix 5 figure 36*. The depth of the colour indicates that how strongly two factors are associated with each other. Overage and revenue

1. Directory-assisted and Revenue-Total
2. DropCalls and Revenue-Total
3. Directory-assisted calls and Outcalls
4. Outcalls and Customer Care

Factors affecting Revenue and Revenue of 6 months:

a) Overage: Our current bundles are not enough for all clusters except cluster 3. This fact is a testament that newer bundles should be made. *See Appendix 5 figure 37.*
b) Directory-assisted calls: Directory-assisted services are quite popular. Directory assisted calls should be converted into a phone application. This would save work for Customer representatives and save cost for customers as well. This will be a step towards innovation by the company. *See Appendix 5 figure 41.*
c) Drop Calls: The fact that network issues exist in our services, customers are paying more to complete their interrupted conversations. Network engineers to be check problematic areas and target them on priority basis. *See Appendix 5 figure 40.*

Factors affecting Outcalls:

a) Directory-assisted calls: Directory-assisted calls are occupying the network; they do provide revenue but we can optimise this for our customer and free up some bandwidth and be available on this service round the clock without involvement of a human representative.
See Appendix 5 figure 38.
b) Customer Care: Since the reasons making customer care calls are not available, we again could convert all the queries in and FAQ and make the customers use that. If they are not satisfied, they can raise their grievances via tickets and avoid having to wait in the call queues. This will again save bandwidth for company and reduce network traffic. *See Appendix 5 figure 39.*

Chapter 6: Appendices

Appendix 1: Data cleaning & Data preparation

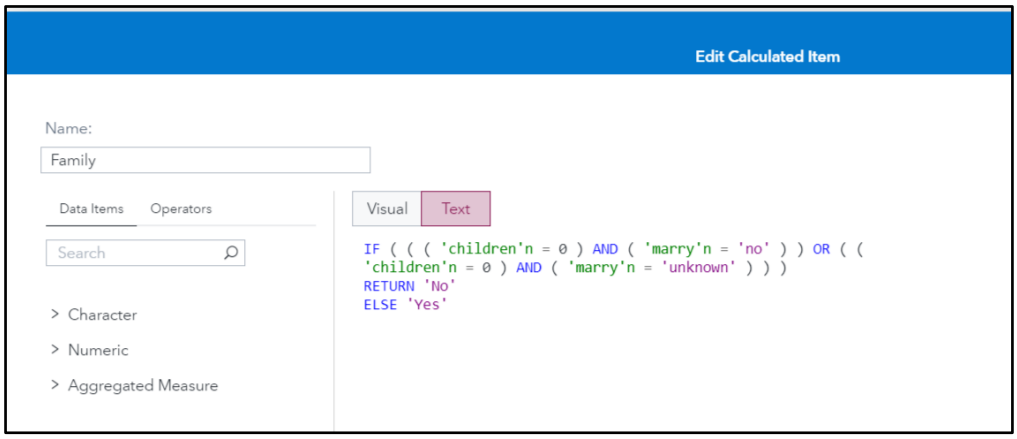

Figure 9: Source code for Column Family

```
/* BEGIN data step with the output table data */
data {{_dp_outputTable}} (caslib={{_dp_outputCaslib}} promote="no");
/* Set the input set */
set {{_dp_inputTable}} (caslib={{_dp_inputCaslib}} );
if 'regionType'n = "r" then 'regionType'n = "rural";
else if 'regionType'n = "t" then 'regionType'n = "town";
else if 'regionType'n = "s" then 'regionType'n = "suburban";
/* END data step run */
run;
```

Figure 10: Code for region type

Appendix 2: Descriptive Statistics

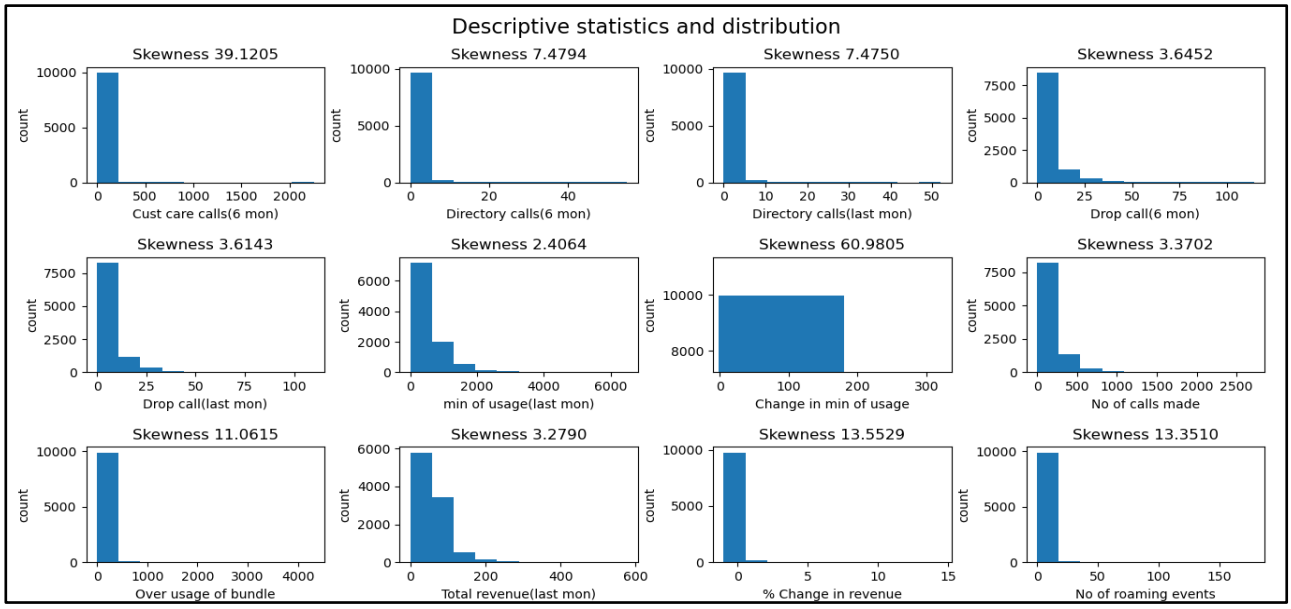

Figure 11: Graphical Representation

	custcareTotal	directas	directasLast	dropvce	dropvceLast	mou	mouChange	outcalls	overage	revenue	revenueChange	roam
count	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	8.090200	0.90930	0.882000	5.985100	5.924600	521.887595	0.727684	158.731000	41.116650	61.482590	0.040231	1.185700
std	31.586415	2.28989	2.252284	8.654488	8.641677	542.907394	23.173149	188.531081	106.262528	43.372238	0.536981	6.105927
min	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	-1.000000	0.000000	0.000000	0.000000	-1.000000	0.000000
25%	0.000000	0.00000	0.000000	1.000000	1.000000	150.999321	-0.064399	37.000000	0.000000	37.909500	-0.016855	0.000000
50%	0.000000	0.00000	0.000000	3.000000	3.000000	361.336280	0.000000	105.000000	0.000000	50.620833	0.000000	0.000000
75%	6.000000	1.00000	1.000000	8.000000	7.000000	708.051098	0.065608	214.000000	42.884316	72.794375	0.016582	0.000000
max	2253.000000	55.00000	52.000000	114.000000	110.000000	6494.324859	1818.528039	2716.000000	4313.142208	577.211667	14.463000	178.000000

Figure 11: Descriptive Statistics

training_descriptive.py | categorial_descriptive.py | simulation.py | Churn2_telecom_V3.py | datamining_project_new.py | grid.py

```
36
37      print(encoded_ds1)
38      num_feature = ['custcareTotal','directas','directasLast','dropvce','dropvceLast','mou','mouChange','outcalls','overage','revenue',
39      'revenueChange','roam']
40      num_feature_label = ['Cust care calls(6 mon)','Directory calls(6 mon)','Directory calls(last mon)','Drop call(6 mon)',
41      'Drop call(last mon)','min of usage(last mon)','Change in min of usage','No of calls made','Over usage of bundle',
42      'Total revenue(last mon)','% Change in revenue','No of roaming events']
43      print(encoded_ds1[num_feature].describe())
44
45      print('median'+ str(statistics.median(encoded_ds1['mou'])))
46      # print('skewness\t')
47      # fun = lambda x: print('oyee'+x+'\t')
48      # print('skewness',end = "  ")
49      print('          ',end = '')
50      skewlist = skew(encoded_ds1[num_feature])
51      for c in num_feature:
52
53          print(c,end='  ')
54      print('')
55      print('skewness  ',end = '')
56      for x in skewlist:
57          print('{:06.4f}'.format(x),end = '    ')
58
59
60      fig , axes = plt.subplots(nrows=3, ncols=4)
61      ax = axes.flatten()
62      i = 0
63      fig.suptitle('Descriptive statistics and distribution',fontsize = 16)
64      for col in encoded_ds1[num_feature].columns:
65          # print(col)
66          ax[i].hist(encoded_ds1[col])
67          ax[i].set_title('Skewness '+'{:06.4f}'.format(skewlist[i]))
68          ax[i].set_xscale('linear')
69          ax[i].set_yscale('linear')
70
71          ax[i].set_xlabel(num_feature_label[i])
72          ax[i].set_ylabel('count')
73          i = i + 1
74
75
76      # fig.tight_layout(pad=0.10)
77      plt.subplots_adjust(hspace = 0.8,wspace = 0.5)
78      plt.show()
79
```

Figure 12: Source Code

Appendix 3: Prediction Model

Variable Name	Label	Type	Role	Level	Order	Comment	Number of Levels	Missing	Minimum	Maximum
children		Numeric	Input	Binary	Default		2	0.0000	0.0000	1.0000
churn		Numeric	Target	Binary	Default		2	0.0000	0.0000	1.0000
credit		Character	Input	Nominal	Default		7	0.0000		
creditCard		Numeric	Input	Binary	Default		2	0.0000	0.0000	1.0000
custcare		Numeric	Input	Interval	Default		49	0.0000	0.0000	376.0000
custcareLast		Numeric	Input	Interval	Default		47	0.0000	0.0000	387.0000
custcareTotal		Numeric	Input	Interval	Default		150	0.0000	0.0000	2,253.0000
customerID		Numeric	ID	Interval	Default		>254	0.0000	1,000,004.0000	1,099,988.0000
directas		Numeric	Input	Interval	Default		31	0.0000	0.0000	55.0000
directasLast		Numeric	Input	Interval	Default		29	0.0000	0.0000	52.0000
dropvce		Numeric	Input	Interval	Default		78	0.0000	0.0000	114.0000
dropvceLast		Numeric	Input	Interval	Default		75	0.0000	0.0000	110.0000
income		Numeric	Input	Interval	Default		10	0.0000	0.0000	9.0000
marry		Character	Input	Nominal	Default		3	0.0000		
mou		Numeric	Input	Interval	Default		>254	0.0000	0.0000	6,494.3249
mouChange		Numeric	Input	Interval	Default		>254	0.0000	-1.0000	967.4179
mouTotal		Numeric	Input	Interval	Default		>254	0.0000	0.0000	38,965.9492
occupation		Character	Rejected	Nominal	Default	The variable exceeds the percentage of missing cutoff value.	7	73.9803		
outcalls		Numeric	Input	Interval	Default		>254	0.0000	0.0000	2,716.0000
overage		Numeric	Input	Interval	Default		>254	0.0000	0.0000	4,313.1422
overageMax		Numeric	Input	Interval	Default		>254	0.0000	0.0000	4,565.2116
overageMin		Numeric	Input	Interval	Default		>254	0.0000	0.0000	4,098.2936
peakOffPeak		Numeric	Input	Interval	Default		>254	0.0000	0.0000	79.3333
peakOffPeakLast		Numeric	Input	Interval	Default		>254	0.0000	0.0000	79.3333
recchrge		Numeric	Input	Interval	Default		>254	0.0000	0.0000	337.9800

Figure 13: Prediction - Selected Columns

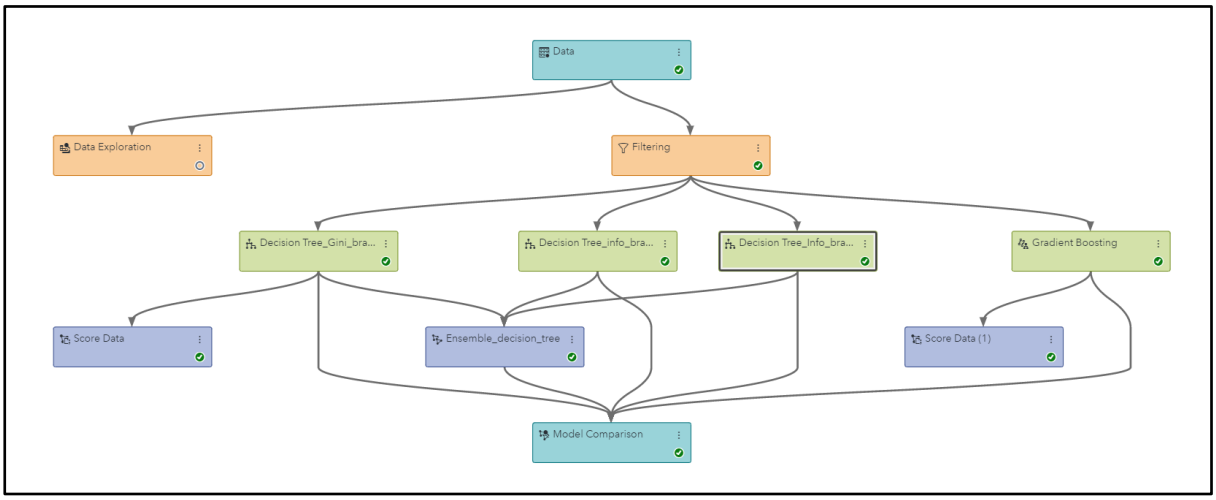

Figure 14: Pipeline

Model Comparison:

Statistics Label	Model Name	Test	Train	Validate
Target Name	Ensemble_decision_tree	churn	churn	churn
Sum of Frequencies	Ensemble_decision_tree	802	4810	2405
Divisor for ASE	Ensemble_decision_tree	802	4810	2405
Formatted Partition	Ensemble_decision_tree	2	1	0
Partition Indicator	Ensemble_decision_tree	2	1	0
Gamma	Ensemble_decision_tree	0.99210375	0.985790586	0.985379564
Area Under ROC	Ensemble_decision_tree	0.988357992	0.983434134	0.982514087
Gini Coefficient	Ensemble_decision_tree	0.976715984	0.966868267	0.965028175
KS (Youden)	Ensemble_decision_tree	0.892255453	0.875582157	0.872190456
ROC Separation	Ensemble_decision_tree	0.892255453	0.875582157	0.872190456
Tau	Ensemble_decision_tree	0.488955514	0.483522622	0.482700696
KS Cutoff	Ensemble_decision_tree	0.4	0.4	0.4
Root Average Squared Error	Ensemble_decision_tree	0.209516847	0.222354779	0.227001151
Multi-Class Log Loss	Ensemble_decision_tree	0.172286726	0.186667992	0.192985032
Misclassification Rate (Event)	Ensemble_decision_tree	0.05361596	0.061954262	0.063617464
Misclassification at Cutoff	Ensemble_decision_tree	0.05361596	0.061954262	0.063617464
Misclassification Rate	Ensemble_decision_tree	0.05361596	0.061954262	0.063617464
Average Squared Error	Ensemble_decision_tree	0.043897309	0.049441648	0.051529523

Figure 15: Model Comparison

Statistics Label	Model Name	Test	Train	Validate
Target Name	Decision Tree_Info_branch2_depth1_leaf4	churn	churn	churn
Sum of Frequencies	Decision Tree_Info_branch2_depth1_leaf4	729	4426	2169
Divisor for ASE	Decision Tree_Info_branch2_depth1_leaf4	729	4426	2169
Formatted Partition	Decision Tree_Info_branch2_depth1_leaf4	2	1	0
Partition Indicator	Decision Tree_Info_branch2_depth1_leaf4	2	1	0
Gamma	Decision Tree_Info_branch2_depth1_leaf4	0.864246334	0.868279484	0.878829515
Area Under ROC	Decision Tree_Info_branch2_depth1_leaf4	0.746007075	0.753825687	0.761038414
Multi-Class Log Loss	Decision Tree_Info_branch2_depth1_leaf4	0.547016862	0.539984051	0.531397078
Gini Coefficient	Decision Tree_Info_branch2_depth1_leaf4	0.49201415	0.507651374	0.522076829
KS (Youden)	Decision Tree_Info_branch2_depth1_leaf4	0.49201415	0.507651374	0.522076829
ROC Separation	Decision Tree_Info_branch2_depth1_leaf4	0.49201415	0.507651374	0.522076829
Root Average Squared Error	Decision Tree_Info_branch2_depth1_leaf4	0.427255709	0.423314629	0.418940156
KS Cutoff	Decision Tree_Info_branch2_depth1_leaf4	0.35	0.35	0.35
Misclassification at Cutoff	Decision Tree_Info_branch2_depth1_leaf4	0.253772291	0.245368278	0.238358691
Misclassification Rate	Decision Tree_Info_branch2_depth1_leaf4	0.253772291	0.245368278	0.238358691
Misclassification Rate (Event)	Decision Tree_Info_branch2_depth1_leaf4	0.253772291	0.245368278	0.238358691
Tau	Decision Tree_Info_branch2_depth1_leaf4	0.246344533	0.253875584	0.261154323
Average Squared Error	Decision Tree_Info_branch2_depth1_leaf4	0.182547441	0.179195275	0.175510854

Figure 16: Model Comparison

Statistics Label	Model Name	Test	Train	Validate
Target Name	Decision Tree_Gini_branches2_depth10_leaf5	churn	churn	churn
Sum of Frequencies	Decision Tree_Gini_branches2_depth10_leaf5	729	4426	2169
Divisor for ASE	Decision Tree_Gini_branches2_depth10_leaf5	729	4426	2169
Formatted Partition	Decision Tree_Gini_branches2_depth10_leaf5	2	1	0
Partition Indicator	Decision Tree_Gini_branches2_depth10_leaf5	2	1	0
Gamma	Decision Tree_Gini_branches2_depth10_leaf5	0.997834596	0.997064969	0.996072147
Area Under ROC	Decision Tree_Gini_branches2_depth10_leaf5	0.995976968	0.996121681	0.993117199
Gini Coefficient	Decision Tree_Gini_branches2_depth10_leaf5	0.991953936	0.992243362	0.986234398
KS (Youden)	Decision Tree_Gini_branches2_depth10_leaf5	0.969795273	0.952426731	0.959381696
ROC Separation	Decision Tree_Gini_branches2_depth10_leaf5	0.969795273	0.952426732	0.959381696
Tau	Decision Tree_Gini_branches2_depth10_leaf5	0.496657321	0.496219208	0.493336157
KS Cutoff	Decision Tree_Gini_branches2_depth10_leaf5	0.4	0.4	0.4
Root Average Squared Error	Decision Tree_Gini_branches2_depth10_leaf5	0.116005819	0.13965612	0.137161733
Multi-Class Log Loss	Decision Tree_Gini_branches2_depth10_leaf5	0.103641983	0.065953936	0.122184352
Misclassification at Cutoff	Decision Tree_Gini_branches2_depth10_leaf5	0.015089163	0.023723452	0.020285846
Misclassification Rate	Decision Tree_Gini_branches2_depth10_leaf5	0.015089163	0.023723452	0.020285846
Misclassification Rate (Event)	Decision Tree_Gini_branches2_depth10_leaf5	0.015089163	0.023723452	0.020285846
Average Squared Error	Decision Tree_Gini_branches2_depth10_leaf5	0.01345735	0.019503832	0.018813341

Figure 17: Model Comparison

Statistics Label	Model Name	Test	Train	Validate
Target Name	Gradient Boosting	churn	churn	churn
Sum of Frequencies	Gradient Boosting	729	4426	2169
Divisor for ASE	Gradient Boosting	729	4426	2169
Formatted Partition	Gradient Boosting	2	1	0
Partition Indicator	Gradient Boosting	2	1	0
Area Under ROC	Gradient Boosting	0.999491946	0.999268974	0.998839404
Gamma	Gradient Boosting	0.999337414	0.998852458	0.998108223
Gini Coefficient	Gradient Boosting	0.998983893	0.998537948	0.997678808
KS (Youden)	Gradient Boosting	0.978067138	0.976541	0.971411081
ROC Separation	Gradient Boosting	0.967040494	0.964639975	0.962105908
Tau	Gradient Boosting	0.500177121	0.499367119	0.499060903
KS Cutoff	Gradient Boosting	0.35	0.35	0.4
Root Average Squared Error	Gradient Boosting	0.132476911	0.133961149	0.140185917
Multi-Class Log Loss	Gradient Boosting	0.086499788	0.088541685	0.092800075
Average Squared Error	Gradient Boosting	0.017550132	0.017945589	0.019652091
Misclassification at Cutoff	Gradient Boosting	0.016460905	0.017623136	0.01890272
Misclassification Rate	Gradient Boosting	0.016460905	0.017623136	0.01890272
Misclassification Rate (Event)	Gradient Boosting	0.016460905	0.017623136	0.01890272

Figure 18: Model Comparison

Gradient Boosting

Accuracy: 98.11%

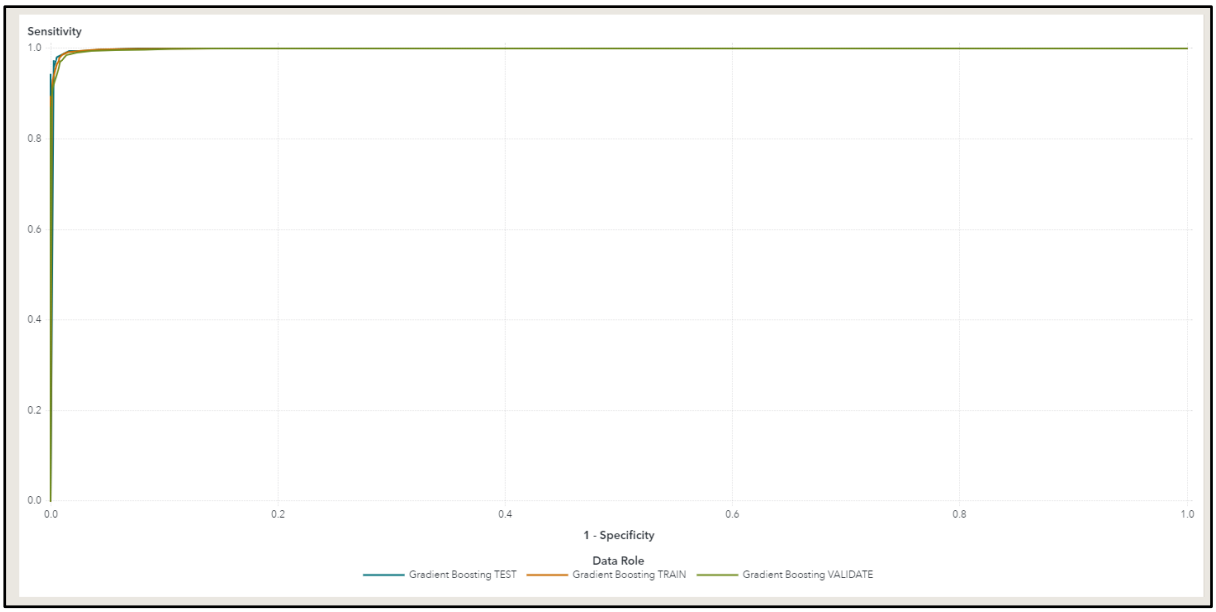

Figure 19: Gradient Boosting Accuracy

Decision Tree: Gini

Accuracy: 97.97%

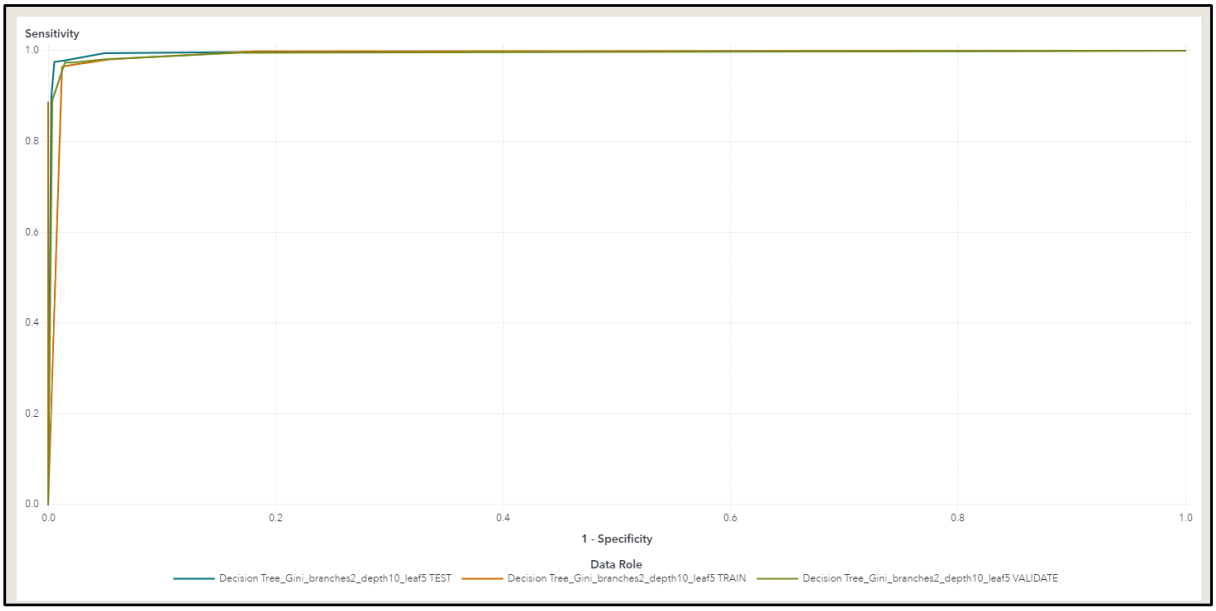

Figure 20: Decision Tree with Gini - Accuracy

Ensemble

Accuracy: 93.64%

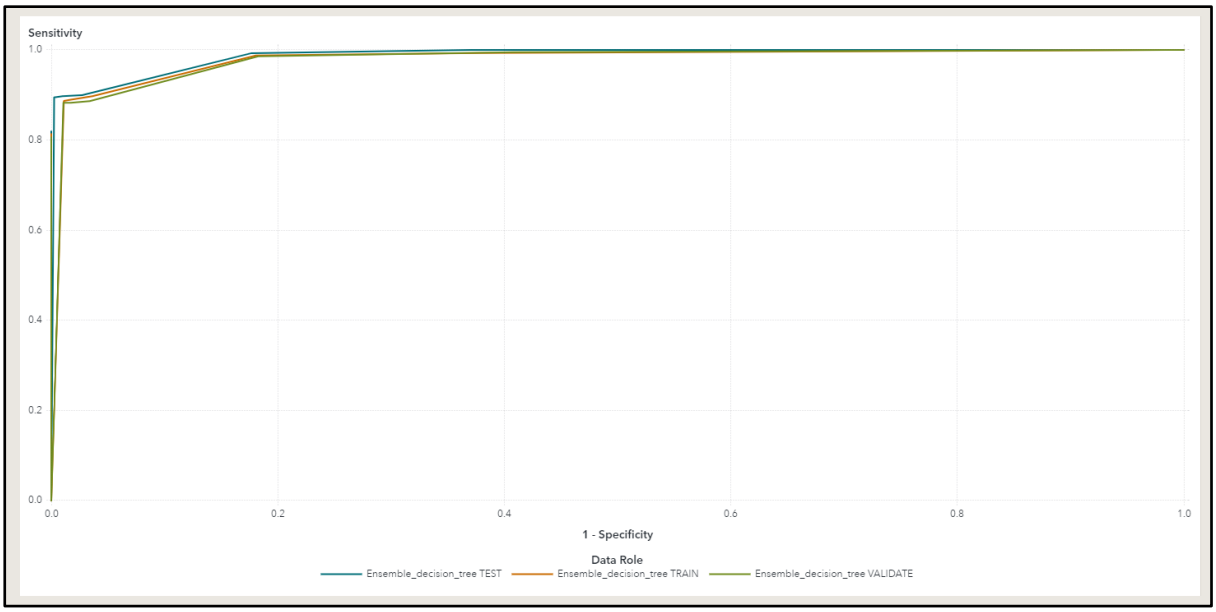

Figure 21: Ensemble: Accuracy

Decision Tree_info_branch2

Accuracy: 76.16%

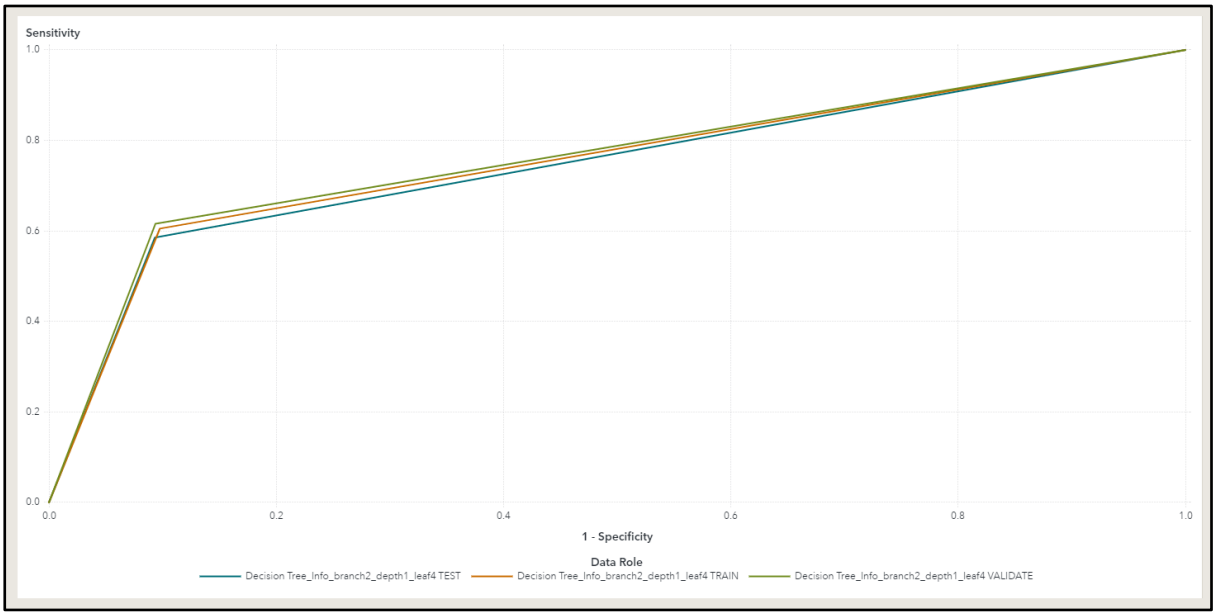

Figure 22: Decision Tree Info Branch 2 – Accuracy

Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
▣	Gradient Boosting	Gradient Boosting	0.9781	0.0165
	Decision Tree_Gini_branches2_depth10_leaf5	Decision Tree	0.9698	0.0151
	Decision Tree_info_branch5_depth10_leaf15	Decision Tree	0.9039	0.0480
	Ensemble_decision_tree	Ensemble	0.8923	0.0536
	Decision Tree_Info_branch2_depth1_leaf4	Decision Tree	0.4920	0.2538

Figure 23: Comparison - All Models

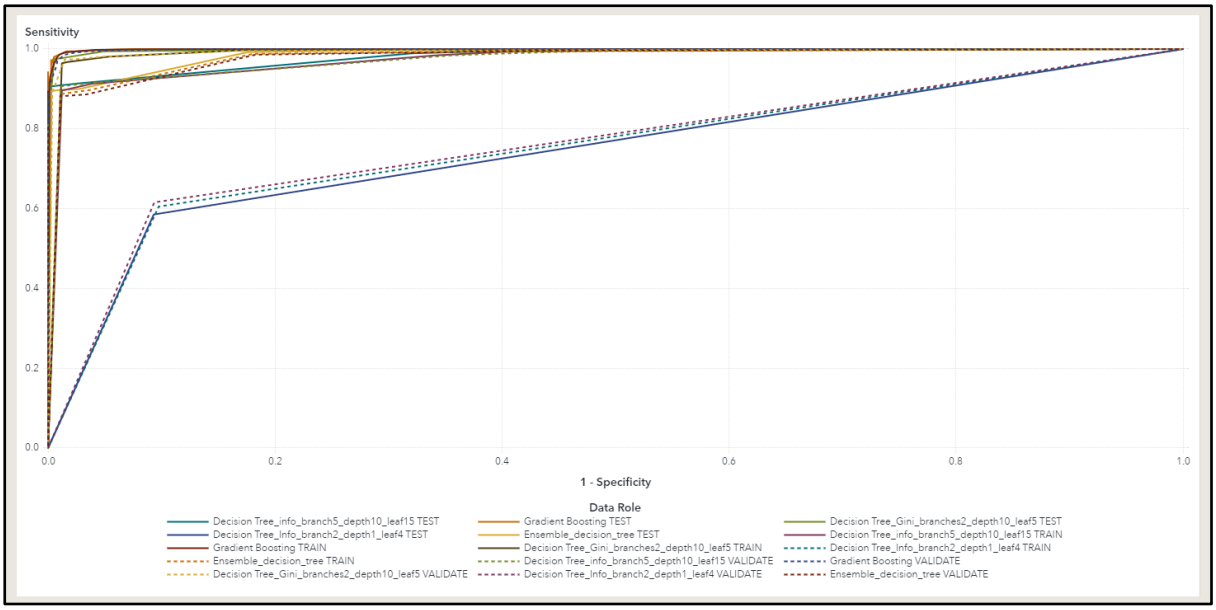

Figure 24: ROC - All models

Appendix 4: Clustering

Manage Variables:

Variable Name	Label	Type	Role ↓	Level	Order
Churn_copy		Character	Target	Binary	Default
children		Character	Rejected	Null	Default
F1		Numeric	Rejected	Interval	Default
marry		Character	Rejected	Null	Default
Number of Records		Numeric	Rejected	Unary	Default
occupation		Character	Rejected	Nominal	Default
children_enc		Character	Input	Binary	Default
churn		Character	Input	Binary	Default
credit		Numeric	Input	Nominal	Default
creditCard		Character	Input	Binary	Default
custcare		Numeric	Input	Interval	Default

Figure 25: Columns Selected for Clustering

Variable Name	Label	Type	Role ↓	Level	Order
custcareLast		Numeric	Input	Interval	Default
custcareTotal		Numeric	Input	Interval	Default
directas		Numeric	Input	Interval	Default
directasLast		Numeric	Input	Interval	Default
dropvce		Numeric	Input	Interval	Default
dropvceLast		Numeric	Input	Interval	Default
family		Numeric	Input	Binary	Default
income		Numeric	Input	Interval	Default
marry_enc		Character	Input	Nominal	Default
mou		Numeric	Input	Interval	Default
mouChange		Numeric	Input	Interval	Default

Figure 26: Columns Selected for Clustering

mouTotal		Numeric	Input	Interval	Default
outcalls		Numeric	Input	Interval	Default
overage		Numeric	Input	Interval	Default
overageMax		Numeric	Input	Interval	Default
overageMin		Numeric	Input	Interval	Default
peakOffPeak		Numeric	Input	Interval	Default
peakOffPeakLast		Numeric	Input	Interval	Default
recchrge		Numeric	Input	Interval	Default
regionType		Character	Input	Nominal	Default
revenue		Numeric	Input	Interval	Default
revenueChange		Numeric	Input	Interval	Default
revenueTotal		Numeric	Input	Interval	Default
roam		Numeric	Input	Interval	Default
customerID		Numeric	ID	Interval	Default

Figure 27: Columns Selected for Clustering

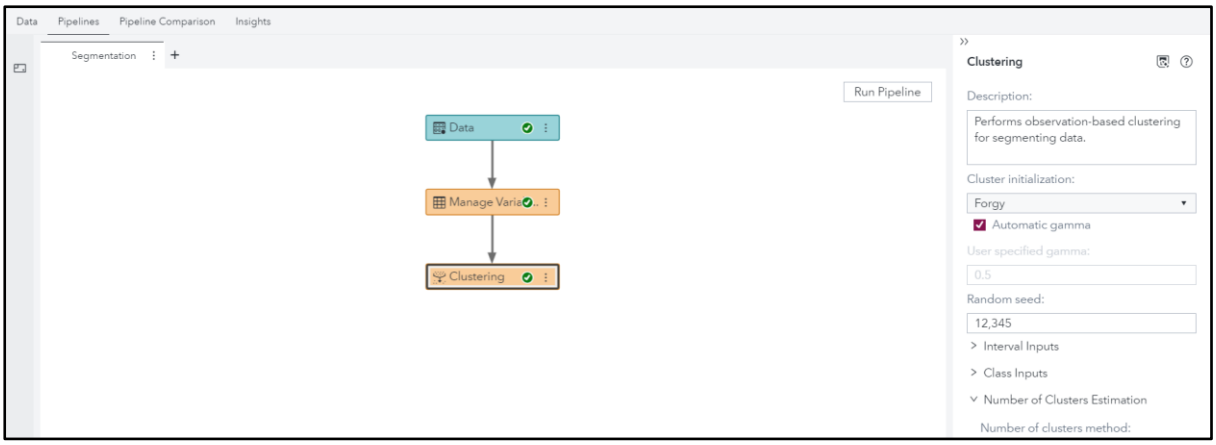

Figure 28: Model Information

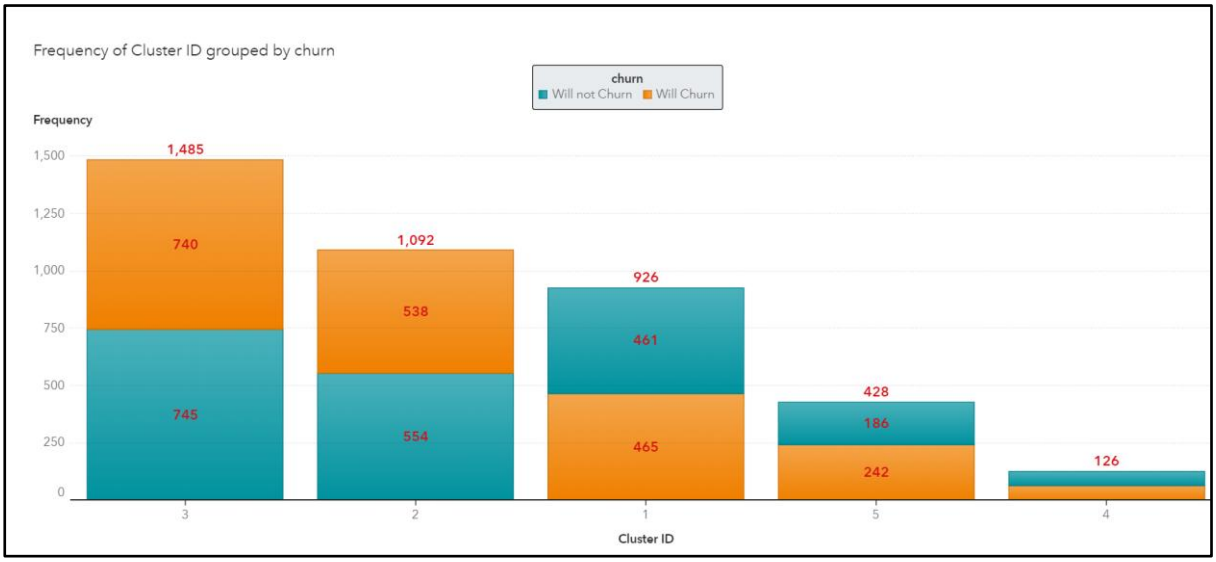

Figure 29: Cluster Id grouped by Churn rate

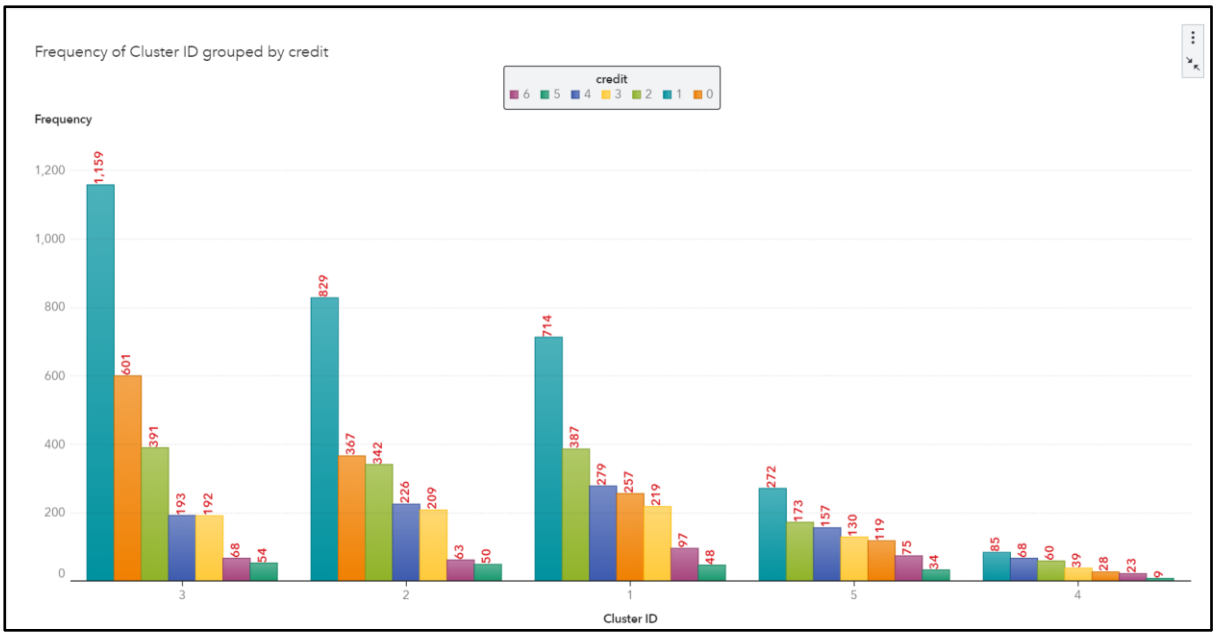

Figure 30: Credit Ratings in each cluster

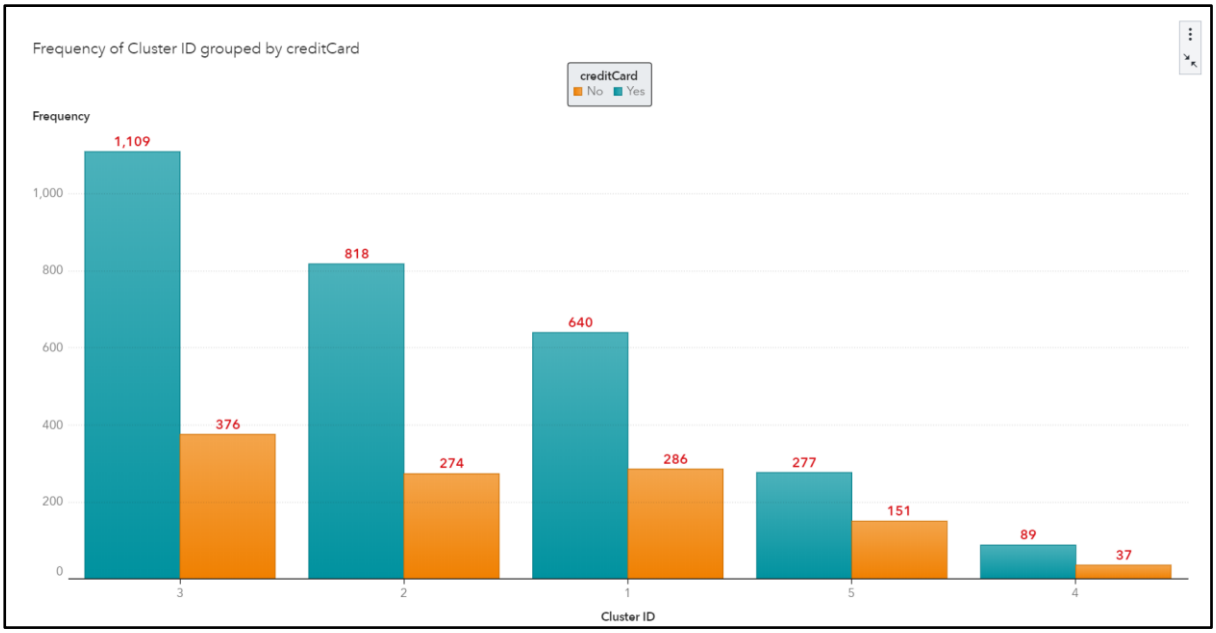

Figure 31: Credit Card holders in each cluster

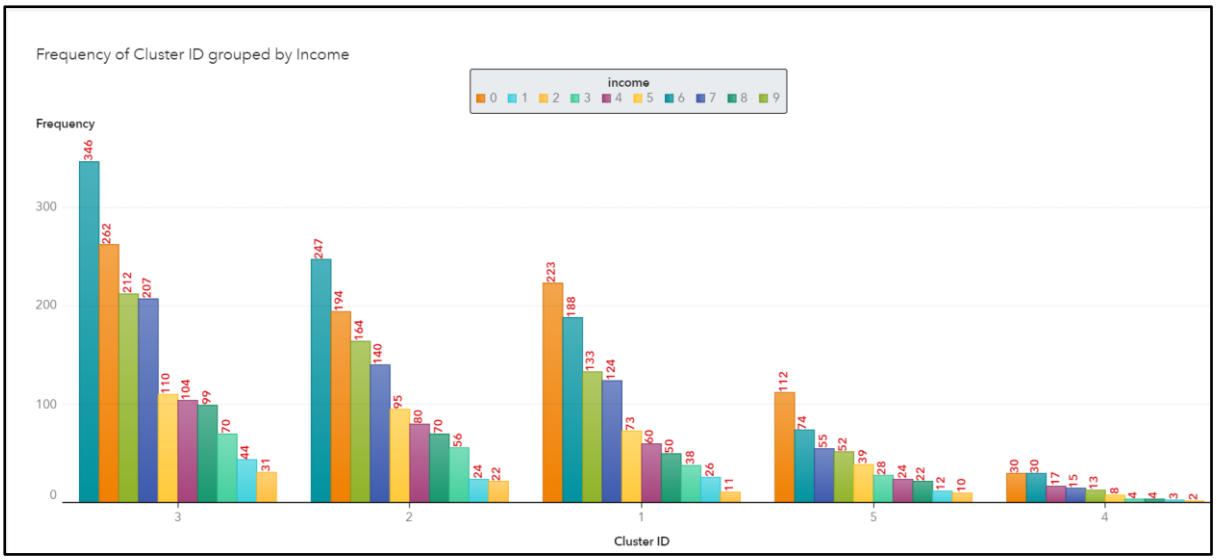

Figure 32: Cluster Id grouped by Income

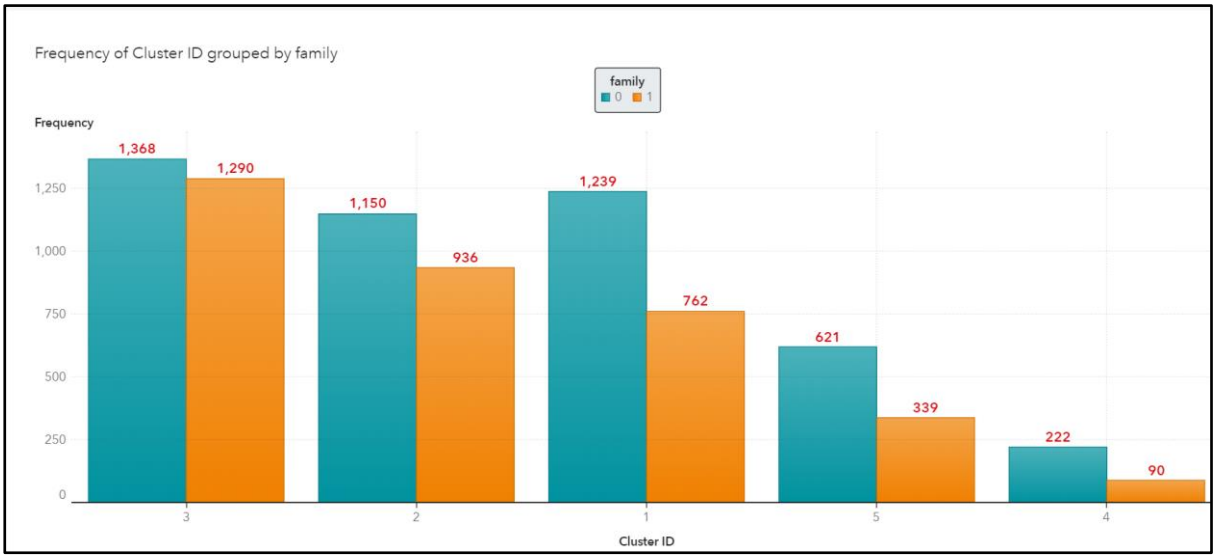

Figure 33: Family & Individuals in each cluster

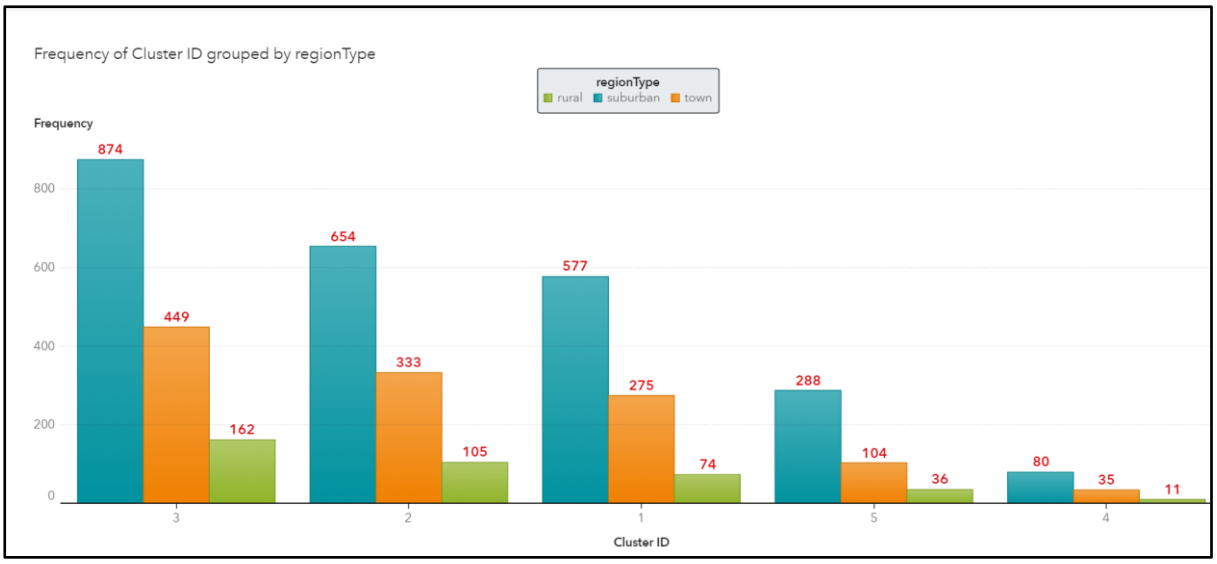

Figure 34: Cluster Id grouped by region type

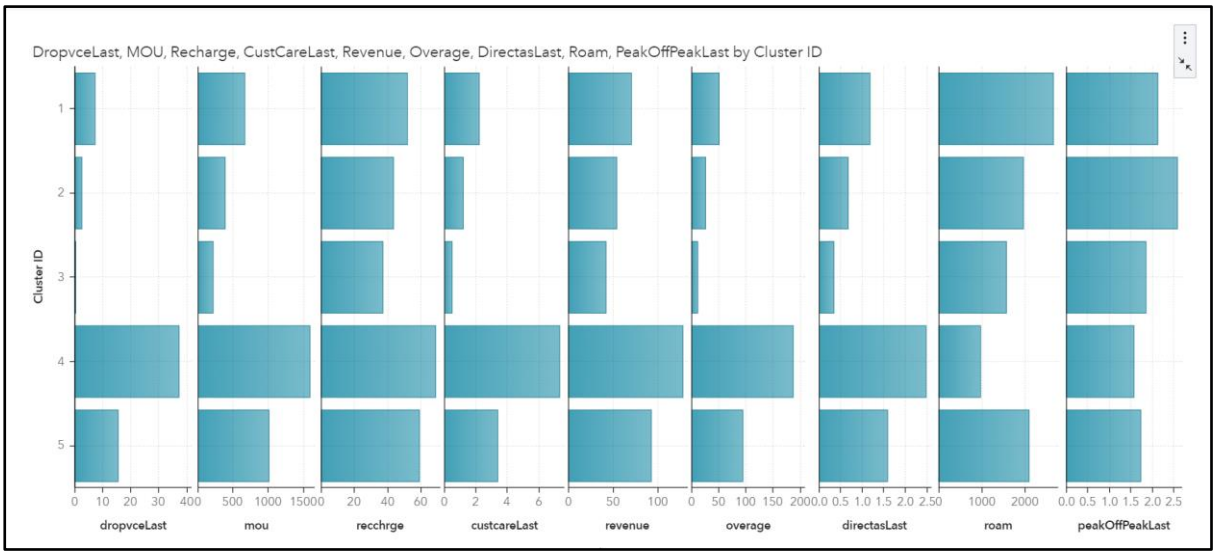

Figure 35: Various Measures analysed in cluster Ids

Appendix 5: Recommendations

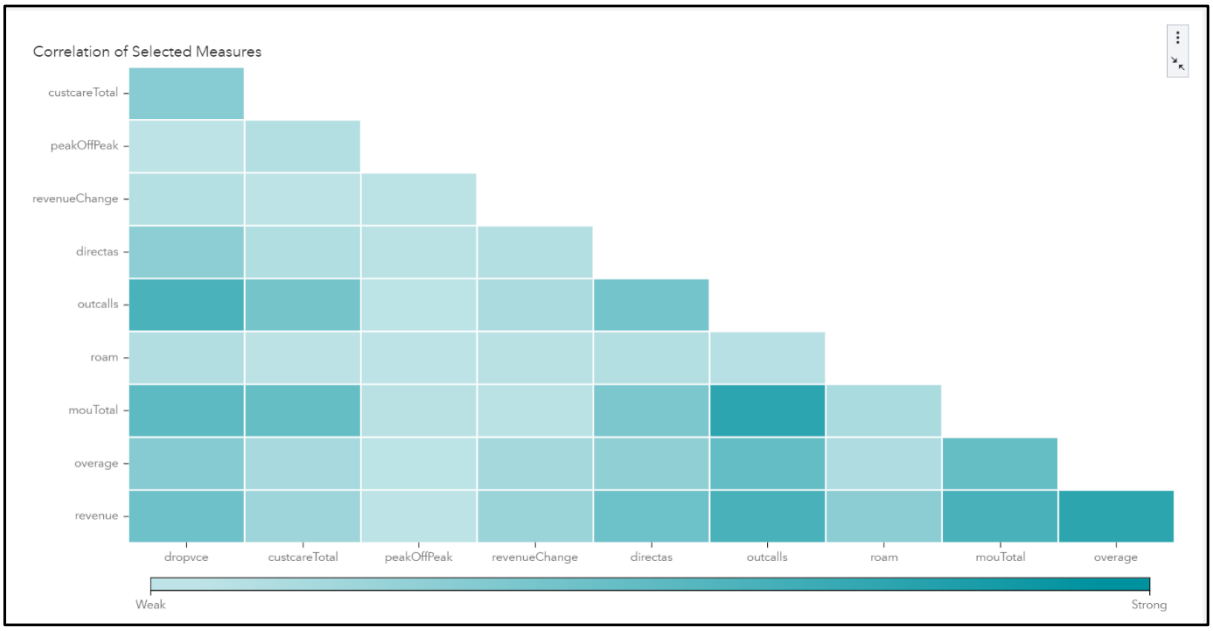

Figure 36: Correlation of selected Measures

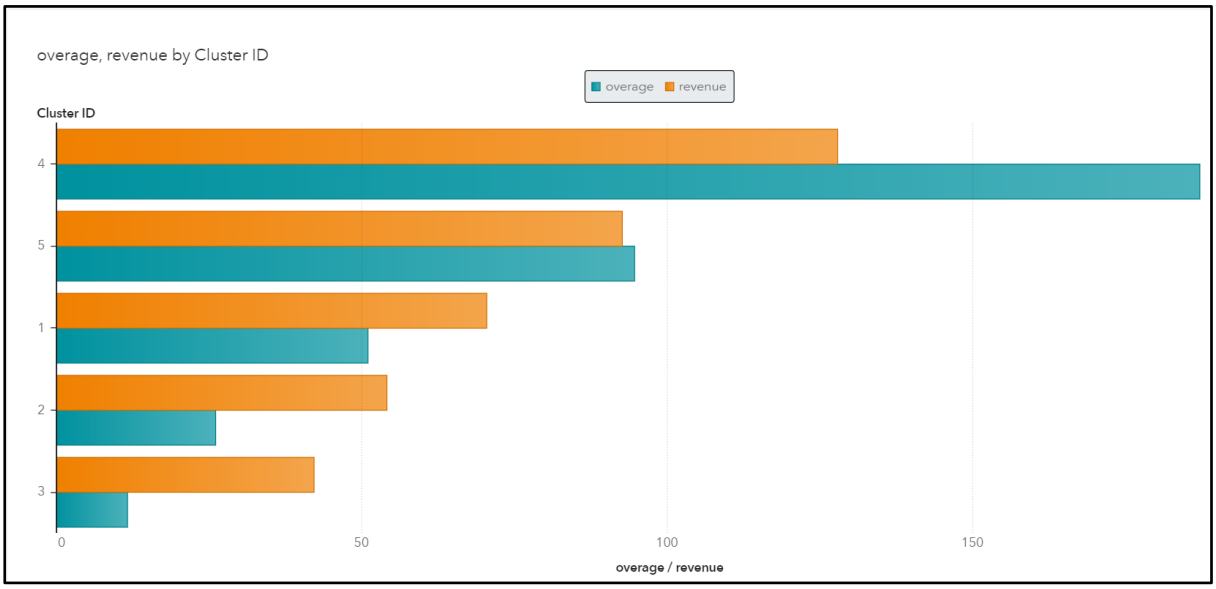

Figure 37: Overage, Revenue in clusters

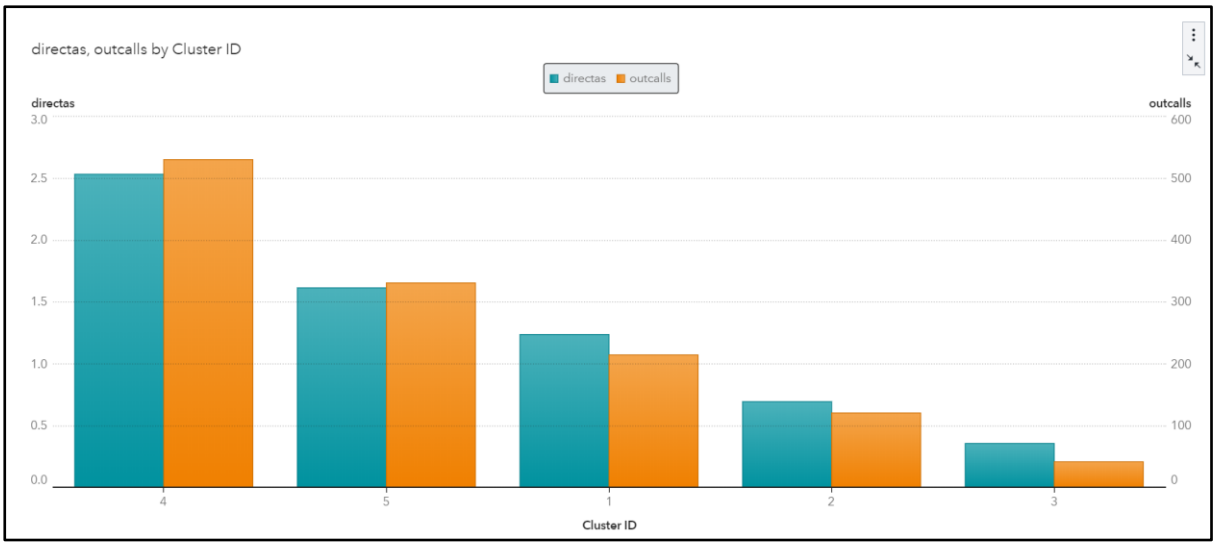

Figure 38: Directas, Outcalls in clusters

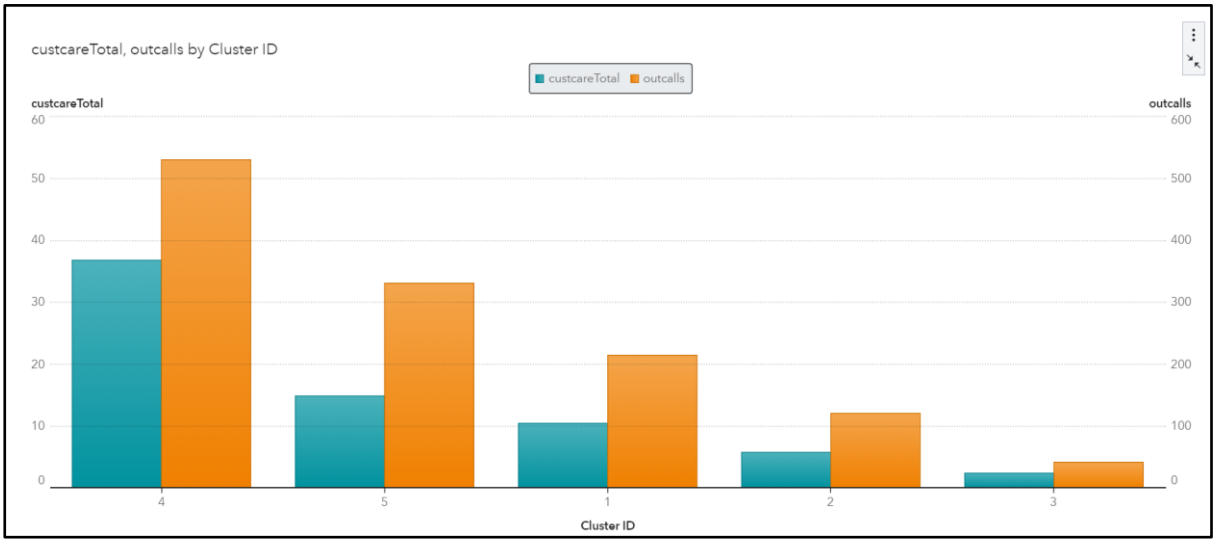

Figure 39: CustcareTotal, Outcalls in clusters

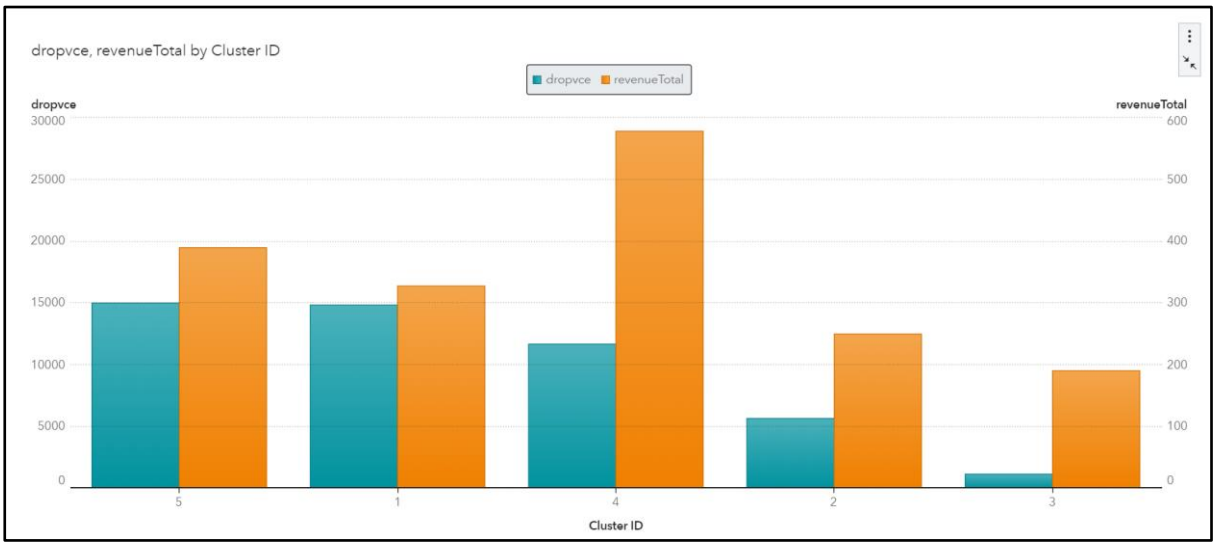

Figure 40: Drop calls, RevenueTotal in Clusters

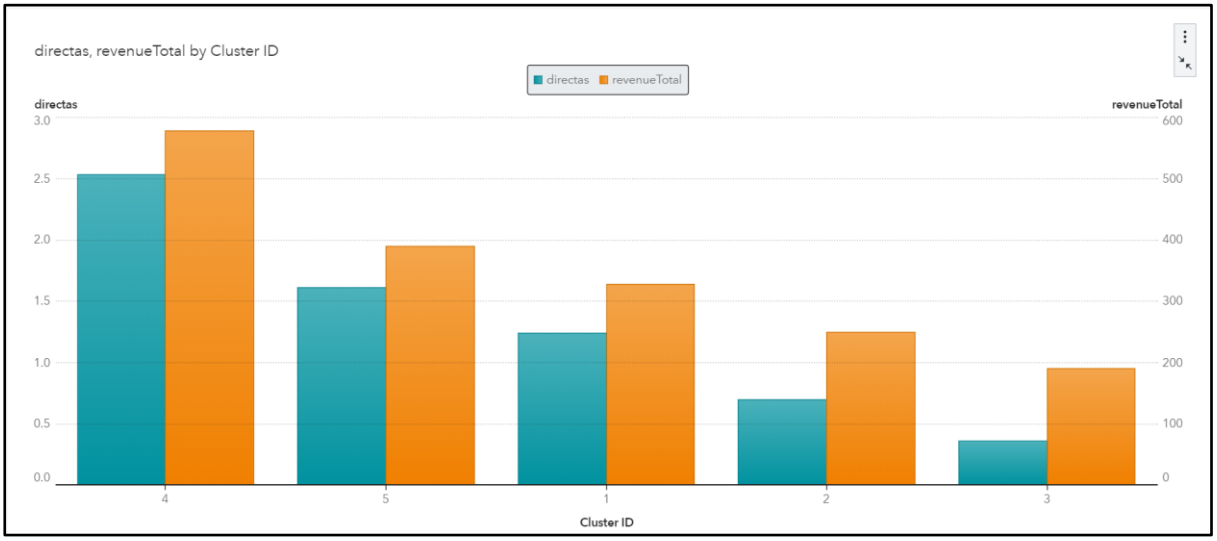

Figure 41: Directas, RevenueTotal in cluster

Appendix 6: Worklog

M.O.M of scheduled calls: Attached separately in submission folder.

Detailed split of tasks among each team member is as below:

Week Ending	Topic	Team Member
08.03.2020	Requirement Understanding & Project Planning	All
15.03.2020	Created high level design of solution & Technical requirements. Team decided to use Python for designing the predictive model.	All
15.03.2020	Data analysis. Data cleaning strategies - Dropping of columns, Replacing abbreviated values, setting metadata limits - Python	Chaitanya, Bhagyashree
15.03.2020	Descriptive Statistics - Python	Kaushal, Siddhesh
22.03.2020	Implementation of decision tree prediction model based on various combinations - Using Python	Bhagyashree, Sidddhesh
22.03.2020	Implementation of Gradient Boosting, SVM, K-Means Prediction model based on various combinations - Python	Kaushal, Chaitanya
29.03.2020	Ensemble Models in Python	All
29.03.2020	Clustering in Tableau	All
05.04.2020	Implementation of prediction model - Using SAS	Bhagyashree, Siddhesh
05.04.2020	Clustering - Using SAS	Kaushal, Chaitanya
12.04.2020	Implementation of prediction model - Using SAS	Kaushal, Chaitanya
12.04.2020	Clustering - Using SAS	Bhagyashree, Siddhesh
19.04.2020	Report Preparation	All
26.04.2020	Report Preparation	All

Appendix 7: References

1. Julien Boudet, Brian Gregg, Jason Heller, and Caroline Tufft,The heartbeat of modern marketing: Data activation and personalization |March 2017, https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-heartbeat-of-modern-marketing
2. Misra, Richa & Mahajan, Renuka & Mahajan, Vishal. (2017). Review on factors affecting customer churn in telecom sector. International Journal of Data Analysis Techniques and Strategies. 9. 122 10.1504/IJDATS.2017.10006960., https://www.researchgate.net/publication/319023470_Review_on_factors_affecting_customer_churn_in_telecom_sector
3. Jain, P. and Surana, k., 2017. Reducing Churn in Telecom Through Advanced Analytics. McKinsey & Company. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/reducing-churn-in-telecom-through-advanced-analytics#>