# Telecom Customer Segmentation and Churn Prediction

Raman Sharma
Ashish Mamgain
Ritesh Sharma
Kartik Aggarwal
Great Lakes Institute of Management
Gurgaon, India

Abstract— This study covers implementation of two data mining techniques using service usage behavior data of a telecom operator to solve Segmentation and Churn Prediction requirements of the business. In this regard it presents steps to achieve customer behavior based segmentation using K-means clustering analytical technique. The study also covers how to validate the results using ANOVA. It also covers ways to identify the correct number of clusters using plots and other metrics. In the other part, it covers creation of an analytical model for predicting Churn subscribers using Logistic Regression technique for different levels of probabilities that could be decided by the business customer. The resultant metrics are benchmarked using Random Forest method to stack performance of Logit model against an ensemble method and use output of one technique in another to improve its performance.

Keywords: Segmentation, Behaviour, Churn, K-means clustering

## INTRODUCTION

Telecommunications is a service business based on a very large number of subscribers. It's imperative for a service oriented business to learn about its customers and tailor services and promotional campaigns according to the customer needs

While this can be easily accomplished in small-scale businesses where the service provider knows its customers directly and can easily identify segments within the customers and serve accordingly it becomes much complicated to know individual customer when the numbers run into millions.

However, since Telecom is a technology enabled business it deals with huge amount of data related to its customers, be it demographic, behavioral or even social. This is where data mining techniques help in exploiting customer data to develop customized customer relationship management strategies such as Customer Segmentation and Churn Management.

Customer segmentation is a process of dividing a large set of heterogeneous customers into groups that have similar characteristics, behavior or needs (Kotler & Armstrong, 2005).Four types of customer segmentations are targeted in telecom customers base - customer value segmentation, customer behavior segmentation, customer life-cycle segmentation & customer migration segmentation.[3]

Most commonly segmentation is based on demographic factors or customers' views and beliefs. However, as the customer demographics and opinions do not correlate well with their actual behavior, there is a strong support for behavioral segmentation (Saarenvirta, 1998).

The behavioral segmentation enables customers to be segmented based on occasion and time, usage rate, benefits sought, user status and loyalty status (Kotler & Armstrong, 2005). Companies benefit from the behavioral segmentation, as they are able to gain better understanding of their customers' actual behavior. Consequently, business and marketing personnel can make better decisions on marketing strategies, new market opportunities can be discovered and companies can differentiate on the market from the competitors.[3]

A typical customer behavior segmentation can yield something like this:

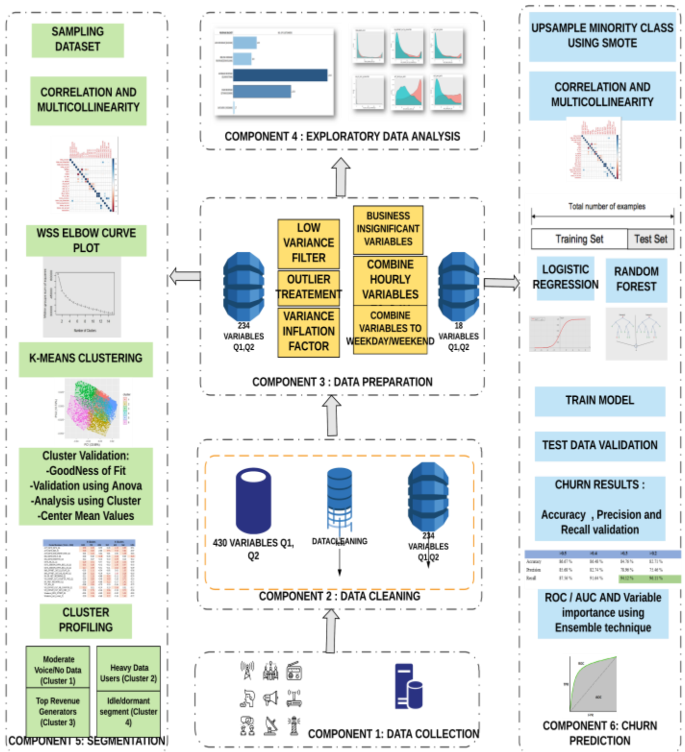
| Segment | Percentage high value |
|---|---|
| Quick talkers | 9 |
| SMS but will talk | 43 |
| Valuable roamers | 4 |
| SMS but high top-ups | 19 |
| Outbound voice sociable | 25 |

Figure 1 Customer Behavior Segmentation

While customer segmentation helps in improving marketing productivity by addressing the customer segment with relevant campaigns, another important aspect of customer retention is Churn Management - the art of identifying the valuable customers, who are likely to churn from a company and executing proactive steps to retain them.[1]

Predicting customer churn is a critical requirement of many if not all companies dependent on customer subscription services. The telecommunication sector is especially impacted due to the rival competition being very high and since tariff rates are maintained at a lower level.[2]

The factors that are considered to have most influence on the customers' likelihood to churn greatly vary between markets and customer segments. In telecom, the most commonly mentioned in recent literature include network quality, handset, cost structure, customer service and brand image. Hence when understanding modeling retention predictions, the technique used depends on the kind of data available.[5]




## BUSINESS PROBLEM

Telecom operators, like other businesses, would like to address their customer pool with tailored segmentation strategies to get more out of their sales & marketing campaigns. That being the starting point during customer acquisition phase what happens with those segments and how do they behave once they are part of the operator's network is equally important. It needs to be learnt what tweaks are required in the segments and how to re-classify subscribers based on their actual behavior than what was initially thought of while segmenting them.

Also, while more and more subscribers join the network a few are planning to leave as well for various voluntary and in-voluntary reasons and would soon become Churned pool.

In order to learn and act on the basis of actual service usage behavior analytical models using data mining techniques needs to be built. The data based segmentation and churn models should address handling huge data exhibiting Big Data properties' such as 5-Vs.

This study looks at two separate analytical models for segmentation and churn prediction. The models are easily interpretable and gives flexibility for business intervention to get desired results. The study has presented an approach to use Churn model output to understand Churn subscriber's segment using clustering.

## DATA MINING TECHNIQUES FOR SEGMENTATION AND CHURN PREDICTION IN TELECOMMUNICTION

Clustering and Prediction are two key subsets of data mining exercise.

The most commonly used data mining technique for solving the problem of customer grouping is cluster analysis. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.

**K-means clustering:** As a partitioning method, the K-means algorithm is the most popular heuristic methods adopted for most business applications. In this method, the sum of discrepancies between a point and its centroid (mean) expressed through appropriate distance is used as the objective function.[

Since we were not looking at granularity beyond cluster characteristics identification we didn't use hierarchical approach.

The advantages of using K-means clustering method are:

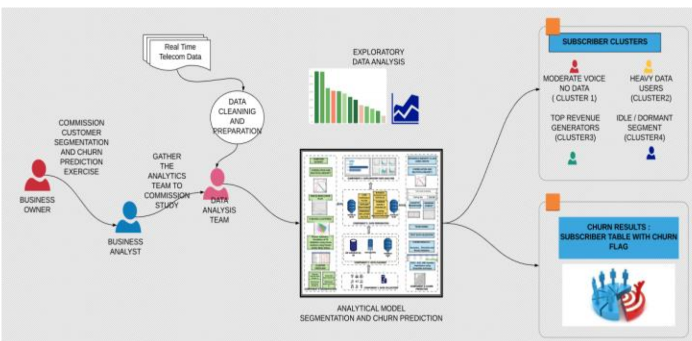
- The method is not sensitive to the input order of data.
- The method is relatively scalable and efficient in processing large data-sets.
- The method is fast in modeling and its results are relatively easier to interpret.

Disadvantages:

- The method is sensitive to the presence of outliers in the data. Outlier treatment is a must.
- The method only works with numerical attributes.

Logistic regression, Neural Networks, Decision Trees, Support Vector Machines are some of the commonly used methods for predicting churn. A review of DM techniques for Churn prediction is avaialble in [6] and [7].

**Logistic Regression:** Logistic Regression is one of the basic and popular algorithms to solve a classification problem. It is named as 'Logistic Regression' because it's underlying technique is quite the same as Linear Regression. The term "Logistic" is taken from the Logit function that is used in this method of classification.
Our reason for choosing logistic regression is logistic classification models are fast and highly interpretable. It doesn't require input features to be scaled, it's easy to regularize and it outputs well calibrated probabilities. It also allows user defined cut-off thresholds to play extraneous inputs to the model to get desired Churn numbers. It also gives an insight into variables that are contributing to churn behavior.




## DATA ACQUISITION AND PRE_PROCESSING

**Data source:**
The study is focused on analyzing subscriber's behavior on the network therefore real-life data was arranged for the analysis.
The dataset used for the study is a sample of 4000 pre-paid subscribers of a Telecom provider. The dataset provides us transactional information around subscribers collated from network nodes, intelligent network system and other OSS/BSS systems such as CRM, Mediation etc.
At a high level dataset consists of following information:

- Weekly subscriber level information (24 Weeks)
- Subscriber last registration Date.
- Subscriber Activation Date.
- Day of Last Activity (churn is considered if DOLA<=90)
- Base station information of first call Latitude and Longitude (including day/night latitude and Longitude details:
- Revenue Voice Offnet – Revenue generated on Offnet calls, outside local network
- Revenue Voice Onnet- Revenue generated on Onnet calls, Inside local network
- Revenue Voice International- Revenue generated in International calls
- Total Incoming/Outgoing revenues from the user
- Voice Onnet calls incoming seconds and counter (Hourly level data/ day Level available)
- Voice Offnet calls incoming seconds and counter (Hourly Level data/Day level available)
- Data used in KB (Hourly Level Data)
- Revenue details at subscriber level (Day level data available)

Raw dataset consisted of over 100,000 rows of weekly data across 430 variables.

**Preprocessing:**
Various data cleaning steps (Fig.1) were performed on the raw dataset to bring it into a format that suited analytical model requirements.

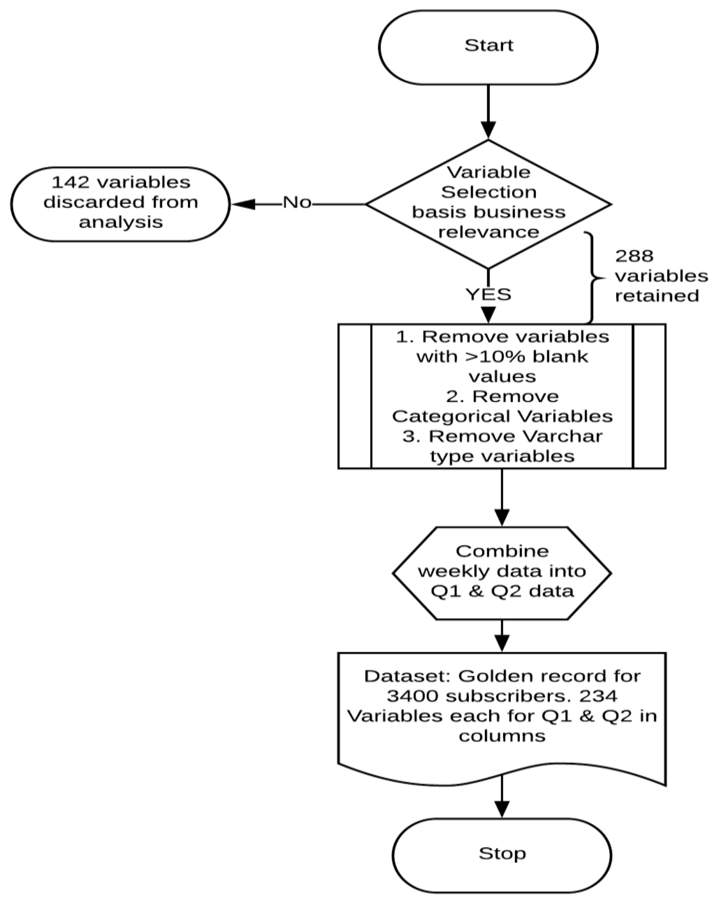



Figure 1 Data Pre-processing tasks

## DATA PREPARATION

Dataset obtained after pre-processing contained golden record i.e. all variable columns listed against subscriber MSIDN number and all subscriber MSISDN numbers sorted in rows. This pre-processing of raw data reduced over 100,000 rows data into 3400 rows ( minus those subscribers that didn't have data for all 24 weeks). Analyzing such a dataset is much easier using statistical tools such as R, Excel & SQL.

**Dimensionality reduction:** Starting with 96 variables after data cleaning and pre-processing, at every stage of data preparation we got rid of redundant and irrelevant variables still present in the data. With the help of VIF method and applying business logic (such as revenue related variable) we finally selected 18 variables, representing all major categories of data including Voice, Data, Roaming, SMS, Revenue, Balance, Weekday & Weekday usage.

Preparing the dataset to be used for analytical models involved many steps included:

- Filtering out variables with Zero values, low variance overall
- Outlier treatment
- Combining interval variables into one variable giving overall quarter number
- Converting all numbers into quarterly number for each MSISDN
- Creating summary variables for Weekday & Weekend usage for Voice, SMS & Data usage
- Removing exact (correlation >99%) collinear variables
- Selecting variables using VIF method, choosing variables with VIF <= 4
- Removing handset indicator variables

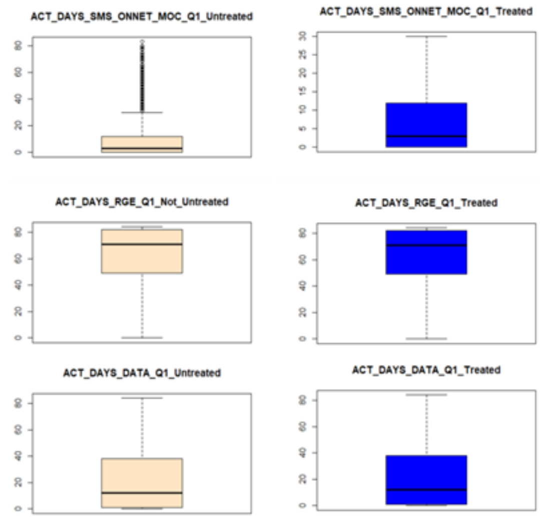



Figure 2 Original distribution (Pink) Vs Outlier treated (Blue)

## CLASS IMBALANCE

After preparing the churn column, we observed only 6% of the total data is flagged as churn which considered very low and data is highly skewed for majority class. This is a case of class imbalance. In case of class imbalance, the ratio of the output categories is one-sided to the extent that the learning algorithm only predicts the majority class. We have also shown the same by running the logistic regression model and obtain a very poor recall of 34% for the churners. One of the methods to deal with the problem of class imbalance is re-sampling. There are two ways in which we can do that: we can either over-sample or under-sample. In under-sampling, we use only a subset of the majority class in order to train our data. On the contrary over-sampling increases the strength of minority class.

| Churn Flag | Subscribers |
|---|---|
| **Churner** | **188** |
| **Non-Churners** | **3402** |
| **Total Subscriber's** | **3590** |

Original Churn

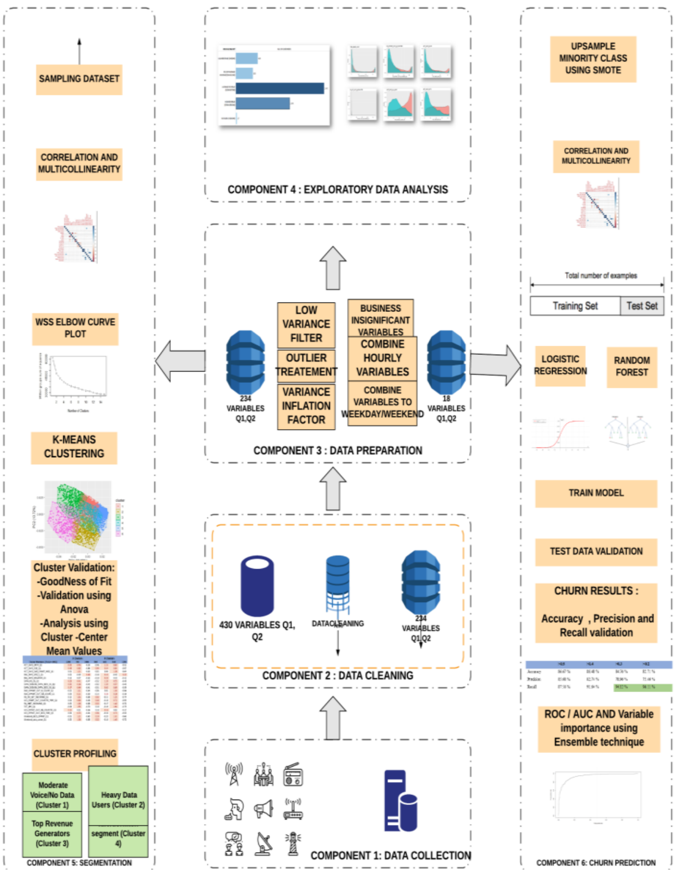
| Churn Flag | Subscribers |
|---|---|
| Churner | 3402 |
| Non-Churner | 3402 |
| **Total Subscriber** | **6804** |

SMOTE Result

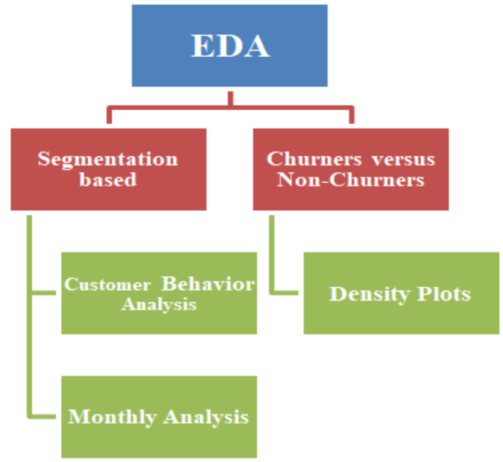
To facilitate over-sampling we used the help of the resampling technique SMOTE (synthetic minority oversampling technique. SMOTE algorithm creates artificial data based on feature space (rather than data space) similarities from minority samples.

### Analytical Model for Segmentation & Churn Prediction

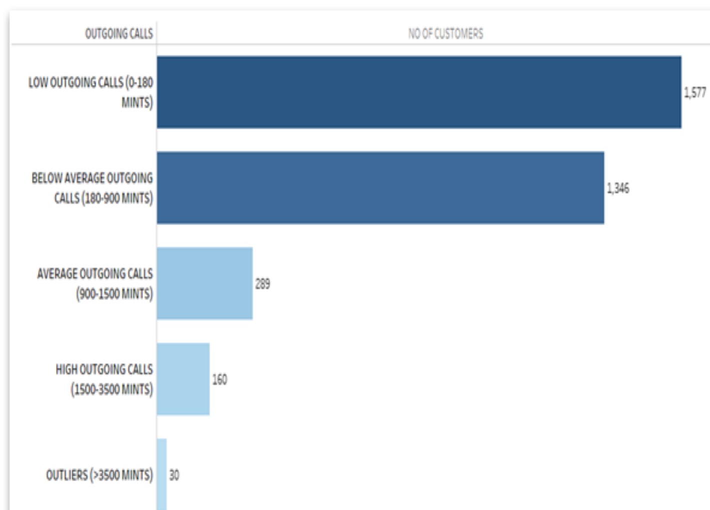
An architectural representation of the models used for Segmentation and Churn prediction is shown below in Fig.




### EXPLORATORY DATA ANALYSIS




- Customer Behaviour Analysis



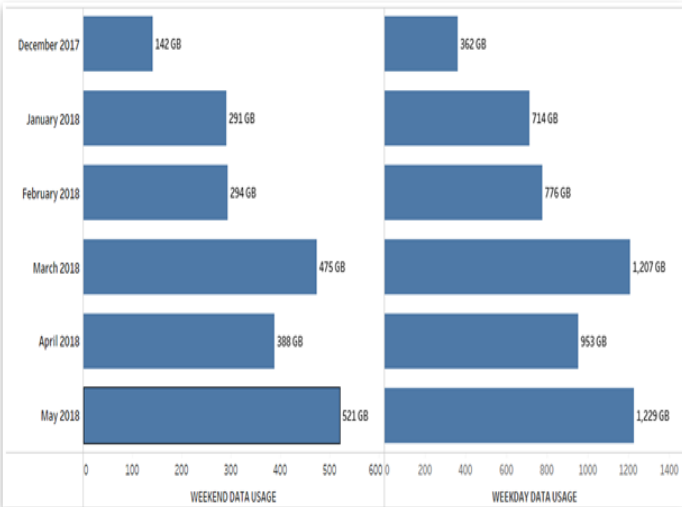

Revenue Buckets

Revenue generation from each user is one of the most important metric for any business. We can see maximum users coming into an average revenue bucket of $1000-$7000. These customers could be the most important to target as with proper tariff rationalization and tariff association, these customers can well be converted to high revenue users.

Voice Calling Pattern Buckets

Below analysis shows how customers are bucketed in terms of outgoing calls. We have included only local and international calls made by the subscriber as of now to see how good the calling capacity of the users is. We can see most users are making calls less than 900 minutes in 6-month period.

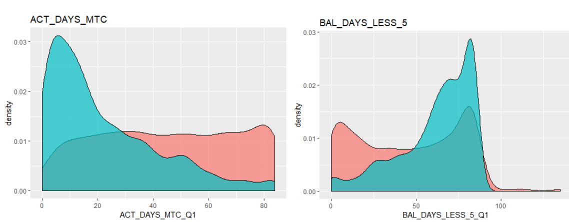
- Weekday/Weekend Analysis




Weekend/Weekday Usage Pattern

On segmentation we can clearly see close competition between weekend and weekday data usage however we are calling weekends only Saturday and Sunday. It is clear from the analysis the data usage increase on weekends and this might also be a good factor for segmentation and targeting.

- Density Plots




We have plotted the churners (red) versus the non-churners (blue) using the density plots. As we can see from our visualizations for some variables, the peaks of the non-churners are higher than those of the non-churners.

## CLUSTERING

- Dataset was normalized using Z-score normalization and converted into a matrix
- Using R, WSS Curve was plotted to make an initial assumption about k value = 3
- K-means technique was applied with k- value of 3,4,5,& 6 to observe different cluster formations and apply other metrics to arrive at final no. of clusters

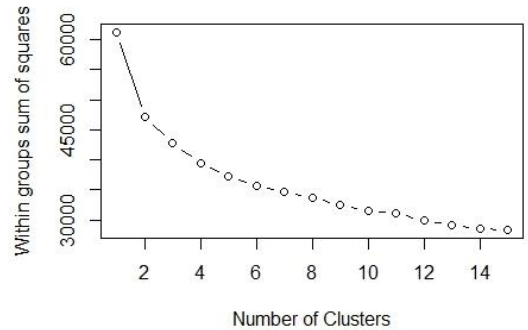



Figure 4 WSS plot to identify clusters in data

- K-means output generates cluster plots representing principal components in a two dimensional space
- Every row of data is labeled with cluster number and cluster means for every variable are calculated and generated as clustering output
- Using cluster mean output table, characteristics of each cluster is understood and cluster profiling is done

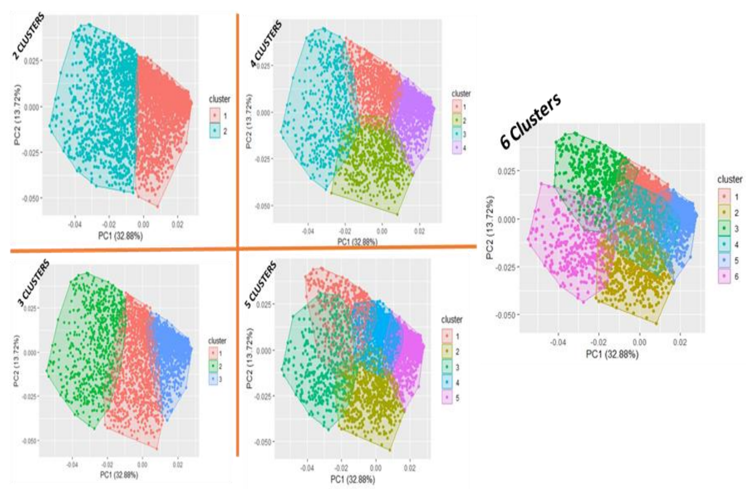



Figure 5 Cluster plots ( 2,3,4,5,6 clusters)

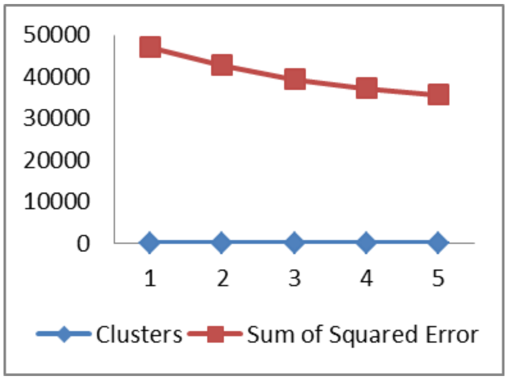
**Cluster plots show that as number of clusters increase, smaller clusters get more compact but cluster boundaries start to get meshed up. Later we will see in cluster goodness of fit table how SSE values drop with higher number of clusters.**

**Cluster goodness of fit (selecting clusters using SSE metrics)**

Cluster Goodness of fit metrics

| Clusters | WSS | % reduction | SSE*#Clusters | Between SS | Between_SS / total_SS |
|---|---|---|---|---|---|
| 2 | 47079.5 | 0% | 94159 | 14138.49 | 0.231 |
| 3 | 42793.5 | -9% | 128380.59 | 18424.47 | 0.301 |
| 4 | 39382.5 | -16% | 157530.072 | 21835.48 | 0.357 |
| 5 | 37175.4 | -21% | 185876.855 | 24042.63 | 0.392737914 |
| 6 | 35696.6 | -24% | 214179.834 | 25521.36 | 0.416893077 |

* WSS-Within Sum of Squares




Since visual inspection of cluster plots are not conclusive in determining the optimum number of clusters, especially when number of clusters are of the order of 100, reduction in SSE metric is a useful metric to analyze number of clusters. To account for natural decrease in SSE value due to splitting of big clusters into small, the number of clusters is multiplied by SSE value. SSE curve helps determining the elbow from where sharp decrease in SSE is noticed, in our study it was observed at 4. We will further do ANOVA test to conclude on the number of clusters.

## CLUSTERING VALIDATION USING ANOVA

To find out if the clusters are created such that the cluster mean values for different model variables are different we used ANOVA method to test the below hypotheses

=> Null (μo)) Hypothesis: Cluster means are equal

=> Alternate (μ1) (Hypothesis): Cluster means are different

As the number of clusters increase, mean of some of the variables across clusters are observed as same. This is not a desirable outcome. One must keep this in mind while weighing the reduction of SSE due to high number of clusters.

| Variable Name | 3 clusters | 4 clusters | 5 clusters | 6 clusters |
|---|---|---|---|---|
| ACT_DAYS_DATA_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| ACT_DAYS_RGE_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| ACT_DAYS_SMS_ONNET_MOC_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| BAL_DAYS_LESS_5_Q1 | Mean is different | Mean is same | Mean is same | Mean is different |
| BAL_DAYS_NEGATIVE_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| DATA_KB_2G_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| DATA_SESSION_DRTN_SECS_2G_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| DATA_SESSION_DRTN_SECS_3G_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| SMS_OFFNET_OUT_B_COUNT_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| SMS_OFFNET_OUT_NB_COUNT_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| SN_ON_NET_INCOMING_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| VOI_ONNET_OUT_COUNTER_FREE_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| SN_XNET_INCOMING_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| TOT_REV_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| VOI_OFFNET_OUT_NB_COUNTER_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| VOI_OFFNET_OUT_SECS_FREE_Q1 | Mean is different | Mean is different | Mean is different | Mean is same |
| Weekend_SECS_OFFNET_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |
| Weekend_secs_onnet_Q1 | Mean is different | Mean is different | Mean is different | Mean is different |

Figure 6 ANOVA test summary

**Cluster profiling**

The clustering exercise produced 4 discernible segments of subscriber's basis their service consumption and other relevant parameters.

- Moderate Voice/No Data (Cluster 1): 954 members,
- Heavy Data Users (Cluster 2): 633 members
- Top Revenue Generators (Cluster 3): 560 members
- Idle/dormant segment (Cluster 4): 1255 members

| | | Moderate Voice/No Data | Heavy Data Users | Top Revenue generators | Idle/dormant |
|---|---|---|---|---|---|
| Cluster Members ( Total = 3402) | | 954 | 633 | 560 | 1255 |
| ACT_DAYS_DATA_Q1 | | -0.46 | 1.15 | 0.63 | -0.51 |
| ACT_DAYS_RGE_Q1 | | 0.55 | 0.34 | 0.84 | -0.97 |
| ACT_DAYS_SMS_ONNET_MOC_Q1 | | -0.01 | 0.06 | 1.38 | -0.64 |
| BAL_DAYS_LESS_5_Q1 | | -0.18 | 0.16 | -0.40 | 0.23 |
| BAL_DAYS_NEGATIVE_Q1 | | -0.10 | 0.29 | 0.10 | -0.12 |
| DATA_KB_2G_Q1 | | -0.52 | 1.28 | 0.55 | -0.49 |
| DATA_SESSION_DRTN_SECS_2G_Q1 | | -0.52 | 1.30 | 0.37 | -0.43 |
| DATA_SESSION_DRTN_SECS_3G_Q1 | | -0.31 | 0.62 | 0.69 | -0.39 |
| SMS_OFFNET_OUT_B_COUNT_Q1 | | -0.05 | 0.01 | 1.33 | -0.56 |
| SMS_OFFNET_OUT_NB_COUNT_Q1 | | -0.14 | 0.28 | 0.28 | -0.16 |
| SN_ON_NET_INCOMING_Q1 | | 0.32 | -0.11 | 1.25 | -0.74 |
| VOI_ONNET_OUT_COUNTER_FREE_Q1 | | 0.40 | -0.18 | 0.72 | -0.53 |
| SN_XNET_INCOMING_Q1 | | 0.22 | -0.17 | 1.42 | -0.72 |
| TOT_REV_Q1 | | 0.10 | -0.16 | 1.64 | -0.73 |
| VOI_OFFNET_OUT_NB_COUNTER_Q1 | | 0.10 | 0.16 | 0.02 | -0.17 |
| VOI_OFFNET_OUT_SECS_FREE_Q1 | | 0.36 | -0.18 | 0.71 | -0.50 |
| Weekend_SECS_OFFNET_Q1 | | 0.19 | -0.20 | 1.31 | -0.63 |
| Weekend_secs_onnet_Q1 | | 0.22 | -0.16 | 1.43 | -0.72 |

Figure 7 Cluster profiling using clusters mean analysis

## CHURN PREDICTION

- Dataset was up-sampled for minority class 'Churn' to correct imbalance in the dataset
- Logit model was run with churn variable as the dependent variable. Confusion matrix was created using 0.5 probability as threshold value for cut-off to differentiate between 0 and 1, where 1 is likely to Churn
- Model was further revised keeping only significant and business relevant variables in the model and 'Recall' was monitored at every iteration. Probability threshold was also changed

| **Probability** | **>0.5** | **>0.4** | **>0.3** | **>0.2** |
|---|---|---|---|---|
| Accuracy | 86.67% | 86.48% | 84.76% | 82.71% |
| Precision | 85.68% | 82.74% | 78.96% | 75.46% |
| Recall | 87.56% | 91.64% | 94.12% | 96.11% |

Recall = TP/TP+FN - Recall is the ratio of correctly predicted positive observations to all observations in actual class - Yes.

As we see, Recall has improved at every iteration indicating improvement in model's Predictive capability.

AUROC curve - a metric that convey balanced tradeoff between True positives and False positives and indicate model's stability when tested with unseen data. We got 94.2% AUC.

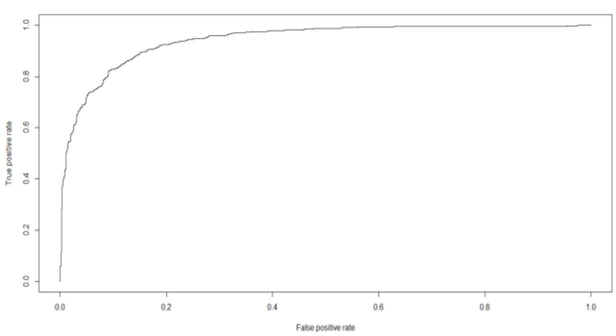



Figure 8 ROC curve

### Comparative Analysis Of Model Accuracy And Variable Importance Using Ensemble Method

**Random Forest:** Random Forest refers to an ensemble of decision trees. In Random Forest, we've collection of decision trees (so known as "Forest"). To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

- Variable importance ranking from RF was used to compare the variable importance of the Logit model output

- Final Logit model was run using variables that were found common in the RF VarImp list

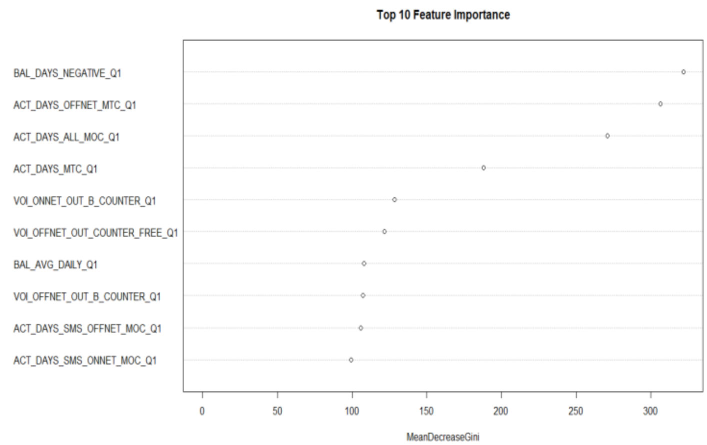



Figure 9Random Forest Top Important Variables

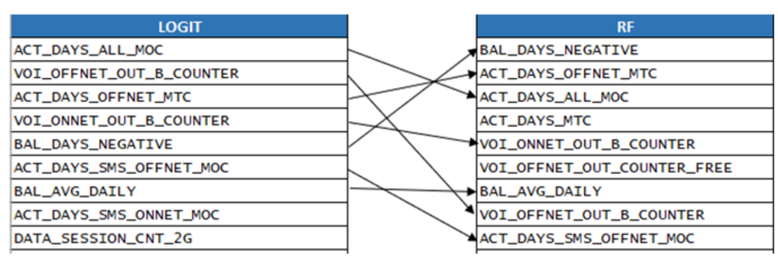



Figure 10 Feature comparisons between Logit and RF

**Logit results using Random Forest variables**

| | >0.5 | >0.4 | >0.3 | >0.2 |
|---|---|---|---|---|
| Accuracy | 86.23 % | 85.48 % | 84.32 % | 81.48 % |
| Precision | 84.96 % | 81.94 % | 78.75 % | 74.02 % |
| Recall | 87.56 % | 91.34 % | 93.33 % | 96.11 % |

**As we see from the results, Logit model (developed with step by step approach of variable selection) and Random Forest have converged and gave almost same Recall results for different probability cut-offs. Other metrics i.e. Accuracy and Precision are also quite similar.**

## BUSINESS VALUE OF ANALYTICAL MODELS

The segmentation model has produced a statistically significant output from a large dataset sample. The model has given a clear view into subsets of subscriber groups in terms of their service usage behaviour on the provider's network.

As the cluster members are labelled with cluster numbers and cluster characteristics are sufficiently explained by cluster means analysis the business users would be able to see the distribution of their subscribers into segments which have distinctly different properties from one another. They can use this information to develop targeted campaigns - whether sales & marketing or customer service related.

This being an automated model it can produce similar results in quick turnaround time on new data.

A value and opportunity matrix as given below can be used to plan customer relationship management strategies for each segment.

Similarly, Churn Prediction model would be helpful in predicting 9.6 out of 10 subscribers that are likely to get Churned out in future. Retention strategies can be deployed beforehand to prevent profitable customers to churn out.

Further, we can use clustering technique to map predicted Churn customers into Segments identified in the Segmentation exercise.


Figure 11 Value Opportunity Matrix for Customer Segments

As we can see from below Value-Opportunity matrix created for Churn subscribers segment, business would adopt different strategies to prevent Churn for each segment and in some cases may satisfy itself with no action at all, for e.g. for dormant/Idle segment subscribers.

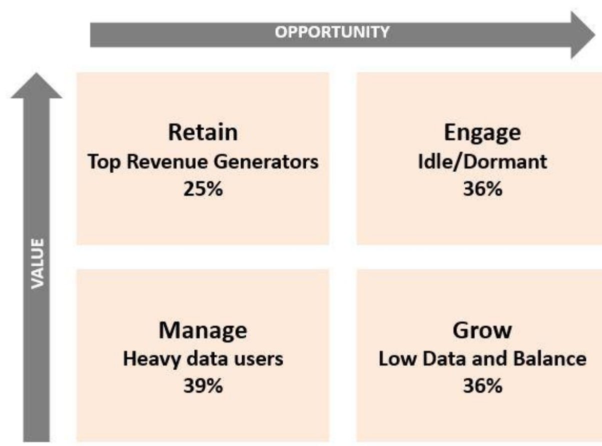



Figure 12 Churn Subscribers Segmentation

## IMPLEMENTATION STRATEGIES

- In Churn Prediction model, percentage of Churn labels in the training dataset is always low and is artificially treated. In a real-time unseen data this can lead to inefficient results
- Missing values in variables or sparsely populated variables even if they are important gets removed from the analysis and it may result in loss of crucial information
- Absence of analytical data mart could be non-starter for such data intensive exercise as data from disparate systems may not be helpful for analytical exercises. In this case lot of crucial time is spent in cleaning up and re-arranging data for every modeling exercise. It defeats business' expectations of quick results from analytics teams.

## CONCLUSION AND FUTURE WORK

** To be done

## REFERENCES

[1] - http://blog.mahindracomviva.com/what-is-churn-management/ - 9/1/19 - 6:30PM
[2] - Churn prediction methodologies in the telecommunications sector: A survey Wanninayake W. M. C. Bandara, A. S. Perera, Damminda Alahakoon
[3] Behavioral Segmentation of Telcommunications Customers - Emilia Mattila
[4] Customer segmentation in the telecommunications industry - Judy Bayer https://link.springer.com/content/pdf/10.1057%2Fdbm.2010.21.pdf
[5] Modeling Customer Lifetime Value in the Telecom Industry- Petter Flordal, Joakim Friberg. pp23

[6] A Survey of Clustering Data Mining Techniques - Pavel Berkhin, sec3.3
[7] Review of Data Mining Techniques for Churn Prediction in Telecom - Vishal Mahajan, Dr. Richa Misra, Dr. Renuka Mahajan-pp188,189
[8] Khakabimamaghani, Sahand & Gholamian, Mohammad & Namvar, Morteza. (2010). Data Mining Applications in Customer Churn Management. Intelligent Systems, Modelling and Simulation, International Conference on. 220-225. 10.1109/ISMS.2010.49. -