Customer Churn Prediction in Telecom Sector

Srishti Lamba[1*]

[1] Symbiosis Centre for Management and Human Resource Development, SCMHRD, SIU, Symbiosis International (Deemed University), SIU, Hinjawadi, Pune, Maharashtra, India

[1] srishti_lamba@scmhrd.edu

ABSTRACT

Purpose: The purpose of this research is to determine factors that lead to customer churn in the telecom sector and to understand how customer response can make an impact on customer churn in the future.
Methodology: It is an exploratory study involving respondents of different age groups and professions selected across India.
Findings: K-Means clustering is the best algorithm that suits the data. It generates 3 clusters and each cluster has a different range of quality index. Thus, companies can use the quality index as an indicator of the Net Promoter Score and identify detractors
Practical Implications: This will help companies identify the customers which might churn in the future. Thus, companies can come up with targeted marketing strategies to curb customer churning.
Originality/Value: This research will focus on unstructured features whereas most of the existing research is based on structured features
Article Classification: Research Paper

Keywords

Customer Churn, K-Means, Support Vector Machine (SVM), Net Promoter Score, Key Performance Indicator (KPI)

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

With the changing consumer mindset, it becomes essential for the companies to identify the key requirements of their consumer and target these consumers with offers that are best suited for them. There are several questions that arise. Are offers enough to retain customers? If not, then how can a company retain its customers?

We are heading towards an era where "data is the new oil". So, with the increasing amount of data in hand, companies can extract knowledgeable information from this data which can help the company to understand the consumer buying behavior and buying pattern in a better way. Big data is an emerging field and is more powerful than the traditional analysis tools being used. The future seems to be driven by big data and machine learning coupled with artificial intelligence and neural networks.

The Telecom industry generates a huge amount of data for every user activity be it calls, SMS, data usage, use of social media apps, games, etc. Using this Big Data, the Telcos can identify the qualitative and quantitative factors that are responsible for customer churn. Customer churn is one of the major reasons for revenue loss for a company because the customer terminates the use of company products or services. It is observed that the cost of acquiring a new customer is nearly 5 times more than the cost of retaining an existing customer. Moreover, the success rate of an incentive offered to an existing customer is 60 - 70% whereas, the success rate for a new customer is nearly 5-20%.

Therefore, the need of the hour for the company is to invest in retaining the existing customers and Analyzing not only their usage patterns but also understanding how the behavioral changes can serve as an early indicator of a churner

$$Cutomer\ churn\ rate = \frac{customer\ lost * 100}{customer\ in\ the\ beginning} \quad \text{eq. (1)}$$

Literature Review

Telecom companies believe that post-paid customers are more loyal as compared to pre-paid customers. Post-paid customers have to inform the company beforehand for subscription cancellation whereas for a pre-paid consumer terminating service is very spontaneous. When a consumer terminates the use of a product or service by his/her own will, it is called voluntary churn. When a company has to terminate the use of service for a particular consumer let's say due to some fraudulent activity, then it is called involuntary churn source exacaster (Customer Retention For Prepaid Base Management) white papers. To prevent customer churn companies, give "incentives" to the customers. But "predicting" the customers which might churn in the future can be a great option. Companies must cross-sell and upsell to their existing loyal customers and roll out proactive campaigns to retain the customers who are losing interest in the company. Shah et al., 2018 have discussed the importance of churn prediction for the organization. This paper provides an overview of what is churn, types of churn like involuntary churn and voluntary churn, causes of churn, applications of churn in various industries namely telecom, insurance, web-based services, etc. It also provides various solutions to mitigate churn, which involves classifying the customers, using targeted emails, smooth onboarding for new customers, and providing offers to the old customers. Most of the available literature carries out a comparative study between the existing models like decision tree, XGBOOST, SVM, etc. Brândușoiu et al.,2016 studied different algorithms like SVM, neural networks, and Bayes classification on a dataset of 3333 users to identify which is the best-suited algorithm. Ahmad et al.,2019 did a comparative study on various algorithms like decision tree, XGBOOST, and random forest. The area under curve is considered as a metric to understand the performance of the model. The use of feature engineering helped to select the prominent features and to carry out an analysis. It is observed that Social Network analysis enhanced the performance of the model. The study concludes that XGBOOST is the best algorithm for churn prediction. A paper by Yıldız & Albayrak,2015 used rotation forest technique and compared it with Antminer and C4.5 decision tree. It is concluded that rotation forest gives better performance over the decision tree algorithms. Ahn et al., 2006 studied the behavior of South Korean consumers for the telecommunication market. They studied various hypotheses based on a number of factors like customer dissatisfaction, switching cost, service usage, and customer status by making use of 5789 large dataset consisting of customer transactions and billing. Iranmanesh et al., 2019 identifies various other industry which are at high churn risk as sports industry, fast food chains, and banking. This paper created a model for churn prediction for the banking sector using artificial neural network on customer relationship management data. Kraljević, G., & Gotovac, S. (2010) stressed about the use of advanced data mining techniques which can help in identifying the features that are important. It emphasized on the use of data mining technique to identify input variables which can predict churn easily. Here it was also identified that involuntary churn can further be classified into the deliberate churn and accidental churn. This paper carried out a comparison between a neural network, logistic regression, and decision tree. The decision tree model is considered to be the best model according to this research where the confusion matrix indicated an accuracy of 90.9% and the error rate as low as 9.1%. Kulkarni et al., 2019 they used random forest for feature selection. After feature selection, they tried different classification techniques like SVC, Decision Tree, Naïve Bayes, and logistic regression. It was observed that Logistic Regression provided with the best accuracy of 80.38%. They analyzed the past customer records of users who have churned and tried to identify patterns through which they predict the customer churn in advance for the existing users. Moreover, it hinted at the use of a recommendation system to provide offers that are most suitable for each customer. These offer recommendations will in turn improve customer experience and keep them loyal towards the company. Umman and Simsek,2010 studied the Turkey telecom sector, to analyse the demographics of the people who are switching to other operators. They made use of quantitative analysis to predict churn. Subscriber usage was one of the key parameters involved in research. Logistic regression was applied to better understand the data. Spanoudes and Nguyen, 2017 in their paper they have majorly discussed about the deep learning model which can be applied to any company belonging to any sector which maintains a user log. This paper proposed the use of recurrent neural network for better results. So, the existing study makes use of feature engineering to identify the features suitable for training the model. Most of the studies aim at carrying out a comparison between various machine learning algorithms. XGBoost, random forest, and decision tree are the ones that are most talked about.

Research Methodology

The Most of the companies today are making use of quantitative features to determine churn prediction. These companies carry out call detail record (CDR) analysis to understand the usage pattern and identify the users that might churn in the future. Are quantitative features sufficient to understand the consumer? Nowadays buying behavior of the consumer is becoming more and more complex, to understand the consumers we need to study the qualitative features. These qualitative features are derived from user psychographics. The user psychographics comprises of "WHY", why would the consumer buy a particular product or service. To understand user psychographics companies are making use of customer support data, call records of customer care executive, feedbacks, complaints data, and most importantly the queries posted by users on social media platforms like Twitter, Instagram, and Facebook.

On average, every second nearly 6000 tweets are tweeted on twitter. Ample number of these tweets are complaints posted for different companies. These social media complaints form an enormous amount of data, which the companies can assess to understand the consumer behavior and consumer perception for the brand. In this paper, tweets are extracted from twitter using web scrapping. These tweets are analyzed by using the hashtags associated with them. On hashtag analysis, few of the KPI's are identified which formed the foundation for primary research. These KPI's are mapped to the user psychographics and the consumers are asked to rate and prioritize all the features. Post data collection, the collected data is analyzed to draw key insights. After data analysis, an ML model is built using K – Means clustering which is a form of unsupervised learning.

The approach taken for research is completely primary in nature. A questionnaire of 11 questions is circulated, which evaluates the 5 key parameters. The sample size of the research is 473 with 169 responses in favor of company A, 123 in favor of company B,163 in favor of company C and 18 responses in favor of other companies

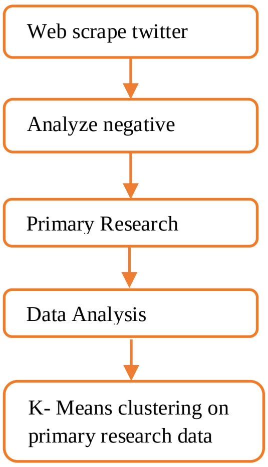

Fig 1: Research Methodology

Web Scraping

Web scraping is a process of retrieving data by using different tools or coding. This helps in extracting data accurately and quickly from various websites.

There are various tools available for web scraping namely-

Octoparse: easy to use data extraction software

Limitations –

- Few features are limited to the premium version
- A user can only extract data 10 times using the free version

ScrapeStorm: helps in extracting data without any prior coding knowledge

Limitations –

- Can't fetch twitter and quora data, it is shown as personal
- Has some privacy issues

Twitter Tags: it makes use of hashtags to fetch tweets

Limitations –

- Can fetch data only for the past 7 days
- Time-consuming as limited data is fetched each hour

GetOldTweets: it is code to extract twitter data by Jefferson-Henrique

Limitations –

- Code can't handle data in regional language

TwitterScraper: python code to web scrape data by Taspinar

Limitations –

- Returns duplicate value
- Returns tweets from around the globe

Hashtag Analysis

This research paper makes use of TwitterScraper to extract tweets for three different companies namely A, B, and C for three consecutive months January, February, and March'20 by using the most relevant company hashtags. These hashtags are analyzed by plotting a frequency v/s hashtag graph. If not graph, word cloud can be used to identify the most relevant hashtags being talked about. Word cloud makes use of frequency and it displays the hashtag with most frequency in the biggest font. Word cloud makes visualization and analysis easy.

On researching some of the hashtags were identified -

- Network - #Network, #horribleNetwork
- Offers - #PrepaidPlans, #billing
- Customer Service -#customercare, #CustomerExperience, #worstService
- Internet - #slowinternet, #internetspeed
- Query Resolution - #wrongcharges, #wrongbilling

So, the Key performance indicators (KPI's) that are identified for the telecom sector using the twitter data are network, internet, customer service, offers, and query resolution

These Key performance indicators are mapped to qualitative features. The network is mapped to network quality, the internet is mapped to internet speed, offers are mapped to offer attractiveness, customer service is mapped to customer service experience and last but not the least, query resolution is mapped to ease of query resolution via social media platforms like Twitter, Instagram or Facebook

Dataset

A survey form which consisted of 11 questions was floated and the consumers were asked to specify their gender, place, phone number (which wasn't mandatory), telecom operator namely A, B, C or Other, ratings for the 5 KPI's (5 being highly satisfied and 1 being highly dissatisfied), prioritize the 5 factors (5 means of utmost priority and 1 means least priority) and consumers viewpoint on their respective telecom service provider. A total of 473 responses were collected of which 455 belonged to service providers A, B, and C and 18 belonged to the category of "others". For company A the total sample size is 169 for company B it is 123 and for company C it is 163 Going ahead the data analysis is based on these 455 data samples. This paper provides an approach with which companies can understand their consumers better and predict churn

Parametric Analysis

To understand the consumer perception % dissatisfaction is calculated for all the five KPI's namely network quality, internet speed, offer attractiveness, customer service experience, and ease of query resolution via social media platforms.

$$\%\ \text{Dissatisfaction}= \frac{\text{Number of Ratings less than 3}}{\text{Total number of ratings}} * 100 \qquad \text{eq. (2)}$$

Company Name	***KPI 1***	***KPI 2***	***KPI 3***	***KPI 4***	***KPI 5***
A	6.51	11.24	10.65	17.75	12.43
B	17.89	30.08	22.76	27.64	24.39
C	1.83	7.32	16.56	20.86	6.75

Table 1: Company Wise % Distribution

In the above table KPI 1 stands for network quality dissatisfaction %, KPI 2 stands for Internet speed dissatisfaction %, KPI 3 stands for Customer service experience dissatisfaction %, KPI 4 stands for ease of query resolution via social media platform dissatisfaction % and KPI 5 stands for Offer attractiveness dissatisfaction %

From the above table, we can conclude that for company A people are most dissatisfied with KPI 4 and KPI 5. For company B consumers are most dissatisfied with KPI 2 and KPI 5 and company C users are most disappointed with KPI 4 and KPI 3. Moreover, from the above table, it is clear that company B is lagging behind its competitors A and C for all the 5 KPI's. This research paper will help in identifying the key improvement areas for company B

K-Means Clustering

Clustering dates back to the 1960's it emerged through numerical taxonomy when a French botanist used numerical similarity and assigned numerical weights to various characteristics used for classifying plants. This "Principle of Taxonomy" was later popularized by Sneath and Sokal. Using this study, they classified plants with similarity under one “taxonomical unit”. Thus, towards the end of the study, they had a variety of “taxas” and each taxas had numerous plants. This is where the idea of clustering came into existence, which says clubbing the similar kinds under one roof.

Clustering is a process by which different data points are segmented based on their similarity and dissimilarity. All the similar data points form a part of the same group, these various groups of collected data points are called clusters. For better clustering results the intracluster similarity should be high, which means the data points belonging to the same cluster are highly associated with each other. Moreover, the inter-cluster dissimilarity must also be high, this means that data points that are a part of group 1 are very different from the data points which are a part of group 2.

As K – Means clustering is an unsupervised form of clustering it doesn’t classify the customers that will churn and customers that won’t churn. Instead, it will form two different groups. Studying the similarities in one group will hint that this group consists of churners whereas the other group consists of non-churner.

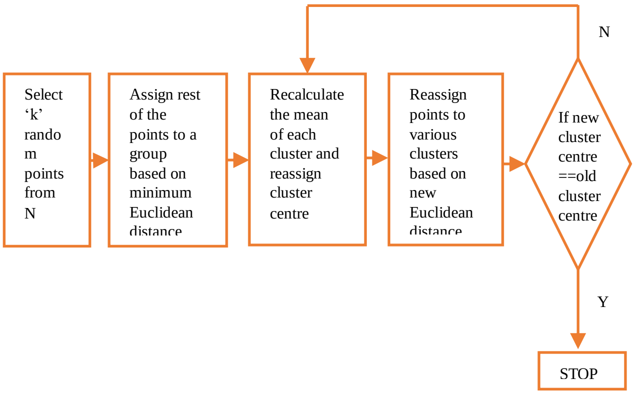
N
Select ‘k’ rando m points from N
Assign rest of the points to a group based on minimum Euclidean distance
Recalculate the mean of each cluster and reassign cluster centre
Reassign points to various clusters based on new Euclidean distance
If new cluster centre ==old cluster centre
Y
STOP

Fig 2: K-Means Clustering Algorithm

Steps to perform K-Means Clustering

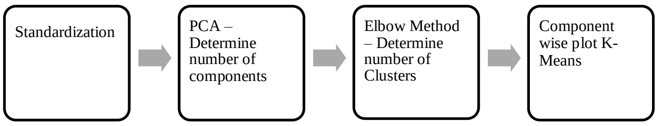
Standardization
PCA – Determine number of components
Elbow Method – Determine number of Clusters
Component wise plot K-Means

Fig 3: Steps to perform K-Means Clustering

Standardization – Standardization helps to homogenize the data. It improves the clustering results as it increases the efficiency of the algorithm. Standardization is important during K - Means clustering because Euclidian Distance is very sensitive to changes in the distance

$$\text{Standardization}=\frac{\text{x-mean}}{\text{standard deviation}} \qquad \text{eq. (3)}$$

$$\text{Euclidian Distance}=\sqrt{(x-a)^2+(y-b)^2} \qquad \text{eq. (4)}$$

Principal Component Analysis - It is always a good habit to perform Principal Component Analysis (PCA) before K – Means clustering. Principal Component Analysis is used because it helps in dimensionality reduction and noise reduction which further improves the result of K-Means clustering. It converts the discrete data points from K -means clustering into continuous solution points. While performing Principal Component Analysis one must ensure to preserve 80% variance to avoid information loss from the data. The dimensionality was reduced from 5 to 3 in this research

Elbow Method – while performing K-Means clustering the number of clusters that are optimal for the research is uncertain. So, to determine the best-suited number of clusters or the best suited “K” which will help in easy segmentation of the data, a method called elbow method is used. Here a graph is plotted against WCSS that is Within cluster squared sum versus number of clusters, an elbow-shaped graph is formed. Where ever the elbow of the graph lies check the number of clusters corresponding to that. This value of the number of clusters is the optimum value of “k” also checks the change in WCSS corresponding to “k+1” as this may represent the best clustering suited for the research data. The ideal value of “k” which is identified in this research is 3. Three clusters are formed at the end of K – Means clustering and data points of each of the clusters are backtracked to find out the similarity in each cluster

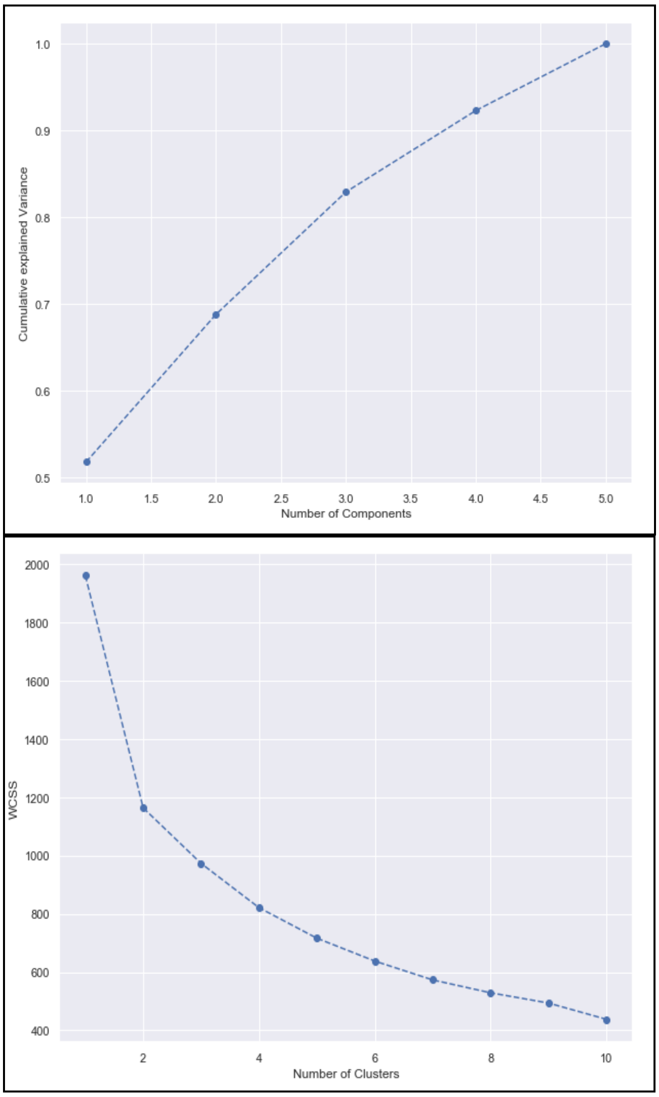

Fig 4: Graph depicting PCA and Elbow Method results respectively

Data Analysis

A quality score is used as a metric to understand each cluster distinctly. The quality score used here is a summation of priority of the key performance indicator multiplied by ratings associated with each of the key performance indicators. So, what will this quality score represent? These quality scores are a summation of priority multiplied by ratings which means that if a KPI is set as priority 5 (highest priority) and is rated 5 (highest rating) then the quality score will be a summation of five plus four plus three plus two plus one that is fifteen and if every KPI is rated 5 then the highest quality score will be fifteen multiplied by five that is seventy-five. If the lowest rating that is one is provided then the quality score will amount to fifteen multiplied by one that is fifteen. So, the range of quality score lies between 15 and 75.

Quality Scrore= Σ (priotity*ratings)

eq. (5)

It was identified that for every cluster each data point belonged to a different set of quality score ranges. Cluster 1 has a quality score below 45, for cluster 2 the quality score belonged to the range of 44 to 59 and for cluster 3 the quality score belonged from 56 to 75. What can be concluded from this? Clearly the cluster 1 with a low range of quality score means customers in this cluster are highly dissatisfied with the services of the telecom operator and these are the ones who might probably churn in the future. Cluster 2 represents a relatively neutral cluster, customers in this cluster are not highly dissatisfied or highly satisfied. The service seems to be okay for them. On the other hand, cluster 3 represents the best clusters, every customer in this cluster seems to be highly delighted with the services that the telecom operator is providing. These are the ones who form the part of the loyal customer base of the telecom operator. Drawing an analogy to better understand the best cluster and worst cluster. Let’s compare this quality score with the net promoter score. The net promoter score helps the company to understand the customers’ perception. It helps in identifying customers who are willing to confide by the company and recommend the company products and services to their peers. Net promoter score has three categories namely promoters, passives, and detractors. On a 1 to10 scale users with ratings 6 or below belong to the category of detractors, users with a rating of 7 and 8 belong to passives, and consumers with a rating of 9 and 10 belong to promoters. Analogous to this cluster 3 consists of users who are promoters, cluster 2 consists of consumers who are passives, and cluster 1 consists of users who are detractors. To prevent churn our main aim is to convert the detractors into passives so that they do not stop using the products and services of the company. Preventing customers from churning plays a vital role in retaining the revenue of the company. Moreover, if the customers churn, the customer lifetime value decreases due to loss in the customer base.

	Cluster 1	Cluster 2	Cluster 3	Total
B	41.46%	39.84%	18.70%	123
A	14.79%	36.69%	48.52%	169
C	11.66%	37.42	50.92%	163
Min	16	35	54	
Max	51	64	75	
Mode	36 and 43	51	59	
Mean	37.4	51.5	63.4	
Range	Below 45	44 to 59	56 to 75	

Table 2: Clustering results for all three clusters

On analysis, it was observed that for company A nearly 15% of consumers belonged to cluster 1 which is the worst cluster. For company C this number is nearly 12%. For company B customers belonging to cluster 1 which is the worst cluster is almost 3 times that of its competitor, it mounted to a huge amount of 41%. Similarly, when we talk about cluster 3

which is the best cluster it is observed that for company A nearly 49% of the consumers belong here, for company C nearly 51% of the consumers belong to this cluster but for company B only 18% of the customers belong to this cluster. This analysis serves as an eye-opener for company B, the company needs to proactively work to improve on all the 5 KPI's being discussed. To prevent customer churn company B must try to satisfy the customers which belong to cluster 1 (worst cluster).

In this paper, we will try to dig further to identify the core problem associated with company B. Hyper segmentation is the key to identify what is the underlying issue. Hyper segmentation means making further smaller segments of the available segments to come up with better outcomes. Here the three clusters already exist, these serve as segments in this research paper. Now diving deeper to understand the problem associated with company B this paper makes use of consumer demographics. Why just demographics? This paper is majorly dealing with the telecom industry. The Telecom industry is bound to have a number of circles and each circle has different issues; some might face network problems while others might be having slow internet. So, it becomes extremely necessary to identify problems associated with each circle. For company B four major regions are identified on the basis of the maximum response received during the survey. These regions are namely – Maharashtra, Gujarat, Madhya Pradesh, and Delhi. To better understand the issue of customer churn, cluster analysis is performed on all the four regions to identify the customer distribution in each cluster in each region

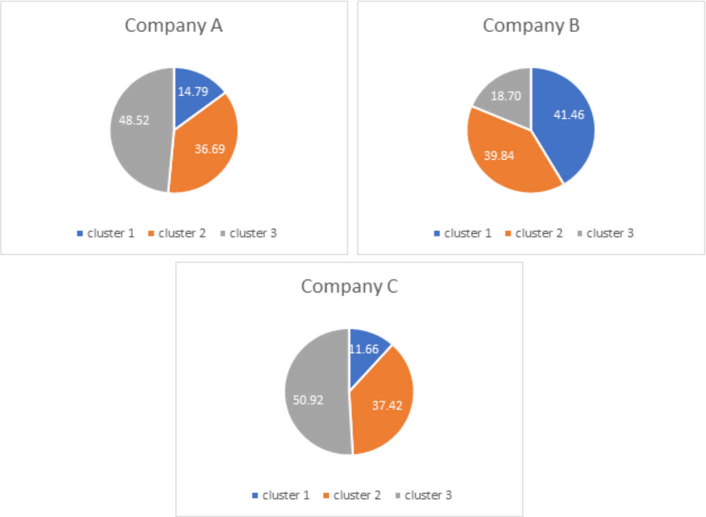

Fig 5: Company wise cluster distribution percentage

Location	Cluster 1	Cluster 2	Cluster 3
Maharashtra	55%	33%	12%
Gujarat	41%	41%	18%
Delhi	24%	47%	29%
Madhya Pradesh	40%	20%	40%

Table 3: Clustering results for all three clusters

The results of cluster analysis show that for Maharashtra region nearly 55% people belong to the worst cluster that is cluster 1, for Gujarat nearly 41% consumers belong to this cluster followed by Madhya Pradesh with 40% consumers lying in cluster 1 and last but not the least is Delhi with 24% consumers in cluster 1. What can be concluded from this? This clearly states that in the Maharashtra region people are highly dissatisfied with company B and its services but in the Delhi region that is not the case. People seem to be highly satisfied with the services of company B because only 24% of people lie in cluster 1. So, the company must identify areas like Maharashtra where there is dire need to improve upon the service provided to the customers. Along with this, the company must not let down the existing services being provided in the Delhi region. Therefore, continuous upgradation and improvement in service is the key to customer retention. Let's try to further understand the importance of each KPI in each of the four regions. To figure this out, first let's calculate the percentage dissatisfaction for all the 5 KPI's namely network quality, internet speed, customer service experience, ease of query resolution via online platforms, and offer attractiveness in all the four regions namely Maharashtra, Gujarat, Madhya Pradesh, and Delhi.

$$\% \text{ Dissatisfaction} = \frac{\text{Number of Ratings less than 3}}{\text{Total number of ratings in that region}} * 100 \qquad \text{eq. (6)}$$

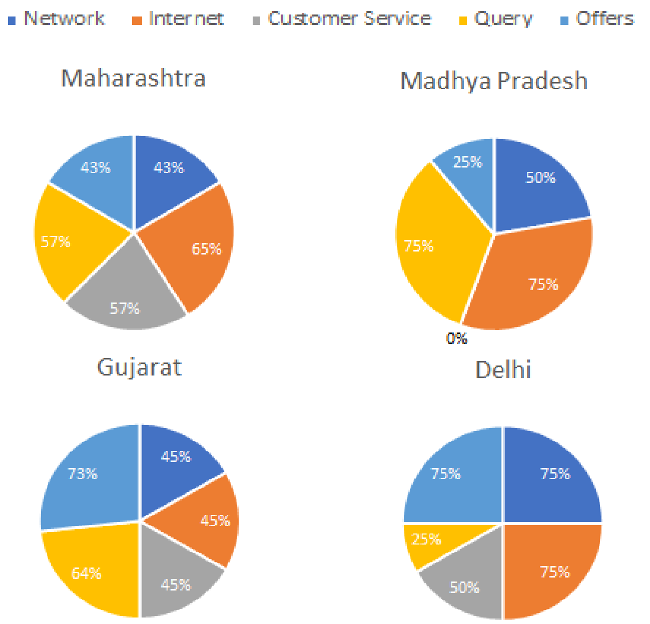

Fig 6: Region wise distribution of KPI dissatisfaction %

From the above graph, one can clearly see that for the Maharashtra region 65% dissatisfaction belongs to the internet, followed by 57% dissatisfaction for customer service experience and 57% for ease of query resolution. Thus, it can be concluded that for the Maharashtra region top three areas of improvement are Internet speed, Customer Service experience, and Ease of query resolution via social media. For the Gujarat region, 73% of people are dissatisfied with Offers being given to them followed by 64% of the consumers are dissatisfied with the ease of query resolution via social media platforms. For the region of Madhya Pradesh, none of the consumers has rated less than 3 on customer service experience, which means that consumers are pretty satisfied with the services or their perception of low service is still not attained by the company, which means that the company is doing decent in terms of customer service experience. The key working areas for this region are internet speed and ease of query resolution via social media platforms. Both these KPI's represent 75% dissatisfaction. For the Delhi region, the company needs to work majorly on offer attractiveness, network quality, and customer service experience. Summing up the above analysis, it is quite clear that internet speed and ease of query resolution via social media platforms are the major concern because 3 out of 4 regions have issues with the above KPI's.

Discussion

The objective of the paper is to understand what all qualitative factors may lead to customer churn particularly in the telecom sector and is the method of K – means clustering a good fit for it. When we talk about qualitative features it becomes extremely important to focus on user psychographics as discussed before. But are these 5 KPI's enough to understand user behavior? The answer to this is No, of course not. Each KPI can further have multiple factors through which one can monitor user behavior. This paper will try to provide a basic understanding of what other factors can influence user behavior. Let's take an example Mr. Sharma who falls under cluster 1 that is the worst cluster has rated customer service experience below 3. By making use of customer journey one can identify the touchpoints which might have hampered his experience. Let say he walked into the store of company B to buy a new SIM, but the in-store staff did not treat him nicely which affected his perception regarding company B. Another instance can be, let say he called a customer care number to seek help, but his query was not resolved in the very first call. He had to repeatedly keep calling the customer care number and explain his issue to a different customer care executive each time. This can lead to Mr. Sharma losing his temper. So, when customer service experience is an issue then there are multiple factors that the company must improve upon. Few of the factors for customer service experience can be first call resolution (means that the query put up by the customer is resolved in the very first call by the customer care representative), service quality and efficiency of response provided by customer care representative, does the customer receive a call from the company whose services he/she is currently using, number of calls and call duration of each call between the customer and competitor company (this can be a sign to identify customer churn if a customer is engaging more with the competitor it can mean that the customer is planning to switch to the services provided by the competitor), nobody likes to tell their complains again and again so it is better if the company maintains a complain record corresponding to each customer. How can a company improve upon this? The company can roll out regular surveys to get consumer response. After each call the customer can be asked to rate their interaction experience with the company executive, the calls which are rated low can be analyzed to identify the areas of improvement that can be done at the companies end. Now let's discuss factors associated with each KPI. If a consumer is dissatisfied with offers what can be the probable reasons? Few such reasons can be: the price of the offer, validity of the offer, data provided in that price, number of calling minutes associated with the offer, offers provided by the competitor and price associated with it, how much is the international roaming charge and many more. Now moving forward to other factors associated with the network. Network strength majorly depends on the location. If there is a company tower nearby, with ample bandwidth to cater to the existing consumer demand; then the network strength is supposed to be good. Moreover, when we talk about network there are other issues like connectivity, call drops, coverage. To solve such problems companies have launched their chatbots, here the consumer can select the option of network issues and provide their pin code. By using this pin code company can identify the areas where the network strength needs to be improved. Apart from this few of the companies show the number of towers located near the consumers' location, all this builds a sense of trust in the mind of the consumer and helps in creating better brand loyalty. Similarly, ease of query resolution via online platforms and internet speed has multiple sub-factors that can help to understand consumer perception better. Let's say Mrs. Mehra is a twitter user and she is having some issue with her old SIM card; the SIM card is not being detected by her phone. She decides to post a tweet tagging the company. Mrs. Mehra is very worried as all her OTPs have stopped coming and she is unable to make any calls. But to her surprise, she doesn't get any response to her tweet within 8 hours, whereas her friend Mrs. Gupta had told her that she posted a tweet tagging the competitor (Mr. Gupta used a different service provider than Mrs. Mehra) and received an instant response. This makes Mrs. Mehra apprehensive. Mrs. Mehra is upset, she thinks that the company she is associated with doesn't value its customer. As a result, Mrs. Mehra plans to change her service provider. Factors that impact the ease of query resolution via social media platforms are response rate on Twitter, Facebook or Instagram, rate of query resolution, quality of response being delivered to the consumer, Average handling time (the lesser the better), the gap between the time of query posted and the first response received. Discussing the factors which affect internet speed. By 2020 nearly 34.8% of the Indian population is using the internet therefore slow internet speed is one of the key concerns. If the company is offering low data at a similar price as compared to its competitor the consumer might switch to the other operator, internet speed depends a lot on location, device specifications (that means whether the device supports 4G or not), whether the SIM is in the primary slot or secondary slot and data consumption of the customer. All these factors contribute to the various touchpoints that impact consumer behavior.

The strategy that any company must follow to determine customer churn must involve the following steps –

1. Identify various factors – these factors can be a combination of both quantitative factors (like the number of calls, number of calls drops, duration of calls, amount of data used, etc.) and qualitative factors (like network quality, service offered, customer experience and many more). Majorly one must keep a track of user activity, engagement of the user with the company, and with the competitor.
2. Check consumer history - to identify certain patterns that can be of help to predict consumer churn. These can be used

to identify an anomaly in consumer Behavior and serve as an early sign of churn. Recharge patterns and outgoing activity change can be monitored. If the outgoing activity decreases drastically, this can indicate that the user is planning to switch to the competitor. If a user recharges the SIM card every month and he/she has not been doing so for the past couple of months, then it's a sign of worry for the company.

3. Identify different target groups – in this research paper K -Means clustering has helped in developing different target groups. Here, there are 3 clusters and each cluster have a different set of users. Therefore, each cluster must be targeted differently
4. Design Campaigns to target the customers – each group should be targeted with a different strategy. Cluster 1 which is the worst cluster should be offered much higher incentives than cluster 2 or cluster 3.
5. Feedback – analyze the consumer response and improvise the existing incentives to serve customers better.

What must the company do post prediction churn? The company must target the riskiest customers. Use tailored offers to address customers' needs and retain customers. Above all, every company must try to gain customer loyalty. Customer loyalty is the key to retain customers and earn revenue. A company gets various other benefits if they focus more on customer retention than acquisition-

1. Word of mouth publicity which leads to positive brand image in the society
2. As talked earlier retention is cheaper than acquisition
3. Increased Customer Lifetime Value
4. Increase in the number of loyal customers which leads to increased revenue for the company
5. Consumers have increased faith in the company and are willing to experiment with the newer products offered by the company

Tools Used –

Jupyter Notebook – All the codes used to analyze data for the above research are coded in python language. Tweets are extracted from twitter using twitter scrapper. K means clustering is visualized using python.

Excel – Microsoft Excel is used to identify % dissatisfaction for different variables. It is used to represent data in the form of a pivot table, which helps to better understand the data.

Conclusion And Recommendations

Customer churn is one of the major issues in the telecom industry. Here, primary research has been carried out on company A, B and C to understand which amongst the 5 qualitative features significantly impact consumer behavior. To identify the qualitative features, Twitter hashtag analysis has been carried out for all the three companies for the month of January, February, and March. After this data pre-processing has been done, followed by Principal component analysis and K – Means clustering. The 5 KPI's being talked here are network quality, internet speed, customer service experience, ease of query resolution via social media platforms like Facebook, Instagram, and Twitter and Offer attractiveness

Sample analysis is done for company B, which is performing worse than its competitors for all the 5 KPI's. Through K-Means clustering three clusters are fetched, these clusters are analogous to the net promoter score. To identify these different clusters, a quality score is used, which is a summation of rating multiplied by priority. It is identified that cluster 1 is the worst cluster, thus it is highly likely that consumers belonging to this cluster might churn in the near future. Further, hyper-segmentation is done to better understand the consumers of cluster 1 of company B on the basis of demographics. 4 major regions are identified and, on each region, further research is carried out to understand % dissatisfaction for each KPI. The top three areas of improvement are identified for each region. Apart from the machine learning model, it also talks about various factors that are responsible for customer churn.

To improve upon the network, the company can show consumers the network strength in their respective areas and give consumers a fair idea of what will the network strength be in the near future. For internet speed and customer care service, the company can carry out live troubleshooting and have selected checkpoints that can help the consumer know exactly which checkpoint did they fail. For query resolution via online platforms, certain rules and regulations must be laid down. Example – The first response should be within 3 hrs; employees should be taught net etiquettes (a direct message shouldn't be replied with a tweet mention). Moreover, any company must start investing more in social listening to understand the consumer perception existing in the market. All these recommendations will provide customers with a seamless journey experience and develop a sense of trust in the company.

Limitations

K-Means clustering is one of the basic unsupervised learning techniques. To get better results an ensemble model can be used, where K - Means clustering can be used in combination with a random forest or decision tree. In this paper, only the hashtags from the tweets are analyzed, rather than that one can focus on the text involved in tweets and perform text analysis along with sentiment analysis to identify what positive and negative trends are being talked about the company on social media. While studying behavioral characteristics it becomes essentially important to add semantic information, give importance to words, and understand word embeddings involved in the sentence. This helps in identifying the tone of the text being communicated and can also help to reduce sarcasm in the text. One must use a large dataset to get better results using this research technique.

References

[1] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 28.

[2] Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications policy, 30(10-11), 552-568.

[3] Kraljević, G., & Gotovac, S. (2010). Modeling data mining applications for prediction of prepaid churn in telecommunication services. Automatika, 51(3), 275-283.

[4] Brânduşoiu, I., Toderean, G., & Beleiu, H. (2016, June). Methods for churn prediction in the pre-paid mobile telecommunications industry. In 2016 International conference on communications (COMM) (pp. 97-100). IEEE.

[5] Dass, R., & Jain, R. (2011). An Analysis on the factors causing telecom churn: First Findings. In AMCIS.

[6] Shah, J. D., Shah, F. D., & Rahevar, M. Customer Churn Prediction Analysis. International Journal of Computer Applications, 975, 8887.

[7] Yıldız, M., & Albayrak, S. (2017). Customer churn prediction in telecommunication with rotation forest method. DBKDA 2017, 35.

[8] Seyed Hossein Iranmanesh, M. H. (2019). Customer Churn Prediction Using Artificial Neural Network: An Analytical CRM Application. International Conference on Industrial Engineering and Operations Management, 23-26.

[9] Kulkarni, A., Patil, A., Patil, M., & Bhoite, S. (2019). Customer Churn Analysis and Prediction. International Journal of Computer Application Technology and Research, 363-366.