IBM - 306: Marketing Research

Customer Churn Management

Under the guidance of :

Prof. Jogendra Kumar Nayak

Assistant Professor

Indian Institute of Technology Roorkee

jogendra.nayak@ms.iitr.ac.in

Presented by:

Divyam Goel (18111009)[1]

Kunal Mohan (18116042)[1]

Akshat Jain (18116006)

Ashutosh Bharambe(18116019)[1]

Pranav Singhal (18116062)

[1] Equal contribution

CONTENT

S.No.	Topic	Page No.
1.	Acknowledgement	3
2.	Abstract	4
3.	Introduction	5
4.	Problem Statement	7
5.	Research Methodology	8
	5.1 Data Collection	8
	5.2 Questionnaire Design	9
	5.3 Measurement Scales	13
	5.4 Sampling	14
	5.5 Modeling	14
6.	Data Analysis and Results	17
7.	Conclusion	23
8.	Limitations of Study	24
9.	References	25

ACKNOWLEDGEMENT

We would like to thank our mentor and guide Prof. J.K. Nayak to give us the opportunity to work on this excellent project. Also, we would like to show our gratitude to him for his constant support and guide which led to us completing this project on time with such efficiency.

We would like to thank our friends and colleagues as well who took out their time to help provide the necessary data needed for its success. Also, kudos to the public who actively participated in our survey and made it possible to create a good database.

Thank You.

ABSTRACT

The Indian telecommunications industry's rapid growth has led to a more significant subscriber base for service providers. With new competitors trying to find their way to tap into the Indian market, new and innovative business models and a promise of providing better services for lower costs to the consumer are increasing customer acquisition costs for the service providers.

With such high levels of entropy defining the Indian telecom sector, it becomes important to analyze the churn patterns of the Indian consumer. In this paper, we utilize data mining techniques for the identification of churn. Based on relevant data, these methods try to find patterns that can point out possible churners.

INTRODUCTION

Customer relationship management (CRM) is a strategic approach that targets the development of profitable, long-term relationships with key customers and stakeholders [1]. As the Indian telecommunications industry hits a saturation point, more and more companies have realized the importance of CRM. This becomes evident as one begins to analyze the telecom giants' advertising strategies in the Indian market, which have moved away from product-centric mass marketing to consumer-centric targeted marketing.

One of the major concerns of building reliable CRM systems is Customer Churn Management [2]. In mobile telecommunications, the term "churn" refers to the loss of subscribers who switch from one provider to another during a given period. Even as the industry hits a saturation point in terms of new customers, the service providers find themselves in a rat race to establish a monopoly over this multi-billion dollar industry. As it is more profitable to retain existing customers than to attract new customers continually, it is crucial to build an accurate churn prediction model to identify customers who are most prone to churn [3]-[5].

Historical work on Customer Churn Prediction and management in various international markets has found that churners only account for a fraction of the customer base. However, if the meteoric rise of Jio has taught us anything, it is that the Indian telecom industry is still extremely volatile. In that respect, we thought it was essential to carry out a specialized study focusing on the Indian telecom

industry to better understand the factors that influence churn. As expected, the data collected in our study reveals that the current churn dynamics of the Indian telecom industry are the polar opposite of those which are generally expected. We find that most Indian customers are inclined towards trying a different network provider. A detailed analysis of the data is presented in the following sections.

Established literature on customer churn relies on data mining technologies, such as Neural Networks [6], Clustering [7], Decision Trees [6], [8], Regression[9], [10], Support Vector Machines [11], and an ensemble of hybrid methods [12], to provide accurate predictions. While getting accurate predictions is the main objective, it is also necessary that the churn models bring with them a certain level of interpretability. This is important with regards to understanding the key drivers that influence customer churn and customer retention. Previous research works indicate that Regression based techniques are generally more interpretable than other data mining techniques [9], [10]. However, with the rapid advancements towards addressing interpretability concerns across the board, it is also essential to re-analyze this norm. Other research explores the power of new features for churn prediction, such as social networks and text information of customer complaints, which is beyond this paper's scope.

In this paper, we present a comprehensive analysis of customer churn in the Indian telecom industry and an empirical analysis of the interpretability of various Customer Churn Prediction models.

PROBLEM STATEMENT

In the Indian telecom industry, public policies and standardization of mobile communication allow customers to easily switch over from one service provider to another, resulting in a strained fluidic market. The advent of new market players has further led to a rise in entropy in the customer's mindset towards their network provider.

As these companies battle to establish some control over the Indian market, customer churn prediction, or the task of identifying network subscribers who are likely to discontinue the use of a service, has become an important and lucrative concern of the Telecom industry.

This project aims to study and analyze features that influence customer churn while also presenting an empirical analysis of historical data mining techniques to build both accurate and interpretable churn prediction models.

RESEARCH METHODOLOGY

We used a basic DESCRIPTIVE TYPE research design to find and explore the various factors that influence customer churn in the telecom industry.

We make use of PRIMARY DATA collection strategies to build a dataset relevant to the Indian Telecom Industry, focusing on both quantitative and qualitative features that influence customer churn. Factors such as effectiveness of customer care services, age of customers as well as network plans were identified as being crucial in predicting customer churn and hence customer retention.

Data Collection

Data collection is a systematic way of collecting and evaluating data collected from different sources of information to provide answers to relevant questions and is required to capture quality evidence that seeks to answer all questions asked. Through data collection a business or management can obtain the quality information that is needed to make informed decisions.

There are many ways to collect data but most used are : published literature sources, surveys (email and mail), interviews (telephone, face-to-face or focus group), observations, documents and records, and experiments.

Our data collection falls under primary data collection and explores both the qualitative, i.e. based on the non-quantifiable elements like the feeling or emotion of the researcher, and the quantitative, i.e. methods that are presented in numbers and require a mathematical calculation to deduce, aspects of the research.

We collected our data through a survey and used a questionnaire for the purpose. Questionnaires provide a relatively cheap, quick and efficient way of obtaining large amounts of information from a large sample of people. Data can be collected relatively quickly because the researcher would not need to be present when the questionnaires were completed. This proved particularly useful to collect quality information during the COVID-19 pandemic, without having to leave the safety of our homes.

However, a major problem with questionnaires is that respondents may lie due to social desirability. Most people want to present a positive image of themselves and so may lie or bend the truth to look good. In order to ensure that we did not have to compromise with the quality of information collected by this instrument, we payed special focus in designing the questionnaire, making use of historically proven techniques and methodologies such as nominal data collection via dichotomous and polychotomous questions, relevant normalization techniques etc. The details of this design are presented in the following section.

Questionnaire Design

Our questionnaire was framed in order to extract maximum information from the user in a minimum number of questions to ensure that we maintain the willingness

of the respondent to answer further questions. Our questionnaire is presented below-

1. What age group do you belong to (years)?
    a. Below 15
    b. 15-25
    c. 25-40
    d. 40-55
    e. Above 55

2. What state are you from?

3. Name of your current Cellular Network Provider
    a. Airtel
    b. Reliance Jio
    c. Vodafone-Idea
    d. BSNL
    e. Other? Please tell _____

4. Your current network subscription is
    a. Personal network
    b. Provided by employer

5. Select your plan type
    a. Pre-paid
    b. Post-paid

6. What is your current network plan?
    a. Less that 99
    b. 100-249
    c. 250-399

d. 400-599
e. More than 600

7. Your internet usage using cellular networks (per day)
    a. Less than 1 GB
    b. 1GB-2GB
    c. 2GB-4GB
    d. More than 4GB

8. How much time do you spend on calls everyday?
    a. Less than 0.5 hour
    b. 0.5-2 hours
    c. More than 2 hours

9. Rate your provider's internet services
    a. Very poor
    b. Poor
    c. Satisfactory
    d. Good
    e. Very good

10. Rate your provider's network coverage
    a. Very poor
    b. Poor
    c. Satisfactory
    d. Good
    e. Very good

11. Are you satisfied with your provider's customer care services?
    a. Yes
    b. No

The questionnaire was designed to extract knowledge about the basic parameters that affect a telecom customer's choice regarding changing their cellular network provider. Here we present a few ideas that allowed us to build a questionnaire that encompassed both qualitative and quantitative features that influence customer churn while ensuring that our study remained cohesive:

- **Information we needed-** In order to gain knowledge about the customer churn trends and even predict them, the most important thing to determine is the usage pattern of customers. Questions related to plan type, plan price, data usage and time spent on calls help us gain knowledge about how heavily a user relies on his network provider in their daily life and would thus determine how important is quality of service to them. We also needed the information about how satisfied the respondent is with the network coverage, internet service and customer care services of their provider which directly affect the fact whether they're willing to switch to a different network provider or not. The region where the respondent resides would also help us because some providers offer better services in some particular geographical areas than others, which affects the hold of the provider on it's customers.

- **Our target audience-** Our target audience is any person that uses a mobile phone and has knowledge about the basic terms used while choosing a plan and paying bills for the same. Hence we mainly targeted people above the age of 15 years. The reason for choosing this interval is that it is highly likely that children below the age of 15 may not be aware of the plan type,

price and network usage since most of them don't pay the bills for their cellular network on their own.

- **Framing the questions-** The questions have been framed in a very simple to understand and easy language for a common man to understand. No questions regarding the respondent's personal interests have been asked and any questions that we felt the respondent may not be willing to answer have been kept optional. These steps have been taken to overcome the respondent's unwillingness to answer or respond to the questions.

Measurement Scales

Feedback on network coverage and internet services- We have used ordinal scale with 5 levels- Very Poor, Poor, Satisfactory, Good, Excellent- to get the respondents feedback on their network provider's network coverage and internet services. This helped us to compare and categorize the users accordingly.

Plan price, data usage and time spent on calls- Since it wouldn't be logical to collect each respondent's exact data usage, plan price and time spent, we used interval scaling to collect data for the same. This helped us to classify the samples as light user, moderate user or heavy user of the cellular network services. We used Min-max normalization on the collected data to normalize the interval scale to have uniform significance of data points across all intervals.

Sampling

We adopted the method of convenience sampling for data collection. This was done by circulating our survey form via a number of social media platforms, so that anyone who comes across it and is available and willing to participate in the survey can do so by filling the form.

Modeling

We tried a few different models to find which one of the statistical data mining techniques was best fitted to the problem statement. Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. The features that we used are derived from the answers of our survey and the categories of the features were adopted from a research paper [13],which are as follows:

1) Customer care service details.
2) Customer demography and personal details.
3) Customer credit score.
4) Bill and payment details.
5) Customer usage pattern.
6) Customer value added services.

The questions were framed such that we could interpret the answers to get our desired features and the features that we used are :

1) Age of the user
2) Usage statistics of data and call time of the user
3) The network provider
4) The satisfactory measure of the service provided

5) The demographic area(state) they belong to

Below is a brief description of the models that we implemented and their results can be found in the results section, the best results were achieved using logistic regression, so, the website cited in this report only predicts based on the logistic regression model.

1) **Random Forest Classifier :**

Random forest is an ensemble algorithm that employs random decision trees to make classification boundaries. Random Forest is a classifier that creates decision trees based on a random subset of data and features. Random forests combine the simplicity of decision trees with flexibility. The variety is what makes random forests more effective than individual decision trees. Another important thing is that the random forest classifiers provide us with the feature importance matrix denoting the importance of each feature used.

While such ensemble methods are often preferred for predictive modeling, they have been widely criticized for being less interpretable to their regression based counterparts [16], [17]. By interpretability we mean the ability to identify risk factors which can be addressed by the retention process to prevent a customer from leaving. Hence, while the importance matrix provides valuable information, we can not take this information as a standard and need to conduct more tests to establish concrete statistical evidence.

2) **Logistic Regression Classifier :**

Logistic regression is a simple classifier, that is a special case of the generalised linear models. The hypothesis function of the regression is

replaced by the sigmoid function given by:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

where

$$g(z) = \frac{1}{1 + e^{-z}}$$

The y|x thus follows the Bernoulli distribution and is a special case of the GLMs. We have found in the existing literature that logistic regression has performed well in determining the churn probability. And the same is in our case as well.

Regression based models are historically preferred in churn prediction, as well as other marketing tasks, mainly because of their feature interpretation abilities. The information gathered from feature sets, as well as the final results often account for the essential risk factors overlooked in ensemble learning methods.

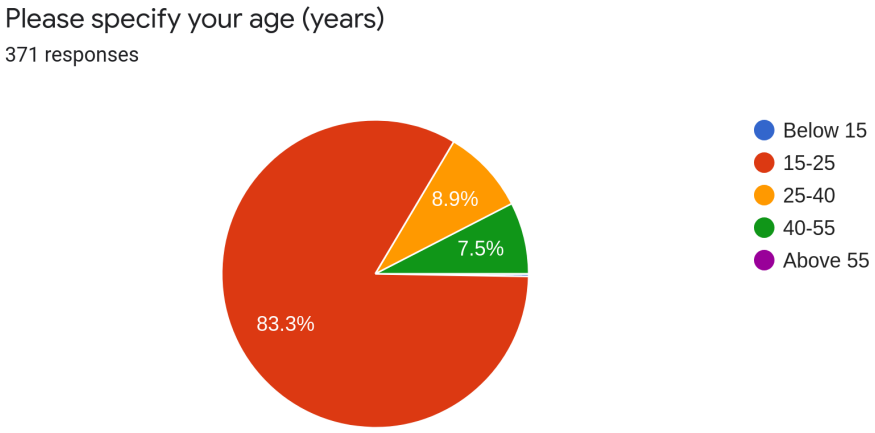
3) Naive Bayes Classifier :

As the name suggests Naive Bayes classifier belongs to the 'probabilistic classifier' family is a basic classifier which is based upon Bayes theorem. It has a basic assumption that all features are pairwise independent of each other. It is highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. It works by maximizing the maximum likelihood of the features.

DATA ANALYSIS AND RESULTS

The data collected by means of the Google form and the website presents a study of nearly 400 responses. The respondents were widely dispersed across the country as the questionnaire was floated online. An exploratory data analysis revealed the following:

- A large portion (83.3%) of the respondents belonged to the 15-25 age group. In fact, all the respondents actually fall within the age of employment, i.e., between 15 and 55 years. The following diagram shows the exact distribution of the data collected by the google form.



- Consumer distribution across various telecom service providers in our dataset is presented below. The data closely resembles the actual customer distribution in the Indian telecom sector and identifies three major market players: Airtel, Reliance Jio and Vodafone Idea. Unsurprisingly, local providers, as well as government providers attract an extremely small

proportion of the populace.

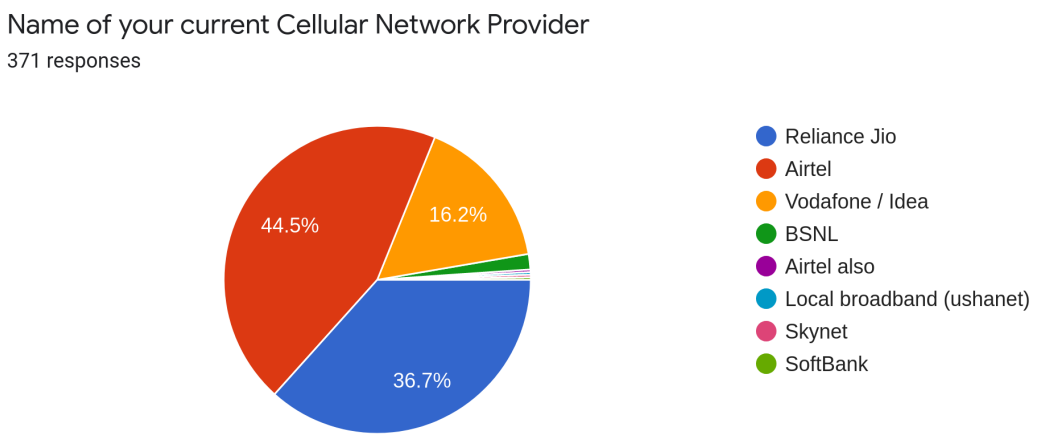


Fig. 1

- Interestingly, the data collected reveals a disparity between the current state of the Indian telecom industry and the expected norm. Historically, churners have accounted for an extremely small fraction of the overall user base of a given network provider. Our study reveals that more than 75% of the current customer base are attracted to the idea of switching to a different network provider. Further, more than 2/3rd of these possible churners pay little heed to the cost variations, citing a lack of overall service quality as the main reason for their churn behaviour.

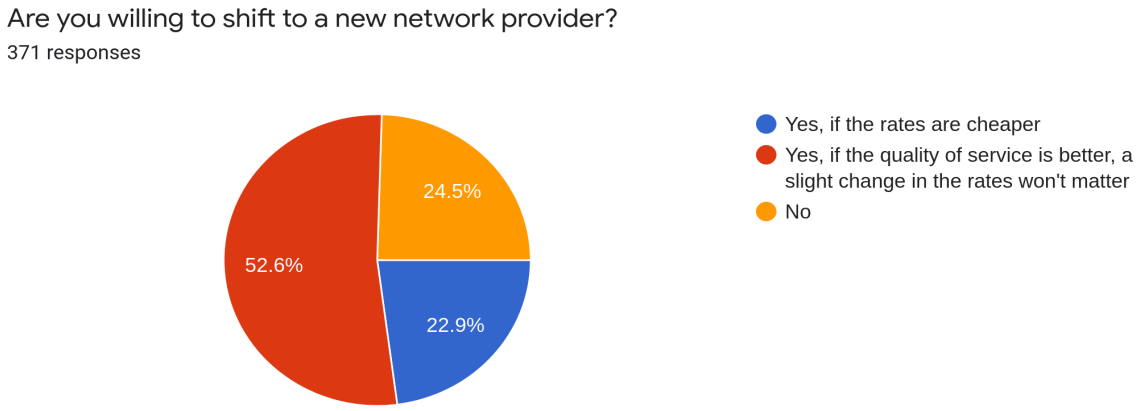


Fig. 2

- The region that the customer belongs to has historically been seen as one of the most important features influencing customer churn. The data collected was initially extremely sparse for most states in India and would have hence led to a poor churn prediction system. In order to deal with this problem, we decided to group these data points based on the region that the customer belonged to rather than the state. As an example, Delhi became part of North India, Maharashtra became part of West India and so on. We decided to limit this bifurcation to 5 categories: North India (NI), South India (SI), West India (WI), East India (EI) and Central India (CI). The final distribution is presented below. As can be seen, while the sample does not present a normalized distribution over the different regions, we have been able to get rid of the data sparsity issues observed in the previous set up.

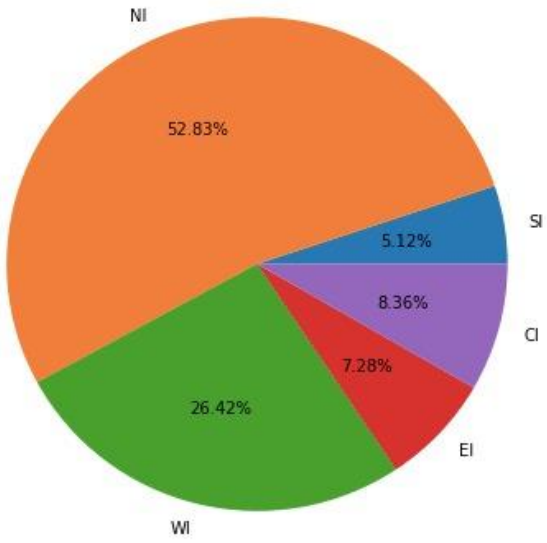


Fig. 3

Pearson Correlation Coefficient

For the features extracted from the data collected, we calculated the Pearson's Correlation Coefficient. This is a statistical measure that measures the linear coefficient between any two variables. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. The conclusions drawn from the result are given below:

Fig. 4

- There is a high degree of positive correlation between the fields corresponding to the name and quality of internet services of a network provider. This combined with the fact that internet services are almost uncorrelated to the regional variations in our data sample, shows that certain network providers have stronger infrastructure capacities to support better cellular services.
- There is a high degree of negative correlation between churners and the age of the churners. This shows that the younger populace is more prone towards switching network subscriptions. The implication of this observation is two-fold:
    - Companies have been successful in coming up with marketing policies to ensure customer retention (older population prefers not to switch).
    - The results indicate that there has been almost too much focus at customer retention rather than aggressive, innovative and disruptive marketing strategies that would help build a new customer base among the younger users.
- Possible churners show a high degree of negative correlation with customer service satisfaction hence highlighting customer service satisfaction as the most important feature that influences churn and customer retention.

Results

Machine learning algorithms are almost always optimized for raw, detailed source data. Thus, the data environment must provision large quantities of raw data for discovery-oriented analytics practices such as data exploration, data mining,

statistics, and machine learning. This however is not always the case. Often, it is difficult to collect a large volume of reliable data samples.

In order to ensure the validity of prediction models trained using machine learning techniques in low resource settings, one can adopt one of the following strategies:

1. Data cleaning and preprocessing
2. Cross Validation

To confirm the validity of our churn prediction models, we made use of a technique known as 10-fold cross validation. The technique comprises of the following:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    a. Take the group as a hold out or test data set
    b. Take the remaining groups as a training data set
    c. Fit a model on the training set and evaluate it on the test set
    d. Retain the evaluation score and discard the model
    e. Summarize the skill of the model using the sample of model evaluation scores.

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times. In our case, k = 10.

The evaluation metric used for our models is the weighted F1-score. It is a measure of a model's accuracy on a dataset. It is used to evaluate classification systems,

which classify examples into ‘positive’ or ‘negative’, especially those with inherent class imbalances such as those presented in Churn Management problems.

The final results of the Customer Churn Prediction Models are presented below:

Model	10-Fold F1-Score
Random Forest Classifier	74%
Logistic Regression	75%
Gaussian Naive Bayes	71%

Fig. 5

Hence, our study reveals that Logistic Regression presents the best approach towards modelling customer churn in the Indian Telecom Industry, predicting customer churn with a weighted F1-score of nearly 75%.

CONCLUSION

From all the data that we collected, we have developed a statistical and mathematical model to predict the probability of a person switching their network provider based on their responses to a set of questions. The model has been hosted on a website that we created for this purpose. The link of the website is given below.

Website Link- https://telecom-churn-pred.herokuapp.com/

LIMITATIONS OF STUDY

This research study, like any other, has several limitations.

1) **Unreliable data:**

   Descriptive methods mainly depend on the responses of people. There are chances that people might not act their true selves if they know they are being observed. In the case of the survey method, there are chances that some people don't answer the questions honestly, which makes the output of the descriptive research study invalid. Because the results derived from this type of data will not be accurate.

2) **Limited data:**

   Since the study targets the general public, we try to reach as many people as possible. But the reach is limited to people of the same social groups and geographical areas as researchers, mostly. In our case, we had only around 400 data points. With such a small database, extrapolation is partial according to this database and does not reflect the opinion of the general public.

3) **Error in sampling:**

   In descriptive research methods, participants are picked randomly. The randomness of the sample can't represent the whole population accurately. In this research, survey forms were circulated via social media platforms. Thus, the reach was our colleagues, family and friends. Hence, its results are partial towards this set of population.

REFERENCES

[1] A. Payne and P. Frow, “A strategic framework for customer relation-ship management,”J. Marketing, vol. 69, no. 4, pp. 167–176, Oct.2005.

[2] A. Berson, K. Thearling, and S. Smith, Building Data Mining Applica-tions for CRM. New York: McGraw-Hill, 1999.

[3] K.CoussementandD.V.Poel,“Churnpredictioninsubscriptionser-vices: An application of support vector machines while comparing two parameter-selection techniques,”Expert Syst. Appl., vol. 34, no. 1, pp.313–327, Jan. 2008.

[4] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, “New Insights into churn prediction in the telecommunication sector: A profit driven data mining approach,”Eur. J. Oper. Res., vol. 218, no. 1, pp.211–229, Apr. 2012.

[5] W. J. Reinartz and V. Kumar, “The impact of customer relationship characteristics on profitable lifetime duration,”J. Marketing, vol. 67,no. 1, pp. 77–99, Jan. 2003.

[6] P. Datta, B. Masand, D. R. Mani, and B. Li, “Automated cellular mod-eling and prediction on a large scale,”Artif. Intell. Rev.,vol.14,no.6,pp. 485–502, Dec. 2000.

[7] D. PopovićandB.D.Bašić, “Churn prediction model in retail banking using fuzzy C-means algorithm,”Informatica, vol. 33, no. 2, pp.235–239, May 2009.

[8] C.-P. Wei and I. T. Chiu, “Turning telecommunications call details to churn prediction: A data mining approach,”Expert Syst. Appl., vol. 23,no. 2, pp. 103–112, Aug. 2002.

[9] M. Owczarczuk, “Churn models for prepaid customers in the cellular telecommunication industry using large data marts,”Expert Syst. Appl.,vol. 37, no. 6, pp. 4710–4712, Jun. 2010.

[10] J. Burez and D. V. Poel, “Handling class imbalance in customer churn prediction,”Expert Syst. Appl., vol. 36, no. 3, pp. 4626–4636, Apr.2009.

[11] N. Kim, K.-H. Jung, Y. S. Kim, and J. Lee, “Uniformly subsampled ensemble (USE) for churn management: Theory and implementation,”Expert Syst. Appl., vol. 39, no. 15, pp. 11839–11845, Nov. 2012.

[12] D. A. Kumar and V. Ravi, “Predicting credit card customer churn inbanks using data mining,”Int. J. Data Anal. Tech. Strategies, vol. 1,no. 1, pp. 4–28, Aug. 2008.

[13] Umayaparvathi, V. and Iyakutti, K., 2016. A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. International Research Journal of Engineering and Technology (IRJET), 3(04).

[14]https://www.businesstoday.in/current/corporate/vodafone-agr-subscribers-trai/story/396938.html

[15] Ng, A., Machine Learning (CS229).

[16] N. Meinshausen, "Node harvest,"Ann. Appl. Stat., vol. 4, no. 4, pp.2049–2072, Dec. 2010.

[17] J. Dutkowski and A. Gambin, "Research on consensus biomarker se-lection,"BMC Bioinformatic., vol. 8, no. 5, pp. 1–13, May 2007.