ijiTECH International Journal of Innovative Technologies

WWW.IJITECH.ORG

ISSN 2321-8665
Volume.07, Issue.01,
January-June, 2019,
Pages:25-28

A Detailed Analysis Customer Churn in Telecommunication Industry: Datasets, Methods and Metrics

D. GEERVANI[1], T.S SANDEEP[2]

[1]PG Scholar, Dept of CSE, SV Engineering College for Women, Tirupati, AP, India.
[2]Assistant Professor, Dept of CSE, SV Engineering College for Women, Tirupati, AP, India.

Abstract: Predicting customer churn in telecommunication industries becomes a most important topic for research in recent years. Because its helps in detecting which customer are likely to change or cancel their subscription to a service. Now a days the mobile telecom market has growing market rapidly and all the telecommunication industries focused on building a large customer base into keeping customers in house. So it is very important to find which customers are wants to switch to a other competitor by cancel their subscription in the near future. Analysis of data which is extracted from telecom companies can helps to find the reasons of customer churn and also uses the information to retain the customers. So predicting churn is very important for telecom companies to retain their customers. The paper reviews the relevant studies on Customer Churn Analysis on Telecommunication Industry in literature to present general information to readers about the frequently used data mining methods used, results and performance of the methods and shedding alight to further studies. To keep the review up to date, studies published in last five years and mainly last two years have been included.

Keywords: Churn Analysis, Telecommunications, Data Mining.

I. INTRODUCTION

Studies revealed that gaining new customers is 5 to 10 times costlier than keeping existing customers happy and loyal in today’s competitive conditions, and that an average company loses 10 to 30 percent of customers annually (Kotler 2009).Many companies, being aware of this fact, are engaged in satisfying and retaining the customers. Especially in the subscription oriented industries, such as telecommunications, banking, insurance, and in the fields of customer relationship management, etc., companies working with numerous customers, the revenues of the companies are provided by the payments made by these customers periodically. It is very important to be able to keep customers satisfied in order to be able to sustain this revenue with the least expenditure cost.

A. Objectives

- Reviewing the relevant studies about churn analysis on telecommunications industry presented in the last five years, particularly in the last two years, and introducing these up-to-date studies in the literature,
- Determining the data mining methods frequently used in churn implementations,
- Shedding a light on methods that can be used in further studies.

B. Data Mining and Customer Churn Analysis

In today’s technological conditions, new data are being produced by different sources in many sectors. However, it is not possible to extract the useful information hidden in these data sets, unless they are processed properly. In order to find out this hidden information, various analyses should be performed using data mining, which consists of numerous methods. The Churn Analysis aims to predict customers who are going to stop using a product or service among the customers. And, the customer churn analysis is a data mining based work that will extract these possibilities. Today’s competitive conditions led to numerous companies selling the same product at quite a similar service and product quality. In the midst of this competition, the cost of gaining new customers is more than retaining the existing customers. For this reason, existing customers are very valuable. With the Churn Analysis, it is possible to precisely predict the customers who are going to stop using services or products by assigning a probability to each customer. This analysis can be performed according to customer segments and amount of loss (monetary equivalent). Following these analyses, communication with the customers can be improved in order to persuade the customers and increase customer loyalty. Effective marketing campaigns for target customers can be created by calculating the churn rate or customer attrition. In this way, profitability can be increased significantly or the possible damage due to customer loss can be reduced at the same rate (Argüden 2008).

For example, if a service provider which has a total of 2 million subscribers, gains 750.000 new subscribers and losts 275.000 customers; churn rate is calculated as 10%. The customer churn rate has a significant effect on the financial market value of the company. So most of the companies keep an eye on the value of the customer at monthly or quarterly periods (Seker 2016). Churn can be called as voluntary and involuntary. Voluntary churn occurs when an existing customer leaves the service provider and joins another service provider; but in involuntary churn, customer is asked by the service provider to leave due to reasons like non-payments etc. (Mahajan2015). Voluntary churn can be sub-divided into: incidental churn and deliberate churn (Gotovac 2010). Incidental churn occurs because of the unplanned changes in the customers' lives like a change in financial conditions, change in living location. Deliberate churn occurs for reasons of technology (customers that want a newer or better technology, price sensitivity, service quality factors, social or psychological factors and convenience reasons) (Mattison 2005).

II. LITERATURE REVIEW

In a study by Gursoy, customers who tend to leave a large company operating in the telecommunication sector in Turkey have been identified to develop special marketing strategies for these customers. Logistic Regression Analysis and Decision Tree classification techniques have been used on a 4-month data set consisting 1000 records with24 variables, and the results have been presented (Gursoy2010). In the churn analysis study by Brandusoiu and Toderean, 4different core functions have been used in the Support Vector Machines model and performances have been compared by using a data set consisting of 3333 customer records with21 variables provided by a telecommunications company. And among these models, the one with the polynomial core function has been reported to have the best result by 88.56% (Brandusoi 2013).

Yildiz has conducted a study to predict the customer churn using data mining classification techniques. In order to reduce the run-time of the classification techniques and to increase the performance, they have reduced the number of features, used different classification techniques and measured their performances. In addition, outlier analysis has been performed to observe the effects on the classification results. These classifications have been tested on 2 different data sets containing 5000 subscribers with 20 variables and51306 subscribers with 172 variables, and Recall Ratio and Precision Ratio have been used as the performance criteria (Yildiz 2015).

Mahajan and Som present a study on analyzing customer behaviors on the customers' pre-paid recharge data, voice and SMS usage data to identify patterns in user behavior for intelligent and targeted promotions and churn prediction over the dataset taken from BSNL telecommunications company in India. The number of records of the dataset is not clear as data about different types are included. But generally 25 variables on customer details, recharge details, outgoing and incoming voice calls and sms sent are used. And a logistic model on predicting customer churn has been offered (Mahajan 2016).

III. PROBLEM DESCRIPTION

In a business setting, the term, client attrition merely refers to the purchasers exploit one business service to a different. Client churn or subscriber churn is additionally kind of like attrition that is that the method of shoppers shifts from one service supplier to a different anonymously. From a machine learning perspective, churn prediction could be a supervised (i.e. labeled) downside outlined as follows: Given a predefined forecast horizon, the goal is to predict the longer term churners over that horizon, given the info related to every subscriber within the network. The churn prediction downside diagrammatic here involves three phases, namely,

- The training part
- Testing part
- Prediction section.

The input for this downside includes the info on past necessitate every mobile subscriber, along with all personal and business data that's maintained by the service supplier. Additionally, for the training section, labels are provided within the type of an inventory of churners. When the model is trained with highest accuracy, the model should be able to predict the list of churners from the important dataset that doesn't embody any churn label. Within the perspective of information discovery method, this downside is categorized as prognostic mining or prognostic modeling.

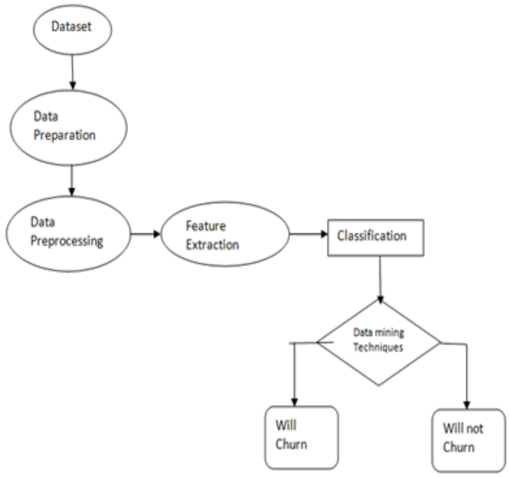


Fig1. Churn Prediction Framework.

This is where the churn prediction model can help the business to identify such high risk customers and there by helps in maintaining the existing customer base and increase in revenues. Churn prediction is also important because of the fact that acquiring new customers is much costly than retaining the existing one. As the telecom users are billions in number even a small fraction of churn leads to high loss of revenue. Retention has become crucial especially in the present situation because of the increasing number of service providers and the competition between them, where everyone is trying to attract new customers and lure them to switch to their service. With a large customer base and the information available about them data mining techniques proves to be a viable option for making predictions about the customers that have high probability to churn based on the historical records available.

IV. PROPOSED SYSTEM

KDD (Knowledge Discovery in Databases) is defined as the non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns of in data". The problem of our discussion deals with the discrete valued target variable and our ultimate aim is to declare each subscriber as „potentially churner" or "potentially non churner", so the KDD function for our problem is defined to be the classification problem.

A. MapReduce

A MapReduce job usually splits the input data-set in to independent chunks which are processed by the map tasks in a completely parallel manner. The frame work sorts the outputs of the maps, and later they are used as input to the reduce tasks. The frame work takes care of scheduling tasks, monitoring them and re-executes the failed tasks. We use divide and conquer algorithm in this particular Hadoop process. Divide and Conquer is an algorithmic paradigm. A typical Divide and Conquer algorithm solves a problem using following three steps.

Divide: Break the given problem into sub problems of same type.
Conquer: Recursively solve these sub problems
Combine: Appropriately combine the answers

A classic example of Divide and Conquer is Merge Sort demonstrated below. In Merge Sort, we divide array into two halves, sort the two halves recursively, and then merge the sorted halves.

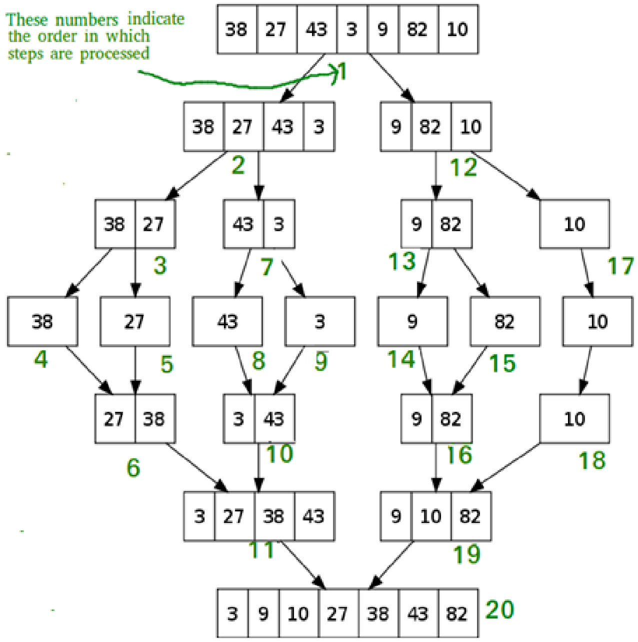


Fig 2. MAP Reduce.

B. Data Preprocessing

Data preprocessing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand. The data gathered from multiple sources first is aggregated and then cleaned as the complete data collected is not suitable for modeling purposes. The records with unique values do not have any significance as they do not contribute much in predictive modeling. Fields with too many null values also need to be discarded.

C. Data Extraction

The attributes are identified for classifying process. In our work, we have worked with numerical and categorical values.

V. RESULTS AND DISCUSSION

The data which is present in MySQL is imported to hive using Sqoop. Steps that are involved in hive are,

- Start installation.
- Preparing to use a MySQL streaming result set.
- Beginning code generation.
- Transferred the data in certain time.
- Retrieving the records.
- Execute SQL statement.
- Loading uploaded data in to Hive.

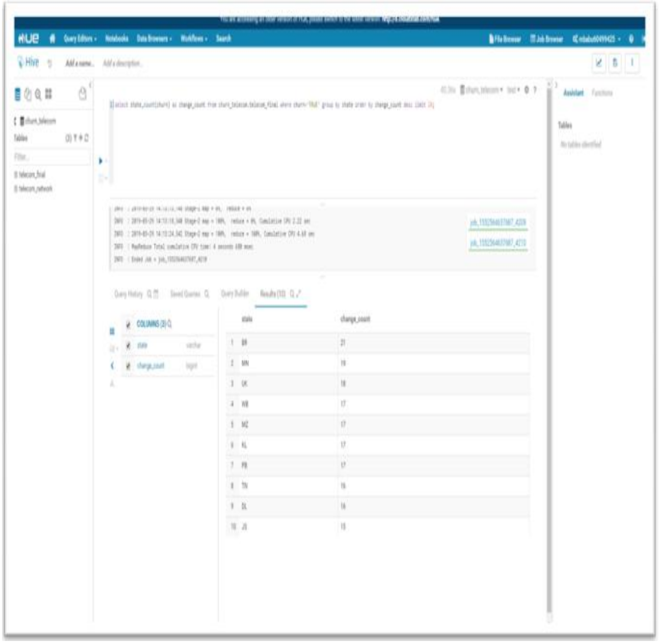


Fig 3. Query 1.

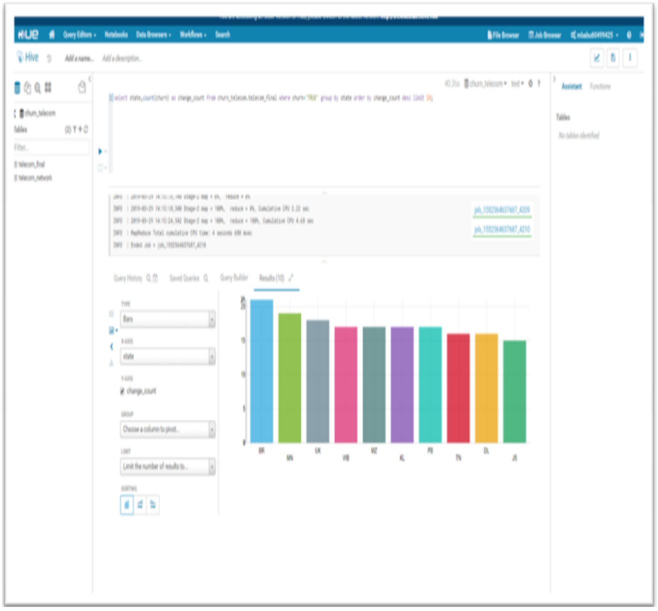
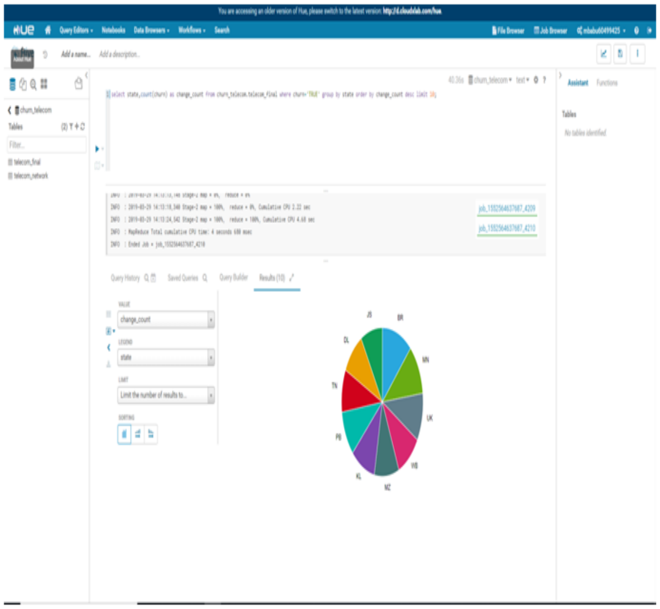


Fig 4. Query 1 Bar Graph.



```
select state,count(churn) as change_count from churn_telecom.telecom_final where churn='TRUE' group by state order by change_count desc limit 10;
```

```
INFO  : 2019-03-29 14:13:13,148 Stage-2 map = 0%,  reduce = 0%
INFO  : 2019-03-29 14:13:18,340 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.22 sec
INFO  : 2019-03-29 14:13:24,542 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.68 sec
INFO  : MapReduce Total cumulative CPU time: 4 seconds 680 msec
INFO  : Ended Job = job_1552564637687_4210
```

job_1552564637687_4209
job_1552564637687_4210

Fig 5. Query 1 Pie Chart.

Apache Hive is a component of Horton works Data Platform(HDP). Hive provides a SQL-like interface to data stored in HDP. In the previous tutorial, we used Pig, which is a scripting language with a focus on data flows. Hive provides a database query interface to Apache Hadoop.

VI. CONCLUSION

Today, Big Data is influencing IT industry like few technologies have done before. The massive data generated from sensor-enabled machines, mobile devices, cloud computing,social media, satellites help different organizations improve their decision making and take their business to another level. "Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives," Susan Hauser, corporate vice president of Microsoft. Data is the biggest thing to hit the industry since PC was invented by Steve Jobs. As mentioned earlier in this paper, every day data is generated in such a rapid manner that, traditional database and other data storing system will gradually give up in storing, retrieving, and finding relationships among data. Big data technologies have addressed the problems related to this new big data revolution through the use of commodity hardware and distribution. Customer complaint analysis is important to find and there's no better way to collect direct feedback from your customers and improve your product or service. However, the way you handle a complaint is the difference between keeping a customer or losing one. So, the next time you receive a customer complaint, listen to what the customer has to say, apologize find a solution and follow up to see if he or she is happy with the way you are handling it. In doing so, you are on your way to creating more loyal customers, improving your product and delivering a better quality of customer service. As earlier loading large amount of data is very difficult. By using Big data complexity of loading large amount of data can be reduced. The proposed tool enables agencies too easily and economically clean, characterize and analyze the data to identify actionable patterns and trends.

VII. REFERENCES

[1] Anderson, R. E., "Consumer Dissatisfaction: The Effect of Disconfirmed Expectancy on Perceived Product Performance," Journal of Marketing Research, February 1973, pp. 38-44.

[2] Barbara, S., "Consumer Complaint Handling as a Strategic Marketing Tool," The Journal of Consumer Marketing (2:4), Fall 1985, pp. 5-17.

[3] Betrand, K., "Marketers Discover What 'Quality' Really Means," Business Marketing (72), 1987, pp. 58-72.

[4] Blodgett J. G., Donald. H. G., and Walters, R. G., "The Effects of Perceived Justice on Negative Word-of-Mouth and Repatronage Intentions," Journal of Retailing (69), Winter 1993, pp. 399-428.

[5] Cho, Y., Im, I., Ferjemstad, J., and Hiltz, R., "Causes and Outcomes of Online Customer Complaining Behavior: Implications for Customer Relationship Management (CRM)," Proceedings of the 2001 Americas Conference on Information Systems, Boston,
August 2001.

[6] Cho, Y., Im, I., Ferjemstad, J., and Hiltz, R., "An Analysis of Pre- and Post-Purchase Online Customer Complaining Behavior," Proceedings of Conference on Customer Satisfaction, Dissatisfaction & Complaining Behavior, Jackson Hole, Wyoming, June 2001.

[7] Day, Ralph L., "Modeling Choices Among Alternative Responses to Dissatisfaction," in Advances in Consumer Research, 11, Thomas C. Kinner ed., Provo, UT: Association for Consumer Research, 1984, pp. 469-499.