A Data Mining Approach for Analyzing Customer Churn Rate in the Telecom Industry

Sabarish Subramaniam A.V
Department of Electronics and Communication Engineering
Sri Venkateswara College of Engineering (SVCE)
Sriperembudur, India

Abstract- **Churners have always posed a major problem for every business that offers services. Churning drives up a company's expenses while also lowering its profit margin. Customers who request service termination is typically exhibiting customer attrition. The people and organizations who contribute the data that is stored in government databases, as well as the organizations who fund the collection of such data, are simultaneously becoming more aware of how tools that increase analytical capabilities also provide threats to the privacy of data records. Nonetheless, it is possible to forecast if a consumer wishes to cancel a service using predictive analysis based on historical service usage, service performance, expenditure, and other behavioral patterns. This study analyzes the topic of churn analysis while imagining a situation in which a business with private databases wants to use a churn analysis method on the combination of those databases without disclosing any unwanted information. The point of the research would be to forecast if a client would leave a telecom provider soon or not, using predictive analysis of billing data.**

Keywords- Churn prediction, Churn analysis, Customer attritions analysis, Client defection analysis

I. INTRODUCTION

DATA MINING

Recognizing patterns in data is commonly referred to as mining. Data mining is becoming a more crucial method to turn this data into information as more data are collected, with the amount of data doubling every three years. It is often applied to several profiling techniques, including marketing, surveillance, fraud detection, and scientific research. The three fields of databases, artificial intelligence, and statistics are connected by the field of data mining. The procedure must be fully automated or, more frequently, semiautomated. The patterns that are found must be significant in that they result in some benefit, often an economic benefit. The information era has made it possible for numerous organizations to collect enormous amounts of data. However, if "meaningful information" or "knowledge" cannot be drawn from this data, its value is minimal. Data mining, sometimes referred to as knowledge discovery, tries to meet this demand. Data mining techniques look for useful information without requiring a priori preconceptions, in contrast to conventional statistical methods. As a discipline, it has brought novel concepts and techniques such as association rule learning. In the context of very big datasets, it has also utilized well-known machine-learning methods like inductive-rule learning (for example, using decision trees). Data mining methods are employed in both industry and research, and their use is growing.

There are many uses for data mining in both the public and business sectors. Data mining is frequently used in the banking, insurance, healthcare, and retail sectors to save costs, improve research, and boost sales. For example:

- Businesses may create models that forecast whether a consumer is a good credit risk or if an accident claim may be fraudulent and needs further investigation using customer data gathered over a few years.

- Retailers can utilize the data gathered through affinity programs such as shoppers' club cards, frequent flyer miles, and competitions to evaluate the success of their judgments on product placement and selection, discount offers, and which goods are frequently bought in combination.

- Applications for data mining are used by the banking and insurance sectors to spot fraud and help evaluate risks (e.g., credit scoring).

- Data mining is occasionally used in the medical industry to forecast a procedure's or drug's success. Pharmaceutical companies employ data mining of genetic information and chemical substances to assist in direct research on novel illnesses therapies.

- Businesses like music venues and phone service providers may use data mining to generate a "churn analysis", which determines which consumers are more likely to remain with them and which ones are more likely to transfer to a competitor.

Churn analysis is one distinctive component of data mining. It is the calculation of the rate of client attrition for any business. Finding customers who are most likely to stop using a service or product includes doing this. The development of a solid and long-lasting plan for client retention in a firm benefits greatly from churn analysis. When a business is aware of the percentage of customers that discontinue doing business with them within a specific time frame, it is simple to utilize churn analysis to develop a complete study of the reasons behind the churn rate. This aids in the company's creation of efficient client retention initiatives.

Churn rate normally pertains to a variety of businesses, but subscription services like long-distance phone service or periodicals are one among them. Churn analysis aids in understanding the actions of customers who unsubscribe and switch to a rival as well as in estimating the chances of this happening. In addition to estimating staff attrition, there are other uses

II. RELATED WORK

An article on a case study of churn analysis was written by Marco Richeldi and Alessandro Petrucci. In this work, the churn analysis tool Mining Mart is used. The preparation of data for analysis using Mining Mart is primarily covered. [4]

In a case study, according to Shyam V. Nath, a database of 50,000 consumers from the cellular telecommunications sector was evaluated to identify potential churners. The analysis was carried out utilizing the Naive Bayes method and supervised learning in the study utilizing JDeveloper tools. [3]

Customer attrition was discussed in a case study by Teemu Mutanen. The methodologies for the prediction, the data utilized, as well as the outcome, were all thoroughly explained in the study. The author provided two churn analysis techniques. Logistic regression is the first one. A discrete result can be predicted using logistic regression using continuous and/or categorical data. Only one dependent variable may exist in this methodology. After converting the dependent variable into a logistic variable, this technique uses maximum likelihood estimation. The second approach examines the logistic regression's estimation outcomes. The lift curve is what it is called. This curve is connected to the precision-recall curve and the ROC curve of the signal detection theory. [2]

The authors describe data mining-based predictive modeling for churners. The report also goes into great depth on how to apply the decision tree analysis approach. In the article, client churning was primarily examined from a commercial standpoint. But they also included case studies, process flows, and modeling strategies near the end of the paper. [1]

III. METHODOLOGY

Churn analysis mostly uses a large amount of historical data. This information is available from the relevant company's data warehouse. The following data is required for a thorough examination of a telecom company's customer attrition:

- Customer demographics, such as location, age, gender, marital status, etc.
- Call stats, including the number of local and long-distance calls, as well as call lengths at various times of the day.
- Each customer's billing information, including the cost of local and long-distance calls.
- Additional service information, which includes the extra plan for which the consumer is signed up, such as discounted long-distance rates.
- The customer-purchased voice and data products and services, such as broadband services, private virtual networks, dedicated data transport lines, etc.
- Information on complaints: how many calls to customer care are made over disputed invoicing, dropped calls, sluggish service provisioning, malfunctioning special services, etc.
- Credit status.

INPUT DATA

An established telecom company made available the data used in this research. The data, however, only includes billing records. We imported the data into MySql to perform some preprocessing. Then run SQL queries to learn more about the data. One could use the phone number, bill payment date, payment location, and payment amount attribute sets from the data gathering. Let's give them the appropriate names, ph, date, loc, and sum. Having learned a lot about the data by running queries on the collection. The list of aspects we discovered is shown below, along with the search terms that were used to locate each aspect.

- The data set has 6938 records. query: *select count(1) from bill*
- The data set contains record of 26 months. query: *select count(distinct substring(date,1,7)) from bill*
- It has 880 phone numbers. query: *select count(distinct phone) from bill*
- Bills were paid from 25 locations. query: *select count(distinct loc) from bill*

CHURN ANALYSIS METHODS:

Churn analysis may be done in a variety of methods. However, these techniques may be broadly divided into two groups. Those are:

- Supervised techniques
- Unsupervised techniques

i) Supervised Techniques:

In supervised approaches, the algorithm essentially picks the best classification for the data based on what it discovers from the training set. A collection of training examples makes up the training data. Each example in supervised learning is a pair that includes an input item and the desired output value. In this instance, the rows of our data would serve as the input object, and the output result would be a binomial value indicating whether or not the customer has left. Our dataset, however, lacks any characteristic that may predict who would churn. So, with our dataset, a supervised technique is not an option.

ii) Unsupervised Techniques:

Unsupervised approaches, on the other hand, approach the issue by looking for hidden structures in unlabeled data. Unsupervised techniques can be used to cluster our data set. The churners might then join a different cluster in this manner.

Again, we may employ techniques that help the data self-learn. Rule-based classifiers are one type of unsupervised learning approach. According to pre-established principles, the technique learns the data. As a result, the technique learns for itself. Extracting features from our dataset may be used to build the rules.

IV. ANALYSIS OF RESULTS

The last examination of our dataset reveals certain instances that can be categorized as distortion or impurity. We must exclude these impure records from our data collection to improve the analytical results. One example of this is the billing record for the phone number, which appeared just once in the whole dataset. We could find them with the following query:

select count(1), phone from bill group by phone having count(1) = 1

The total number of records that only occurred once in the dataset can be found with the following query:
select count(1) from bill where phone in (select phone from bill group by phone having count(1) = 1)

With only one occurrence of each phone number in the dataset, we now have 303 records overall. Therefore, 303 of the 880 phone numbers may be categorized as noise. Consequently, 577 of the dataset's entries are legitimate. Due to the fact that we have data for 26 months and that bills are paid once per month, the client who paid bills for the full 26 months would have one frequency of payment. The lower frequency may indicate irregular bill payments by the consumer. Therefore, the likelihood that a given consumer will leave increases with decreasing frequency. We added the frequency number for each phone number to a new database called frequency.

Using the following query, we can get the customers' maximum and minimum frequency:

select max(frequency), min(frequency) from frequency

A higher frequency rate indicates that some customers have paid their bills more than once in a given month. The greatest frequency value from this query is 1.1923, and the minimum frequency value is 0.0385. This is conceivable since a consumer can pay their entire balance in many instalments.

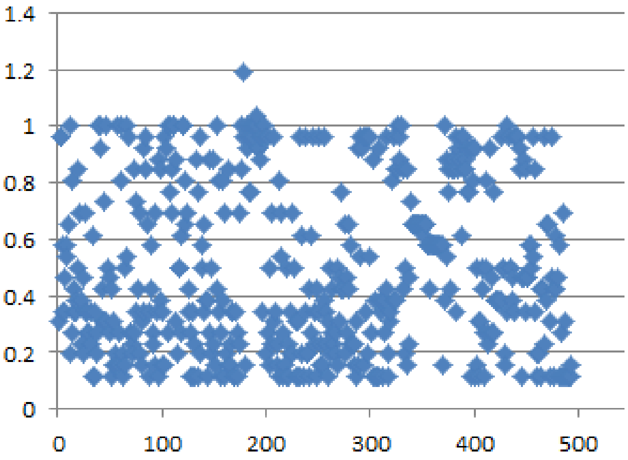

Figure 1: Scatter chart of frequency. [x-axis: frequency, y-axis: record no.]

Fig. 1 shows a scatter plot of each customer's frequency, which is higher than or equal to 0.10. Figure 1 shows that there is just one client with a frequency greater than 1. Around line 0.6, there is a definite gap that is visible. A decent place to set our boundaries is at this moment. Therefore, a client that churns is one whose frequency is less than 0.6. The resulting rule is as follows:

r1: (frequency < 0.6) → Churner

The decision tree for this rule will look like this:

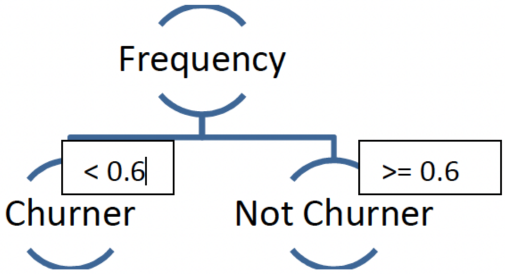

Figure 2:Decision tree for frequency rule

By separately charting quantity versus date for each client in a line curve, this rule may be verified. The likelihood of the client leaving is lower if the line slopes. While a dropping line indicates that customers are gradually ceasing to use the service, this might indicate customer churn. We have selected two clients and plotted their payments here to provide a better view. One client has a frequency of 0.9651, whereas the other has a frequency of 0.5769.

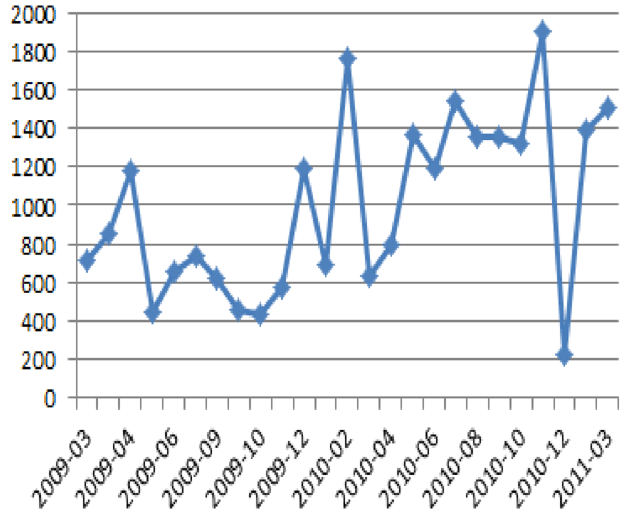

Figure 3:Line graph of a single customer with frequency 0.9651

By examining the two graphs, it is clear that the client with the greater frequency has an inclined line over time, indicating that they are utilizing the service more frequently. In contrast, the line graph of the client who uses the service less frequently displays a falling line that over time indicates decreased usage of the service. Therefore, according to our criterion, the first customer is not a churner and the second customer is a churner.

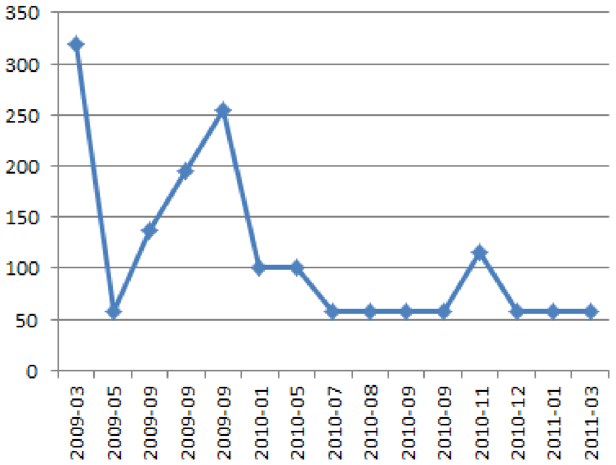

Figure 4:Line graph of a single customer with frequency ***0.5769***

We can simply identify the churners based on the rule. The estimated number of churners based on this rule is shown in the table below. To cluster our dataset, we employed the RapidMiner data mining tool. We gave the tool both the bill and the frequency table. The technology, however, was unable to appropriately cluster them. Two clusters, one for each class, were what we were hoping for. Classes are churners, not churners at all. The tool did, however, cluster the dataset into more than two segments in both instances.

Churners (frequency < 0.6)	Not Churners (frequency>=06)	Not Classified (frequency<0.1)
315	187	387

Table 1: Results

V. OBSTACLES

1) *Incomplete Data:* The information we used was insufficient. Consequently, the outcome is inaccurate. However, it provides an illustration of how churn analysis functions.
2) *Time:* Churn analysis is a difficult and drawn-out procedure. As a result, there was not enough time to do a thorough churn analysis and produce a workable result.
3) *Concerns about confidentiality:* are a major issue that occurs with any mass collection of data. When privacy is required by law (such as for medical databases), corporate interests may also be at play. However, there are some circumstances where exchanging data might benefit both parties. Nowadays, research, whether it be scientific or commercial and market-oriented, is a major use of massive datasets. Therefore, industries like medicine have a lot to gain by sharing data for research, as do rival companies with compatible goals. Despite the apparent benefit, confidentiality concerns make this frequently impractical.

VI. FUTURE WORKS

The outcome that is obtained is not encouraging. We have an incomplete dataset, which is why this is the case it failed because we were unable to manage a suitable dataset for churn analysis. There are a lot of things we would like to accomplish with this project in the future. Following is a list of some of them:

- When implementing the churn analysis approach, use a complete dataset.
- Use data mining tools properly to forecast churning.
- Use a variety of tools to assess churn.
- Compare many approaches to choose the best one.
- Make use of and contrast various data mining tools.

We will test the efficacy of a noise addition algorithm that will conceal the data in order to maintain privacy.

VII. SOFTWARE USED

1) *MySQL database:* The dataset was stored in a MySQL database. In general, the database contained the bill and frequency records.
2) *HeidiSQL IDE*: This program functions as a front end for MySQL. We connected to our database using it and ran queries. This tool also assisted everything in exporting data in the CSV format needed by data mining software.

VIII. CONCLUSION

Having attempted to give churn analysis in the telecom business in this research. The churn prediction analysis is based on rule-based classification. However, it was challenging to develop a predictive model due to the dataset's incompleteness. The results show that if we want to obtain any level of accuracy, we must employ a comprehensive and substantial dataset.

Churn analysis is a very significant problem in many applications of data mining. In this field, many approaches are likely to be crucial. However, this research highlights some of the difficulties that these methods have while doing churn analysis.

It shows that, under some circumstances, it is rather simple to evade the privacy protection provided by the various strategies. It offered detailed experimental findings with many sorts of data, demonstrating the seriousness of the issue. The research also raises this issue and provides a model churn analysis approach that may be used more widely in creating a fresh viewpoint for creating an improved churn analysis technique.

IX. REFERENCES

[1] K. B. Oseman, S.B.M. Shukor, N. A. Haris, F. Bakar, “Data Mining in Churn Analysis Model for Telecommunication Industry”, Journal of Statistical Modeling and Analytics, Vol. 1 No. 19-27, 2010.

[2] T. Mutanen, Customer churn analysis – a case study, from:http://www.vtt.fi/inf/julkaisut/muut/2006/customer_churn_ case_study.pdf, April 12, 2014.

[3] S.V. Nath, Customer Churn Analysis in the Wireless Industry: A Data Mining Approach, from: http://download.oracle.com/owsf_2003/40332.pdf, April 14, 2014.

[4] M. Richeldi and A. Perrucci, “Churn Analysis Case Study”,from:http://sfb876.tudortmund.de/PublicPublicationFiles/richeldi_perrucci_200 2b.pdf, April 12, 2014.

[5] Wei CP, Chiu IT. Turning telecommunications call details to churn prediction: a data mining approach. Expert Syst Appl. 2002;23(2):103–12.

[6] S. Gupta and V. Zeithaml, “Customer Metrics and Their Impact on Financial Performance,” Mark. Sci., vol. 25, no. 6, pp. 718–739, Nov. 2006.

[7] Z. Qian, W. Jiang, and K.-L. Tsui, “Churn detection via customer profile modelling,” Int. J. Prod. Res., vol. 44, no. 14, pp. 2913–2933, Jul. 2006.

[8] F. Robert Dwyer, “Customer lifetime valuation to support marketing decision making,” J. Interact. Mark., vol. 11, no. 4, pp. 6–13, 1997.

[9] J. Franklin, “The elements of statistical learning: data mining, inference, and prediction,” Math. Intell., vol. 27, no. 2, pp. 83–85, Nov. 2008.

[10] S. A. Neslin, S. Gupta, W. Kamakura, J. Lu, and C. H. Mason, “Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models,” J. Mark. Res., vol. 43, no. 2, pp. 204–211, May 2006.

[11] M. Karnstedt, M. Rowe, J. Chan, H. Alani, and C. Hayes, “The Effect of User Features on Churn in Social Networks,” presented at the ACM WebSci’11, Koblenz, Germany, 2011, pp. 1–8.