Data Mining in Telecommunication Industry

Hiren H Darji[1]
[1]Assistant Professor

Abstract— Telecommunication companies today are operating in highly competitive and challenging environment. Vast volume of data is generated from various operational systems and these are used for solving many business problems that required urgent handling. These data include call detail data, customer data and network data. Data Mining methods and business intelligence technology are widely used for handling the business problems in this industry. The goal of this paper is to provide a broad review of data mining concepts.

Key words: Data mining, Telecommunications, Fraud Detection, Neural Networks, Churn management

I. INTRODUCTION

The concept of Data Mining has gained a common market acceptance. Telecommunication is one of the most data intensive industries in the world. One of the first industries to accept Data Mining is the telecommunications. Companies in the telecom industry are making use of Data Mining technologies to improve their marketing techniques, for identification of customer fraud and for the better management of their networks.[1]

Most of the telecom companies have realized that the vast volume of data they collect and possess could be effectively utilized for solving their business problems by converting them into information and knowledge. Data Mining can be viewed as a technique automatically generating this knowledge from the data available. One of the first industry to experience the benefits from the application if Business Intelligence (BI) and Data Mining technologies in the telecommunications industry. [1]

II. DATA MINING: AN OVERVIEW

A. Definition

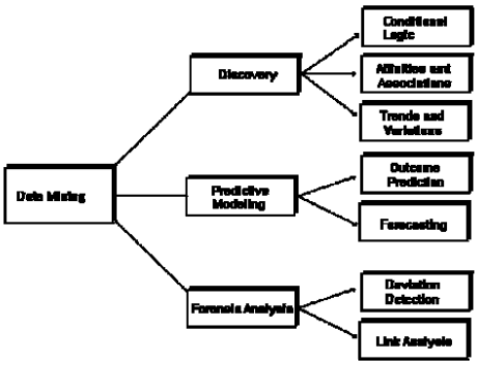
"Data mining" is defined as a sophisticated data search capability that uses statistical algorithms to discover patterns and correlations in data. Data mining finds and extracts knowledge ("data pieces") hidden in corporate data warehouses, or information that visitors have dropped on a website, most of which can lead to improvements in the understanding and use of the data. Data mining discovers patterns and relationships hidden in data, and is actually part of a larger process called "knowledge discovery" which describes the steps that must be taken to ensure meaningful results. Data mining helps business analysts to generate suggestions, but it does not validate the hypotheses. [7]

B. The evolution of data mining

Data mining techniques are the result of a long research and product development process. The origin of data mining lies with the first storage of data on computers continues with improvements in data access, until today technology allows users to navigate through data in real time. In the evolution from business data to useful information, each step is built on the previous ones. [7]

C. Data mining techniques

Data mining tools take data and construct a representation of reality in the form of a model. The resulting model describes patterns and relationships present in the data. From a process orientation, data mining activities fall into three general categories [7]



1) Discovery:
The process of looking in a database to find hidden patterns without a fixed idea or hypothesis about what the patterns may be.

2) Predictive Modeling:
The process of taking patterns discovered from the database and using them to predict the future.

3) Forensic Analysis:
The process of applying the extracted patterns to find strange or unusual data elements. Data mining is used to construct six types of models expected at solving business problems: classification, regression, time series, clustering, association analysis, and sequence discovery. The first two, classification and regression, are used to make predictions, while association and sequence discovery are used to describe behavior. Clustering can be used for either forecasting or description.

III. TYPES OF TELECOMMUNICATION DATA

The different kinds of data used in this industry are mainly grouped into 3 different types.

A. Call detail data

This is the information about the call, which stores as the call detail record. The number of call detail records generated is huge since every call is placed on the network, the details are stored. Call detail record includes information like making and ending phone numbers, date, time and duration of call. Usually these call detail records are not directly used for Data Mining. A list of features can be generated from the call detail data such as [1]

- Average call duration
- Average number of call created per day
- Average number of call received per day
- Percentage of no-answer calls

- Percentage of day time calls
- Percentage of weekday calls

B. Network data

Telecommunication networks contain thousands of components, which are interconnected. These components are capable of generating error and status messages which leads to a large volume of network data. These network data are used for network management functions like fault detection. Expert systems have been developed to analysis these messages automatically, since the huge volume of network messages generated cannot be handles by technicians. Hence Data Mining technologies are used in identification of network faults by automatically extracting knowledge from network data. Network data is also generated in real time which can be accomplished by applying a time window to the data. [1]

C. Customer data

Like any other business, telecommunication companies also have millions customers. Hence it is very much essential to have a database for storing the information about these customers. Information about the customer will include:

- Name of the customer
- Address details
- Payment history
- Service plan and so on

Group customer data is used to provide call detail data in order to identify fraud. [1]

IV. CHURN PREDICTION - PROBLEM DESCRIPTION

In a business environment, the term, customer attrition simply refers to the customers leaving one business service to another. Customer churn or subscriber churn is also similar to attrition, which is the process of customers switching from One service provider to another anonymously. From a machine learning perspective, churn prediction is a managed problem defined as follows: Given a predefined forecast horizon, the goal is to predict the future churners over that horizon, given the data associated with each subscriber in the network. The churn prediction problem represented here involves 3 phases, namely, I) training phase, ii) test phase, iii) prediction phase. The input for this problem includes the data on past calls for each mobile subscriber, together with all personal and business information that is maintained by the service provider. In addition, for the training phase, labels are provided in the form of a list of churners. After the model is trained with highest accuracy, the model must be able to predict the list of churners from the real dataset which does not include any churn label. In the viewpoint of knowledge discovery process, this problem is categorized as predictive mining or predictive modeling.

Churn Prediction is a occurrence which is used to identify the possible churners in advance before they leave the network. This helps the CRM department to prevent subscribers who are likely to churn in future by taking the required retention policies to attract the likely churners and to retain them. Thereby, the potential loss of the company could be avoided. This study utilizes data mining techniques to identify the churners. [3]

V. METHODOLOGY

KDD (Knowledge Discovery in Databases) is defined as the "nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns of in data". The first step in predictive modeling is the acquisition and preparation of data. Having the correct data is as important as having the correct method. [3]

A. Data Acquisition

It is a difficult problem for the researchers to acquire the actual dataset from the telecom industries. This is because the customer's private details may be misused. Since churn prediction models requires the past history or the usage behavior of customers during a specific period of time to predict their behavior in the near future, they cannot be applied directly to the actual dataset. Therefore, it is the usual practice to perform some kind of aggregation on the dataset. During the process of aggregation, in addition to the actual variables, new variables will be generated which display the periodic consuming behavior of the customers. These variables have vital information to be used by the prediction models in forecasting the behavior of customers in advance. The dataset used here was aggregated for 6 months duration.

B. Data Preparation

In data mining problems, data preparation consumes considerable amount of time. In the data preparation phase, data is collected, integrated and cleaned. Integration of data may require extraction of data from multiple sources. Once the data has been arranged in tabular form, it needs to be fully characterized. Data needs to be cleaned by resolving any ambiguities, errors. Also redundant and problematic data items are to be removed at this stage.

C. Derived Variables

Derived variables are new variables based on original variables. The most effective derived variables are those that represent something in the real world, such as a description of some original customer behavior. There are some general classes of derived variables, like total values, average values, and ratios. Some examples are:

- The average number of calls in last 6 months
- The average number of late in last 6 months
- The ratio of incoming and outgoing calls
- Average payment amount for last six months
- Average late count in last 6 months

D. Variable Extraction

The selected variables are grouped under 4 categories and are described below

1) Customer Demography:

- Age: It is found that the customers between the age group of 45 – 48 have high probability to churn.
- Line_Tenure: Customers with 25 – 30 months of tenure period are about to churn.
- Customer_Class : Generally the churn probability of the corporate account holders is high. This is due to the fact that their account will be maintained by the company and customers who

quit the company would churn. The Customer Class can be any one of VIP /Individual/Corporate.

- Days_to_Contract_Expiry: Most of the customers would subscribe to a new service with the intention

Acquiring new HAND_SET. These people would leave the network after the contract expires.

2) Bill and Payment

- Average_Bill_Amount
- Avg_Pay_Amount
- Overdue_Payment_Count

3) Call Detail Record

- Avg_Min_OB: If the average out bound call is less than 168 minutes they will churn.
- Tot_Past_Delink: If the count of total past delink is greater than 3 then they will churn.
- Tot_Dis_Int: If the customers who make more number of distinct international calls then they will churn. If the count is greater than 6 they may churn.

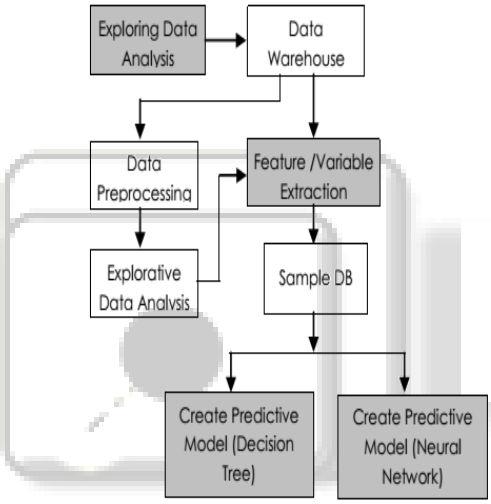


Figure 1: Churn prediction model

VI. CONCLUSION

Data Mining play a significant role in the telecommunication industry due to the availability of large volume of data and the rigorous competition in the sector. The primary application areas include marketing and Customer Relationship Management, Fraud detection and Network Management. The recent developments in the Data Mining and the implementation and enhancement of existing techniques and methods ensure the continuous growth and compatibility of telecommunication companies that make use of them.

REFERENCES

[1] Madhuri V. Joseph, "Data Mining and Business Intelligence Applications in Telecommunication Industry".
[2] D. Ćamilović*, "DATA MINING AND CRM IN TELECOMMUNICATIONS".
[3] V. Umayaparvathi, K. Iyakutti, "Applications of Data Mining Techniques in Telecom Churn Prediction ".
[4] Khalida binti Oseman, Sunarti binti Mohd Shukor, Norazrina Abu Haris, Faizin bin Abu Bakar, "Data Mining in Churn Analysis Model for Telecommunication Industry".
[5] N.Kamalraj, Dr.A.Malathi, "Applying Data Mining Techniques in Telecom Churn Prediction".
[6] Wiktor Daszczuk, Piotr Gawrysiak, Tomasz Gerszberg, "Data Mining for Technical Operation of Telecommunications Companies: a Case Study".
[7] Anita B. Desai, Dr. Ravindra Deshmukh, "Data mining techniques for Fraud Detection".
[8] Pareek, D.: Business Intelligence for Telecommunications. Auerbach Publications, Taylor & Francis Group LLC. (2007).
[9] Yu-Teng Chang, "Applying Data Mining To Telecom Churn Management", IJRIC, 2009 67 – 77.