Churn Prediction Using Hadoop

Avishkar Dalvi[1], Bhushan Bhor[2], Hiren Chauhan[3], Prof. Prashant Sawant[4]

[123]*Student, Dept. of Information Technology, K.J.S.I.E.I.T, Mumbai, Maharashtra, India*
[4]*Assistant Professor, Dept. of Information Technology, K.J.S.I.E.I.T, Maharashtra, India*

---***---

Abstract - *with the rapid increase in mobile telecommunication market and with numerous service providers stepping in the market and makes the customer think to switch from the service provided by one service provider and move to another service provider for a more satisfactory experience. This project is an attempt to design and implement an application that can take Customer Records as input and give* ***Customer churn prediction*** *details as output. Thus it will help the service provider to have a somewhat clear idea about the existing customers who might be the potential churners in the near future. By merely giving customer data records as input, user can get the desired customer behavior pattern, which is the churn output. The output obtained will distinguish the churners and the non-churners. The system is built using* ***Apache HBase****,* ***Apache Hadoop*** *and a Data Mining Algorithm called* ***C4.5****. The processing of large datasets containing the information of customers is made easier because of the use of the* ***Hadoop framework****.*

1. INTRODUCTION

The telecommunication industry in India has shown a consistent overall growth rate of more than 35 percent over the past decade in terms of subscribers, thus it is a great economic success [1]. Therefore, the aim of these telecommunication companies mainly is to retain the existing customers rather than increasing the number of customers. Therefore, it becomes crucial for these companies to have knowledge of the customers who might leave their services and switch to the competitor. A customer who leaves one service provider and joins the services given by another service provider is known as churn customer [1]. The analyzation of customer behavior based on several attributes such as his number of calls per day or by the usage of services provided to him, helps in determining the churners. Churn prediction has currently proven to be an appropriate subject in data mining and has shown useful applications in the field of banking, mobile telecommunication, life insurances, and others [1]. This type of prediction allows the companies to focus their resources on the customers who are about to churn. In addition, it will avoid the loss to the company [1]. Telecommunications companies create enormous amounts of data every day. The data these companies have includes call history of the customer, details about various plans enabled by the customers, etc. This data helps in determining whether the customer is a potential churner or not. As the data is large and unorganized, it becomes necessary to have an efficient system which can handle such data. For this reason, the open source framework, Apache Hadoop along with MapReduce and HBase are used. The components of Hadoop provide efficient processing and mining of data. The Hadoop Distributed File System is appropriate for storage of large datasets. Here in this project, we are going to use decision tree approach using C4.5 Data Mining Algorithm, which provides great accuracy in predicting churners.

2. PROBLEM DEFINITION

The mobile telecommunication market is increasing rapidly as numerous service providers are stepping in the market. Because of this, the customer may think of leaving the service provided by one service provider and switch to another service provider in order to get a more satisfactory experience. This project is an attempt to design and implement an application that can take Customer Records as input and give Customer churn prediction details as output. It will help the service provider to know in advance about the valuable customers who have the highest possibility to churn. By merely giving customer data records as input, user can get the desired customer behavior pattern, which is the churn output. The output obtained will help to distinguish the churners from the non-churners. The system is built using Apache Hadoop, Apache HBase and a Data Mining Algorithm called C4.5. The processing of large datasets containing the information of customers is made easier because of the use of the Hadoop framework.

The main aim of this project is to help to predict the churn in telecommunication domain using Hadoop and C4.5. Customer churn has been identified as one of the major issues in Telecom Industry. Telecom research indicates that it is more economical to retain an existing customer than to gain a new one. In order to retain existing customers, Telecom providers need to know the reasons of churn, which can be realized through the knowledge extracted from Telecom data.

The efficiency and reliability of handling Big Data increases with Hadoop. This system can handle Big Data easily as compared to the existing system. Minimum grabbing of data will fetch more than 1 million records. The system is reliable to handle this data efficiently. The system can be scaled as per the user requirement. The proposed system meets all the requirements of the users and scales as per user needs.

Graphs and charts will provide a clear picture of the latest trends. Reading a graph is much easier and clear as compared to a whole database. They provide comparative results of the grabbed data, which makes the system user friendly. Because of the use of Hadoop, which has HDFS, storage the data is more reliable since many duplicates are made of any file according to the replication factor. Since Hadoop is easily scalable, there is no need to worry about the availability issues.

3. LITERATURE REVIEW

Clustering, Neural Networks, Regression, Decision Tree and Support Vector Machine are some of the different approaches used for churn prediction under data mining technologies. This project uses decision tree approach using C4.5 Data Mining Algorithm, which shows a greater accuracy in giving results than any other algorithm mentioned above. The following points will help us understand why C4.5 algorithms is better than the aforementioned algorithms.

3.1. NEURAL NETWORK:

In the field of artificial intelligence, neural network models are generally called as artificial neural networks (ANNs); these are basically simple mathematical models that define a function or a distribution over or both and, but sometimes models are also intimately associated with a particular learning algorithm or learning rule [2]. ANN consists of topological structures of nodes which process information and distribute that information in a parallel manner [2][4]. Combinations of nonlinear transfer functions helps us to obtain the mappings of inputs and estimated output responses [2]. Past experience, neural cells, memory and association etc [2]. can be used along with self-adaptive information pattern recognition methodology to analyze the training algorithms of neural networks [2][4].

3.1.1. Disadvantages of Neural Network: -

Neural networks cannot serve as a substitute for understanding the problem deeply and they are not probabilistic. Neural networks are too much of a black box: This has several consequences such as it makes them difficult to train: the training outcome can be nondeterministic and crucially dependent on the choice of initial parameters, e.g. the starting point for gradient descent when training back propagation networks [2][4]. Neural networks are not magic hammers: many still believe the myth that, neural networks are magic hammers which can solve any machine learning problem, and as a result, people apply them to problems which are not suited for neural networks to solve [4]. Although neural networks have been proven to be successful for certain domain specific problems, as a consumer of machine learning technology, it is more advisable to go for approaches having comparatively more robust theoretical foundation, rather than just blindly using a general-purpose neural network at your problem and hoping for the best [4].

3.2. SUPPORT VECTOR MACHINES:

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

3.2.1 Disadvantages of SVMs:

The support vector approach is based on the kernel and that is one major limitation. A second limitation is that of speed and size in training and testing. Another problem is that the SVMs can be terribly slow in test phase, even though they have a good general performance. SVMs - from a practical point of view - have some drawbacks. Another important unsolved practical question is the selection of the kernel function parameters. Another drawback of SVMs is the high algorithmic complexity and large memory requirements of the required quadratic programming in large-scale tasks.

3.3. CLUSTERING:

Cluster analysis or clustering is the task of grouping a set of similar objects in the same group (called a cluster) than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Cluster analysis is a general task to be solved, instead of one specific solution. This can be achieved by using various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

3.4. REGRESSION ANALYSIS:

In statistics, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors') [5]. More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quintile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function, which can be described by a probability distribution. Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables [5]. However, this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation.

3.5. ID3 ALGORITHM:

ID3 (Iterative Dichotomizer 3) is an algorithm used to generate a decision tree from a dataset. The data split up more purely by some attributes than others. That means that their values show greater and more consistent correspondence with instances having specific values of the target attribute (the one we want to predict) than those of another attribute [5]. The ID3 algorithm begins with the original set S as the root node. On each iteration, the algorithm iterates through every unused attribute of the set S and calculates the information gain of that attribute, then the attribute which has the largest information gain value is selected .The set S is then split by the selected attribute to produce subsets of the data [5]. The algorithm continues to recur on each subset, considering only attributes never selected before.

3.6. ENTROPY:

The entropy of a dataset is the degree of uncertainty of that data or how disordered that dataset is. It is related to information, in the sense that if that data has higher entropy or higher uncertainty then the amount of information, which will be required to completely describe that data will also be more. While building a decision tree, our aim is to minimize the entropy of the dataset until we reach leaf nodes at which point the subset that we are left with represents instances all of one class and has zero entropy (all instances have the same value for the target attribute). The entropy of a dataset S, with respect to the target attribute, with the following calculation: Entropy(S) =PC i=1 pi log2 pi where Pi is the proportion of instances in the dataset that take the ith , value of the target attribute, which has C different values.

3.7. INFORMATION GAIN:

The reduction in entropy is called as information gain (Gain in information) that would result in splitting the data on an attribute, A. Gain (S, A) =Entropy(S)-P|Sv| |S| Entropy(Sv) where v is a value of A, |Sv| is the subset of instances of S where A takes the value v, and |S| is the number of instances. The nodes in the tree represent the features, while the possible values connecting the features are represented by the branches. A leaf terminates a series of nodes and branches. Initially, an attribute with best information gain at root node is searched by the method the tree is divided into sub-trees.

Similarly, the same rule is followed and each sub-tree is further separated recursively. Once the leaf node is reached or there is no information gain, then the partitioning stops. ,The rules can be obtained by traversing each branch of the tree, once the tree is created.

3.8. ALGORITHM

C4.5 has following advantages over ID3: Employ information gain ratio instead of information gain as a measurement to select splitting attributes. Not only discrete attributes, but also continuous ones can be handled handling incomplete training data with missing values Prune during the construction of trees to avoid over-fitting Handling attributes with different costs. Handling training data with missing attribute values- C4.5 allows attribute values to be marked as '?' For missing. [3] Missing attribute values are simply not used in gain and entropy calculations. Handling both continuous and discrete attributes- in order to handle continuous attributes. [3] C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it [3]. Pruning trees after creation- C4.5 goes back through the tree once it has been created and attempts to remove branches that do not help by replacing them with leaf nodes [3].

4. METHODOLOGY

Hadoop: Apache Hadoop is an open source software framework. Hadoop consists of two main components: a distributed processing framework named MapReduce and a distributed file system known as the Hadoop distributed file system, or HDFS [7]. One of the most important reasons for using this framework in this project is to process a large amount of data and do its analysis, which is not possible with other system [11]. The storage is provided by HDFS and MapReduce does the analysis. Although Hadoop is best known for MapReduce and its distributed file system, the other subprojects provide complementary services, or build on the core to provide high-level abstractions [9].

HDFS: The Hadoop Distributed File System (HDFS) is the storage component. In short, HDFS provides a distributed architecture for extremely large-scale storage, which can easily be extended by scaling out. When a file is stored in HDFS, the file is divided into evenly sized blocks. The size of block can be customized or the predefined one can be used. In this project, the customer dataset is stored in HDFS. The dataset contains many customer records, which are the main constraint of this project. Also, the output of is written into HDFS [8].

MapReduce: MapReduce is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster MapReduce works by breaking the processing into two phases: the map phase and the Reduce phase. Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer. The programmer also specifies two functions: the Map function and the Reduce function. The input to our map phase is the raw data of customers. We choose a text input format that gives us each line in the dataset as a text value. The key is the offset of the beginning of the line from the beginning of the file. The output from the map function is processed by the MapReduce [12] framework before being sent to the reduce function. This processing sorts and groups the key-value pairs by key [10]. HBASE: HBase is a distributed column-oriented database built on top of HDFS. HBase is the Hadoop application to use when you require real-time read/write random access to very large datasets. It provides full consistency of data, which means the database is quickly updated [10]. As HBase has been built on top of Hadoop, it supports parallel processing. HBase can be used as data source as well as data sink. It can used be used to retrieve a particular customer's detail by writing a query.

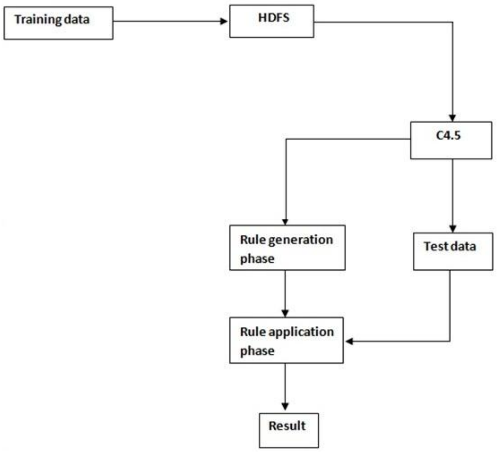

Fig -1: System Architecture

4.1 ALGORITHM:

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set of already classified samples [14][6]. Each sample consists of a p-dimensional vector, where they represent attribute values or features of the sample, as well as the class in which falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other [6]. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision [6]. The C4.5 algorithms then recurs on the smaller sub lists. This algorithm has a few base cases. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class. None of the features provide any information gain [6]. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class. Instance of previously unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Pseudo code-In pseudo code, the general algorithm for building decision trees can be given by this following code:

```
Training Data1 (Customer_id, Week number, Incomingcalls, Incomingcalldrop, Outgoingcalls, Outgoingcalldrop, %failed Incoming, %failed Outgoing);
Test Data (Customer_id, Week number, Incomingcalls, Incomingcalldrop, Outgoingcalls, Outgoingcalldrop, %failed incoming, %failed Outgoing);
Map(Training Data1, Test Data);
Prediction: [6]
Average%ofCallFailures > 30% = Yes = Churner
Average%ofCallFailures <= 30% = No = Non-Churner
CallFailures=Yes=ChurnersCallFailures=No=Non-
```

5. CONCLUSION

The result obtained after applying the Data Mining Algorithm is a list of churners and non-churners. This information will be shown in the output folder of HDFS. The results with corresponding customer records are added to a text file located in HDFS.

The main reason to write the customer records into HBase so that it can be accessed easily at times other than processing. For example, if the company wants to know the list of customers who have their international plan enabled then this can be done by writing a simple query to the HBase table to get the information. Also, searching a record by its unique id or by name of the customer is very easy through HBase. In other words, data about customers can be easily viewed from HBase. This is one of the benefits of using HBase over any other database systems

ACKNOWLEDGEMENT

We are grateful to be associated with the University of Mumbai and our college, K.J. Somaiya Institute of Engineering and Information Technology, Mumbai for providing us the necessary resources and facilities to accomplish our work on the project “Churn prediction using MapReduce and HBase.”

It is a pleasure that we find ourselves penning down these lines to express our sincere thanks to the people who helped us along the way in completing our project. We find inadequate words to express our sincere gratitude towards them.

First and foremost, we would like to express our gratitude towards our Project guide, Prof. Prashant Sawant for placing complete faith and confidence in our ability to carry out this project and for providing us his time, inspiration, encouragement, help, valuable guidance, constructive criticism and constant interest.

We would also like to thank faculty members of our department for their guidance and wholehearted devotion that inspired us to take this project as well as helped us in fulfilment of our desired task. We would also like to thank our family members and friends for their constant trust and support, and their valuable suggestions that has helped us in making our project better.

REFERENCES

[1] Churn Prediction using MAPREDUCE S. Ezhilmathi Sonia, Prof. S. Brintha Rajakumar, Prof. Dr. C. Nalini , International Journal of Scientific Engineering and Technology , Department of CSE, B. I. S. T. (Bharath University), Chennai, TN, India Volume No.3 Issue No.5, pp : 597-600

[2] A Survey On Data Mining Techniques in Customer Churn Analysis for Telecom Industry Amal M. Almana*, Mehmet Sabih Aksoy**, Rasheed Alzahrani***, Amal M. Almana et al Int. Journal of Engineering Research and Applications, Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia., Vol. 4, Issue 5(Version 6), May 2014, pp.165-171

[3] [3] A MapReduce Implementation of C4.5 Decision Tree Algorithm, University, International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.49-60.

[4] A Neural Network based approach for predicting customer churn in cellular network services, Anuj Sharma, Dr. Prabin Kumar Panigrahi, International Journal of computer application (0975-8887), Vol 27-11, August 2011

[5] Comparison of classification methods based on the type of attributes and sample size, Reza Enteleki-Maleki, Arash Rezaei and Behrouz Minaei-Bidgoli, Dept. of Computer Engineering, Iran University of Science and Technology (IUST).

[6] A Map Reduce Implementation of C4.5 Decision Tree Algorithm, Wei Dai and Wei Ji, School of Economics and Management, Hubei Polytechnic University, Huangshi 435003, Hubei, P.R.China , International Journal of Database Theory and Application , Vol.7-1 (2014), pp 49-60 .

[7] Tom White, Hadoop: The Definitive Guide, 3rd ed., O‟Reilly Media, Inc.

[8] The Apache Hadoop website. [Online],2008.

[9] The Apache HBase website,2012. [Online]. Available:http://hbase.apache.org/

[10] (2010) The Apache HBase website. [Online]. Available:http://gethue.com/

[11] Apache Hadoop Wikipedia page: http://en.wikipedia.org/wiki/Apache_Hadoop

[12] MapReduce: http://en.wikipedia.org/wiki/MapReduce

[13] C4.5: http://en.wikipedia.org/wiki/C4.5_algorithm

[14] Hue: http://en.wikipedia.org/wiki/Hue_%28Hadoop%29

[15] http://a4academics.com/final-year-be-project/11-be-itcsecomputer-scienceproject/560-data- mining-byevolutionary-learning-dmel-using-hbase

[16] http://www.informit.com/articles/article.aspx?p=2253412