A Graph-Based Churn Prediction Model for Mobile Telecom Networks

Saravanan M.[1] and Vijay Raajaa G.S.[2]

[1] Ericsson R & D, Chennai, India
{m.saravanan@ericsson.com}
[2] Thiagarajar College Of Engineering, Madurai, India
{gsvijayraajaa@gmail.com}

Abstract. With the ever-increasing demand to retain the existing customers with the service provider and to meet up the competition between various telecom operators, it is imperative to identify the number of visible churners in advance, arbitrarily in telecom networks. In this paper, we consider this issue as a social phenomenon introduced to mathematical solution rather than a simple mathematical process. So, we explore the application of graph parameter analysis to the churner behavior. Initially, we try to analyze the graph parameters on a network that is best suited for node level analysis. Machine learning and Statistical techniques are run on the obtained graph parameters from the graph DB to select the most significant parameters towards the churner prediction. The proposed novel churn prediction methodology is finally perceived by constructing a linear model with the relevant list of graph parameters that works in a dynamic and a scalable environment. We have measured the performance of the proposed model on different datasets related to the telecom domain and also compared with our earlier successful models.

Keywords: Call Graph, Churn Prediction, Hadoop Framework, Graph DB, Graph Parameters, Dynamic and Scalable environment.

1 Introduction

Churn in the telecom industry refers to the movement of customers from one operator network to the other. It is an interesting social problem which relates not only to surviving the competition among telecom service providers but also to better understanding of their own customers [1, 2]. It is all the more important as customer churn leads to diminished profits for the operator and enhanced business for the telecom operator's competitor. Moreover, it is more important for the operator to retain an existing customer than to get a new one. With the continuous addition of new operators in the market and with the availability of mobile number portability service, churners are increasing at a higher rate than before. Churn being a predictive model, there is no generalized scalable approach to capture the probable churners effectively in the telecom data.

Several methods and machine learning algorithms have been proposed for predicting churners in different domains [2,3,4]. Existing approach to churn prediction

S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 367–382, 2012.

pertains to attribute based analysis which has proven to be relatively time consuming because the process has to be rerun every time the dataset is fed or updated. Moreover the classification model proposed related to this has been proved to face issues with respect to skewness of the churn data [1]. The churn data tends to be imbalanced because the churners tend to be far less in number in the order of (2% - 5%) compared to the non-churners. Due to the existence of the class imbalance problem [5], the high accuracy value derived from a model in churn prediction analysis provides no useful result in real time. Also it has difficulty in parallelizing certain aspects of the traditional algorithms, poses difficulty in applying them over large telecom dataset. Another interesting aspect is that certain attribute based analysis was found to be specific to a particular dataset such as data from a developed country where in the same model failed miserably for the developing country [7].

In order to tackle the above core problems, this paper examines the close relationship between the graph parameter analysis and in understanding the churn behavior. In this study the telecom data is visualized in the form of graph and several graph parameters are inferred from the same. The graph parameters are computed from the vertex and edge pairs ((V (G), E (G)), visualized from the telecom dataset stored in a scalable graph DB framework. The graph DB falls in the class of NOSQL database technologies. The idea of using a NOSQL DB rather than the traditional relational DB is that the NOSQL data technologies supports scalable and schema less structure that helps in analyzing and storing huge datasets [10].

The graph parameters considered for node level analysis are as follows: In-Degree, Out-Degree, Closeness centrality, Call weight, Proximity prestige, Eccentricity centrality, Clustering coefficient, In Degree and Out degree prestige. In addition to this, Game theory approach using Shapley value is calculated to find influential members (most important members) in the network. The graph parameters chosen specifically indicate the active participation of a customer and thus aid in studying the churn behavior over a period of time. The existing call graph implementation for churn analysis related to telecom network behavior study was confined only to the degree module and participation coefficient in a network [2, 3]. It is difficult to predict the churners accurately using a confined set of parameters chosen arbitrarily. Thus we propose a novel idea to analyze graph parameters exhaustively during the training phase of the model, which can help in understanding the factors contributing to churn behavior.

Even though the considered graph parameters hold close relationship to the churn behavior analysis, evaluation of all the graph parameters over the huge dataset on a dynamic environment tends to be a costly process. Thus we need to run predictive machine learning methods like multivariate Discriminant and Regression Analysis that can aid in finding out specific graph parameters that contribute significantly to the discrimination of churn behavior. The corresponding analysis can help in bringing down the list of graph parameters for identifying the visible churners quickly.

Call Detail Record (CDRs) is generated for every transaction made in the telecom domain. With billions of CDR records to be processed, it's virtually impossible to manipulate the data over a single machine. Therefore a map reduce based parallelized framework using HADOOP architecture [11] is employed for pre-processing of CDRs over a cluster environment. The efficiency of combining the map reduce based framework with NOSQL based storage makes this innovative model work in a

scalable and a dynamic platform with ease and minimal cost. Eventually, the telecom service providers can use the proposed model in identifying churners efficiently on a streaming environment and in launching retention campaign based on their priorities.

1.1 Our Specific Contributions

- **Geo-spatial data processing on a distributed environment:** The huge CDR data set is pre-processed by splitting them based on specific locations. It is further processed by splitting them into periodic chunks for graph parameter analysis using Hadoop based Map-Reduce framework.
- **Usage of predictive machine learning models to identify specific graph parameters for churn behavior analysis:** We have written the code for predictive models such as: Multivariate Discriminant Analysis and Logistic Regression to extract specific graph parameters for churn analysis that can aid in reducing the computational cost and thus it improve the effectiveness of probable churner identification on a distributed environment.
- **Proposal of a final model that can work in a scalable and dynamic environment for churn prediction:** The proposed graph-oriented model which is a replacement to the traditional approaches can be easily extended to any other domains for probable churner prediction that can work in a scalable and dynamic environment.

2 Related Work

Prediction of probable churners was analyzed by different levels of various studies relevant to the domains such as Telecom service providers, Insurance companies [12], Pay - TV service providers [13], banking and other financial service companies [14], Internet service providers [15], newspapers and magazines [16]. Existing models for churn prediction pertains to supervised and semi supervised methods. Such methods have been designed using different data mining techniques [17]. For instance, the predictive performance of the Support Vector Machine method is benchmarked to Logistic Regression and Random Forest in a newspaper subscription context for constructing a churn model [16]. Another study dealt with the prediction model built for a European pay-tv company by using Markov chains and a Random Forest model benchmarked to a basic logistic model [13]. The general issue with these approaches is that the models don't scale well in a dynamic environment and they tend to be relatively time consuming. Few studies such as [18] employ more than one method based on cluster analysis and classification but they failed to present a standardized model for churn prediction that can be applied to a generalized telecom dataset. The other issue with respect to the classification models is that they face class imbalance problem due to skewedness in the telecom data [1]. Churners correspond to a minor class and therefore building the classification model would bias the model trained towards the majority or the non-churner class [5]. The similarity in these approaches is that the model measures the overall prediction accuracy instead of measuring the accuracy of

predicting churners separately. The accuracy can be improved by predicting the non-churners with a high degree of correctness. This is possible because the models trained will be good at predicting non-churners because of the relatively huge number of non-churner samples.

The usage of graph-based techniques for data analysis has been employed in social network analysis [19], analyzing the network structure in the telecom circle and World Wide Web Hyperlink graph analysis [8]. One of the first studies on graph for telecommunication was performed on a graph of land line phone calls made on a single day data. The generated graph consisted of approximately 53 million nodes and 170 million edges [20]. The graph inferred that most of nodes being pairs of telephones that called only each other. Most of the existing graph models are based on the node distributions [8]. Analysis based on confined set of parameters would lead to diminished results. The usage of graph as a means to analyze *big data* has been a successful entity in studying the usage of websites in the internet which employs the ingestion of massive data feed from World Wide Web [8]. Interesting analysis such as Page Rank for search results has been performed out of the same [9]. The telecom data holds close resemblance to the internet feed where in the real time data generated trails a power law graph and the size tends to be huge.

In this paper, we have considered exhaustive list of graph parameters that suits for node level churn analysis which includes centrality and prestige measures. Centrality is based on the choices or participation made by a user whereas prestige depends on the choices that a given user received from others. Churn is a specific business case wherein the telecom carrier would like to identify chunks of users who are likely to churn. We have analyzed the call graph properties specific to customer churn behavior on a telecom domain.

Two of the most widely used statistical methods for analyzing categorical outcome variables related to consumer behavior analytics are linear discriminant analysis and logistic regression [23]. The goal of Logistic Regression is to find the best fitting and most parsimonious model to describe the relationship between the outcome or response variable and a set of independent variables [23]. Linear discriminant analysis can be used to determine which variable discriminates between two or more classes, and to derive a classification model for predicting the group membership of new observations. These methods are more suitable to be employed for graph parameter analysis in our approach. Also one of them can be extended to generate a final linear model for predicting visible churners.

3 Graph Parameter Analysis

The structural properties of call graph are calculated for every node in the network which are analyzed over two time frames in the churn window namely: before the period of churning and during the period of churning. It is noted that the customer is likely to churn out after the second time frame. We try to understand the churn behavior pattern over different phases of the churn window and use the corresponding results for further analysis. The graph parameters chosen for the study depicts

different aspects of participation by a customer in the given network. The call graph G is generated by ingesting the CDRs to create ((V(G), E(G)) pairs, where V(G) represents the vertices in the call graph and E(G) represents the edge connecting two vertices. The edge weight represents the number of calls made in the CDR. Fig 1 illustrates the nodes with specific graph parameter measures. The filled vertex (in red color) signifies that the node has high influence in the network which is determined by the Shapley value metrics. The graph parameters considered for the node level analysis are described here.

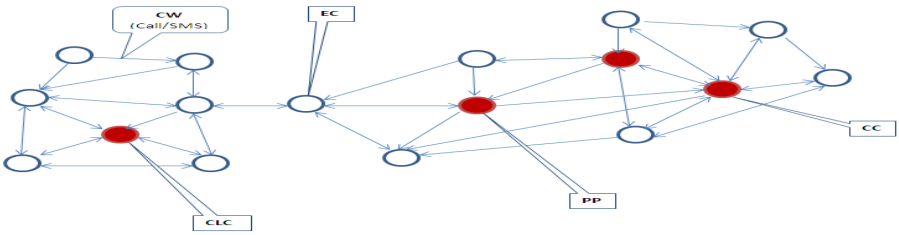

Fig. 1. Sample network with vertices with high graph measures are marked

In-Degree (ID): In-degree measures the number of incoming connections to a given user [19]. The incoming connections can represent the incoming calls or SMS received. To measure the in-degree of a given user (v_i), we count the number of unique users who have communicated to a given user in the network.

Out-Degree (OD): Out-degree measures the number of outgoing connections from a given user (v_i) [19]. The out-degree signifies the active participation of a customer in the network. We find the measure by counting the number of unique users that the user v_i has communicated to.

Closeness Centrality (CC): Closeness centrality measures the importance of a user in a network based on their location in the call graph [21]. A central user will tend to have a high closeness centrality; i.e. if a central user was thought of an information spreader, then rumors initiated by him will spread to the whole network quicker [21]. Let $d_{i,j}$ be the length of the shortest path between vertex v_i and other vertices v_j . Then the average distance between vertex v_i and all other vertices v_j which is given by:

$$l_i = \frac{1}{|V|} \sum_{j \in v} d_{i,j} \tag{1}$$

The closeness centrality is defined as the inverse of l_i.

$$cc_i = \frac{1}{l_i} \tag{2}$$

Degree Prestige (DP): It is based on the in-degree (ID) and the out-degree (OD) of a node in the graph, which takes into account the number of nodes that are adjacent to a particular node in the graph [19]. Prominent customers in the network can be found using this factor.

$$DP_i = \frac{f_i}{|V|-1} \tag{3}$$

where f_i- is the number of first level neighbors adjacent to node v_i.

Proximity Prestige (PP): Reflects how close all the nodes are present in the graph with respect to a given node x in the network [19]. It signifies the ease of reaching a specific customer in the network. If k_i be the number of nodes in the network who can reach member v_i then PP is given as

$$PP_i = \frac{\frac{k_i}{|V|-1}}{\frac{1}{k_i}\sum_{j=1, j\in V}^{k_i} d_{i,j}} \tag{4}$$

Eccentricity Centrality (EC): It states the most central node in the network [21]. The node with high EC value is the one that minimizes the maximum distance to any other node in the network. It signifies the closeness of the neighbor's to a given customer in the network.

$$EC(x) = \frac{1}{\max\{d_{i,j} : j \in V\}} \tag{5}$$

Clustering Coefficient (CLC): The clustering coefficient represents the density of community accruing from a given node n in the network [19]. When a customer from a highly clustered community is likely to churn then there is a possibility that he will induce other members in the community to churn as well. It is represented as:

$$CLC = \frac{|\text{Actual edges between neighbors' of n}|}{|\text{Possible edges between neighbors' of n}|} \tag{6}$$

Shapley Value (SV): The Shapley value represents the influential score for a given node in the network [22]. Influential nodes are the one who are not only active in participation but also holds strong influence among their neighboring nodes. The telecom carriers must target the influential churners with their retention scheme first to prevent them from becoming an influential churn spreader. It is represented as:

$$SV_i = \sum_{v_j \in v_i \cup N(v_i,d)} \frac{1}{1+\deg(v_j)} \tag{7}$$

where, $N(v_i, d)$ represents nodes with d degree of separation from node v_j.

There exists several other graph parameters for graph analysis but we have restricted our analysis to specific parameters that is applicable to node level analysis which have closed association to the events happening in telecom domain. The algorithmic representation of Shapley value is given in Section 4.3.

4 System Overview

The CDR data is visualized as a call graph which consists of vertices and edges based on the activities of individual customers in the network. Exhaustive graph parameter analysis is inferred from the ingested graph database (InfiniteGraph [25]). The visualization and the corresponding analysis are made over a period of time. Specific graph parameters are chosen by employing two different multivariate methods that contributes more to extract churn behavior. Finally, we arrive at a linear model with more specific graph parameters using logistic regression to be employed for probable churner prediction on a dynamic environment. The overall system throws light on a novel way of churn prediction with ease and minimal cost as illustrated in Fig 2. The detailed description of individual components of the system is discussed in the following sections.

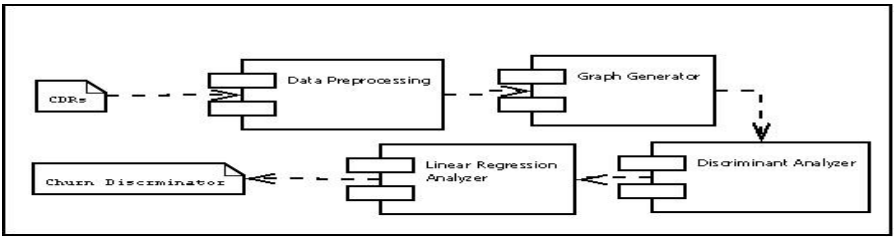

Fig. 2. Overall System Flow Diagram

4.1 Data Pre-processing

In a typical billing system of a mobile operator, for every operation performed by the customer varying from Voice, SMS usage to GPRS usage and each individual event is recorded and stored as Call Detail Records (CDRs).The dataset initially used for the model employs CDRs from leading telecom operators of a developing and a developed country respectively. The time span of the churn window is taken for a period of 3 months. One of the datasets is initially preprocessed to generate urban and rural region datasets to explore deeper analysis of telecom data related to churn problem. The CDR data is later processed by splitting them into weekly window chunks. We split the dataset to analyze the behavior of a customer over a period of time to comprehend the churn behavior. In earlier churn prediction models the attributes were aggregated over entire time period [22]. The disadvantages of other approaches is that in case of new user who has joined in later part of the month may be classified as churner due to his low usage and also it's difficult to generate a

predictable pattern from those data, as differences in usage pattern cannot be derived [22]. The dataset generated in the telecom industry tends to be of huge size and hence processing those takes a lot of computational time. Thus we employ Hadoop-based Map Reduce framework to preprocess the CDR data by converging them to location-wise details and use them for graph generation and parameter computations.

4.2 Data Ingestion and Graph Generation

The graph is generated using a distributed graph database implemented in java. It is from a class of NOSQL (or Not Only SQL) data technologies focused on graph data structure. Graph data typically consist of objects (nodes) and various relationships (edges) that may connect two or more nodes. The graph is generated by ingesting the CDRs to create vertex and edge pairs. The edge weight represents the number of calls made or SMS sent. The connections among the node are represented using directed edges. Location wise call graphs are generated for every split window of data and each graph is stored in a graph DB for further graph parameter analysis. The graph parameter chosen ranges from simple computations such as degree calculation to understand the incoming and outgoing activities from a customer to complex centrality evaluation to comprehend the closeness or a closed community formation. The detailed explanation on the graph parameters are already discussed in Section 3. We further discuss the influence exerted by a customer within a network based on the game theoretic centrality approach implemented using the Shapley value.

4.3 Game Theoretic Network Centrality: Using the Shapley Value

The game theoretic network centrality assists in finding out the importance of each node in terms of its utility when combined with the other nodes [22]. In telecom network it wouldn't suffice to find the importance of a node as a mere standalone entity as in other centrality measures. Other works related to the finding of influential nodes in the network [7] didn't address the influence of a node as a combination of several nodes in a network. Given a telecom network, the game theoretic network centrality indicates the *coalition value* of every combination of nodes in the network. We have introduced *Dijkstra's algorithm* to efficiently track the shortest distance between a given node and its neighbor's in calculating Shapley value for each node.

Program: Computing SV by running a game.
Input: Graph Network ingested from CDR.
Output: SVs of all nodes in network

```
foreach node v in Network do
   DistanceVector D = Dijkstra(v,Network);
   kNeighbours(v) = null;
   kDegrees(v) = 0;
   foreach node u<= v in Network do
      if D(u) <= k then
         kNeighbours(v).push(u);
         kDegrees(v)++;
      end
end
```

```
foreach node v in Network do
   ShapleyValue[v] = 1
      kDegrees(v)++;
      foreach node u in kNeighbours(v) do
         ShapleyValue[v]+= 1
         kDegrees(u)++;
      end
   end
   return ShapleyValue;
end
```

4.4 Linear Discriminant Analysis

Linear discriminant analysis is used to determine which attribute discriminates between two or more naturally occurring groups [23]. The linear step-wise discriminant model is used to extract graph parameters that contribute more to the churn behavior. A discriminant function that is a linear combination of the components of graph parameters x can be written as:

$$g(x) = w^T x + w_0 \tag{8}$$

where w is the weight vector for different graph parameters and w_0 is the threshold weight. In our analysis we define two groups namely: Probable churners and non-churners. In case of the LDA, a linear discriminant function is of the form

$$y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + + a_n x_n \tag{9}$$

with x_i being the graph parameter derived from the CDR data set. The parameters a_i has to be determined in such a way that the discrimination between the groups is at its best.

4.5 Modified Logistic Regression Model

The usage of logistic regression model in our study is applicable as a two folds process: First, the logistic regression aids in listing out the graph parameters based on their significance towards the contribution of churn behavior. Next, we derive a linear model using the selected list of graph parameters for the churn analysis. The linear model derived using the logistic regression y is usually defined as similar to Eqn. (9).

The intercept is a constant value of y, when the value of all graph parameters is taken to be zero. Each of the regression coefficients describes the size of contribution of the graph parameter towards the churn behavior. The logistic regression is a useful way of describing the relationship between the extracted graph parameters and for predicting the churners. Using the linear model we derive a logistic function which takes on values between zero and one [23]:

$$f(y) = \frac{e^y}{e^y + 1} = \frac{1}{1 + e^{-y}} \tag{10}$$

The input is y and the output is $f(y)$. The logistic function is useful because it can take a graph parameter input value ranging from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1 which indicates the probability to churn or not .

5 Analysis and Results

5.1 Datasets

The graph parameters are examined over three different datasets obtained from the leading telecom service providers of two different countries. The first dataset corresponds to a rural base whereas the second one corresponds to an urban region of a particular country and the third set corresponds to a data from a developed country. The idea of using three different datasets helps to make the model more generic and a standard one. The dataset spans to a time period of three months record. Hence during this period, who ever disconnected from the service is considered as churners. The results given in this section are related to third set for example. Since the evaluation presented in section 6 not shows much difference on the accuracy of all three models generated.

5.2 Weekly Analysis of Graph Data

Fig 3 illustrates the windowing frame used to analyze the churn behavior over a period of time. The windowing frame currently shows a spread of 6 week splits.

- **First time frame:** Time frame before the period of churning.
- **Second time frame:** Time frame during the period of churning.
- **Churn Window:** The customer is likely to churn out after the second time frame.
- **Churn Slider:** The slider moves over the windowing frame over different period of time. The size of the churn slider is a period of one week. It captures the variation in graph parameters. Significant variation in the graph parameters shows the presence of churn behavior.

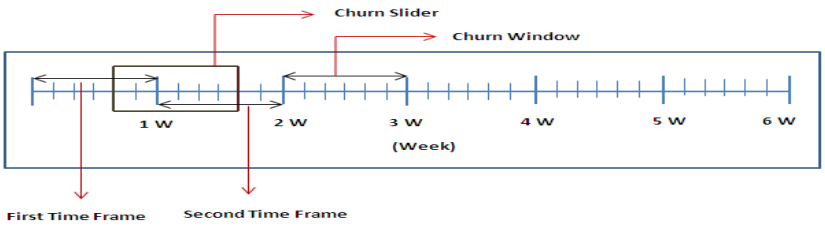

Fig. 3. Analyzing the churn behavior pattern over a period of time using the Windowing frame

The windowing frame is initially split into 3 weeks. The probable churner exists in the first two weeks and churns out in the 3rd week or in churn window. Once the slider crosses the first churn window, the three frames present in the windowing frame progresses by one frame in the forward direction leaving the first week behind. As the slider progresses in the windowing frame, graph parameter variation is captured from one time frame to the other and radical changes in graph parameters are marked as probable churners in the churn window.

5.3 Selection of Best Graph Parameters for Churn Prediction

The idea of selecting specific graph parameters contributing effectively to the churn behavior rather than considering all the parameters can be achieved by running the machine learning algorithms such as Logistic regression and Multivariate Discriminant analysis. These machine learning approaches are used to analyze the graph parameters over the two time frames to highlight a specific list of graph parameters that contribute significantly for discriminating churners from non-churners. The analysis results using the logistic regression and multivariate discriminant analysis are described here.

5.3.1 Implementation of Predictive Data Mining Model: Logistic Regression Model and Multivariate Discriminant Analysis

Logistic regression and Multivariate Discriminant analysis are run over the graph parameters calculated for the different datasets as a training model. Discriminant analysis is used to find out the canonical weighted score for the graph parameters that contribute to the churn behavior. Based on the step wise analysis the graph shown in Fig 4 shows the significant contribution of few graph parameters for churn behavior.

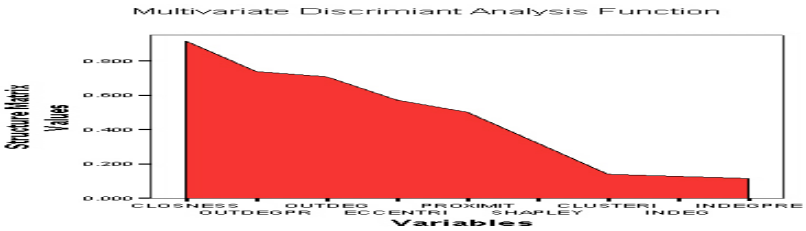

Fig. 4. Multivariate Discriminant analysis Function for graph parameters

The Logistic regression model is first run to calculate the W*ald statistics measure* with its significance for the best contributing graph parameters as illustrated in Table 1. The Wald statistic for a coefficient is the square of the result of dividing the coefficient by its standard error. The logistic regression is later used to derive a linear model for predicting churners for a specific threshold level (0.60) as elaborated in Section 6.

Table 1. Wald Statistics Scores for top four parameters

Variables	Wald	Sig
OD	9.246	0.002
SV	8.354	0.005
CC	23.550	0.000
PP	39.703	0.000

Fig 5 illustrates the discrimination of churners and non-churners based on the selective graph parameters derived from the multivariate discriminant and logistic regression models. The variation is studied by analyzing the Out degree, Shapley value, closeness centrality and proximity prestige before the period of churning. The discriminating functions based on the selected variables clearly segregate the churners and non-churners separately.

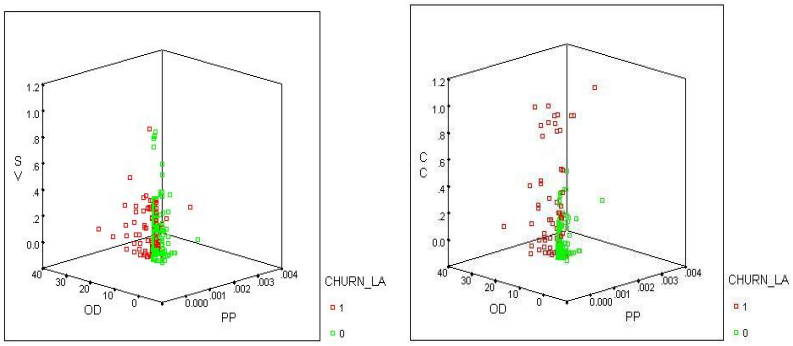

Fig. 5. Discrimination between the churners and non-churners is clearly visible based on selected graph parameters

5.4 Statistical Testing: Using T Test

In addition to Multivariate model, we have used paired t statistics [24] to measure the mean differences of graph parameters between two different time windows. The customer churns out in the third slot or window are considered as real churners. The variation in graph parameter is carefully analyzed over the period of time. The mean value is calculated to give an accumulated score for the graph parameters in terms of churners and non-churners separately. The standard deviation finds the variation in graph parameter from the mean score. The mean and the standard deviation value for the first time frame are compared with the second one for the churners and non-churners respectively. The gradual reduction in specific graph parameters shows that the churners are slowly losing interest in using the corresponding network. This is verified with paired t statistics.

In Table 2, the values given in the bracket denotes that the analysis is carried out during the second time frame in the windowing frame. We find that there is a significant variation for certain graph parameters as in the case of churners. The variation for the non-churners is minimal which proves that the usage by non-churners in the network is almost constant. Based on the t statistics we found that the out degree, closeness, proximity, eccentricity and shapely value has showed relatively high significance for probable churner identification. The drop in out degree of churners clearly illustrates their intention and changes in other important parameters depict the value of their presence in the network. Losing some of the influential users will create rickety in the present network. This result cross verifies the predictive model outputs.

Table 2. Mean and Standard deviation scores for churners and non churners

	CHURNERS		NON-CHURNERS	
	MEAN	**STD. DEV**	**MEAN**	**STD. DEV**
ID	0.70000(0.682)	1.10000(0.95)	1.00000(1.00200)	1.60000(1.65500)
OD	**2.26848(1.8571)**	**3.33069(2.7708)**	1.00000(0.99195)	3.00000(2.98535)
IDP	0.00001(0.00001)	0.00002(0.00001)	0.00002(0.00002)	0.00002(0.00002)
ODP	0.00004(0.00003)	0.00006(0.00006)	0.00002(0.00002)	0.00005(0.00004)
CC	**0.68345(0.74389)**	**0.39080(0.37275)**	0.87667(0.87579)	0.29137(0.29301)
PP	**0.00006(0.00003)**	**0.00018(0.00010)**	0.00002(0.00002)	0.00011(0.00010)
EC	**0.91679(0.93018)**	**0.21142(0.19249)**	0.96912(0.96784)	0.13192(0.13512)
CLC	0.00863(.00865)	0.06270(0.06724)	0.00211(0.00222)	0.02654(0.02797)
SV	**0.30590(0.32772)**	**0.21750(0.22551)**	0.45759(0.46464)	0.21243(0.22581)

6 Evolving a New Methodology for Churn Prediction

The Venn diagram representation given in the Fig 6 illustrates the effectiveness of graph parameters contributing for the churn behavior over a period of time. We found that the graph parameters contributing commonly from all the three datasets mention in Section 5, based on the analysis inferred from the predictive machine learning models and statistical models are: Out Degree, Shapley Value, Proximity Prestige, and Closeness Centrality. We propose that when there is significant variation in the above parameters as a whole then there is a high probability that the customer will churn out. On the other way, the considered parameters will run much faster to generate model and even it can process the data on a streaming environment.

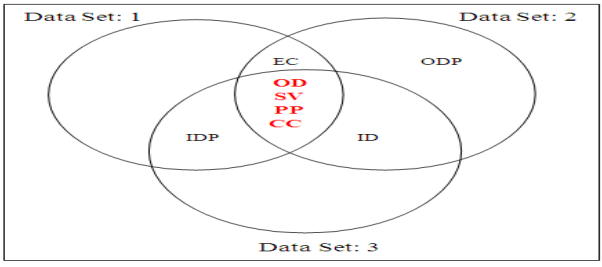

Fig. 6. Significant graph parameters contribution to churn behavior in various datasets

The proposed linear model is used for predicting the probable churners in dynamic environment. We compare the results of predicted churners with the actual churners in the test datasets to find the accuracy of the proposed model. The model was tested for three different datasets over a period of three month time scale. The average accuracy

for churn prediction using the proposed model was found to be 81.67 % as shown in Fig 7. The maximum accuracy reached in our previous model using hybrid learning is 72.18% for the same dataset used in this study [22]. The improvements in results highlight the significance of the proposed graph-based model for the churn prediction on mobile telecom networks. Also we have seen the decent improvement in F-measure value from 0.46 to 0.55.

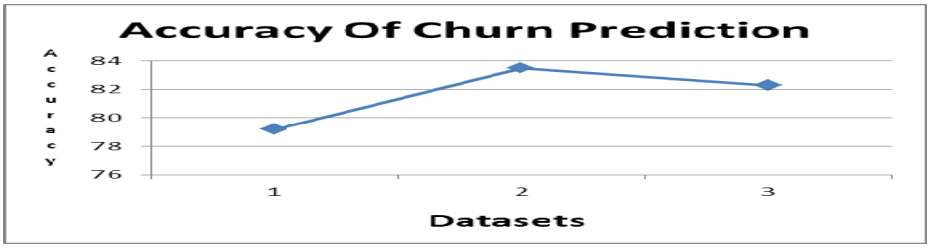

Fig. 7. Testing the accuracy of the churn prediction model

7 Conclusion

The Graph-based analysis for churn prediction is a novel idea proposed for efficient churn prediction in the telecom domain. The graph-based visualization aids in better understanding of the behavior of the customers. Also the simplicity in graph traversal aids in quick manipulation of the graph parameters. The machine learning methods and statistical test chosen for analysis have effectively worked out in choosing specific graph parameters contributing to churn behavior analysis. This approach has helped to make the study cost effective with respect to time and computation. The model also works on a distributive, parallel and scalable platform. The proposed model has overcome the class imbalance problem with graph-based solution which is prevalent in the other existing models. The new model is tuned to work on a generalized dataset. It is framed by collecting the graph parameters over a period of time, analyzing the variation in the parameters over the time period of churning and finally extracting the graph parameters that contribute more to churn behavior. Thus the new model proposed has opened a new way for finding visible churners with ease and cost effectively on a streaming environment.

References

1. Hung, S.-Y., Yen, D.C., Wang, H.-Y.: Applying data mining to telecom churn management. Expert System Applications 31(3), 515–524 (2006)
2. Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A.A., Joshi, A.: Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th International Conference on Extending Database Technology, EDBT 2008, New York, USA, pp. 668–677 (2008)

3. Lazarov, V., Capota, M.: Churn Prediction in the Business Analytics Course. TUM Computer Science (2007)
4. Lu, J.: Predicting Customer Churn in the Telecommunications Industry – An Application of Survival Analysis Modeling Using SAS. In: Berry, M.J.A. (ed.) Data Mining Techniques (2004)
5. Burez, J., Vandenpoel, D.: Handling class imbalance in customer churn prediction. Expert Systems with Applications 36(3), 4626–4636 (2009)
6. Kurucz, M., Benczúr, A., Csalogány, K., Lukács, L.: Spectral Clustering in Telephone Call Graph. In: Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007 (WebKDD/SNA.KDD 2007), San Jose, California, USA (2007)
7. Yeshwanth, V., Saravanan, M.: Churn Analysis in Mobile Telecom Data using Hybrid Paradigms. In: Second Conference on the Analysis of Mobile Phone Datasets and Networks, NetMob 2011. MIT, Cambridge (2011)
8. Broder, A.Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.L.: Graph structure in the web. The International Journal of Computer and Telecommunications Networking 33(1-6), 309–320 (2000)
9. Wicks, J., Greenwald, A.R.: Parallelizing the Computation of PageRank. In: Bonato, A., Chung, F.R.K. (eds.) WAW 2007. LNCS, vol. 4863, pp. 202–208. Springer, Heidelberg (2007)
10. DeCandia, G., Hastorun, D., Jampani, Kakulapati, M., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., Vogels, W.: Dynamo: Amazon's highly available key- value store. In: Proceedings of Twenty-First ACM SIGOPS Symposium on Operating Systems Principles (SOSP 2007), pp. 205–220. ACM, New York (2007)
11. Apache Hadoop, `http://hadoop.apache.org`
12. Morik, K., Kopcke, H.: Analysing customer churn in insurance data a case study. In: Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, New York, USA, pp. 325–336 (2004)
13. Burez, J., Van den Poel, D.: CRM at a Pay –TV Company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications 32(2), 277–288 (2007)
14. Halling, M., Hayden, E.: Bank failure prediction: A two-step survival time approach (2006), `http://ssrn.com/abstract=904255`
15. Khan, A.A., Jamwal, S., Sepehri, M.M.: Applying Data Mining to Customer Churn Prediction in an Internet Service Provider. In: IACSIT Hong Kong Conferences, IPCSIT, vol. 30. IACSIT Press, Singapore (2012)
16. Coussement, K., Van den Poel, D.: Churn prediction in subscription services: An application of sup-port vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34(1), 313–327 (2008)
17. Wei, C.-P., Chiu, I.-T.: Turning telecommunications call details to churn prediction: a data mining approach. Expert Systems with Applications 23(2), 103–112 (2002)
18. Tsai, C.-F., Lu, Y.-H.: Customer churn prediction by hybrid neural networks. Expert Syst. Appl. 36, 12547–12553 (2009)
19. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, New York (1994)
20. Abello, J., Pardalos, P., Resende, M.: Maximum clique problems in very large graphs. DIMACS Series, vol. 50, pp. 119–130. American Mathematical Society (1999)
21. Karnstedt, M., Rowe, M., Chan, J., Alani, H., Hayes, C.: The Effect of User Features on Churn in Social Networks. In: WebSci 2011, Koblenz, Germany (2011)

22. Yeshwanth, V., Vimal Raj, A., Saravanan, M.: Evolutionary Churn Prediction in Mobile Networks using Hybrid Learning. In: Proceedings of 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24), Palm Beach, Florida, USA (2011)
23. Pohar, M., Blas, M., Turk, S.: Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodološki zvezki 1(1), 143–161 (2004)
24. Donal Zimmerman, W.: A note on interpretation of the Paired-Samples t Test. Journal of Educational and Behavioral Statistics 22(3), 349–360 (1997)
25. InfiniteGraph, http://www.infinitegraph.org