Customer Churn Prediction, Segmentation and Fraud Detection in Telecommunication Industry

Conference Paper · December 2016

CITATIONS
5

READS
3,548

2 authors:

Ahsan Rehman
IBM
4 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Abbas Raza Ali
Bournemouth University
13 PUBLICATIONS **87** CITATIONS
SEE PROFILE

Customer Churn Prediction, Segmentation and Fraud Detection in Telecommunication Industry

Ahsan Rehman[1], Abbas Raza Ali[2]
Advanced Analytics and Big Data [1], Advanced Analytics and Big Data[2]
IBM Pakistan[1], IBM United Kingdom[2]
s.ahsanrehman@gmail.com[1], abbas.raza.ali@gmail.com[2]

Abstract

Every telecommunication market player is launching innovative business models and offering better services which in turn have increased the cost of retaining customers. Realizing the importance of customer retention, service providers are now putting more efforts in prediction and prevention of customer churn. This work being divided into two phases, with phase one presenting some of the commonly used data-mining techniques for the identification of churners and customer segmentation based on various Key Performance Indicators (KPIs). While phase two phase discusses how social network analysis extracts relationships between different subscribers to improve the results produced by the traditional learning algorithms at individual subscriber level. Satisfactory results have been achieved from Base model which were significantly enhanced by applying the social network analysis technique.

Keywords: Call Detail Record; Churn Prediction; Customer Segmentation; Fraud Detection; Social Network Analytics.

1. Introduction

Over the last few decades, mobile telecommunication has emerged as the dominant medium of communication across the world. In several countries, market saturation has reached a level where every potential customer has to be won over from competitors. At the same time, standardization of mobile infrastructure and public regulation allowed the customer to port easily from one network to another, resulting in a fluid market. Since the cost of winning a new customer is greater than the cost of retaining an existing one [1], mobile carriers have been shifting their attention from customer acquisition to customer retention. As a result churn prediction and prevention have become one of the most crucial Business Analytic applications aimed at identifying potential customers who are about to transfer their business to a competitor (i.e. churn) [2].

A good churn prediction system should not only identify the potential churners successfully, but also provide reasons for their churn and forecast such results for a sufficiently long horizon, i.e. six months. Once potential customers are identified as likely to churn, the marketing and retention department tries to retain their business by luring them with attractive and well-designed campaigns. Thus, a long forecast horizon is an obvious advantage because the further away the customer is from actually making the decision to churn, the easier it is to prevent the decision at a significantly lower cost.

Since the retention efforts are constrained due to limited resources a fraction of the subscriber-base can be contacted at any given time. With this constraint in mind, churn prediction models are usually measured by their ability to identify actual churners amongst the top 0.1% to 5% of the customers predicted that have the highest risk of churning.

The telecom operators wanted to divide the existing revenue generating subscriber-base into multiple segments with distinct usage patterns. The base needs to be grouped efficiently using clustering models. Each segment needs to have its own mean values [3]. This allows the service providers to target customers in particular segments with better offers, thereby resulting in improved customer satisfaction and revenue generation.

The ever increasing competition for a larger customer base means setting tougher sales targets for which multiple steps are taken to promote new activations. Commission on new subscriber activation is one major incentive which has given rise to fraudulent sales. Such sales involve no activity and therefore need to be identified and separated from normal churn scoring process.

The traditional learning algorithms process each subscriber as an independent entity. This results in hiding the inherent relationship between subscribers. A subscribers Call Detail Record (CDR) contains various information pertaining to each call, e.g., caller, called, timestamp of the call etc. Based on this information a call graph, known as Social Network Analysis (SNA), can be constructed with customer mobile numbers as identity of nodes and the calls as edges. The weight of an edge captures the strength of the relationship between two nodes, e.g. call duration, frequency, etc. [12]. In this study, SNA has been used to model this relationship of the subscribers and embed it on top of the traditional base-model in order to enhance overall accuracy.

The subsequent sections of the paper are organized as follows. Section 2 briefly discusses existing churn prediction and customer segmentation systems. Section 3 describes business challenges faced during this research. Section 4 gives an overview of gathering and preparing data from data-sources. Section 5 presents detailed analysis performed on input data to make it compatible for modeling, with some preliminary analysis that led to final system methodology, which is discussed in Section 6, supported by results that are recorded in Section 7. Finally, the paper is concluded in Section 8 followed by possible directions for the future in Section 9.

2. Literature Review

This section reviews different research approaches related to churn prediction and customer segmentation with a focus on the kind of techniques used to solve business problems pertinent to Telecom companies worldwide.

Gopal et al [5] has used ordinal regression for customer churn time prediction for modeling tenure of mobile customers. The dataset used in this work consisted of 100,000 customers with 169 independent features. At minimum 7 months old customers were included in the sample and the study period consisted of 25 months, with a total of 5 respective ranks. Top 50 features were picked for modeling, with dataset being randomly partitioned into two sets. Ordinal regression produced 86.21% accuracy while multi-class

 ISBN: 978-1-62561-003-4

classification gave an accuracy of 83.8%. Later it was compared with Survival Analysis technique, a state of the art method for tenure modeling which was able to capture 20% of churners in the top deciles (7 to 11 months period) while ordinal regression was able to capture 20-45% of churners in the top deciles (25 months period).

Another research to predict customer churn, [8] applied Survival Analysis techniques to predict customer churn and compared it with the existing conventional statistical methods. The technique is aimed to help telecom service providers understand customer churn risk and its hazard in a timely manner. The objectives of this study were two-fold: 1) to estimate customer survival and customer hazard function in order to gain knowledge of customer churn and 2) to demonstrate how Survival Analysis techniques can be used to identify high risk customer and their expected time of churn. The dataset used consisted of about 40,000 active high-value customers, who were randomly selected from the entire customer-base over a period of 15 months. Final results were quite promising where 90% churners were correctly identified.

Another study for predicting customer churn in mobile networks, [11] used SNA technique called 'Group First Churn Prediction.' This technique exploited the structure of customer interactions to predict which groups of subscribers are most prone to churn. Further, it used second order social metrics to analyse interactions within each group that helped in identifying social leaders. Two datasets from different service providers were used; one dataset used 7 days call data consisting of 16 million instances while the second dataset used 28 days call data consisting of CDRs of 26 million subscribers. The results were integrated to a prior churn score, which significantly improved the operator's previous results.

In another study, social ties and their relevance to churn in mobile telecom networks, [12] used simple diffusion-process which exploits social influences affecting churn. Their diffusion model was based on Spreading Activation (SPA) which examines millions of mobile phone users' patterns. This technique predicts potential churners by examining the current set of churners and their underlying social network. The dataset used in this work consisted of aggregated call usage, frequency and duration of each user over the span of one month. After performing several experiments the best predictor was obtained by using spreading factor of 0.72 and SPA was successful in making correct predictions about 50-60% of future churners.

In the domain of Customer segmentation for telecom companies, a study was based on service usage behavior [3] which discussed various clustering techniques to understand customer behavior. Customers were clustered using k-Means based on their CDR and categorized in four loyalty groups. By focusing on call duration of 6 months data with the value of *k* as 5, it labeled each cluster according to its distinct characteristic. The results defined a threshold for loyal and revenue generating customers in terms of active time on network, where 14% churners were identified.

3. Business Challenges

During the last decade, a rapid growth in the telecom industry has led to a bigger subscriber-base for service providers and handheld phones are becoming the dominant communication medium world-wide. The telecom operator discussed in this study has a subscriber-base closer to 20 million. The operator maintains a Data-Warehouse (DWH) which consolidates the appropriate transactional and customer data records and is planning to use it for various analyses. The key challenge is to build a data-mining application for churn prediction that can reduce gross churn rate by 5% for both post-paid and pre-paid subscribers.

Another major challenge stems from the current prepaid subscriber-base, compromised of different segments. The customers have been consuming different packages and bundles according to their usage. These packages offer unlimited calls, messages and Value Added Service (VAS) hence the subscriber-base was quite stretched with user's average spending ranging from $1 to $12 per month. The operator wanted to identify distinct segments according to subscribers behavioral and usage attributes.

The operator started off its operations with a subscriber-base of less than a million and planned to capture a substantial market share in the first few years. That target was achieved by setting aggressive sales targets and offering lucrative incentives to the sales and distribution channels. Consequently, while the subscriptions were rapidly increasing, a new problem emerged known as "fraudulent sales." The fraudulent sales phenomenon is described as the consumption of free credit which comes on new subscriber activation and is valid for a few days, but is never recharged afterwards. Hence no revenue is generated from the particular new sales activation; on the other hand the sales channels earn lucrative commissions on it. Currently the operator plans to build fraudulent sales prediction model which can predict subscribers who will not be generating revenue after the initial activation period and also identifying sales channels that are involved in massive fraudulent sales.

4. Data-Sources

A variety of data-sources are required to build churn prediction and customer segmentation models including customer demographic, geographic information and call detail records, etc. In the context of current work, very limited customer demographic and geographic information was available, as most of data-sources comprise of basic information i.e., CDR, operations, etc.

4.1 Data Gathering

For this research the training dataset was gathered from three main data-sources namely, 1) customer call detail record, 2) call center contact details and 3) customer personal information. Table 1 shows detailed list of features from these data-sources.

Most of the features are continuous, except a few, i.e. package plan, city, region, franchise, etc. The granularity of source data is at customer level with daily transaction details. According to the retention policy of the telecom operator, a total of 6 months prepaid and 12 months postpaid historical transactional data is being maintained. The total prepaid base is around 20 million, postpaid is around 0.1 million and on daily basis around 20,000 new subscribers join the network. From the 20,000 new subscribers, around 25% utilizes the initial free balance offered with new activation for the first 7 days and then never recharge, hence churn.

Table 1: Data-sources and their features

	Features
	Customer call detail record
1.	Outgoing and incoming call usage, minutes, and revenue
2.	Outgoing and incoming short message service (SMS) usage and revenue
3.	VAS details i.e. internet (GPRS), interactive voice response (IVR), ring-back tones, Multimedia Message Service (MMS) etc. usage and revenue
4.	Recharge count and amount
5.	Current package plan details
	Call center contact details
6.	Customers call center contact details which include call duration and

	count.
Customer personal information	
7.	Customer location including city and region
8.	Customer purchase details including franchise and default package
9.	Customer gender

4.2 Data Preparation

Data preparation is an important process dealing with a number of key issues related to the organizational data environment such as data quality, availability, loading etc. Given the importance of such issues a structured approach of the data preparation process was adapted in this work.

For each predictive model, one single structured table was prepared for building, assessing and scoring models in a real-time environment. The existing features were later used to build various derived features, which were able to give a better meaning to the customer behavior. The following derived features added a whole new dimension for predictive modeling:

1) Active days: total revenue generating days
2) Average call distance: time between two consecutive calls
3) Maximum inactive days: consecutive number of inactivity days
4) Network Age: scoring date - activation date

Initially, the data for the prepaid and post-paid models was limited to 6 months, but later it was observed that post-paid had a very small base (0.1 million). However, the predictive modeling demands more records to prepare a comprehensive rule-set. So, 12 months historic data was gathered for post-paid. The prepaid model already had sufficient records (20 million), therefore, just six months data was used. While for early churn model where data was limited to seven days, predictive modeling was a challenge. Keeping in mind the DWH retention policy, a maximum of 7 months historic data was extracted and later used for early churn modeling.

5. Data Preprocessing and Analysis

This section discusses transformations that have been applied to make input data compatible for predictive modeling. The data was filtered to predict accurate results, as the raw data usually includes instances which are loosely controlled and can have out-of-range and missing values. Therefore the representation and quality of data was enhanced before further analysis. Data pre-processing involves a number of steps, which can take a considerable amount of time and effort. These steps are explained as follows:

5.1 Data Audit

This process helps in analyzing the quality of dataset features. It provides a comprehensive first look at the raw data which is often helpful for the initial data exploration. The descriptive statistics of three datasets are listed in Table 2. This table lists the key features of each dataset with respect to four statistical measures.

Mean indicates the average distribution for each feature while Standard Deviation gives a precise measure of the spread of data. Skewness indicates the degree to which a features distribution departs from symmetry about its mean value and outlier percentage indicates if features have any extreme values. An initial analysis of each dataset has been driven after this step (Bold items in Table 2).

Table 2: Data Audit

Features	Mean	Std. Dev.	Skewness	Outlier (%)
Early Churn				
Total incoming minutes	**54.7**	**181.09**	7.28	1.24
Total incoming revenue	2.67	12.94	**39.56**	0.45
Total outgoing minutes	**46.64**	**135.59**	7.31	1.34
Total outgoing revenue	34.74	25.22	-0.21	0
Total VAS revenue	8.24	14.03	2.30	3.09
Active days	3.80	2.46	**-0.28**	0
Average call distance	0.72	0.51	0.57	1.01
Total inactive days	1.06	1.22	1.36	2.02
Post-paid				
Active days	201.82	135.26	-0.03	0.00
Average call distance	1.27	9.12	23.57	0.31
Maximum inactive days	19.88	45.36	3.63	2.17
Network age days	**486.83**	544.17	1.91	2.73
Outgoing calls minutes avg.	13.19	24.09	7.65	1.27
Outgoing calls revenue avg.	16.26	32.93	37.43	0.57
VAS revenue average	2.72	17.61	**110.82**	0.27
Revenue line-rent average	5.74	10.41	3.36	**3.08**
Payment amount average	23.44	44.48	31.46	0.65
Prepaid				
Active days	88.37	61.33	0.20	0.00
Average call distance	2.89	10.33	9.92	0.75
Total inactive days	24.32	33.99	2.19	2.89
Network age days	364.77	276.99	0.77	0.00
Outgoing calls minutes avg.	4.08	6.21	8.70	1.14
Outgoing calls revenue avg.	10.34	22.29	5.61	1.58
VAS revenue average	1.44	2.92	**25.20**	0.79
Recharge value average	9.61	11.94	11.41	0.92
Total count of recharge	**0.05**	0.23	4.04	**5.35**

5.2. Sparseness Elimination

The initial data-audit phase (see Table 2) highlighted that the datasets had lots of missing values. The key features of prepaid and post-paid included outgoing calls minute average, outgoing calls revenue average, VAS revenue average, recharge value average and total count of recharge. While for the early churn dataset all key features had missing values.

After carefully analyzing the DWH business rules the sparseness was eliminated from the input datasets by replacing missing values with zeroes. This balanced out the odd effect and gave a smooth transitional pattern. Sparseness eliminated data also provided key insights about the early churn model, which will be discussed in the later sections of this paper.

5.3. Outlier Detection and Fixation

Another observation from the initial data-audit phase was the percentage of outliers against every feature needed to be carefully handled. This process identified top 1% records having outliers or extreme values and could in turn affect the overall accuracy of the predictive model. These records (top 1%) were strictly adhered to avoid minimum number being excluded from the scoring process.

Rank based anomaly detection algorithm [16] was used to search for unusual cases based on deviations from the norms of their cluster groups. It is designed to quickly detect unusual cases for data-auditing purposes in the exploratory data analysis step, prior to any inferential data analysis. This algorithm can be divided into three stages:

1) Modeling based on the similarities of input feature set (shown in Table 2), cases are placed into cluster groups. The cluster groups are then identified using clustering model (Groups for datasets in Table 3) while sufficient statistics are used to calculate the norms of the cluster groups

2) Scoring model is applied to each case to identify its cluster groups and some indices are created for each case to measure the unusualness of the case with respect to its cluster group. All cases are sorted by the values of the anomaly indices and the top portion of the case list is identified as the set of anomalies

3) Reasoning for each anomalous case where the variables are sorted by its corresponding variable deviation indices. The top variables, their values and the corresponding norm values are presented as the reasons why a case was identified as an anomaly.

Table 3 lists three groups for post-paid and early churn datasets while prepaid dataset has 10 groups which indicate subscriber segments with distinct attribute properties. The anomaly detection process generated a model which was later used to detect anomalous records in scoring data based on patterns found in the original training data.

Table 3: Outliers Grouping

Groups	Early Churn	Prepaid	Postpaid
1	834,288	4,128,817	20,863
2	1,165,790	1,765,918	11,654
3	1,785,386	1,545,174	6,059
4		1,572,276	
5		2,740,655	
6		1,226,066	
7		1,060,808	
8		1,732,170	
9		1,718,279	
10		1,113,252	

5.4. Correlation Analysis

During the initial data preparation, a total of 125 features were loaded for data pre-processing. After initial data audit (Table 2), sparseness elimination and outlier's removal (Table 3) the next step was feature selection, i.e. use correlation to identify the top features having a strong relationship with the target variable (churn status).

The correlation measures the strength of relationship between target and independent fields with values ranging between –1.0 and 1.0. Values close to +1.0 indicate a strong positive association so that attributes have high relationship with target. Values close to –1.0 indicate a strong negative association so that high values for features are associated with low values in target and vice versa. Values close to zero indicate a weak association so the values for the two fields are more or less independent. In Table 4 the correlation strength for labels has been computed by importance (1 - p), where *p* is defined as significance/probability that the difference in means could be explained by chance alone.

Table 4 shows top six attributes for the three datasets in which most of them have a strong relationship with the target. Active day was a key feature which proved vital for all three churn prediction models. After correlation analysis we were able to deduce top 25 features which were later used in churn prediction modeling.

Table 4: Correlation Analysis

Features	Early Churn	Prepaid	Postpaid
Active days	0.50 (Strong)	-0.13 (Strong)	-0.19 (Strong)
Total incoming minutes	0.14 (Strong)	-0.01 (Strong)	-0.04 (Strong)
Total incoming revenue	0.2 (Strong)	0.02 (Strong)	-0.013 (Strong)
Total outgoing minutes	0.42 (Strong)	0.07 (Strong)	-0.009 (Medium)
Total outgoing revenue	0.24 (Strong)	0.07 (Strong)	0.01 (Strong)
Total VAS revenue	0.35 (Strong)	-0.002 (Weak)	-0.07 (Strong)

5.5. Data Balancing

After applying initial pre-processing steps, the dataset still had one major problem of 'imbalance classes' which imposed another level of difficulty for data-mining algorithms to extract meaningful patterns from the data. In case of imbalanced classes in the dataset, the ratio of sizes of output categories becomes biased to the extent that the learning algorithm only predicts the majority class in results [9]. For example, the post-paid dataset consists of 31,964 (94%) active subscribers, whereas there were only 2,203 (6%) churners, which presents a typical case of imbalance classes.

One of the methods to deal with this problem is re-sampling which can be applied in two ways: 1) under-sampling [7] and 2) over-sampling [6]. In case of pre-paid and post-paid datasets, a random selection of entries from the active customers (around 10%) were removed to balance out the ratio of subscribers who would churn versus the subscribers who would stay active to 16:84 respectively. This is a case of random under-sampling. On the contrary, over-sampling increases the strength of minority class by replicating a random selection of this class, which can result in over-fitting.

5.6. Data Quality and Analysis

This section covers the detailed analysis of pre-processed datasets of each model separately and highlights useful insights.

Early Churn/Fraud Detection: After the pre-processing of the early churn historical data, it was deduced that a total of 670,000 (18%) subscribers out of 3,700,000 had zero activity during the first 7 days of subscription. These zero activity subscribers were already part of the subscriber-base but were distinctively identified after preprocessing of the data. Figure 1 shows the users with zero active days which were discovered after sparseness elimination.

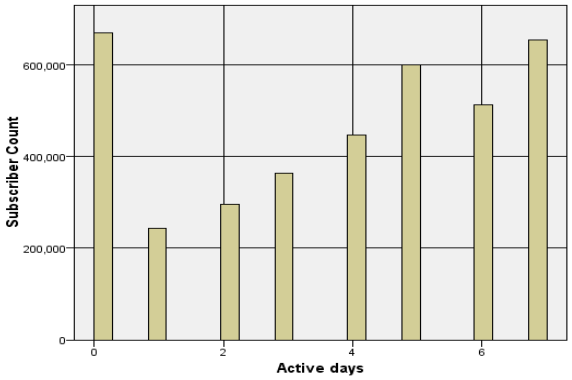

Figure 1: Pre-processed early churn data

The above finding was actually an enhancement to the scoring model where all the false activations being made by distributors to earn commission based on total activations were eliminated. These subscribers were producing noise in the data which were later removed from the early churn prediction model. Secondly, these subscribers were tagged to franchise for further action, i.e. eliminate such sales.

Post-paid Churn: The pre-processed post-paid dataset highlighted certain insights, which required necessary action. It can be observed from Figure 2 that most of the spikes are where subscribers have high outgoing call revenue with high total recharge and network age days being very low. These subscribers, not being part of the major post-paid subscribers, could badly influence the modeling process. Further analysis concluded that a total of 36 high usage and revenue generating numbers were present in the data which were non-churn corporate numbers and had to be excluded from the list as to balance the overall mean of the subscriber-base.

Pre-paid Churn: The pre-processed dataset for pre-paid subscribers has been grouped into 10 distinct clusters (as shown in Table 3)

which indicate that the overall subscriber-base activities were quite diverse. Further analysis of these segments revealed that such high diversification was due to the lucrative packages with multiple bundle offers being offered by the telecom operator. These bundles could be subscribed on daily, weekly or monthly basis and offered unlimited calls, messages and VAS services. Subscribers were using these bundles with non-consistent subscription routine and due to such varying usage their activity was split into high number of segments.

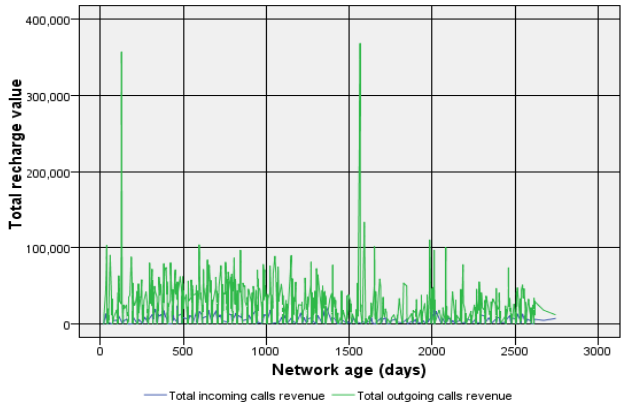

Figure 2: Post-paid dataset

Another behavior in this pre-paid base was of customers who generated outgoing revenue activity for a very limited time period. These subscribers had low active days and high maximum inactive days, which meant they were using the services of telecom provider after a gap of every few weeks. On further analyzing these re-activations it was discovered that these subscribers resumed services after getting extensive promotional offers every few weeks for its dormant subscriber base. After extensive study of all such subscriber behaviors in these 10 segments, the pre-paid dataset was split into two models. The first model included all those subscribers which satisfied the following conditions:

1) Active days > 75
2) Total inactive days < 60
3) Maximum Inactive days < 30
4) Total Recharge > 100

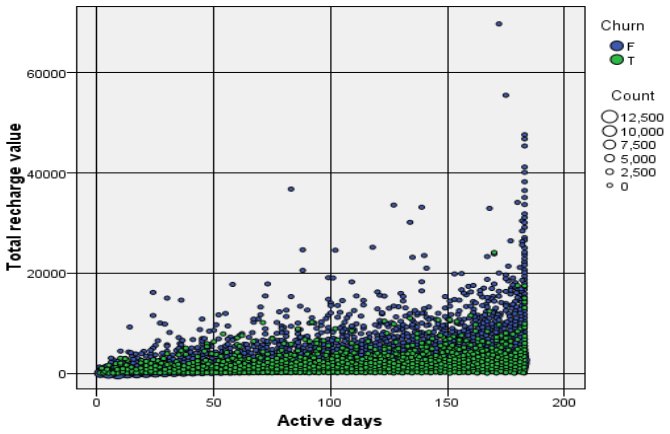

Figure 3: Prepaid dataset

The second model included all those subscribers which were discarded by first model.

All active subscribers having high active days and recharge can be seen in Figure 3, while subscribers with low active days and recharge are mostly churners. The challenge was to predict subscribers who have less number of active days but do not churn, and this was taken care of in the second model.

Customer Segmentation Model: The customer segmentation model was specifically designed for pre-paid subscriber-base (see Table 2). After rigorous analysis on source data (from Table 1), 44 distinct features where extracted which proved to be vital in producing distinct customer segments. The segmentation model built on these features could now split the subscriber-base according to combinations of average revenue per user (ARPU), dormancy, gender, package and usage which allowed handling each segment as per its subscribers' usage pattern.

Social Network Analysis Model: Social Network Analysis (SNA) is applied on top of traditional base models to extract relationships between the subscribers. The traditional models process each subscriber as an individual entity, which hides the impact of a subscriber's churn on other connected users. SNA based model consisted of 3 major features having churn details of last 90 day's subscribers' activity:

1) Caller Number: The person making the call
2) Called Number: The person being called
3) Duration of Call: The duration of call (in minutes)
4) Target variable: List of churned subscribers

Using these features, a network or graph is built where the caller and called numbers are presented as nodes and duration of calls are the weights of the edges. There was some basic pre-processing applied on the network to minimize the noise, i.e. eliminating the edges with less than 1 minute call (such calls were found to be 10% of all the calls made in a day on an average), discard calls where called number information was missing (such calls were found to be 1% of the daily calls), etc.

6. Methodology

The methodology section explains the steps involved in building the churn prediction, customer segmentation, and early churn (fraud detection) models. Churn prediction, being a supervised learning model, has the target variable (churn/no-churn) available in the training data. This model is further divided into two parts: 1) data modeling and 2) scoring. On the contrary customer segmentation model initially constituted of customer segments based on the 44 features (which are discussed in Section 4.4) and later used to generate monthly subscriber segments as well as track whether the subscriber revenue generation activity has had any effect.

In this case study there are three different models developed for churn prediction, with each model having its own set of subscriber-base. For postpaid churn model, the 0.1 million subscribers have been split according to their billing cycle and all subscribers who have been active in the last 12 months are included in the model's base. For prepaid churn model, the subscribers who have been active in the last 6 months are included in the model's base. The scoring of prepaid model has been scheduled on weekly basis while the postpaid model is scheduled on the basis of billing cycle. The resulting churn scores are later used to design new campaigns for subscribers.

The historical and predicted windows of pre-paid and post-paid model are shown in Figure 4. The historical data consisted of multiple attributes aggregated on a monthly basis against every distinct subscriber. Historical data window is followed by a 14 days marketing window where all subscribers are performing outgoing revenue generating activity. This gap has been kept as per the service

provider's marketing requirements, to ensure that the subscriber is active in this period and can be pitched with a retention campaign. The marketing window was followed by a 30 days predictive window where the subscriber's churn status is marked as inactive if no outgoing activity is found during this period, which indicates the start of their dormancy period. Furthermore, the dormancy period was followed by a 90 days inactivity window to indicate that the subscriber actually churned and moved out physically from his connection. On the other hand, the subscribers who were carrying out any activity (i.e. even if they are only receiving incoming off-net calls, and not making outgoing calls) till the end of the predictive window would have an active status and can be marked as no-churn at the end of the inactivity window. In the modeling phase, the churn model is trained to predict the subscribers churn status using only the

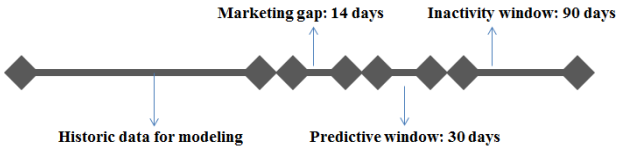

Figure 4: Pre-paid and post-paid model

historical data window. Thus, on every scoring cycle the model predicted subscribers churn status till the state of the predictive window. This churn status enabled the marketing team to target campaigns on subscribers especially those who became inactive during the predictive window.

Figure 5 shows different windows sizes for fraudulent sales/early churn model design. The historical data-modeling window included 7 days aggregated data with multiple features for distinct subscribers. The subscriber-base comprises of newly activated customers, who as policy of the telecom are eligible to consume free balance in the first 7 days of activation with no recharge activity during those 7 days. The subscriber's recharge status (target variable) is derived from the next 90 days beyond the modeling window. This model will predict newly activated subscribers who haven't recharged in the first seven days and their activation credit has been confiscated (i.e., defrauded) after the 7th day of activation.

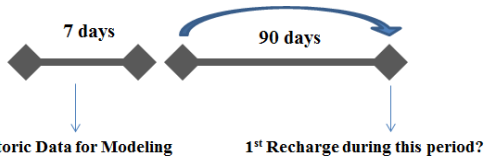

Figure 6: Early Churn (fraud detection) Model

For early churn (fraud detection) modeling, 7 months dataset for newly activated subscribers with no recharge in the first 7 days was extracted from the DWH. For each subscriber, a 7 day aggregated data along with some behavioral features were used to build the model. From the pre-processed data, the subscribers with new activations who had not performed any revenue generation activity were identified. These activations were having no data available because of zero activity, therefore were excluded from the modeling data and were declared as fraudulent sales.

The system architecture with detailed steps used to perform modeling and scoring for churn prediction and customer segmentation is shown in Figure 6. The telecom service provider is maintaining its transactional and usage data in a DWH whose sources are Customer Relationship Management (CRM) and Mobile

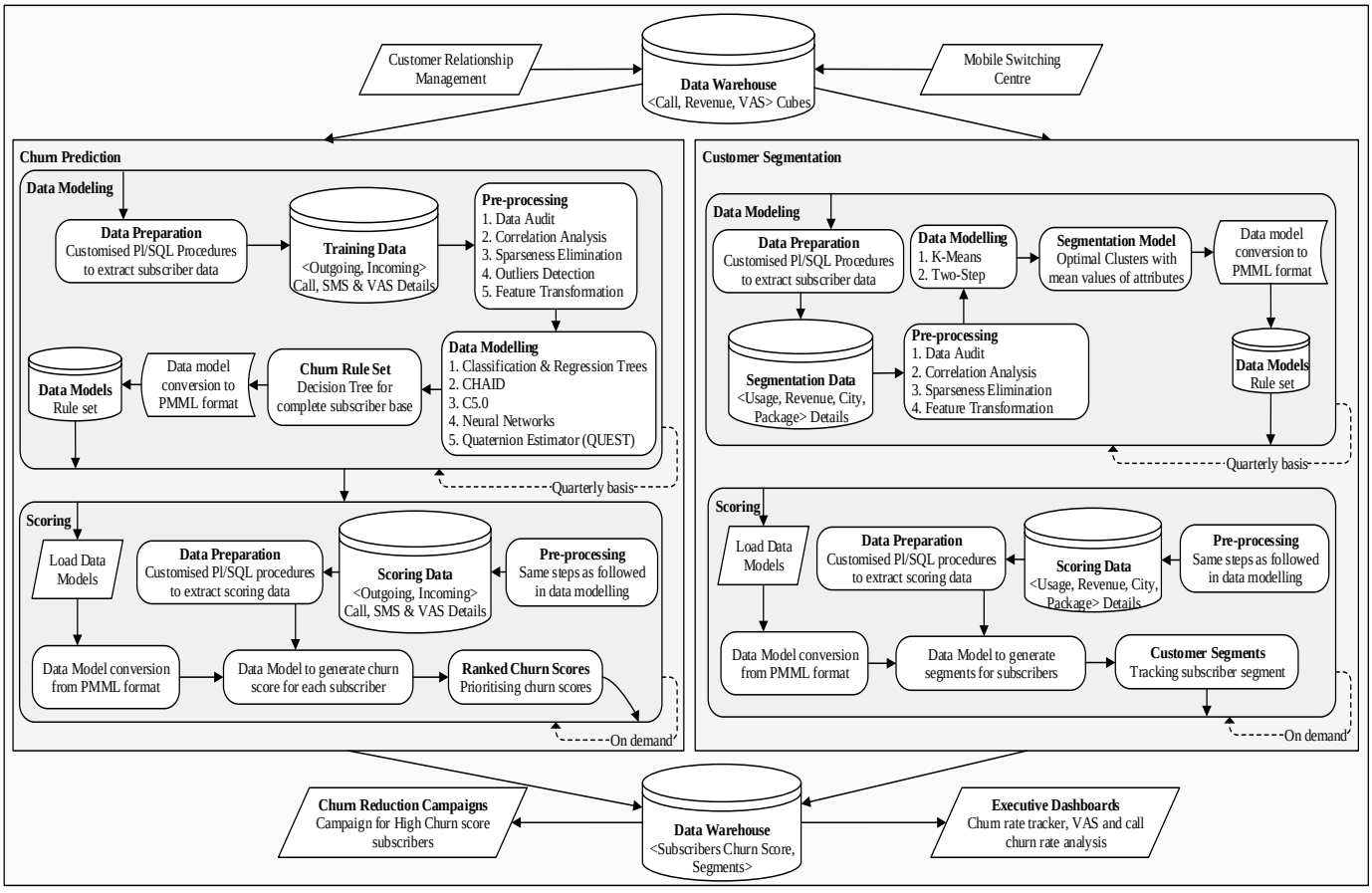

Figure 5: Detailed architecture of churn prediction and customer segmentation models

Switching Centre (MSC), which are updated on daily basis. The DWH is also maintaining daily, weekly and monthly level aggregated summary tables on top of transactional records.

The data is initially extracted from DWH through customized PL/SQL procedures for each subscriber. These procedures are executed in various query blocks and consolidated at the later stage to produce a single table with unique subscriber number and aggregated records as per model requirement, i.e. on monthly-basis for pre-paid and post-paid models while weekly aggregation for early churn. The aggregated data includes subscribers' usage and revenue information of outgoing and incoming calls, short message service (SMS), VAS, and some derived features. This data is pre-processed following the steps that are discussed in Section 5. For all three models the raw data was initially audited to identify the refinements required to minimize noise from the data and then partitioned into training and testing sets for data modeling and evaluation purpose. The training set comprises of 70% instances while the testing set composed of remaining dataset. The training data is used to build the model and generates different patterns which are later used to evaluate the model using testing set.

Data modeling has a pool of algorithms for churn prediction consisting of regression-based, decision trees (CHAID, C5.0, QUEST & C&R Tree) and Neural Networks algorithms. These algorithms are used to build models on training set where their accuracies are compared with each other and the model with maximum accuracy is selected and converted to Predictive Model Markup Language (PMML). The PMML formatted model is later used to score churn probabilities on quarterly basis and updated in DWH. These scores are ranked to identify the subscribers with highest churn score and reason of churn which is used to pitch campaigns on predictable churners.

The customer segmentation model, shown in Figure 6, is different from the churn prediction, as it has no target value associated with it. The customers are grouped into clusters with similar spending patterns and behaviors which allow the telecom service provider to focus on the more appropriate product and service offerings for a particular subscriber. The pre-processed data is used for customer segmentation modeling, where different features, i.e., usage, revenue, active days, etc. are used to form distinct clusters of subscribers. The modeling process uses two clustering algorithms: 1) K-means, and 2) Two-step [17, 18] where algorithm which generates highest accuracy is selected. These segments are passed to marketing team as to design different campaigns for the potential churners. The model is scheduled to execute on monthly basis to ensure the maintenance of positive revenue segment tracking.

Social Network Analytics: Social network analysis was used to analyze the relationship between the subscribers to form the interrelated groups that cannot be extracted from the traditional techniques. The traditional data-mining algorithms use process the subscribers data as i.i.d. (independent and identical distributed) which ignores relationship (call-in and out) between the subscribers which is a key in this context. Also it is observed that the traditional algorithms are not able to compute accurate predictions for the least and most frequent callers. Mostly the least frequent callers are predicted as churners due to insufficient calling patterns and revenue generation, while in actual these callers are not churners. On the other extreme the most frequent callers are usually predicted as no-churn by the traditional models but in actual these subscribers can churn.

As a result the most and least frequent subscribers are excluded from the traditional churn analysis to social network analytics on those set of subscribers. SNA is further divided into two techniques: 1) Diffusion analysis [12] and 2) Group analysis. Diffusion analysis is applied on the data to predict churn while group analysis is used to segment customers. Figure 7 shows the detailed architecture of SNA based customer segmentation and churn prediction.

Churn Prediction using Diffusion Analysis: A subscriber's calling pattern has an impact on other peer nodes in its social structure. The

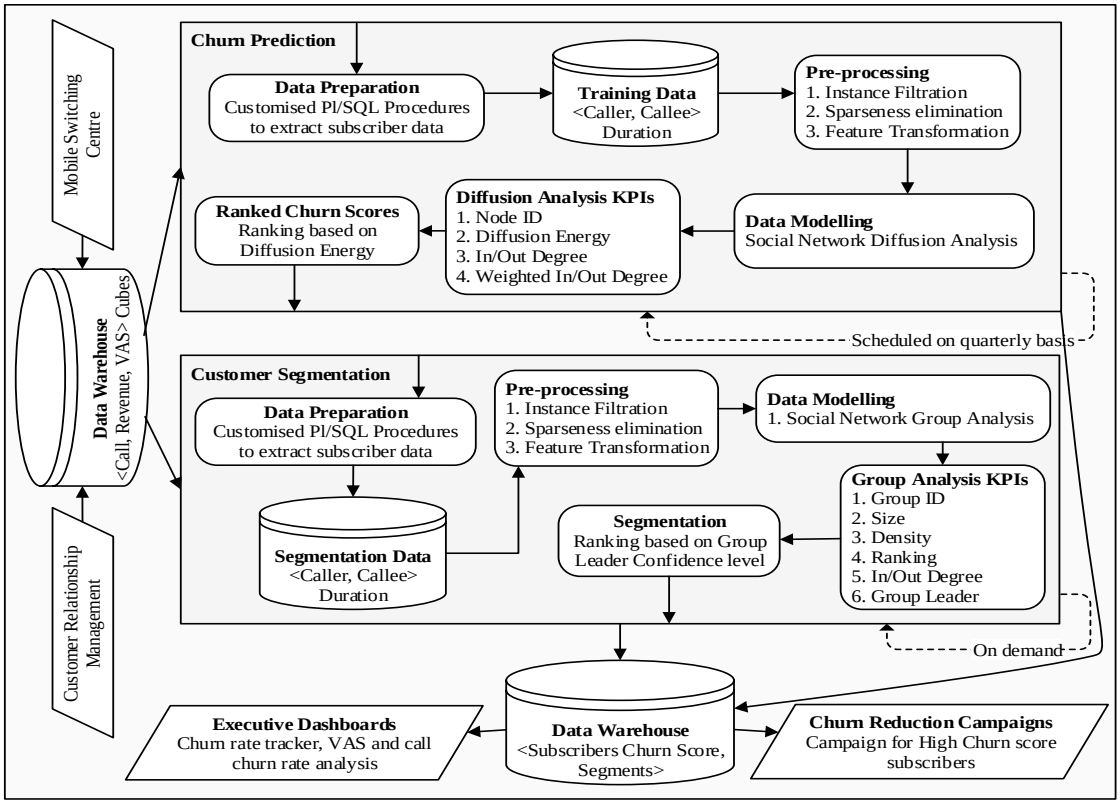

Fig. 7. Detailed architecture diagram of Social Network Analytics

diffusion analysis technique is specifically applied on such individuals who are affected due to the churners in the group. This technique uses four features extracted from the DWH CDR table which are discussed in Section 5. The caller and called numbers of CDR are used to form nodes of the network while duration of calls represents weights of the edges. Using this network, diffusion analysis technique generates a set of KPIs, i.e. diffusion energy which is used as churn score, number of in and out degree for calls etc. The subscriber churn scores are ranked and the top scorers (having high diffusion energy) are marked as probable churners. These subscribers are to be immediately targeted with effective campaigns and offers.

Customer Segmentation using Group Analysis: Group analysis technique is applied on CDR to form customer segments where the node (which represents a subscriber) containing the maximum in and out edges (calls) is marked as leader of that group. This model was initially supposed to target group leaders as they are highly influential over other group members. These group leaders are subscribed for service offerings and their influence helps in promoting those offerings to wider group members likely to subscribe similar offering. Another objective of this model is trying to attract a group leader from a telecom competitor. If this leader successfully ports-out, it can also help in increasing the churn rate of the group members associated with the particular competitor as well as reducing the churn rate of group members associated with this telecom service provider. The group analysis algorithm uses this data to generate a set of KPIs which include size, density, ranking for each group based on incoming and outgoing calls (degree), and call duration (weight) to identify group leaders.

7. Results

The accuracy of churn prediction model is computed using a confusion matrix where the actual and predicted churn values are used to predict churn accuracy. The accuracy of customer segmentation model is computed using Silhouette measure of Cohesion and Separation [10] measure.

The churn prediction model has been evaluated on traditional (base) algorithms initially which are reported in Table 5. Additionally, SNA were applied on the instances which were unable to predict correctly using base algorithms (most and least frequent callers) to boost the overall accuracy of the system. The SNA technique shows a boost of 10% on pre-paid and 8% on post-paid models.

Table 5: Comparison of Machine Learning algorithms used for churn prediction

Algorithms	Scores of traditional approaches		
	Early-churn (%)	Pre-paid (%)	Post-paid (%)
C5.0	61.20	70.66	74.27
C&R Tree	65.12	63.17	68.26
CHAID	63.51	60.35	73.12
QUEST	63.35	73.78	65.23
Social Network Analytics Boost	-	9.67	7.98
Total Accuracy	65.12	83.45	82.25

Similarly, traditional clustering algorithms are used as base models where the results of one of the segmentation algorithms is sufficient enough to meet the target of at least 0.5 silhouettes of cohesion and separation [10] (as shown in Table 6).

Table 6: Comparison of Unsupervised Learning algorithms used for Customer Segmentation

Algorithms	Scores of traditional approaches		
	Silhouette	Number of Clusters	Largest Cluster (%)
K-Means	0.516	10	18
Two-Step	0.469	7	25

The total clusters, shown in Figure 8, still account for a large set of subscribers who were impossible to be targeted at the same time, and therefore the top subscribers are listed using SNA group analysis technique. The results list top 300 subscribers who are social leaders with high ARPU. A high influence over several other subscribers (means) they can easily spread new offers and promotions to their peers, thereby saving operator's expenses.

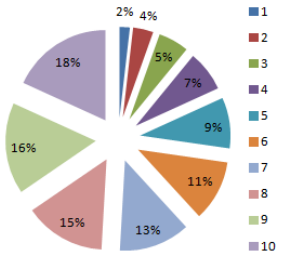

Figure 8: Segments and their proportions generated with K-means algorithm

8. Discussion and Conclusion

This research work proposed a combination of traditional and SNA based approach to the conventional telecom problems, churn prediction and customer segmentation. The telecom services provider discussed here were facing high churn rate and fraudulent sales problems resulting loss in terms of revenue and subscriber-base to the operator. The problems required some advanced analytical techniques to minimize these problems. Every data-mining problem requires sufficient amount of historical data to build predictive models but the available data consisted of limited number of features containing noisy data. This data was pre-processed using in-depth analysis to minimize noise and ensure better understanding of customer behavior. On the other hand, the traditional learning algorithms were not able to find the relationships between different subscribers which extract useful patterns for churn prediction and customer segmentation models. As a result, SNA based algorithms were used on top of the base (traditional) models to incorporate multi-dimensional analysis and to boost the overall accuracy of the system.

9. Future Work

The analysis of the telecom data using different data-mining techniques was particularly aimed to predict subscribers' churn behavior, finding leaders and churn effected subscribers in a group, improve customer relationship management, and develop various campaigns strategies for customer retention and loyalty. The local telecom market offers lucrative offers to subscribers with easy carrier switching offer, therefore customers needed to be profiled into specific groups and only targeted campaigns and packages would allow such customers to be retained.

In the next step using more features such as call quality, customer complaint types including severity level and resolution time can further help in boosting the performance of churn prediction models. Also the demographics information of the subscribers can further help in segmenting customers in a much efficient and accurate way.

Using text analytics for converting unstructured data, i.e. customer feedback, to structured data can further improve the models. These enhancements could add a new dimension in customer churn and segmentation models where customer behavior and attitude can be learnt and predicted.

Acknowledgment

The author wishes to thank Sanket Jain, Zunaira Rasheed, Shamyl Bin Mansoor and Syed Yasir Hassan for taking part in subject discussion and review of this work.

References

[1] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Computer Assisted Customer Churn Management: State-of-the-Art and Future Trends," *Journal of Computers and Operations Research*, vol. 34, issue 10, Oct. 2007, pp. 2902-2917.

[2] R. Fildes, "Telecommunications Demand Forecasting - A Review," *International Journal of Forecasting*, vol. 18, 2002, pp. 489-522.

[3] S. Aheleroff and M. R. Gholamian, "Customer Segmentation For a Mobile Telecommunications Company Based on Service Usage Behavior," in *Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications*, Macau, China, Oct. 2011, pp. 308-313.

[4] Z. Zhongding, M. Xuemei and L. Guangcan, "Customer Segmentation Algorithm of Wireless Content Service Based on Ant K-Means," in *Proceedings of the 2009 International Forum on Computer Science-Technology and Applications*, vol. 1, 2009, Pages 267-269.

[5] R. K. Gopal and S. K. Meher, "Customer Churn Time Prediction in Mobile Telecommunication Industry Using Ordinal Regression," in *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining*, 2008, pp. 884-889.

[6] Vicente Garcia, J. Salvador Sanchez, Ramon A. Mollineda, Roberto Alejo, and Jose M.Sotoca. "The class imbalance problem in pattern classification and learning," in *Proceedings of Francisco J. Ferrer-Troyano et al, editor, II Congreso Espanol de Informatica*, 2007, pages 283–291.

[7] X. Y. Liu, J. Wu, and Z. H. Zhou. "Exploratory under-sampling for class imbalance learning," In *ICDM IEEE Computer Society*, 2006, pages 965–969.

[8] J. Lu, "Predicting Customer Churn in the Telecommunications Industry - An Application of Survival Analysis Modeling Using SAS," in *AUGI 27*, Orlando, Florida, Apr. 2002, pp. 732-741.

[9] C. Drummond and R. C. Holtel. Severe class imbalance: "Why better algorithms arent the answer," in *Proceedings of 16th European Conference of Machine Learning*, 2005.

[10] P. J. Rousseeuw. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, 1987, pages 53–65.

[11] Y. Richter, E. Yom-Tov, and N. Slonim, "Predicting Customer Churn in Mobile Networks through Analysis of Social Groups," in *Proceedings of the SIAM International Conference on Data Mining (SDM)*, Columbus, Ohio, 2010, pp. 732-741.

[12] K. Dasgupta, R. Singh, B Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati and A. Joshi, "Social Ties and their Relevance to Churn in Mobile Telecom Networks," in *Proceedings of 11th Conference on Extending Database Technology*, Nantes, France, Mar. 2008.

[13] Pushpa and G. Shobha, "An Efficient Method of Building the Telecom Social Network for Churn Prediction," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 2, issue 3, May 2012, pp. 31-39.

[14] E. Xevelonakis and P. Som, "The impact of social network-based segmentation on customer loyalty in the telecommunication industry," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, issue 2, May 2012, pp. 98–106.

[15] K. Coussement and D. Van den Poel, "Churn Prediction in Subscription Services: An Application of Support Vector Machines while Comparing Two Parameter-Selection Techniques," *Journal of Expert Systems with Application*, vol. 34, issue 1, Jan. 2008, pp. 313-327.

[16] H. Huang, "Rank Based Anomaly Detection Algorithms". *Electrical Engineering and Computer Science – Dissertations*, 2013. Paper 331.

[17] T. Zhang, R. Ramakrishnon, and M. Livny. BIRCH: An efficient data clustering method for very large databases. In: *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1996, Montreal, Canada.

[18] T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In: *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, 2001. San Francisco, CA.

Ahsan Rehman is a Data Analyst working at IBM Business Analytics and Optimization, Global Business Services. He received his B.S. degree in Information Technology from the National University of Sciences and Technology in 2012. His current research interests include predictive analytics, social network analytics and big data.

Abbas R. Ali is a Data Scientist working at IBM Business Analytics and Optimization Center of Competence. He received his B.S. degree in computer science and mathematics from the Institute of Management Sciences in 2004, M.S. degree in artificial intelligence and natural language processing from the National University of Computers and Emerging Sciences in 2009 and currently doing his PhD in machine learning and predictive analytics from Bournemouth University. His current area of research is Meta-level Learning in the Context of Multi-component, Multi-level Evolving Predictive Systems.