



Internal Report 2013–03

Universiteit Leiden

Computer Science

Extending traditional telecom churn prediction
using social network data

Name: Palupi Diah Kusuma
Student-no: 1056247

Date: 15/02/2013

1st supervisor: Dr. P.W.H. (Peter) van der Putten
2nd supervisor: F.W. (Frank) Takes, MSc
In-company supervisor: D. (Dejan) Radosavljevik, MSc

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Churn, which is defined as the loss of customers to another company, is a crucial problem in the telecommunication industry. The telecom market has matured and opportunities for growth are limited, so retaining the existing customer base has a high priority. Cost of acquiring a new subscriber is 5 times higher than of retaining one (Jacob & Kerremans, 2010). In order to minimize the churn rate, mobile telecom players have to form defensive strategies to identify and present appropriate incentives to subscribers with high propensity to churn. The conventional churn prediction models are typically constructed using variables that characterize customers, such as subscribers' contractual, demographic or usage information (Ferreira et al., 2004), complaints data (Hadden et al., 2006) or Customer Experience Management data (Radosavljevik et al., 2009). Recently, detailed information retrieved from Call Detail Records (CDRs), such as the incoming and outgoing numbers, duration and timeframe, has also been exploited to predict churn and to reveal the viral effect of churn (Dasgupta et al., 2008). This social network data is projected to a graph containing a set of vertices and a set of edges, which respectively refer to subscribers and the interaction flows among them.

Commonly, a churn model focuses either on the traditional variables, such as demographics, contractual, handset and usage, or on the social interactions. We believed that the traditional variables and social factors, such as a strong tie with churned neighbors, could mutually contribute to churn. Therefore, we proposed to combine both aspects using the feature-based scoring and the propagation technique. We constructed seven distinct models to evaluate the added value of the new approaches. The models were built separately for the postpaid and prepaid telecom segments because churn was defined differently in each of those segments. The first three models were feature-based models built using scoring algorithms. *Model 1* was trained using traditional predictors, *Model 2* used social network features and *Model 3* used combination of both features. A spreading activation algorithm was applied to build the remaining four models. In these models, negative energy of churn was propagated through a directed or an undirected social network. *Model 4* was based on the simple propagation model proposed by Dasgupta et al. (2008), which assigns energy of 1 to all churners and 0 to all non-churners in the beginning of the propagation process. *Model 5* is an extended propagation model, which incorporated the result of *Model 1* as the initial energy of the non-churners. *Model 6* and *Model 7* are respectively similar to *Model 4* and *Model 5* aside from the fact that these last 2 models propagated energy through an undirected graph instead of a directed graph.

We first expected that the hybrid models combining traditional predictors and the social ties information would have the best performance (*Model 3*, *Model 5* and *Model 7*). Our analysis returned similar results for both the prepaid and postpaid segment. The extended propagation models were clearly superior to the simple propagation models. The predictive accuracy of those models, however, was still lower than the benchmark model, *Model 1*. The results showed that social network features do not add any substantial performance gain. Although correlation of churn probability with the number of churned neighbors showed promising results, *Model 2* and *Model 4*, which are based exclusively on social network, proved otherwise. Those models consistently made the most incorrect predictions. We could conclude that social network alone is not an appropriate measure to predict churn because the performance gain was minimal, while the computational cost was relatively high. The contribution of traditional predictors to churn prediction is substantially higher than that of the social network behavior.

Acknowledgements

First and foremost, I would like to thank Peter van der Putten and Dejan Radosavljevik for their guidance and support. I am very grateful that I have had the opportunity to learn from them. I could not ask for better supervisors. I would also like to thank Frank Takes for his feedbacks and advice as the second supervisor of my thesis. Next, I would like to show my special gratitude to Miruna Anastasoae and Kim Larsen, who have continually inspired and motivated me from the very beginning of the internship period.

Special thanks to my colleagues, especially in the Network Economics department, for their knowledge, expertise and all wonderful lunch discussions. It has been a great pleasure to be a part of a fantastic team.

Last but not least, I would like to thank my family and friends for their endless love and encouragement. I would also like to thank Jeroen for his understanding, technical and moral supports. Finally, I would like to express my sincere appreciation to everyone who has contributed to this thesis in one way or another and facilitated me for completing the project.

Table of contents

Abstract	2
Acknowledgements.....	3
Table of contents	4
List of Figures	6
List of Tables	7
1. Introduction	8
1.1. Background	8
1.2. Research Problem.....	8
1.3. Related Work.....	9
1.4. Research Methodology.....	11
1.5. Thesis Outline.....	11
2. Theoretical Background.....	13
2.1. Social Network Graph.....	13
2.1.1. Nodes.....	13
2.1.2. Edges	13
2.1.3. Network dynamics.....	14
2.1.4. Graph centrality: Degree	14
2.2. Telecom social network.....	15
2.2.1. Nodes.....	15
2.2.2. Edges	15
2.2.3. Network dynamics.....	16
2.3. Machine learning.....	16
2.3.1. Scoring algorithm	16
2.3.1.1. Logistic Regression.....	17
2.3.1.2. Decision Tree	17
2.3.2. Social Network Mining using Spreading Activation	18
3. Implementation setup.....	22
3.1. Operational definition of telecom churn	22
3.2. Dataset.....	23
3.2.1. Population and outcome definition.....	23
3.2.2. Call Detail Records.....	24
3.2.3. Subscriber information.....	26
3.2.3.1. Intrinsic variables.....	26
3.2.3.2. Extrinsic information.....	27
3.3. Implementation Scenarios.....	27
3.3.1. Scoring models.....	27
3.3.2. Propagation models	28
4. Implementation Results	29
4.1. Modeling results.....	29
4.1.1. Prepaid dataset	30
4.1.1.1. Prepaid: Scoring and Propagation Models	31
4.1.1.2. Prepaid Scoring: Predictor Details	33
4.1.2. Postpaid EOC6 dataset	34
4.1.2.1. Postpaid EOC6: Scoring and Propagation Models.....	34
4.1.2.2. Postpaid EOC6 Scoring: Predictor Details.....	35
4.1.2.3. Postpaid EOC6: Scoring Models without Contract.....	36
4.1.3. Postpaid EOC2 dataset	38

4.1.3.1. Postpaid EOC2: Scoring and Propagation Models.....	38
4.2. Effect of the weight propagation function	39
5. Conclusions and Future Works	40
5.1. Conclusions	40
5.2. Future works	41
References	43
Appendix A. Abbreviations	45
Variable Description.....	46
B1. Intrinsic variables	46
B2. Extrinsic variables.....	47
Scoring Predictors	48
C1. Predictor Description.....	48
C2. Scoring Model Predictors.....	49

List of Figures

Figure 1. Social network graph	13
Figure 2. Telecom social graph	15
Figure 3. Initial energy of the simple and extended propagation technique	19
Figure 4. Spreading activation in a weighted graph.....	20
Figure 5. Population settings.....	23
Figure 6. The distribution of postpaid churners based on the EOC month	24
Figure 7. Illustration of weight decay applied on 7 observation days between a pair of nodes	26
Figure 8. Implementation scenarios	27
Figure 9. Coefficient of Concordance (Chordiant, 2009).....	29
Figure 10. Illustration of a gain chart and a lift chart.....	30
Figure 11. Correlation of churn rate with churner degree ratio	30
Figure 12. Gain and Lift chart of the prepaid models	32
Figure 13. Prepaid predictor: CONTRACT_STARTDATE	33
Figure 14. Prepaid predictor: DEGREE2ND.....	33
Figure 15. Gain and Lift chart of postpaid EOC6 models.....	35
Figure 16. Postpaid EOC6 predictor: RATEPLAN_GROUP.....	35
Figure 17. Postpaid EOC6 predictor: LIFETIME	36
Figure 18. Postpaid EOC6 predictor: CONTRACT_END_TOLAST	36
Figure 19. Gain and Lift chart of EOC6 scoring models without contract related variables.....	36
Figure 20. Postpaid EOC6 predictor: AGE.....	37
Figure 21. Gain and Lift chart of postpaid EOC2 models.....	39

List of Tables

Table 1. Contingency table of GENDER and CHURN variable	18
Table 2. Chi-Square distribution table (Ryan, n.d.)	18
Table 3. Churned subscribers on March 2012 for selected EOC month	24
Table 4. Churn rate of the sample sets	24
Table 5. Performance of the prepaid scoring models.....	31
Table 6. Performance of the prepaid propagation models	32
Table 7. Prepaid predictor: CONTRACT_STARTDATE	33
Table 8. Prepaid predictor: DEGREE2ND.....	33
Table 9. Performance of the postpaid EOC6 scoring and propagation models.....	34
Table 10. Postpaid EOC6 predictor: RATEPLAN_GROUP	35
Table 11. Postpaid EOC6 predictor: LIFETIME	36
Table 12. Postpaid EOC6 predictor:	36
Table 13. Performance of EOC6 scoring models without contract related variables	37
Table 14. Postpaid EOC6 predictor: AGE.....	37
Table 15. Performance of postpaid EOC2 scoring models	38
Table 16. Performance comparison of prepaid using linear and non-linear propagation function.....	39

1. Introduction

This research study explores the extent to which social network information could be used to predict telecom churn and also to which it could potentially improve the predictive performance of the conventional churn prediction method. The first section of this chapter introduces the background information and the motivation of the research topic. The second section presents the problems and the questions that can potentially be solved and answered by the research study. A brief description of the research methodology is covered in the third section. The literature review, which is presented in the fourth sections, discusses related research studies addressing similar topic. It also explains how our research could have an added value to these previous studies. The rest of the thesis is outlined in the last part of the chapter.

1.1. Background

A social network is a dynamic relational system of entities that are connected by one or more interaction types, such as friendships, kinships, voice calls, emails or citations. A social network is often projected as a graph with nodes/points and ties/lines, which respectively represent actors and relationships. The study of a social network, which is referred to as social network analysis (SNA) has emerged as an interdisciplinary field in the recent years. It has been conducted in many research fields including sociology, anthropology, psychology, statistics and mathematics. SNA focuses mainly on exploring the relationship of network actors with their neighbors rather than the characteristics of the individuals. It aims to reveal the underlying relationship structures and patterns of the interacting units, for instance the network growth and evolution (Backstrom et al., 2006), the discovery of communities (Steinhaeuser & Chawla, 2008) and the influential role of a social network in the decision making processes (Hill et al., 2006; Kiss & Bichler, 2008).

Mobile interactions among telecom subscribers are also a form of a social network. In a telecom graph, subscribers are depicted as nodes, which are connected by either directed or undirected ties representing voice calls or messages among them. This type of graph does not rely on users to explicitly define links among them (as in a network constructed from an online social network). Therefore, it could provide a fairly straightforward insight of the subscribers' relationships. As a consequence, a telecom operator could gain a better understanding on the subscribers behavior and it could use the knowledge as a support for the decision making processes, such as cross-selling (Kiss & Bichler, 2008) or predicting churn (Dasgupta et al., 2008), in order to ensure the profit margin stability. Kiss and Bichler (2008), who sought to measure the role of influencers in the telecom viral marketing, state that highly influential subscribers are able to diffuse messages quicker throughout the network compared to the rest of the network members. A churn study conducted by Dasgupta et al. (2008) claims that the churn decision can be diffused through the social network. They show that the propensity of a subscriber to churn correlates with the number of friends that have already churned. Previous studies demonstrate that a social network, in some extent, has an impact on the customer decisions. Hence, it is worthwhile to further investigate the role of social network in influencing customer behavior.

1.2. Research Problem

As the telecom market is maturing, it is more difficult yet costly to acquire a new subscriber than to retain one (Jacob & Kerremans, 2010). Therefore, the focus of telecom players has shifted from customer acquisition to maximizing the value of the existing customer base and retention. The key objective of this research study was to support the business so that the loss of customers to another company, which is simply referred to as churn, can be minimized. In order to manage churn, it is

important for the telecom provider to identify the subscribers who are more likely to churn. Marketeers could then form appropriate defensive strategies to prevent churn.

Many research studies have been conducted to build churn models and to optimize model performance. A *conventional* or a *feature-based churn model* can be constructed by utilizing *intrinsic* and/or *extrinsic* variables (Karnstedt et al., 2010). Since the variables are presented in a tabular form, the conventional churn model is also called the *tabular churn model*. Intrinsic variables are related to the customer profile and the products/services provided by the telecom network, such as demographic information (e.g., age, gender or location), contractual details (e.g. package plan type, contract duration or price), usage facts (e.g., voice call duration, the frequency of sending or receiving text-messages) and/or other service-related information (e.g., number of interaction with customer service or number of dropped calls). These *intrinsic variables* are also referred to as *traditional predictors*. The *extrinsic variables* represent features extracted from social interactions among subscribers, such as the number of neighbors or the interaction frequency with the neighbors. The term *extrinsic variables* and *social network features* are used interchangeably throughout this paper.

Although much work has been done in the churn prediction field, these studies generally exploit either intrinsic or extrinsic variables separately. The predictive power of a churn model based merely on the traditional predictors might potentially be reduced in the case of many missing values (e.g., in the prepaid base). On the other hand, the social network features might not be enough to influence the churn decision of a subscriber. Therefore, we proposed a new tabular model, which combines both intrinsic and extrinsic parameters to predict churn, aiming to gain significant lift. We also assessed the extent to which social network features can be used to improve the churn prediction performance compared to the model built using only traditional predictors.

The feature-based model, however, does not take into account the influential effect of an individual decision to his/her social network. A recent work by Dasgupta et al. (2008) has been able to address this problem by constructing a churn model based on a traditional social network mining technique, i.e., propagation. The churn model propagates the negative churn influence from one subscriber to another in a cascade manner. Besides building a traditional tabular churn model using combination of the traditional predictors and the social network features, we also proposed a novel approach, which extends the traditional propagation model to include the traditional predictors' information.

The research project strived to answer the following questions:

1. What are the characteristics of subscribers with high propensity of churn?
2. How to extend the conventional tabular mining as well as the traditional social network mining techniques to combine both intrinsic and extrinsic features?
3. What is the added value of incorporating social network features into the traditional tabular churn model?
4. Is there any performance gain on the extended models that take into account the traditional predictors as well as social ties information, and if so, does the performance gain justify the computation cost?

1.3. Related Work

The conventional churn prediction models are typically simple, robust and have a relatively good performance. Many machine learning techniques, such as decision trees, naïve bayes, logistic regression, neural networks and genetic algorithms, are used to build these feature-based models. Churn has been widely analyzed not only in the telecommunication industry (Ferreira et al., 2004; Hadden et al., 2006;

Dasgupta et al., 2008; Radosavljevik et al., 2009; Kisioglu & Topcu, 2011), but also, among others, in the online gaming (Kawale et.al, 2009), insurance markets (Soeini & Rodpsyh, 2012) and banking (Prasad & Madhavi, 2012).

Ferreira et al. (2004) utilized contractual and demographic information of a Brazilian mobile telecommunication service provider to build several postpaid churn models using neural networks, decision trees, genetic algorithms and hierarchical neuro-fuzzy systems. Besides evaluating the predictive power, they also assessed the profitability value of those tabular models. They proved that even the churn models with the worst performance were still able to save significant cost in the postpaid segment. Since the demographic and contractual data might be limited in some cases, Hadden et al. (2006) exploited provisions, complaints and repair interaction data to build the churn models. They claimed that the regression tree model performs better than the one with neural networks or logistic regression. However, there was no further information regarding the performance comparison between the complaints-based model and benchmark model (based on demographic and contractual variables). Radosavljevik et al. (2009) investigated the extent to which Customer Experience Management (CEM) data could improve prepaid churn prediction. Several Key Performance Indicators (KPI) of service quality combined with other subscriber data were used to train the decision tree models. Since the CEM data is always available, the constraint on lacking demographic information on the prepaid subscribers could be eliminated. Although the CEM data was predictive, the empirical study showed that there was no significant gain on the churn model performance compared to the benchmark model.

Kisioglu and Topcu (2011) applied a slightly different approach to analyze churn. Using the Bayesian Network technique, they investigated the interdependent causal relationships of variables that determine churn probability in the postpaid segment. Consider two variables, age and call duration. Age might have an effect on call duration to some degree. For example, a teenager group might initiate fewer calls than the older segment because they prefer other type of communication. Since call duration contributes directly to churn, it means that age has an indirect effect on churn. Using a Bayesian network model, the variable causal relationship could be visualized. Besides defining the churn probability of subscribers, the effect of multiple variables on churn could also be examined.

Social network analysis recently became a popular method for predicting future churners. Dasgupta et al. (2008) analyzed the influential impact of the churned neighbors to their social circle by applying a spreading activation-based technique similar to trust metric computations (Ziegler & Lausen, 2004). They were able to show that churn can be propagated through a social network. Although the study was limited to use social ties information only, reasonable predictive accuracy could still be achieved. The analysis identified that the churn propensity of a subscriber correlates positively with the number of churned neighbors. Kawale et.al (2009) conducted a similar study using social network data from a popular online gaming community. They proposed a new twist to the existing churn propagation model proposed earlier by Dasgupta et al. (2008) by combining the social influence and user engagement in the game. The user engagement property, which referred to the length of the playing session during the observation period, can be classified as an intrinsic variable. The research showed that the models trained using a combination of social factors and this user engagement property performed better than a traditional propagation model.

We believed that a churn model, which exploits both traditional predictors and social relationships among subscribers, could outperform the simple propagation model and the conventional churn model built exclusively using traditional predictors. Customer's decision to churn not only depends on the social influences but also on how they perceive the products/services. Therefore, this research project aims to investigate the added value of models incorporating both aspects.

1.4. Research Methodology

This research study was conducted in a network telecom provider in the Netherlands. The social network information was derived from the Call Detail Records (CDRs), whereas the traditional predictors, i.e., demographic, contractual, handset and usage information, were obtained from the marketing data warehouse. CDRs, which contain detailed facts about mobile interactions, such as source phone number, destination phone number, the duration and the timestamp, were aggregated to include only distinct caller-callee information due to limited resources. This information was then mapped as a social graph made up of nodes denoting subscribers and edges representing interactions (voice calls and text-messages).

Based on service-payment, telecom subscribers are categorized into two groups, postpaid and prepaid. Postpaid subscribers have a specific service arrangement with a mobile service provider, which specifies the limit of provided services, such as calls, SMS, MMS and internet data that the subscribers could initiate. These subscribers pay a flat rate at the end of each service-month as long as their usage is less than or equal to the limit. Additional charges are applied for any usage above the predefined limit. Unlike postpaid subscribers, prepaid subscribers make an advanced payment method by means of a voucher purchase. If they do not have sufficient voucher credit, they could not initiate any calls, send SMS/MMS or connect to internet. They could re-enable the service by recharging or topping-up their voucher credit. In this study, we evaluate and discuss churn models for prepaid and postpaid segment separately. Unlike postpaid subscribers that are bound by a contract, prepaid subscribers have no switching barriers, so they can discontinue their subscription or move to another mobile carrier anytime. The operational definitions of churn are also defined separately because they differ per segment.

For each segment, we constructed and assessed the performance of several models using the conventional technique, propagation technique and combination of both. We used the conventional model constructed exclusively using traditional predictors as the benchmark model. We compared the predictive performance of all models to this benchmark model to determine the added value of social network information and to evaluate the most appropriate method for the project. Detailed description of the implementation scenarios is given in the chapter 3.

The proposed research study used Neo4j technology to store the graph structure and the content of graph elements. Neo4j database differs from the relational database management systems, as it is oriented to store semi-structured and network data, which makes it appropriate to store social graphs (Neo4j, 2012). It also provides an intuitive representation of the graph and it is easy to traverse through the graph's nodes and relationships. The scalability of this system presents a great advantage because its functionality can be easily extended to perform a large scale social network analysis. Eclipse IDE (Eclipse Foundation, 2012) was used to automate the construction of the social graph and to perform spreading activation algorithm. The Chordiant Predictive Analytics Director software (Chordiant, 2009) was primarily used to train the scoring models and to evaluate the model performance with the benchmark value. This software provides an automated yet intelligence model building and elegant visualization, which results in saving a reasonable amount of time.

1.5. Thesis Outline

The rest of the paper is outlined as follows. Chapter 2 covers the theoretical background of the social network analysis and techniques to analyze social graphs. The basic structure of the network graph is described here, along with the temporal aspect of the graph properties. It also explains how to construct a social network graph using the information gathered from mobile telecommunication field. Moreover,

the last section of this chapter covers the detailed description of the existing machine learning methodologies that were utilized in this research study.

Chapter 3 discusses the basic concept and operational definition of telecom churn for each customer segment. It also provides the dataset description and the necessary activities performed in the data preprocessing stage. This chapter also introduces the implementation scenarios that were applied to build the churn prediction models.

Chapter 4 comprises the implementation results of all scenarios presented in the Chapter 3. The evaluation of the model performances for different experimental settings in the postpaid as well as the prepaid segment is also presented in this chapter. Moreover, it highlights some important issues and limitations of these techniques.

Finally, chapter 4 summarizes the paper, the limitation of the project and presents some suggestions for the future work.

2. Theoretical Background

The following chapter discusses the graph theory and the basic data mining techniques that can be applied to identify churners. The chapter begins with the definition of the general attributes of the social graph and the telecom graph. Next, a brief description of the scoring technique using logistic regression and decision tree algorithms is presented. The spreading activation method, which was applied for building the churn propagation model, is also discussed in the last section of the chapter.

2.1. Social Network Graph

The social network graph theory in this section will be based mainly on text by Wasserman and Faust (1994), Borgatti (1994) and Hanneman and Riddle (2005). Social network analysis aims to understand the dynamic interaction between network entities. The structure of a social network can be formally defined as graph $G = (V, E)$ with a set of vertices V and a set of edges E (Borgatti, 1994). Two entities of a certain social network share an edge if there is a common relation between them. In an online social network, the network entities are connected by means of friendship relations, whereas in the telecom social network, members are connected by their mobile interactions. Figure 1 shows an example of a directed social network graph with nodes representing graph entities and lines representing ties.

Based on the construction type, a social network graph can be distinguished into two groups, which are ego-based network and complete network (Hanneman & Riddle, 2005). In an ego-based network, the creation of the graph is started from a set of predefined entities (ego). Then, the ego's neighbors (alter) are added to the graph. The process continues by examining the neighbors' alters and so on. This type of network places ego as the center of community. The second type, which is the complete network, is relatively simpler because it does not require defining a set of egos in the beginning of network creation. It is simply adding all network actors along with their ties to build the graph.

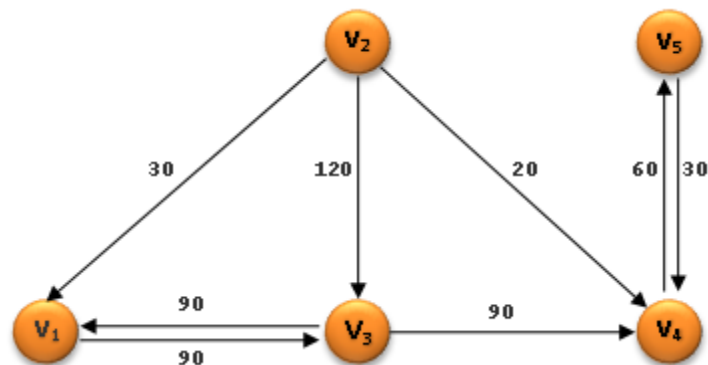


Figure 1. Social network graph

2.1.1. Nodes

As mentioned previously, a social network graph consists of a set of unique vertices or nodes $V = 1, \dots, n$. Each node corresponds to an actor or an interacting component within the network. The node can have 0 or k attributes that represents its characteristics.

2.1.2. Edges

The set of edges E is the collection of $m = |E|$ ties between network actors. The presence of a tie between two actors is indicated by an edge e_{uv}^t with $u, v \in V$ and $u \neq v$ (no self-loop). t can be used to identify an edge if there are duplicate edges between two nodes. t could simply be a numeric identifier

($t = 1, \dots, l$) but it could also be other types, such as timestamp when the edge is introduced. The notion e_{uv}^1 and e_{uv}^2 , for example, represent two identical edges of node u and node v . We can ignore the notion t if there are no duplicate edges within the graph.

Two nodes that share an edge are called neighbor or adjacent. In general, an edge could be either directed or undirected. The direction of an edge corresponds to the orientation interaction of the paired nodes. In a directed graph, an edge e_{uv} refers to an interaction initiated by node u towards node v but not vice versa. Since the paired vertices in directed graph are ordered, the notion e_{uv} in the directed graph is not the same as e_{vu} . Both edges could exist within the same graph.

The graph's edges can either be weighted or unweighted. w_{uv}^t corresponds to the weight of the edge e_{uv}^t . The weight of an edge could imply the tie strength, the interaction intensity or the distance between the paired nodes. If there are no duplicate edges, we can simply abandon the notion t . In the directed graph, the weight of an edge of u towards v does not correspond to the weight of edge from v to u ($w_{uv} \neq w_{vu}$). Besides having weight and direction, the graph's edges can also have attributes, which describe its characteristics, such as the edge type or the time frame when the interaction occurs. As mentioned previously, several identical edges could exist between the same pair of nodes. For our research purposes, we treated duplicate edges between two nodes as a single edge, by aggregating the weight values. The aggregation method applied in this research is explained in the chapter 3.

2.1.3. Network dynamics

The graph structure is dynamic, which means that it can either grow or shrink depending on how the graph evolves. Thus, nodes and edges could be introduced or removed within the lifetime of the network graph. If a certain node is removed, the corresponding edges that connect it with neighboring nodes are also eliminated. Similar to nodes and edges, weight can also change over time.

2.1.4. Graph centrality: Degree

The simplest method to calculate the centrality of a node is by assessing its degree (Wasserman & Faust, 1994). The degree of a node denotes the count of the ties shared with other nodes in the network. A node with degree of 0 is called an isolate, while a node with degree of 1 is known as a pendant. The degree centrality reflects the probability of a node influencing its adjacent neighbors. In a friendship network, the higher the degree, the more popular a node is. This type of node can act as a central hub because it connects many nodes in the network.

The degree centrality of a node in an undirected graph can be calculated by dividing the actual degree of that node by the maximum number of degree that the node could have. It can be formally formulated as follows:

$$C_D(u) = \frac{\text{degree}(u)}{(n - 1)}$$

$C_D(u)$ is the normalized degree of the node u , $\text{degree}(u)$ is the degree of the node u , and n is the total count of nodes in the network ($n = |V|$). In the directed graph, degree centrality can be calculated separately for incoming and outgoing edges and thus resulting in the so-called in-degree and the out-degree, respectively. This degree centrality metric is suitable for large networks because it requires only simple computation. It takes into account only the local view of the adjacent neighbors.

2.2. Telecom social network

Telecom Call Detail Records (CDRs) can also be seen as a form of a social network. Telecom service information retrieved from the CDR, such as the incoming and outgoing phone number, voice calls duration, timeframe, SMS and MMS count, can be mapped as a social graph with nodes denoting subscribers and edges describing relationships among subscribers. Similar to the definition of social graph above, the building blocks of telecom social network $G = (V, E)$ also consists of a set of nodes and a set of edges, which respectively refer to subscribers and the interaction flows among them (see Figure 2). The structural properties of the graph could provide a fundamental insight into the communication pattern and customer sentiment within the network. For example, it could reveal the calling patterns of subscribers with their social circle and it could also show how information is distributed throughout the network. These can be useful insights for the formation of offensive strategies, such as up-/cross-sell targeting campaign (Kiss & Bichler, 2008) or defensive strategies, such as identifying the influential churners (Dasgupta et al., 2008).

2.2.1. Nodes

The nodes in the telecom graph denote all subscribers from the local/observed mobile network provider (on-net) and subscribers from other provider (off-net), who are having interactions with the local subscribers. The profile information of the local subscribers can be assigned to the corresponding node as node attributes. These characteristics describing attribute can range from 0 up to hundreds of variables (e.g., subscription type, payment, socio-demographic, network experience log). Nodes with no known variables refer to subscribers of other telecom networks. Since we focused only relationships within prepaid and postpaid segment in this research study, we did not include the off-net subscribers while constructing the social graph.

2.2.2. Edges

An edge is established if there is a communication between two subscribers, such as a voice-call, an SMS or a MMS. In a directed network, an edge is directed from the source of the calls (caller) and ends in the destination (callee). The edge weight can be calculated from one variable or a combination of interaction variables, for example voice call duration or SMS frequency. The weight could provide an indication of the interaction intensity and how strong the relationship between two nodes is.

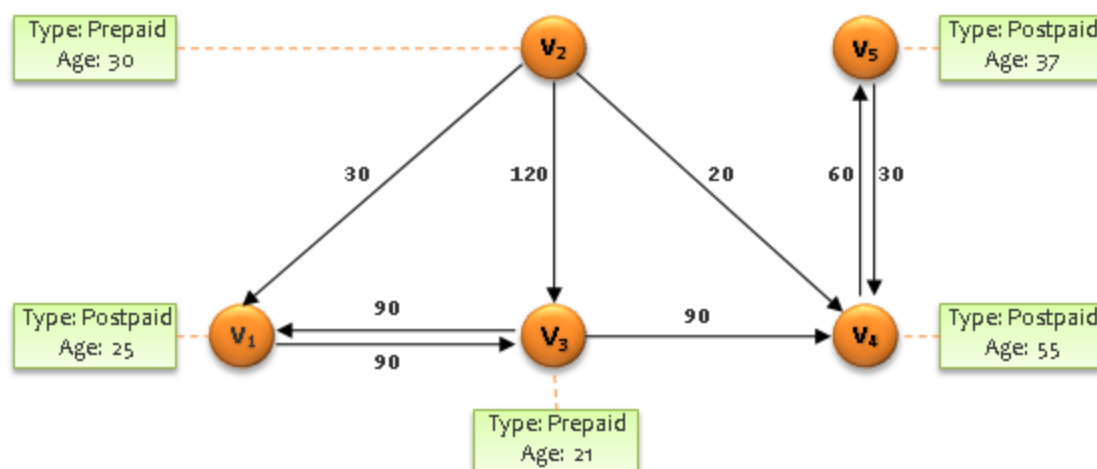


Figure 2. Telecom social graph

2.2.3. Network dynamics

A telecom network graph is constructed from records of subscribers' interactions happening over time. A single edge can be constructed from several mobile communications between a pair of subscribers within the observation period. As a dynamic network, nodes and edges can be added to or removed from the graph. A change on the subscriber's characteristics, such as a new subscription, reflects to an update on the node's attributes. Depending on the goal of the research study, it might be necessary to update the edge's weight because the intensity or duration of the interactions between a pair of nodes can change over time.

Different techniques can be applied to integrate weight dynamics on the graph. A simple approach would be summing up the existing weight with the new one. The time element could also be considered to determine the new weight, for example by emphasizing the importance of the more recent event. In the case of two interactions with the same strength, the more recent one will account more to the total edge weight.

2.3. Machine learning

The following section describes the algorithms, i.e., scoring and spreading activation, which were applied to build churn models in the prepaid as well as the postpaid segment. The scoring algorithms were used for training the feature-based models, whereas the spreading activation algorithm was applied to build the churn propagation models.

2.3.1. Scoring algorithm

The scoring algorithm, which is classified as a supervised machine learning technique, learns patterns and constructs a statistical model from a labeled training data, e.g., churner (1) or a non-churner (0). We applied Logistic Regression and the CHAID Decision Tree algorithm to derive churn models, which estimate the probability of a subscriber to churn. These churn probability scores are distributed into score intervals (Chordiant, 2009). The upper interval groups contain more subscribers with high churn propensity behavior compared to the lower interval groups. Therefore, with this scoring technique, marketers can easily determine their target groups. Since the scoring algorithm returns only the churn probabilities, we could define a cut-off point to perform a classification. Using the threshold score-based technique, the subscribers with churn scores above a predefined threshold score can be labeled as 'churner'. As an alternative, the cut-off point can also be determined by specifying the target group size.

The Chordiant Predictive Analytics Director software was utilized for preparing the variables prior to modeling and also for training the scoring models. The software provides automated computation for data preparation, such as variable discretization and grouping. The discretization process of a given numeric variables begins by grouping the numeric values into a number of bins (default = 200) with an equal number of cases per bin. The bins with statistically insignificant differences in behavior are then merged resulting in a fewer number of bins. The values within a bin have more similar behavior or a homogeneous churn rate compared to values of neighboring bins (Chordiant, 2009).

After discretization process, the variables are grouped and ranked based on their statistically significant differences. A more robust model can be achieved because the model's inputs do not provide the same information. Chordiant also indicates the predictive performance of the variables. Therefore, we can exclude variables with a very low predictive power (e.g., variables with no variation in values) and the ones with extremely high predictive power (e.g., variables with all unique values).

2.3.1.1. Logistic Regression

Rather than having continuous target variables as in linear regression, the logistic regression algorithm generates an inferential function to fit a binary or dichotomous output variable. The idea of both algorithms is nevertheless similar. Both predict a class output by calculating a score of a linear function that is constructed from a set of selected predictors with predefined coefficients. A simple linear regression works as follows: it sums the multiplication of the selected variables with the corresponding coefficients with a goal to minimize the error between the newly calculated Y and the actual known output. Suppose that there are n input variables. The linear regression equation is expressed as

$$Y = C_0 + C_1x_1 + C_2x_2 + C_3x_3 + \dots$$

Here, Y represents the numeric output, C_i represent the coefficients and x_i symbolize the input variables (Witten & Frank, 2011).

Suppose that there are two classes with binary representation, 1 for churner and 0 for non-churner. Although linear regression can be used to produce a function that fits these two classes, the score Y can fall outside the valid 0-1 range. Therefore, logistic regression is more suitable for handling this type of binary classification problem. Logistic regression applies the logistic transformation function in order to measure the membership probability of an instance to a certain class.

$$\log\left(\frac{Y}{1-Y}\right) = C_0 + C_1x_1 + C_2x_2 + C_3x_3 + \dots$$

The logit transformation can take infinity until negative infinity input but still confine the output value between 0 and 1. The resulted model after applying logit transformation is defined as follows:

$$Y = \frac{1}{1 + e^{(C_0 - C_1x_1 - C_2x_2 - C_3x_3 - \dots)}}$$

Instead of selecting models with the lowest sum of squared errors, the model fit of the logistic regression algorithm can be calculated by using the maximum likelihood approach. The algorithm iteratively estimates the coefficient value of the input variables which can maximize the model likelihood.

2.3.1.2. Decision Tree

Decision tree learning is a divide-and-conquer method of predicting a class value from a set of independent variables (Witten & Frank, 2011). This learning method represents the classification model in a tree-like structure with leaves indicating the class value and branches representing the splitting conditions of the input variables. The construction of a decision tree is started from selecting a predictor variable as the root node and then continued by splitting it to create branches. The process is repeated until the branch node contains instances with a single target value, in our case either churner or non-churner.

There are many splitting measures that are used to determine the purity or significance of the split, for example Gini Index, Information Gain ratio and Chi-square value. In this project, we utilized the CHAID decision tree algorithm to classify the churners and non-churners class. Chi-squared Automatic Interaction Detection (CHAID) is a decision tree algorithm that uses the Chi-Square statistic method to select the significant split point of the tree. Ryan (n.d.) explains how a Chi-Square test (*Pearsons X^2*) and a *p-value* evaluation are applied to all potential splitting conditions in order to measure the goodness of fit. A split is constructed if the corresponding chi-square and p-value is considered to be

statistically significant. The CHAID algorithm, however, is not suitable for training model with a small sample set.

The following example illustrates how to statistically calculate the goodness of fit of a split for a variable GENDER, which has 2 values, male and female.

	Churner	Non-Churner
Male	a = 10	c = 550
Female	b = 20	d = 450

Table 1. Contingency table of GENDER and CHURN variable

The chi-square value of the mentioned split can be calculated as follows:

$$X^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

Obs is the actual count of the correlated values in the dataset (see Table 1: Male_Churner, Female_Churner, Male_Non-Churner and Female_Non-Churner). *Exp* represents the expected count of those correlated values. In the contingency table, *Exp* can be calculated by multiplying the row total by column total and divided by the total population. For example, the *Exp* value of Male_Churner is as follows:

$$Exp(Male_Churner) = \frac{(a + b)(a + c)}{(a + b + c + d)} = \frac{30 * 560}{1030} = 16.31$$

After calculating all *Exp* values, we can determine the X^2 . The X^2 value of the split for our example case is 6.13. Next, we need to determine the degrees of freedom *Df* and significance level α .

$$Df = (row\ count - 1)(column\ count - 1)$$

For our example, this gives $Df = (2 - 1)(2 - 1) = 1$. We assume the significance level $\alpha = 0.05$. Next, we can look up the corresponding p-value of our calculation result in the Chi-Square distribution table (see Table 2). Our X^2 value lies between p-value 0.02 and 0.01. Since it is less than the significant level ($p\text{-value} < 0.05$), we can conclude that there is a relationship between GENDER and CHURN. We can test other variables using the same criteria and choose split with the lowest p-value.

<i>Df</i>	<i>p-value</i>					
	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

Table 2. Chi-Square distribution table (Ryan, n.d.)

2.3.2. Social Network Mining using Spreading Activation

Spreading activation algorithm simulates how brain neurons spread information to each other (Collins & Loftus, 1975). A brain neuron that receives information transmits it down to its neighboring neurons through the axons. The information is then spread out throughout the network in a cascade manner. In

the diffusion process, the propagated information is typically defined as energy or weight. Spreading activation algorithm is first used in cognitive psychology to study semantic networks (Collins & Loftus, 1975) and is later used for other research purposes, such as trust propagation (Ziegler & Lausen, 2004).

The spreading activation consists of two processes, activation and spreading (Rodriguez, 2011). The process is initiated by activating a set of predefined source nodes ("activation"), which then propagate a fraction of their energy out to neighboring nodes ("spreading"). If the neighboring nodes receive energy more than a predefined threshold value, it will be enabled or activated for propagation in the next cycle. The process continues to the next adjacent nodes and so on. The propagation process carries on until a termination condition is reached, i.e., the network reaches convergence and/or number of maximum iterations is exceeded.

In this research study, we applied the spreading activation to measure how churn is diffused around telecom social network (Dasgupta et al., 2008). Differs from other popular SNA algorithms, such as PageRank or HITS algorithms, the spreading activation algorithm does not compute the centrality or the social influential power of entities within the network. Since it measures how much influence entities received from its social network, it is more suitable for our research purposes. PageRank or HITS algorithm can be used for other research studies, such as cross/up-sell targeting and product diffusion campaigns, which focus more on identifying the influential network members within the social network (Nanavati et al., 2006).

The churn propagation process begins by initialization of all nodes. The initial energy of the churned nodes is set to 1. In this study, we set the energy of non-churners using two different values (see Figure 3). In the simple or traditional propagation approach, the initial energy of non-churners is set to 0; in the extended approach it is set to the churn score previously calculated using featured-based scoring techniques. Detailed explanation of both methods is presented in the next chapter.

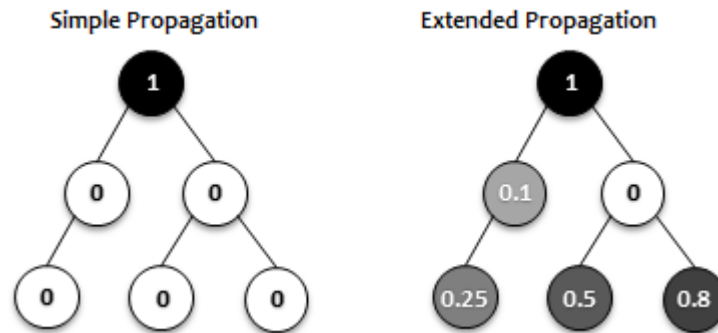


Figure 3. Initial energy of the simple and extended propagation technique

A node is denoted by x and $x \in V$. The $En(x)$ represents the amount of energy of a node is currently having and the $En(x, i)$ represents the amount of energy or social influence transmitted to the node x via one or more its neighbors (Dasgupta et al., 2008). After energy initialization, a set of previous churners (seeds) is activated. In this first activation stage, the current energy of the seeds $En(x)$ is used as initial spreading value. Therefore, the current energy value $En(x)$ becomes 0 and amount of energy in a node x at step 0 or $En(x, 0)$ becomes 1.

The activated nodes are then transferred the portion of their energy to the adjacent neighbors and keep certain portion for itself. The spreading factor δ controls the proportion of the transmitted energy $\delta \cdot En(x, i)$ and the retained energy $(1 - \delta) \cdot En(x, i)$. A spreading factor value $\delta=0.8$ means that 20% of the activated energy is retained by the node and 80% of that energy is transferred to the neighboring

nodes. This factor value could also be seen as a decay measure because the transferred energy will decline as it gets further away from the source. It implies that the direct neighbors will receive more influence than second degree neighbor and so on. In the trust propagation study conducted by Ziegler and Lausen (2004), it is shown that people tend to trust individuals trusted by own friends more than individuals trusted only by friends of friends.

Since nodes can have multiple neighbors, the amount of the distributed energy from an active node to each neighbor depends on the tie strengths between the node pair. On Figure 4, for example, the amount of energy transferred from node 1 to node 2 might not the same as the one transferred from node 1 to node 3 because the edge weight w_{12} is not equal to the edge weight w_{13} .

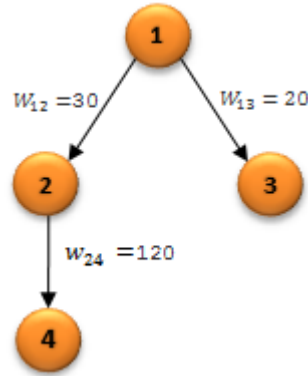


Figure 4. Spreading activation in a weighted graph

The amount of transferred energy depends on the relative tie weight of the paired nodes. This is determined by a transfer function F . Let y be a neighboring node of an active node x and $x, y \in V$. The amount of energy transferred through the edge e_{xy} is formulated below.

$$En_{x \rightarrow y} = \delta \cdot En(x, i) \cdot F_{xy}$$

The amount of energy of node x after spreading computation is as follows:

$$En(x) = En(x) + (1 - \delta) \cdot En(x, i)$$

There are multiple functions to determine the relative weight between two nodes. The simplest method is using linear edge weight normalization function (Ziegler & Lausen, 2004).

$$F_{xy} = \frac{w_{xy}}{\sum_{e_{xz} \in E} w_{xz}}$$

Here, F_{xy} denotes the relative weight of the edge e_{xy} , w_{xy} represents the weight of the edge e_{xy} , the $\sum_{e_{xz} \in E} w_{xz}$ represents the total weight of all edges connecting node x to its adjacent nodes.

In addition to the linear weight function, Ziegler and Lausen (2004) also proposed a non-linear weight method to calculate transferred energy. This new function introduces penalty to edges with small weight. The energy transferred between nodes with weak relationship will be smaller than using simple linear normalization. Moreover, it transmits more energy between nodes with strong relationship.

$$F_{xy} = \frac{w_{xy}^2}{\sum_{e_{xz} \in E} w_{xz}^2}$$

We propagated the churn energy through a directed and an undirected graph. In a directed graph, energy is propagated only to outgoing edges. For churn propagation, the remaining energy after termination determines the probability of a network member to churn.

Initialization:

- Define churners as the node seeds
- Assign initial propagation value, for example energy value 1 for the seeds and 0 for the rest of the nodes.
- Define the spreading factor δ , range 0.0 – 1.0
- Define a threshold value th , e.g., $th = 0.01$

Steps:

1. Activate the node seeds
2. Transfer energy of the activated node to the direct neighbors.
3. Mark neighbors with a new activation energy more than the firing threshold th for activation in the next spreading activation cycle
4. Go to step 2 unless there are no more nodes that can be activated or the spreading activation cycle has reached the maximum iterations number.

To demonstrate the spreading activation process, we will calculate the energy distribution based on the Figure 4. We assign node 1 as the seed node with energy value of 1. The rest of the node has initial energy of 0. In the initial activation of the seed node, we can calculate the amount of energy that can be distributed to the neighboring nodes, which is node 2 and node 3. With a spreading factor $\delta = 0.8$, $th = 0.01$, we can calculate the distributed energy $En_{1 \rightarrow 2}$ and $En_{1 \rightarrow 3}$. The linear edge weight normalization function is applied in this exercise.

$$En_{1 \rightarrow 2} = \delta \cdot En(1,0) \cdot F_{12} = 0.8 \cdot 1 \cdot \frac{20}{20+30} = 0.32$$

$$En_{1 \rightarrow 3} = \delta \cdot En(1,0) \cdot F_{13} = 0.8 \cdot 1 \cdot \frac{30}{20+30} = 0.48$$

The energy retained in the node 1 is as follows:

$$En(1) = En(1) + (1 - \delta) \cdot En(1,0) = 0 + 0.2 \cdot 1 = 0.2$$

Since the activation energy received by node 2 and node 3 is more than the threshold th , node 2 and node 3 are activated and the spreading computation will take place for these nodes.

3. Implementation setup

This chapter describes the implementation setup and the churn models used in this research study. The first section presents the operational definition of churn for the postpaid and the prepaid telecom segment. Next, the dataset description and the data preparation activities taken prior to data modeling are also covered. The last part presents the description of the seven different scenarios that were used to construct the churn models.

3.1. Operational definition of telecom churn

There are two segments in the mobile telecommunication market, postpaid and prepaid. The postpaid subscribers have a service contract arrangement with a mobile service provider. They can typically only churn after their contract has expired. In contrast to postpaid subscribers, prepaid subscribers are not bound by a contract, which make it easier for them to churn. Therefore, the operational definition of prepaid and postpaid churn should be constructed separately. Beside prepaid and postpaid churn, there are generally two general distinction of churn, voluntary and involuntary. Unlike the voluntary churn, where the subscribers initiate the action to leave to other competitor or stop using provided services, involuntary churn occurs when the company terminates the subscriber contract. Involuntary churn occurs due to payment problems or fraud activities.

For the postpaid churn study, we considered the disconnection date, which is the date when the subscription is deactivated from the network, as the churn date. We did not distinguish voluntary from involuntary churn in this segment because involuntary churn does not occur frequently.

Postpaid churn operational definition: Postpaid subscribers are marked as churners the moment when they are disconnected from the network.

Similar to churn definition in the postpaid segment, the disconnection date was considered as the churn date in the involuntary prepaid churn. A prepaid subscriber was marked as a churning and he/she was disconnected from the network after 6 consecutive months of inactivity. For our purpose, the prepaid activity was defined as one of the activities below:

- Outbound voice call
- Inbound voice call
- Outbound SMS
- Data usage
- Commercial voucher recharge, also known as top-up

As churn should be detected as early as possible, the disconnection date might not be the appropriate churn date measure (Kraljevic & Gotovac, 2010). The prepaid subscribers might be long gone before they are actually disconnected from the network. Therefore, we used a new operational definition for the prepaid churn and considered a smaller number of the inactivity months, which is two months.

Prepaid churn operational definition: Prepaid subscribers are marked as churners after two consecutive months of inactivity.

3.2. Dataset

3.2.1. Population and outcome definition

For this study, the experimental datasets were obtained from one of the largest telecom provider in the Netherlands. The dataset collection is presented on Figure 5. We used the Call Detail Records from the whole month of February 2012, which is roughly about 700 million records, to construct the social graph. The subscribers exploited in the research study, no matter whether it is a churning or a non-churner, were all active during the CDR recording period. They should have at least one CDR data in the chosen time range because we based our social network graph on the interactions occurred in that month.

The traditional predictors of the on-net subscribers or subscribers of the local telecom provider were also obtained in the same month. The churn models could also be constructed using dataset incorporating traditional predictors of February, March and April 2012. Since we chose not to do so, the observation period is not completely the identical, which makes it rather complicated to compare the models trained using solely the traditional predictors and the ones constructed only using social network information.

We recorded churn for the postpaid as well as prepaid segment in two separate periods, in the end of April 2012 (observation 1) and in the end of June 2012 (observation 2). The observation 1 comprised of March and April's churners, while observation 2 contained churners of May and June. We labeled the nodes/subscribers that are found in the observation 1 as seeds/churners of the social graph. The end goal of the research study is to use the traditional predictors as well as the social network information obtained in February-March-April 2012 to predict churn of June 2012.

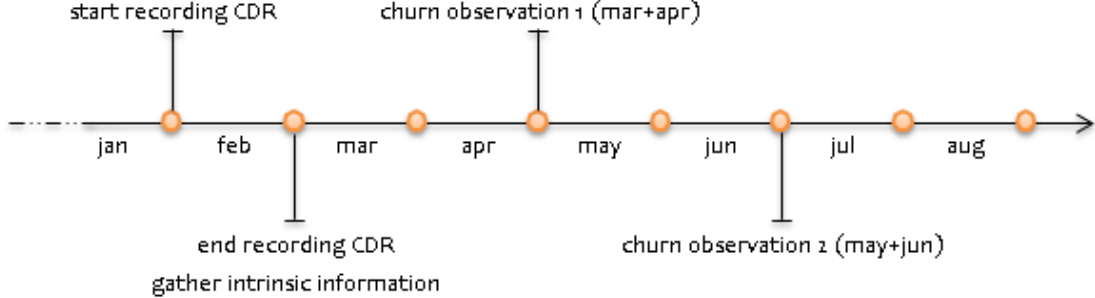


Figure 5. Population settings

For the prepaid segment, we further limited the sample set to only include subscribers registered earlier than 1 December 2011 with intention to eliminate temporary subscribers, e.g., tourists.

In the postpaid segment, the churners count was strongly correlated with the end of contract (EOC) period (Figure 6). Table 3 presents that 49.3% of March 2012's churners has EOC month equal to March 2012. About 61.7% churners of March 2012 are having February 2012 or March 2012 as the EOC month. We concluded that the EOC month is a strong churn predictor in the postpaid segment and it is hard for other variables to compete with this predictor. As consequence, we limited the experimental sample into two different sets based on the end of contract month, EOC2 and EOC6. EOC2 consisted of postpaid instances with EOC of March and April 2012, whereas EOC6 comprised postpaid subscribers with EOC from October 2011 until April 2012.

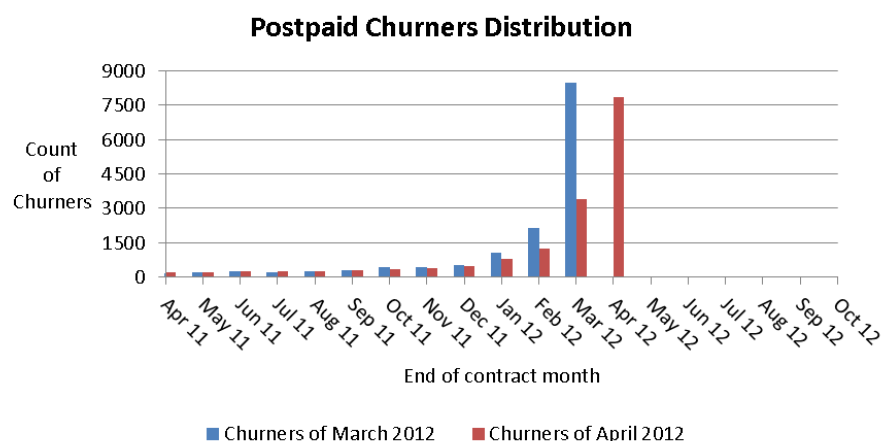


Figure 6. The distribution of postpaid churners based on the EOC month

Churners of March 2012			
EOC month	Number of churners		
...	...		
Oct 2011	417	75.9%	
Nov 2011	430		
Dec 2011	523		
Jan 2012	1077		
Feb 2012	2133		
Mar 2012	8495	49.3%	61.7%
Apr 2012	41		
...	...		
Total	17721		

Table 3. Churned subscribers on March 2012 for selected EOC month

Table 4 describes the dataset sizes and the actual churn rate measure in June 2012 for each sample set. The modeling dataset contains subscribers that are active during the predictor period. 50% of the dataset sample was used as the training set, 25% as the validation set and the rest 25% as the test set.

Sample set	Total rows	Churn rate based on the 2 nd observation churners (class output) on		
		Training set	Validation set	Test set
Prepaid	387762	5%	5%	4.88%
Postpaid EOC6	151690	3.44%	3.32%	3.41%
Postpaid EOC2	74111	3.45%	3.64%	3.61%

Table 4. Churn rate of the sample sets

3.2.2. Call Detail Records

As mentioned previously, we obtained the social network information from telecom Call Detail Records. A single CDR contained detailed information about the incoming or outgoing voice call, SMS and MMS as well as the volume of the data/internet usage. A data CDR only contains volume information and it does not record internet activities information, such as visited websites or downloaded contents. In this research study, we focused only on the voice call and SMS. We could not explore mobile interactions

utilizing the data connection, i.e., using over the top (OTT) service, due to legal issues (Schellevis, 2011). In addition, the MMS interactions were excluded as there were relatively a small number of records.

A single CDR for a voice call or a SMS contains the following information:

- A source MSISDN or a phone number
- A destination MSISDN number
- Type of communication (Voice or SMS)
- Timestamp of the mobile communication
- Duration of the call in seconds (applicable for voice call)

The raw CDRs consisted of average 25 million records daily with approximately 12.7 million individuals in total. It contained 3.5 million local network subscribers and 9.2 million subscribers from other network providers. Processing dataset of this magnitude was complex and it required a lot of computing power. Due to limited resources, we focused on exploring the role of postpaid and prepaid churners within the local network and the study was constrained only to local subscribers.

For our analysis, the CDRs were grouped by directed edge such that only distinct edges present within the telecom graph. Recall that an edge of a graph G was denoted as e_{uv}^t and the corresponding weight as w_{uv}^t (edge: $e_{uv}^t \neq e_{vu}^t$ and weight: $w_{uv}^t \neq w_{vu}^t$). The edge weight w_{uv}^t indicated the intensity and duration of interactions of the pair. The identifier t identified, in our case, the hourly timestamp when the interaction starts. t ranged according to our CDR's observation period, which starts from 1 February 2012 until 29 February 2012.

By assuming that a text-message was equivalent to 30 seconds of voice calls, we could generalize the edge weight to include both types of mobile communication, voice calls and SMS. Therefore, all interactions could all be measured uniformly in seconds.

$$w_{uv}^{t'} = w_{uv}^t \cdot \begin{cases} 1, & \text{if the interaction type is voice call} \\ 30, & \text{if the interaction type is SMS} \end{cases}$$

Furthermore, during preprocessing, we took into account the start time of the interactions. Interactions that occurred outside working hours were assigned twice the weight to emphasize the importance of it. The underlying assumption here is that interactions within working hours denote communication of professional nature, while interactions outside working hours denote communication of more personal nature (e.g., friends, family), which could have higher influence on the decision to churn. Therefore, we introduced a weight scale $\rho(t)$, which is defined as follows:

$$\rho(t) = \begin{cases} 1, & \text{if } t = \text{weekdays between 08:01} - 17:00 \\ 2, & \text{otherwise} \end{cases}$$

$$w_{uv}^{t''} = \rho(t) \cdot w_{uv}^{t'}$$

Besides taking into account the time of day of the interaction, we also assumed that a recent interaction should carry more weight than older ones. Therefore, the daily decay rate α was introduced to the dataset. The weight value of an edge that was measured in a certain day exponentially decayed according to a predefined rate as follows:

$$w_{uv}^{t'''} = w_{uv}^{t''} \cdot e^{-\alpha \cdot i}$$

Here, the symbol i corresponds to the gap measured in days between the interaction timestamp and the end of the observation period. In our case, i is equal to 28 measured from 1 February until 29 February

2012. If we want to measure the weight value in the end of 7 February 2012, we can simply use $i = 6$ in the equation. Figure 7 demonstrates the effect of implementation of weight decay between 2 nodes for 7 days. There are 4 interactions occurred during those 7 observation days. It is illustrated that the weight value decreased exponentially per day with a decay rate $\alpha = 0.2$. Applying this decay rate, only 127 seconds remained for the interaction occurred in the 1 February 2012 after 7 days.








weight on day =	1-Feb	2-Feb	3-Feb	4-Feb	5-Feb	6-Feb	7-Feb
							
interaction on 1 Feb	420	344	282	231	189	155	127
interaction on 2 Feb	-	90	74	60	49	40	33
interaction on 3 Feb	-	-	-	-	-	-	-
interaction on 4 Feb	-	-	-	450	368	302	247
interaction on 5 Feb	-	-	-	-	30	25	20
interaction on 6 Feb	-	-	-	-	-	-	-
interaction on 7 Feb	-	-	-	-	-	-	-
daily sum of weight	420	434	355	741	637	521	427

Figure 7. Illustration of weight decay applied on 7 observation days between a pair of nodes

In the end of the observation period, the weight values are aggregated. As a result, each node pair has only one edge per direction. The equation below formulated the aggregation process of the weight values.

$$w_{uv} = \sum_t w_{uv}^t'''$$

For an undirected graph, we could simply add up the weights for both directions together as follows:

$$w_{uv} = \sum_t w_{uv}^t''' + \sum_t w_{vu}^t'''$$

3.2.3. Subscriber information

3.2.3.1. Intrinsic variables

The subscriber intrinsic variables, also referred to as traditional predictors, that includes demographic, contractual, handset and service usage were collected in the end of February 2012. The service usage variables included the monthly aggregated information, such as duration of calls, SMS count and mobile data volume consumption. In general, distinguishing on-net nodes from the off-net nodes within the telecom graph was quite straightforward. The properties of the on-net nodes were known, while the off-net nodes contained no properties. A dimension reduction technique, correlation matrix, was performed as to eliminate redundant and irrelevant variables in the dataset. The selected attributes (appendix B1) could be categorized as follows:

- Demographic characteristics, such as age
- Contractual information, such as type of subscription and package plan
- Handset information, such as handset model and manufacturer
- Service usage, such as voice call duration, SMS count and data usage
- Churn identification, i.e., churner or non-churner

A churn identification variable was added to classify whether the specific subscriber is a churner or a non-churner in a point of time. It was defined as a binary flag, where 1 represented a churner and 0 represented a non-churner. The analysis was based on both postpaid and prepaid churners. Since all subscribers should be active on February 2012, postpaid churn was recorded monthly starting from March and prepaid churn was captured from April 2012.

3.2.3.2. Extrinsic information

Social network properties were extracted from the telecom graph. These properties measured the connectivity of nodes within a social network and described the neighborhood relationships. See Appendix B2 for more detailed description of the features extracted from the telecom graph.

The social network properties could be categorized into the following 2 groups:

- Connectivity represents the interpersonal relationship of a node with its neighbors, namely the number of in-degree and out-degree. The connectivity variables are measured using the CDR information extracted on February 2012.
- Churn connectivity is similar with the general connectivity measures but it focuses more on the relationships with neighbors that were labeled as churned in observation 1. Several examples of the variables are the number of churners in the first degree neighborhood and in the second degree neighborhood.

3.3. Implementation Scenarios

To investigate to which extent social network data could be used to predict churn and possibly could improve prediction performance, we trained 7 different models using scoring algorithms and spreading activation techniques. The implementation scenario is illustrated on Figure 8. Each model was trained using three different datasets mentioned previously, i.e., postpaid EOC2, postpaid EOC6 and prepaid.

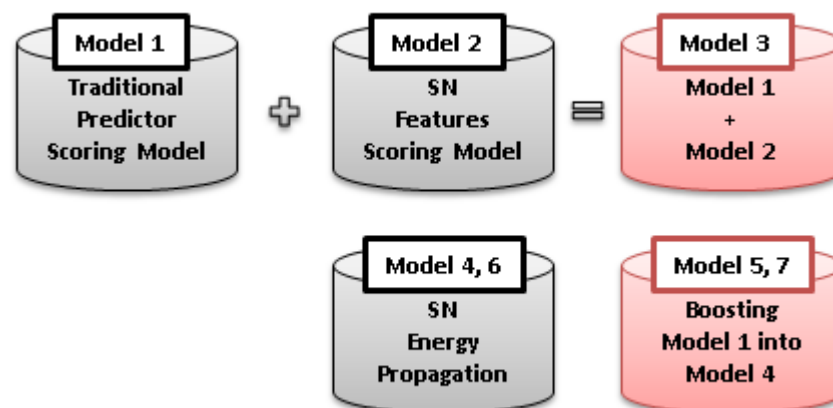


Figure 8. Implementation scenarios

3.3.1. Scoring models

We applied a logistic regression and a CHAID decision tree algorithm to train the scoring models. The models are as follows:

- Model 1: simple scoring model
- Model 2: social network (SN) scoring model
- Model 3: extended scoring model

Model 1, a simple scoring model, was trained using the traditional predictors, such as demographic, contractual, handset and usage information. We employed this model as the benchmark model. *Model 2* was a social network scoring model, which focused solely on the extrinsic social network attributes, such as the number of incoming and outgoing ties of the first and second degree neighbors. The extended scoring model, *Model 3*, combined the dataset of the first and the second model. This last model was learned from both social network features as well as the traditional intrinsic variables.

3.3.2. Propagation models

The remaining four models were trained using energy propagation technique based on spreading activation algorithm.

- Model 4: simple propagation model
- Model 5: extended propagation model
- Model 6: simple propagation model undirected
- Model 7: extended propagation model undirected

Nodes labeled as churned in the observation 1, i.e., March and April's churners, were used as the source or the root of the energy propagation. Each churned node was given an initial energy of 1. The main difference between *Model 4* and *Model 5* is the initial energy value of the non-churned nodes. *Model 4*, which was a simple propagation model, set the initial energy of non-churners to 0. *Model 5* was actually boosting of *Model 1* into *Model 4*. It indirectly incorporated subscribers' intrinsic information into the propagation model. Instead of setting the energy of non-churners to 0, this model assigned the churn score obtained from *Model 1* as the initial energy of the non-churner nodes. The intuition behind this idea is that a subscriber might already have a certain tendency to churn due to his/her experience with the provided service. *Model 6* and *Model 7* were similar to *Model 4* and *Model 5*, except that those models were trained using an undirected instead of a directed graph.

The total energy value, which is remained after termination, was equivalent to the probability of a network member to churn. To study the influential effect of churned neighbors in the social network, we then compared the propensity values of non-churners to the actual/known churn class. The firing threshold of the propagation models was set to threshold $th = 0.01$ and the spreading factor to $\delta = \{0.2, 0.8\}$. Moreover, we tested the linear as well as the non-linear normalization function to determine the amount of energy that should be moved from one node to another.

4. Implementation Results

We next proceed to present the modeling results of the seven different scenarios that presented in the section 3.3 for the prepaid and the postpaid telecom segment. This chapter also discusses the performance and limitation of those seven models.

4.1. Modeling results

The models were implemented for all three datasets, postpaid EOC2, postpaid EOC6 and prepaid. The default evaluation statistic that was used for measuring the performance of the predictors and model was Coefficient of Concordance (CoC). According to Chordiant (2009), CoC measures the area under Lorenz curve formed by the percentage of cases with positive behavior against the percentage of cases with negative behavior for each unique score (see Figure 9). Since we wanted to predict churn, the churn target label was assigned as the positive case, while non-churn label was used as the negative case. Therefore, CoC shows the correctness of the relative ordering of predictions among the churn and non-churn cases. The CoC value varies between 50% and 100% and the larger the CoC value the better.

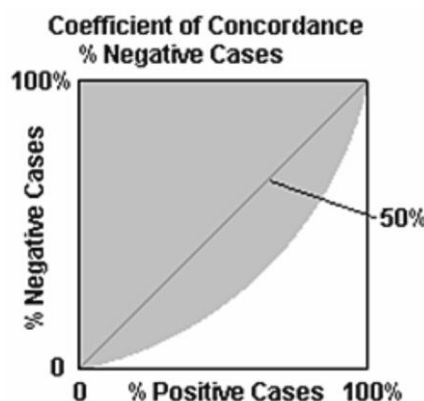


Figure 9. Coefficient of Concordance (Chordiant, 2009)

As mentioned previously, each modeling dataset were divided into three sample sets, training, validation and test set with 50%, 25% and 25% proportion, respectively. The training set was used to build the churn models and the validation set was primarily used to avoid overfitting. We needed to make sure that the models are not too much adjusted to fit the training data only. One simple way to detect whether overfitting has occurred is by constructing two models using the training and the validation set and then assesses their performance. Overfitting has occurred if the predictive accuracy of the model built over the training set increases but the one built over the validation set stays the same or even decreases. In addition, the performance of each model was assessed by using an independent test set that was not used previously to train the model.

For model evaluation, the instances of the test set were ranked based on their probability of being a positive case. The probability scores were then visualized as a gain and a lift chart. The gain chart visualizes the cumulative percentage of positive cases, whereas the lift chart shows the improvement in cumulative behavior over the average behavior (Chordiant, 2009). In the gain chart, the horizontal axis represents the percentage of the cases, whereas the vertical axis depicts the percentage of the true positive cases. This true positives percentage is counted after comparing the probability score of each instance with the actual class value. The gain chart illustrated in the Figure 10 has two lines, a solid and a dotted line. The solid line shows the true positives returned by a model, while the dotted line represents the expected true positives by random guess (baseline). The solid line indicates that the model is able to

correctly identified about 45% churners in the 20% of the sample set. The trained model clearly gives better result than random, which can only predict 20% churner within the same sample set.

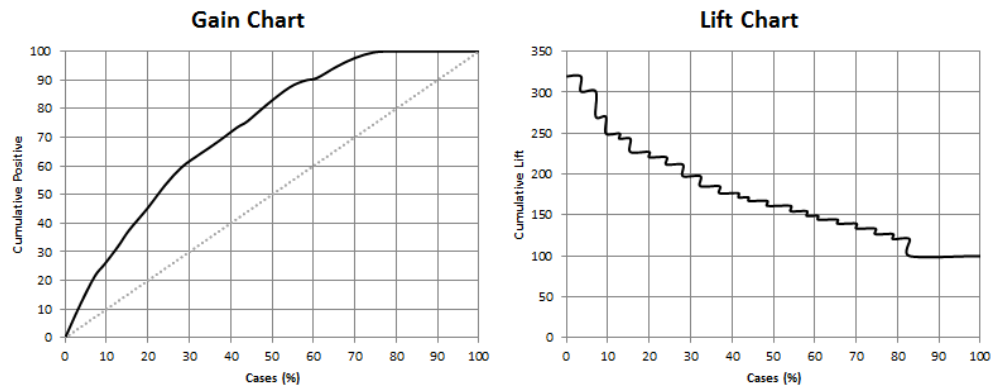


Figure 10. Illustration of a gain chart and a lift chart

Similar to the gain chart, the horizontal axis of the lift chart also displays the percentage of the sample set. The vertical axis maps the true positives ratio of the predicted model and random. Figure 10 shows that the model could return 225 lift in the 20% of the cases, which means that it can correctly predict churn 2.25 times better on that sample set than without using the model.

Many models with different sets of variables were learned from the predefined dataset. However, only models with the best predictive performance are presented in this paper. We discussed only the scoring models using decision tree for both postpaid and prepaid segment because the models have a slightly better predictive performance compared to the ones built using logistic regression. Moreover, we included propagation models with a normalization function that has the best prediction results.

4.1.1. Prepaid dataset

Since one of the research objectives was to assess whether we can predict churn based on previous churning decision of other nodes in the social network, we first should determine whether there is any evidence that showed any correlation between social network data and churn. The figure below shows the ratio of the immediate churned neighbors to the number of adjacent neighbors (degree) and the corresponding churn behavior in the training set.

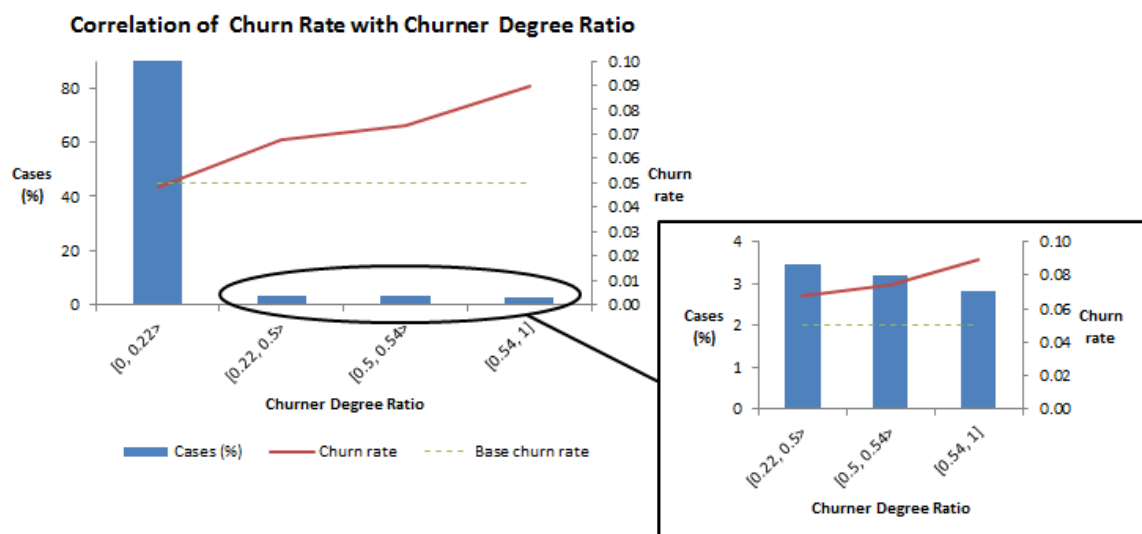


Figure 11. Correlation of churn rate with churner degree ratio

Figure 11 shows that the churn probability of the subscribers in the target period, which is May and June 2012, correlates with the ratio of adjacent churners to the number of the first degree neighbors in the social circle. The baseline churn rate of the training set for the prepaid segment is 5%. The churn rate is higher than the baseline rate once the churning degree ratio is higher than 22%. When the ratio value equals to 22-50%, the churn probability is 1.4 times the base rate. It becomes almost 2 times higher when there are more than 54% churners in the neighborhood. It implies that the social network behavior has a certain impact on the subscribers' churning decision. However, we still needed to justify whether it is sufficient for predicting churn, as about 90% of the sample does not have churned neighbors.

4.1.1.1. Prepaid: Scoring and Propagation Models

	Performance on			Count	Predictor*	
	Training set	Validation set	Test set		Name	CoC (%)
Model 1: Simple Scoring	65.48	64.47	64.88	6	contract_startdate	62.39
					value_segment	57.49
					arpu	56.00
					cnt_mtc	55.70
					cnt_moc_sms	54.19
					amt_moc_data	52.90
Model 2: SN Scoring	57.93	56.72	56.57	4	degree2nd	54.73
					inweight_sum	53.34
					outweight_avg	53.06
					freq_avg_undirect	52.40
Model 3: Extended Scoring	65.65	64.45	64.98	7	contract_startdate	62.39
					value_segment	57.49
					arpu	56.00
					cnt_mtc	55.70
					degree2nd	54.73
					cnt_moc_sms	54.19
					amt_moc_data	52.90

*see Appendix C for further description of the predictors

Table 5. Performance of the prepaid scoring models

For the scoring models, Model 3 has the highest CoC score, which is 64.98%, whereas the CoC score of Model 1 and Model2 is respectively 64.88% and 56.57% (see Table 5). Model 3, which includes both traditional predictors as well as social network variables in the modeling stage, clearly outperforms Model 1 and Model 2. However, since the performance difference between Model 1 and Model 3 is so small (only 0.10 CoC value difference), we can conclude that adding social network features on top of the traditional predictors does not provide any substantial improvement for our scoring model. This implies that social network information extracted from the prepaid CDRs does not substantially improve churn prediction.

Although the initial churn degree correlation analysis showed a promising result that there is a correlation between churn probability and social network, the scoring model based exclusively on the social network variables returns the lowest predictive accuracy for the training, validation as well as the test set.

Next, the performance of the social network propagation models that were constructed using directed as well as undirected graph is presented in Table 6. As expected, the extended propagation models

(Model 5 and Model 7), which incorporate churn scores of the simple scoring model as the initial energy value in the propagation process, outperform the simple propagation models (Model 4 and Model 6). These extended or hybrid models provide better predictive accuracy than the simple propagation models for the directed and the undirected graph.

	Performance on		
	Training set	Validation set	Test set
Model 4: Simple Propagation	53.34	53.43	53.04
Model 5: Extended Propagation	55.26	54.58	55.24
Model 6: Simple Propagation Undirected	52.07	52.15	52.26
Model 7: Extended Propagation Undirected	58.39	57.66	58.30

Table 6. Performance of the prepaid propagation models

The initial CoC value of the extended propagation models before applying the spreading activation algorithm is identical to the Model 1. After spreading, however, the CoC value of the models regardless the graph types, directed or undirected, drops lower than the CoC of Model 1. Also, the simple propagation models based on the study conducted by Dasgupta et al. (2008) have considerably bad predictive performance. Since the propagation process conducted on a telecom graph containing prepaid subscribers only, the propagation results indicates that prepaid subscribers have a low influence among them.

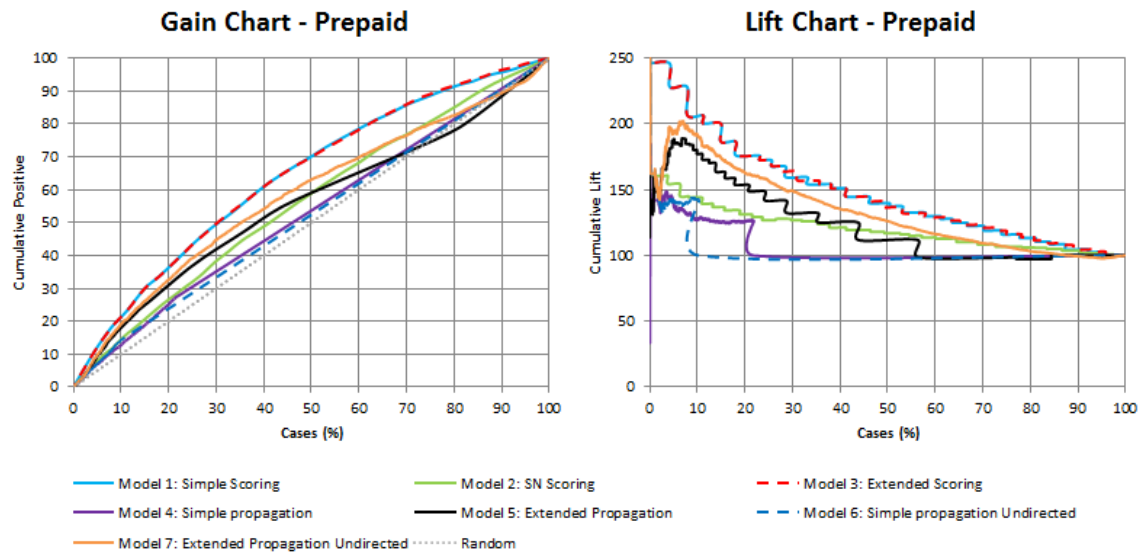


Figure 12. Gain and Lift chart of the prepaid models

From the gain and lift chart in Figure 12, it is clear that social network scoring (Model 2) returns the worse predictive performance compared to the other scoring models. By targeting the top 30% of the subscribers, Model 2 can find only 37% of the churners, while Model 1 and Model 3 are able to return about 50% of the churners. The lift chart shows that Model 2 performs 1.3 times better than random in the top 30% of the cases, whereas other scoring models predict 1.6 times more. Model 3 outperforms Model 1 with the same sample size by 0.05 point.

Targeting 30% subscribers, Model 7 is able to correctly predict about 45% churners. It returns 5% less than the Model 1 and Model 3. Although Model 7 incorporated the traditional predictor elements in the propagation, the predictive power is still lower than the performance of Model 1, the simple scoring model.

The simple propagation models, Model 4 and Model 6, have even lower performance compared to Model 2. Unlike Model 2, the simple propagation model uses only the previous churner information within the social network without considering the individual churn propensity. This leads us to believe that the churning behavior of neighbors does not have enough influential effect on other members within a prepaid telecom subscriber social network.

4.1.1.2. Prepaid Scoring: Predictor Details

The following section discusses the factors contributing to churn in the prepaid segment.

CONTRACT_STARTDATE	Cases (%)	Behavior
[197001,200906>	20.68	0.03
[200906,201005>	30.86	0.04
[201005,201101>	17.88	0.07
[201101,201106>	13.34	0.06
[201106,201108>	9.10	0.07
[201108,201111]	8.14	0.10

Table 7. Prepaid predictor: CONTRACT_STARTDATE

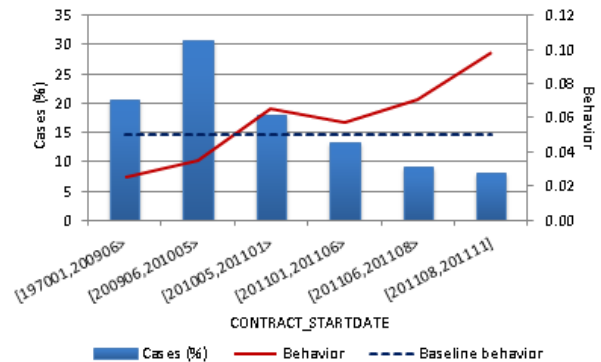


Figure 13. Prepaid predictor: CONTRACT_STARTDATE

The CONTRACT_STARTDATE variable is the strongest predictor in this churn prediction (CoC = 62.39%), followed by VALUE_SEGMENT (CoC = 57.49%). A model built solely using CONTRACT_STARTDATE is able to outperform models constructed from all other variables. From Table 7 and Figure 13 above, we can see that the churn rate of subscribers who started using prepaid subscription from May 2010 onwards is higher than the 5% baseline churn rate for prepaid. About 10% of the subscribers have churned in cases where the start of the subscription is \geq August 2011. This represents the highest churn rate of the entire cases.

DEGREE2ND	Cases (%)	Behavior
[0, 1>	39.08	0.04
[1, 3>	33.55	0.05
[3, 5>	13.06	0.06
[5, 9>	9.19	0.06
[9, 179]	5.12	0.07

Table 8. Prepaid predictor: DEGREE2ND

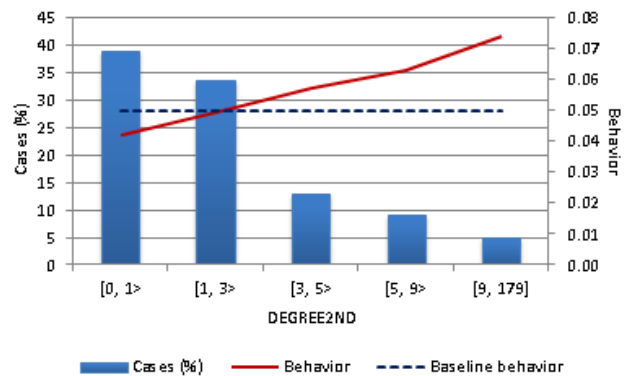


Figure 14. Prepaid predictor: DEGREE2ND

The main difference of Model 3 and Model 1 is variable DEGREE2ND, which symbolizes the count of the second degree neighbors. This variable, which has CoC value of 54.73%, is the strongest predictor out of all other social network features (see Table 8 and Figure 14). Subscribers with at least 3 second degree neighbors are more likely to churn. Although it clearly shows that the number of second degree neighbors aligns with the churn rate value, the DEGREE2ND variable adds an unsubstantial value on the model performance.

4.1.2. Postpaid EOC6 dataset

Since postpaid is always bound by contract, most subscribers choose to churn right on or after their subscription contract is ended. As consequence, a churn model trained using single variable representing the end of contract date could have high accuracy. However, it does not explain why churn occurs and what other variable that could influence it. In this section, we include only subscribers with end of contract subscription between November 2011 and April 2012, which accounts for 75% of all churners measured in the first observation period.

4.1.2.1. Postpaid EOC6: Scoring and Propagation Models

In the Table 9, the result of Model 3 is left blank because the result is the exactly same as Model 1. DEGREE2ND variable, which has the highest CoC value (57.52%) out of all social network features, could not be used to train model because it correlates with the RATEPLAN_GROUP. Correlated variables contain the same information. Hence, adding these variables will only create redundancy. Any other social network features are either correlated with the traditional predictors or having low predictive power. Therefore, the social network features does not improve the model's prediction quality. Model 1 outperforms Model 2 by roughly 10% CoC in all three sample sets.

	Performance on			Count	Predictor*	
	Training set	Validation set	Test set		Name	CoC (%)
Model 1: Simple Scoring	71.47	69.01	69.94	5	rateplan_group	65.04
					contract_end_tolast	60.93
					lifetime	58.69
					handset	57.59
					cnt_moc	53.66
Model 2: SN Scoring	61.13	59.25	59.86	5	degree2nd	57.52
					outweight_sum	54.83
					inweight_avg	53.26
					churner_degree_ratio	52.60
					smsfreq_sum	52.46
Model 3: Extended Scoring						
Model 4: Simple Propagation	50.99	50.45	51.12			
Model 5: Extended Propagation	61.06	60.48	62.57			
Model 6: Simple Propagation Undirected	52.09	52.26	53.21			
Model 7: Extended Propagation Undirected	62.81	63.09	64.19			

*see Appendix C for further description of the predictors

Table 9. Performance of the postpaid EOC6 scoring and propagation models

Similar to prepaid, the extended propagation models, i.e., Model 5 and Model 7, return better accuracy than the simple propagation models, i.e., Model 4 and Model 6. These better results are consistently shown in the training, validation as well as the test set in the **Error! Reference source not found..** Model 1 or the simple scoring model has the best predictive accuracy compared to the rest of the scoring and propagation models. Although churn scores correlate with the churn degree ratio, the majority of the

subscribers do not have churned neighbors. This might be the reason why propagation models consistently have the worst performance compared to the scoring models.

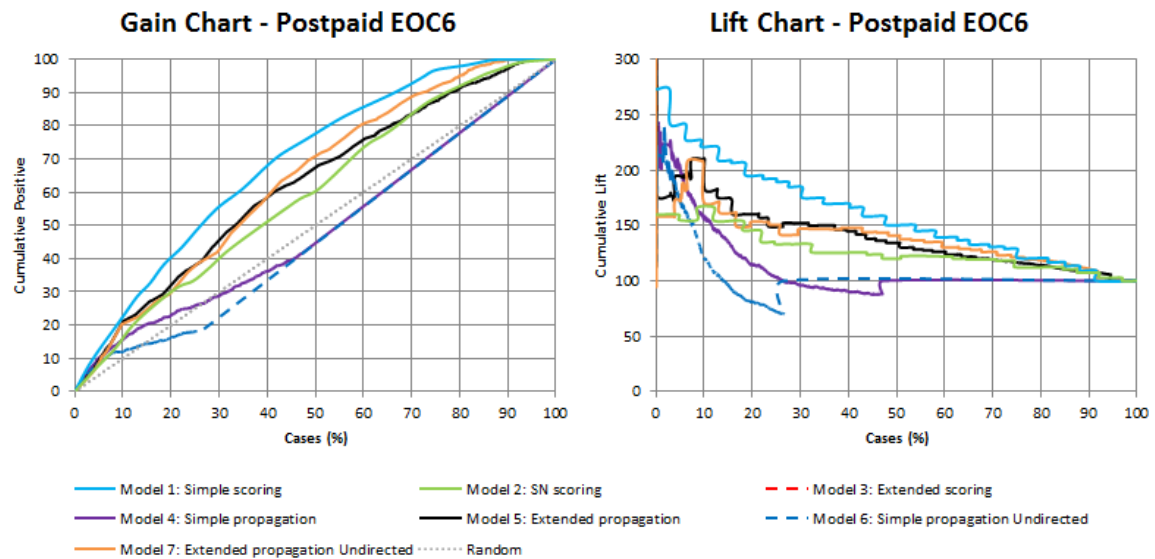


Figure 15. Gain and Lift chart of postpaid EOC6 models

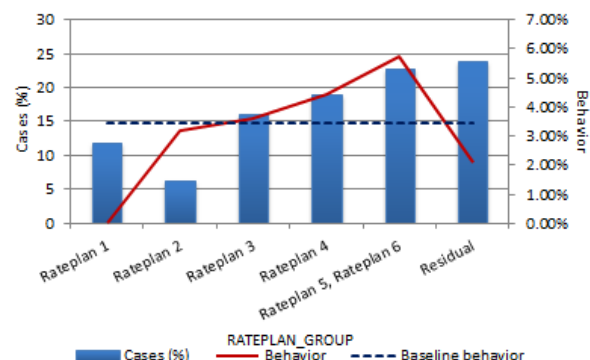
Figure 15 above shows that Model 1 has 55% gain and 180% lift in the top 30 percentile; whereas Model 2 is only able to have 40% gain and 130% lift. By contacting 20% and 30% cases, Model 5 and Model 7 are able to correctly predict about 38% and 55% of future churners, respectively. Model 4 and Model 6 can be considered unstable because they make correct predictions only for a relatively small fraction of the sample. These two simple propagation models have a lift about 200-220% in the top 5 percentile, which are even better than the extended ones. However, there is a big performance decay after the 25 and 15 percentile (Model 4 and Model 6, respectively), where the results are worse than random prediction.

4.1.2.2. Postpaid EOC6 Scoring: Predictor Details

The variable with the highest predictive power for the scoring models is RATEPLAN_GROUP (CoC = 65.04%). Users with Rateplan 4, Rateplan 5 and Rateplan 6 subscription have the highest probability to churn. It is alarming because Rateplan 5 and Rateplan 6 subscription alone comprises for about 23% of the cases.

RATEPLAN_GROUP	Cases (%)	Behavior
Rateplan 1	11.80	0.00
Rateplan 2	6.35	0.03
Rateplan 3	16.13	0.036
Rateplan 4	19.08	0.044
Rateplan 5, Rateplan 6	22.82	0.06
Residual	23.82	0.02

Table 10. Postpaid EOC6 predictor:
RATEPLAN_GROUP



In this postpaid case, the variable RATEPLAN_GROUP indirectly correlates with LIFETIME, which is the length of subscriber's relation with the mobile provider (see Table 11 and Figure 17). It shows that subscribers with LIFETIME value between 22-26 months are having the high tendency to churn. It

supports our earlier observation that subscribers would churn in the end of contract date (typically postpaid contracts have duration of 12 or 24 months).

LIFETIME	Cases (%)	Behavior
[0, 22>	9.88	0.04
[22, 26>	18.19	0.05
[26, 34>	16.18	0.03
[34, 62>	25.12	0.04
[62, 157]	30.64	0.02

Table 11. Postpaid EOC6 predictor: LIFETIME

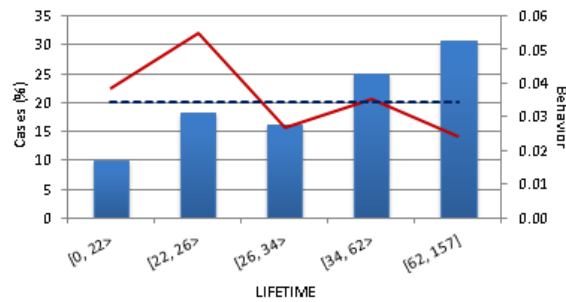


Figure 17. Postpaid EOC6 predictor: LIFETIME

This statement is supported by variable CONTRACT_END_TOLAST, which represents the days between subscribers' last activity date and their end contract date. As shown in Table 12 and Figure 18, in about 20% of the cases, churners are already inactive for more than 40 days before their actual subscription is expired. The missing value can be ignored due to low number of cases.

CONTRACT_END_TOLAST	Cases (%)	Behavior
Missing	0.03	0.05
[-123, -47>	20.87	0.02
[-47, -1>	18.74	0.04
[-1, 8>	8.95	0.02
[8, 22>	13.37	0.02
[22, 39>	18.62	0.03
[39, 50>	9.96	0.05
[50, 1827]	9.47	0.07

Table 12. Postpaid EOC6 predictor: CONTRACT_END_TOLAST

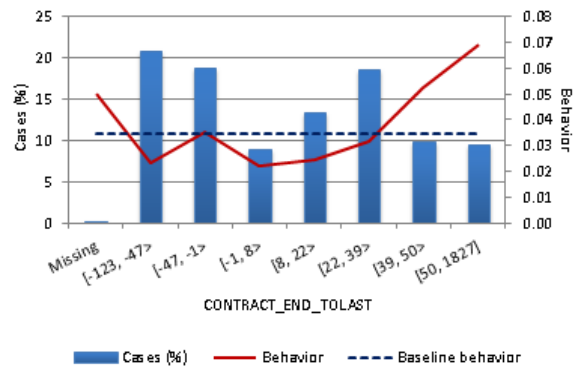


Figure 18. Postpaid EOC6 predictor: CONTRACT_END_TOLAST

4.1.2.3. Postpaid EOC6: Scoring Models without Contract

Churn in postpaid segment is highly related to the end of contract date and subscription type. Therefore, we also trained a model with dataset that excludes the three variables above, RATEPLAN_GROUP, LIFETIME and CONTRACT_END_TOLAST, with intention to investigate other attributes that might contribute to churn.

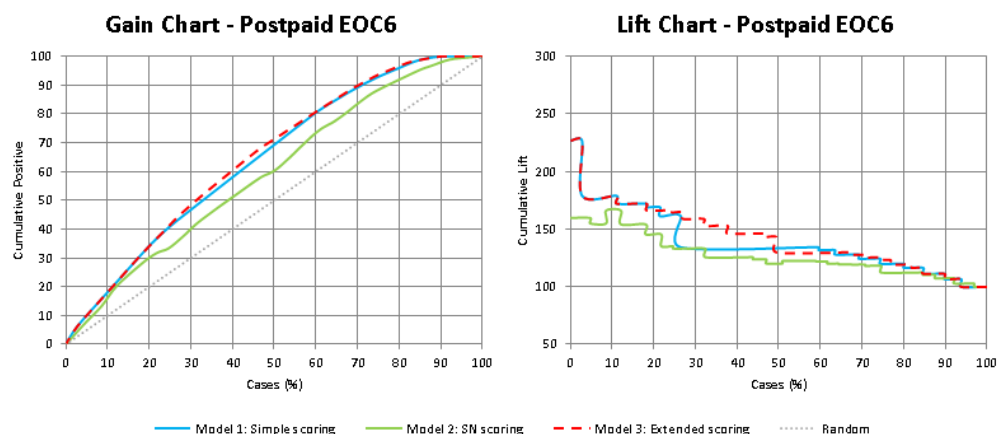


Figure 19. Gain and Lift chart of EOC6 scoring models without contract related variables

	Performance on			Count	Predictor*	
	Training set	Validation set	Test set		Name	CoC (%)
Model 1: Simple Scoring	64.85	63.27	64.82	5	age	59.45
					handset	57.59
					cnt_moc_sms	54.85
					bill_sms_usage	54.84
					interconnect_cost	54.53
Model 2: SN Scoring	61.13	59.25	59.86	5	degree2nd	57.52
					outweight_sum	54.83
					inweight_avg	53.26
					churner_degree_ratio	52.60
					smsfreq_sum	52.46
Model 3: Extended Scoring	65.64	63.42	65.25	6	age	59.45
					handset	57.59
					cnt_moc_sms	54.85
					bill_sms_usage	54.84
					interconnect_cost	54.53
					churner_deg_ratio_und	52.79

*see Appendix C for further description of the predictors

Table 13. Performance of EOC6 scoring models without contract related variables

Given the results of all three sample sets in the table above, it can be concluded that social network features do not drastically improve the predictive performance of models built using traditional predictors. Model 3 is having higher lift and positive prediction count in the top 25-60 percentile (see Figure 19). However, the improvement is still minimal.

Without contract related variables, AGE appears to have the highest predictive performance out of all variables. Based on the table and figure below, the young age group of less than 26 years old, which comprises about 10% of the cases, is more likely to churn in the postpaid segment than the older one. Although the age group 30-46 is dominating the segment (40% of the cases), the churn rate is on the same level as the base population.

AGE	Cases (%)	Behavior
Missing	10.17	0.00
[1, 26>	9.48	0.06
[26, 30>	9.75	0.04
[30, 47>	41.64	0.04
[47, 51>	9.74	0.03
[51, 112]	19.21	0.03

Table 14. Postpaid EOC6 predictor: AGE

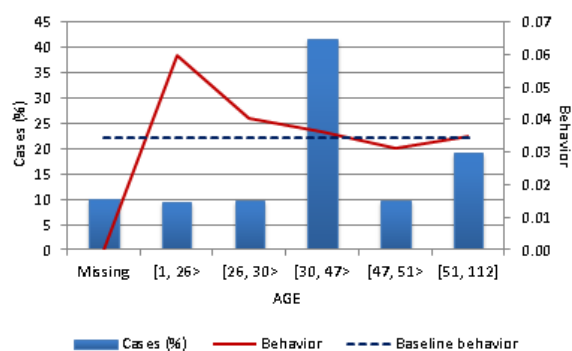


Figure 20. Postpaid EOC6 predictor: AGE

4.1.3. Postpaid EOC2 dataset

Due to the strong predictive power of the contract end date variable, we further limit the size of the postpaid dataset. This section presents the results of postpaid EOC2 dataset, which includes subscribers with end of contract subscription on March 2012 and April 2012. It accounts for 55% of all churners measured in the first observation period.

4.1.3.1. Postpaid EOC2: Scoring and Propagation Models

Table 15 presents the model performance in the training, validation and test set for postpaid EOC2 segment.

	Performance on			Count	Predictor*	
	Training set	Validation set	Test set		Name	CoC (%)
Model 1: Simple Scoring	74.57	74.05	73.83	5	rateplan_group	65.40
					contract_end_tolast	62.18
					lifetime	61.69
					handset	58.88
					dur_mtc	55.41
Model 2: SN Scoring	63.98	60.80	60.91	5	degree2nd	59.18
					outvoicefreq_avg	55.19
					freq_avg_undirect	53.76
					churner_degree_ratio	53.71
					smsfreq_sum	53.53
Model 3: Extended Scoring						
Model 4: Simple Propagation	51.28	50.96	51.22			
Model 5: Extended Propagation	64.88	65.68	68.00			
Model 6: Simple Propagation Undirected	50.86	53.92	52.76			
Model 7: Extended Propagation Undirected	65.27	66.62	68.98			

*see Appendix C for further description of the predictors

Table 15. Performance of postpaid EOC2 scoring models

Generally, performance of models in the postpaid EOC2 dataset is higher than in the postpaid EOC6. However, a similar result to the postpaid EOC6 dataset is obtained. Once again, there is no result for Model 3 because social network features do not have substantial added value on the conventional model. The variables with high contribution to churn prediction are still similar. Although the overall CoC performance of the social network features, e.g., DEGREE2ND and CHURNER_DEGREE_RATIO, is higher than the previous dataset, it remains inadequate for improving churn prediction.

The performance of models trained on the postpaid EOC2 dataset is still similar to the prepaid and postpaid EOC6 dataset. Overall, models built on the EOC2 dataset are having better accuracy compared to models built on the EOC6 dataset. Knowledge of previous churners simply does not have enough influential effect on a subscriber. However, combination of this knowledge with the predicted churn propensity results to model with a reasonable predictive accuracy. Although the CoC values of model

with directed graph have slightly lower performance than the models with undirected graph, the gain and lift chart shows that the difference is insubstantial.

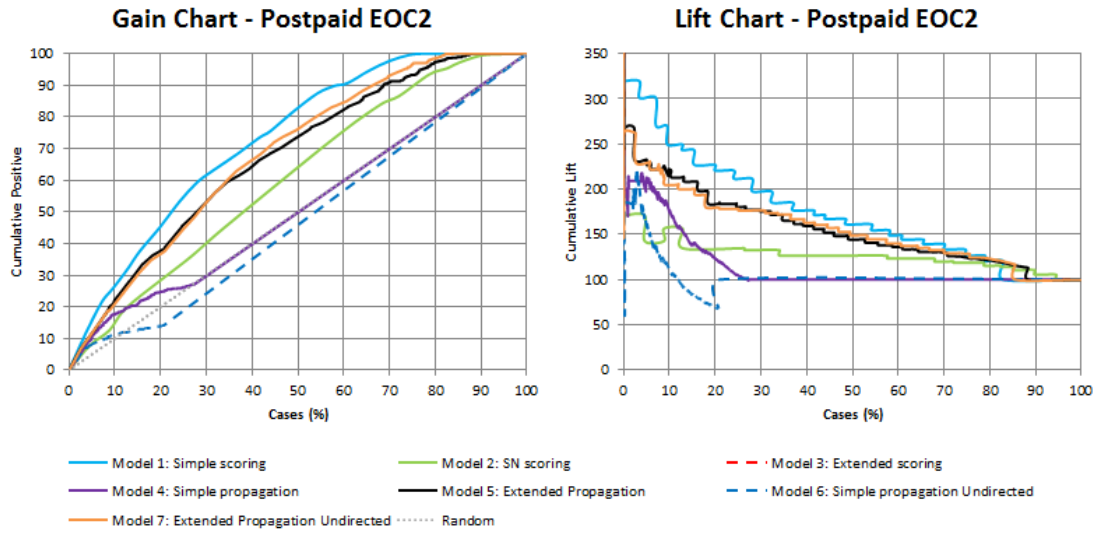


Figure 21. Gain and Lift chart of postpaid EOC2 models

As presented in the figure above, 60% of all churners can be identified by Model 1 in only 30% of all the cases and 80% of churners are in the top 48% of all cases. Additionally, Model 1 has a lift 300% in the top 5 percentile, and 250% in the top 10 percentile. It is about twice as much as the lift of Model 2.

Model 7 successfully predicts about 53% potential churners in the 30 percentile. The model gives about 225% lift on the top 10% and 175% lift on the top 30% of cases. Again, the simple propagation works only on the small fraction of the cases. Overall, it resulted to low quality models and the performance becomes worse after the 10th percentile.

4.2. Effect of the weight propagation function

In the propagation models, we used two different normalization functions, linear and non-linear. The linear function uses the relative weight of an edge shared between two nodes to determine the transferred energy value. The non-linear is simply squared the relative edge weight. In this case, nodes connected to an energy activation source through an edge with small weight will receive less energy compared to using linear function. On the contrary, they will receive more energy if connected with a high weight edge. Based on Table 16, the weight propagation function does not have a high effect on the quality of our predictive models. A very small gain achieved on the linear function, but it is not adequate to conclude that one function is better than the other.

	Performance on test set using	
	Linear function	Non-linear function
Model 4: Simple Propagation	53.06	53.04
Model 5: Extended Propagation	55.73	55.24
Model 6: Simple Propagation Undirected	52.27	52.26
Model 7: Extended Propagation Undirected	58.66	58.30

Table 16. Performance comparison of prepaid using linear and non-linear propagation function

5. Conclusions and Future Works

The following section provides a summary of the most important results of this research study. In addition, the directions of future research are also discussed here.

5.1. Conclusions

Since high churn level could significantly contribute to a substantial revenue loss, churn management and retention is in the focus of the telecommunication industry. Many techniques have been developed to predict future churners as early as possible with the main objective to keep churn within an acceptable rate. The conventional churn models that exploit traditional predictors, such as a subscriber's usage profile, demographic and contractual variables, are typically simple and have a good predictive accuracy (Ferreira et al., 2004; Hadden et al., 2006). However, the predictive accuracy of these models could be reduced if there is few data available, namely in the prepaid segment. Dasgupta et al. (2008) proved that this issue could be solved by using social network information. They showed that models, which are built exclusively using this information, could achieve good prediction accuracy. This research study combined traditional predictors and social network information to build churn models using scoring and propagation techniques and examined its predictive accuracy compared to the conventional churn models.

We constructed and observed 7 churn models for prepaid as well as postpaid segment using scoring and propagation techniques. *Model 1* is a feature-based churn model based on traditional predictors. *Model 2* is based on features extracted from social network graph, such as the average incoming call duration, degree and ratio of churners in the first degree neighborhood. *Model 3* is learned from the dataset of the first and the second model. These first three models are feature-based models, which are built using scoring techniques. The remaining four models are constructed using energy propagation technique based on the spreading activation algorithm. *Model 4* is a simple propagation model, which is based exclusively on social network data. In this model, the initial energy of churners and non-churners are set respectively to 1 and 0. *Model 5*, which is an extended propagation model, is actually boosting of *Model 1* into *Model 4*. The model construction is initialized by calculating the churn propensity score of all subscribers using traditional predictors similar to *Model 1* dataset. Then, these churn score are assigned as the initial energy of non-churners nodes in the social graph. *Model 6* and *Model 7* are similar to *Model 4* and *Model 5*, respectively, except that they are applied to an undirected instead of a directed graph.

Since the number of churners in the neighborhood aligns with the churn propensity, social network information might potentially be a good source for churn prediction. In some extent, the previous churners' behavior has an influential effect on the neighboring decision. Dasgupta et al. (2008) showed similar results and provided evidence that churn models derived from social relationship could have a good predictive accuracy. The models were able to correctly predict 50-60% of future churners by targeting only 10-20% of subscribers. Our experimental results, however, demonstrated the opposite. *Model 2*, which is a feature-based model built exclusively using social network data, has the worst performance compared to the model built using traditional predictors (*Model 1*). *Model 4*, which is a propagation model similar to the Dasgupta's propagation model, has even lower accuracy than *Model 2*. These results showed that social network information alone is not sufficient to predict future churners in both the prepaid as well as the postpaid segment. In our case, *Model 1* consistently outperformed *Model 2* and *Model 4*. Answering the first question, "What are the characteristics of subscribers with high propensity of churn?" we can conclude that the traditional predictors characterize churn in both postpaid and prepaid segment.

The extended propagation model or *Model 5* answers the second question, “How to construct a churn model combining the conventional model with the propagation model?” It incorporates the churn scores of the conventional model (*Model 1*) as the initial values of the propagation model (*Model 4*). This churn scores represent the subscribers’ tendency to churn as a result of previous experiences with the product/service. The social influence might further boost the subscribers’ churn propensity.

This section provides answer of the third question, “What is the added value of incorporating social network features into the conventional churn model?” and the last question, “Is there any performance gain on the models that take into account the traditional predictors as well as social ties information and if so does the performance gain justify the computation cost?” In the prepaid segment, adding social network features on top of the dataset of the traditional predictors’ model has resulted to a slight improvement. However, the performance gain is not substantial enough to justify the computational costs. In the postpaid models, the social network features do not have any added value on the churn prediction. Traditional predictors have apparently a stronger influence on churn compared to social relationships. This statement also aligns with the results of the propagation models. *Model 5*, which indirectly incorporates traditional predictors through social network, outperforms *Model 4*, which only takes into account social network information. In some cases, propagations through an undirected graph return better result than through a directed graph. Since the propagations were done within prepaid and postpaid separately, there might be missing links between subscribers. The directed graph might be negatively influenced more than the undirected graph during the spreading activation process.

5.2. Future works

Our analysis was only able to obtain a limited call detail history of subscribers due to resources constraint. It might be possible that one month of Call Detail Records is not sufficient to project the actual relationships structure of subscribers. Additionally, our research included only traditional mobile interactions, i.e., voice calls and SMS. We could not capture information generated from communications via OTT services (e.g., WhatsApp, Viber, Facebook, Skype etc.), which are popular among the customer base. Therefore, more history might help to cover the lack of information of communications done using OTT services.

Within the postpaid segment, there are variables which strongly correlate with churn, such as the end of contract date. Postpaid models containing only end of contract information could easily beat the performance of models constructed using any other variables. Therefore, we limited the subscriber base to include only subscribers with a certain end of contract period, which is explained in the section 3.2. As consequence, we lose links to subscribers with end of contact period other than defined. It might have an impact on the social network spreading process. Postpaid subscribers with the predefined end of contact period might not directly influence each other. They could have an indirect contact through other subscribers that are not included within the dataset, prepaid subscribers or even off-net subscribers. The future work should further investigate this issue and include the whole subscriber base in order to better view the relationships structure.

The current research study only explored the negative effect of previous churners within social network. Future research could potentially be directed from this limitation. We could take into account influences from both churners and non-churners because both might spread messages on how they perceive the product/service quality. Assuming bad news can have a stronger influence effect and it travels further than the good news, positive influence from non-churners to stay within the network might not as strong as negative influence from churners. Another potential direction is to include both on-net as well as off-net subscribers in the telecom graph because they might influence each other in some extent.

Subscribers from other networks might persuade local subscribers to churn to have a better grouped package plan and vice versa.

Our energy propagation model is purely derived from one-on-one relationships. The spreading computation is done locally and subscribers do not have knowledge beyond their direct neighbors. Other algorithms, such as community detection, are capable identifying the role of subscribers within the social network, such as influencer or adopter. Rather than targeting all future churners, we can minimize our resources by focusing only on churners with high influential power.

References

- [1] Alberts, L.J.S.M. (2006). *Churn Prediction in the Mobile Telecommunications Industry: An application of Survival Analysis in Data Mining*. (Master's thesis), Maastricht University, Maastricht, The Netherlands.
- [2] Angeli, E., Wagner, J., Lawrick, E., Moore, K., Anderson, M., Soderlund, L., & Brizee, A. (2010). *General format*. <http://owl.english.purdue.edu/owl/resource/560/01/>
- [3] Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 44-54. ACM Press: Philadelphia, PA, USA.
- [4] Bohn, A., Walchhofer, N., Mair, P., & Hornik, K. (2009). Social Network Analysis of Weighted Telecommunications Graphs. *Research Report Series/Department of Statistics and Mathematics*, 84. Department of Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna.
- [5] Borgatti, S.P. (1994). A quorum of graph theoretic concepts. *Connections*, 17(1), 47-49.
- [6] Cebrian, M., & Frias-Martinez, E. (2009). Word-of-Mouth algorithms: What you don't know will hurt you. *Proceedings of the ICMI-MLMI '09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*, 5-11. ACM Press: New York, NY, USA.
- [7] Chordiant. (2009). Chordiant Predictive Analytics Director (Version CDM 6.3) [Software]. Chordiant Software, Inc.
- [8] Collins, A.M., & Loftus, E.F. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, 82(6), 407-428.
- [9] Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjee, S., Nanavati, A.A., & Joshi, A. (2008). Social Ties and their Relevance to Churn in Mobile Telecom Networks. *Proceedings of the 11th international conference on Extending database technology*, 668-677.
- [10] Eclipse Foundation. (2012). Eclipse IDE for Java Developers (Version Juno Release) [Software]. Canada: The Eclipse Foundation.
- [11] Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006). Churn prediction using complaints data. *International Journal of Intelligent Technology*, 13, 158-163.
- [12] Ferreira, J.B., Vellasco, M., Pacheco, M.A., & Barbosa C.H. (2004). Data Mining Techniques on the Evaluation of Wireless Churn. *Proceedings of European Symposium on Artificial Neural Networks*, 483-488.
- [13] Greene, W., & Lancaster, B. (2007). *Over the Top Services*. LTC International Inc.
- [14] Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Riverside, CA: University of California, Riverside. <http://faculty.ucr.edu/~hanneman/>
- [15] Hill, S., Provost, F., & Volinsky, C. (2006). Network-Based Marketing: Identifying Likely Adopters via Consumer Networks. *Statistical Science*, 21(2), 256-276.
- [16] Hung, S-Y., Yen, D.C., & Wang, H-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3), 515-524.
- [17] Jacob, R., & Kerremans, P. (2010). *Social Network Analysis: Decrease Churn Rate at Telecom Operators* [White paper]. http://www.social-3.com/downloads/NimzoSTAT_Social_Network_Analysis_white_paper.pdf
- [18] Karnstedt, M., Hennessy, T., Chan, J., & Hayes, C. (2010). Churn in Social Networks: A Discussion Boards Case Study. *Proceedings of the 2nd IEEE International Conference on Social Computing (SocialCom2010)*, 233-240.
- [19] Kawale, J., Pal, A., & Srivastava, J. (2009). Churn Prediction in MMORPGs: A Social Influence Based Approach. *Proceedings of the 2009 International Conference on Computational Science and Engineering*, 4, 423-428.

- [20] Kisioglu, P., & Topcu, Y.I. (2011). Applying Bayesian Belief Network Approach to Customer Churn Analysis: A Case Study on the Telecom Industry of Turkey. *Expert Systems with Applications*, 38(6), 7151-7157.
- [21] Kiss, C., & Bichler M. (2008). Identification of influencers - Measuring influence in customer networks. *Decision Support Systems*, 46(1), 233-253.
- [22] Kraljevic, G., & Gotovac, S. (2010). Modeling Data Mining Applications for Prediction of Prepaid Churn in Telecommunication Services. *Automatika*, 51(3), 375-283.
- [23] Nanavati, A.A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., & Joshi, A. (2006). On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications. *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, 435-444.
- [24] Neo4j. (2012). Neo4j: Community Edition (Version 1.8.M05) [Software]. <http://Neo4j.org/>
- [25] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., & Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences of the United State of America*, 104(18), 7332-7336.
- [26] Prasad, U.D., & Madhavi, S. (2012). Prediction of churn behavior of bank customers using data mining tools. *Business Intelligence Journal*, 5(1), 96-101.
- [27] Radosavljevik, D. (2009). *Prepaid Churn Modeling Using Customer Experience Management Key Performance Indicators*. (Master's thesis), Leiden University, Leiden, The Netherlands.
- [28] Radosavljevik, D., van der Putten, P., & Larsen, K. (2010). The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience Matter?. *Transactions on Machine Learning and Data Mining*, 3(2), 80-99.
- [29] Richter, Y., Yom-Tov, E., & Slonim, N. (2010). Predicting customer churn in mobile networks through analysis of social groups. *Proceedings of the 10th SIAM International Conference on Data Mining (SDM)*, 732-741.
- [30] Rodriguez, M.A. (2011). Graphs, Brains, and Gremlin. *Supporting the Emerging Graph Landscape*. 20 Sept 2012. <http://markorodriguez.com/>
- [31] Ryan, J. (n.d.). The Chi Square Statistic. 1 April 2012. <http://math.hws.edu/javamath/ryan/ChiSquare.html>
- [32] Schellevis, J. (2011). Tweede Kamer neemt Telecomwet met netneutraliteit en cookieregels aan. *Tweakers*. 20 Oct 2012. <http://tweakers.net/nieuws/75223/>
- [33] Soeini, R.A., & Rodpsyh, K.V. (2012). Applying Data Mining to Insurance Customer Churn Management. *International Proceedings of Computer Science and Information Technology*, 30, 82-92
- [34] Steinhaeuser, K., & Chawla, N. V. (2008). Community Detection in a Large Real-World Social Network. *Social Computing, Behavioral Modeling, and Prediction*, 168-175.
- [35] Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- [36] Witten, I. H., Frank, E., & Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). San Francisco: Morgan Kaufmann.
- [37] Ziegler, C-N., & Lausen, G. (2004). Spreading Activation Models for Trust Propagation. *In Proceedings of the IEEE International Conference on e-Technology, e-Commerce, and e-Service*, 83-97.

Appendix A. Abbreviations

Abbreviation	Description
CDR	Call Detail Records
CEM	Customer Experience Management : Measuring how customer experience the products or services
EOC	End of Contract : A given time where the contract is ended
IMEI	International Mobile Identity : A mobile phone unique identifier
IMSI	International Mobile Subscriber Identity : A unique identifier of the SIM card. It is used to identify a subscriber in a network.
KPI	Key Performance Indicator
MO	Mobile Originating : Outbound activity. MO-call means that the subscriber initiates the phone call
MSISDN	Mobile Station Integrated Services Digital Network : A unique number to identify a SIM card in a mobile telecommunication network. It is used to route a call.
MT	Mobile Terminating : Inbound activity. MT-call means that the local subscriber is receiving a phone call
OTT	Over the Top : An over the top service is utilizing the telecom network to perform. However, it does not require any explicit affiliation with the network provider. Examples of over the top application are WhatsApp, Skype or Viber application.
SIM	Subscriber Identity Module : A SIM or SIM card is used to identify a mobile subscriber with a mobile network subscription.
SMS	Short Message Service
SNA	Social Network Analysis
WEKA	Waikato Environment for Knowledge Analysis : A Java-based tool to visualize variable distributions and to perform correlations and machine learning.

Variable Description

B1. Intrinsic variables

Category	Variable
Demographic Information	Age and age group
	Location
Contractual information	Subscriber type, Postpaid or Prepaid
	Package plan, type and group
	Contract start and end date*
	Customer type, Consumer or Business*
	Value segment
	Length of the subscription
	Average revenue per user
Handset information	Handset model
	Manufacturer
Service usage	Count and duration of outbound voice calls
	Count and duration of inbound voice calls
	Count and duration of outbound international voice calls
	Count and duration of inbound international voice calls
	Count and duration of outbound national voice calls
	Count and duration of inbound national voice calls
	Count of outbound and inbound SMS
	Count of outbound and inbound MMS
	Total volume of data usage
	Count and duration of roaming outbound voice calls
	Count and duration of roaming inbound voice calls
	Count of roaming outbound and inbound SMS
	Count of roaming outbound and inbound MMS
	Monthly fixed recurring payment*
	Total payment for each of the service usage, e.g., voice, SMS, MMS, data
	Other payment, e.g., insurance, sales reductions
	Roaming costs and revenues
	Interconnect costs and revenues
	Count and amount of commercial and non-commercial voucher recharge**
	Last activity date

* Only applicable for postpaid

** Only applicable for prepaid

B2. Extrinsic variables

Category	Variable
Connectivity	Count of in-degree and out-degree
	Sum and average of in-weight and out-weight
	Total voice, SMS and voice + SMS frequency to/from neighbors
	Average voice and SMS frequency to/from neighbors
	Total and average of weight*
	Total interaction frequency with neighbors*
	Total and average frequency with neighbors for voice and SMS separately*
	Degree count*
	2 nd degree count*
	3 rd degree count*
Churner connectivity	Count of in-degree and out-degree churners
	Sum and average of in-weight and out-weight with churners
	Total frequency of voice, SMS and voice + SMS to/from churners
	Total weight and average weight of interactions with churners*
	Total interaction frequency of voice, SMS and voice + SMS with churners*
	Ratio of in/out-degree churners to the total in/out-degree
	Ratio of in/out-weight churners to the total in/out-weight
	Ratio of in/out voice, SMS and voice + SMS frequency with churners to the total in/out-weight
	Ratio of churner weight to the total weight*
	Ratio of interaction frequency with churners to the total interaction frequency*
	Churner degree count*
	Churner 2 nd degree count*
	Churner 3 rd degree count*
	Ratio of churner degree out of the total degree*
	Ratio of 2 nd churner degree out of the total 2 nd degree*
	Ratio of 3 rd churner degree out of the total 3 rd degree*

* Direction is not taken into account

Scoring Predictors

C1. Predictor Description

Predictors	Description
age	Age of subscriber
amt_moc_data	Amount of the mobile internet data usage
arpu	Average revenue per user in euro
bill_sms_usage	Amount of payment for the SMS usage
churner_degree_ratio	Total churners in the adjacent neighborhood divided by the total adjacent neighbors
churner_degree_ratio_und	Total churners in the adjacent neighborhood divided by the total adjacent neighbors regardless the direction
cnt_moc	Number of outbound voice calls
cnt_moc_sms	Total outbound SMS
cnt_mtc	Total inbound voice calls
contract_end_days_to_last	The number of days between the end date of subscription and the last activity date
contract_startdate	Start date of subscription
degree2nd	Total number of the second degree neighbors
dur_mtc	Duration of inbound voice calls
handset	Handset model information
freq_avg_undirect	Average of interaction count to neighbors regardless the direction
interconnect_cost	The cost charged by another network to enable connection with their subscribers
inweight_sum	Total weight of incoming edges
inweight_avg	Average weight of incoming edges
lifetime	Relationship duration of the observed telecom network with a subscriber
outvoicefreq_avg	Average of outbound voice interaction
outweight_sum	Total weight of outgoing edges
outweight_avg	Average weight of outgoing edges
rateplan_group	Subscription plan group
smsfreq_sum	Total number of SMS interactions with neighbors regardless the direction
value_segment	Information about how valuable the subscriber is. It consists of New Joiner, Unknown and it also ranges from A to F.

C2. Scoring Model Predictors

Predictors	Prepaid			Postpaid EOC2		Postpaid EOC6	
	Model1	Model2	Model3	Model1	Model2	Model1	Model2
	(6)	(4)	(7)	(5)	(5)	(5)	(5)
amt_moc_data	x		x				
arpu	x		x				
churner_degree_ratio					x		x
cnt_moc						x	
cnt_moc_sms	x		x				
cnt_mtc	x		x				
contract_end_days_to_last				x		x	
contract_startdate	x		x				
degree2nd		x	x		x		x
dur_mtc				x			
handset				x		x	
freq_avg_undirect		x			x		
inweight_sum		x					
inweight_avg							x
lifetime				x		x	
outvoicefreq_avg					x		
outweight_sum							x
outweight_avg		x					
rateplan_group				x		x	
smsfreq_sum					x		x
value_segment	x		x				