UNIVERSITEIT GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

IMPROVING CUSTOMER RETENTION IN FINANCIAL SERVICES USING KINSHIP NETWORK INFORMATION

Dries F. Benoit

Dirk Van den Poel

May 2012

2012 / 786

Dries F. Benoit, Postdoctoral researcher, Department of Marketing, Faculty of Economics and Business Administration, Ghent University, Tweekerkenstraat 2, B-9000 Ghent, Belgium, e-mail: Dries.Benoit@UGent.be
Dirk Van den Poel, Professor of Marketing Modeling, Department of Marketing, Faculty of Economics and Business Administration, Ghent University, Tweekerkenstraat 2, B-9000 Ghent, Belgium, e-mail: Dirk.VandenPoel@UGent.be

D/2012/7012/19

IMPROVING CUSTOMER RETENTION IN FINANCIAL SERVICES USING KINSHIP NETWORK INFORMATION

Abstract

This study investigates the advantage of social network mining in a customer retention context. A company that is able to identify likely churners in an early stage can take appropriate steps to prevent these potential churners from actually churning and subsequently increase profit. Academics and practitioners are constantly trying to optimize their predictive-analytics models by searching for better predictors. The aim of this study is to investigate if, in addition to the conventional sets of variables (socio-demographics, purchase history, etc.), kinship network based variables improve the predictive power of customer retention models. Results show that the predictive power of the churn model can indeed be improved by adding the social network (SNA-) based variables. Including network structure measures (i.e. degree, betweenness centrality and density) increase predictive accuracy, but contextual network based variables turn out to have the highest impact on discriminating churners from non-churners. For the majority of the latter type of network variables, the importance in the model is even higher than the individual level counterpart variable.

Keywords: network based marketing, CRM, predictive analystics, social network analysis (SNA), kinship network, financial services, random forests

1. INTRODUCTION

1.1 Customer Relationship Management (CRM)

In the past, companies had close relationships with their customers. They knew each customer individually and offered them personal customized service. As a result, they earned loyalty and a large share of their customers' business. Over the years, through increased competition and mass marketing, customers interchanged personalized service for anonymity, reduced variety and lower prices (Peppard, 2000).

The current business environment is characterized by intense competition and saturated markets. Mutanen et al. (2006) remarks that the mass marketing approach, where each customer gets the same treatment of the company, cannot succeed in the diversity of consumer business today. Therefore, companies are practicing an approach to marketing that uses continuously refined information about current and potential customers to anticipate and respond to their needs. This marketing strategy is called Customer Relationship Management (CRM) (Peppard, 2000).

CRM is about structuring and managing the relationships with customers (Kim, Suh and Hwang, 2003). CRM covers all the processes related to customer acquisition, customer cultivation, customer retention and the reactivation of defected customers. This study can be situated in the customer retention domain. The goal is to identify the customers with a high churn probability in order to target them with appropriate actions and consequently try to keep them within the company. These actions may include targeting these customers with appropriate “next-product-to-buy” (NPTB) as shown in Prinzie & Van den Poel (2006) for financial services.

1.2 Customer attrition in financial services

Personal retail banking is characterized by customers who typically spread their assets over only one or two companies and stay with a company for long periods of time (Mutanen et al., 2006). From the point of view of the financial services company, this produces a stable environment for CRM. It is argued that these companies need to operate on a long-term “cradle-to-grave” customer management strategy (Li et al., 2005). This means that they recognize that young customers are often unprofitable in their earlier years, but become profitable at a later stage. The longer customers stay with the bank, the more they become tied to such an extent that the perceived cost of defection outweighs the benefits of shifting their banking business to another provider.

Although the process of attracting new customers is important, most financial services companies make customer retention a top priority for several reasons: in general, the longer a customer stays with a bank, the more that

customer is worth (Benoit & Van den Poel, 2009). Long-term customers buy more, take less of a company's time, are less sensitive to price differences, and bring in new customers (Ganesh et al. 2000; Reichheld, 1996). Long-term customers become less costly to serve because of the banks' greater knowledge of the existing customer base and reduced servicing costs (Ganesh et al., 2000). In addition, the cost of winning a new customer is about five times greater than the cost of keeping an existing one (Colgate & Danaher, 2000). A study by Reichheld & Sasser (1990) showed that reducing defections by just 5% can generate 85% more profits for a bank. The latter findings corroborate the results of a study of Van den Poel & Larivière (2004), which illustrated how increasing retention by just one percent resulted in substantial profit gains.

1.3 Network based marketing

A limitation of traditional direct marketing is that it assumes that customers act independently. In reality, a customer's decision to buy a product is strongly influenced by his or her friends, family, business partners, etc. (Domingos and Richardson, 2001). Ignoring these network effects when deciding which customers to market to can lead to suboptimal decisions. For example, an unprofitable customer may be worth marketing to when this customer is likely to influence a lot of peers. In contrast to traditional direct marketing, network based marketing recognizes that links between consumers exist. As a result of the availability of gigantic databases of customer information today, companies now are able to target their customers taking into account their interrelatedness. Traditional marketing research does not reveal these social connections between consumers and thus cannot take advantage of links between customers.

Network based marketing assumes some kind of interdependency among customer preferences (e.g. purchase patterns, shopping habits,...). These interdependencies are measured through implicit links (e.g. matching on demographic attributes, geographic links, etc.), or through explicit links (e.g. communications between actors, family ties, etc.) (Hill et al., 2006).

Although network based marketing offers clear advantages over direct marketing, the use of social network information in prediction modelling is a very recent phenomenon (e.g. Hill et al., 2006; Manchanda et al., 2008, Subelj et al., 2011). This study contributes to the literature by investigating if social network information can improve the accuracy of churn detection. Moreover, this is, to the best of our knowledge, the first study that investigates different types of network effects in the same research setting.

The reminder of this paper is organized as follows: Section 2 delves into the methodological aspects of social network analysis. In order to get acquainted with the prevailing concepts and terminology, we first give a brief introduction to the field. Next, we show how the different effects that come into play in a social network can be quantified and how this data can be used in a modeling context. Finally, Section 2 is concluded with a discussion of the classifier and the evaluation criteria used in this study. Section 3 explains the dataset that was used to test the proposed methodology and gives an overview of the results that were obtained. Finally, Section 4 concludes the study with a discussion on the main findings.

2. METHODOLOGY

2.1 Social networks

A crucial insight in network analysis is that actors and their actions are viewed as interdependent rather than as independent and autonomous units (Wasserman and Faust, 1994). Typically, a cross-sectional CRM dataset contains a single row for every customer and columns for the information on that customer, where we assume that all rows are independent of each other. However, the information embedded in social networks is not of this standard form where attributes can easily be linked to individuals. To make this clear, consider the simple graphical representation of a kinship network in Figure 1.

[INSERT FIGURE 1 HERE]

This way of representing a network is called a graph. Several dots (or '*nodes*') can be seen, which correspond to the individuals or any other unit of analysis. Some nodes are linked to other nodes by lines (or '*ties*'). Two nodes sharing a link are '*adjacent*' nodes. Together, all ties and nodes form a graph.

Nowadays we are facing a new trend in network research that is largely driven by the availability of powerful computers and the fast growing number of relational databases available to researchers (Chen et al., 2009). The last couple of years, the focus is shifting away from the analysis of small-scale networks and the properties of individual ties towards large-scale statistical properties of networks (Newman, 2003). Previous studies used to look at small networks of only ten to several hundreds of nodes. However, in recent studies, it is not unusual to see

networks with millions of nodes (e.g. Hill et al., 2006). Due to the dimensions of these new datasets, some specific approaches have emerged.

The data warehouse of the anonymous financial services company used for this study contains information on three categories of kinship links, i.e. parent-child relations, sibling relations and finally spouse relations. Using this information on the ties, the kinship networks of the customers were constructed. More specific, we built the networks by means of the egocentric network approach (e.g. Bar-Yossef et al., 2008). This means that a given customer or '*ego'* is focused on and then all other customers with whom the '*ego'* shares a kinship link (the '*alters'*) are identified (see Figure 2). The network for this given ego is now defined. Next, we zoom in on another customer (who now becomes '*ego'*) and construct his/her network. This process continues until all customers' egocentric networks are identified.

[INSERT FIGURE 2 HERE]

The egocentric network approach has the distinct advantage that its analysis is related to the traditional attribute-based methodology, in that the typical predictors (socio-demographics, purchase history, etc.) are augmented with network measures that are deduced from the ego network (Knoke & Yang, 2007). Moreover, other methods that emerged from social network analysis are only suitable for networks up to a few dozen to a few hundred customers, whereas the egocentric network approach is able to handle the typical CRM datasets with hundreds of thousands of customers (Hill et al., 2006). The egocentric network created in this research, contains all alters no more than two ties removed from ego. The network that emerges from this method is thus a 2nd order egocentric network.

This strategy has various advantages. First, in social network analysis, it is well recognized that individuals who are more than two links away do not exert a significant influence on the focal customer (Knoke & Yang, 2007). Second, since the company database only includes information on immediate family, the 2nd order ego network extends the 1st order ego network with other relevant family, while distant family members are excluded. For example (see Figure 2), ego can now be influenced by his/her grandfather, but not by his/her granduncle, since the latter is three links away from ego. Finally, the 2nd order egocentric network approach has the additional advantage of increasing data quality. Network data is very labor-intensive to collect and missing information on the ties is

likely to occur (Wasserman and Faust, 1994). By using a 2nd order egocentric network, this can be partly overcome. In Figure 2, when the link between ego and his/her mother is missing in the data warehouse, ego's mother would not be included in ego's network. By considering all alters no more than two ties removed from ego, his/her mother will also be included in the ego network and the resulting ego network did not suffer from the missing data.

2.2 Social network metrics

At this point we have defined how we can indentify the kinship network of a given customer (i.e. using the 2nd order egocentric network approach). The question, however, still remains what the effects are that play in this network environment and how these effects can be quantified and measured. Earlier work on economic and social theory gives guidance in this respect.

According to Manski (2000) three types of network influence may occur. A first type is endogenous interaction, meaning that the propensity of an agent to behave in some way varies with the behavior of the group. This is the most intuitive network effect and is often central in studies of peer influence. Recently, a number of studies have corroborated the existence of such an effect in different situations. Nair et al. (2006) found that physicians are influenced by the prescription behavior of their colleagues. In a similar setting, Manchanda et al. (2008) also found evidence for the existence of endogenous interaction. Finally, Hill et al. (2006) showed that cell phone users are more inclined to upgrade their account when they call to people using such an upgraded account. Most often, this efect is represented by a dummy variable that flags one of the behavior of interest is already present in the network of the focal customer.

A second type of network influence is contextual interaction (Manski, 2000). Here, the propensity of an actor to behave in some way varies with the exogenous characteristics of the group members. To clarify this, consider this example: A variable indicating social class is often based on the geographic area where a customer lives. This variable only is an approximation of the customers' social class and thus can be inaccurate. A variable returning the average social class score of the network members might be more accurate in indicating social class or at least give some additional information about the this customer. The effect on the individual customer of this latter variable is what is called a contextual interaction. Although other methods than averaging for summarizing the attributes of the network members are possible, taking the mean of the attributes of the network members is a standard procedure in the extant literature (Manski., 2000).

Finally, a third type of network influence is network structure effects. Centrality measures are some of the most fundamental and frequently used measures of network structure (Newman, 200). Examples are found in Liu (2011) and Kim et al. (2011). Centrality measures address the question: "Who is the most important or central person in this network". There are many answers to this question, depending on what we mean by 'important', giving rise to many different centrality metrics. Kiss and Bichler (2008) investigated which of the various centrality measures are best able to select influential customers. They found out that degree and betweenness centrality are good describers of the capability of a customer to influence others.

- *Degree centrality*

Probably the simplest of the centrality measures is degree centrality (also called degree). It measures the importance of a node by the number of ties that are connected to a given node (Wasserman and Faust, 1994). Degree centrality is illustrated by network A and network B, shown in Figure 3. Both networks have $m = 5$ actors. Network A has the property that exactly one actor, n_1, has ties to all $m - 1$ other actors. It is clear that the first actor is the most central. In network B, all actors are, from a structural point of view, interchangeable. This means that all actors have the same centrality index. Degree centrality is often interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network, for example a virus, some information, the risk of churn behaviour, etc..

[INSERT FIGURE 3 HERE]

Degree centrality can be computed by creating an *m* x *m* matrix where the cells contain a1 value when there is a link between n_i and n_j and a 0 value when no such link exists. The degree, $C_d(n_i)$, of a node n_i is then:

$$C_d(n_i) = \sum_{j=1}^{m} x_{ij} \qquad (1)$$

Although degree is pretty simple, it is often a highly effective measure of the influence or importance of a node: in many social settings people with more connections tend to have more power (Newman 2007).

- *Betweenness centrality*

Interactions between two nonadjacent actors depend on the other actors in the set of actors, in particular the actors who lie on the paths between those two. These 'in-between actors' potentially have some control over the interactions between the two nonadjacent actors (Wasserman and Faust, 1994). Node n_1 of Figure 4 illustrates the concept of betweenness. Although this node is only connected with two other nodes, it serves as a bridge between two groups of nodes and therefore it has a high betweenness centrality.

[INSERT FIGURE 4 HERE]

Betweenness is based upon on the concept of network paths. Newman (2007) defines a path in a network as a sequence of nodes traversed by following ties from one to another across the network. A geodesic path is the shortest path through the network from one node to another. Note that there might be (and often is) more than one geodesic path between two nodes (Newman, 2003). The betweenness of a node is calculated as the fraction of shortest paths between node pairs that pass through this node (Freeman, 1979). The betweenness centrality index is defined mathematically by Freeman (1977) as

$$C_B(n_i) = \frac{\sum_{i \neq j \neq l} g_{jl}(n_i)}{g_{jl}} \qquad (2)$$

Where $g_{jl}(nl)$ is the number of shortest paths linking the two nodes j and l containing node i. A node with high betweenness will, in many social contexts, exercise most of its influence by virtue not of being in the middle of the network, even though this is possible, but by lying between other nodes in this way (Newman, 2007).

- *Density*

Density is a widely used concept that describes the general level of linkage among the nodes in a network (Scott, 2000). A 'complete' network, from density point of view, is a network in which all the nodes are adjacent to another, meaning that each node is connected directly to every other node. The concept of density thus summarizes the overall distribution of ties in terms of how far the state of the network is from a complete network.

Density depends upon two parameters of network structure: First, the inclusiveness of the network and second, the sum of the degrees of its nodes. Inclusiveness stands for the number of nodes that are included within the various connected parts of the network. In other words, the inclusiveness of a network is the total number of nodes minus the number of isolated nodes. An isolated node has no ties so can contribute nothing to the density of the network. Thus, the more inclusive is the graph, the more dense it will be. Those nodes that are connected to one another, however, will vary in their degree of connection. Some nodes will be connected to many other nodes, while others will be less well connected. The higher the degree of the nodes in a network, the denser, it will be.

These two parameters are included in the formula of density. This involves comparing the actual number of ties present in a network with the total number of ties that would be present if the network were incomplete. The density of a network is defined as the number of ties in a network, expressed as a proportion of the maximum possible number of ties. The formula for the density is:

$$D(n_i) = \frac{l}{m(m-1)/2} \quad (3)$$

Where l is the number of ties present and m is the number of nodes in the network.

2.3 Classification technique: Random Forests

As the problem we are dealing with in this research is of binary form (will a customer leave the company, yes or no), we argue to use a modeling technique that has some unique properties when applied in this context, i.e. random forests. Random forests is Breiman's (2001) extension of the decision tree method. Decision tree methods build a collection of rules to use as a predictive model (Quinlan, 1986). Decision trees have become a popular classification technique because of its simplicity and interpretability. Moreover, they can deal with predictors measured at different measurement levels. The downside is that these models often suffer from suboptimal performance (Hu, 2005). Random forests is an answer to this shortcoming that overcomes the instability of traditional decision trees by creating an ensemble of trees and letting them vote for the most popular class (Breiman, 2001). In this paper, we select random forests as proposed by Breiman (2001), which uses the strategy of a random subset selection of m predictors to grow each tree, where each tree is grown on a bootstrap sample of the training set. This subset of variables is then used to create splits for the nodes. Luo et al (2004) argue that the predictive power of random forests is among the best of the available techniques. This has led to a wide area of applications of the technique, ranging from bioinformatics (Deng et al., 2004) to marketing (Larivière & Van den Poel, 2005). An interesting by-product of these ensembles of trees is their importance measures for each variable.

The importance measures are calculated as follows: for each tree, the node impurity (based on AUC, see Section 2.3) on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two predictive performances are then averaged over all trees, and normalized by the standard error. Random forests require only two parameters to be set by the researcher. These are the number of variables, *m*, to be randomly selected and the number of trees to be grown. In accordance with the instructions of Breiman (2001), we pick a large number for the number of trees to be grown (i.e. 500) and we set *m* to the square root of the number of variables.

2.3 Evaluation criteria

Many different evaluation criteria are possible for investigating the predictive performance. Evaluation criteria for the predictive performance of classification models are often confusing because of the cut-off value that has to be chosen to discriminate between the predicted events and non-events. The Area Under the Receiver Operating Curve (AUC) avoids this difficulty by considering all possible thresholds on the predicted probabilities. It presents a two-dimensional graph of the sensitivity of the confusion matrix (the number of true positives versus the total number of defectors) and one minus the specificity of the confusion matrix (the number of true negatives versus the total number of non-defectors) for all possible cut-offs (Egan, 1975). The area under the resulting curve lies between 0.5 and 1. The closer this value is to 1, the better the model is at discriminating events from non-events. AUC can be interpreted as the probability that the predicted churn probability of a churned customer is higher than the predicted probability of a retained customer. AUC evaluation for predictive accuracy is extensively used in CRM (e.g. Lemmens & Croux, 2006; Hill et al., 2008, Coussement et al., 2010) and other data-mining contexts (Takahashi et al., 2009). Comparing the predictive performance of two models using AUC is then conducted using the non-parametric test proposed by Delong et al. (1988).

The other performance measure used in this study is “lift”. This evaluation criterion focuses exclusively on the top *x* percent of most critical customers. The top *x* percent riskiest customers (i.e. the group of customers with the highest predicted churn probabilities) represents an ideal segment for targeting in a retention-marketing campaign (Lemmens & Croux, 2006). This performance measure is very attractive because it incorporates somewhat the fact that marketing budgets are limited. As a result, actions to reduce churn (e.g. direct mail campaigns) are limited to a segment of customers that is at high risk. In practice, the metric is calculated by ordering the customers on decreasing predicted churn probability. Next, the proportion of real churners in the top *x* percent is compared with

the proportion of churners in the total dataset. The higher the lift, the better is the model. For example, a top-10% lift of 2 means that the model under investigation identifies twice as many churners in the top 10% than a random assignment would do.

AUC and lift are measuring different aspects of the predictive accuracy of the models. Both evaluation criteria provide complementary information. A model can be good at identifying the most risky segment but less effective at recognizing less risky customers. Combining the two metrics provides a thorough evaluation of the performance.

- **RESULTS**

3.1 Data

A European financial services company provided the data for this research project. All active customers at the end of June 2006 were selected, a group of 244,787 clients in total. Information about the customers was extracted from the company data warehouse from the moment they joined the company until June 29th 2006. This information was then captured into explanatory variables (both traditional and social-network-based variables). The dependent variable in this setting is whether a given customer churns or not. Here a churned customer is defined as someone who closed all his/her bank accounts with this company. The dependent variable is based on the churn behavior in the period from June 30th 2006 until December 31st 2006.

[INSERT FIGURE 5 HERE]

Figure 5 gives a graphical representation of the modelling process. For a given individual all traditional predictors are extracted from the data warehouse. Next, the 2nd order egocentric network is identified for that customer. Using this network, the different network effects that might play are calculated. This is done for every customer in the database and the result is one large table with both traditional as social network based variables for every customer. These predictors then are used as inputs for the random forests and the different types of network effects are evaluated in terms of predictive performance and variable importance.

Table 1 gives an overview of both traditional and network based-variables used in this study. Traditional churn models usually take socio-demographic variables (age, gender, etc.) as predictors in addition to past behavior that

is summarized in terms of recency, frequency and monetary value (RFM). This information is represented in variables 1 to 12 in Table 1. In the current setting, the traditional churn predictors are augmented with network-based variables. Variables 13 to 24 contain information on the exogenous characteristics of the network members. Variable 25 accounts for the endogenous network effect, while variables 26 to 31 capture the network structure influence.

[INSERT TABLE 1 ABOUT HERE]

To avoid overfitting of the model, the database is divided into a training set and a validation set. The model is trained on the training set and tested on the validation set. The training set is composed by randomly assigning 70% of the customers, while the other 30 % are assigned to the validation set. In order to get an equal churn rate in both of the sets, stratification is performed on the churn variable. Note that only a very small percentage, i.e. 2 %, of the dataset churns in the six month follow-up period.

2.4 Results

In order to provide evidence for the hypothesis of additional predictive performance of social network based variables on top of the traditional variables, two different churn models are built. The first prediction model makes use of traditional variables (hereafter called the 'traditional model'). The second model (or the 'extended model') augments the traditional model with the social network indicators.

2.4.1 Predictive performance

Table 2 provides an overview of the predictive performance of both models in terms of lift (measured at different percentiles) and AUC. All metrics of predictive performance are generated on the validation datasets. Table 2 shows that the extended model always has a considerable higher predictive performance than the corresponding traditional model, irrespective of the performance measure used.

[INSERT TALBE 2 ABOUT HERE]

The AUC metrics demonstrate that the extended model is better in discriminating churners from non-churners. The difference in AUC between the two models is almost 0.04 and this difference is significant ($p < 0.001$) according to the test of Delong et al. (1988). This means that the model augmented with the social network variables does 4 percent points better in discriminating customers at risk versus the others. In a churn context within the financial services industry, where even a small change in churn rate strongly affects profit (see Section 1.2), this is a very encouraging result.

The lift values show that the extended model is better at predicting customers at high risk compared to the traditional model. In a marketing context this aspect of predictive accuracy is highly important. Since managers always have to deal with limited budgets, not all customers at risk can be targeted. Therefore they have to restrict the recovery attempts to the customers having the highest churn risk. For example, the top 5% lift shows that the traditional model identifies 5.85 times as many real churners in the top 5% of highest predicted probabilities than a random assessment would do. The extended model however, is able to identify 6.77 times as many real churners. These results show that taking into account the social network based variables leads to a substantial increase in efficiency of the retention program.

2.4.2 Variable importance

In Table 3, the average normalized importance of each predictor for the random forest method is presented. As an importance value of zero means that there is no predictive power in the variable, the table shows that all variable in the model have an impact on the accuracy of the predictions. The sociodemographic variable age exerts the largest impact, while whether a customer uses home banking has the lowest impact.

[INSERT TABLE 3 ABOUT HERE]

The importances show that the most important variables in the model are sociodemographic variables together with the RFM variables (i.e. recency, frequency, monetary value, interpurchase time). This finding corroborates previous research in this context (e.g. Baesens et al., 2002, Buckinx & Van den Poel, 2005). Apart from the home banking variables, the least important variables turn out to be the network structure measures. However, still having a considerable impact on predictive performance, they do not contribute to the same extent as the socio-demographics and the RFM variables to the performance of the model. This result does not support the results of

Hill et al. (2006) where the network structure variables were among the most impactful variables. An interesting and new result from the table of importances is that the aggregate versions of the sociodemographic, RFM and other behavioral variables, i.e. the contextual network variables, have an important impact in discriminating churners from non-churners. In general it is the case that when a given variable on the individual level exerts a large impact on the dependent variable, the network based counterpart variable turns out also to be important compared to the other variables. Note that this effect cannot be linked to possible multicollinearity between the predictors, as the random forests approach (and the method for computing the importances) is not influenced by this phenomenon (Sandri & Zuccolotto, 2006).

[INSERT TABLE 4 ABOUT HERE]

Table 4 compares every individual level variable with the network based counterpart variable, i.e. the contextual network effects. The figures show that for seven out of twelve variables, the network based variable has a higher impact on the churn probability than the individual level variable. This leads to the remarkable insight that for those variables it is more beneficial to have the information of the network members of the customer than knowing the value of the variable of the customer him or herself. E.g. Nbr_insurances measures the number of different insurance services the individual customer owns. N_Nbr_insurance represents the number of insurances of the network members of the individual customer. The figures show that it is almost double as important to know the number of insurances of the network members compared to knowing the number of insurances of the individual customer when predicting the individual customers churn probability. This again shows that network based variables not only are important in predicting customer churn, but often they turn out to be even more impactful than the individual counterpart variables.

- **DISCUSSION**

This study presents the benefits of integrating kinship network based information in churn management. It adds to the small but growing literature that investigates the opportunities of network data emerging from individual

consumers. To the best of our knowledge, this is the first study that compares the predictive performance of the different network effects in the same context.

This study shows that it is beneficial for database marketers to store network information in their data warehouses (if this is not already the case) and to include network based information in their churn prediction models. Three different types of network variables were investigated. Contextual effects turned out to have the highest impact on the predictive performance of a traditional churn model, followed by the endogenous effect. Also the network structure effects significantly increased predictive performance. Together, these effects considerably improved the traditional churn model. The importances of the contextual network effects showed that the majority of these variables are even more important than the individual-level counterpart variable.

Noteworthy is that the current results are both similar and different compared to the main findings of the study of Hill et al. (2006) which also focused on the predictive potential of social network based variables. Similar in the sense that both studies found that network variables improve predictive performance. Different because the improvement in predictive performance was much larger in the study of Hill et al. (2006) than in the current study. Although both studies investigated the influence of social network based variables, Hill et al. (2006) differs in some key aspects from the current study. First, the type of social network is different: the study of Hill et al. (2006) makes use of telecom data (who calls whom), while the current research deals with kinship network information. Moreover, the former study investigates an up-sell context, while here, the focus is on churn behavior. Finally, the two studies deal with a very different industry setting. The differences in effect sizes do confirm Lessig & Park (1982) and Childers & Rao (1992) in that the degree of reference group influence varies for products, consumed on different occasions (public versus private) and for reference groups (family versus peers).

Nonetheless, in highly competitive and saturated markets, such as the financial services industry, customer retention is crucial. Better identification of customers at risk and subsequent actions towards those customers has a substantial impact on profits. Incorporating network information turned out to be a viable strategy to achieve this goal.

REFERENCES

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G., 2002. Bayesian neural network learning for repeat purchase modelling in direct marketing. European Journal of Operational Research, 138 (1), 191–211.

Bar-Yossef, Z., Guy, I., Lempel, R., Maarek, Y.S. & Soroka, V. (2008). Cluster ranking with an application to mining mailbox networks. Knowledge and Information Systems, 14 (1), 101-139.

Benoit, D.F. & Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. Expert Systems with Applications, 36(7), 10475–10484.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Chen, C., Yan, X.F., Zhu, F.D., Han, J.W., Yu, P.S. (2009). Graph OLAP: a multi-dimensional framework for graph data analysis. Knowledge and Information Systems, 21(1), 41-63.

Childers T.L, & Rao A.R. (1992). The influence of familial and peer-based reference groups on consumer decision. Journal of Consumer Research, 19(2), 198-211.

Colgate, M. R., & Danaher, P. J. (2000). Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. Journal of the Academy of Marketing Science, 28(3), 375-387.

Coussement, K., Benoit, D.F. & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. Expert Systems with Applications, 37(3), 2132-2143.

DeLong E.R., DeLong D.M., & Clarke-Pearson D.L. (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach. Biometrics, 44 (3), 837-845.

Deng, Y. P., Chen, H. S., Tao, L., Sha, Q. Y., Chen, J., Tsai, C. J., et al. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics, 5(81), 1–12.

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. International Conference on Knowledge Discovery and Data Mining, 57-66.

Egan J.P. (1975). Signal detection theory and ROC analysis. New York: Academic Press.

Freeman, L. C. (1977). Set of Measures of Centrality Based on Betweenness. Sociometry, 40(1), 35-41.

Freeman, L. C. (1979). Centrality in Social Networks Conceptual Clarification. Social Networks, 1, 215-239.

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. Journal of Marketing, 64(3), 65-87.

Hill S., Provost F., & Volinsky C. (2006). Network-based marketing: identifying likely adopters via consumer networks. Statistical Science, 21(2), 256-276.

Hu, X. H. (2005). A data mining approach for retailing bank customer attrition analysis. Applied Intelligence, 22(1), 47-60.

Kim, J., Suh, E., & Hwang, H. (2003). A model for evaluating the effectiveness of CRM using the balanced scorecard. Journal of Interactive Marketing, 17(2), 5-19.

Kim, S., Suh, E., & Jun, Y. (2011). Building a knowledge brokering system using social network analysis: A case study of the Korean financial industry. Expert Systems with Applications, 38, 14663-14649.

Kiss, C. & Bichler, M. (2008). Identification of Influencers – Measuring Influence in Customer Networks. Decision Support Systems, 46(1), 233-253.

Knoke, D. & Yang, S. (2007). Social Network Analysis. Los Angeles: Sage Publications.

Larivière, B. & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications, 29(2), 472-484.

Lemmens A., & Croux C. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276-286.

Lessig V.P., & Park C.W. (1978). Promotional perspective of reference group influence: Advertising implications. Journal of Advertising, 7(2), 41-47.

Li, S., Sun, B. & Wilcox, R.T. (2005). Cross-Selling Sequentially Ordered Products: An Application to Consumer Banking Services. Journal of Marketing Research, 42(2), 233-239.

Liu, C.H. (2011). The effects of innovation alliance on network structure and density of cluster. Expert Systems with Applications, 38, 299-305.

Luo, T., Kramer, K., Goldgof, D. B., Hall, L. O., Samson, S., Remsen, A., et al. (2004). Recognizing plankton images from the shadow image particle profiling evaluation recorder. IEEE Transactions on Systems Man and Cybernetics Part B—Cybernetics, 34(4), 1753–1762.

Manchanda, P., Xie, Y., & Youn, N. (2008). The Role of Targeted Communication and Contagion in Product Adoption. Marketing Science, 27(6), 961-976.

Manski, C.F. (2000). Economic Analysis of Social Interactions. Journal of Economic Perspectives, 14(3), 115-136.

Mutanen T., Ahola J., Nousiainen S. (2006) “Customer churn prediction - a case study in retail banking”, ECML/PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges, Berlin.

Nair, H. S., Manchanda, P., & Bhatia, T. (2006). Asymmetric Social Interactions in Physician Prescription Behavior: the Role of Opinion Leaders: Stanford University, Graduate School of business.

Newman, M. E. J. (2003). The structure and function of complex networks. Siam Review, 45(2), 167-256.

Newman, M. E. J. (2007). The mathematics of networks. Working paper: Center for the Study of Complex Systems, University of Michigan, Ann Arbor.

Padgett, J. F., & Ansell, C. K. (1993). Robust Action and the Rise of the Medici, 1400-1434. American Journal of Sociology, 98(6), 1259-1319.

Peppard, J. (2000). Customer Relationship Management (CRM) in financial services. European Management Journal, 18(3), 312-327.

Pekalski, A. (2001). Ising model on a small world network. Physical Review E, 64(5), art. no.-057104.

Prinzie A. & Van den Poel D (2006), Investigating Purchasing Patterns for Financial Services using Markov, MTD and MTDg Models, European Journal of Operational Research, 170 (3), 710-734.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

Reichheld, F. F. (1996). Learning from customer defections. Harvard Business Review, 74(2), 56-69.

Reichheld, F. F., & Sasser, W. E. (1990). Zero Defections - Quality Comes to Services. Harvard Business Review, 68(5), 105-111.

Scott, J. (2000). Social network analysis : a handbook (2nd ed.). London ; Thousand Oaks, Calif., SAGE Publications.

Sandri, M. & Zuccolotto, P. (2006). Variable selection using Random Forests. In: Data Analysis, Classification and the Forward Search, Springer Berlin Heidelberg.

Sigman, M., & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon. Proceedings of the National Academy of Sciences of the United States of America, 99(3), 1742-1747.

Subelj, L., Furlan, S., & Bajec, M. (2011). An expert system for detecting automobile insurance fraud using social network analysis. Expert Systems with Applications, 38, 1039-1052.

Takahashi, K., Takamura, H. & Okumura, M. (2009). Direct estimation of class membership probabilities for multiclass classification using multiple scores. Knowledge and Information Systems, 10(2), 185-210.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research, 157(1), 196-217.

Wasserman, S. & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

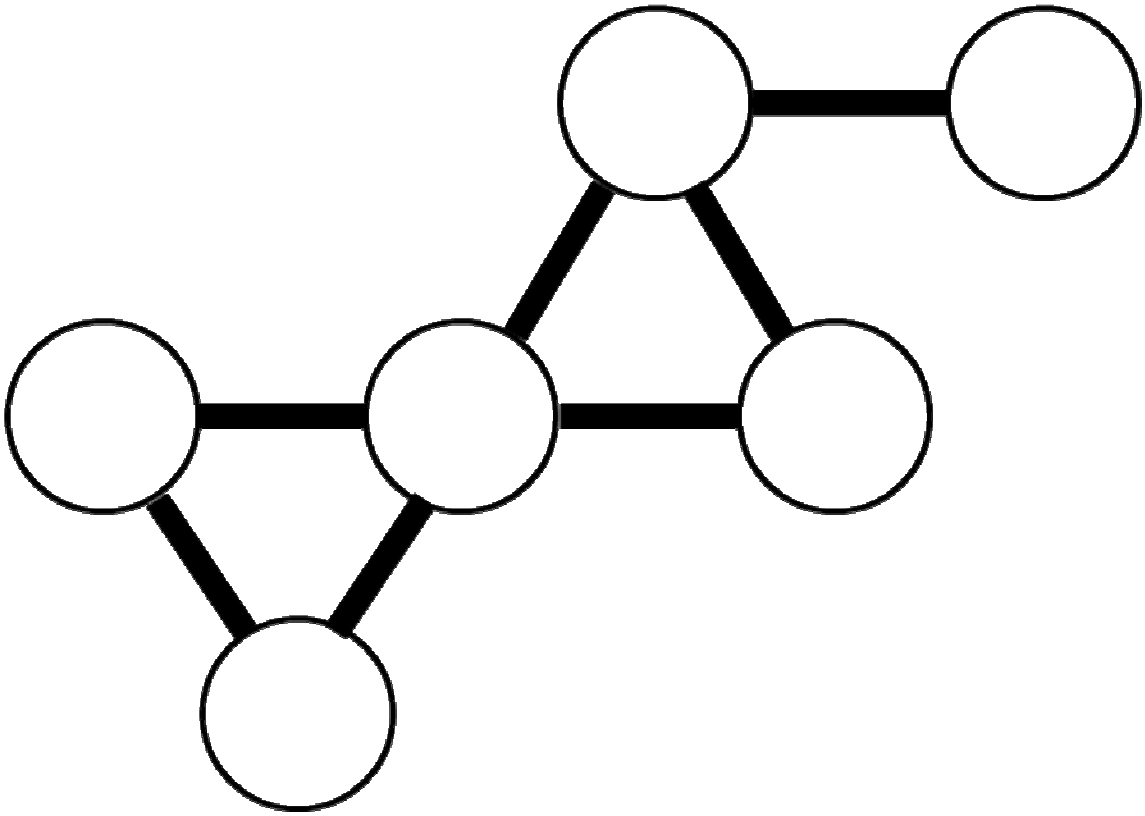

Figure 1: Simple network graph

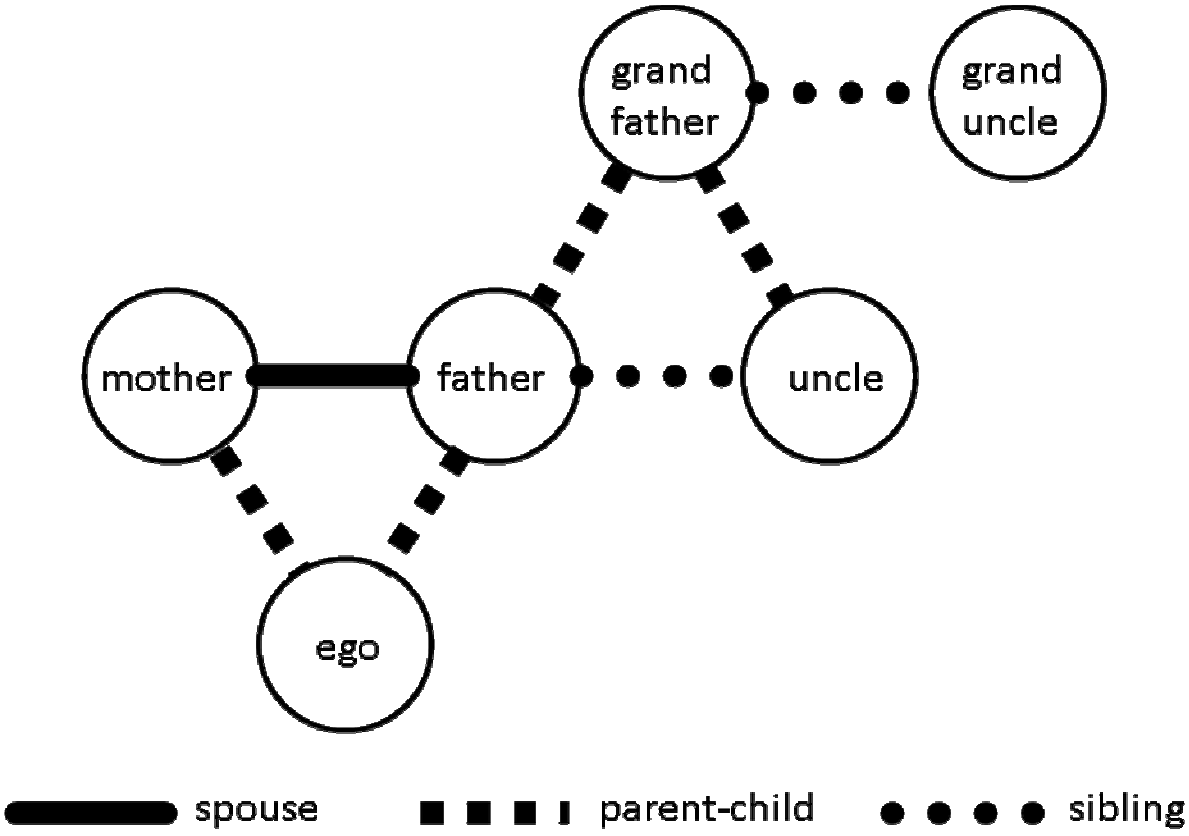

Figure 2: Construction of the ego centric network

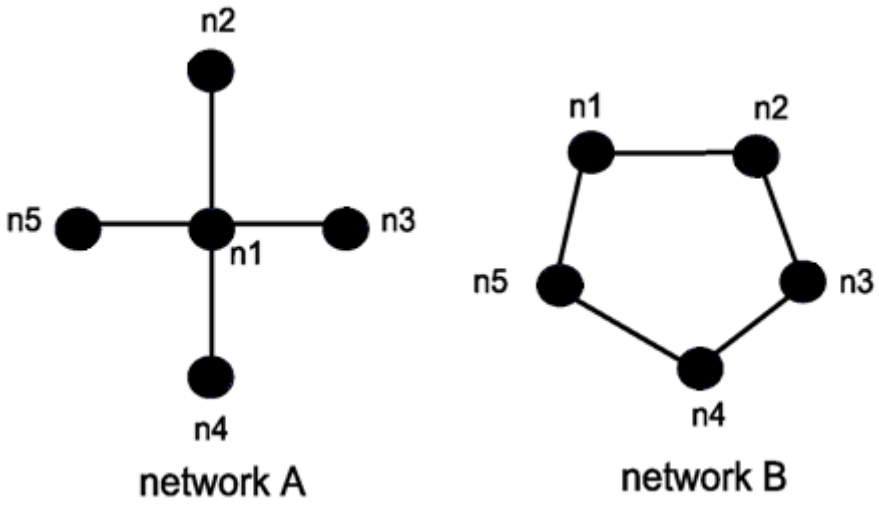

Figure 3: Degree centrality

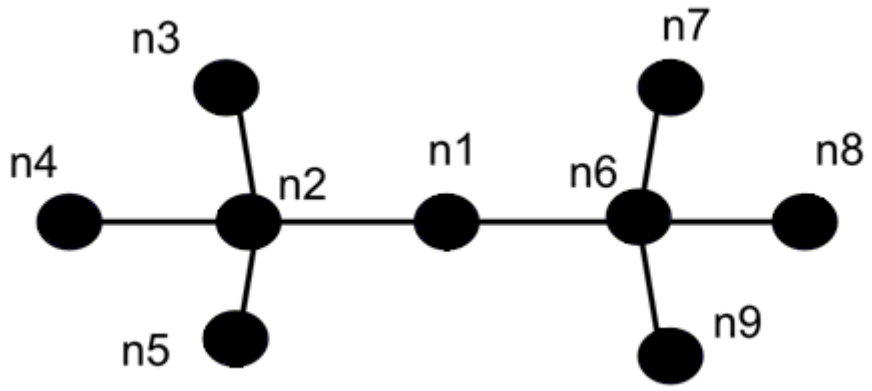

Figure 4: Betweenness centrality

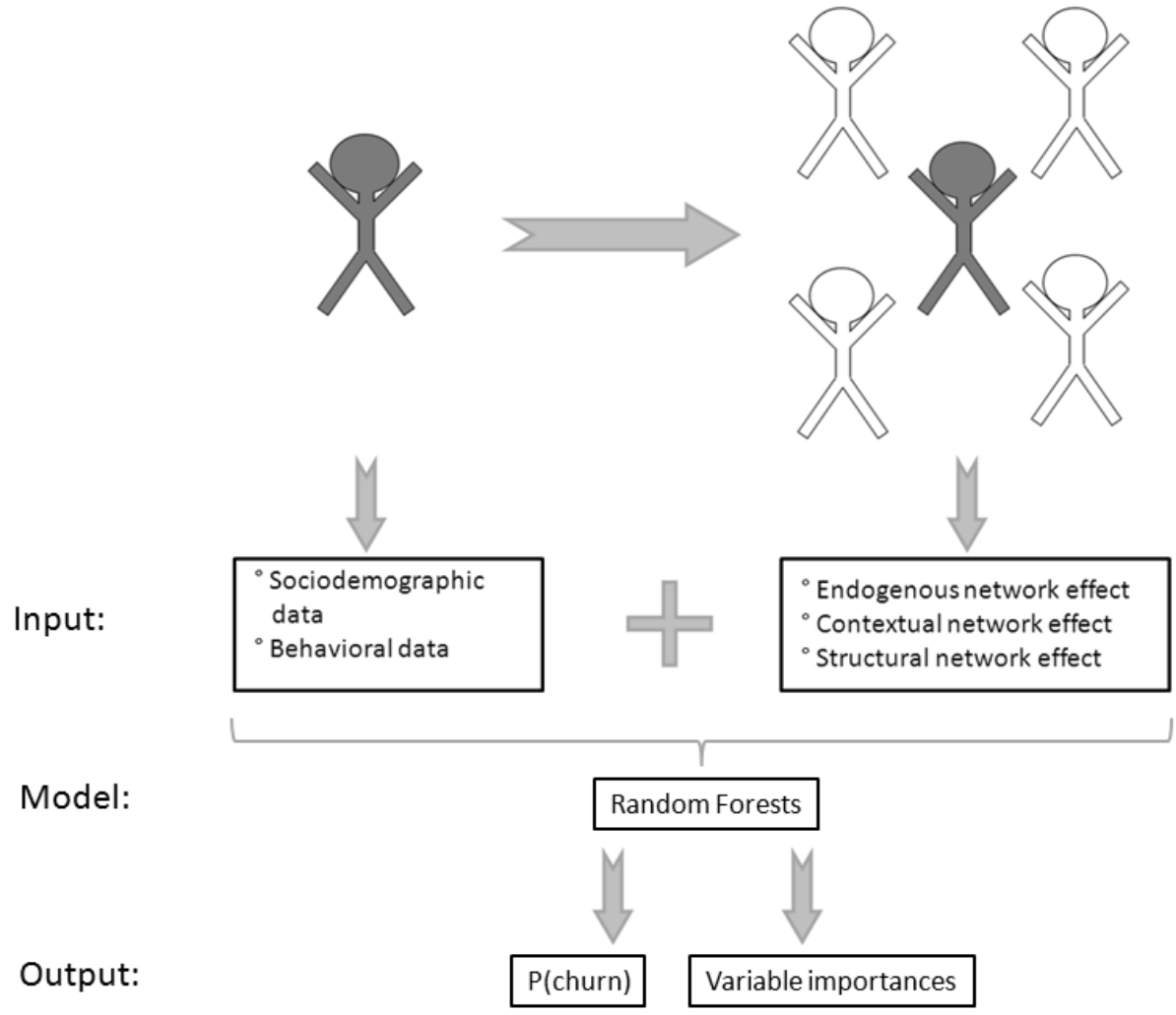

Figure 5: Graphical representation of the methodology

	Variable Name	Description
1	**Age**	
2	**Sex_m**	(male = 1, female = 0)
3	**Social_class_score**	Social class (minimum = 0, maximum = 1000)
4	**Lor**	Length of relationship
5	**Freq**	Amount of purchases in the past
6	**Nbr_cred**	Number of loans
7	**Nbr_insurances**	Number of insurances
8	**Recency**	Time since last purchase
9	**Int_pur_time**	Average time between two purchases
10	**Home_banking**	(yes = 1, no = 0)
11	**Total_passiva**	Total amount borrowed
12	**Total_activa**	Total amount of savings
13	**N_age**	Average age of the network
14	**N_sex_m**	Proportion of male customers in the network
15	**N_Social_Class_Score**	Average social class score of the network
16	**N_lor**	Average length of relationship of the network
17	**N_freq**	Average number of purchases in the past of the network
18	**N_nbr_cred**	Average number of loans of the network
19	**N_nbr_insurances**	Average number of insurances of the network
20	**N_recency**	Average time since last purchase of the network
21	**N_int_pur_time**	Average time between two purchases of the network
22	**N_Home_Banking**	Proportion home banking of the network
23	**N_Total_Passiva**	Average amount borrowed by the network
24	**N_Total_Activa**	Average amount of savings of the network
25	**Churn_1y**	Tests if there is there at least one person in the network who churned last year (1 = true)
26	**Degree_Centrality**	Number of persons a customer is directly connected with
27	**Degree_Centrality2**	Number of persons a customer is indirectly connected with (max. path length = 2)
28	**Density**	Number of links in the second degree of the network of the focal customer
29	**Density_rel**	Number of links in the second degree of the focal customer , divided by the total number of possible links in the second degree network.
30	**Betweenness_c**	Number of times that the focal customer lies on the geodesic path between two other actors of the focal customer's second degree network
31	**Betweenness_c_rel**	Number of times that the focal customer lies on the geodesic path between two other actors of the focal customer's second degree network, divided by the total amount of geodesic paths between two other actors of the focal customer's second degree network

Table 1: Description of variables

Model	Lift			AUC
	5%	10%	12.5%	
Traditional	5.85	4.01	3.48	0.7572
Extended	6.77	4.56	4.02	0.7958

Table 2: Predictive performance

Variable	Importance	Variable	Importance
age	1048.86	**freq**	285.07
recency	989.74	**N_nbr_cred**	254.05
N_age	953.85	**N_sex_m**	249.01
lor	942.94	**nbr_cred**	183.16
Social_Class_Score	900.55	**nbr_insurances**	169.03
N_recency	788.40	**Degree_Centrality2**	159.18
N_lor	780.56	**density**	127.73
N_Social_Class_Score	776.63	**density_rel**	125.43
Total_Passiva	567.65	**sex_m**	121.05
N_int_pur_time	559.08	**betw_c_rel**	104.36
N_Total_Passiva	524.23	**betw_c**	94.12
int_pur_time	517.15	**Degree_Centrality**	87.24
N_Total_Activa	402.52	**N_churn_1y**	67.09
N_freq	392.78	**mis**	43.26
N_nbr_insurances	332.14	**N_Home_Banking**	42.21
Total_Activa	326.77	**Home_Banking**	23.06

Table 3: Importance of variables

Variables compared			Imp_{SNA} vs Imp_{trad}
N_recency	**vs.**	**recency**	0.7966
N_lor	**vs.**	**lor**	0.8278
N_Social_Class_Score	**vs.**	**Social_Class_Score**	0.8624
N_age	**vs.**	**age**	0.9094
N_Total_Passiva	**vs.**	**Total_Passiva**	0.9235
N_int_pur_time	**vs.**	**int_pur_time**	1.0811
N_Total_Activa	**vs.**	**Total_Activa**	1.2318
N_freq	**vs.**	**freq**	1.3778
N_nbr_cred	**vs.**	**nbr_cred**	1.3870
N_Home_Banking	**vs.**	**Home_Banking**	1.8309
N_nbr_insurances	**vs.**	**nbr_insurances**	1.9649
N_sex_m	**vs.**	**sex_m**	2.0570

Table 4: Comparison of importances of contextual variables