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

Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework

Jeroen D'Haen[1]

Dirk Van den Poel[2]

November 2013

2013/863

[1] PhD Candidate at Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Gent, Belgium, http://www.crm.UGent.be
[2] Professor of Marketing Analytics at Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Gent, Belgium, http://www.crm.UGent.be

D/2013/7012/34

Abstract

This article discusses a model designed to help sales representatives acquire customers in a business-to-business environment. Sales representatives are often overwhelmed by available information, so they use arbitrary rules to select leads to pursue. The goal of the proposed model is to generate a high-quality list of prospects that are easier to convert into leads and ultimately customers in three phases: Phase 1 occurs when there is only information on the current customer base and uses the nearest neighbor method to obtain predictions. As soon as there is information on companies that did not become customers, phase 2 initiates, triggering a feedback loop to optimize and stabilize the model. This phase uses logistic regression, decision trees, and neural networks. Phase 3 combines phases 1 and 2 into a weighted list of prospects. Preliminary tests indicate the good quality of the model. The study makes two theoretical contributions: First, the authors offer a standardized version of the customer acquisition framework, and second, they point out the iterative aspects of this process.

Keywords: customer acquisition, sales funnel, prospects, nearest neighbor, decision tree, neural network

1. Introduction

The phrase *customer relationship management* (CRM) is often used in contemporary marketing literature. Although it has been in use since the beginning of the 1990s, researchers have reached no consensus with regard to its definition (Buttle, 2009a; Ngai, 2005; Richards & Jones, 2008). Most definitions have, however, some core features in common; for example, CRM consistently deals with the acquisition and retention of customers and the maximization of long-term customer value (Jackson, 2005; Ngai, Xiu, & Chau, 2009). Prior literature also distinguishes four types of CRM: strategic, operational, analytical and collaborative (Buttle, 2009a). This paper focuses on analytical CRM, which involves mining customer-related data for strategic purposes (Ang & Buttle, 2006; Buttle, 2009a; Ngai et al., 2009), centered on the process of acquiring new customers, and how data mining techniques can facilitate this process.

Most CRM literature neglects customer acquisition in favor of other topics, such as retention (Sohnchen & Albers, 2010), because retention strategies are typically cheaper than acquisition strategies (Blattberg, Kim, Kim, & Neslin, 2008a; Wilson, 2006). However, as important as customer retention might have become, customer acquisition is and should be a crucial focus for companies and researchers for several reasons (Ang & Buttle, 2006; Buttle, 2009b; Kamakura et al., 2005). Startups and companies aiming to exploit new markets need new customers, because they lack existing customers. Even existing companies in a mature market will lose some customers and must replace them (Wilson, 2006). Acquiring new customers is a multistage process, in which only certain suspects (for a definition of the terms used herein, see Section 2) become actual customers, also referred to as the "sales funnel" (Cooper & Budd, 2007; Patterson, 2007; Yu & Cai, 2007). During this process, it is often difficult for sales representatives to cope with all available data (Yu & Cai, 2007). Monat (2011, p. 192) indicates that many companies face this issue:

> "Sales leads are the lifeblood of industrial companies, yet determining which leads are likely to convert to bookings is often based upon guesswork or intuition. This results in a waste of resources, inaccurate sales forecasts, and potential loss of sales. A quantitative model that may be used to predict which leads will convert, based on information inherent in the leads themselves, would be highly valuable."

In response, this article presents a quantitative model, designed to be used as a tool to assist sales representatives in customer acquisition—that is, a sales force automation tool.

Moreover, it is designed to be implemented in a web application, giving it certain specific characteristics and advantages. First, it should be usable regardless of specific company characteristics such as size and industry. Whether for a large company in the automotive sector or a small company in the food sector, the model should render high-quality predictions. Second, it must be fully automated and run without the need for human interference. Third, it must be fast and inexpensive. Because it is a web application, users typically want results immediately.[3] When the algorithm is embedded into a web application, the cost to the user is limited. The user (i.e., a business-to-business [B2B] company) only needs to pay a membership fee to obtain access to the application and does not need to pay for the whole database of prospects, which can be expensive. Moreover, the company does not need in-house experts to analyze the data, as the algorithm performs this step and provides intuitive, ready-to-use output.

Sales representatives must sometimes make arbitrary decisions in selecting prospects from a list of suspects and further qualifying them into leads. Thus, time is lost pursuing bad prospects and leads, violating the famous “time is money” corporate mantra. A model with high predictive power in forecasting the right prospects to pursue can save a company time and, ipso facto, money. Research indicates that approximately 20% of a sales representative’s time is spent selecting prospects (Trailer, 2006) and depicts prospecting as the most cumbersome part of the selling process (Moncrief & Marshall, 2005). Furthermore, making ineffective decisions in the customer acquisition process decreases the overall value of the company over time (Hansotia & Wang, 1997). The proposed algorithm is designed to make the decision-making process less arbitrary by providing model-based prospects.

Although the algorithm should work regardless of the company using it or the industry in which it is situated, note that the proposed sales force automation tool will work best in markets that are highly saturated, in which market penetration is strategically crucial. Moreover, we expect the highest efficiency in markets in which the pool of potential customers is large. In those markets, the selection process is often costly and arbitrary, due to information overload. In contrast, in industries in which customers are large organizations, well-known, and few in number, the proposed algorithm will not provide a significant advantage, because the selection of prospects is limited (Long, Tellefsen, & Lichtenthal, 2007). The algorithm functions in a B2B environment and uses the current customer base of a

[3] We ran the algorithm discussed herein on a 3.40 GHz Windows server containing 16 GB of RAM.

company to predict prospects. It also contains a feedback loop that iteratively improves its overall predictive performance.

There is a limited amount of research on customer acquisition (Blattberg et al., 2008a). With this research, we aim to fill this void and also stimulate further research. The theoretical contributions are twofold. First, we offer a standardized version of the customer acquisition framework. Second, we point out the iterative aspect of this process, which has been neglected in research. The remainder of this article is structured as follows: We present a literature review on customer acquisition, then describe the different stages of our model. After we elaborate on the data, we report the results of the model and finally discuss the conclusions, implications, limitations, and further research suggestions.

2. Customer acquisition framework

The sales funnel conceptualization offers a way to describe the customer acquisition process, dividing it into different stages (Ang & Buttle, 2006; Coe, 2004a; Patterson, 2007; Yu & Cai, 2007). These divisions vary from study to study, as do the definitions they use to characterize each part. A main difference, however, is where the studies place a prospect and a lead in the sales process: Some put the prospect before the lead (e.g., Coe, 2004a; Metzger, 2005), whereas others put the lead before the prospect (e.g., Gillin & Schwartzman, 2011; Patterson, 2007). For the sake of clarity and as a way of creating a standardized framework, we first describe our vision on the sales funnel and define each stage. The emphasis is not on where the different terms are placed but on their definitions.

The darker portion of Figure 1 illustrates the sales funnel. The beginning is a list of suspects. *Suspects* are all potential new customers available. In theory, they could include every other company in a B2B context, apart from the current customer base. In practice, they boil down

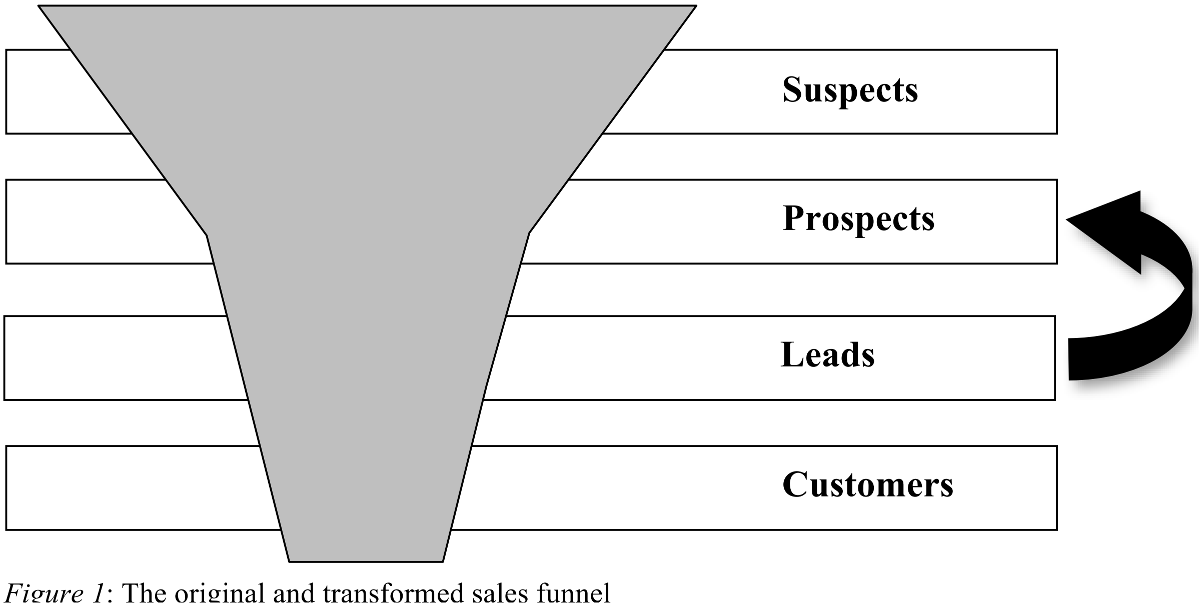


Figure 1: The original and transformed sales funnel

to a limited list of companies (perhaps purchased from specialized vendors; Buttle, 2009b; Rygielski, Wang, & Yen, 2002; Wilson, 2006). The vast amounts of information in those lists tends to overwhelm B2B marketers (Wilson, 2003). As a result, marketers often make selections using a set of arbitrary rules. The outcome of this selection is the list of prospects. *Prospects* are suspects who meet certain predefined characteristics. The next step is to qualify these prospects. *Leads* are prospects that will be contacted, after they have been qualified as the most likely to respond. This qualification is often driven by gut feeling or self-claimed competence. Finally, leads who become clients of the company are *customers*.

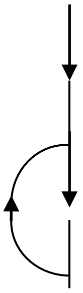
However, current theories and models fail to acknowledge the iterative nature of these stages, which implies none of the different stages is static. Yet the dynamics of this process influence the process itself. First, if customer acquisition is successful, the customer base is altered as new customers get added to it. As a result, these new customers are excluded from the next iteration in the sales funnel. Second, knowledge from a previous iteration should be incorporated in consecutive iterations. The successes and failures in each stage fine-tune the overall process. Here, we focus on the interplay between prospects and leads. The created model alters on the basis of the conversion from prospect to lead, including learning from the new information generated in each iteration. Incorporating the iterative aspect will improve the quality of customer acquisition models.

The procedure we propose radically alters the shape of the sales funnel (the lighter portion of Figure 1), forming an isosceles trapezoid. More prospects are selected, but they are of higher quality. As a result, a greater proportion will be converted into leads and ultimately customers. Furthermore, the algorithm integrates a feedback loop that, over time, further elevates the quality of the prospects. Note that Figure 1 is an exaggerated representation; reality should be somewhere between the graphs, because sales representatives will most likely select a smaller proportion of leads due to time constraints. It is nearly impossible for companies to increase their number of sales calls, assuming sales representatives work close to capacity (Coe, 2004b). The only alternative is to improve the quality of these calls, which is what the proposed algorithm aims to do. It provides high-quality prospects that are easier to convert, as recommended by research showing that call productivity can be improved by the use of information technology tools (Ahearne, Hughes, & Schillewaert, 2007; Eggert & Serdaroglu, 2011).

Traditionally, the conversion rate from prospects to qualified leads is approximately 10% on average (Coe, 2004b). Thus, getting a good list of prospects saves time that then can be spent qualifying them. Moreover, better qualified leads should lead to a higher customer conversion rate. Usually, a conversion rate from prospects to customers of 1%–5% on average can be expected (Coe, 2004b). Research shows that a lower conversion rate increases the cost of customer acquisition (Blattberg et al., 2008a). Thus, raising the conversion rate will also lower the cost of customer acquisition.

3. Proposed model

The model contains three phases that must be executed chronologically. Phase 1 runs when there are data only on the current customers. The model must indicate hidden structures in the data without the presence of feedback data (i.e., a dependent variable). Therefore, unsupervised learning is necessary. The input of phase 1 is data on a list of suspects and the current customers of a company. The output is a list of ranked prospects. As soon as there are data on which prospects were or were not qualified as leads, phase 2 initiates. The model uses this feedback data and supervised learning methods, such as logistic regression, decision trees, and neural networks. Phase 3 combines phases 1 and 2 into a weighted list of prospects. The output of phase 3 generates more feedback data, which in turn are fed into phase 2, initializing a feedback loop. That is:



Phase 1: This is the base (default) model. It uses the current customer base to predict new potential customers.

Phase 2: When the outcome of phase 1 has been used to qualify leads, there is a possibility for feedback. Phase 2 uses this feedback to improve the predictions.

Phase 3: This phase is a combination of phases 1 and 2.

Every model uses an estimation and validation sample to prevent overfitting and to calculate the area under the receiver operating characteristic curve, also known as the AUC (Blattberg, Kim, Kim, & Neslin, 2008c, 2008d). The AUC is a common metric to evaluate the accuracy of a model (Chen, Hsu, & Hsu, 2011). It can vary from 0.5 (random model) to 1 (perfect model) (Baecke & Van den Poel, 2011; Blattberg et al., 2008d). The data set is randomly distributed over the estimation and validation sample, with a ratio of two-thirds and onethirds, respectively, as Blattberg et al. (2008d) suggest. The estimation sample is used to compute the models, whereas the validation sample tests the predictive performance of these models.

3.1 Phase 1

The key problem of customer acquisition is that, in the beginning, the current customer base in combination with a suspect list represents the only inputs, so no supervised learning can be applied. A solution is to conduct a profiling model, also known as a look-alike model (Blattberg et al., 2008a; Jackson, 2005; Setnes & Kaymak, 2001; Wilson, 2006). To acquire new customers, sales representatives must know in detail who their own customers are (Ngai et al., 2009). Profiles are created according to the current customer base, and these profiles are subsequently used to predict prospects (Bose & Chen, 2009; Chou, 2000). This method is a type of clustering, in which identical prospects are put in the same cluster rather than the center of the cluster being a current customer. The cluster continues to expand with less similar prospects, with a measure of (dis)similarity assigned to these prospects. This procedure creates concentric circles, and in each circle, we find prospects that have the same similarity to the center (being a specific current customer). The more distant a circle is, the more dissimilar the prospects are on that circle. Prospects in the same cluster or circle share comparable preferences and behaviors (Bruckhaus, 2010). As a result, we assume that finding prospects that are similar to the current customer base increases the probability that these prospects become future clients of the company, compared with less similar prospects, because they share the same company preferences. Kim et al. (2005) use this profiling method in a business-to-consumer (B2C) environment. They build a model on their current customers and use that model on potential prospects to rank them from most to least likely to respond. Here, we apply it in a B2B environment.

A profile is composed of a combination of variables (Hansotia & Wang, 1997). The profiles of the prospects are compared with those of the current customers. The technique used here to search for similar profiles is the nearest neighbor algorithm. This method is conceptually simple; it involves calculating the distance between observations using a set of variables. The more similar the cases are, the lower the distance. The advantage of this algorithm is that it is powerful yet easy to understand (Weinberger & Saul, 2009). Figure 2 is a simplified presentation of the nearest neighbor algorithm. In the two-dimensional space shown (representing a profile of two variables), company C is closer to company A than company B is to A, which means that company C is more similar to company A than company B is. The reality is more complex though, in that there is a multidimensional space rather than a two-dimensional one. The method we apply here is a k-nearest neighbor algorithm, meaning that for each current customer, it ranks the k-nearest prospects. We set k arbitrarily to 10,000. The size of k is not that important, as long as it is set high enough. The larger the number, the larger the list of outputted prospects. However, this list can be reduced, such as by selecting only prospects with a similarity higher than a predefined threshold. A different, more recommended strategy is to rank the list first on similarity and then on a different variable of interest (e.g., company size). More ranking variables can be added to further refine this ranking. Next, the top *n* prospects of the list are selected, with *n* being the maximum amount of prospects that the sales representatives are able to handle. Thus, we advise practitioners to set k > *n* and refine the ranking by adding variables that are relevant to the company of interest.

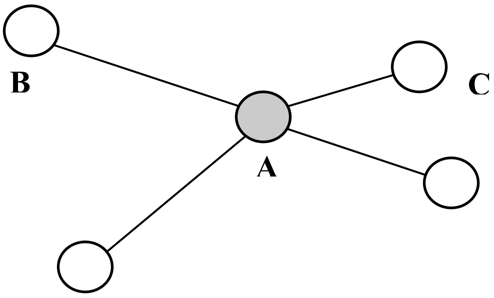


Figure 2: Nearest neighbor

The most important element of a nearest neighbor analysis is the distance metric, which calculates how similar companies are. Thus, it is crucial for the quality of the model. Distance metrics are data type specific: there is no easy way to combine categorical and numeric data types in one nearest neighbor. Because most of the variables are categorical, numeric ones are converted into categories (for more information, see Section 4). The Jaccard and Hamming distance measures are two possible distance metrics for categorical data (Ichino & Yaguchi, 1994). The Jaccard similarity coefficient is obtained by dividing the size of the set of variables that have the same value by the size of the set of variables that do not have the same value (Charikar, 2002). The formula is as follows (where A and B signify companies):

$$\mathrm{S}_{\mathrm{Jaccard}}\,(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$$

The Hamming metric is similar (Steane, 1996). However, the Jaccard metric ignores variables that have a zero for both companies, whereas the Hamming metric does not (Zytynska, Fay, Penney, & Preziosi, 2011). Because in the used data, a zero usually stands for a missing value, this comparison should be ignored; thus, we prefer the Jaccard metric. Although several distance metrics exist for numeric variables, such as the Euclidean and Mahalanobis distances, there is no generally accepted preference. Aggarwal (2001) suggest that fractional distance metrics work better than others when dimensionality is high. The output of phase 1 is a list of prospects with their respective similarities (ranging from 0 to 1, with 1 being completely the same with regard to the set of variables and 0 being completely different).

3.2 Phase 2

As mentioned previously, phase 2 can only be implemented after phase 1 has rendered positive and negative feedback (see the feedback loop in Figure 1), which is used as a dependent variable. Thus, the model in Phase 2 uses the prospect list of phase 1, including the feedback data on those prospects, and the reference database (for more information, see Section 2). By adding this second phase to the algorithm, we incorporate an iterative customer acquisition process. Each time its output has been evaluated, the feedback is inserted into the algorithm, and it re estimates the model. The process gradually optimizes and stabilizes the model.

A basic model to predict customer acquisition is (logistic) regression (Bose & Chen, 2009; Gupta et al., 2006; Hansotia & Wang, 1997), the formula for which is as follows:

$$F(z) = \frac{1}{1+e^{-z}} \quad \text{where} \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots +$$

(Blattberg et al., 2008a; Hansotia & Wang, 1997; Pampel, 2000; Van den Poel & Buckinx, 2005). Because there is a danger of overfitting the model when using all possible independent variables, we apply a stepwise selection (i.e., the combination of a forward and backward selection) (Blattberg et al., 2008d; Kim et al., 2005). We also include variable transformations (e.g., taking the square of variables) to take nonlinearity and skewed distributions into account. A problem with logistic regression is that it cannot use categorical variables, only continuous ones (Pampel, 2000). We solve this problem using dummy variables. However, the large number of categorical variables could lead to an overload of dummies, which is a computational burden (Bose & Chen, 2009); moreover, no a priori knowledge is available about which categorical variables are likely crucial to include in the model. Thus, the logistic model only incorporates continuous variables.

Therefore, we estimate a model using the categorical variables as well, including categorized versions of the continuous variables. We created the categories using equal frequency binning: the different categories of a variable have the same size, and they are based on the ranking of the values of this variable, the preferred technique for discretizing the variables for commercial data, which are often unbalanced or contain outliers (Cantu-Paz, 2001). We apply a decision tree to estimate the model, an efficient method for estimating categorical input variables (Bose & Chen, 2009). It involves dividing a data set into subsets, using the values of the independent variables as selection criteria to predict the dependent variable (Blattberg, Kim, Kim, & Neslin, 2008b). It then involves dividing the data into homogeneous subsets that are heterogeneous to each other, while minimizing the cost of this division (Danielson & Ekenberg, 2007). The top of a decision tree is called the *root node* (Berk, 2008). This root node contains the full data set. The outcome of a decision at each node is called a *split* (Duda, Hart, & Stork, 2001). Splits after the root node are termed *branches*, and the final splits are the *terminal nodes*. All splits after the initial split imply interaction effects, unless they use the same predictor (Berk, 2008). We use pruning to find the right size of the tree to avoid the omnipresent problem of overfitting: the bigger a tree is, the fewer cases there are in the terminal nodes and the more chance there is of having an overfitted tree. Pruning a tree begins at the terminal nodes and works up to the top (Berk, 2008). It eliminates nodes that do not reduce heterogeneity enough compared with the complexity they add to the tree. Occam's razor prescribes that researchers should prefer the simplest model that explains the data (Baesens, Mues, Martens, & Vanthienen, 2009; Duda et al., 2001). The decision tree and its pruning method are based on Breiman et al. (1984). We use a majority voting scheme to calculate the probabilities of the decision tree. We then calculate the probabilities by taking the percentage of ones in each ending node. Figure 3 presents a simple tree.

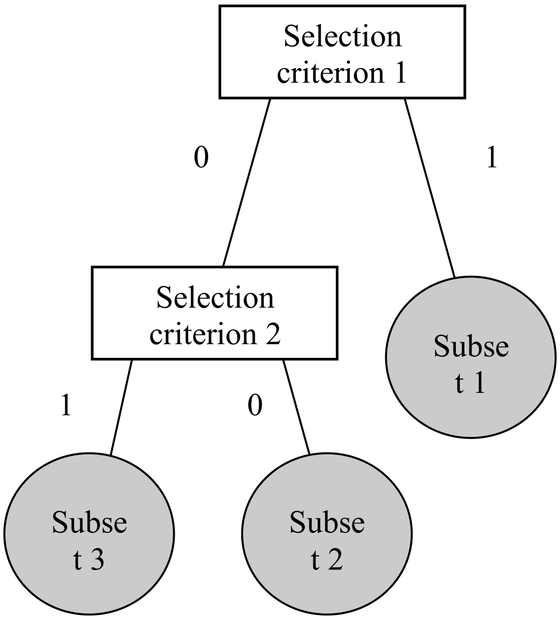


Figure 3: Decision tree

We calculate the AUC to determine whether to use the logistic model or the decision tree and choose the model with the highest AUC. We include both logistic regression and a decision tree, because there is no a priori hypothesis for which model works best. Furthermore, it might be company or industry specific. (Recall the stipulation that the algorithm must run fully automatically without human interference.)

We incorporate a backup model that runs if both the logistic model and the decision tree fail to produce a model that predicts better than a random ranking of prospects (i.e., a model of an AUC of 0.5): a neural network. The reason we use it only as a back-up is that it is relatively slow and unstable (Rygielski et al., 2002), and the algorithm must provide fast and reliable results. A neural network is a nonlinear nonparametric regression model that mimics the structure and function of the brain (Ha, Cho, & Mela, 2005). It is a black box method, in that it provides no information on the estimated model. The input generates a certain output, and the way this output is generated remains hidden from the user. The main advantage of neural networks is that they are capable of estimating very complex relationships.

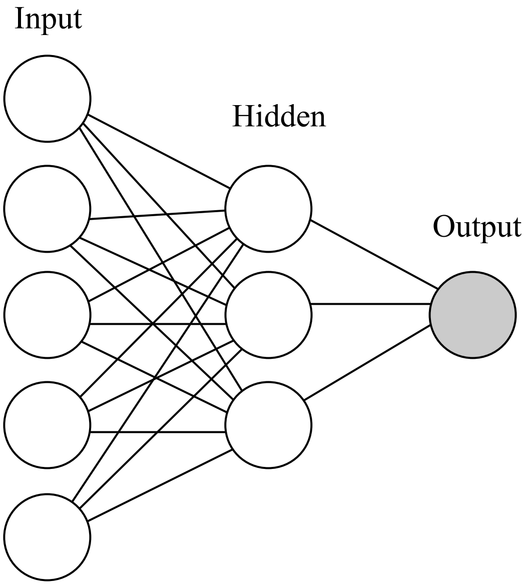


Figure 4: Neural network

A neural network usually contains an input layer, a hidden layer, and an output layer (Figure 4). The input layer corresponds to the independent variables, and the output layer is the dependent variable. The hidden layer represents the nonlinearity of the model. Multiple hidden layers can be introduced, but one hidden layer is deemed enough to obtain quality estimations (Ha et al., 2005). The neural network is implemented in Matlab and is a feed-forward network. For the input, hidden, and output layers, the purelin, tansig, and purelin transfer functions are applied, respectively. The hidden layer size is varied from 1 to 10 neurons selecting the one rendering the highest AUC. The output of phase 2 is the list of prospects of phase 1 with their respective predicted probability.

3.3 Phase 3

Prior literature indicates that predictability can be improved by weighting the predictions from different models (Gupta et al., 2006). Combining models can partially eliminate the bias inherent in each model (Bose & Chen, 2009). The AUC calculated in phase 2 (i.e., the AUC of the best model) is used to assign the weights in phases 1 and 2. We apply the following linear function to calculate the weight of phase 2:

$$\omega_{Phase\ 2} = (AUC - 0.5)\ *\ 2$$

The weight of phase 1 is naturally computed as follows:

$$\omega_{Phase\ 1} = 1 - \ \omega_{Phase\ 2}$$

Table 1 portrays some AUCs between 0.5 and 1 and their respective weights for phases 1 and 2. The function used to calculate the weights is conservative in the sense that it requires a relatively high AUC before phase 2 weights more than phase 1. The output of phase 3 is the prospect list generated in phase 1 and the weighted similarity.

AUC phase 2	0.5	0.6	0.7	0.8	0.9	1
Weight phase 2	0	0.2	0.4	0.6	0.8	1
Weight phase 1	1	0.8	0.6	0.4	0.2	0

Table 1: AUC and weights

In summary, phase 1 generates a list of prospects with their similarity; some prospects will be qualified as leads, while others are not; this feedback is entered into phase 2, and the algorithm calculates a new similarity (probability); phase 3 defines the weights of phase 1 and 2 and produces a final prospect list (see Figure 5).

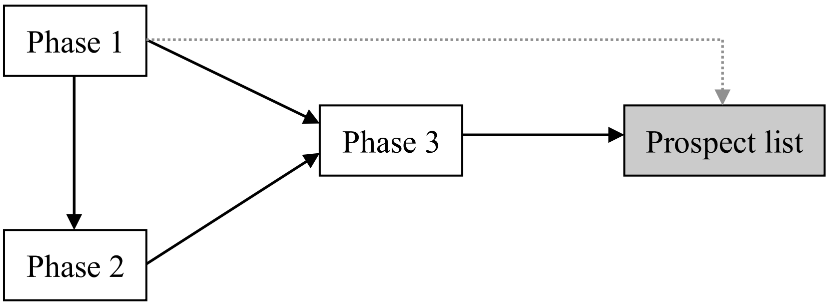


Figure 5: Overview of the algorithm

Phase 3 combines the similarities of phase 1 and the probabilities of phase 2. Even though they are not the same measure, they represent the same idea. More specifically, the higher a prospect is ranked in the list, the more likely this prospect is to become a customer. This justifies combining two different measures because they measure the same content. Furthermore, they both range between 0 and 1, making a combination simple and straightforward.

4. Data

We leased a database of more than 16 million U.S. companies from an international data provider (hereinafter referred to as the reference database). It represents the list of suspects, after excluding current customers. It contains a selection of 4 numerical and 24 categorical variables (see the Appendix, Table A.1). Moreover, we created four additional variables, representing the discretized versions of the numeric variables. Some literature exists on which variables are relevant in profiling models. Industrial demographic data are often used to prospect new potential customers (Bounsaythip & Rinta-Runsala, 2001). Two basic demographic variables of companies are industry type and company size (Coe, 2004c). However, to our knowledge, no research addresses the full range of relevant industrial demographic variables, and it is likely that these variables will be industry or even company specific. Therefore, we included as many variables as possible, because the algorithm must perform well regardless of the company using it. We used three criteria to exclude variables:

1. Redundant variables: Redundant variables are highly correlated with other variables. Including them in a nearest neighbor analysis would artificially assign them more weight. Because we make no hypotheses about variable importance, this would be detrimental to the quality of the analysis.
2. Name-based variables: Name-based variables are mainly general company variables that have no predictive power, such as the chief executive officer name and company name. For example, the fact that a company is called Apple has no predictive performance as such. The connotation and familiarity of the name might influence customer acquisition, but the specific letters do not. If we were to run a nearest neighbor algorithm with Apple as a current customer, Applebee's would be a relatively good match based on the name variable. It is however unlikely that Applebee's will be evaluated as a good prospect, due to the large difference between the two companies.
3. Variables containing a high percentage of missing values: We excluded variables with more than 50% of missing values. For the retained variables, we did not infer missing values, which might insert bias in the data (Han & Kamber, 2006).

The B2B company that serves as a test case for the algorithm is active in telecommunication services and was founded in 1997. It is based in the United States and is one of the leaders in its market. The platform the firm developed handles more than 1 billion calls a year and has deployed more than 750 tailored solutions for customers. The telecom company has 389 active current customers, of whom we selected 107 as input for the algorithm. We deleted companies that had a large amount of missing data or that could not be matched in the reference database. The matching with the reference database is necessary because we extracted the variables of the current customers from this database.

In summary, we used two types of data: the reference data set and the telecom company data set. The reference data set is a database containing variables on more than 16 million U.S. companies. We used this database as a list of suspects, which is the input of the algorithm (excluding the current customers of the company). The telecom data set contains the customers of the telecom company, without any variables on these customers. We extracted variables of the telecom company customers from the reference database.

5. Results

The sales funnel of B2B companies is more complex than that of B2C companies (Yu & Cai, 2007). More processes are needed to complete transactions, and, as a result, deals take longer to close. Thus, it is difficult to conduct an extensive real-life test in a B2B setting. We were, however, able to do a (limited) real life test of the algorithm.

We inserted the current customers of the company in the algorithm as input for phase 1. This rendered a list of prospects sent to the telecom company. The company reviewed this list and qualified prospects into "good" and "bad" leads. The list was first ranked on similarity and then on company sales volume, which was a relevant ranking variable according to the company sales manager. The sales representatives selected the top 356 prospects to evaluate. Of these, 56 companies were qualified as good leads, corresponding to a conversion rate from prospect to lead of 15.73% [= 56/(56 + 300)], higher than the overall conversion rate of 10% on average (Coe, 2004b). Next, we administered two pseudo tests to determine the quality of the algorithm. Although they are not real-life tests, they use real data.

In the first test, we used the positively qualified prospects as input to the algorithm. Here, we employed a reverse logic to test the model. We used the 56 positively qualified prospects received from the telecom company as input to find the 107 original telecom company customers. Using an (arbitrary) selection rule of retaining prospects with the highest similarity (to the 56 positively qualified prospects that we used as input), we retained 228 potential prospects, of which 8 were original telecom company customers. Assuming that these 8 prospects would become company clients, we obtain a conversion rate of prospect to customer of 3.5% (= 8/228), similar to what can be expected on average (Coe, 2004b). However, this is obtained by only running phase 1. We expect that running phases 2 and 3 will elevate this conversion rate by including feedback data.

The second test assesses the combination of the three phases and their ability to find specific companies, mainly as a test of the efficiency of the feedback loop. This test does not use the telecom company data, only the reference data. We selected companies with the following random profile from the reference database (the interpretation of the variables is not relevant here): sales volume > $100 000 and ≤ $190 000; number of employees > 4 and ≤ 50; square footage estimator >2 210 and ≤ 3319; import export indicator = 2; population code > 4; and active in the accommodations and food services industry. This rendered a list of 36 companies. We then randomly selected 10 companies as current customers and ran the model to search for the other 26 profile customers. In other words, these 26 companies are "hidden" in the reference database, and the goal is to find them. In each run, the algorithm chose prospects that had the highest similarity, regardless of how big this selection was. The first run only used the nearest neighbor algorithm, because no feedback data were available yet (Table 2). Ten prospects had the highest similarity, and all of these were part of the 26 profile customers. In the second run, again only the nearest neighbor algorithm ran, because the previous run gave only positive feedback points and no negative ones. Of the 123 prospects with the highest similarity, none were profile customers. Run 3 rendered 10,176 prospects, of which 12 were profile customers. The selected method in phase 2 was a decision tree (Appendix, Figure A.1). The AUC of the decision tree was 1. Run 4 provided two more prospects, both of which were profile customers. A decision tree again represented the selected method (Appendix, Figure A.2), with an AUC of 0.99985. Additional runs did not reveal the remaining two customers, most likely because run 4 did not add a great deal of feedback data to the model (only two feedback points).

Run	1	2	3	4
Number of prospects	10	123	10176	2
Number of profile customers	10	0	12	2
Phase 2: Selected method	Only Phase 1	Only Phase 1	Decision Tree	Decision Tree
AUC	/	/	1	0.99985

Table 2: Results of profile searching

6. Conclusions and implications

This article presents a procedure to facilitate the customer acquisition process in a B2B environment. The algorithm contains three phases, and the output is a ranked list of prospects. Sales representatives could select a top percentage of these ranked prospects to qualify further as leads to pursue. Because these prospects are higher quality, it is easier for sales representatives to qualify and, in turn, convert them into customers. Real-life and pseudo tests show positive results. The real-life test suggests a conversion rate from prospect to lead that is higher than average. The first pseudo test produced a conversion rate from prospect to customer similar to the average conversion rate by only using the first phase of the algorithm. The second pseudo test needs only four runs to find 24 of 26 companies in a suspect list that contains more than 16 million companies.

This study provides several managerial implications. First, the proposed sales force automation tool operates in a fully automated way, but human intervention remains possible, when necessary. As a result, the tool can work in a broad range of situations. It supports sales managers from a starting position, in which there is merely a basic set of current customers and no information on the acquisition process, to a situation in which the customer base is more mature and a vast amount of data is available on the history of this process. However, human intervention might be preferable in some cases. Look-alike models tend to overlook opportunities in other segments (Blattberg et al., 2008a), which is inherent to the method, in that it searches for new prospects similar to the current customers. As a result, it is not always optimal to include the full set of variables. For example, the industry (NAICS code) can be withheld from the algorithm to find prospects in different industries as well.

Second, the output of the algorithm can be used straightforwardly without any knowledge of the statistical models running in the background. Thus, its applicability does not rely on any human expertise, such that it lowers the threshold for sales representatives to use this tool. Furthermore, research has shown that the efficiency of sales representatives using sales force automation tools is only augmented when it is accompanied by user training and support (Ahearne, Jelinek, & Rapp, 2005). Because this tool can intuitively be used and no significant training is necessary, the cost and time of such support is marginal, making it more likely that B2B sales managers will implement it and that this implementation is fluent.

Third, the tool could help sales managers negotiate with a data vendor to pay for only the prospects indicated by the sales force automation tool and not the whole list of suspects. The tool can also be embedded into a web application, limiting the costs (see Section 1). However, even if a data vendor was already willing to sell a selection of prospects on the basis of some arbitrary rules instead of a list of suspects, sales managers or the vendors themselves could improve the selection using the proposed algorithm.

Fourth, this study offers an explicit iterative view of the customer acquisition process. Each iteration provides useful information for the next. Therefore, there is a need for an extensive documentation when sales managers attempt to acquire new customers. Information on decisions made, steps taken, strategies employed, and so on must be recorded and analyzed periodically. This way, new customer acquisition can be improved incrementally.

This iterative view is also a theoretical implication. The shift from a static to a dynamic framework is a more accurate conceptualization of reality. When designing models using a customer acquisition framework, modelers should take the iterative aspect into account, which has been neglected to date. A different but related theoretical implication is the need for a standardized customer acquisition framework. We provide a personal, though literature-based, view on the flow between the different acquisition stages and their respective definitions. It is by no means meant as an ultimate framework, but rather as a tool tailored for our purposes.

7. Limitations and further research

The main limitation of this study is that it was not possible to run a full, real-life test of the algorithm. Such a test is necessary to fully validate the model. Therefore, further research should first involve an extensive, real-life test using the suggested algorithm. If these tests prove the model valid, adjustments can be made to improve it further. A possible avenue of study is to make a distinction within the current customer base between good and bad customers and give corresponding weights to them when running the algorithm. The distinction between good and bad customers might be based, for example, on profitability, because research shows that customers are not equally profitable (Jacobs, Johnston, & Kotchetova, 2001). Another possible avenue is to include other data sources into the model, because the success of a model depends partly on the data input (Baecke & Van den Poel, 2012a, 2012b). For example, web data have proven to be strong predictors of profitable customers (D'Haen, Van den Poel, & Thorleuchter, 2012; Thorleuchter, Van den Poel, & Prinzie, 2012).

References

Aggarwal, C., Hinneburg, A., & Keim, D. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Lecture Notes in Computer Science* (pp. 420-434). London: Springer-Verlag.

Ahearne, M., Hughes, D. E., & Schillewaert, N. (2007). Why sales reps should welcome information technology: Measuring the impact of CRM-Based IT on sales effectiveness. *International Journal of Research in Marketing, 24,* 336-349.

Ahearne, M., Jelinek, R., & Rapp, A. (2005). Moving beyond the direct effect of SFA adoption on salesperson performance: Training and support as key moderating factors. *Industrial Marketing Management, 34,* 379-388.

Ang, L. & Buttle, F. (2006). Managing for successful customer acquisition: an exploration. *Journal of Marketing Management, 22,* 295-317.

Baecke, P. & Van den Poel, D. (2011). Data augmentation by predicting spending pleasure using commercially available external data. *Journal of Intelligent Information Systems, 36,* 367-383.

——— (2012a). Improving customer acquisition models by incorporating spatial autocorrelation at different levels of granularity. *Journal of Intelligent Information Systems*, forthcoming.

——— (2012b). Including spatial interdependence in customer acquisition models: A cross-category comparison. *Expert Systems with Applications, 39,* 12105-12113.

Baesens, B., Mues, C., Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: upcoming trends and challenges. *Journal of the Operational Research Society, 60,* S16-S23.

Berk, R. A. (2008). Classification and regression trees (CART). In *Statistical learning from a regression perspective* (pp. 103-166). London: Springer Verlag.

Blattberg, R. C., Kim, P., Kim, B. D., & Neslin, S. A. (2008a). Acquiring customers. In *Database marketing: Analyzing and managing customers* (pp. 495-514). London: Springer Verlag.

——— (2008b). Decision trees. In *Database marketing: Analyzing and managing customers* (pp. 423-441). London: Springer Verlag.

——— (2008c). The predictive modeling process. In *Database marketing: Analyzing and managing customers* (pp. 245-286). London: Springer Verlag.

——— (2008d). Statistical issues in predictive modeling. In *Database marketing: Analyzing and managing customers* (pp. 291-321). London: Springer Verlag.

Bose, I. & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research, 195,* 1-16.

Bounsaythip, C. & Rinta-Runsala, E. (2001). *Overview of data mining for customer behavior modeling*. Otaniemi, Finland: VTT Information Technology.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Covington, KY: Wadsworth International Group.

Bruckhaus, T. (2010). Collective intelligence in marketing. In J.Casillas & F. J. Martínez-López (Eds.), *Marketing intelligence systems using soft computing: Managerial and research applications* (pp. 131-154). London: Springer Verlag.

Buttle, F. (2009a). Introduction to customer relationship management. In *Customer relationship management: concepts and technologies* (2nd ed., pp. 1-23). London: Taylor & Francis.

——— (2009b). Managing the customer lifecycle: Customer acquisition. In *Customer relationship management: Concepts and technologies* (2nd ed., pp. 225-254). London: Taylor & Francis.

Cantu-Paz, E. (2001). Supervised and unsupervised discretization methods for evolutionary alghorithms. *Proceedings of the Genetic and Evolutionary Computation Conference.* San Francisco: Association for Computing Machinery.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on theory of computing* (pp. 380-388). New York: Association for Computing Machinery.

Chen, W. C., Hsu, C. C., & Hsu, J. N. (2011). Optimal selection of potential customer range through the union sequential pattern by using a response model. *Expert Systems with Applications, 38,* 7451-7461.

Chou, P. B. (2000). Identifying prospective customers. In *Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 447-456). New York: Association for Computing Machinery.

Coe, J. M. (2004a). Segmentation for communications. In *The fundamentals of business to business sales and marketing* (pp. 71-94). New York: McGraw-Hill.

——— (2004b). The integration of direct marketing and field sales to form a new B2B sales coverage model. *Journal of Interactive Marketing, 18,* 62-77.

——— (2004c). The start: Profiling and targeting the market. In *The fundamentals of business to business sales and marketing* (pp. 51-69). New York: McGraw-Hill.

Cooper, M. J. & Budd, C. S. (2007). Tying the pieces together: A normative framework for integrating sales and project operations. *Industrial Marketing Management, 36,* 173-182.

D'Haen, J., Van den Poel, D., & Thorleuchter, D. (2012). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert systems with applications,* forthcoming, DOI: 10.1016/j.eswa.2012.10.023.

Danielson, M. & Ekenberg, L. (2007). Computing upper and lower bounds in interval decision trees. *European Journal of Operational Research, 181,* 808-816.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Non-metric methods. In *Pattern Classification* (2nd ed., pp. 1-66). New York: Wiley.

Eggert, A. & Serdaroglu, M. (2011). Exploring the impact of sales technology on salesperson performance: A task-based approach. *Journal of Marketing Theory & Practice, 19,* 169-186.

Gillin, P. & Schwartzman, E. (2011). Lead generation. In *Social Marketing to the Business Customer: Listen to your B2B market, generate major account leads, and build client relationships* (pp. 156-175). Hoboken, NJ: Wiley.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N. et al. (2006). Modeling customer lifetime value. *Journal of Service Research, 9,* 139-155.

Ha, K., Cho, S., & Mela, C. F. (2005). Response models based on bagging neural networks. *Journal of Interactive Marketing, 19,* 17-30.

Han, J. & Kamber, M. (2006). Data preprocessing. In *Data mining: Concepts and techniques* (2nd ed., pp. 47-104). Amsterdam: Elsevier.

Hansotia, B. J. & Wang, P. (1997). Analytical challenges in customer acquisition. *Journal of Direct Marketing, 11,* 7-19.

Ichino, M. & Yaguchi, H. (1994). Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE transactions on Systems, Man and Cybernetics, 24,* 698-708.

Jackson, T. W. (2005). CRM: From "art to science." *Journal of Database Marketing & Customer Strategy Management, 13,* 76-92.

Jacobs, F. A., Johnston, W., & Kotchetova, N. (2001). Customer profitability: Prospective vs. retrospective approaches in a business-to-business setting. *Industrial Marketing Management, 30,* 353-363.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R. et al. (2005). Choice models and customer relationship management. *Marketing Letters, 16,* 279-300.

Kim, Y. S., Street, W. N., Russel, G. J., & Menczer, F. (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science, 51,* 264-276.

Long, M. M., Tellefsen, T., & Lichtenthal, J. D. (2007). Internet integration into the industrial selling process: A step-by-step approach. *Industrial Marketing Management, 36,* 676-689.

Metzger, M. (2005). Using water testing to convert prospects into leads and leads into customers. *WC&P International, 47,* 7-8.

Monat, J. P. (2011). Industrial sales lead conversion modeling. *Marketing Intelligence & Planning, 29,* 178-194.

Moncrief, W. C. & Marshall, G. W. (2005). The evolution of the seven steps of selling. *Industrial Marketing Management, 34,* 13-22.

Ngai, E. W. T. (2005). Customer relationship management research (1992-2002): An academic literature review and classification. *Marketing intelligence & planning, 23,* 582-605.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36,* 2592-2602.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage Publications.

Patterson, L. (2007). Marketing and sales alignment for improved effectiveness. *Journal of Digital Asset Management, 3,* 185-189.

Richards, K. A. & Jones, E. (2008). Customer relationship management: Finding value drivers. *Industrial Marketing Management, 37,* 120-130.

Rygielski, C., Wang, J., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society, 24,* 483-502.

Setnes, M. & Kaymak, U. (2001). Fuzzy modeling of client preference from large data sets: An application to target selection in direct marketing. *IEEE Transactions on Fuzzy Systems, 9,* 153-163.

Sohnchen, F. & Albers, S. (2010). Pipeline management for the acquisition of industrial projects. *Industrial Marketing Management, 39,* 1356-1364.

Steane, A. M. (1996). Error correcting codes in quantum theory. *Physical Review Letters, 77,* 793-797.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications, 39,* 2597-2605.

Trailer, B. (2006). Understanding what your sales manager is up against. *Harvard Business Review, 84,* 48-55.

Van den Poel, D. & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research, 166,* 557-575.

Weinberger, K. Q. & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research, 10,* 207-244.

Wilson, R. D. (2003). Using online databases for developing prioritized sales leads. *Journal of Business & Industrial Marketing, 18,* 388-402.

——— (2006). Developing new business strategies in B2B Markets by combining CRM concepts and online databases. *Competitiveness Review: An International Business Journal incorporating Journal of Global Competitiveness, 16,* 38-43.

Yu, Y. P. & Cai, S. Q. (2007). A new approach to customer targeting under conditions of information shortage. *Marketing Intelligence & Planning, 25,* 343-359.

Zytynska, S. E., Fay, M. F., Penney, D., & Preziosi, R. F. (2011). Genetic variation in a tropical tree species influences the associated epiphytic plant and invertebrate communities in a complex forest ecosystem. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366,* 1329-1336.

APPENDIX

Variable Name	Type
Sales_volume	Numeric
Employees_total	Numeric
Employees_here	Numeric
Status_indicator_0	Categorical

Naics_1 to Naics_5	Categorical
Veteran_indicator	Categorical
Women_owned_indicator	Categorical
Minority_owned_indicator	Categorical
Minority_type	Categorical
Cottage_indicator	Categorical
Import_export_indicator	Categorical
Manufacturing_indicator	Categorical
Public_private_indicator	Categorical
Legal_status_code	Categorical
Owns_rents_indicator	Categorical
Small_business_indicator	Categorical
Population_code	Categorical
Fortune_1000_indicator	Categorical
Non_profit_indicator	Categorical
8a_disadvantage_indicator	Categorical
Square_footage_estimator	Numeric
Franchise_indicator	Categorical
Territory_covered	Categorical
Hierarchy_code	Categorical
Sales_cat	Categorical
Emp_here_cat	Categorical
Emp_total_cat	Categorical
Square_footage_cat	Categorical

Table A.1: Variable list

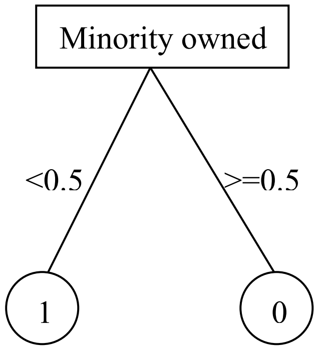


Figure A.1: Decision tree round 3

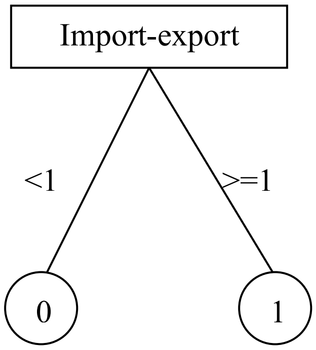


Figure A.2: Decision tree round 4