MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

MINING MASSIVE FINE-GRAINED BEHAVIOR DATA TO IMPROVE PREDICTIVE ANALYTICS[1]

David Martens
Applied Data Mining Research Group, Department of Engineering Management, University of Antwerp,
S.B. 516, 2000 Antwerp BELGIUM {David.Martens@uantwerpen.be}

Foster Provost
Department of Information, Operations, and Management Sciences, Stern School of Business, New York University,
44 West Fourth Street, New York, NY 10012 U.S.A. {fprovost@stern.nyu.edu}

Jessica Clark
Department of Information, Operations, and Management Sciences, Stern School of Business, New York University,
44 West Fourth Street, New York, NY 10012 U.S.A. {jclark@stern.nyu.edu}

Enric Junqué de Fortuny
Department of Marketing Management, Rotterdam School of Business, Erasmus University Rotterdam,
Burgemeester Oudlaan 50, 3062 PA Rotterdam, THE NETHERLANDS {junquedefortuny@rsm.nl}

Organizations increasingly have access to massive, fine-grained data on consumer behavior. Despite the hype over "big data," and the success of predictive analytics, only a few organizations have incorporated such fine-grained data in a non-aggregated manner into their predictive analytics. This paper examines the use of massive, fine-grained data on consumer behavior—specifically payments to a very large set of particular merchants—to improve predictive models for targeted marketing. The paper details how using this different sort of data can substantially improve predictive performance, even in an application for which predictive analytics has been applied for years. One of the most striking results has important implications for managers considering the value of big data. Using a real-life data set of 21 million transactions by 1.2 million customers, as well as 289 other variables describing these customers, the results show that there is no appreciable improvement from moving to big data when using traditional structured data. However, in contrast, when using fine-grained behavior data, there continues to be substantial value to increasing the data size across the entire range of the analyses. This suggests that larger firms may have substantially more valuable data assets than smaller firms, when using their transaction data for targeted marketing.

Keywords: Behavioral similarity, big data, response modeling, banking, payment data, customer analytics

Introduction

This paper studies the use of massive fine-grained data when applying predictive analytics. Specifically, the paper focuses on settings where the consumers' fine-grained financial transactions can be observed, with our particular application being the identification of prospective customers for marketing offers in banking. Financial firms increasingly are using predictive modeling to target offers to cross- or up-sell to existing customers and for customer retention (Hormozi and Giles 2004; Hu 2005; Van Den Poel and Lariviere 2003). This is not a paper describing a new application area. Predic-

[1]Bart Baesens, Ravi Bapna, James R. Marsden, Jan Vanthienen, and J. Leon Zhao served as the senior editors for this paper.

tive analytics has been used extensively for targeted marketing, and large banks use such methods routinely. Indeed, the bank that is the subject of this study has a sophisticated targeted marketing operation, using data on the customers' demographics, geographic location, and prior activity with the firm (tenure, lifetime value so far, services used, etc.). As a shorthand, we will call this type of data *structured data*.

Instead this paper examines expanding the data used in the modeling to "big data" and, specifically, to massive fine-grained data on consumer transaction behavior, in a non-aggregated manner. Does it add value? If so, can it be done simply and in a scalable manner?

In particular, this paper focuses on taking advantage of fine-grained data on customer payments to merchants, which banks collect routinely. Such money-transfer data currently are not being used (broadly) for targeted marketing, either because the data are too big or unwieldy for traditional methods to handle or because the modelers are not convinced of the value of changing their methods. However, intuitively, observing that a consumer makes payments to a certain squash center in Brussels, a student restaurant in Leuven, and a high-end fashion store online provides substantial information about the consumer's interests. As we will demonstrate empirically, such fine-grained data are remarkably predictive of which consumers will be good prospects for particular offers. The data set for this study, described in detail in the "Results" section, contains over 21 million payments made by 1.2 million customers to 3.2 million merchants.

The main contributions of this paper are:

- The demonstration that incorporating measures of behavioral similarity based on massive fine-grained transaction data indeed can improve predictive analytics in a real application.[2]
- An overview and comparison of different modeling techniques to mine such data in terms of predictive performance and scalability.
- The careful examination of when exactly fine-grained behavioral similarity adds value, whether it is complementary to existing methods, and whether it is subject to improvement with increasingly bigger training data.
- The associated demonstration that firms with larger data assets can have a significant advantage when applying predictive analytics.

[2]The behavioral similarity method introduced and assessed in this paper has been used in practice by a leading bank to improve its targeting of customers. The actual production results are proprietary.

In what follows, we start by reviewing prior work on data-driven targeted marketing. We then introduce a behavioral similarity measure that allows the building of predictive models incorporating massive, fine-grained behavior data. This technique is evaluated empirically on the banking data in the "Results" section, including a comparison with traditional targeting based on structured data. The final section concludes the paper and raises issues for future research.

Prior Work

Sophisticated marketing modelers use predictive analytics to build models to estimate which potential prospects will be good prospects for product offers, as well as which customers will be likely to abandon the company (attrition or churn prediction). In this paper we focus on the sorts of data used, so it is worthwhile spending some time reviewing current practices and related research. The most sophisticated data-driven marketers use a wide variety of data to create features that summarize consumers' demographics, geographic location, and related characteristics (see Hill et al. 2006; Hormozi and Giles 2004; Hu 2005; Van Den Poel and Lariviere 2003). In cases where the consumers are (or have been) customers of the firm, to these features are added summaries of the individuals' prior activity with the firm (tenure, lifetime value so far, services used, etc.). Furthermore, when available, features also can summarize product position and product use. The predictive analytics community has gotten used to targeted marketing and attrition/churn applications as bread-and-butter examples of predictive analytics in action, and even have benchmark data sets representing this sort of data (e.g., the data sets from KDDCUP 1997, 1998, 2009[3]).

In some cases, traditional targeted marketing does incorporate data from transaction behavior. However, transaction data traditionally are aggregated into a relatively small set of variables summarizing properties such as the transactions' recency (e.g., when was the most recent transaction), frequency (e.g., what is the frequency of the transactions), and monetary value (e.g., what is the monetary value of the transactions) (Fader et al. 2005). Many variants of such RFM variables can be engineered, including the average amount of payment, the median, or some other percentile. Such variables are included in our structured data. A main point of this paper is to demonstrate that it also is important to view transaction data as very detailed and fine-grained information on a consumer's behavior, and that such a view can lead to improving predictive performance.

One case where fine-grained transaction data have been taken into account in traditional targeted marketing is in social

[3]http://www.kdd.org/kdd-cup.

network-based marketing (Hill et al. 2006; Verbeke et al. 2014). Social network targeting has been justified based on theories of homophily (McPherson et al. 2001) and social influence (Aral et al. 2009). Unfortunately, neither of these theories is sufficient for justifying the use of general behavioral similarity for targeting. Social influence requires an actual social connection between the individuals in order that there be correlations in their preferences, not simply similarity in their behavior. The concept of homophily also is used as theoretical justification for data-driven targeting in a slightly more circuitous manner. Social theory has long held that social connections are more likely to be made between people who are similar, along a wide range of dimensions—and the stronger the similarity the more likely a connection will be made (McPherson et al. 2001). Thus, people who are connected are likely to be similar, which justifies targeting the social network neighbors of people who are observed to exhibit the desired characteristics (such as having purchased the product; Hill et al. 2006). Such targeting can be done by observing fine-grained behaviors indicating a social connection (Hill et al. 2006), but unfortunately as a theoretical justification, as with social influence it requires that there be a social connection. Aral et al. (2009) provide intriguing evidence that a surprising proportion of the observed correlation in product adoption of social-network neighbors is due simply to their similarity rather than to social influence. From a predictive analytics point of view, the fact of being social-network neighbors could be considered a particularly useful similarity measure (as suggested by Hill et al. 2006). However, even the sophisticated targeted marketing study of Hill et al. (2006) did not consider massive fine-grained data on payments or merchant visitation.

One area of study that has examined prediction using fine-grained behavior data, albeit for an application with important differences, is recommender systems/collaborative filtering, where one tries to predict the utility of items for a particular user based on the items previously purchased or rated by other users (Adomavicius and Tuzhilin 2005). A typical example is predicting which movies a consumer is likely interested in, based on the ratings provided by the user and the set of all other users' ratings. The main differences with our setting are (1) that recommender systems make predictions within the same domain of behavior as the data used to make the predictions, and (2) recommender systems try to predict for a very large number of products simultaneously. Within our domain, where the features are merchants transacted with, a recommender system approach would tell which other merchants a consumer is likely to transact with. A targeted marketing model will predict the value of one specific target variable for the consumer, outside the domain of merchants. (In this case, for example, would the consumer buy a pension fund product if it were to be offered?)

Modeling Techniques for Fine-Grained Payment Data

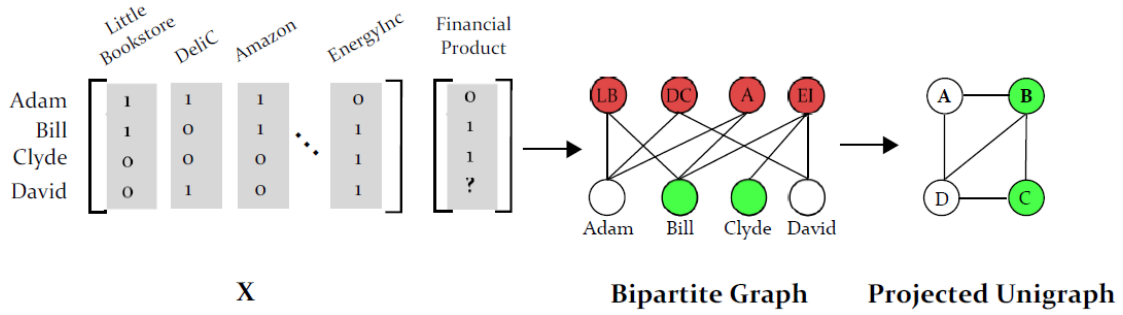
Payment Data Representation

Consider a (possibly anonymized) transaction log containing payments[4] from a set of consumers to various entities, such as firms, institutions and other persons, which we will loosely call merchants. We distinguish between three data representations, as shown in Figure 1, each opening up a separate set of applicable prediction techniques. First, we can represent the data as a very large matrix $\mathbf{X}$ where element x_{ij} indicates whether consumer i made a payment to merchant j ($x_{ij} = 1$) or not ($x_{ij} = 0$). A separate binary vector represents the *target variable* (i.e., the quantity to be estimated and for which training labels will be provided; Provost and Fawcett 2013). Note that for this paper we limit ourselves to a binary matrix $\mathbf{X}$ and binary target variables. Given the encouraging results we present below, the extension to a weighted matrix that represents the frequency, recency, or monetary value of the payments and other types of target variables are interesting future directions.

When dealing with behavior data, where people interact with massive sets of entities like merchants (by choice), the data usually are extremely sparse. The intuitive explanation for this is that people can only make a finite number of such choices in a limited amount of time (Junqué de Fortuny et al. 2013). This is amplified in the present application by the fact that these choices involve monetary transactions. In our setting, this translates to the fact that no one consumer will transfer money with a large fraction of possible payment receivers. Therefore, the data set representing behaviors is very large dimensionally yet extremely sparse. Techniques that take advantage of the sparsity can be useful even in settings where analysts are not accustomed to building models from massive data matrices or using specialized big data computing architectures.

Specifically, we can exploit the extreme sparsity of the matrix to design a direct similarity comparison that is particularly scalable. Consider that such data also can be represented as a bipartite graph (bigraph). In bigraphs, sometimes also referred to as affiliation or two-mode networks (Borgatti and Everett 1997; Breiger 1974; Latapy et al. 2008), there are two types of nodes with edges only between nodes of different types. In our example, we have consumers as one type of node, merchants as another, and edges defined by the payment transaction data. The matrix representation discussed previously corresponds to the adjacency matrix of the bigraph.

[4]The payment transactions can be viewed broadly, including debit and credit transactions, check payments, etc.



Note: Payments can be represented as a matrix **X**, as an attributed bigraph from consumers to merchants, or as a projected, attributed pseudo-social network.

Figure 1. Data Representation of a Transaction Log of Payments

A Network-Based Behavioral Similarity Score

Consider a transaction log containing money transfers, as illustrated for the very simple example in Figure 1, and repeated in Table 1. Let us introduce a behavioral similarity score based on a two-step approach: (1) define a weighted pseudo-social network that represents the similarity in payment behavior between consumers, and (2) based on the resulting network, compute predictive *behavioral similarity* (BeSim) scores for each consumer for a selected target variable.

Defining a Weighted Pseudo-Social Network

The similarity network will be constructed such that two consumers are considered similar if they make payments to the same entities, and are more similar the more such connections they share. This leads to the *pseudo-social network* (PSN), where a data network among consumers is built by linking two consumers if they send a payment to the same merchant. We call the inferred network model among consumers a *pseudo-social network* because, as in a true social network, strongly connected consumers demonstrate a strong similarity, at the very least in the particular merchants with which they transact. The key underlying assumption is that, as with to a true social network, if two consumers are strongly linked, they will be similar in other ways as well—such as affinity for a marketing offer. It is a *pseudo*-social network because, by and large, the linked consumers probably have no true social relationship with one another.

In the PSN, each link provides some evidence of similarity. For example, in Figure 1, both Adam and Bill made a payment to Little Bookstore, and hence are linked in the PSN. Now the question becomes how to assign a similarity weight to the links, incorporating two aspects: (1) the more merchants the linked consumers share, the higher the weight should be, and (2) the more popular a merchant, the lower the weight should be.

The latter aspect is motivated by the fact that there will be companies that many consumers pay, such as telecommunication operators or energy providers. These may provide little information on the similarity between two consumers and could swamp more informative links based on, for example, the fact that two consumers shop at the same small store. Very popular merchants can be omitted or down-weighted. As a simple example of such a "micro-affinity" measure, one could weight a merchant by the inverse of the number of customers: $1/NC_j$. We will introduce more sophisticated metrics later.

The resulting weight between two consumers X and Y is then defined as the sum of the micro-affinity values (in this simple case, the $1/NC_j$ values) of the shared merchants j. This satisfies the desire that having more shared merchants leads to stronger links, and also takes into account the micro-affinity.

$$w(X,Y) = \sum_{j \in [merchants(X) \cap merchants(Y)]} \frac{1}{NC_j} \qquad (1)$$

Predictive BeSim Scores Based on the Labeled PSN

Given the weighted PSN with a label for each consumer (having purchased the product in the past or not), we can now

Table 1. Example Payment Data to Merchant-Specific Metrics*

merchant *j*	Consumers	NC_j	NS_j	$1/NC_j$
LittleBookStore	A, **B**	2	1	.5
DeliC	A, D	2	0	.5
Amazon	A, **B**	2	1	.5
EnergyInc	**B**, **C**, D	3	2	.33

*Number of customers NC_j, number of known buyers NS_j, and inverse frequency $1/NC_j$. The known buyers (label 1) among the consumers are denoted in boldface.

apply a modification to the standard weighted-vote relational neighbor (wvRN) procedure for predictive inference with network data (Macskassy and Provost 2003, 2007). Let's call the consumers that are positively labeled (the known buyers) the "positive seed customers." The BeSim score for each consumer is simply the sum of the weights to the seed customers. For example, the score of David is the sum of the weight of the links in the PSN to Bill and Clyde (these are the only neighbors who are seed customers; see also Figure 2). The differences from the prior work using wvRN are important for processing the massive data:

1. Here only the positive class is considered (thereby avoiding most of the links in the massive network).

2. The weights of the links between consumers are computed as the sum of the micro-affinity values ($1/NC_j$ in the example) for all shared merchants.

3. The final scores are not normalized across all neighbors, but instead are simply the sum across the relatively small set of positive neighbors.

To understand the BeSim score more deeply, consider the following: using the example definition of the link weights as sums of micro-affinity scores, the terms in the resultant BeSim score can be regrouped algebraically by merchant. For each merchant j, the corresponding BeSim term is the empirical probability E_j—the ratio of the number of seed customers (known buyers) that made a payment to the merchant (NS_j) divided by the total number of (unique) consumers making a payment to the merchant (NC_j). When we have n consumers and m unique merchants, the (heuristic) score of a consumer X_i is defined as the sum of the empirical probabilities E_j of the merchants to which X_i has made payments ($j|x_{ij}=1$):

$$S_{BeSim}(X_i) = \sum_{j|x_{i,j}=1} E_j \tag{2}$$

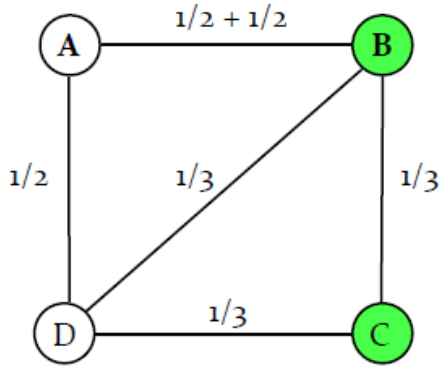
$$E_j = \frac{NS_j}{NC_j} \tag{3}$$

Note that a normalization factor per data instance can be added equal to the sum of the weights of the links in the pseudo-social network: $Z_X = 1/\left(\sum_Y w(X,Y)\right)$.

The BeSim calculation thus provides a measure of behavioral similarity between a consumer X and the set of known buyers (seed customers). The time complexity of computing the BeSim measure is $O(n \cdot \bar{m})$, where n is the number of training points, and $\bar{m}$ is the average number of merchants consumers pay (i.e., the average number of nonzero elements per row). This corresponds to one pass over the complete log of payment data, where we count the number of times a certain merchant is paid, by the seed customers versus all consumers. Once the empirical probabilities are all calculated, the actual scoring requires summing just those that correspond to nonzero elements, as given by Eq. (2).

The calculation is illustrated further with the simplified example shown in Table 1. The score for each consumer is obtained by summing the $1/NC_j$ scores across each consumer's set of merchants. In the example, consumer A (Adam) made payments to three merchants: *LittleBookStore*, *DeliC,* and *Amazon*. Hence, the score for A is given by the sum of the scores for these three merchants as calculated in Eq. (4), where each of these scores is determined by the empirical probability NS_j/NC_j. For consumer D (David), the scores for *DeliC* and *EnergyInc* are summed, providing a score of 0.67. We can easily verify these results by applying the modified wvRN on the PSN given in Figure 2. The formulation given by Eq. (2) has the advantage that it scales very well to a huge number of merchants and consumers (as compared to running the actual wvRN on the PSN).

$$\begin{aligned} S_{BeSim}(A) &= S_{LittleBookStore} + S_{DeliC} + S_{Amazon} \\ &= \frac{1}{2} + \frac{0}{2} + \frac{1}{2} \\ &= 1 \\ S_{BeSim}(D) &= 0.67 \end{aligned} \tag{4}$$



Note: Positively labeled customers are indicated as green (shaded) nodes, and the weights are taken as the sum of the $1/NC_j$ values of the shared merchants.

Figure 2. The Weighted PSN of the Running Example (Table 1)

Relation to Other Neighbor-Based Methods

While we already discussed prior work, we can now provide a deeper comparison based on the technical details presented in the previous section. As just noted, drawing inferences with the BeSim calculation produces a neighbor-based inference method most closely related to wvRN, as applied to the induced pseudo-social network. It is useful to clarify the difference from the most popular neighbor-based inference method, *k*-nearest-neighbor inference (*k*NN). Although there are many variants of *k*NN, the principle is always the same: a similarity metric is chosen between instances, and the inference for a new instance is calculated based on the target values of the *k* training instances (those with the value of the target variable known) most similar to this new instance. These target values are combined via some combining function, possibly taking the instances' similarities into account. A traditional *k*NN classifier, however, using (for example) Euclidean or cosine distance, is not nearly as efficient as the BeSim calculation as the amount of data grows. The time penalty for massive data sets stems from the fact that *k*NN must calculate the distance between all inference data and the training data. This results in time complexity $O(n \cdot n_{te} \cdot \bar{m})$ (with n the number of training points, n_{te} the number of inference/test points, and $\bar{m}$ the average number of features a vector has). In our specific scenario, *k*NN would take about 100,000 times longer to compute than BeSim.

Within the general neighbor-based prediction framework, another main difference with the present method (as with wvRN) is that the number of neighbors can be different for each instance, and this number is determined implicitly by the data—specifically, by the number of seed customers who share a payment receiver with the focal consumer. Further, BeSim only considers positive examples as candidate neighbors. The similarity function and combining function then comprise the novel BeSim calculation (including the link importance weighting). One contribution of this paper, thus, is the introduction of this scalable similarity computation that can be used for fast inference from other massive fine-grained behavioral data coming from transactions, web monitoring, social networking sites, etc. In theory, the BeSim metric could also be introduced to *k*NN classifiers, especially when behavior data are used.

Alternative Calculations of BeSim

So far, to introduce the use of fine-grained behavioral similarity, we presented one particular behavioral similarity calculation. Specifically, we used the sum of supervised components (*NS*/*NC*) over all merchants a consumer paid to (Eq. (6), where m is the total number of unique merchants). In the empirical evaluation below we consider several variants.

First, we consider adding an additional micro-affinity penalty. We take inspiration from the well-known IDF relevance measure used in information retrieval, where terms occurring in many documents receive low weights and terms occurring in fewer documents receive higher weights. The inverse consumer frequency (ICF), defined by Eq. (5), provides an indication of the inverse popularity of the merchant, as a function of the number of customers that made a payment to

the merchant (NC_j), and the total number of consumers in the dataset (n). This additional penalty leads to the heuristic score defined by Eq. (7).

We can also remove the micro-affinity weighting altogether, leading to Eq. (9), which scores a merchant based on the absolute number of seed customers only. Taking this even further, Eq. (10) shows the calculation where each merchant receives a binary value indicating whether any seed customer (known buyer) has made a payment to it or not.

We can also consider replacing this ICF weighting metric with a more sophisticated alternative. (The reason for further attention on this weighting factor is that it works well, as will be shown later.) Employing ICF is based on the presumption that merchants to which fewer consumers make payments should contribute more influence in the similarity calculation than merchants who receive payments from many consumers. Using ICF, we discount non-informative merchants such as the tax authority. However, ICF might (also) amplify non-informative noise: merchants to which only a very few customers make payments will receive a very high weight, even though this may be simply due to random chance (e.g., due to how the data were sampled).

To assess the importance of this concern, we replace the ICF measure with a frequency-weighting based on the Beta distribution, a continuous probability distribution defined by two parameters, α and β, which define its shape. By tuning the parameters, we can determine the empirically optimal shape of the weight distribution as a function of the normalized degree (number of unique consumers that made a payment) of a merchant. For the analyses that follow, these parameters are tuned on a validation set (one third of the training data) with a grid search procedure, separately using the different target evaluation measures, AUC ($S_{B,AUC}$) and lift ($S_{B,Lift}$), discussed below. This leads to empirically optimal values α^* and β^*. Note that more advanced approaches can be considered to learn these hyper-parameters, by optimizing a loss or likelihood function, using a Bayesian, gradient descent or random search approach (Bergstra and Bengio, 2012).

The flexibility of the Beta distribution (see Figure 3) gives a learning-based method to penalize low-frequency merchants, based on the degree to which they provide informative signal or mainly noise. As will be described later, the Beta distribution with parameters determined to be optimal empirically on these data does indeed resemble closely the shape of ICF and thereby confirms that the low-frequency merchants contribute more valuable information than deleterious noise (see Figure 4). However, Figure 4 also shows that while the shape of the best Beta distribution conforms to that of the shape of the ICF measure, the resultant weights are significantly different. Thus it makes sense to compare the two empirically to judge their relative generalization performance.

$$ICF_j = \log_{10}\left(\frac{n}{NC_j}\right) \tag{5}$$

$$S_{NSNC}(\mathbf{x}) = \sum_{j|x_j=1} \frac{NS_j}{NC_j} \tag{6}$$

$$S_{ICF}(\mathbf{x}) = \sum_{j|x_j=1} \frac{NS_j}{NC_j} \cdot ICF(j) \tag{7}$$

$$S_B(\mathbf{x}) = \sum_{j|x_j=1} \frac{NS_j}{NC_j} \cdot B(\alpha^*, \beta^*)(j) \tag{8}$$

$$S_{NS}(\mathbf{x}) = \sum_{j|x_j=1} NS_j \tag{9}$$

$$S_1(\mathbf{x}) = \sum_{j|x_j=1} I(S_{NS}(\mathbf{x}) \neq 0) \tag{10}$$

Results

We now present results comparing different methods for purchase prediction, based on fine-grained payment data from a major international bank.[5] The goal of this study is specifically to assess whether predictive modeling based on fine-grained payment data holds value, and to compare different methods with an eye toward both predictive and time performance when processing massive data. Later in this section, we will assess whether modeling fine-grained data actually adds value over traditional approaches (using traditional structured data) for predictive modeling for targeted marketing.

The Data

The data for the analyses are from a period of 11 months, comprising over 21 million (debit) transactions made by 1.2 million of the bank's customers (anonymized) to a total of 3.2 million unique, anonymized merchants. We built predictive

[5]These are data from a European office, which is noteworthy because European consumers and American consumers have different general credit- and debit-account habits, with European consumers employing non-card debit transfers substantially more frequently.

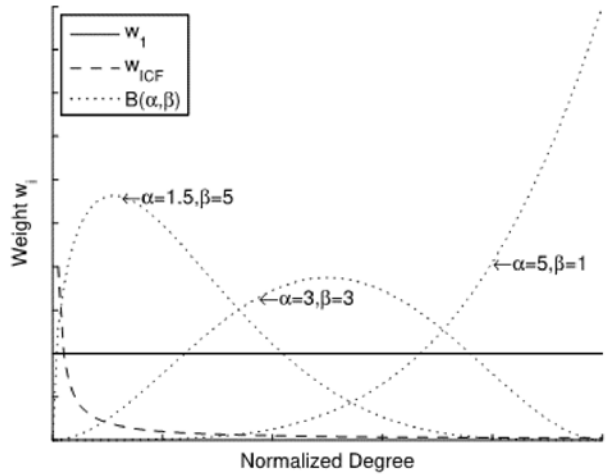


Figure 3. Weighting Schemes for a Merchant Based on the Normalized Degree

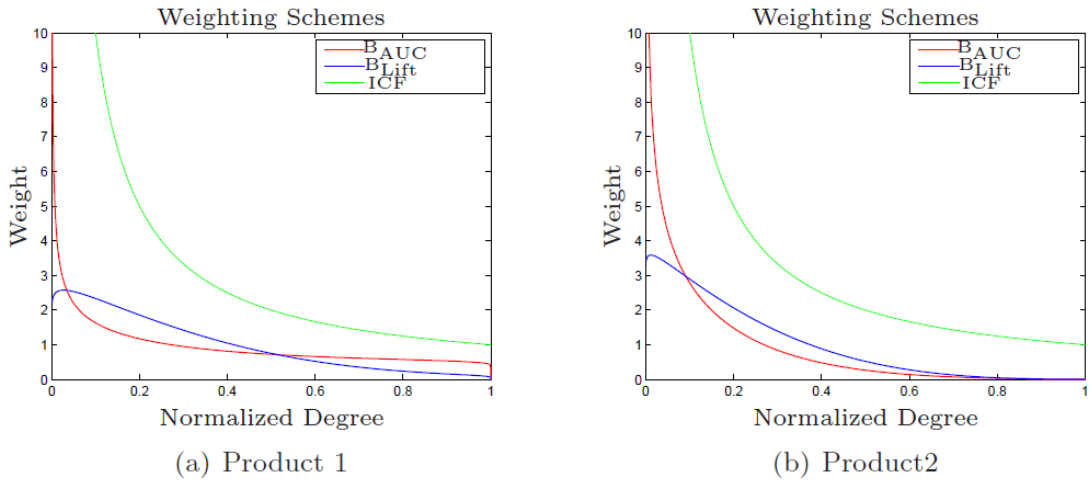
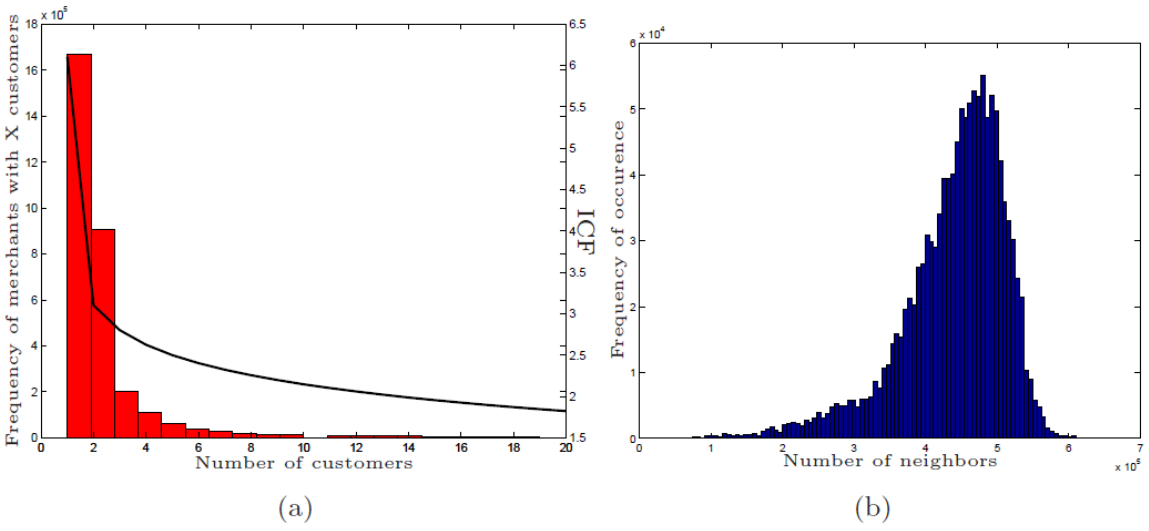


Figure 4. ICF Weighting Versus Beta Distribution Optimal Weighting (optimized for AUC and Lift at 1%)

models for two tasks, targeting the purchasing of two different financial products: a pension fund product and a long-term deposit product, with, respectively, 20 percent and 3 percent of the customers having bought the product (the "class priors"). The binary target variable for each product represents whether or not the consumer purchased the product. No targeted campaign had taken place beforehand.

Figure 5 shows some characteristics of the data. Figure 5(a) shows a histogram of the number of customers per merchant (the merchant's degree in the bigraph; bars, left vertical axis). We see that most merchants receive few payments (the average number of customers per merchant is 6.7).[6] However, there exist a few merchants to which almost all consumers make payments. These are likely monopoly-like companies such as energy suppliers, large telecommunication operators, or the tax authority. Since for this study the data were anonymized, we are not able to know exactly what they were. The ICF weight for each merchant is given by the black line (right vertical axis), showing how merchants with

[6]Although the number of customers for a payment receiver goes up to several hundred thousand, the vast majority of the distribution is given in the range shown, up to 20.



Notes: Figure 5(a) shows a histogram of the number of customers per merchant (the bars); the corresponding ICF weight is shown by the black line. In Figure 5(b), the distribution of the number of neighbors for consumers in the projected graph (the pseudo-social network) is shown.

Figure 5. Banking Data Characteristics

many customers are down-weighted more severely than merchants with few customers.

Figure 5(b) plots the distribution of the number of neighbors in the projected graph. Most consumers have more than 100,000 neighbors, showing that the pseudo-social network indeed has a different structure than a typical, true social network (people rarely have 100,000 friends, for example). The merchants with very many customers implicitly link very many consumers, leading to this high connectivity. (Note: Since the BeSim calculation is based only on the positive neighbors, the number it must process is much smaller.)

In addition to the payment transaction data, 289 traditional variables were obtained from the bank for the consumers in the data set. These are the variables used by the bank for their own targeting. For confidentiality reasons, a complete enumeration of all variables is not possible, but these variables can be categorized as follows:

- Demographics, such as age and gender
- Location, such as postal code, province, and main bank office
- Prior products, such as financial funds, savings accounts, or other products
- Product usage, such as amount of use for a product
- Tenure, such as time with the bank
- RFM, recency, frequency, and monetary value of payments

Such data accord with those traditionally used by large banks and other large companies for their customer analytics applications (see Hill et al. 2006; Hormozi and Giles 2004; Hu 2005; Van Den Poel and Lariviere 2003). We will investigate the complementary value of these data later in this section.

Predictive (Generalization) Performance: Purchase Prediction Using Payment Data

We now present the generalization performance of the different BeSim variants for predicting consumer purchase likelihood, as measured by the area under the ROC curve (AUC) (Fawcett 2006) and the lift over random selection (Linoff and Berry 2011). The AUC is equivalent to the Mann–Whitney–Wilcoxon statistic, and measures how well a predictive model ranks a binary outcome. The lift over random selection is defined for a particular targeting threshold—usually the approximate percentage of consumers who will be targeted, those receiving the highest predicted likelihoods. Given a targeting threshold (e.g., the top 1 percent), the lift is the ratio of the target purchase rate to the average purchase rate (corresponding to random selection). For example, consider the situation where in the population as a whole, we expect 5 percent of the consumers to be purchasers; if for the 1 percent of consumers with the highest model scores the observed purchase rate is 15 percent, then we have obtained a lift of three. AUC and lift are both exam-

ined because together they give a more nuanced assessment of the predictive performance of the model: the AUC measures performance over the entire range of predictions and the lift measures performance for those with the highest predicted purchase likelihoods. In practice, we only really care about the latter, as typically targeting budgets allow one only to target the very top prospects on the ranked list; however, for completeness we would also like to assess the overall generalization performance of the model.

Besides the various BeSim measures, we also consider estimating the relative purchase likelihoods via support vector machines (SVMs), which are generally thought to work well on high-dimensional data. We choose a linear kernel, as linear models have been reported to perform as well or even better than nonlinear alternatives for sparse data, since soft linear separability is more likely to occur (Hastie et al. 2001). (Supplementary studies verify this for the data used in this paper.) Equally importantly, using a linear kernel, the resulting models are generally faster to train and evaluate than nonlinear models. We use L1 regularization (on the L2 loss) for several reasons. Ng (2004) observes that L1 regularization has been shown to be better than L2 regularization if the number of input instances is smaller than the number of features, as it is in these data. Also, L1 regularization promotes sparsity in the model (Hastie et al. 2001), which is in line with our goal of finding particular merchants that are informative. Furthermore, based on the guidance of Chapelle and Keerthi (2008), in order to allow for operationally reasonable convergence rates on the massive data during the processing-intensive cross-validation phase, we relaxed the ϵ-stopping criterion level of the quadratic program to be solved in the SVM procedure during cross-validations, since a stringent error tolerance level is not necessary in this phase to reach good final models, especially when dealing with high volumes of data. Based on preliminary analysis, this setup achieved similar (not significantly different) generalization performance at a more than ten-fold speed increase.

For the analyses in this section, the data are split into a training set (90 percent) and a (held out) test set (the remaining 10 percent). All consumers in the training set who bought the product are labeled as known buyers; scores are computed using the different methods for the consumers in the test set. As discussed above, we evaluate all of the models in terms of predictive performance using AUC and lift. This process was then repeated 10 times, each with a different random training/test split. The averages and standard deviations of the results over the 10 analyses for each of the techniques and each of the measures are shown in Table 2.

The best models in terms of AUC are the variants of BeSim where the Beta distribution is optimized (on a validation portion of the training data) to maximize AUC. Interestingly, as noted above, inspection of the optimal Beta distribution parameters reveals that it has a similar general shape to ICF (Figure 4). The SVM-based methods are not competitive for AUC using these massive, sparse data.

In terms of lift, the best AUC model (BeSim with the Beta distribution tuned for AUC) does not perform well. However, as we would hope, BeSim with the Beta distribution trained for each lift threshold performs quite well for the corresponding lift threshold—in four of six cases being the best-performing method (or tied for that distinction), and in all cases being a close competitor. The main disadvantage of the Beta distribution version of BeSim is that it takes a very long time to optimize since two parameters need to be cross-validated on a separate part of the huge data set (Table 3). Although it is never the best-performing method, the SVM using the fine-grained behavior data performs much better for lift than it does for AUC, especially for Product 2. However, it is even more computationally expensive than the Beta-based BeSim calculations (Table 3).

The biggest surprise is that the *ICF*-based BeSim scores perform comparably to or better than any of the other techniques, including the predictive modeling methods, for top-segment prediction (lift)—the main measure of interest. This is especially remarkable because it is also one of the fastest techniques to run on massive data; it could easily scale up to much larger data sets due to its space efficiency and sparse linear run-time complexity.

Seeing that the heuristic S_{ICF} approach performs best in terms of predictive performance and scalability, we examine it in more depth next, where we assess the complementarity of the fine-grained payment data and the structured data.

Behavioral Similarity Versus Traditional Structured Modeling

We will now replicate the procedure employed by this bank for modeling using traditional (structured) data, compare it to the BeSim scoring, and examine combining the two. Note that as part of its normal practice, the bank performs studies of different learning/prediction methods using standard predictive analytics evaluation procedures.

Analytical Setup

To replicate the bank's standard practice, we build a linear support vector machine (SVM) (Vapnik 1995) model using the 289 traditional variables on a balanced sample from the

Table 2. Results for the Empirical Study Comparing the Different Versions of BeSim and SVM-Based Modeling*

(a) Product 1				
Method	**AUC**	**lift1**	**lift5**	**lift10**
ICF	62.9 (2.3)	**11.6** (2.5)	**3.5** (0.4)	2.3 (0.2)
NSNC	62.9 (2.5)	11.5 (2.4)	**3.5** (0.4)	2.3 (0.2)
NS	57.8 (2.7)	1.25 (0.3)	1.4 (0.3)	1.4 (0.2)
S_1	58.1 (3.6)	3.0 (0.9)	1.9 (0.3)	1.6 (0.2)
$S_{B,AUC}$	**70.9** (2.5)	1.3 (0.6)	1.3 (0.2)	1.3 (0.2)
$S_{B,Lift}$	62.88 (2.6)	11.0 (2.2)	**3.5** (0.5)	**2.4** (0.3)
SVM	50.7 (3.2)	7.3 (2.8)	2.0 (0.5)	1.3 (0.3)
(b) Product 2				
Method	**AUC**	**lift1**	**lift5**	**lift10**
ICF	63.5 (4.1)	**19.9** (7.4)	**5.1** (1.6)	**3.2** (0.7)
NSNC	63.1 (4.1)	19.2 (7.3)	**5.1** (1.5)	3.1 (0.7)
NS	51.1 (3.3)	1.3 (0.3)	1.1 (0.2)	1.1 (0.2)
S_1	51.7 (2.7)	3.1 (0.9)	1.7 (0.4)	1.4 (0.2)
$S_{B,AUC}$	**78.9** (4.2)	1.0 (0.8)	1.0 (0.4)	0.9 (0.3)
$S_{B,Lift}$	63.3 (4.0)	18.2 (7.1)	**5.1** (1.5)	**3.2** (0.7)
SVM	67.1 (4.2)	18.8 (7.5)	4.9 (1.7)	**3.2** (0.9)

*Each value is an average over the 10 test folds, with standard deviation in parentheses.

Table 3. Averaged Time Duration to Learn the Different Models and Score the Test Data*

Method	Duration (sec.)
ICF	55
NSNC	40
NS	35
S_1	34
Beta	4097
SVM	6919

*As the preprocessing time is the same for all models, it is not included.

training set, including all the seed customers and just as many randomly selected non-buyer consumers.[7] A forward input selection procedure based on the area under the ROC curve (AUC) (Fawcett 2006) was conducted to select a maximum of 30 variables.[8] A validation set (chosen as a third of the training set) is held out to determine the optimal number of variables and to optimize the SVM's regularization parameter using a grid search, using the grid [0.001 0.01 0.1 1 10 100 1000]. The resulting model is called the *structured data* (SD) model.

For further comparison, we also created a new model that (as described next) is a combination of the BeSim and the SD models. This will help us to assess the extent to which the two models incorporate complementary information. Specifically, for this analysis we produce a linear combination of the two models' scores. The BeSim output score is rescaled to the interval [0,1] by subtracting the minimum and dividing by the range. All positive examples and just as many negative

[7]This sampling was conducted to enable the large number of analyses with the structured data. A smaller-scale comparison using the full data yielded similar results.

[8]A 30 variable plateau is visible in terms of AUC, as shown by Figure B1 in Appendix B.

examples are chosen to create a balanced sample (for scalability). Since the BeSim score is only available for the test data, we are limited to estimating the combined model on test data. Therefore, we hold out a randomly selected 10 percent of the previously defined test set to estimate the weights that combine the two output scores. The results are then computed over the remaining 90 percent of the test set to evaluate the performance of all models. The model resulting from combining the BeSim and linear SVM models is denoted **BeSim + SD.**

Empirical Comparison

For the two products, Tables 4 and 5 report the AUCs and lifts at 1, 5, and 10 percent averaged over 10 randomizations, each time using 80 percent of the data as training data and the remaining 20 percent as test data, following the procedures described above, and using the BeSim model as a benchmark. In the tables, for each of these performance metrics the cell with the best value is shown in boldface. The BeSim model alone seems to perform comparatively quite poorly when viewed in terms of AUC. However, it does extraordinarily well when comparing the lift at 1 percent. BeSim's lift degrades as the threshold becomes more liberal: it is comparable with the other models at 5 percent, and worse at 10 percent and higher. This pattern is observed for both products.

The attentive reader will have noticed a difference in performance between Table 2 and Table 4. This is explained by the fact that in the BeSim-only calculations (Table 2), the validation set was set apart from the training set for all metrics and not used for training. This was done to keep the comparison conservative since one method needed to use a validation set (viz., the Beta function). Using more data yields better results (Tables 4 and 5), to which we will return in the next section.

For all comparisons, the BeSim calculation plays a part in the best-performing model. Using a one-sided paired t-test over the 10 randomizations, we find that the combined BeSim + SD model performs significantly better than the individual BeSim and SD models for the AUC, lift at 5 percent and lift at 10 percent (all p-values < 1e-5). For the lift at 1 percent, the BeSim model alone performs best, significantly outperforming the SD model (p-value 7e-6), and the combined method (at a lower significance level, p-value 0.02). For product 2, the same techniques perform best, always significantly outperforming the two others (all p-values < 1e-5).

Thus, BeSim does a very good job when predicting which consumers have the highest likelihood of purchasing: the consumers ranked most highly by BeSim are comparatively very likely to buy the product themselves. As mentioned above, given budget limitations, marketing campaigns are often limited to targeting only the high-percentile prospects.

Examining the ROC and lift curves shown in Figure 6 illustrates why we see the comparison numbers that we do. The curves show the model generalization performance across all thresholds (for a chosen representative randomization). Indeed, BeSim performs very well at the top of the rankings, but once the high percentiles have been passed, the BeSim model (solid line) performs quite badly. Notice that the ROC curve becomes almost a straight line, indicating that, in this region, BeSim does not discriminate among these consumers at all. Looking even more deeply, the reason for this performance is that BeSim only provides a nontrivial score to a small number of consumers (which seemingly are indeed very likely candidates for the product). In the calculation of BeSim, only the neighbors of existing (seed) customers in the pseudo-social network are provided with a nonzero score; most of the PSN remains unscored (see Appendix A for details). More advanced network inference schemes, including collective inference (Macskassy and Provost 2007), thus may further improve the performance of inference as compared to scores based only on immediate neighbors in the pseudo-social network.

We can contrast BeSim's performance with the performance curve for the SD model (lighter, dotted line). The latter exhibits the ROC curve shape one normally sees for typical predictive models. Notably, it performs worse than the BeSim model at the very high score range and better everywhere else.

The combined model (BeSim + SD) performs strikingly well over the entire score range. Thus scoring using similarity based on fine-grained behavior data indeed has complementary predictive power to the traditional structured, data-based scoring. This result is particularly encouraging given the simplicity of the method we used to combine the two different scores. Designing a more sophisticated combining strategy may give additional lift.

Big Data and Generalization Performance

It is important to consider that these results are generated for a particular data set of a particular size. There is extreme variance in the number of customers patronizing different banks. There is also a large variance in the number of customers for different banking products. Thus, it is interesting to ask whether "data assets" of different sizes confer different improvements in decision-making ability. Further, standard

Table 4. Results for Product 1 with 80% Training Data, Showing the Averages and Standard Deviation (in parentheses)

	AUC	Lift 1%	Lift 5%	Lift 10%
BeSim	63.9 (0.6)	**14.9** (1.0)	4.1 (0.2)	2.6 (0.1)
SD	75.5 (0.7)	4.9 (0.3)	3.9 (0.3)	3.3 (0.2)
BeSim + SD	**78.2** (0.9)	12.7 (3.0)	**5.5** (0.5)	**4.0** (0.2)

Table 5. Results for Product 2 with 80% Training Data, Showing the Averages and Standard Deviation (in parentheses)

	AUC	Lift 1%	Lift 5%	Lift 10%
BeSim	71.7 (0.7)	**31.8** (0.6)	7.6 (0.1)	4.4 (0.1)
SD	86.6 (0.5)	10.1 (0.5)	7.8 (0.2)	6.0 (0.1)
BeSim + SD	**89.0** (0.7)	18.2 (4.1)	**9.7** (0.7)	**6.7** (0.2)

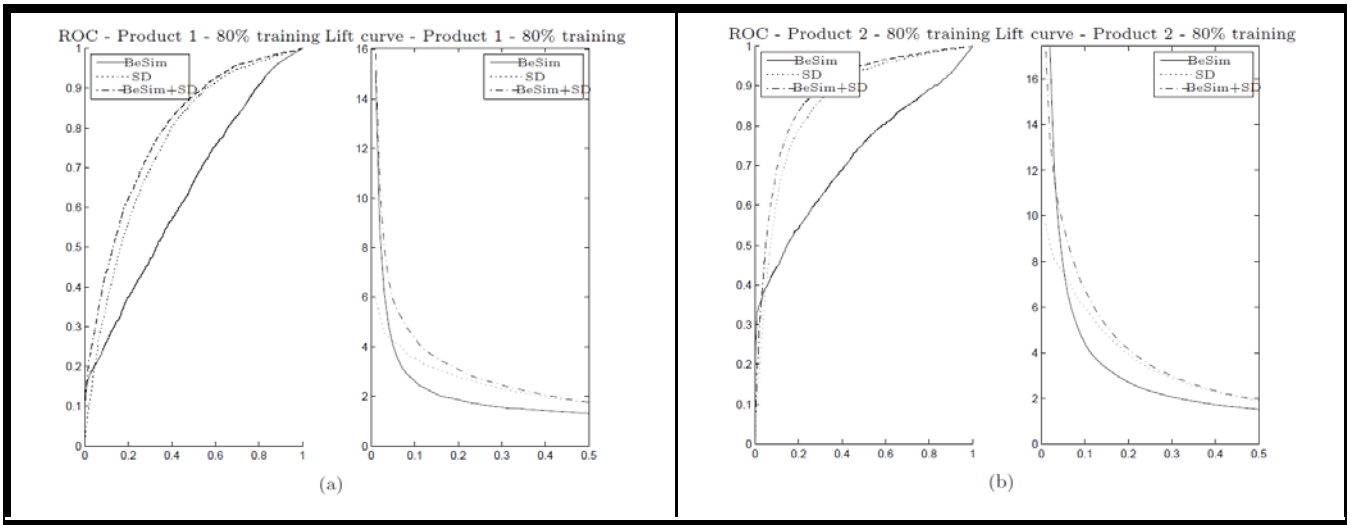


Figure 6. ROC and Lift Curves for the Behavioral Similarity (BeSim) Model, the Model Using Structured Data (SD), and the Combined Model (BeSim + SD)

practice in this application (as we are told) generally is to build targeting models from subsamples of the entire customer base, as the analysts believe that learning from a moderate-sized sample will confer all of the predictive power they are going to get. For both of these reasons, it is important to examine the relationship between the amount of training data and the generalization performance of the different models.

Because of its design, the BeSim scoring should be affected strongly by the amount of data available: larger training data size means more connections among consumers as well as more seed customers becoming available for inference; both should tend to lead to lower estimation variance (and thus lower error) in the scores. Most importantly, with more data, prospects for whom the BeSim score would previously have been zero due to no connection to any prior (seed) buyer would increasingly receive nonzero scores. Thus, it may be that these results are conservative compared with what might be expected across a large bank's entire customer base (which could be one or two orders of magnitude larger), and especially for products with large sets of prior purchasers (seed customers).

We can assess the effect of the data size within the range of our sample by plotting learning curves (Provost and Fawcett 2013), simulating the availability of different amounts of data. In Figure 7 we show the evolution of the performance metrics for the models (BeSim, SD, and BeSim + SD) as we increase

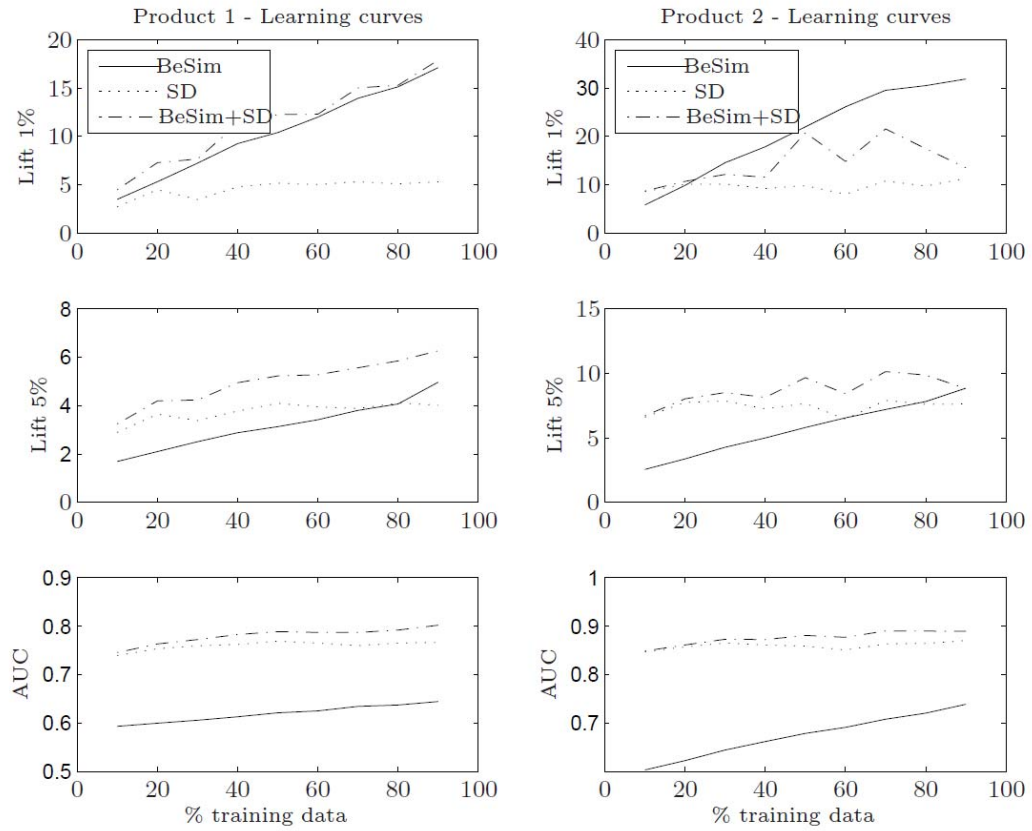


Figure 7. Learning Curves: Change in Lift and AUC on the Test Set for Product 1 (left) and Product 2 (right) as the Amount of Training Data is Increased

the amount of data available for training.[9] The AUCs and lifts for the BeSim model (solid line) indeed increase as we increase the amount of training data, and do so relatively constantly across the entire range. As noted above, as more seed customers become available, more consumers in the network will receive a nontrivial score. This improvement trend is not observed for the SD (traditional) model for any of the performance metrics, for either product. As is typically observed in predictive modeling applications (see Perlich et al. 2003), after a certain point the marginal performance improvements obtained by adding more training data become small.

This indicates that we should expect further model improvements for the BeSim model (and the combined BeSim + SD model) with larger data sets (with more seed customers), even though here we already have data on a fairly large number of customers. If this trend were to continue for orders of magnitude more data, it would argue that the largest banks have a remarkable data asset from which they could get significant competitive advantage over banks with smaller customer bases.

Considering these big data arguments, it is important to assess whether computational cost would prohibit the practical use of the new techniques. In fact, inference using the BeSim-based techniques is quite fast (an analysis of the computational requirements of the BeSim calculation was provided earlier). For all of the analyses in this paper, inference over the entire data set took about a minute (detailed measures are provided in Table 3). All analyses were conducted on an Intel Core i5-2400 CPU @ 3.10Ghz machine with 4Gb RAM. The BeSim procedure is implemented in Matlab, so the run time could likely be improved substantially. The linear SVM models were built using the LIBLINEAR package (Fan et al. 2008).

[9]For this analysis, for each training size we use all remaining data as testing data. Note that we are unable to assess the performance with 100% of the data used for training, as no more test data would be available to score.

As the data size grows, expected BeSim inference time shows a linear increase. Most time is spent on the initial preprocessing of the data, going from payment transaction data to the lists of customers for each merchant (as previously shown in Table 1). For our analyses, this required about a day of computation to incrementally read in and process the transactions.[10] Note that the BeSim model can be built incrementally; as new data become available the model can be updated in one fast operation.

Expert Feedback

The bank with which we worked for this study was particularly happy with the well-performing BeSim + SD model for the following reason: From the point of view of scoring consumers, we can consider the BeSim score to be just another variable. Then the combined BeSim+SD model is itself simply a linear model, where each component variable has a simple explanation. The structured variables were the variables already in use. The BeSim score has the intuitively satisfying interpretation as the similarity of the consumer prospect to the prior customers of the product in question. The fact that the variable had a calculation behind it was not problematic, as the other variables in use also do (e.g., RFM variables). Furthermore, the calculation is easy to explain intuitively. To automate the individual explanations of why the BeSim measure classified a particular consumer as being a good prospect, the method introduced by Martens and Provost (2014) can be applied, where an explanation would be defined as the set of merchants that a consumer paid, such that removing these payments would lead to the consumer not being predicted to be a good prospect. An example explanation for a consumer that is predicted to be interested in a student loan might be "if this consumer had not made a payment to online_course_XYZ, then the predicted class would change from interested to not interested in a student loan."

As anecdotal support for this line of work, BeSim-based targeting indeed was deployed by the bank. The production results are proprietary, but the bank claims that their own "A/B" evaluations support our conclusions: the prospects identified by the BeSim-based models actually purchased the product significantly more frequently than the prospects identified by the traditional targeting models.

[10]Presumably, a large bank with one or two orders of magnitude more data would not be running Matlab on a desktop PC. Moreover, modern big data architectures could speed up this sort of processing substantially.

The scalability and ease of interpretation and implementation of our method are of importance in the deployment of new techniques for targeting consumers. As described by Michael Wexler, Director of Digital Insights and Marketing Effectiveness at Citibank (Wexler 2014):

> Because of (perceived regulatory restrictions), for many decisions touching consumers or capital allocation, banks keep preferring to use models that are well understood in the community and are well supported out-of-the-box in the software they use (SAS, primarily), and are able to be easily examined around a) specific and clear impact of each input variable and b) the model consistency and stability. This works fine with regressions and other generalized linear models…but there is little guidance from the regulators on how to similarly review and judge machine learning models, many which have complex nonlinear impacts from predictors, and may be continuously changing (and therefore seen as non-stable). Even well regarded and documented models…are kept in the R&D wing, and are usually not part of mainstream bank decision processes around marketing, offer selection, or other optimizations for consumers. Some exceptions include the fraud-detection groups, who are given more leeway to experiment, and some of the specialized quant trading groups, who have less liquid markets requiring more advanced math to manage. At the end of the day, there is acceptable risk in the use of any model, and, while banks don't avoid all risk, they do tend to prefer measurable risk, and this is considered easier to do with traditional models.

A somewhat surprising aspect of behavioral similarity targeting is that it can be remarkably privacy friendly, in contrast to how it may seem *prima facie*. For the BeSim component, the only data required are

1. An anonymized transaction log: a list of anonymized payment transactions, denoting for each transaction the following attributes:
   - Consumers (anonymized, but reversible for targeting)
   - Nonconsumer merchants (anonymized; no need to be reversible)

2. The target values for a set of anonymized consumers for training

Each consumer as well as each merchant in the network can have her identity encrypted, not needing any name or account

number. In the case of the consumers, the encryption would be reversible in order actually to target a subset of them. However, the decryption could be limited to a protected, task-specific environment. This privacy friendliness is a very attractive feature in a banking setting as it does not allow modelers and analysts to view customers' names and payment profiles. In addition, the data would be useless to almost any recipient in the case of a data breach. As we have seen, additional data on consumer characteristics, as in the SD data, can improve predictive performance. These also could be encrypted for modeling.

Another operational advantage of this type of data concerns data quality (Moges et al. 2012): typically this is a major challenge when working with structured data. With payment data, however, no such issues arose: a simple "dump" from the transaction log is all that is needed for the BeSim method to be applied. The account number of the customer or the merchant was never missing or invalid. This might seem obvious and of little importance, but this has tremendous implications for the time saved on data preprocessing.

Before continuing to the conclusion, it is worth briefly discussing the proprietary nature of such behavior data and its implications for scientific research. The research area comprising data science and big data for business is highly dependent on close collaboration with industry, because understanding effectiveness depends on the actual characteristics and distributions of real data. When analyzing behavior data, as we do, privacy is always a major issue; such data contain information of consumers' everyday actions. Many studies (including this one) only receive such data in an encrypted form. For example, the data we received contained identifying information neither on the individual nor on the merchant. This limits the degree to which we can dig into the results to understand them more deeply.

A separate but related issue is that such sensitive company data rarely if ever can be shared with other researchers, changing the possibilities for replicability and follow-up studies. Thus, replicability is limited to other researchers applying the methods to similar data sets from the same or other organizations. One might argue that this is a more interesting sort of replicability than simply making sure that mistakes were not made in running the code on a particular data set, since replicating the results on similar data sets would test generality. Nevertheless, as a scientific community, we may want to elevate the discussion of the tradeoffs between being able to do certain sorts of studies at all and limitations based on restrictions to sharing data.

Such an examination might elevate a little-trodden research avenue: (how) can we create synthetic behavioral data sets that mimic the true data sufficiently to satisfy research needs, yet guarantee the confidentiality of the original data to satisfy our organizational partners? Related to this issue is the fact that, for our study, we relied on the bank for the preprocessing of the structured data (the 289 traditional variables) and we are not allowed to list the exact meaning of each of them. Therefore, we rely on the bank to have used correct data science practices, which again limits the reproducibility of our findings.

Conclusion

This paper provides an in-depth study of the use of a particular sort of big data—massive, fine-grained data on consumer behavior—to improve targeted marketing. Specifically, we examined the use of behavioral similarity for predictive modeling using fine-grained data on payments to specific individual merchants, in the context of targeting product offers to banking customers. We first isolated the computation of similarity from massive, fine-grained data by defining direct measures of behavioral similarity (BeSim). For two different banking products, the results show that the BeSim method is substantially better at placing consumers who purchase at "the top of the list" (i.e., score them as the highest-ranked individuals) than a traditional targeting model. Further, the BeSim model and the traditional model comprise complementary information. Combining the two produces a very robust model that gives better lift for almost any targeting budget. (The pure BeSim model still does better for the smallest selections of candidates.)

The BeSim calculation identifies those consumers most similar to key individuals of interest. The payment behavior data allow this similarity to be broadly based on the tastes, interests, and latent socioeconomic constraints represented by shared payment recipients and sources of money transfers. In our banking applications, the individuals of interest were prior customers of the products, so the BeSim found other consumers who were very similar along these dimensions to the existing customers. However, the BeSim design is not specific to targeting marketing offers. If the individuals of interest were chosen to be different tranches of total credit exposure, the BeSim method may be helpful for predicting wallet share. If the individuals of interest instead were chosen to be particularly good (or bad) known credit risks, the BeSim models ought to find other very similar consumers. Thus, behavior similarity could improve another very common modeling application: estimating creditworthiness.

The analyses also show a striking result of particular relevance to the current discussions of using big data for

improving business decision making (Provost and Fawcett 2013). In this application, predictive modeling using traditional structured data does not seem to be enhanced by increasing the amount of data to a massive scale. In contrast, targeting based on fine-grained behavioral similarity is indeed enhanced substantially by increasing the data size to a massive scale. This suggests that the already telling results presented in this paper may underestimate—possibly by a lot—the potential lift achievable by calculating behavioral similarity from all the data available to a huge bank. This provides one of the clearest illustrations (of which we are aware) of how large institutions have an important asset in the data they have collected, an asset from which they can get substantial competitive advantage over institutions without as much data—in this example, smaller banks.

The analyses showing that the alternative BeSim calculations perform very well also provide broader evidence that direct behavior similarity calculations indeed should be considered for research and practice when massive fine-grained behavioral data are available. The point of this paper was not to design the best method for this application or this sort of data; it seems likely that future work will show how to achieve even larger performance improvements from such data. (The alternative BeSim calculations based on the Beta distribution calculation are not suitable to large data on a desktop platform, but may be implemented feasibly using big data architectures.) Nonetheless, the basic BeSim calculation is quite simple to calculate and to implement, which is not a trivial factor when trying to deploy the model.

A first avenue for future research is defining and testing a weighted input matrix or bigraph, using RFM indicators of the payments. The timing of payments, both in terms of the time horizon to use, as well as the specific day and time of day of payment, might also be interesting further improvements to examine. Using different time horizons would also allow one to investigate the dynamics of the resulting network, for example, using the network with data of different quarters, years, etc. to create the models. Analyzing the resulting models, both in the workings of the models (e.g., the coefficients of the linear model) and the performance thereof could lead to additional insights into the domain and when to apply which technique.

From a technical point of view, other avenues for future research include defining extended heuristics, applying other large-scale classification techniques, and performing dimensionality reduction on the input matrix using singular value decomposition (Clark and Provost 2016).

Within a banking setting, other applications of our methodology include churn prediction, fraud detection, and default prediction, where for the latter we assume that the payments of consumers are likely also predictive for their creditworthiness. Note that such use is also applicable for credit card companies. Finally, behavioral similarity for predictive modeling is applicable well beyond the banking setting. Other companies (telecommunications companies, Amazon.com, online advertising companies, payment processors such as VISA and Paypal) have data on the specific merchants with which consumers transact. In this age of increased analysis of massive data, we hope that this paper can add to the new line of thinking into how firms best can use their data assets for consumer analytics in banking and beyond.

Acknowledgments

We are very grateful to the bank for providing us the data and working with us to achieve these results. We thank our reviewers for their constructive feedback. We give a special thanks to the editors of this special issue and the authors of the other articles in this issue, who provided us with valuable and constructive feedback during the workshop that took place in the summer of 2015. Foster Provost thanks NEC and Andre Meyer for Faculty Fellowships.

References

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering* (17:6), pp. 734-749.

Aral, S., Muchnik, L., and Sundararajan, A. 2009. "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences* (106:51), pp. 21.544-21-21.549.

Bergstra, J., and Bengio, Y. 2012. "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research* (13), pp. 281-305.

Borgatti, S. P., and Everett, M. G. 1997. "Network Analysis of 2-Mode Data," *Social Networks* (19:3), pp. 243-269.

Breiger, R. L. 1974. "The Duality of Persons and Groups," *Social Forces* (53:2), pp. 181-190.

Chapelle, O., and Keerthi, S. 2008. "Large Scale Support Vector Machines," presentation at the ICML Workshop on Large Scale Learning, July 9.

Clark, J., and Provost, F. 2016. "Matrix-Factorization-Based Dimensionality Reduction in the Predictive Modeling Process: A Design Science Perspective," Working Paper No. CBA-16-01, Stern School of Business, New York University.

Fader, P. S., Hardie, B. G. S., and Ka, L. L. 2005. "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," *Journal of Marketing Research* (42:4), pp. 415-430.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J. 2008. "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research* (9), pp. 1871-1874

Fawcett, T. 2006. "An Introduction to ROC Analysis," *Pattern Recognition Letters* (27:8), pp. 861-874.

Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, New York: Springer.

Hill, S., Provost, F., and Volinsky, C. 2006. "Network-Based Marketing: Identifying Likely Adopters Via Consumer Networks," *Statistical Science* (22), pp. 256-276.

Hormozi, A.M., and Giles, S. 2004. "Data Mining: A Competitive Weapon for Banking and Retail Industries," *Information Systems Management* (21:2), pp. 62-71.

Hu, X. 2005. "A Data Mining Approach for Retailing Bank Customer Attrition Analysis," *Applied Intelligence* (22:1), pp. 47-60.

Junqué de Fortuny, E., Martens, D., and Provost, F. 2013. "Predictive Modeling with Big Data: Is Bigger Really Better?," *Big Data* (1:4), pp. 215-226.

Latapy, M., Magnien, C., and Vecchio, N. 2008. "Social Networks: Basic Notions for the Analysis of Large Two-Mode Networks," *Social Networks* (30:1), pp. 31-48.

Linoff, G. S., and Berry, M. J. 2011. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, Hoboken, NJ: Wiley Computer Publishing.

Macskassy, S. A., and Provost, F. 2003. "A Simple Relational Classifier," in *Proceedings of the Second Workshop on Multi-Relational Data Mining,* Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, pp. 64-76.

Macskassy, S.A., and Provost, F. 2007. "Classification in Networked Data: A Toolkit and a Univariate Case Study," *Journal of Machine Learning Research* (8), pp. 935-983.

Martens, D., and Provost, F. 2014. "Explaining Data-Driven Document Classifications," *MIS Quarterly* (38:1), pp. 73-99.

McPherson, M., Smith-Lovin, L., and Cook, J.M. 2001. "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology* (27:1), pp. 415-444.

Moges, H., Dejaeger, K., Lemahieu, W., and Baesens, B. 2012. "A Total Data Quality Management for Credit Risk: New Insights and Challenges," *International Journal of Information Quality* (3:1) (http://dx.doi.org/10.1504/IJIQ.2012.050036).

Ng, A. Y. 2004. "Feature Selection, L_1 vs. L_2 Regularization, and Rotational Invariance," in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada.

Perlich, C., Provost, F., and Simonoff, J. S. 2003. "Tree Induction vs. Logistic Regression: A Learning-Curve Analysis.," *Journal of Machine Learning Research* (4), pp. 211-255.

Provost, F., and Fawcett, T. 2013. *Data Science and its Relationship to Big Data and Data-Driven Decision Making*, Sebastapol, CA: O'Reilly Media.

Van Den Poel, D., and Lariviere, B. 2003. "Customer Attrition Analysis for Financial Services Using Proportional Hazard Models," *Journal of Operational Research* (157), pp. 196-217.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.

Verbeke, W., Martens, D., and Baesens, B. 2014. "Social Network Analysis for Customer Churn Prediction," *Applied Soft Computing* (14:Part C), pp. 431-446.

Wexler, M. 2014. Private communication.

About the Authors

David Martens is a Professor at the University of Antwerp, where he heads the Applied Data Mining research group. His research focuses on the development and application of data mining techniques that lead to improved understanding of human behavior, and the use thereof in marketing and finance. In 2014, David won the "Best EJOR Application Paper Award" (*European Journal of Operational Research*), and in 2008 was a finalist for the prestigious international KDD doctoral dissertation award.

Foster Provost is Professor of Information Systems and Data Science, and Andre Meyer Faculty Fellow at New York University's Stern School of Business. He is coauthor of the best-selling data science book, *Data Science for Business*. His research has won many awards, including best paper awards at KDD, and the INFORMS Design Science Award. It also formed the basis for several data-science-oriented companies. Foster previously was editor-in-chief of the journal *Machine Learning*. His latest album, *Mean Reversion*, will be released in 2016.

Jessica Clark is a doctoral candidate in the Information, Operations, and Management Sciences department at New York University's Stern School of Business, concentrating in Information Systems. Her research focuses on applied data science, particularly in the domain of advertising.

Enric Junqué de Fortuny is an Assistant Professor at the Rotterdam School of Management, Erasmus University Rotterdam (The Netherlands). He holds a Master's degree in Computer Science Engineering from the University of Ghent (Belgium), a Ph.D. in Applied Economics from the University of Antwerp (Belgium), and was previously a Senior Research Fellow at INSEAD's eLab for Big Data (France/Singapore). His research interests include the development of machine learning algorithms with a specific focus on predicting human behavior and improving comprehensibility in data science.

Appendix A

Output Score

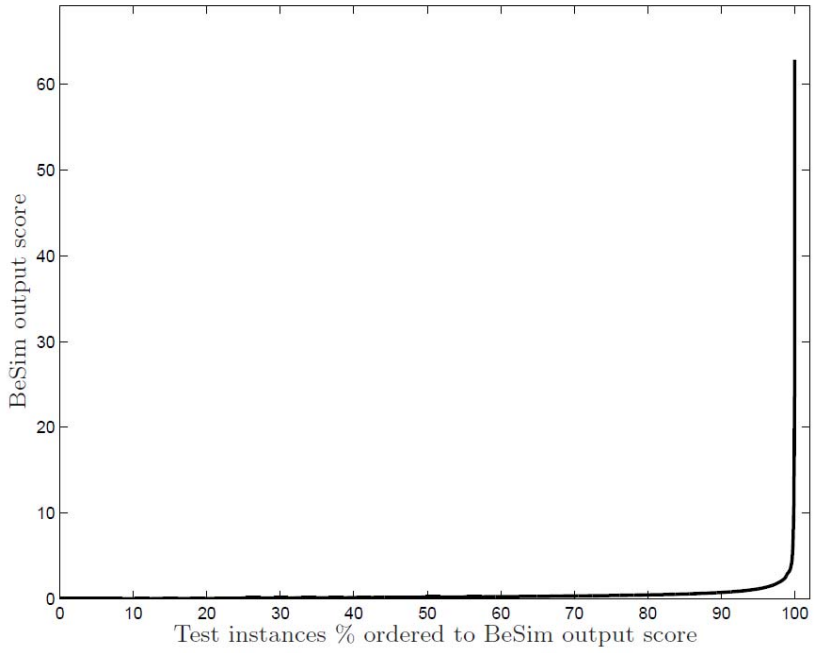


Figure A1. Output Score of BeSim Model for Product 1, with the Consumers Ranked According to the Output Score (Similarity for Product 2). Most consumers receive a (near-) zero score while a few receive a high score.

Appendix B

Feature Selection

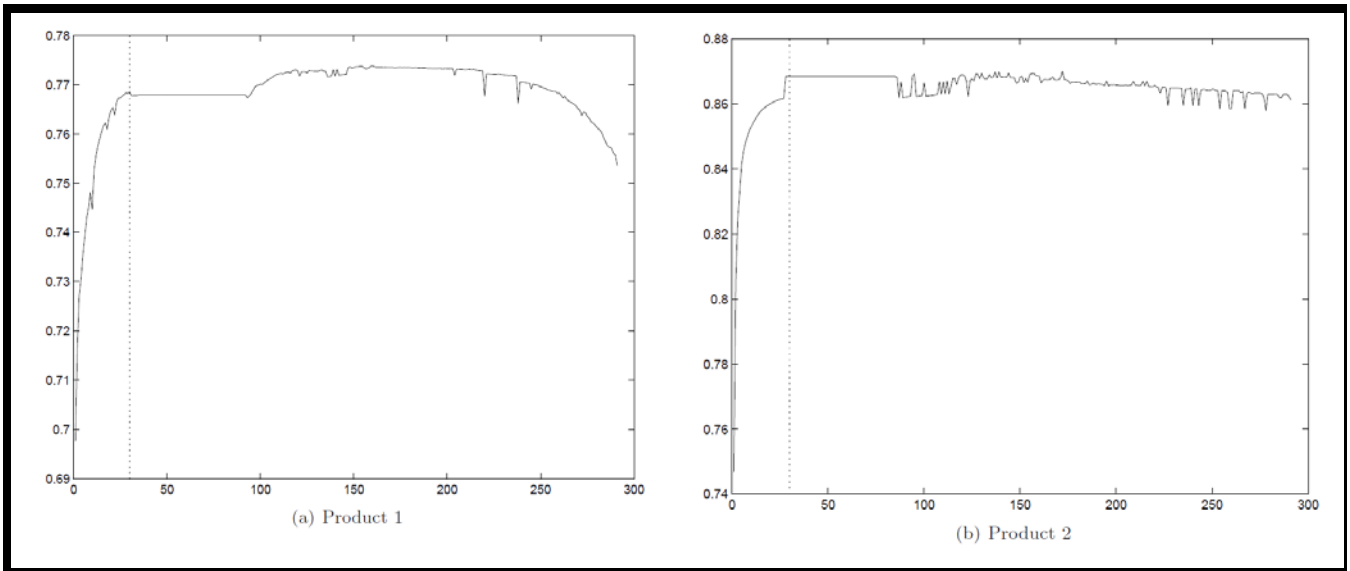


Figure B1. AUC (Y-axis) for an Increasing Number of Features (X-axis). We chose to use a maximum of 30 features, as a plateau is reached at that point for both products (marked with the dotted line).