eShopper Modeling and Simulation

Valery A. Petrushin

Center for Strategic Technology Research, Accenture
3773 Willow Rd., Northbrook, IL 60062

ABSTRACT

The advent of e-commerce gives an opportunity to shift the paradigm of customer communication into a highly interactive mode. The new generation of commercial Web servers, such as the Blue Martini's server, combines the collection of data on a customer behavior with real-time processing and dynamic tailoring of a feedback page. The new opportunities for direct product marketing and cross selling are arriving. The key problem is what kind of information do we need to achieve these goals, or in other words, how do we model the customer? The paper is devoted to customer modeling and simulation. The focus is on modeling an individual customer. The model is based on the customer's transaction data, click stream data, and demographics. The model includes the hierarchical profile of a customer's preferences to different types of products and brands; "consumption" models for the different types of products; the current focus, trends, and stochastic models for time intervals between purchases; product affinity models; and some generalized features, such as purchasing power, sensitivity to advertising, price sensitivity, etc. This type of model is used for predicting the date of the next visit, overall spending, and spending for different types of products and brands. For some type of stores (for example, a supermarket) and stable customers, it is possible to forecast the shopping lists rather accurately. The forecasting techniques are discussed. The forecasting results can be used for on-line direct marketing, customer retention, and inventory management. The customer model can also be used as a generative model for simulating the customer's purchasing behavior in different situations and for estimating customer's features.

Keywords: Customer modeling, shopping behavior simulation, agent-based simulation, consumption model, step stream data, click stream data, price sensitivity.

1. INTRODUCTION

The strategic objective of traditional data mining for Customer Relationship Management (CRM) is to create customer segmentation and build response models for targeted marketing for a particular product or service. A response model has a set of features and a set of explicit or implicit rules that selects customers who most likely can buy a product or a service [1]. The response model is based on some generalized data about a customer's recent transactions and demographics. The weakness of the approach is that the response model targets a "typical or average representative" of a particular segment of customers, but it may not fit any individual customer. Individual customer modeling promises essential improvements to CRM.

As an example, let us consider data on customer spending for a small supermarket chain. About 25,000 preferred customers were sorted in descending order according to their annual spending and divided into ten groups (deciles). The percentage of total sales spent by each decile is presented in Figure 1. It is evident that the customer value is different for different deciles. The top ten percent of the customers spent 47% of the whole money spent, and the top 40% accounted for about 90% of the supermarket chain's revenues. These 10,000 top customers deserved to be treated individually to improve their satisfaction and increase retention. Building individual customer models means creating a basis for establishing a learning relationship between a customer and an enterprise, when the customer teaches the enterprise about her needs, and the enterprise remembers these needs and adapts its behavior to satisfy the customer. Imagine a grocery store that cares about a customer's pantry and reminds her to check her supply of salt or sugar. The learning relationship builds a wall between the customer and the enterprise's competitors – even if a competitor provides the same level of service, the customer has to reteach the competitor to reach the same level of satisfaction. Building individual customer models is the key element of the enterprise's aftermarketing strategy [2] and the only way to build a customer-oriented enterprise or enterprise 1:1 [3].

The advent of e-commerce intensifies the interaction with the customer and allows to build the learning relationship. The new generation of commercial Web servers, such as the Blue Martini server [4], collects a new kind of data – click stream data. It also combines the collection of data with real-time processing of customer models for dynamic tailoring of a feedback page. The new opportunities for effective aftermarketing strategies are arriving.

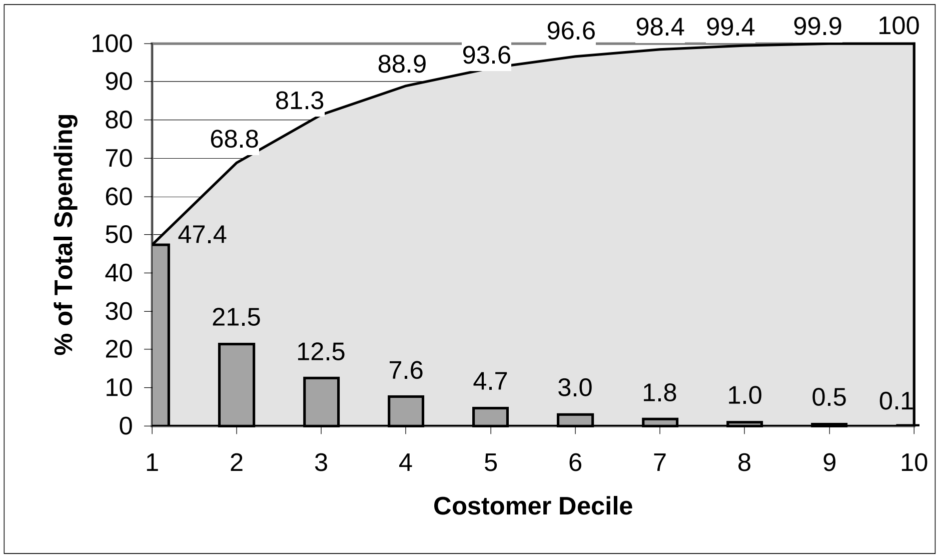


Figure 1. Distribution of customer spending by decile and cumulative spending.

2. CUSTOMER MODELING

In this section, consideration is given to the types of data that can be used for creating individual customer models, and types of models and modeling techniques that can be useful for achieving CRM objectives. Concepts are illustrated with examples from retail industry, and in particular, the on-line grocery business.

2.1 Data sources

The most important source of data about a customer is her *transactions*, i.e. what the customer bought and how much money she spent. A typical transaction record consists of date and time of purchase, name of a product, weight or count of items, and amount of money spent. Sometimes it may also have information about current promotions or discounts. All records related to the same purchasing session form a *basket*. From the viewpoint of a store manager, the transaction record is a description of purchase of a basic element of the store inventory that is specified by its SKU (a unique bar code). The typical inventory of a grocery store consists of 50,000 – 80,000 SKUs. Assume that an inventory is represented as a hierarchy of categories, subcategories, brands and SKUs (Figure 2). For example, "Country Fresh 2% Reduced Fat Milk" belongs to the category "Dairy", subcategory "Milk", brand "Country Fresh" and has the SKU #7160000901.

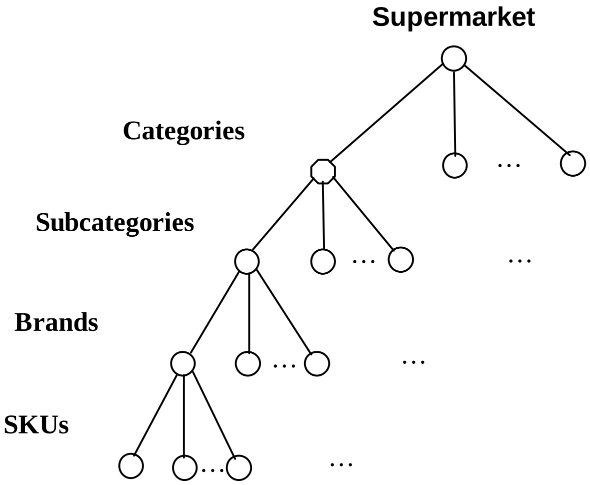


Figure 2. Supermarket inventory representation.

Another source of data is *demographics*. If you know the name and address of a person, you can purchase these data from customer data integration companies such as Claritas [5], Equifax [6], Donneley Marketing [7] or Acxiom RTC Inc.[8]. A record of demographic data can have from twenty to sixty fields that specify the customer's gender; age; occupation; length of residence at the current address; average income; presence of children and seniors in the household; how many and what kind of vehicles are used; what kind of credit cards the customer has; the type of property the customer owns; how much the property is priced; and some other data related to the customer's lifestyle and wealth. More detailed financial information can be obtained from the customer's *credit history* which collected from the Credit Bureau's report. Additional sources of data are *questionnaires* in the form of registration cards or on-line registration forms. Usually a questionnaire contains demographic questions and other questions, which were designed to reveal the customer's lifestyle and brand preferences.

The advent of e-commerce brought to life a new source of data – *click stream* data. Click stream data is a sequence of Web pages visited by the customer during an on-line session. Each element of the sequence is a record that contains the name or identification number of the current page and time the customer spent reading this page. This data is a trace of the customer's path through the commercial Web site. This data can be used to learn customer's characteristics based on her behavior and to determine brands and products that she could be potentially interested in. The analog of the click stream data in the physical world is *step stream* data, which is the sequence of the customer's coordinates and timing in a physical store. Imagine a shopping cart that has a tiny built-in transmitter and many small receivers built into the ceiling and shelves of the store. They can communicate with each other using the Bluetooth and WAP technologies [9]. A computer system can trace the cart, and later at the cash register, it can associate the step stream data with the basket and the customer, if she is a preferred customer, or if she uses a credit card to pay for products.

If we compare the above-mentioned sources of data, we can see that transactional data is very precise but too finely grained. These data need to be aggregated and associated with brands and categories of products to form an appropriate customer profile. The demographic data are less precise and often outdated, but could be very useful for estimating some customer characteristics, such as the purchasing power, and can provide some insight on the customer's lifestyle. The credit history is rather precise but very specific data. It may tell nothing interesting about the customer to a retailer such as a grocery store. The questionnaires are not very reliable sources of data. Usually they contain many missing values because customers try to avoid or minimize registration time. Step stream data are not in use yet, but, probably will be available in the near future. Click stream data can be available at each commercial Web site. Moreover, taking into account that only about 1-3% of the virtual store customers purchase products, the volume of click stream data is greater than the volume of transactional data in two or more orders of magnitude. The click stream data can be ambiguous and hard to interpret, but they are definitely the most valuable addition to transactions and demographics for creating richer, more accurate and robust customer models.

2.2 Modeling

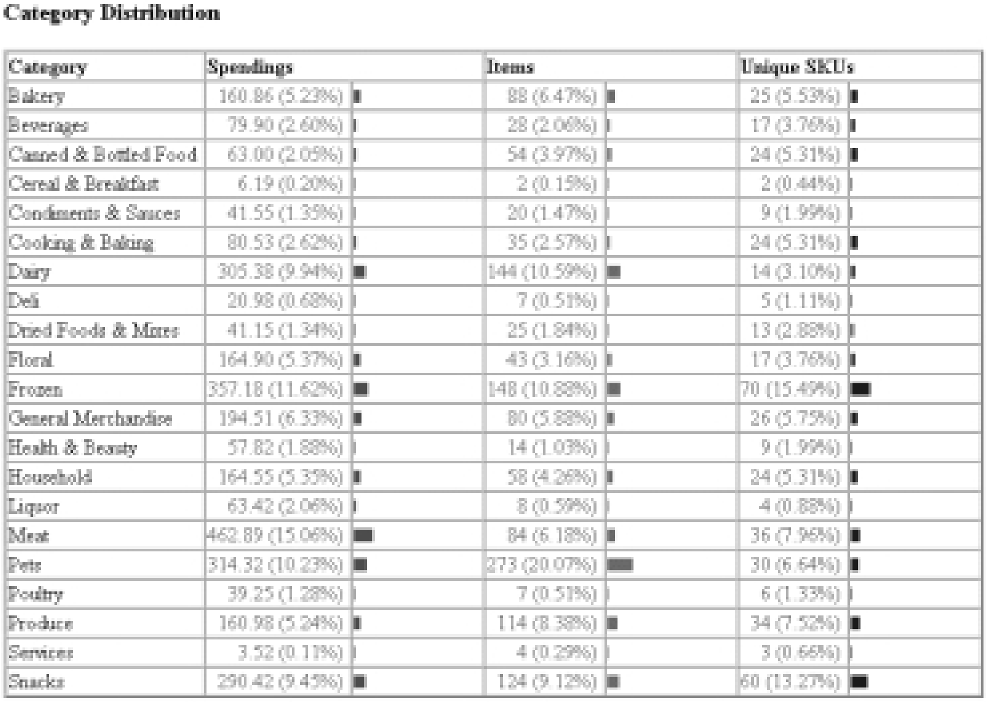
Data mining is traditionally defined as "the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules" [1]. The ultimate objective of data mining is to find some phenomena based on the data, but usually, data mining discovers some relationships among data and leaves the interpretation to people. Some researchers, including the author of this paper, believe that data mining can do more than just discover the relations among data. We believe that data mining techniques, coupled with knowledge management techniques, can discover some simple facts about the world and relationships among them. Eventually, these techniques will be able to discover more sophisticated phenomena and relationships including completely new ones. This school of thought is known as phenomenal data mining [10].

Consider an ideal customer model. For many retail businesses, a customer model is a household model. For example, for a supermarket shopper, the ideal model should contain knowledge about the customer's intentions, preferences, and shopping habits; some important assets the household has such as a refrigerator or a freezer;, consumption rates for different categories of products; purchasing power and number of people in the household along with the gender and age of each. It should also have knowledge about the customer's lifestyle and some customer's features such as price and advertising sensitivities, and sensitivity to mass behavior and fashion. It could be very useful to have knowledge about the size of the customer's network – the number of people with whom the customer communicates daily – and the customer's ability to communicate news and complaints through the network. Assume that we have transactional, demographical, and possibly click stream data for a customer for some period of time. How we can use these data to build a customer model?

There are three types of models: extensional, stochastic, and analytical.

An *extensional* model keeps all the customer data. For example, supermarket transactions can be represented as a sequence of baskets with each basket represented as a list of purchases. This representation could be useful for visualizing the history of the customer's purchases, but it has limited value because it does not allow us to see the forest behind the trees. It is preferable to have a more general view on the customer's preferences to different categories of products. A hierarchical or tree representation works much better. The root of the tree is marked as "supermarket" and contains aggregated data such as the total customer spending, the number of items bought, and the number of different SKUs bought. The nodes on the next level are marked by the names of categories ("Produce", "Dairy", "Meat", "Poultry", etc.) and contain aggregated data (spending, items, SKUs) for each category. The number of categories depends on the business and belongs to range from 3 to 30. Each category node has subcategory nodes (for example, for the category “Meat” may have the following subcategories: “Beef”, “Pork”, “Ham”, “Lamb”, etc.) and each subcategory has brand nodes. The number of subcategories/brands can be different for each category/subcategory. Usually, the number of subcategories ranges from 100 to 300, and the number of brands from 300 to 1000. The next level of nodes represents individual products (SKUs). The number of SKUs ranges from 20,000 to 80,000. For the supermarket problem, 23 categories, 193 subcategory, 380 brands and about 50,000 SKUs were used.

This representation, which we shall call a *purchase tree*, shows the customer’s preferences to categories/subcategories of products and brands. Building a purchase tree for each basket shows the *focus* of the visit to the store. The aggregated data at each node form time series for customer spending, number of items and number of SKUs. Normalizing data for categories, we can obtain the customer’s *basket profile*. In the same way, we can get the customer profiles for each category and brand. Building the purchase tree for all baskets purchased so far creates the customer’s *general profile* (Figure 3). Profiles form multidimensional time series that can be used to learn how the customer’s preferences change over time, and to predict customer spending. Abrupt changes in the customer profile signalize important events in the household, such as a marriage or a new baby arrival.



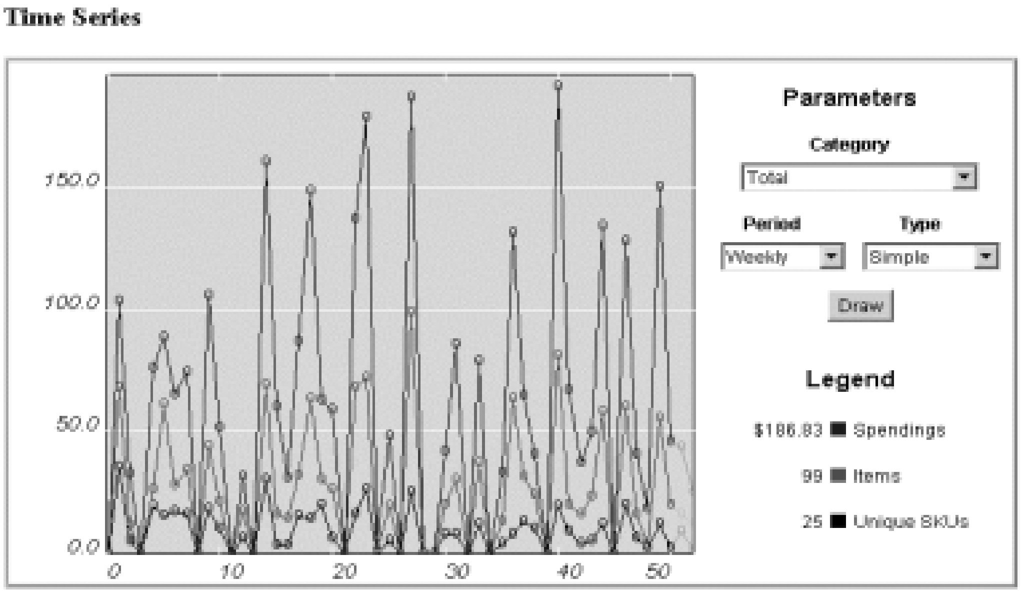
Category Distribution

Category	Spendings	Items	Unique SKUs
Bakery	160.86 (5.23%)	88 (6.47%)	25 (5.53%)
Beverages	79.90 (2.60%)	28 (2.06%)	17 (3.76%)
Canned & Bottled Food	63.00 (2.05%)	54 (3.97%)	24 (5.31%)
Cereal & Breakfast	6.19 (0.20%)	2 (0.15%)	2 (0.44%)
Condiments & Sauces	41.55 (1.35%)	20 (1.47%)	9 (1.99%)
Cooking & Baking	80.53 (2.62%)	35 (2.57%)	24 (5.31%)
Dairy	305.38 (9.94%)	144 (10.59%)	14 (3.10%)
Deli	20.98 (0.68%)	7 (0.51%)	5 (1.11%)
Dried Foods & Mixes	41.15 (1.34%)	25 (1.84%)	13 (2.88%)
Floral	164.90 (5.37%)	43 (3.16%)	17 (3.76%)
Frozen	357.18 (11.62%)	148 (10.88%)	70 (15.49%)
General Merchandise	194.51 (6.33%)	80 (5.88%)	26 (5.75%)
Health & Beauty	57.82 (1.88%)	14 (1.03%)	9 (1.99%)
Household	164.55 (5.35%)	58 (4.26%)	24 (5.31%)
Liquor	63.42 (2.06%)	8 (0.59%)	4 (0.88%)
Meat	462.89 (15.06%)	84 (6.18%)	36 (7.96%)
Pets	314.32 (10.23%)	273 (20.07%)	30 (6.64%)
Poultry	39.25 (1.28%)	7 (0.51%)	6 (1.33%)
Produce	160.98 (5.24%)	114 (8.38%)	34 (7.52%)
Services	3.52 (0.11%)	4 (0.29%)	3 (0.66%)
Snacks	290.42 (9.45%)	124 (9.12%)	60 (13.27%)

Figure 3. Customer’s profile.

Three *time series* are associated with every non-terminal node of the purchase tree. These time series present customer’s spending, number of items and number of unique SKUs per basket (Figure 4). For some categories these data are very sparse, for the others they are rather rich. For the time series, we can do qualitative analysis distinguishing between sparse and rich series. For the rich series we can do trend analysis to classify them into stable, increasing and decreasing series. Applying quantitative analysis we can predict how much money the customer will spend and how many items she will buy during the next visit to the store.

Click stream data can serve both to improve the structure of a commercial Web site and to learn about customers’ preferences. If we extend the purchase tree with products of description pages that were visited during a session, we obtain a *wishing tree*. This tree contains both purchased and visited items. Each item has a positive weight, which is equal to 1 for the purchased item, and is less than 1 for the visited item, depending on how much time the customer spent at its description page. The weights can be considered as fuzzy membership coefficients. Analysis of the wishing tree for the current session can determine the focus of the session, and comparison of the focus to the general wishing and purchase trees can detect the change in preferences. For example, take a customer that usually buys Coca-Cola’s pop soda products, but her wishing tree for the last session contains many highly ranked PepsiCo’s products. Then it might mean that the customer is about to change her preferences.



Time Series

Parameters

Category: Total

Period: Weekly

Type: Simple

Draw

Legend

$186.83 ■ Spendings

99 ■ Items

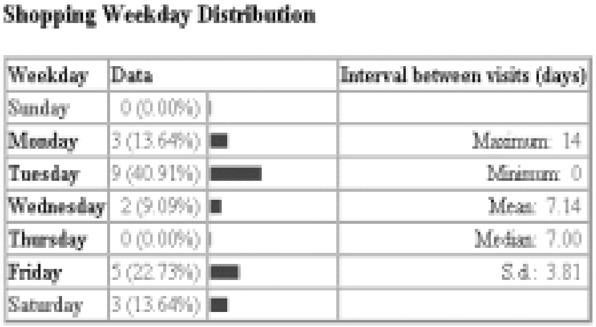
25 ■ Unique SKUs

Figure 4. Time series for customer spending, items and SKUs.

Another piece of information that might be useful and could be extracted from transactional and click stream data is which products are purchased together in the customer's baskets. Each customer has her own relationships among products. In data mining this problem is known as *affinity analysis*, and the result of analysis is a set of association rules on the product level with certainty factor assigned to each rule. Aggregating the rules, the relationships among brands, subcategories and categories are obtained.

A *stochastic* model represents concepts as random variables and uses data to estimate probabilistic distributions for the variables and relations among them. Consider the above-mentioned customer profiles as simple stochastic models of the customer's preferences. A number assigned to a category is considered as a probability of a product of this category to be found in customer's basket. Assigning probability distributions to all nodes of the purchase tree, it is possible to obtain the stochastic customer model that can be used as a generative model to imitate the customer's purchasing behavior. This type of model proved to be very useful in customer simulation research. Unfortunately, it is never enough data to estimate true distribution for 50,000 SKUs. This is why some heuristics should be used to assign probability to a SKU. For example, by assigning a small probability for products of brands that were not purchased, and larger probabilities to the products of brands that were purchased, more large probabilities to products from the wishing list, and probabilities that are close to estimated frequencies for purchased products. Taking into account the relationship among products, brands, subcategories and categories, we can extend the purchase or wishing tree to the *purchase* or *wishing network*, which can be specified in a form of a Bayesian belief network [12] and serve for predicting customer's shopping list.

Let us consider the days of week when the customer visits the store. The simple histogram can be generalized to a probability distribution (see Figure 5), which in combination with average interval between visits gives us a predictive model for the visiting days.



Shopping Weekday Distribution

Weekday	Data		Interval between visits (days)
Sunday	0 (0.00%)		
Monday	3 (13.64%)	■	Maximum: 14
Tuesday	9 (40.91%)	■■■■	Minimum: 0
Wednesday	2 (9.09%)	■	Mean: 7.14
Thursday	0 (0.00%)		Median: 7.00
Friday	5 (22.73%)	■■	S.d.: 3.81
Saturday	3 (13.64%)	■	

Figure 5. Days of week distribution.

A first-order Markov chain model, that gives the probabilities of the pairs of visiting days, allows detecting such events as change of the typical day of visit or finding a typical sequence of visiting days. For example, suppose there are two sequences of day of visits:

(1) Wed, Wed, Wed, Wed, Wed, Sat, Sat, Sat, Sat, Sat, Sat and

(2) Sat, Wed, Sat, Wed, Sat, Wed, Sat, Wed, Sat, Wed, Sat.

The first sequence indicates that the customer decided to change the day of visits to the store. The second sequence demonstrates regularity in days of visits, probably two members of the household visit the store on different days. The histograms for these sequences look exactly alike (Wed: 5, Sat: 6). But first-order Markov chain models look quite different (see Table 1, where tables 1a and 1b present the model for the first and the second sequences correspondingly). Detecting the second event allows separating data from different members of the household, learning them in-depth, and tailoring an individual promotional program for each member.

Table 1. First-order Markov chain models for day of visits.

(1a)

	Wed	**Sat**
Wed	0.4	0.1
Sat	0.0	0.5

(1b)

	Wed	**Sat**
Wed	0.0	0.5
Sat	0.5	0.0

An *analytical* model represents relationships between concepts in the form of equations and rules. Creating production rules is the most general approach to expert knowledge representation [11]. Rules can be used for representing relations among data, among data and concepts (phenomena), among concepts, and for deriving phenomena from the data. For example, if a customer is buying a lot of frozen products, then she has a freezer. On the other hand, if a customer's demographics tell that she has pets, it can be derived that she is supposed to buy pets' products. If such products are not found in the baskets, then it can be concluded that a) she uses another shop for buying pets' products or b) the demographic data is outdated. This hypothesis may be confirmed by issuing the customer coupons for pets' products or in a virtual store by directly asking the customer. Another, more complex, example is if a household's real monthly spending at supermarket is significantly less than the projected consumption norm, then this customer is a partial customer. It means that the customer uses this supermarket for satisfying only some of her needs, say for purchasing produce or dairy products. The projected consumption norm is an estimate of a household monthly spending. It depends on the number of people in the household, their ages and the household's income. For example, for a family of two adults and one child under 5, with an income in the range from $50,000 to $70,000, the estimate is 2*$200 +$150 = $550. It is interesting to note that a large portion of the second decile customers and practically all higher decile customers belong to this category. The relationship with a partial customer could be very fragile and depends on the supermarket's inventory and location of competitors' stores. Knowing this fact, the supermarket management could develop a retention program for this customer or could try to extend the customer's interests.
Another important element of customer model is a *consumption model*. A consumption model gives an analytical representation of the consumption rate for a particular product or a category of products. Consumption models can be different for different products. Some products that are used every day, such as cereal, milk, and bread, can be described by the linear function of time. Other products, such as meat, poultry, snack bars, and candy, can be better represented by the exponential function. To estimate the parameters of consumption models, the exact amount (weight, volume) of the purchased product is needed. Sometimes these data are not available in transactions and should be retrieved from the product database. Figure 6 shows a customer's pattern for purchasing butter during one year. A package of butter weights 250 grams. The first pack was bought on 4th day, the next two packs on 38th day, two more on 62nd day, another two on 94th day, and so on. Table 2 contains the current and average consumption rates for butter, assuming that the linear function is used. It can be seen that this customer purchases butter regularly and consumes about one pound of butter per month. The two-month gap in purchases can be explained that the customer bought butter of different brands or another product which substitutes butter. Assigning a consumption model to each node of the purchase tree allows to estimate the consumption rate for the whole hierarchy of product categories. Moreover, the amount of products that is kept in the customer's pantry can be estimated for every category at any moment of time. This knowledge can be used to make more precise prediction of the customer's next purchase.

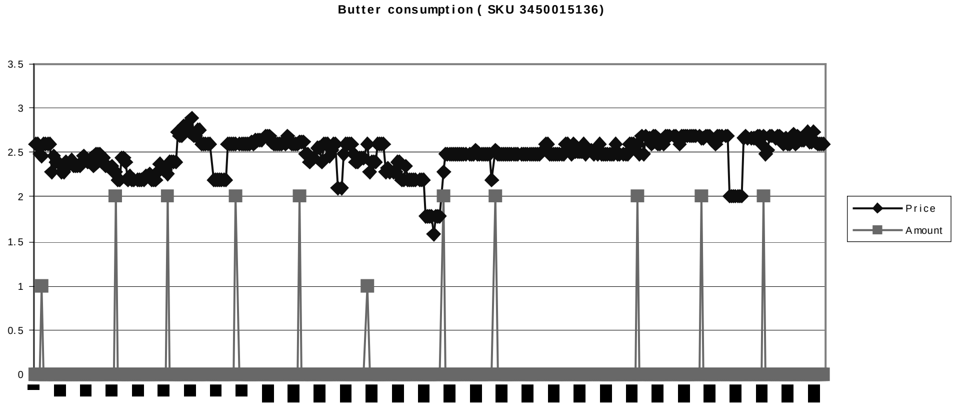


Figure 6. Butter consumption pattern.

Table 2. Consumption rates for butter.

Day	4	38	62	94	124	152	190	212	280	308	338
Amount	250	500	500	500	500	250	500	500	500	500	500
Current CR	?	7.35	20.83	15.63	16.67	17.86	6.58	22.73	7.35	17.86	16.67
Average CR	?	7.35	12.93	13.89	14.58	15.20	13.44	14.42	12.68	13.16	13.47

We can use transactional data to derive some customer characteristics such as price sensitivity. A customer is price sensitive if she buys larger amounts of a product when the price is low and smaller amounts of the same product when the price is high (see Figure 7). Let us characterize customer price sensitivity by a number from 0 to 1, where 0 means that the customer is indifferent to the price and 1 means that the customer is a true bargain finder. A customer can be very sensitive to some categories of products and indifferent to the others. To estimate the customer price sensitivity, choose a product our customer cares about. If there are enough transactional data, the price history for this product can be derived. Then we can compare the customer's purchases to the price history and estimate the feature. One of the metrics that can be used is the coefficient of linear correlation between amount of product and price. However it does not work when the amount is stable. We use the following metrics:

$$s = 1 - \frac{\sum_k p_k v_k - p_{\min} \sum_k v_k}{(p_{\max} - p_{\min}) \sum_k v_k}$$

where s is price sensitivity;
v_k is amount of product purchased at *k*-th visit;
p_k is the price of the product at time of *k*-th visit;
p_{max} is the maximal price;
p_{min} is the minimal price.

According to the above formula, if the customer paid minimal price for all purchases of the product then her price sensitivity is 1; if she paid maximal price then her price sensitivity is 0. Taking into account the dynamic nature of prices and regularity of purchases of most important products, it is unlikely that somebody will have price sensitivity that equals 0 or 1. The sensitivity for the customers depicted on Figure 7 is 0.73, 0.29, and 0.5 correspondingly. The optimal time interval for calculating price sensitivity depends on price dynamics and frequency of purchases, but for most grocery products, 3-5 month time intervals give robust estimations. Unfortunately, the above formula does not take into account the current amount of product that the customer has in her pantry and it needs further refinements.

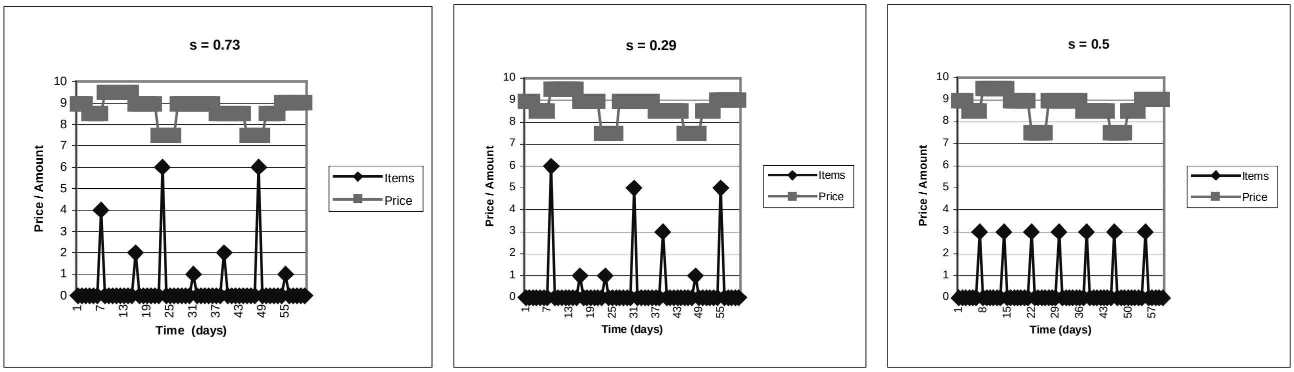


Figure 7. Estimating price sensitivity.

2.3 Models at work

Now consider how the above data and techniques work to produce valuable customer models.

The first step is to collect enough transactional and click stream data to built an initial model. Usually, ten or more baskets are needed. Taking into account that the average interval between visits to a supermarket is less than 6 days, roughly two-month data is needed. From these data we can estimate a very important customer's feature – an interval between visits. First, estimate the interval and its standard deviation based on the date of visits; then, aggregate two physical visits (and their baskets) into one logical visit (logical basket) if the difference between them is less than standard deviation rounded to the whole number. After this recalculate the interval. For example, we have 12 visits with the following differences between them: 6, 6, 1, 5, 6, 6, 7, 7, 2, 4, 6; the average interval equals 5.1 and standard deviation equals 1.97. Applying the above procedure we aggregate the third and forth visits and the ninth and tenth visits to get the following sequence of intervals: 6, 6, 6, 6, 6, 7, 7, 6, 6 with the average 6.2 and standard deviation 0.44. This estimate can be used to predict the next visit of the customer to the physical or virtual store. Also, the distribution of visits on days of week can be estimated, and a first-order Markov chain can be created to learn in depth the customer's pattern of visits (see Table 1 above).

The next step is to analyze the customer's transactional and click stream data. This step includes building the customer's purchase and wishing trees, creating profiles on the level of brands, subcategories and categories, doing affinity analysis to learn relationships between different brands, subcategories and categories, and doing time series analysis to predict the customer's spending, number of items bought and number of new SKUs. We found that traditional approaches to time series forecast [13] for spending and number of items for non-sparse categories give acceptable results, which can be used for predicting the customer's shopping list (see below). Customer profiles on categories level can be used for customer segmentation using clustering techniques such as self-organizing maps, and combining the result with demographic data to learn more about groups of supermarket customers [14]

The next step is to obtain the customer's demographic data, verify them using the transactional and click stream data, and derive some phenomena of interest. Demographic data contain mostly phenomenal data that tells about what the customer has, such as a house, cars, boats, credit cards, babies, children, dogs, etc., and some estimates, such as the customer's annual income. Using these data in combination with transactions and click streams can help to derive some other phenomenal data, such as the possession of a freezer, microwave or a grill. But demographics could be unreliable or outdated, and that's why some evidence that confirms or rejects the fact should be collected. For example, a demographic record shows that the customer has both cats and dogs, but transactions show the purchases of only dog food. Unlikely, the customer buys dog food in our store and cat food in another store (unless she buys a very unique item available only in this store). This might confirm that the piece of demographic data regarding cats is wrong. The conclusion gets additional confirmation if the customer ignored an on-line coupon for cat food that was issued to her by our personification system. On the other hand, if a piece of phenomenal data cannot be confirmed or rejected, then it can be decided that the customer uses another store to buy related products. It gives the store an opportunity to extend the customer's interests by advertising and issuing coupons for this kind of product.

The next step is to build consumption models and to estimate some customer's characteristics, such as price and advertising sensitivity. A consumption model can be built for any product that was bought at least twice, but consumption models on subcategory level, are usually much more reliable than models on brand and product level. To estimate the customer's price sensitivity we need to pick up products, which both have variability of price and are really important to the customer. Using the above metrics we can estimate the price sensitivity for each product and then calculate average, maximal and minimal price sensitivity. It is very hard to estimate reliably the customer advertising sensitivity in a physical supermarket, but in a virtual store we can use the proportion of accepted individually tailored advertisements of the total number of advertisements.

The next step is to build a stochastic model for predicting the customer's shopping list. This model uses probability distributions on category, subcategory, brand and product level to choose a set of products that satisfies the restrictions imposed by forecasts for spending, number of items and number of new SKUs for each level and the current state of the customer pantry estimated by consumption models. We used a strictly hierarchical model, but it is possible to create a network model taking into account the relationships among concepts on each level imposed by affinity analysis. The predicted shopping list is used for personalizing interaction with the customer, such as presenting individually tailoring advertisements to introduce a new product, issuing coupons, and asking questions to collect more evidence for confirming some conclusions. The predicted shopping list can be presented to the customer as a recommendation. The shopping lists for all customers can be used for store inventory management. After each session the customer model is updated, the current basket is compared to the model to infer the tendencies and events, and predictions for the next session are generated.

3. CUSTOMER SIMULATION

Customer's purchasing behavior simulation is a part of more general problem of social simulation. Much of the work in social simulation involved studying macroeconomic, environmental and demographic issues. Social systems involve humans, organizations and their interactions. Executives dealing with complex social systems often make decisions that can lead to a number of unintended consequences. One of the important new trends in simulation technologies involves building tools for executives to help them navigate through the consequences of a decision, a strategy or a policy in a complex system.

In this research, we used the above customer models to generate shopping lists under different states of customer's pantry and to simulate customer's interactions with the system varying customer's features. The main objective of research is to develop strategies that achieve such goals as extending customer's interests to a category of products that is not listed in her profile, retaining a customer by improving the customer's satisfaction, and confirming some elements of knowledge about the customer by using direct and indirect questions.

Now we are in transition to extend our research agenda to the simulation of the whole physical or virtual store. The objective of the simulation is to observe the consequences of a decision or a strategy that related to the store depending on the population of the store. For example, the store management made decision to upscale the quality of products by introducing organic products. How will this decision may influence the demographics of the store population, attrition rate, and store's revenues?

For this research we intent to follow agent-based methodology and use the above described individual customer models as agents [15]. Compared to traditional approach to simulation, when parameters of agent models are made up or derived from made-up distributions, our approach has an advantage that agent models are derived from real data and better represent the target population.

4. REFERENCES

1. Michael J.A. Berry and Gordon Linoff *Data Mining Techniques for Marketing, Sales, and Customer Support.* John Wiley & Sons, Inc.: 1997.
2. Terry G. Vavra *Aftermarketing. How to keep customers for life through relationship marketing.* McGraw-Hill. 1995.
3. Don Peppers and Martha Rogers *Enterprise one to one. Tools for competing in the interactive age*. Currency-Doubleday. 1999.
4. Blue Martini http://www.bluemartini.com
5. Claritas http://www.claritas.com/
6. Equifax http://www.equifax.com/
7. Donneley Marketing http://www.donnelleymarketing.com/
8. Axciom RTC Inc. http://www.acxiom.com/
9. Bluetooth technology: http://www.bluetooth.com/
10. John McCarthy Phenomenal data mining: from data to phenomena. http://www-formal.stanford.edu/jmc/phenomenal.html
11. Patrick H. Winston *Artificial Intelligence*. Addison-Wesley Pub Co: 1992.
12. Richard E. Neapolitan *Probabilistic reasoning in expert systems: theory and algorithms*. John Wiley & Sons, Inc.: 1990.
13. George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel *Time series analysis: forecasting and control*. Prentice-Hall: 1994.
14. Marie Cottrell, Patrice Gaubert, Patrick Letremy, Patrick Rousset Analysing and representing multidimentional qualitative data: Demographic study of the Rhone valley. The domectic consumption of the Canadian families. In E. Oja and S. Kaski (eds.) *Kohonen Maps*. Elsevier:1999. pp. 1-14.
15. Joshua M. Epstein and Robert Axtell *Growing Artificial Societies: social science from the bottom up*. Brookings Institution Press: 1996.