Models of Customer Behavior: From Populations to Individuals

Tianyi Jiang
IOMS Department
Leonard N. Stern School of Business
New York University
44 West 4th Street
New York, NY 10012 USA
tjiang@stern.nyu.edu

Alexander Tuzhilin
IOMS Department
Leonard N. Stern School of Business
New York University
44 West 4th Street
New York, NY 10012
atuzhili@stern.nyu.edu

Abstract

There have been various claims made in the marketing community about the benefits of 1-to-1 marketing versus traditional customer segmentation approaches and how much they can improve understanding of customer behavior. However, few rigorous studies exist that systematically compare these approaches. In this paper, we conducted such a systematic study and compared the performance of aggregate, segmentation, and 1-to-1 marketing approaches across a broad range of experimental settings such as multiple segmentation levels, multiple real world marketing datasets, multiple dependent variables, different types of classifiers, different segmentation techniques, and different predictive measures. Our results show that, overall, 1-to-1 modeling significantly outperforms the aggregate approach among high-volume customers and is never worse than aggregate approach among low-volume customers in our experimental settings. Moreover, the best segmentation techniques tend to outperform 1-to-1 modeling among low-volume customers.

1. Introduction

Customer segmentation, such as customer grouping by the level of family income or education, is considered as one of the standard techniques used by marketers for a long time [1]. Its popularity comes from the fact that segmented models usually outperform aggregated models of customer behavior [2]. More recently, there has been much interest in the marketing and data mining communities in building individual models of customer behavior within the context of 1-to-1 marketing [3] and personalization [4]. Although there have been many claims made about the benefits of 1-to-1 marketing [3], there has been little scientific evidence provided to this regard and no systematic studies comparing individual, aggregate and segmented models of customer behavior have been reported in the literature.

In this paper, we address this issue and provide a systematic study in which we compare performance of individual, aggregate and segmented models of customer behavior across a broad spectrum of experimental settings. We found that in general, there exists a tradeoff between the sparsity of data for individual customer models and customer heterogeneity in aggregate models: individual models may suffer from sparse data, while aggregate models suffer from high levels of customer heterogeneity.

We studied this tradeoff across different experimental settings. Our results show that the individual level models significantly outperform aggregate and segment level models for high-volume customers and are never worse than aggregate models for low-volume customers across these experimental settings. Also, the best segmentation techniques perform significantly better than the aggregate and individual level models for low-volume customers. In addition, we present various other results of comparison among aggregate, segmentation and individual approaches.

Before presenting the results, we formulate the problem setting in Section 2 and provide a literature review in Section 3. We present the experimental setup and results in Sections 4 and 5.

2. Problem Formulation

To build predictive models on customer behaviors, we used panelist datasets that track a set of customers' transaction histories over time. The unit of analysis is an individual customer. Predictive models of customer purchase behavior, such as total price of purchase or time of the day of the purchase, are generated from customer's purchase history. We also calculated customer-specific summary statistics, such as average purchase price, purchase frequencies for day of the week, etc. These summary statistics and customer's demographic data are then used to build progressively refined segmentations of the customer base.

More formally, let C be the customer base consisting of N customers, each customer C_i is defined by the set of m demographic attributes $A = \{A_1, A_2, \ldots, A_m\}$, k_i transactions $Trans(C_i) = \{TR_{i1}, TR_{i2}, \ldots, TR_{ik}\}$ performed by customer C_i, and h summary statistics $S_i = \{S_{i1}, S_{i2}, \ldots, S_{ih}\}$, computed from the transactional data $Trans(C_i)$. Moreover, each transaction TR_{ij} is defined by a set of transactional attributes $T = \{T_1, T_2, \ldots T_p\}$. The number of transactions k_i per customer C_i varies from high-volume customers with thousands of purchase transactions to low-volume customer with only a few purchase transactions (we restrict low-volume customers to have at least 10 transactions to ensure enough data for 10-fold cross validation on our models). For example, a customer C_i can be defined by attributes A= {Name, Age, Income, and other demographic attributes}, by the set of purchasing transactions $Trans(C_i)$ she made at a Web site, each transaction defined by such transactional attributes T as an item being purchased, when it was purchased, and the price of an item. Finally, summary statistics S_i can be computed for a purchasing session and can include such statistics as the average amount of purchase per session, the average number of items bought, and the average time spent per purchase session.

Given this data, we learn predictive models of customer behavior of the form

$$Y = \hat{f}(X_1, X_2, \ldots, X_p) \qquad (1)$$

where $X_1, X_2, \ldots, X_p$ are some of the demographic attributes from A and some of the transactional attributes from T. Function $\hat{f}$ is a predicative model learned via different types of machine learning classifiers from the transactional and demographic data described above, as will be explained below.

Various models of customer behavior can be built at different levels of analysis when customers can be grouped into different *segments* based on some of their demographic and behavioral characteristics. Moreover, we can have different levels of analysis depending on how finely we want to partition the customer base into various segments. In this paper, we consider the following three levels of analysis:

- *Aggregate level* – when the unit of analysis is the *whole* customer base, and only *one* predictive model of customer behavior (1) is built for the whole customer base. Moreover, this model is learned from *all* the transactional and demographics data of all the customers aggregated into one dataset.
- *Segmentation level* – when “similar” customers are grouped into progressively finer *segments,* and the model(s) of customer behavior are built at each segment level based on the transactions and the demographic data of *that* particular grouping of customers. In this case, we still use the model of type (1) but learn it from the data pertaining *only* to the selected segment of customers. Moreover, we do this for *each* customer segment. In our study, the degree of customer similarity is determined via different clustering methods.
- *Individual (or 1-to-1) level* – when the unit of analysis is an individual customer, the model of customer behavior is built based *only* on the purchase transactions of that particular customer and his or her demographic data. In other words, we build model of type (1) for *each* customer C_i in the customer base using the transactional history $Trans(C_i)$ and the demographic data of that customer. Such customer-specific models capture idiosyncrasies of the purchase behavior of individual customers.

As we progress from the aggregate to the segmented and then to the individual models of customer behavior, as described above, we would create increasingly more “homogenous” customer groups for which predictive models are theorized to be more accurate. However, while we consider more and more refined segments containing fewer and fewer customers, the less data is contained in each customer segment, and the estimation of function $\hat{f}$ in (1) is based on fewer data points thus, potentially, resulting in less accurate estimates.

Thus the general research question is to *determine which level of analysis would provide better prediction of customer behavior*, as defined by some measure of predictive performance of models of type (1). The answer to this question depends on the tradeoff between the sparsity of data for individual customer models and customer heterogeneity in aggregate models. In those applications where the customer base is homogeneous and the customer’s transactional data is sparse, aggregate models should dominate; and in those applications where customer base is heterogeneous and the customer’s transactional data is abundant, individual models should dominate.

In this paper, we study this tradeoff experimentally by comparing predictive models of type (1) across the three levels of analysis (i.e. individual vs. aggregate, individual vs. segmentation, and segmentation vs. aggregate) and six dimensions of different

- Types of data sets
- Types of customers (high vs. low-volume)
- Types of predictive models (classifiers)
- Dependent variables
- Performance measures
- Segmentations techniques (clustering algorithms)

We explain each dimension of comparison below.

Types of dataset. Few real world marketing datasets are publicly available for research use. In our study we focused on two marketing datasets: ComScore panelist dataset from Media Metrix on Internet browsing and buying behaviors of one hundred thousand users across United States for a period of 6 months (available via Wharton Research Data Services - http://wrds.wharton.upenn.edu/); and the Nielson panelist dataset on beverage shopping behaviors of 1,566 families for the year 1992.

The two marketing datasets are very different in terms of the type of purchase transactions (Internet vs. physical purchases), variety of product purchases, number of individual families covered, and the variety of demographics. Compare to Nielson's beverage purchases in local supermarkets, ComScore dataset covers a much wider range of products and demographics and is more representative to today's large marketing datasets.

Types of customers. Since we are only interested in purchase behaviors, we reduced our datasets to families with at least 10 transactions (see footnote 1). We partitioned our datasets into high-volume and low-volume customers in order to study the effect of data sparsity. Ideally, we would also like to experiment across the entire customer population for both ComScore and Nielson datasets, but the sheer size of ComScore dataset and our computational requirements across all dimensions of analysis make this nearly impossible. Thus we created 5 datasets of high and low-volume customers for ComScore, and high, low, and all-volume customers for Nielson. Table 1 shows the breakdown of the number of transactions for each customer type.

Note that the top 5% of ComScore customers in terms transactional frequency conducted more purchase transactions than the entire Nielson dataset. From the transaction totals of the high and low-volume customers, it is also evident that the disparity between the two types of customer is much greater in the ComScore dataset. The difference in average transaction frequencies among the low-volume customers of different datasets will play a role in our results.

Table 1. Customer Types and Transaction Counts

DataSet	Customer Type	% of Total Population	Families	Total Transactions
ComScore	High	5%	2,230	137,157
ComScore	Low	5%	2,230	24,344
Nielson	High	10%	156	28,985
Nielson	Low	10%	156	5,007
Nielson	All	100%	1,566	132,210

Types of predictive models. We use three different types of classifiers in building predictive models: C4.5 decision tree [5], Naïve Bayes [6], and rule based RIPPER [7]. The three classifiers are chosen because they represent different and popular approaches to predictive model building, and they are fast in execution time. We generated a total of 216,159 unique predictive models across all dimensions of analysis in this study, and the amount of computational effort makes classifier speed a practical concern. Computational time constraint is also a critical factor behind our decision to not use other high performance classifiers such as support vector machines. Since our goal in this research is to observe and study various factors that could influence the relative performance of various customer segmentation levels, it is not critical to use the "best" possible classifiers, nor necessary to specifically tune classifier parameters to achieve the "best" performance. In fact, to achieve consistency, we ran all three classifiers with the same default parameters for all predictive modeling tasks.

Dependent variables. We built various models to make predictions on transactional variables since our goal is to compare discussed approaches across different experimental settings. The data we used to train any one model are customer C_i's demographic data as well as all other transactional variables not used for prediction in that specific model. We used 8 ComScore transactional attributes as dependent variables in our models: Internet purchase session duration, number of webpage viewed, time of the day, day of the week, category of the website, product category, product price, and basket total price. We also used 5 Nielson transactional attributes: category of drinks bought, primary shopper's gender, day of the week, quantity of drinks bought, and total price.

Performance measures. We used Weka 3.4, from the University of Waikato [8], for all predicative modeling tasks. Each classifier generates a model via ten-fold cross validation. The predictive power of the model is then evaluated via three performance measures: percentage of correctly classified instances, root mean squared error, and relative absolute error that are defined as [8]:

- Correctly classified instances (CCI) = $\frac{\sum_{1}^{n} TP_i}{n}$

  where $\left\{ \begin{matrix} TP_i = 1 & if\ p_i = a_i \\ TP_i = 0 & if\ p_i \neq a_i \end{matrix} \right\}$, p_i is the predicted value, a_i is the actual value, and n is the total number of observations in a customer segment.
- Root Mean-Squared Error (RME)
  $= \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$
- Relative Absolute Error (RAE) $= \frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \overline{a}| + \ldots + |a_p - \overline{a}|}$

  where $\overline{a}$ is the average value of the predicted class.

Given models α and β, α is considered "better" than β when:

$$CCI_\alpha > CCI_\beta$$
$$RME_\alpha < RME_\beta$$
$$RAE_\alpha < RAE_\beta$$

Segmentation techniques. We segment the customer base using standard clustering techniques. In particular, we generate progressively smaller groupings of customers via five levels of segment/sub-segment hierarchy. With the exception of random clustering, all clustering algorithms try to find similarity among customers from customer summary statistics and demographics. So to split group g_j in sub-segment level *l*, we input to a clustering algorithm the set of summary statistics and demographic information (XC) for all customers c_i in group g_j (σ denotes a select operator, and ⋈ denotes a join operator):

$$XC = \sigma_{C_i \in g_{jl}}(S) \bowtie \sigma_{C_i \in g_{jl}}(A)$$

Random clustering, where the customer base gets segmented into random groups regardless of customer "similarity", is used as the control group. Predicative models of customer behavior based on random clustering of customers are used to gauge how well a particular clustering technique segmented the customer base. A segmentation technique is considered "bad" if the resultant performance measures are statistically equivalent to that of random clustering of customers.

For each new level *l*+1 in the segment/sub-segment hierarchy, we created *k* new customer groups from a single customer group g_j in level *l* of the hierarchy. The branching factor *k* is used to control the granularity of clusters. If *k* is set too high, we would approach near individual level clustering as we create increasingly smaller customer groupings at subsequent sub-segment levels. Due to different dataset sizes, we used branching factors of 3 and 2 for ComScore and Nielson data respectively.

We used the following clustering methods to create different segmentations of the customer base (methods 2, 3, and 4 are supported by Weka [8]):

1. *Random Clustering (Random)* – To create *k* new groups on sub-segment level *l*+1 from a set of customers in group *j* on sub-segment level *l*, the probability of customer C_i belonging to a new group g_j out of possible *k* new groups in sub-segment level *l*+1 is 1/k and is the same for all g_j's.
2. *SimpleKMeans (SMean)* – *k* local minimum cluster centers in the XC instance space are chosen via a random start iterative approximation strategy. Completely different clusters centers can be returned due to the initial random cluster center selections [8].
3. *FarthestFirst (FFirst)* [9] - A greedy k-center algorithm that is guaranteed to produce clustering results within a constant factor of two of the optimum.
4. *Expectation Maximization (EM)* – An iterative approach to approximate the cluster probabilities and distribution parameters that converges to a local maximum.

In this paper, we want to study how predictive models of customer behavior vary across the six dimensions of different types of datasets, customers, predictive models, dependent variables, performance measures, and segmentation techniques. Before we get into the details of our experiments, we first explain how the problem explored in this paper is related to the previous work on segmentation and personalization.

3. Related Work

The problem of building individual and segmented models of customer behavior is related to the work on (a) user modeling and customer profiling in data mining, (b) customer segmentation in marketing, and (c) building local vs. global models in statistics. We examine the relationship of our work to these three areas of research in this section.

There has been much work done in data mining on modeling customer behavior and building customer profiles. Customer profiles can be built in terms of simple factual information represented as a vector or as a set of attributes. For example, in [10], a user profile is defined as a vector of weights for a set of certain keywords. Customer profiles can be defined not only as sets of attributes but also as

- **Sets of rules** defining behavior of the customer. For example, we may store the rule "John Doe prefers to see action movies on weekends" (i.e., *Name="John Doe" & MovieType="action"* → *TimeOfWeek="weekend"*) as a part of John Doe's profile. [11] describes a method for generating and validating such rule-based profiles.
- **Sets of sequences**, such as sequences of Web browsing activities or movie watching sequences. For example, we may store in Jim's profile his popular Web browsing sequences, such as "when Jim visits the book Web site XYZ, he usually first accesses the home page, then goes to the Home&Gardening section of the site, then browses the Gardening section and then leaves the Web site" (i.e., *XYZ: StartPage* → *Home&Gardening* → *Gardening* → *Exit*). Such sequences can be learned from the transactional histories of consumers using frequent episodes and other sequence-learning methods [12] and have been extensively used in the web usage mining literature [13-15].
- **Signatures**, i.e., the data structures that are used to capture the evolving behavior learned from large data streams of simple transactions [16].

There has also been some work done on modeling personalized customer behavior by building appropriate probabilistic models of customers. For example, [17] builds customer profiles using finite mixture models and [18] use maximum entropy and Markov mixture models for generating probabilistic models of customer behavior. However, all these approaches focus on the task of building good profiles and models of customers and do not study the performance of individual vs. segmented and vs. aggregate models of customer behavior.

Comparison of segmentation vs. aggregate models of customer behavior has also been done by marketing researchers who demonstrated that segmented models of customer behavior exhibit better performance characteristics than aggregate models [2]. However, this work has not been extended to the 1-to-1 case and no comparison has been made between aggregate and individual, and between individual and segmented models.

Our work is also related to the work on clustering that partitions the customer base and their transactional histories into homogeneous clusters for the purpose of building better models of customer behavior using these clusters [19]. In our work, we use various clustering method for the very same purpose. However, we go beyond this partitioning and compare performance of aggregated vs. segmented and vs. individual models of customer behavior.

Finally, our work is related to the problem of building local vs. global models in data mining and statistics [12, 20, 21]. Rather than building one global aggregated model of customer behavior, it is often better to build several local models that would produce better performance results. Furthermore, this method can be carried to the extreme when a local model is built for *each* customer, resulting in 1-to-1 customer modeling. In this paper, we pursue this approach and compare the performance of aggregate, segmented and individual models of customer transactions.

4. Comparing Individual vs. Aggregate Levels of Customer Modeling

In this section, we compare individual vs. aggregate levels of customer modeling. More specifically, we compare predictive accuracy of function (1) estimated from the transactional data *TRANS(C_i)* for all the *individual* customer models and compare its performance with the performance of function (1) estimated from the transactional data for the *whole* customer base. In particular, we explore the aforementioned tradeoff between the heterogeneity of customer base and the sparsity of data.

4.1 Experimental Setup

As a first step, we discretized Nielson and ComScore data to improve classification speed and performance [22]. Transactional attributes, such as product categories, were discretized to roughly equal representation in sample data to avoid overly optimistic classification due to highly skewed class priors. We also discretized continuous valued attributes such as price and Internet browsing durations based on entropy measures via our implementation of Fayyad's [23] recursive minimal entropy partitioning algorithm.

To determine whether individual modeling performs statistically better than aggregate level modeling, we use a variant of the none parametric Mann-Whitney rank test [24] to test whether the accuracy score of the one aggregate model is statistically different from a random variable with a distribution generated from individual accuracy results of the individual level models. The null hypothesis for each of the performance measures of CCI, RME, RAE is then:

(I) H_0: The aggregate level performance measure *is not* different from the set of individual level performance measures.

H_1+: The aggregate level performance measure *is* different from the set of individual level performance measures in the *positive* direction.

H_1-: The aggregate level performance measure *is* different from the set of individual level performance measures in the *negative* direction.

To illustrate what we have done, consider the following example.

Example: For the 156 Nielson low-volume customers, we generate a NaiveBayes model on all low-volume customers' demographic and purchase data via ten-fold cross validation. Model α is used to predict the day of the week a purchase transaction is likely to occur for a customer C_i given his/her demographic information and other transactional data, such as store location and primary shopper's gender, from that particular purchase trip. Ten-fold cross validation during the model generation give us three performance measures of α: CCI_α, RME_α, and RAE_α.

To compare α's performance against individual level models, we generate 156 separate NaiveBayes models for each of the 156 low-volume families. Let this set of models be model set β, where each model i predicts the day of week a customer C_i would conduct his/her shopping trip. From each model i in model set β, we also have three performance measures: CCI_i, RME_i, and RAE_i.

Let CCI_β, RME_β, and RAE_β be 3 random variables with distributions CCI_i, RME_i, and RAE_i respectively for all model i in model set β. Then to test for H_0 (I) along the performance measure CCI, we would compare CCI_α against CCI_β and determine whether CCI_α is statistically different from CCI_β using a variant of the Mann-Whitney rank test mentioned earlier.

The above scenario is repeated for all customer type datasets listed in Table I, across 8 ComScore transactional variables and 5 Nielson transactional variables listed in Section 2, and three different classifiers.

4.2 Results

Table 2 lists the number of statistical tests that rejects the null hypothesis (I) at 95% significance level for all ComScore and Nielson customer type datasets (we note that by counting the number of statistically significant distribution tests on results generated from 10-fold cross validations is not a case of the pathological multiple comparison procedure[25]).

From Table 2, we can draw the following conclusions:

- None of the statistical tests accepts H_1+, which means that the performance measures at the aggregate level is never greater than that of the individual level.
- The number of significant results drops as we move from the high-volume customer dataset to low-volume dataset.
- ComScore data, which has ten times more families in each customer type dataset, has the highest number of significant results in the high-volume dataset and the greatest discrepancies between the high and low volume datasets.

Table 2. Aggregate vs. Individual Level Customer Models for Hypothesis Test (I)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	72	0	29
ComScore	Low	72	0	4
Nielson	High	45	0	5
Nielson	Low	45	0	2
Nielson	All	45	0	5

* numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0

Performances of classifiers vary, but the overall trend is clearly visible in Table 2: *for high-volume customers, modeling customer behavior at the individual level will yield significantly better results than the aggregate case*. In fact, modeling low-volume customers at the individual level will not be worse off than the aggregate level approach.

5. Comparing Individual vs. Segmentation vs. Aggregate Levels of Customer Modeling

In this section, we compare individual vs. segmentation and aggregate vs. segmentation levels of customer modeling. More specifically, we compare predictive accuracy of function (1) estimated from the transactional data *TRANS(C_i)* for the segmentation level models, and compare its performance with the performance results obtained in Section 4.

As explained in Section 2, we generate progressively finer customer sub-segment levels using different hierarchical clustering techniques. Moreover, this hierarchical clustering generates 5 levels of sub-segments where the number of customer groups within each sub-segment level is determined by a branching factor *k*.

As was also explained in Section 2, the factors that influence the prediction accuracies of different sub-segment levels include the quality of segmentation, the levels of refinements, data sparsity, and customer heterogeneity.

5.1 Segmenting Customer Base Using Clustering Methods

Once we determined the new groupings of our customers within each of the 5 sub-segment levels, we generate predictive models for each of groups as described in Section 2.

To compare the clustering algorithms against aggregate and individual level models, we first compute the *best performing segmentation level* for a clustering algorithm as follows:

Best Segment Level = $\arg\max\left(\overline{CCI_l} - \overline{RME_l} - \overline{RAE_l}\right)$,

where $l = 1\ldots 5$ levels, and $\overline{CCI_l}, \overline{RME_l}, \overline{RAE_l}$ are the average CCI, RME, and RAE for all the groups at level *l* as defined in Section 2. We took the difference between these performance measures for the reasons explained in Section 2.

Then we compare aggregate model to the best segment level in the same manner as for the aggregate versus individual model comparison in Section 4. Thus, the null hypothesis for comparing best clustering level for each clustering algorithm against the aggregate model is:

(II) H_0: The aggregate level performance measure *is not* different from the set of best segment level performance measures.

H_1+: The aggregate level performance measure *is* different from the set of best segment level performance measures in the *positive* direction.

H_1-: The aggregate level performance measure *is* different from the set of best segment level performance measures in the *negative* direction.

To compare best segment level against individual models, we again use the Mann-Whitney rank sum test as our statistical comparator [24] because of the none normal distribution of performance measures and different sample sizes across segment levels. The null hypothesis for comparing best segment level for each clustering algorithm against individual level models then becomes:

(III) H0: The distribution of individual model performance measure *is not* different from that of the best segment level model.

H1+: The distribution of individual model performance measure *is* different from that of the best segment level model in the positive direction.

H1-: The distribution of individual model performance measure *is* different from that of the best segment level model in the negative direction.

To compare clustering algorithms against Random clustering, we compute the average performance measures from all models in each level *l,* and then compare the distribution of mean performance measures of each clustering algorithm against Random clustering across all 5 levels. Similar to the above null test (III), the null hypothesis for comparing clustering algorithm performance versus Random is then:

(IV) H0: The distribution of mean performance measure across all levels of Random clustering *is not* different from that of the clustering algorithm.

H1+: The distribution of mean performance measure across all levels of Random clustering *is* different from that of the clustering algorithm in the positive direction.

H1-: The distribution of mean performance measure across all levels of Random clustering *is* different from that of the clustering algorithm in the negative direction.

5.2 Results

Table 3 lists the number of statistical tests that rejects the null hypothesis (II) at 95% significance level for all ComScore and Nielson customer type datasets. Similar to our analysis for aggregate versus individual level models, there are 75 statistical comparisons for each ComScore data clustering scheme, which gives us a total of 288 comparisons aggregated across all 4 clustering algorithms. Likewise, the 45 statistical comparisons for each Nielson data clustering scheme gives us 180 comparisons aggregated across all 4 clustering algorithms.

Table 3. Aggregate vs. Best Segment Level Models Hypothesis Test (II)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	288	1	171
ComScore	Low	288	14	97
Nielson	High	180	14	55
Nielson	Low	180	25	65
Nielson	All	180	8	50

* numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0

From Table 3, we can draw the following conclusions:

- Best Segment Level significantly dominates aggregate level models across all customer types.
- There is a significant number of instances where the aggregate level models performed better than best segment level models. We will see in the clustering performance analysis that this occurred because of some of the clustering algorithms resulted in poor performance.

Table 4 lists the number of statistical tests that rejects the null hypothesis (III) at 95% significance level for all ComScore and Nielson customer type datasets.

Table 4. Individual vs. Best Segment Level Models for Hypothesis Test (III)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	288	141	27
ComScore	Low	288	66	72
Nielson	High	180	45	8
Nielson	Low	180	36	6
Nielson	All	180	71	20

* numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0

From Table 4, we can draw the following conclusions:

- Individual level models significantly dominate best segment level models for high-volume customers.
- For low-volume customers, especially for ComScore bottom 5% dataset, we see that best segment level models performed better in more instances than individual level models. Similar to observations made on Table 2, we will see that in the clustering analysis section, the best segment level models in the best performing clustering algorithms significantly dominate individual level models.

Table 5 lists the number of statistical tests that rejects the null hypothesis (IV) at 95% significance level for all ComScore and Nielson customer type datasets.

Table 5. Random vs. Other Clustering algorithms (CA) for Hypothesis Test (IV)

DataSet	Type	CA	Tests	H_1+	H_1-
ComScore	High	SMean	72	2	4
ComScore	High	FFirst	72	8	29
ComScore	High	EM	72	3	24
ComScore	Low	SMean	72	3	0
ComScore	Low	FFirst	72	0	57
ComScore	Low	EM	72	1	7
Nielson	High	SMean	45	2	0
Nielson	High	FFirst	45	12	9
Nielson	High	EM	45	17	10
Nielson	Low	SMean	45	9	4
Nielson	Low	FFirst	45	15	6
Nielson	Low	EM	45	21	5
Nielson	All	SMean	45	7	2
Nielson	All	FFirst	45	10	12
Nielson	All	EM	45	8	2

* numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0

From Table 5, we can draw the following conclusions:

- FarthestFirst (FFirst) clustering algorithm performs the best out of all four clustering schemes across all customer types.
- Expectation Maximization (EM) performs well in high-volume customer datasets and poorly in low-volume customer datasets.
- SimpleKMeans (SMean) produced roughly the same level of performance as that of Random clustering because the small number of significant counts in columns H_1+ and H_1-

Table 5 clearly shows that there are significant differences in performance among the three non-random clustering algorithms. To truly test hypotheses II and III, we ought to take the best segment-level results from the best clustering algorithm. Table 6 lists the number of statistical tests that rejects the null hypothesis (II) at 95% significance level for all ComScore and Nielson customer type datasets for FarthestFirst clustering scheme *only*.

The results of Table 6 confirm our expectation from earlier findings that for "well-behaved" clustering algorithms (clustering that performed significantly better than Random clustering), such as FarthestFirst, best segment level model performs significantly better than aggregate model across all customer types.

Table 6. Aggregate vs. Best FFirst Segment Level Models for Hypothesis Test (II)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	72	1	45
ComScore	Low	72	0	17
Nielson	High	45	3	12
Nielson	Low	45	9	18
Nielson	All	45	4	9

* numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0

Table 7 lists the number of statistical tests that rejects the null hypothesis (III) at 95% significance level for all ComScore and Nielson customer type datasets for FarthestFirst clustering scheme only.

Table 7. Individual vs. Best FFirst Segment Level Models for Hypothesis Test (III)

DataSet	Customer Type	Tests	H_1+	H_1-
ComScore	High	72	22	9
ComScore	Low	72	1	37
Nielson	High	45	6	1
Nielson	Low	45	8	1
Nielson	All	45	18	8

* numbers in columns H_1+ and H_1- indicate the number of statistical tests that reject hypothesis H_0

From Table 7, we see a clear reversal of relative performance between individual and best segment level models amongst the ComScore datasets. While individual level outperforms best segment level for the *high-volume* customers, best segment level clearly dominates for the *low-volume* customers. As mentioned in Section 4, there is general tradeoff between customer heterogeneity and data sparsity when building customer segmentation models. The results of Table 7 clearly show that for good clustering methods, *aggregation of idiosyncratic customers with insufficient data outperforms individual level models*.

5.3 Performance Curves

As we have shown in Section 4, the performance at the individual level outperforms the aggregate level of customer behavior analysis. In Section 5.2, we showed that the best segment level models would outperform both aggregate and individual level models and that different clustering algorithm can give us significantly different patterns of results. But if we draw the *performance curve* that plots a performance measure across different segmentation levels, an important question is how would the *shapes* of the curves change under different conditions. For instance, would the performance grow monotonically when the customer segments are refined or would the performance reach a maximum at a certain sub-segment level and then drop when segments are refined further? Alternatively, could the performance curve become concave due to "bad" clustering (i.e. clustering algorithms that yield statistically equivalent performance results as that of Random clustering)?

To gain a better understanding of the various factors that influence the performance of our models across the three levels of analysis, we plot the performance curves for the average CCI measure. Figure 1 shows the performance curve predicting primary shopper's gender for all Nielson data under Random clustering. The X-axis denotes the level of segmentation, which runs from aggregate level near the origin, through the 5 segmentation levels, to the individual family level on the far right. The Y-axis specifies the $\overline{CCI}$ measure. The three different curves represent the performance of three classifiers across all levels of predictive models.

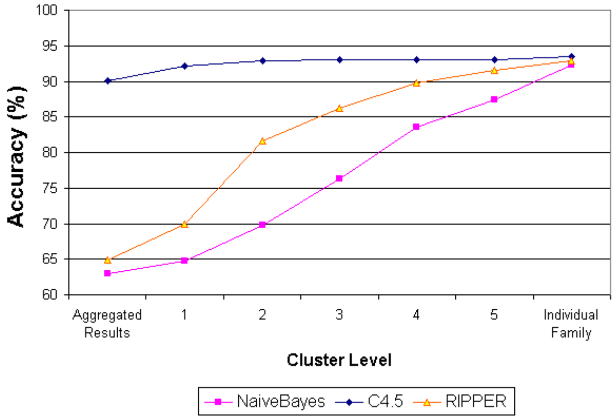

Figure 1. Nielson All Data Random Clustering on Primary Shopper's Gender

We plotted such performance curves for all 5 types of customer datasets, across 4 clustering schemes and 13 transactional dependent attributes. From the 260 performance curves of $\overline{CCI}$, we observed three dominating patterns. For *high-volume* customers and "well-behaved" clustering algorithms, we see a monotonically increasing curve as represented by Figure 1. This occurs primarily for high-volume customer datasets because with sufficient data, we can build models of idiosyncratic customer behavior all the way to the individual level without running into the problem of data sparsity.

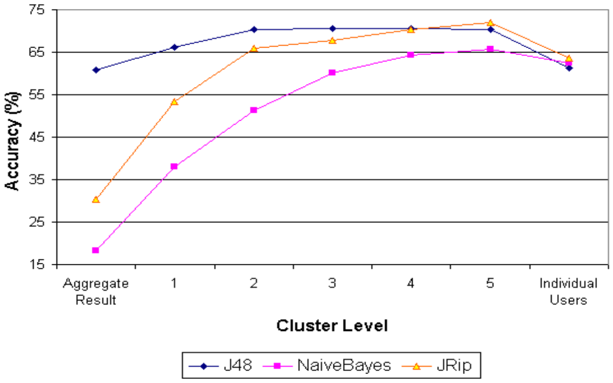

Figure 2. ComScore Low-Volume Customer, FFirst Clustering on Day of the Week

Figure 2 shows the second general pattern, that of convex performance curves. This is observed for *low-volume* customer datasets and "well-behaved" clustering algorithms. Our discussion of FarthestFirst clustering algorithm for low-volume ComScore customers fits well into this category. This pattern shows that for low-volume customers, we will eventually run into the problem of data sparsity while trying to build progressively finer models of customer behavior.

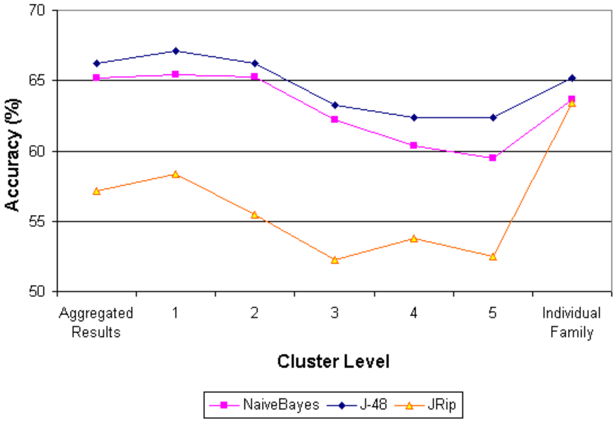

Figure 3. Nielson Low-Volume Customer, EM Clustering on Category Count

Figure 3 shows the third general pattern, that of concave performance curves. This pattern is observed mainly for low-volume customer datasets and "badly-behaved" clustering algorithms. This "concave" pattern occurs because heterogeneous customers are grouped into same segments by "badly-behaved" clustering algorithms.

As follows from these discussions, there are certain factors driving the overall shapes of the performance curves and we explained these factors for these curves. While we experimented with 2 marketing datasets, we believe our findings are application independent, and that our results provide good insights into performance analysis of various segmentation techniques.

6. Conclusions

We conducted an extensive comparative study of aggregate, segmentation, and individual level modeling across multiple dimensions of analysis such as different types of datasets, customers, predictive models, dependent variables, performance measures, and segmentation techniques. We identified four factors that significantly influence the prediction outcomes of customer behavior models: customer heterogeneity, data sparsity, quality of segmentation techniques, and levels of segmentation.

Our results show that, given sufficient transactional data, 1-to-1 modeling significantly outperforms other types of models of customer behavior. However, when modeling customers with very little transaction data, segmentation dominates individual modeling for the best segmentation techniques and the best level (granularity) of segmentation. What is surprising, however, is that 1-to-1 modeling is never worse than aggregate level modeling in our experiments, even in the case of sparse data. We also showed that poor segmentation techniques could lead to poor performance results that are comparable to the random segmentation method.

We performed further analysis of the four influencing factors by plotting *performance curves* across all levels of customer segmentation and observed three dominating patterns presented in Figures 1 – 3. The first monotone pattern presented in Figure 1 occurs for high-volume customers and "well-behaved" clustering algorithms, and shows that we can build models of idiosyncratic customer behavior all the way to the individual level without running into the data sparsity problem. The second convex pattern presented in Figure 2 occurs for low-volume customers and "well-behaved" clustering algorithms, and shows that we will eventually run into the problem of data sparsity while trying to build progressively finer models of customer behavior. The last concave pattern presented in Figure 3 occurs primarily for low-volume customers and "poorly-behaved" clustering algorithms (i.e. clustering algorithms that yield statistically equivalent performance results as that of Random clustering). It occurs because heterogeneous customers are grouped into same segments by "poorly-behaved" clustering algorithms.

In the future, we would like to study the problem of predicting customer behaviors via different levels of segmentation under a more general class of experimental settings. We would also like to gain a better understanding on the nature of the tradeoff between customer heterogeneity and data sparsity at a more theoretical level.

7. References

[1]. Kotler, P., *Marketing Management*. 2003: Prentence Hall.

[2]. Allenby, G.M. and P.E. Rossi, *Marketing Models of Consumer Heterogeneity.* J. of Econometrics, 1999. **89**.

[3]. Peppers, D. and M. Rogers, *The One-to-One Future*. 1993, New York, NY: Doubleday.

[4]. *Comm. of ACM*, in *Spec. Issue on Personalization*. 2000.

[5]. Quinlan, R., *C4.5: Programs for Machine Learning*. 1993: Morgan Kaufmann.

[6]. John, G.H. and P. Langley. *Estimating Continuous Distributions in Bayesian Classifiers*. in *11th UAI Conf.* 1995.

[7]. Cohen, W.W. *Fast Effective Rule Induction*. in *12th Int. Conf. on Machine Learning*. 1995.

[8]. Witten, I.H. and E. Frank, *Data Mining: practical machine learning tools and techniques with Java implementations*. The Morgan Kaufmann Series in Data Management System, ed. J. Gray. 2000, San Francisco: Morgan Kaufmann. 147-150.

[9]. Hochbaum, S.D. and B.D. Shmoys, *A Best Possible Heuristic for the K-Center Problem.* Mathematics of Operational Research, 1985. **10**(2): p. 180-184.

[10]. Pazzani, M. and D. Billsus, *Learning and revising user profiles: The identification of interesting web sites.* Machine Learning, 1997. **27**(313-331).

[11]. Adomavicius, G. and A. Tuzhilin, *Expert-driven validation of rule-based user models in personalization applications.* Data Mining and Knowledge Discovery, 2001. **5**((1/2)): p. 33-58.

[12]. Hand, D., H. Mannila, and P. Smyth, *Principles of Data Mining*. 2001: MIT Press. Sections 6.3.2-6.3.3.

[13]. Mobasher, B., et al. *Discovery of Aggregate Usage Profiles for Web Personalization*. in *Web Mining for E-Commerce Workshop (WebKDD'00)*. 2000.

[14]. Mobasher, B., et al. *Using Sequential and Non-Sequential Patterns for Predictive Web Usage Mining Tasks*. in *IEEE ICDM Conf.* 2002.

[15]. Spiliopoulou, M., et al., *A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis.* INFORMS Journal of Computing, 2003. **15**(2).

[16]. Cortes, C., et al. *Hancock: a language for extracting signatures from data streams*. in *6th ACM SIGKDD Conf.* 2000.

[17]. Cadez, I.V., P. Smyth, and H. Mannila, *Predictive profiles for transaction data using finite mixture models*. 2001, University of California, Irvine.

[18]. Manavoglu, E., D. Pavlov, and C.L. Giles. *Probabilistic User Behavior Models*. in *IEEE ICDM Conf.* 2003.

[19]. Yang, Y. and B. Padmanabhan. *Segmenting Customer Transactions Using a Pattern-Based Clustering Approach*. in *IEEE ICDM Conf.* 2003.

[20]. Atkeson, C.G., A.W. Moore, and S. Schaal, *Locally Weighted Learning.* Artificial Intelligence Review, 1997. **11**.

[21]. Fan, J. and R. Li, *Local Modeling: Density Estimation and Nonparametric Regression*, in *Advanced Medical Statistics*, J. Fang and Y. Lu, Editors. 2003, World Scientific. p. 885-930.

[22]. Dougherty, J., R. Kohavi, and M. Sahami. *Supervised and Unsupervised Discretization of Continuous Features*. in *12th Int. Conf. on Machine Learning*. 1995.

[23]. Fayyad, U.M. and K.B. Irani. *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning*. in *13th IJCAI Conf.* 1993.

[24]. Mendenhall, W. and R.J. Beaver, *Introduction to probability and statistics*. 9th ed. 1994, Pacific Grove, California: International Thomson Publishing. 591-598.

[25]. Jensen, D.D. and P.R. Cohen, *Multiple Comparisons in Induction Algorithms.* Machine Learning, 2000. **38**.