Classification of Proxy Labeled Examples for Marketing Segment Generation

Dean Cerrato
Akamai Technologies
8 Cambridge Center
Cambridge, MA 02142
617-444-3000
dcerrato@akamai.com

Rosie Jones
Akamai Technologies
8 Cambridge Center
Cambridge, MA 02142
617-444-3000
rejones@akamai.com

Avinash Gupta
Akamai Technologies
8 Cambridge Center
Cambridge, MA 02142
617-444-3000
avgupta@akamai.com

ABSTRACT
Marketers often rely on a set of *descriptive segments*, or qualitative subsets of the population, to specify the audiences of targeted advertising campaigns. For example, the descriptive segment "Empty Nesters" might describe a desirable target audience for extended vacation package offers.

While some segments may be easily described and generated using demographic data as ground truth, others such as "Soccer Moms" or "Urban Hipsters" reflect a combination of demographic and behavioral attributes. Ideally, these attributes would be available as the basis for ground truth labeling of a classifier training set or even direct member selection from the population. Unfortunately, ground truth attributes are often scarce or unavailable, in which case a proxy labeling scheme is needed.

We devise a method for labeling a population according to criteria based on a postulated set of shopping behaviors specific to a descriptive segment. We then perform supervised binary classification on this labeled dataset in order to discover additional identifying patterns of behavior typical of labeled positives in the population. Finally, the resulting classifier is used to perform selection from the population into the segment, extending reach to cookies who may not have exhibited the postulated behaviors but likely belong in the segment.

We validate our approach by comparing a descriptive segment trained on ground truth to one trained on behavioral attributes only. We show that our behavior-based approach produces classifiers having performance comparable to that of a classifier trained on the ground truth data.

Categories and Subject Descriptors
I.5.2 [**Pattern Recognition**]: Design Methodology – *Classifier design and evaluation.*

General Terms
Algorithms, Experimentation.

Keywords
Marketing segments, classification, computational advertising, rules, RIPPER.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD'11, August 21–24, 2011, San Diego, California, USA.
Copyright 2011 ACM 978-1-4503-0813-7/11/08...$10.00.

1. INTRODUCTION
We describe our efforts to design a supervised classification algorithm that determines membership for a group of descriptive marketing segments in the absence of ground truth demographic data.

1.1 Motivation
Descriptive marketing segments are a familiar way to specify targeted audiences for on-line advertising campaigns; for example, a purveyor of fast food might find it desirable to display coupon ads to young people of a certain age who attend college, i.e., members of the College Students descriptive segment.

Ideally, comprehensive demographic data would be available from which to deduce segment membership for members of the population, but often this is not the case. In fact, some descriptive segments have a significant subjective component (e.g., Trend-setters) that cannot be inferred directly from hard quantitative data.

In the absence of ground truth, it becomes necessary to define a proxy definition for segment membership derived from available data, which in an on-line context is presumed to be related to web surfing behavior. Once such a definition is created, it can be used to proxy label the population.

A naïve proxy labeling scheme would be to restrict segment membership to visitors to certain websites or pages; indeed, this is often the approach taken by marketing operations.

Our work refines and extends this approach, recasting it as a supervised classification problem that yields a variety of solutions having a range of performance characteristics relating to descriptive segment population size and targeting precision.

1.2 Business Context
The Advertising Decision Systems (ADS) group at Akamai maintains a confidential cooperative network of advertising partners for whom we track visitor traffic and target on-line advertisements. In particular, our Descriptive Segments product offering enables partners to target ads to specific types of on-line shopper [2].

Users are identified using anonymous web browser cookies. For the purpose of identifying descriptive segment members, we have access to web shopping data collected at both the website and department levels. For the purpose of this study, we also use a set of limited demographic data. Each user is represented by an anonymous identifier, and no attempt is made to connect any identifier with a real-world person. Browsing history for each cookie comprised a set of every distinct department-level web location they had visited, from our confidential cooperative

network of shopping sites. All data was collected in accordance with Akamai's privacy policy [4] and the privacy policies of the sites visited by the cookies.

Marketing intuition affords us a starting point in identifying websites and departments that are relevant to the various descriptive segments, particularly the murkier ones.

A significant hard constraint on our solution arises from the targeting system: segment membership is bestowed via disjunctive normal form (DNF) rulesets, in terms of antecedents of the form "this cookie visited Site A" or "this cookie visited Department B" or their negations.

Other, softer constraints on the ruleset solutions are: the number of rules per ruleset should not be too large; and as few as possible rules should fire for any given cookie (this would adversely affect the ability of operations to optimize the system on-the-fly).

Finally, the ability to "tune" the *reach* (number of segment members) and targeting precision of the rulesets is highly desirable.

1.3 Related Work

Market segmentation refers to the differentiation of sub-types of customers who would prefer different kinds of products. Most applications of machine learning to market segmentation have focused on predicting purchase behavior. In our work, we are using as our target psycho-demographic descriptions of target audiences. Thus in most cases we do not have labeled instances.

Machine learning has been successfully applied to learning customer segmentation when the target is to identify responders. One piece of previous work focused on using SVMs to learn user preferences in the taste of beef [7] – here a kernel-based similarity function based on clustering finds groups of users with similar tastes. Other work has combined machine learning and expert opinions to perform market segmentation of performing arts audiences [1]. Florez-Lopez et al [8] showed that decision trees out-performed statistical approaches on their task of predicting caravan insurance purchasers. This approach [10] of automatically finding users with similar behavior is referred to in marketing as "look-alike" modeling. Büchner and Mulvenna [5] combine marketing domain expertise with on-line data sources to discover user patterns. In [11], Perlich and Huange analyze relational database structure to discover relationships relevant in CRM contexts.

1.4 Roadmap

Consistent with the requirements for papers submitted under the Industry Track: Emerging area, our work represents the development of a KDD-based prototype solution designed to address a specific industry problem subject to a special set of constraints arising from data availability and legacy system limitations.

This Section has introduced the problem in the particular context of our business operations and constraints. Section 2 presents a more thorough explication of the KDD-based binary classification approach using proxy labeled examples. In Section 3, we define our modeling and evaluation methodologies. In Section 4, we display some resulting rulesets and present performance metrics, which are discussed in Section 5. Sections 6 and 7 offer thoughts for future work and a recap of our conclusions.

2. SOLUTION APPROACH

The primary challenge is to define a subjective yet reasonable proxy labeling scheme to identify positive examples for segment membership. From there, the supervised classification problem may be solved and the solutions evaluated.

2.1 Proxy Labeling Scheme

At the heart of our definition of segment membership is the *indicator set*, or the set of websites (*indicator sites*) and departments (*indicator departments*) deemed relevant to the descriptive segment. For example, a Luxury segment member might be relatively likely to visit a *haute couture* website or the fine jewelry page of a department store website. The indicator set should span the shopping domain of the typical segment member.

Candidates for inclusion in the indicator set are initially drawn from marketing intuition but may extend to sites and departments found during, say, searches for relevant substrings in site and department titles.

Cookies with a history of activity in the indicator set are candidates for segment membership. A minimum *threshold* is defined that can be set in order to vary the strictness of the labeling criteria, i.e., cookies who have visited at least *threshold* distinct indicator sites (including parent sites of indicator departments) receive a positive label. In general, one would expect that higher thresholds result in a smaller but more precise set of positive labels.

The indicator set criterion can be supplemented by demographic criteria if such data are available. For example, an augmented labeling scheme might select as positives those cookies who meet the indicator set criteria *and* fall within a certain desired demographic.

2.2 Ruleset Generation

A supervised binary classification algorithm is needed to create a rule-based model as described. Initially, we considered an association rule approach for itemset discovery and subsequent mapping to descriptive segments; in fact, the Apriori algorithm [2] was run on some preliminary trial data, but found to be inappropriate to the problem, since a) we're defining the segments up front and wish to model them explicitly, and b) Apriori seeks to discover all itemsets satisfying a set of minimum coverage and accuracy requirements, which turns out to be rather memory intensive past a certain modest data set size (~10,000 examples).

The definitive algorithm for this task is the RIPPER algorithm [6]. RIPPER is a supervised binary classification algorithm that generates DNF rulesets. During training, RIPPER iteratively grows and prunes a set of rules in order to maximize coverage of the minority class on a *grow* fold of the training data while minimizing classification error on a *prune* fold; an optimization phase then replaces and revises underperforming rules. RIPPER has been shown to be robust and fairly efficient. Our work utilizes JRip, the Weka [9] implementation of RIPPER.

2.3 Performance Metrics

In addition to good prediction performance on the test dataset, the rulesets generated should satisfy a set of business requirements.

2.3.1 Prediction Performance

The typical metrics derived from the confusion matrix serve as our benchmarks. We mainly consider true-positive-rate (TPR) and false-positive-rate (FPR) on the test data:

$$TPR = \frac{\#\ \text{positives selected (true positives)}}{\#\ \text{positives}}$$

$$FPR = \frac{\#\ \text{negatives selected (false positives)}}{\#\ \text{negatives}}$$

TPR should significantly exceed FPR, an indication of preferential selection of actual positive examples over negatives:

$$require\ \ TPR >> FPR$$

An (FPR, TPR) duple maps to a point on an ROC curve. Thus, when TPR >> FPR, the model identifies positives better than random selection. (If we were able to vary the strictness of selection by a ruleset, we could sweep out an entire ROC curve; unfortunately for rulesets, selection is boolean and this is not the case.)

Over the population as a whole, the marketing department will find targeting precision and reach to be of most interest, where

$$precision = \frac{\#\ true\ positives}{\#\ selected}$$

serves as the "accuracy" measurement, and reach is simply the number of cookies selected for membership in the descriptive segment.

We find it helpful to compute the *normalized* precision, or *gain*, relative to the baseline precision of the random selection method, which simply equals the frequency of the positive class in the population, or *response rate*:

$$gain = \frac{precision}{response\ rate}$$

2.3.2 Precision vs. Reach Trade-off

Depending on the descriptive advertising campaign, the focus might be on precise targeting or extensive reach or some of both. In general, the two are mutually exclusive: to increase reach, you must select less suitable cookies, which decreases precision.

Therefore, it is important to provide a way to tune the precision/reach trade-off. Unfortunately, rulesets perform a binary selection; thus, there is no segment membership score that can be thresholded to vary the selection. This is addressed by generating multiple rulesets with varying precision and reach statistics; it is then up to the marketing department which ruleset they shall use.

2.3.3 Rule Overlap

Because of legacy constraints on the way descriptive segments are operationalized, it is important that the number of rules that fire for a typical cookie not exceed one or two (preferably one); multiple rules firing means multiple campaign treatments for that cookie. By matching one cookie with multiple rules, there can be unintended negative consequences; e.g., the daily limit on ad impressions for that cookie may be exceeded, causing them to saturate on ad creative and thus become unresponsive; and decreased reach as ad impressions intended for distinct viewers are exhausted on repeat impressions to the same people.

We have no direct control over rule overlap, although RIPPER rulesets generally begin with large-coverage rules followed by additional, smaller-coverage rules that fill in the gaps. Therefore, we check overlap of all rulesets to verify that this requirement is satisfied.

2.3.4 Intuitive Appeal

Just as the indicator set is based on subjective descriptive criteria, so must every ruleset pass intuitive muster by discovering behaviors typical of the (intended) members of that segment, and hopefully avoiding any that are contraindicated.

3. METHODOLOGY

3.1 Data Sources

A data set of 20 million browser cookies and their browsing history was drawn randomly from the (much larger) overall population. The bulk of the modeling effort focused on this data.

In addition, an independent, anonymous subset of roughly 280,000 cookies and their browsing history was associated with a small set of demographic features (procured from an external source and joined to our data via an automated cookie-matching process). Separate modeling runs involving demographically-enhanced positive definitions were thus possible.

Browsing history was redacted for all sites and departments opting to exclude their data from analysis.

3.2 Definition of Positives

Five descriptive segments of interest were modeled, each of a different type: Brides-To-Be, Moms-With-Kids, Luxury, Trend-setters, and College Students.

3.2.1 Indicator Set Selection

For each descriptive segment, an indicator set was selected based initially on marketing expertise and then extended and filtered based on various methods.

For each of the marketing indicator sites, cosine similarity was computed for every other site in the network (based on co-occurrence of visitors), with the most similar sites also considered for indicator set membership based on visitor volume and relevance to the descriptive segment.

Sets of terms related to the descriptive segment concept were enumerated (e.g., "kids", "wedding", "diamond") and used to perform substring matching on the names of sites and departments; matching sites and departments were added to the indicator set depending on visitor volume.

Finally, sites and departments whose data was excluded from analysis were removed from the indicator sets.

3.2.2 Threshold Variation

Positives were defined for various threshold values, creating a range of strictness criteria for labeling. Threshold values typically ran from 1 through 3 or 4 depending on the descriptive segment.

3.2.3 Enhancement Using Demographic Data

Demographic age range data was used to enhance the positive definition for the College Students descriptive segment. An additional criterion – Age between 18 and 24 – was added to the indicator site criterion. The smaller demographic dataset was thus labeled for generation of College Students rulesets.

A purely demographic Moms-With-Kids positive definition was constructed from Age, Gender, and Child-In-Household features. This labeling was applied to the demographic dataset for generation of Moms-With-Kids rulesets. Positives met the criteria of Age between 20 and 49 + Gender = Female + Child-In-Household = True; negatives were those for whom feature values were available but who failed one or more criteria; the remaining cookies were ignored for training purposes.

3.3 Modeling Procedure

3.3.1 Down-Sampling

After proxy labeling the datasets, the data was down-sampled in order to reduce total dataset size and to bring the number of negatives closer to the number of positives.

It was observed that large variations in down-sampling ratios (neg:pos) can affect ruleset size, as larger ratios afford more opportunity to prune rules, resulting in smaller, more precise rulesets.

Higher down-sampling ratios were generally allowed for higher (i.e., stricter) thresholds where the number of positives was generally low, thus magnifying the precision of the resulting rulesets.

For lower thresholds, the down-sampling ratio was automatically lowered by the algorithm in order to control the total dataset size and to increase reach.

In all cases, the number of positives was forced to be less than the number of negatives, so that the positive class remained the minority class for purposes of modeling with RIPPER. Thus, the down-sampling ratio never fell below 1.0.

3.3.2 Feature Set Preparation

The feature set consisted of ~20,000 sites and departments possibly having been visited, with the exception of those that were opted-out of analytics use.

Site and department features (and their children) appearing in the indicator set were stricken from the feature set.

Finally, the feature set was reduced to the top 1000 features ranking highest in terms of InfoGain score. (This was performed by Weka.)

3.3.3 RIPPER Modeling in Weka

The down-sampled dataset was passed to Weka for JRip ruleset generation. All parameters were left at their default values, except for "Tweights" (the minimum allowed coverage for a rule) which was varied automatically by the outer modeling loop until a ruleset having between 7-10 rules was found.

Weka applied a stratified 2/1 train/test split to the data and returned train and test performance statistics.

3.3.4 Generation of Multiple Rulesets

Multiple rulesets were generated for the current descriptive segment with the goal of providing a range of precision and reach options. An outer loop varied the pre-specified thresholds, whereas down-sampling ratios were varied automatically (as described previously). In a few trials, the pre-defined indicator sets were iteratively augmented with newly "discovered" sites/departments from the latest ruleset to generate the next ruleset, stopping when gain fell too low or reach stopped increasing sufficiently.

Added to each ruleset was a "Rule 0", equivalent to a positive definition based on the ruleset's indicator set with a threshold of 1. This is a necessary addition prior to production, since the rulesets as generated will never have perfect recall, yet we do not want to exclude any labeled positives from the descriptive segment.

3.4 Performance Measurement

TPR >> FPR was checked for the Weka test split of every ruleset to ensure good discrimination between positives and negatives.

The rulesets were "sniff tested" for intuitive appeal by looking for newly discovered behaviors typical of the descriptive segment and verifying that no suspicious rules had been created. Rulesets were also verified to be between 7-10 rules.

The rulesets were applied to the entire 20 million-cookie datasets; TPR and FPR were confirmed to be approximately equal to those for the Weka test split. Performance statistics were extrapolated upwards to the entire large population, and gain and reach computed for each ruleset. The set of rulesets was confirmed to provide a range of precision/reach trade-off options.

Rule overlap was computed for each ruleset, confirming that the large majority of cookies fired only one or two rules per ruleset.

4. RESULTS

4.1 Proxy Labeling Using Demographic Data

4.1.1 Pure Demographic Ground Truth

The Moms-With-Kids indicator set-based rulesets were compared with that segment's demographic-based ruleset, which can be viewed as a best-case ruleset having been trained on approximate ground truth data.

In terms of TPR and FPR on the demographic ground truth data, performance was only slightly worse than best-case for the lower-precision rulesets, i.e., those generated from threshold values of 1 and 2 (Table 1):

Table 1. Comparison of Low-precision vs. Best-Case

Ruleset	TPR	FPR
thresh = 1	47.20%	38.93%
thresh = 2	43.53%	35.60%
best-case	44.45%	30.21%

Specifically, the threshold (1,2) indicator set-based rulesets had comparable TPR but somewhat elevated FPR, suggesting that, at least for this segment, the indicator set was a reasonable proxy for ground truth as long as it was not too strictly applied.

For the threshold (3,4) rulesets, both TPR and FPR fell significantly, revealing that it was unrealistic to expect those segment members to have visited that many distinct indicator sites (Table 2):

Table 2. Comparison of High-precision vs. Best-Case

Ruleset	TPR	FPR
thresh = 4	3.34%	1.73%
thresh = 3	7.71%	5.56%
best-case	44.45%	30.21%

When applied to the indicator set-based data, the demographic ruleset showed significantly higher reach than the indicator set-based rulesets (Table 3):

Table 3. Comparison of Low-precision vs. Best-Case. Reach is normalized.

Ruleset	Reach
thresh = 2	1.00
thresh = 1	2.04
best-case	4.28

This supports the suspicion that reach may be fundamentally limited by indicator set visitor volume, whereas ground truth labeling does not suffer from this effect.

4.1.2 Demographically Enhanced Proxy Labeling

The addition of the demographic age range criterion to the College Students indicator set-based labeling scheme appears to have shifted the precision vs. reach trade-off towards higher precision (Table 4):

Table 4. Adding Age Criterion Effect on Precision/Reach. Reach is normalized.

	no Age-range	w/Age-range
Reach	1.46	1.00
Precision	48.4%	54.2%
Gain	15.8	17.6

This should be expected, since the additional demographic criterion makes for a stricter positive definition.

4.1.3 True Positive Rate vs. False Positive Rate

For all descriptive segments, we observed that TPR >> FPR for the Weka test splits for all proxy labeled rulesets, indicating a high degree of discrimination between positive and negative examples. For the Brides-To-Be segment (Table 5):

Table 5. TPR >> FPR on Weka Test Split

	TPR (Weka)	FPR (Weka)
Ruleset 1	50.30%	0.30%
Ruleset 2	57.60%	0.80%
Ruleset 3	52.30%	2.40%
Ruleset 4	72.30%	7.70%

This pattern persisted when the rulesets were applied to the full non-down-sampled set of 20 million examples (Table 6):

Table 6. TPR >> FPR on Population

	TPR (population)	FPR (population)
Ruleset 1	46.45%	0.15%
Ruleset 2	53.69%	0.32%
Ruleset 3	44.56%	0.79%
Ruleset 4	59.93%	2.63%

However, the difference in TPR and FPR was much less pronounced whenever the Moms-With-Kids ground truth-labeling was used either as the training data for the ruleset (Table 7) or as the test data for any of the rulesets (cf. Tables 1 and 2).

Table 7. Ground Truth Ruleset Applied to Indicator Set Test Data

Test Data	TPR	FPR
threshold = 4	61.27%	10.26%
threshold = 3	58.91%	10.25%
threshold = 2	52.03%	10.21%
threshold = 1	22.07%	10.07%

It would appear that the ground truth-agnostic case (where a ruleset trained on proxy labeled positives is evaluated on proxy labeled test data) is an easier problem than either a) selecting proxy labeled positives using a ruleset trained on ground truth, or b) selecting ground truth positives using either type of ruleset.

4.1.4 Precision vs. Reach

For each descriptive segment, it was possible to generate multiple rulesets exhibiting a range of precision and reach when applied to the entire population. As expected, high precision was associated with low reach and vice versa. For the Brides-To-Be segment, for example, precision and reach were varied by using different combinations of threshold and down-sampling ratio (Table 8):

Table 8. Precision vs. Reach in Brides-To-Be Rulesets

Brides-To-Be	Ruleset 1	Ruleset 2	Ruleset 3	Ruleset 4
Threshold	3	3	2	2
Down-Ratio	50	20	5	2
Reach Ratio	1.0	2.1	5.1	17.3
Precision	7.89%	4.60%	2.41%	0.99%
Gain	180.2	104.9	54.9	22.6

Reach is defined as the number of users matching the segment with a given ruleset. We use the first ruleset as the baseline, and give reach relative to it.

The higher threshold (3) results in higher precision and reduced reach. For a given threshold, the higher down-sampling ratio makes more negatives available to be used in pruning and therefore increases precision and reduces reach also. (Precision is computed for the dataset where threshold = 2 and down-sampling ratio = 2.)

For Luxury, precision and reach tend to behave similarly as threshold and down-sampling ratio are varied; for some reason, however, threshold = 4 actually becomes *less* precise (Table 9):

Table 9. Precision vs. Reach in Luxury Rulesets

Luxury	Ruleset 1	Ruleset 2	Ruleset 3	Ruleset 4
Threshold	4	3	2	1
Down-Ratio	50	50	8.5	1
Reach	1.0	0.69	4.6	21.0
Precision	43.29%	51.27%	23.33%	8.34%
Gain	23.3	27.5	12.5	4.5

Regardless of segment, reach becomes difficult to increase beyond a certain point where threshold = 1, down-ratio is small, and the indicator set is using all available relevant sites and departments.

4.1.5 Rule Overlap

For all rulesets, multiple rules tend not to fire for the vast majority of cookies. A typical distribution of #rules-fired for a ruleset shows that ~80% of cookies fire only a single rule (Figure 1):

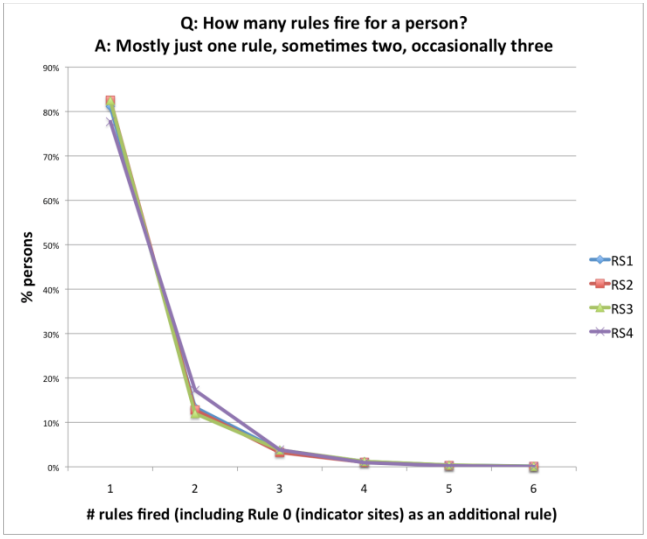

Figure 1. Rule overlap is minimal

4.1.6 Intuitive Ruleset Appeal

The degree to which the rulesets made intuitive sense varied across descriptive segments.

For the Brides-To-Be and Moms-With-Kids segments, at least some of the rules could be interpreted as behaviors expected from members of the descriptive segment. In particular, their more precise rulesets found relevant sites and departments that weren't included in the indicator set, or combinations of sites and departments that characterized a relevant behavior. This was less true for the lower precision rulesets, as more of the rules "cast a wide net" by selecting from sites that weren't strongly relevant to the segment, but weren't irrelevant either.

The most precise ruleset for Brides-To-Be appears to describe shopping for invitations, floral arrangements, and wedding favors; there's even a rule for bride who is adding housewares to her registry (Table 10).

Table 10. Ruleset for Brides-To-Be (Rule 0 = indicator set)

Rule	Antecedents
0	Crafts-Hobbies-1, Speciality-and-Gifts-8, Seasonal-1
1	Personalized-Gift-Site-1 ∧ Housewares-1
2	Housewares-1 ∧ Digital-Printing-4:10 Great Product Offers
3	Housewares-1:Tabletop ∧ ¬Housewares-1:Gift Registry
4	Personalized-Gift-Site-1:Home page
5	Digital-Printing-3 ∧ Housewares-1
6	Flowers-Gifts-1 ∧ Digital-Printing-4
7	Department-Store-1:Classic Collections ∧ Housewares-1:Furniture
8	Housewares-1: Search ∧ Housewares-1: Gift Registry ∧ Housewares-1: Home page
9	Digital-Printing-4: Any page ∧ Digital-Printing-3: Home page

This indicator set for Brides-To-Be contains three very strongly wedding-oriented sites; this situation is somewhat ideal.

The indicator set for the Moms-With-Kids descriptive segment is likewise composed of relevant kid-oriented sites. In this case, we had more sites to select into the indicator set. The resulting ruleset discovers some kid-related departments, and contains rules for moms who are shopping for personalized items that aren't business cards, and moms who are shopping for health and wellness items (Table 11).

Table 11. Moms-With-Kids Ruleset (Rule 0 = indicator set)

Rule	Antecedents
0	Toys-1, Digital-Printing-2, Kids-Clothing-1, Toys-2, Toys-3, Kids-Gear-1, Furniture-1, Kids-Gear-2, Speciality-and-Gifts-1, Kids-Clothing-2
1	Housewares-1: Home Page
2	Clothing-1 ∧ Shoes-1: Home Page
3	Clothing-2: Home page
4	Clothing-1: Girls
5	Housewares-1 ∧ Digital-Printing-4
6	Clothing-3: Kids
7	Clothing-1: Boys
8	Shoes-1: Home page ∧ Department-Store-1
9	Personalized-Gift-Site-2 ∧ Digital-Printing-4 ∧ ¬Digital-Printing-4:Business Cards
10	Health-and-Wellness-1 ∧ ¬Health-and-Wellness-2: Home page

Keep in mind that the indicator set elements are excluded from the feature set, so ruleset generation is restricted to building rules from a less richly relevant set of features.

For the ground truth-labeled Moms-With-Kids examples, all sites and segments are included in the feature set. Two of the indicator sites are discovered when training with demographic data and a common clothing brand (Table 12), but overall many of the rules differ.

This suggests that ground truth demographic data for Moms-with-Kids leads to different shopping behavior than our Platonic Concept of a Mom-with-Kid, who shops for children's clothing and toys. Both had similar precision, when evaluated on the demographic data, but focus on different parts of the space. It calls into question whether we should be marketing to the true demographic or the population whose shopping patterns we wish to target.

Note rules 12 and 13 in the demographic based ruleset for Moms-With-Kids in Table 11. Negations are included for Home Improvement and Electronics, perhaps ruling out Dads or Handy-Moms. On the other hand, the indicator site method does not exclude these patterns – shopping for children's items is sufficient.

The concept behind the Luxury descriptive segment is somewhat diffuse; consequently, applying the indicator set method to this segment was more problematic, as "luxurious" behavior can take many different forms.

Table 12. Moms-With-Kids Ruleset (Ground Truth)

Rule	Antecedents
1	Clothing-2
2	Telecom-1
3	Kids-Clothing-1
4	Clothing-4
5	Clothing-5
6	Shoes-2
7	Fast-Food-1
8	Automotive-1 ∧ ¬Home-Improvement-1 ∧ ¬Specialty-and-Gifts-2
9	Clothing-6 ∧ ¬Specialty-and-Gifts-2
10	Clothing-14 ∧ ¬Specialty-and-Gifts-2
11	Toys-3 ∧ ¬Specialty-and-Gifts-2
12	Toys-1 ∧ ¬Electronics-1 ∧ ¬Specialty-and-Gifts-2 ∧ ¬Home-Improvement-1
13	Department-Store-3 ∧ ¬Home-Improvement-1 ∧ ¬Electronics-2 ∧ ¬Clothing-1 ∧ ¬Specialty-and-Gifts-2

The highest-precision ruleset for Luxury discovers a few departments that are vaguely indicative of extravagance or expensiveness (Table 13); perhaps the nicest surprise is the exclusion of shoppers at Clothing-6:Apartment (a youthful budget-oriented site) from Rule 5. However, the ruleset as a whole isn't as convincing as those for Brides-To-Be or Moms-With-Kids (Table 4).

Table 13. Luxury Ruleset (Rule 0 = indicator set)

Rule	Antecedents
0	Cosmetics-1, Speciality-and-Gifts-4, Shoes-5, Speciality-and-Gifts-5, Speciality-and-Gfts-6, Clothing-9, Speciality-and-Gifts-7, Clothing-10, Clothing-11, Clothing-12, Department-Store-2
1	Department-Store-1:Sales & Values
2	Housewares-1:Check out
3	Department-Store-1: Jewelry
4	Clothing-6 ∧ Shoes-4
5	Clothing-6 ∧ Department-Store-1 ∧ ¬Clothing-6:Apartment
6	Clothing-7 ∧ Shoes-3
7	Clothing-13 ∧ Housewares-1
8	Department-Store-1 ∧ Clothing-7
9	Clothing-8
10	Housewares-2:Gifts

The College Students segment and the Trend-setters segment yield similarly vague but not entirely off-the-mark rulesets.

5. DISCUSSION

5.1 Successes

5.1.1 Performance Metrics

Given our method for generating proxy labels for the training and test data, the remainder of the model building process has been fairly successful in terms of the prediction performance of the rulesets that were generated. All rulesets rated well in terms of TPR >> FPR and small ruleset overlap. We generated rulesets of the desired size and provided multiple rulesets per descriptive segment so that a choice of precision vs. reach was available.

5.1.2 Validation on "Ground Truth"

We were able to validate that, for the Moms-With-Kids segment, some of the rulesets created using the indicator set positive criteria performed almost as well on approximate ground truth labeled data as did the "best-case" ruleset that was trained on the same type of data.

5.2 Shortcomings

5.2.1 Upper Limit on Reach

Maximum reach for each ruleset was gotten by reducing the threshold and down-sampling ratio to their most lenient values. Beyond that, this method only allows reach to be increased by either making the rulesets larger (i.e., increasing the maximum allowed number of rules) or by increasing the size of the indicator set (to create more positives).

Increasing the maximum ruleset size was not an option; however, some attempts were made to increase indicator set size both manually and automatically (by adding discovered sites and departments in rulesets to the indicator set), but the affects on reach were small.

It may be that, if indeed there are more ground truth positives in the population than we are finding, our method is fundamentally limited by the numbers of visitors to the relevant websites. This would be especially true for descriptive segments that are less behaviorally oriented towards on-line shopping, or at least to the categories of e-commerce sites appearing in our network.

5.2.2 Believability of Rulesets

Despite the good predictive performance of the rulesets on the labeled data, the bottom line of this exercise is whether or not the rulesets are selecting ground truth positives. The scenario as postulated precludes ever knowing for sure, at least until the rulesets are operationalized and other streams of performance data are created. In the absence of any approximate ground truth to help validate future rulesets, a subjective assessment of the rulesets themselves will hold much weight.

As we have seen, some descriptive segments received unsuprising rulesets and some did not. It must be reiterated that a surprising ruleset does not necessarily indicate an invalid one – Luxury shoppers may indeed pay frequent visits to the somewhat more down-market Shoes-4 and Shoes-3; but this subjective assessment may in fact determine whether a ruleset is accepted by marketing for production. In the end, we must *believe* in our solution.

6. FUTURE WORK

The several challenges encountered during this exercise may be remedied somewhat by some different approaches to the problem.

6.1 Alternative Labeling Schemes

The way in which indicator sets were created and used to label examples could be improved.

6.1.1 Categorized Indicator Sets

Indicator sets would be chosen as before but categorized; thresholding would be employed across the categories to require that shopping behavior span a number of categories.

6.1.2 Weighting of Examples

In addition to assigning a binary class to each example, a weight could also be assigned based on the degree to which the example satisfies the selection criteria. For example, the number of indicator sites a cookie has visited could serve as the weight.

The RIPPER algorithm in Weka allows training with weighted examples; rule coverage is computed as the sum of the weights of the examples that fire the rule. Consequently, examples that were more confidently labeled as positives and negatives would drive ruleset creation.

6.2 Extending the Set of Positives

Increasing reach will probably involve increasing the number of positive examples in the data.

6.2.1 Iterative Labeling by Ruleset Positives

Using the current indicator set approach for initial labeling, generate a ruleset. Next, redefine the set of labeled positives in the data as the union of the existing positives and the ruleset positives. Generate another ruleset and repeat. Stop when reach is sufficiently high or precision falls too low. Start with a bootstrap indicator set that is conservative and precise.

7. CONCLUSIONS

We have shown that shopping data is a reasonable proxy for demographic and behavioral data for the task of identifying customer segments for marketing. For computational efficiency, RIPPER is a more appropriate tool than Apriori, when our task is to learn DNF rules for a pre-defined target. We also find that when evaluated against ground truth demographic data, our rules trained on positives labeled with shopping data have similar true and false positive rates. We may learn an interesting lesson when we look at the details of rules learned, and contrast rules learned from demographic training data with rules trained on shopping positives: the demographic data may indeed focus a little too precisely on the technical definition of segment membership, while shopping data allows us to find others with similar shopping behavior, even if their demographics step outside stereotypes.

8. ACKNOWLEDGMENTS

Thanks to Brian Mancuso for help with data processing. Thanks also go out to our friends on the sales and marketing side: Avi Spivack, Amy Deneson, and Kym Insana. Special thanks to Danner Stodolsky for brainstorming and feedback.

9. REFERENCES

[1] Maria M. Abad-Grau, Maria Tajtakova, Daniel Arias-Aranda. 2009. Machine Learning Methods for the Market Segmentation of the Performing Arts Audiences. *International Journal of Business Environment, Volume 2, Number 3 / 2009.* 356-375.

[2] ADS Descriptive Segments. Akamai Technologies, Advertising Decision Systems. http://www.akamai.com/html/solutions/ads_descriptive_segments.html

[3] Rakesh Agrawal, Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Databases.* 487-499.

[4] Akamai Technologies Privacy Policy. http://www.akamai.com/html/policies/index.html

[5] A. G. Büchner, M. D. Mulvenna. 1998. Discovering Internet Marketing Intelligence Through Online Analytical Web Usage Mining. *ACM SIGMOD Record (special issue on electronic commerce), ISSN 0163-5808, 27(4).* 54-61.

[6] William W. Cohen. 1995. Fast Effective Rule Induction. *International Conference on Machine Learning, 1995.* 115-123.

[7] Jorge Díez, Juan José Del Coz, Carlos Sañudo, Pere Albertí, Antonio Bahamonde. 2005. A Kernel Based Method for Discovering Market Segments in Beef Meat. *Proceedings of the 16th European Conference on Machine Learning – 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD 2005.* 462-469.

[8] Raquel Florez-Lopez, Juan Manuel Ramon-Jeronimo. 2009. Marketing Segmentation Through Machine Learning Models: An Approach Based on Customer Relationship Management and Customer Profitability Accounting. *Social Science Computer Review (2009), Volume 27, Issue 1.* 96-117.

[9] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations, Volume 11, Issue 1.*

[10] D. Iacobucci, P. Arabie, A. Bodapati. 2000. Recommendation Agents on the Internet. *Journal of Interactive Marketing, 14(3), Summer.* 2-11.

[11] Perlich, C., and Z. Huang. Relational Learning for Customer Relationship Management. *International Workshop on Customer Relationship Management: Data Mining Meets Marketing.* NYU 2005