Pleasing the advertising oracle: Probabilistic prediction from sampled, aggregated ground truth

Melinda Han Williams, Claudia Perlich, Brian Dalessandro
Dstillery
470 Park Ave South
New York, NY, 10016
melinda,claudia,briand@dstillery.com

Foster Provost
NYU/Stern School
& Dstillery Research
44 W. 4th Street
New York, NY, 10012
fprovost@stern.nyu.edu

ABSTRACT

Most video advertising campaigns today are still evaluated based on aggregate demographic audience metrics, rather than measures of individual impact or even individual demographic reach. To fit in with advertisers' evaluations, campaigns must be optimized toward validation by third-party measurement companies, which act as "oracles" in assessing ground truth. However, information is only available from such oracles in aggregate, leading to a setting with incomplete ground truth. We explore methods for building probabilistic classification models using these aggregate data. If they perform well, such models can be used to create new "engineered" segments that perform better than existing segments, in terms of lift and/or reach. We focus on the setting where companies already have machinery in place for high-performance predictive modeling from traditional, individual-level data. We show that model building, evaluation, and selection can be reliably carried out even with access only to aggregate ground truth data. We show various concrete results, highlighting confounding aspects of the problem, such as the tendency for pre-existing "in-target" segments actually to comprise biased subpopulations, which has implications both for campaign performance and modeling performance. The paper's main results show that these methods lead to engineered segments that can substantially improve lift and/or reach—as verified by a leading third-party oracle. For example, for lifts of 2-3X, segment reach can be increased to 57 times that of comparable, pre-existing segments.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition-Applications

ADKDD'14, August 24-27 2014, New York, NY, USA

ACM 978-1-4503-2999-6/14/08$15.00.
http://dx.doi.org/10.1145/2648584.2648587

Keywords

Probability Estimation; Online Advertising; Audience Targeting; Logistic Regression

1. INTRODUCTION

As digital advertising has grown into a sizeable industry, it has also become one of the most exciting playing fields for predictive modeling and machine learning. Today we use very sophisticated optimization to make bidding decisions in 30ms that consider the consumer's web browsing history, search history, purchase patterns, social connections, and the context of the site where the ad is shown [13, 14, 4].

Today's technology allows us to provide truly personalized advertising with individual-level metrics. In contrast, parts of the industry have yet to transition from a view of advertising born in an era where advertising meant print media, billboards, and television. These marketers are accustomed to thinking about consumers not as individuals acting in real time, but as broad target audiences, typically characterized by demographics (gender, age, etc) and maybe some wide interest group descriptions ("middle-aged, affluent soccer mom"). As a result, the demand for demographic-based targeting is still high, particularly in video advertising, where the most common metric for video advertising is still the "percent in audience." At the outset of a campaign, the marketer defines a target audience (e.g., "Female age 18-49"). The performance of such a campaign is measured by a third-party service, which quantifies what percent of the campaign was in the target audience ("in-target"). The firm running the campaign is paid only for the impressions which the third-party service deems in-target, and the firm must assume the cost of all extra impressions. Measurement companies like Experian, Quantcast, Nielsen and Comscore provide this measurement service based on proprietary processes for estimating the demographic composition of any arbitrary set of users [15]. We will call these third-party services "oracles."

A common industry solution to assembling such in-target audiences is to purchase demographic information at the user level from a third-party provider (e.g., Bluekai). However, the oracles use proprietary methods for estimating demographics that typically are based on proprietary datasets (such as survey results, credit reporting information, or social network information). As a result, the demographic estimates from the different oracles vary. Estimates purchased from one company typically do not perform well against an-

other company's measurement [12]. This can lead to competing targeting goals, depending on the desires and constraints of the advertiser. For this paper, we will consider one of those goals: the advertiser wants to deliver ads to a large group of individuals who will get high in-target ratings from a designated oracle, without regard to the methodology that that oracle uses. Specifically, in this case it does not matter whether the person really is female and 30 years old; for better or worse, it only matters if the oracle thinks that the person is.

Another industry solution is to buy oracle reports on each piece of ad inventory (that is, each website or group of websites with available ad space) and place ads only on the websites with high in-target audience rates. This approach mirrors the longstanding approach used in print and television advertising, and carries the same drawbacks: lack of precision and limited volume. Targeting based on inventory alone makes it is impossible to assemble an audience with a higher in-target rate or a larger audience than your purest or largest piece of available inventory. Exceeding these restrictions requires individual-level targeting. We will discuss how such inventory-based selection also can lead to biased audiences, rather than representative in-target audiences.

One might suggest a solution relying on the oracle itself as a feedback mechanism. A bandit strategy could be developed where specific segments are chosen for oracle measurements in such a way that incrementally increases the segment accuracy on each iteration [5]. However, oracle measurements are neither cheap nor quick, so this approach would come at a notable price, with a time delay that would not support rapid iteration.

In this paper we present a solution that translates the oracle-based optimization problem into a predictive learning task. We utilize a number of oracle reports received across many campaigns for small audiences and convert the aggregate feedback into class labels to create training examples for predictive modeling. We evaluate the approach first by taking advantage of data for which we know the individual demographics. Then we report actual advertising performance when using the predictions to target several specific demographic groups at scale, achieving much higher in-target reach than is possible with the pre-existing segments.

2. PROBLEM AND SOME NOTATION

For demographic targeting, targeters generally do not have access to individual ground truth for large numbers of online consumers. A version of ground truth can be purchased from third-party validation services such as the oracles described above. However, rather than labeling individual users, these oracles provide aggregate demographic statistics for groups of users. On the technical side, all this is done through server-to-server communications occurring at some "event" where a targeting firm interacts with an internet browser. An event could be showing an ad or visiting a website that triggers a call to the firm's server. At that point, the request may be forwarded to the oracle along with with a segment ID. To further complicate matters, in our own data at Dstillery we do not know the exact set of requests that were forwarded, as they are sampled to limit communication traffic between the servers. We may have decided to collect the oracle data for browsers visiting a specific website, but we do not know which browsers specifically were sent to the oracle. Thus, we get aggregate data on a sample of people who have visited the website today. We will use the term "segment" to refer to any specified set of online consumers.

Demo	Segment 1	Segment 2
Female 2-11	0.0082	0.013
Female 12-17	0.013	0.019
...	...	...
Female 65+	0.111	0.391
Male 2-11	0.023	0.004
Male 12-17	0.031	0.006
...	...	...
Male 65+	0.38	0.109
Female Total	0.32	0.71
Male Total	0.68	0.29

Table 1: Example of an oracle report for two segments with all possible age-gender buckets.

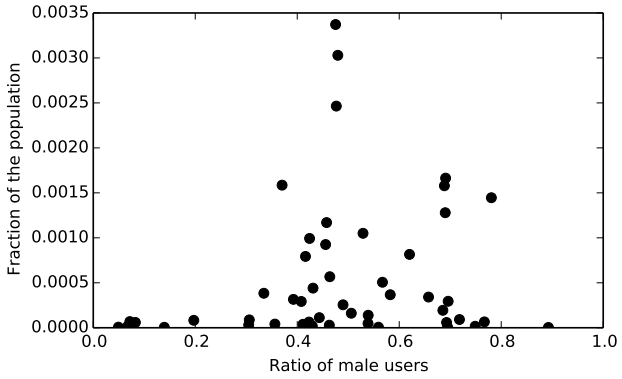

Figure 1: Distribution of gender ratio and size of pre-existing segments for which we have received aggregate audience information from the oracle.

In addition, we have historical reports from entirely unrelated campaigns at our disposal. They cover a variety of segments with different age and gender ratios. The goal is to design a new analytic product that estimates demographic likelihoods at the user level, allowing the creation of high in-target-rate segments for a particular audience.

For example, Figure 1 shows the distribution of gender ratio and size of fifty pre-existing segments for which we have received oracle feedback. The two observations are: 1) the majority of segments have gender ratios close to 50% and 2) the few segments with high or low ratios tend to be small. A naive inventory-based targeting approach would be ineffective for several reasons:

1. A gender ratio close to 0.5 implies that the advertiser must run almost twice as much as theoretically necessary to make up for ads that were not in-target.
2. Exposing too much of the wrong audience to a message can create a negative perception for the publisher as well as the brand.
3. Segments with acceptable precision are for the most part tiny, leading possibly to very high frequency of exposure (the same person sees the same ad 50 times) in order to deliver the desired number of ads in-target.
4. Somewhat less obviously, the very high and low segments are typically biased toward a non-representative subpopulation. For example, the segment with the highest female ratio is a group of users using stamps for scrapbooking; hardly a representative sample of fe-

males in general. Furthermore, using only a few highly biased segments for any targeting purpose is unlikely to lead to satisfactory performance, for which we present evidence later.

In summary, we would like to build a predictive model that for a given audience description (e.g., "Female," or "Female age 18-49") can predict the probability that a browser is in-target, such that a set of browsers with high predictions would form a segment that receives a high oracle score. The target variable is a binary class label: "in-target"/"not in-target." Features are derived from user information, such as website visitation history, location, browser type, etc. At that point it would be a standard classification problem that can be addressed with a number of popular algorithms.

A typical classification problem assumes examples in the form of K-dimensional feature vectors, $X = (x_1, ..., x_K)$. For the purpose of training, a set of N pairs (X, C) of such feature vectors and corresponding class labels $C \in \{C_1, ...C_M\}$ are drawn from some data-generating distribution D. Without loss of generality let us focus on the binary case for M=2 and $C = \{0, 1\}$. The goal of predictive modeling is to estimate a function for that conditional probability $\hat{p}_k = F(X)$ subject to some loss function $L(\hat{p}_k, C)$ over all training examples. To build predictive models one might assume, for example, that the observed class label C is drawn from a Bernoulli distribution with $P(C = k|X) = p_k$.

In the present problem, we define a segment s as any arbitrary collection of users X that we decide to send to the oracle. The oracle can observe the data distribution D and thus can provide information that can help to complete the tuple (X, C). As a policy, the oracle only sends back, for each s and class $C = k$, the segment-conditional probability of class membership $r^s(C_k) = P(C = k|s)$. With this feedback, we define D^s as the oracle's segment conditional data distribution. From D^s we will create the tuples $(X, r^s(C_k))$, which form the basis of our classification problem.

In essence, we attach to each example the oracle probability of the segment it was sampled from as a form of probabilistic label, as a proxy for the unknown true class label. In the next section, we will consider a few alternative approaches of converting these probabilities into class labels for training. We will refer to the inaccessible true class labels $C \in \{0, 1\}$ as "true" ground truth, to the aggregate or probabilistic labels with $r^s \in [0, 1]$ as "fuzzy" ground truth, and to artificially assigned labels $t \in \{0, 1\}$ as "crisp" labels.

3. LEARNING FROM ORACLES

To frame our setting as a generic classification problem, we focus on three key components of the problem: 1) the selection of segments s from which we sample to build the training data, 2) the translation of aggregate labels to individual level labels, and 3) evaluation and model selection using validation data with only aggregate labels.

The last point is potentially the most interesting and challenging: Given that we have no ground truth on the labels, how can we possibly evaluate our work? One slow and expensive option is to create new segments and actually collect oracle feedback. Instead, we focus on whether it is possible to generalize some performance ranking within the limited amount of information we have. To test this, we obtained an external dataset of Facebook users with similar structure to our core dataset for which we have demographic ground truth. We ran experiments on this data to validate our methods and show more detailed results in Section 4.

3.1 Sampling Segments to Build Training Data

Our training data consists of feature vectors X for each user, the segment that each user belongs to, and the per-segment class statistics $r^s(C_k)$ returned by the oracle. Although we know which users belong to which segments, we do not know exactly which users were sent to the oracle. Therefore, we assume that these users are a random sample of the segment s and that in large samples $r^s(C_k)$ converges to $E_S[P(C = k|X)]$, where the expectation is taken over X.

Building a suitable training set from this data raises the following questions:

1. How many examples should we request from each segment? That is, given the lack of ground truth, does the general wisdom of more training data improving the model still hold?
2. Are some segments more informative than others? Are some segments entirely irrelevant and potentially detrimental to model performance?

Some intuitive responses come to mind: segments with very high or low ratios are probably more valuable. However, these can introduce biases that distort our ability to generalize the demographic estimation. This is a subtle but important issue. Most of our segments comprise users who show interest in a particular item. These items are often associated with high gender or age ratios, which we observe in the oracle feedback. Our objective is to generalize the characteristics of different demographic classes unconditional of the specific items that interest the person. Using a few high- and low-ratio segments for training might unintentionally pick up just the behavioral signals that reveal interest in the item. These models would likely have high precision at low scale, but we need models that generalize in order to achieve good recall (i.e., reach). Thus, we can't necessarily limit ourselves to the seemingly most informative (i.e., very high- and low-ratio) segments.

3.2 Creating Labels

At this point we have a training set where we have attached the oracle feedback to each example as a probabilistic fuzzy label that is constant for all examples from a given segment. There are two options that would make direct use of the fuzzy labels: (1) treat the problem as a regression task with a continuous outcome and (2) deal with the proper statistical interpretation of the label in the loss function of the algorithm (e.g., change the calculation of likelihood inside a logistic regression to a probability label).

For reasons primarily of convenience and reuse of our existing large-scale predictive modeling infrastructure, we do not use either of these methods. Rather, we prefer to cast the problem as a classification task with binary labels. Specifically, we have a reliable infrastructure to estimate logistic models with L1 and L2 penalties on millions of features and would prefer by far to utilize this infrastructure for the new demographic targeting product. Once a training set with binary labels has been developed, it can then seamlessly be used with any of our algorithms. We explore the following options for building a training set with crisp labels, given a set of examples with fuzzy labels.

1. Probabilistically labeling: Assign a crisp (binary) label to each instance as a Bernoulli draw with p equal

to the oracle probability $r^s(C_k)$. Additionally we can duplicate each example d times before assigning labels.

2. Maximum a posteriori (MAP) labeling: Fuzzy class labels can be viewed as the probability of membership in each class given the feature vector. As such, a crisp label is given by the maximum a posteriori result: $r_j = 1$ and $r_{i\neq j} = 0$ for $j = \text{argmax}_k r^s(C_k)$.

Of the proposed methods above, the first seems intuitively to best take advantage of the provided information, whereas MAP neglects the information provided by the probabilistic labels: a segment with oracle rating 0.51 would get the same labels (all 1) as a segment with rating 0.99. Furthermore, MAP produces a degenerate training set if segments have $r^s(C_k)$ all above or all below 0.5. Both observations seems to argue against MAP.

However, there is an interesting observation about the difference in label accuracy between the two methods that warrants at least some quantitative analysis of the MAP method. Suppose we have S segments, each representing a portion B_s of the training set, so that $\sum_{s=1}^{S} B_s = 1$, with the fuzzy label for each segment given by $r_1, r_2, ..., r_S$. For probabilistic labeling with a large number of duplicates, the label accuracy is given by:

$$Acc = \sum_{s=1}^{S} B_s(r_s^2 + (1 - r_s)^2) \quad (1)$$

If the labels are assigned by a MAP translation, we have:

$$Acc = \sum_{s=1}^{S} B_s \times \max(r_s, 1 - r_s) \quad (2)$$

Thus, the accuracy of the MAP labels is greater than or equal to the accuracy with probabilistically weighted labeling, with equality when $r_s = 0.5$ for all s.

3.3 Evaluation without Truth?

Model selection and validation pose nontrivial problems in the absence of true class labels. Fortunately, we are interested less in absolute performance of a model than in relative performance between candidate models as we make design decisions. We will return to this question later with the help of a surrogate dataset where the true labels are known. For now, consider the scenario of a training set with crisp labels according to one of the above labeling mechanisms. For any dataset with crisp labels and a corresponding estimator for $P(C_k|X)$, we define AUC_{cr} as the AUC measured on crisp labels and AUC_{tr} as the AUC measured on the actual, but usually unobserved true labels. We can always treat ours as a "regular" classification problem, selecting a subset of the data for validation and measure AUC_{cr} on "holdout" data with crisp rather than true labels. However, using AUC_{cr} leads to a few problems:

1. The selection bias problem still persists. Our holdout set isn't necessarily representative of the general population, so even good measures of AUC_{cr} or AUC_{tr} might not generalize.
2. A high value for AUC_{cr} with MAP labels really means we can predict segment membership for segments with at least a 50% class ratio. Doing so does not guarantee a good AUC_{tr}, leading us to be overconfident in our ability to generalize on the demographic estimation.
3. A subtle but important issue is that with probabilistic labeling we might severely underestimate the performance due to the noise in the labels. For example, a model with near perfect gender discrimination may score only $\text{AUC}_{\text{cr}} = 0.6$. While for this application we only need care about the ranking, if the range of the metric is severely restricted, there is a higher likelihood of missordering due to estimation variance.

Consider a simple scenario where the test set is composed of only two segments of equal size, one with a gender ratio of $r^1(C) = 0.6$ and one with a ratio of $r^2(C) = 0.4$. Imagine in a worst-case scenario that for some reason a perfect predictor of segment membership is available. Under MAP labels the best possible predictions will simply be 1 for all examples in segment 1 and 0 for all examples in segment 2, and this will produce an $\text{AUC}_{\text{cr}} = 1$. In this toy example, we can analytically derive that $\text{AUC}_{\text{tr}} = 0.55$, and we see that AUC_{cr} is severely overestimated. Generally speaking, the better we can predict $P(S|X)$ the more likely we are to run into this problem.

Given the true labels of this test set, any model with an $\text{AUC}_{\text{tr}} > 0.5$ produces a ranking with more true positives than true negatives in the top 50%.[1] In a probabilistic labeled version of this test set, 48% of the positive examples are mislabeled as negative, and 48% of the negative examples are mislabeled as positive. This has the effect of reducing the number of positive crisp labels in the top 50% of the predicted ranking, and reduces the AUC_{cr}.

4. SIMULATING AN ORACLE

To evaluate whether the proposed method indeed can learn well from aggregate ground truth, we generated an oracle setting from data where the ground truth is in fact known. We use a dataset of Facebook users and their characteristics, obtained courtesy of the myPersonality Project.[2] This dataset contains about 220,000 anonymized user IDs, each with a gender label and binary indicators for the Facebook items they have "Liked." One notable point is that this Facebook dataset has no connection to our advertising data; we cannot transfer any of the ground truth to our domain. We use it only to assess potential methods. The Facebook dataset has a natural analogy to the data we use for targeting browsers with ads. Past website visits correspond to Likes and can be used as model features; segments can be defined in various ways, such as the set of users who Liked a particular item.

As mentioned above, sampling via a small set of predefined segments—based for example on consumer interest in particular demographically oriented items—can have the effect of creating selection biases within the training data. Many of our methodological choices are designed specifically to counteract the effects of these biases. We have designed our experiments on the Facebook data to specifically create similar selection biases. Our goal was to recreate conditions in the lab that best mirror the conditions we observe in our production setting.

We created "predefined segments" of Facebook users by assigning users to segments based on specific objects they have Liked. These are similar in structure to the actual predefined segments in our advertising context. We then created a "simulated oracle" feedback mechanism by aggregating the true labels (known for the Facebook data) for each segment s, in order to estimate values for $r^s(C_k)$. We then selected

[1] Since the baserate is 50%.

[2] http://mypersonality.org/

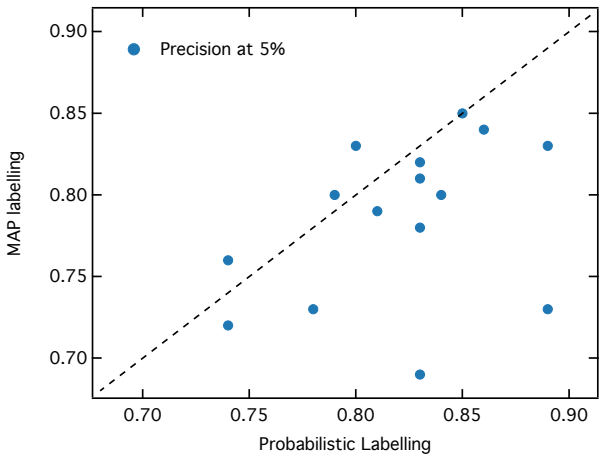

Figure 2: Comparison of the performance of pairs of models each trained on the same set of records, labeled using either MAP or probabilistic labeling, and evaluated on the same test set. All initial datasets were drawn from segments sampled based on Facebook Likes, as described in the text. The dashed identity line shows equivalent performance.

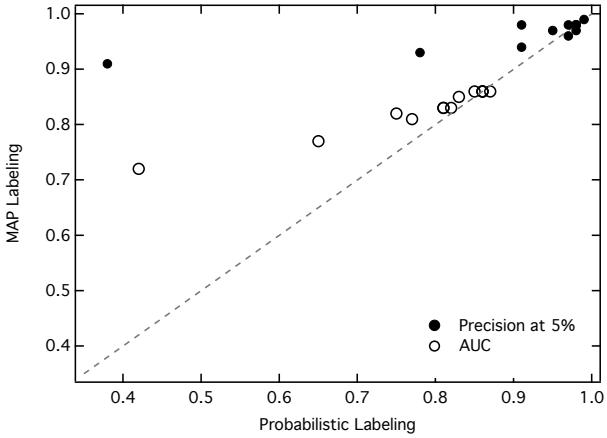

Figure 3: The performance of models trained on datasets labeled with MAP or probabilistic labeling, when the datasets are drawn from unbiased segments—sampled randomly from the population with prescribed gender ratios. Note the difference from the prior figure.

a collection of about 50 specific segments to build our training data. We selected this collection to mimic the sizes and gender ratios we observe in our own data. Compare our own segments in Figure 1 and the corresponding set of Like-based Facebook segments in Figure 8. The difference here, of course, is that we actually know the ground truth for the individuals in the Facebook segments.

4.1 Labels, Segments, and Sampling

Different policies for labeling and training set creation must be evaluated in tandem. Thus, first we explore the effect on model performance of the following design variants: 1) varying the ratio thresholds that drive segment sampling, 2) MAP vs. probabilistic labeling, and 3) varying the duplication of records in the probabilistic labeling. For our evaluation metrics, we will use both AUC and the precision of the predictions at varying percentages. Precision is the more relevant metric for this application because the precision at a given percentage k is equivalent to the in-target rate for the set of users in the top k% of the model predictions.

MAP vs Probabilistic

For the first experiment, we compare MAP labeling and probabilistic labeling. We do this under different segment sampling schemes to show that the presence of selection bias impacts the optimal labeling choice. In all cases, we report metrics evaluated on true labels. Figure 2 shows the results of the main experiment. Here each segment was defined as the set of users associated with a particular Like. We created multiple training sets using this sampling design and in each training set we vary the segments sampled (and thus the demographic ratios of segments included in the training data). For evaluation we created a holdout set that is a random sample of all users with true demographic labels. Each point shows the precision at 5% of the holdout set for each training set and the axes represent the labeling scheme.

A substantial majority of the points shown in this figure fall below the identity line, indicating that probabilistic labeling is a better strategy. This is a somewhat surprising result, since MAP labeling always leads to labels at least as accurate as those produced by probabilistic labeling. We suspect that this is an artifact of the selection bias discussed throughout this work. To confirm this, we ran the same experiment under a different sampling schema, designed to eliminate the bias in the Like-based segments. Instead of defining segments as groups of users that Liked a particular Facebook object, we constructed segments with prescribed gender ratios, but which sampled uniformly from the male and female populations. We then created multiple training sets by drawing from these uniformly sampled segments and did the same labeling test. Figure 3 shows the result of this experiment. We that see for both AUC_{tr} and precision at 5%, MAP is the better strategy when the datasets are drawn from unbiased segments.

We deduce from the above results that if segment assignments are independent of the user features in our data (i.e., $P(s|X) = P(s)$), it would be better to choose MAP labeling. However, when using the biased pre-existing segments, where s is in fact dependent on X, probabilistic labeling is better. In our production system we know for sure that s is not independent of X, and often times this dependency is quite strong. Thus it makes sense to choose probabilistic labeling as the production strategy.

Sampling

The next question is: how should we sample segments? In particular, should we sample segments based on the ratio $r^s(C_k)$ measured by the oracle? To test this, we used 30 segments (defined by groups of people Liking an object), and generated 15 different datasets from them. We wanted to measure how performance metrics change after varying the accuracy of the training data. We start by sampling training examples from two segments: the ones with the highest and lowest ratios, and proceed by adding more segments with ratios closer to 50%. All datasets were the same size $(20,000)$ and drew evenly from each segment used. Observe that these 15 datasets have decreasing accuracy as you go down the list, using either method of label assignment. We applied both labeling strategies and report the performance against ground truth in Figure 4.

We see again that probabilistic labeling tends to perform better despite the fact that the accuracy of the labels is

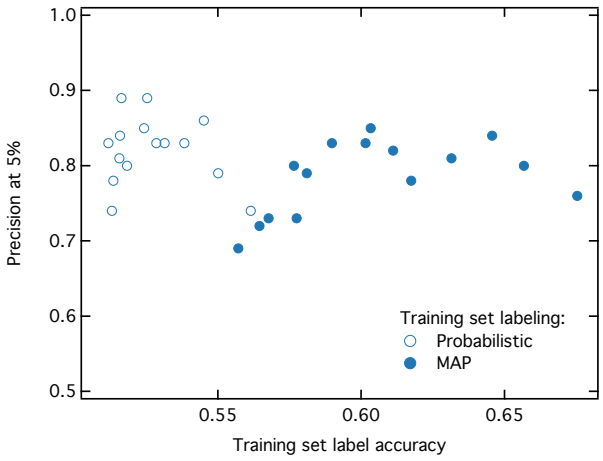

Figure 4: Precision at the top 5%, for the same datasets and models shown in Figure 2. Training sets were sampled from predefined "seed" segments such that those made from more segments had lower accuracy.

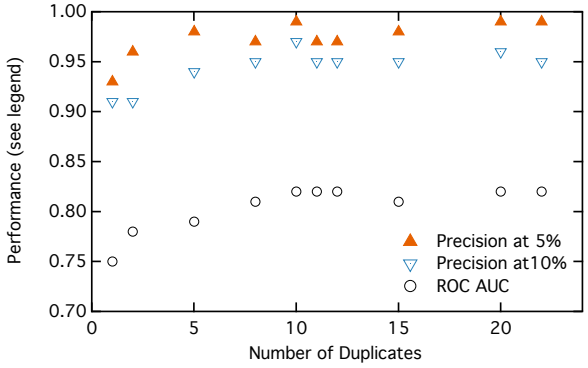

Figure 5: AUC and precision of the top 5 and 10 percent, as a function of the number of duplicates of each example in the training set.

lower. In addition, both labeling strategies create a hump shape, with the optimal performance halfway between uniform sampling from all segments (lowest accuracy) and only sampling from the most extreme segment (highest accuracy).

Duplication

Given that probabilistic labeling seems to be better overall, we now consider whether it helps to include the same example multiple times. Figure 5 shows the performance of models trained on datasets labeled with probabilistic labeling with varying resampling parameter d. This initial dataset was built from two segments selected by sampling from the full dataset based on gender, to create one segment with an aggregate label of 0.8 and one with a label of 0.2. This was done in order to evaluate the dependence on d in general, without the effects of selection bias. Improvements in performance seem to diminish for d greater than about 10. Following these results, we use $d = 10$ for all probabilistically weighted labeling in the remainder of this paper.

4.2 Evaluation Consistency

Recall that our goal for evaluation is a relative performance measure that will allow us to reliably compare two models. We previously identified two reasons why out-of-sample evaluation with crisp labels could be a problem: 1)

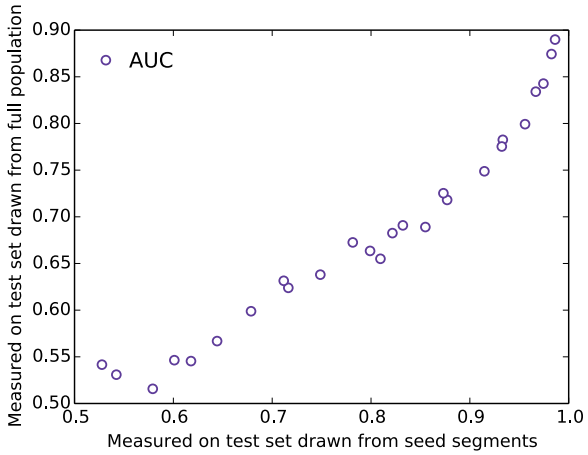

Figure 6: AUC measured using an out-of-sample test set with true labels drawn from the full population (vertical axis) against AUC_{tr} measured using an out-of-sample test set with true labels drawn from the seed segments.

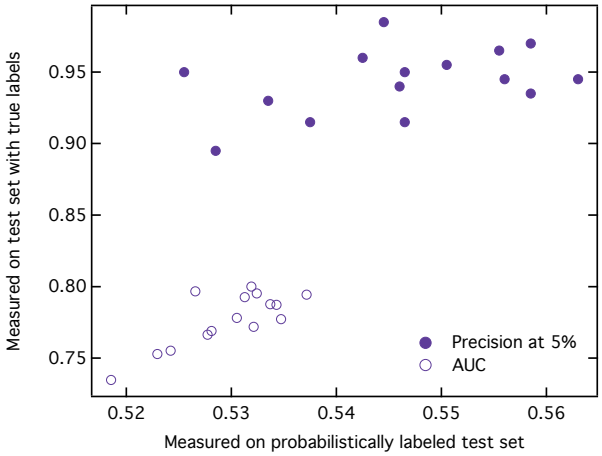

Figure 7: Correlation between performance (AUC and precision at 5%) measured on crisp labeled and true labeled holdout sets, sampled from seed segments. Crisp labels used for this test set were assigned by probabilistic labeling with $d = 10$.

Our holdout set is drawn only from the predefined ("seed") segments, so the results are not guaranteed to generalize to the full population; 2) The translated crisp labels on the holdout set differ from the true individual labels.

Accordingly, we posit that out-of-sample validation on crisp labeled data drawn from the seed segments is only advisable for this purpose if: 1) Evaluation out of sample measured on a true labeled test set sampled from the full population produces a ranking of candidate models which is consistent with the ranking produced from a true labeled test set drawn from seed segments, and 2) Evaluation measured on a true labeled test set drawn from the seed segments ranks candidate models consistently with that measured on a crisp labeled holdout set sampled from seed segments—the data we have access to. We address these two factors in turn.

Initially we examine the bias in the seed segment population with respect to the population at large. In order to see this relationship for models with a wide range of AUC, we built training sets by drawing true labeled samples from the seed segments and adding varying degrees of noise. Models

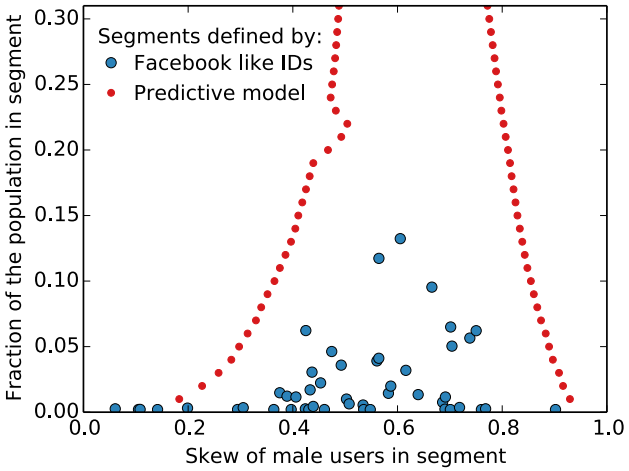

Figure 8: Results from a model to predict male users. Predefined segments, based on Facebook Likes (blue) and engineered segments, created from the model (red) are plotted against the in-target rate and the segment volume, with respect to the full population. Note the fine-grained choice available for the model-generated segments, as well as the much larger size for any in-target ratio.

trained on these datasets were then evaluated against hold-out sets with true labels, drawn from either the full population or from the seed segments. The resulting relationship between AUC_{tr}, measured on a test set drawn from the seed segments, and AUC measured on a test set drawn from the full population is shown in Figure 6. There is clear evidence that the bias affects the model evaluation: the AUC measured on a test drawn from the seed segments is generally higher than that measured on a test drawn from the full population. However, the relationship is monotonic, which suggests that the rankings measured on either set are consistent with each other.

Figure 7 captures the relationship between performance measured on a true labeled set and a crisp labeled set, both sampled from the seed segments. While the precision at 5% does not seem to rank consistently, AUC_{cr} does show a reasonable correlation with AUC_{tr}. It remains open whether there is indeed a significant mis-ranking or a variance problem due to limitations of our test set size.

In summary, although our evaluation method is biased, we are cautiously optimistic that even in the absence of a comparable dataset with true labels, it is possible to make optimization choices on crisp labels alone. We have also experimented with evaluation directly on the aggregate labels, but did not observe generally better performance.

4.3 Finally: Creating New Segments

The results reported so far are in service of our ultimate question: (how) can we use this methodology to build a model that predicts a demographic target population and generates *large* segments with high in-target statistics. Here we will compare the statistics of the original segments (with similar characteristics as ours, based on Facebook Likes) to segments we create from a model built on crisp labels. In the latter case we put a user in a segment if $P(C = k|X) > \delta$.

Figure 8 makes a compelling argument that supports our methods. We chose 48 likes from the Facebook data that result in seed segments with size and ratio statistics resembling those of the actual pre-existing segments we will use in our final online ad targeting application. (The distribution of the actual pre-existing segments is shown in Figure 1; the Facebook Like segments appear in blue in Figure 8.) We pooled these Facebook seed segments and generated a model to predict $P(Male|X)$. This model was then used to estimate the in-target likelihood of individual users. From these predicted likelihoods we created "engineered" segments by varying the cutoff δ. We can see in Figure 8 the accuracy vs. scale tradeoffs that we achieve with our predictive methodology. If high accuracy is a priority, we can create a male segment with 90% in-target covering 3.6% of the population. This is fifteen times larger than the pre-existing segment with 90% in-target. Similarly, if scale is the priority, we can reach 22.5% of the population and still achieve an in-target rate of 80%.

4.4 Summary of Findings

In summary, there is strong evidence that we can achieve our goal of generating demographic predictions that greatly exceed our pre-existing segment information in both purity and scale. More specifically, we learned:

1. When segment membership is not random on the features X, we expect a training set with crisp labels assigned by probabilistic labeling to produce more reliable results than one using MAP labeling.
2. Duplicating each example further improves the probabilistic labeling, and $d = 10$ duplicates is sufficiently high to capture most of this improvement.
3. When selecting segments to include in a training set, segments with high or low ratios lead to improved overall label accuracy. However, it is important to also include a sufficient amount of data from segments with ratios nearer 50% to mitigate the impact of selection bias inherent to high- and low-ratio segments. The definition of "sufficient" here will change with the specific problem and dataset.
4. Comparative model evaluation can be carried out based on oracle data alone, which are translated to crisp labels using probabilistic labeling. This evaluation gives a reliable comparison between candidate models that can be used to fine-tune segment selection.

5. HIGH-REACH ONLINE DEMOGRAPHIC TARGETING

We applied the methodology presented above to our application of interest: targeting demographic groups for display advertising. Here we present results from models we built and deployed in our production setting, targeting the demographic group "Female age 18-49". We use past website visits as features, and we define each pre-existing (seed) segment as the set of browsers that has visited a particular website. Third-party demographic reporting gives us the "oracle" ratings which serve as the aggregate class labels for these seed segments. In our production problem, we never have access to the full ground truth. Model selection must be done using historical oracle ratings as truth, and after choosing a final model we may purchase oracle ratings for the segments we build.

Development of the "Female age 18-49" model represents an additional variant of this problem. For this particular target group, only three of the 50 seed segments at our disposal had an in-target rate of greater than 50%. Having a target

group that is in the minority in all or most seed segments changes the problem in two key ways. First, translation to crisp labels must be done using probabilistic labeling, as MAP labeling would produce a dataset labeled entirely negative. Second, segment selection includes a new trade-off. When all segments have minority labels, the only segments with ratios far from 50% are those with very low ratios in the positive class, and result in very few positive labels. In "simulated oracle" experiments (omitted here for brevity), we found that a balance between overall label accuracy and high label sensitivity (that is, the ratio of correctly labeled positives to all positives) in the training set resulted in better models. Thus, we select more segments with a wider range of ratios.

All models were built using logistic regression with L2 regularization and trained with stochastic gradient descent. For more details on our modeling methodology, see [14, 4]. For all models, final decisions on segment selection and regularization strength were made by evaluating on a hold-out set with crisp labels created by probabilistic labeling. In this model selection process, we face an unavoidable selection bias problem because our test set does not sample evenly from the entire population we work with in production. Accordingly, we use a test set built from all available segments to minimize selection bias. This has the additional effect of producing a test set with relatively low label accuracy, which reduces discrimination between models, but we found we were left with enough discrimination to make decisions.

The final models we selected were deployed in our production environment, and we purchased reports on these from the third-party "oracle", which serves as our ultimate validation. The results for the target demographic "Female age 18-49" are shown in Table 2 and Figure 9. We report the in-target rate of three engineered segments, as well as the lift in the in-target rate of our predictions with respect to the baseline rate of the United States population, which is 0.22 for this demographic [3]. Though we do not claim that randomly targeting the U.S. population is the most relevant alternative to our approach, lift over a random baseline is a standard for comparison within the industry, and for the demographic categories shown here, marketers aim for lifts of at least 1.5 to 2. In Table 2 we also report the relative size of the engineered segments, with respect to the largest set of seed segments that could be combined to give a precision at least as high as that of the engineered segment.

Our predictions provide dramatic increases in both lift and volume. Focusing on Segment 2, which balances gains in lift and volume, we show a lift of 2.54 with a volume that is increased by a factor of 57 compared to the largest combination of seed segments with the same precision. For campaigns that target on demographics, we are typically only paid for impressions shown in-target. Without any predictive modeling, we would need to show an average of 4.5 impressions in order to show one impression in-target. Using our model, this number drops to 1.7 for our most selective segment, a 63% reduction in the number of impressions, and hence the cost of the campaign.

6. RELATED WORK

There have been a number of efforts using this set of Facebook data for modeling personal preferences based on online activity such as likes ([1, 9, 7, 8]). Most of them confirm that there is a strong relationship between the things people do online and core characteristics like gender and age. This is consistent with our ability to use browsing histories to build demographic segments.

Segment	In-Target	Lift	Volume	Rel. Volume
1	0.60	2.73	0.0029	35.2
2	0.56	2.54	0.0072	57.0
3	0.49	2.23	0.017	7.09

Table 2: Results from our demographic prediction models in a production setting, verified by a third-party "oracle".

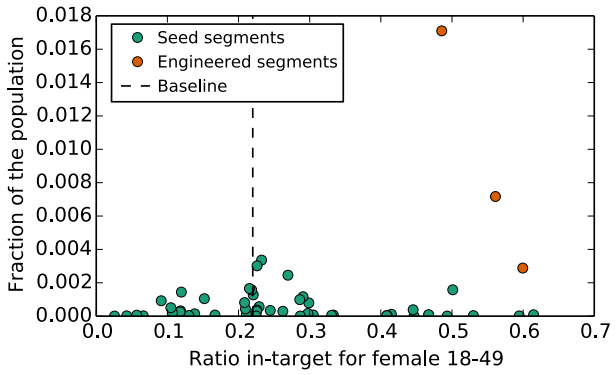

Figure 9: Results from engineered segments verified by a third-party "oracle".

From a technical perspective, learning in scenarios where labels are not available has been considered as an unsupervised learning problem [2]. However, our learning problem is not unsupervised since we do have a form of ground truth information. A classification problem where probabilistic estimates serve as ground truth labels is considered in [17, 16]. In this problem, images of the surface of Venus are assigned a rating by domain experts, indicating the degree of evidence of the presence of a volcano in the image. These papers include a thorough discussion of the problem of assigning labels, though it differs from our problem in that the authors are not constrained to a modeling algorithm that takes in standard class labels. The notion of learning from an oracle was explored for example by [11]. As in our problem, this problem makes use of a type of approximate label, in this case a prediction from an existing and very complex model. The goal was to generate a lower complexity model. The works in both of these scenarios differ from our problem in that individual labels are available, where in our case we only have labels in aggregate, and seek to build a higher complexity model. In that respect, work on modeling team sports is closer to our setting. [10] and [6] infer individual skill level of players based on the wins and losses of the team in aggregate.

7. CONCLUSIONS

In response to a demand from video advertisers, we have created targeted, "engineered" demographic segments as a new product. We designed a process that takes advantage of aggregated labels that are available for purchase from third parties. By framing demographic prediction as a form of probability estimation in very high dimensions, we demonstrate that it is possible to create demographic segments that have sufficient scale to satisfy even large campaigns and

show economically relevant performance improvements over industry standards. For smaller populations utilizing the engineered segments can reduce costs by two-thirds, and improve the customer experience by limiting the frequency of repeated ad views. The engineered segments also do not suffer from the severe bias incorporated by (some) traditional segmentations of online audiences.

In the process of developing this product, we derive interesting methodological insights that we expect to generalize to other applications involving learning from aggregated samples. In particular we consider sampling, labeling, and evaluation on noisy labels. In real problems, segments are biased, and not random cross-sections of the population. As a result, utilizing even segments with very little skew in addition to more informatively skewed segments is beneficial.

Evaluation when ground truth is unavailable adds to the challenge. We observe that the noisy labels provide sufficient ranking consistency to guide the modeling process. In particular, we can use classical holdout methodology. In absolute terms, models tend to look much worse under such evaluation than they actually are. In fact, we've seen that a model which seems nearly random with AUC = 0.54 under evaluation with noisy labels can have AUC = 0.80 on the true class labels.

8. ACKNOWLEDGEMENTS

We would like to thank the Dstillery Tech team under the direction of Rod Hook for all the amazing support that was essential to bring this work to life. This work was conducted while all authors were at Dstillery. We thank the myPersonality Project for generously allowing us the use of their Facebook data. Foster Provost thanks NEC for a Faculty Fellowship.

9. REFERENCES

[1] Y. Amichai-Hamburger and G. Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289–1295, November 2010.

[2] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[3] U.S. Census Bureau. Current population survey, annual social and economic supplement, 2012.

[4] B Dalessandro, D Chen, T Raeder, M Williams, C Perlich, and F Provost. Scalable Hands-Free Transfer Learning for Online Advertising. In *Proceedings of ACM SIGKDD*. ACM, 2014.

[5] J. C. Gittins. Bandit processes and dynamic allocation indices. In *Journal of the Royal Statistical Society. Series B*, volume 41, pages 148–177, 1979.

[6] T. Huang, C. Lin, and R. Weng. Ranking individuals by group comparisons. In *23rd International Conference on Machine Learning*, pages 425–432. ACM, 2006.

[7] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, pages 1–24, 2013.

[8] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[9] D. Markovikj, S. Gievska, M. Kosinski, and DS. Stillwell. Mining facebook data for predictive personality modeling. In *7th International AIII Conference On Weblogs And Social Media*, 2013.

[10] J. Menke and T. Martinez. A bradley–terry artificial neural network model for individual ratings in group competitions. *Neural computing and Applications*, 17(2):175–186, 2008.

[11] J. Menke and T. Martinez. Artificial neural network reduction through oracle learning. *Intelligent Data Analysis*, 13(1):135–149, 2009.

[12] J. Neff. Nielsen, comscore pitted in ratings race, 2012.

[13] C. Perlich, B. Dalessandro, R. Hook, O. Stitelman, T. Raeder, and F. Provost. Bid optimizing and inventory scoring in targeted online advertising. In *Proceedings of SIGKDD*, pages 804–812. ACM, 2012.

[14] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, and F. Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine Learning*, pages 1–25, 2013.

[15] M. Shields. Google and Nielsen Partner on Online Campaign Ratings, 2013.

[16] Padhraic Smyth. *Learning with Probabilistic Supervision*. MIT Press Cambridge, MA, USA, 1995.

[17] Padhraic Smyth, MC Burl, UM Fayyad, and Pietro Perona. Knowledge Discovery in Large Image Databases : Dealing with Uncertainties in Ground Truth. *KDD Workshop*, pages 109–120, 1994.