Broad vs Narrow: Modelling Strategies for Online Behavioural Targeting

Markus Svensén[1], Qing Xu[2], David Stern[1], Steve Hanks[2], Christopher M. Bishop[1]

[1]Microsoft Research Cambridge
Roger Needham Building
7 J J Thomson Avenue
Cambridge CB3 0FB,UK

[2]One Microsoft Way
Redmond, WA 98052-6399
USA

{markussv, qingxu, dstern, sthanks, cmbishop}@microsoft.com

ABSTRACT

In this paper we consider different strategies for constructing click-prediction models that can subsequently be used for audience segmentation and behavioural targeting. In particular, we address the question whether one should build separate models for each audience segment or instead build a single model that simultaneously predicts membership in multiple segments. We discuss the pros and cons of both strategies and then investigate which yields the best results empirically. We use a recently developed Bayesian model that is capable of combining traditional feature-based modelling with collaborative filtering based techniques. We apply this model to a large set of web log data, harvested from a collection of linked, large commercial websites. In our experiments, multiple Bayesian logistic regression models, each built for a single segment topic, generally produce better results than a single model built against all topics simultaneously. But there are indications that, at least for some segment topics, allowing the feature representation to depend on the topic can improve the performance of multi-topic models.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models–*statistical*; I.6.5 [**Computing Methodologies**]: Model Development; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Algorithms, Performance, Experimentation.

Keywords

Matchbox, audience segmentation, online advertising, predictive modelling, click-through rate (CTR)

1. INTRODUCTION

The world wide web (WWW) has become an important vehicle for reaching consumers of goods and services through online advertising, and it has been estimated that in 2011, this market will be worth $28.5 billion [6]. Not only does online advertising allow the advertiser to reach a large number of consumers, but more important, it offers the potential of reaching the *right* consumers, i.e. those most likely to purchase the product being advertised. There are several mechanisms of online advertising that serve to achieve this aim. In *sponsored search*, users of search engines such as Google or Bing, are shown not only a list of documents, but also a number of advertisements, in response to their query. By choosing the advertisements to be relevant to the content of the query, the search engine presents the user with advertisements that they are likely to be interested in, thereby improving the chance that an advertisement gets clicked and generates revenue for the search engine.

A complementary form of online advertising is *display ads*, which are shown on pages of web sites along with the main content of such sites, which could be anything from news to social network content to more or less general information about some topic. For many such sites, the main content of pages can be informative when it comes to selecting ads; for example, a page that contains information about gardening is likely to be visited by people with the corresponding interest and so it makes sense to show ads for gardening products. This particular form of adverting is commonly referred to as *contextual advertising*. In situations where the web page does not provide clear and monetizable content, or when the advertiser wants to reach interested customers regardless of where they browse, *behavioural targeting* (BT) becomes relevant. Behavioural targeting uses users' online behaviours, sometimes along with other user features, to predict user interests, and uses those predictions to choose ads that are relevant to the users. This is complementary to both sponsored search and contextual advertising and can thus be used in combination with these and other techniques for online advertising campaigns. Typically, BT will be used for dividing a set of users into topic-specific segments, where users in a segment are selected on the basis of prior behaviours or other features to have an interest in or affinity to the topic. Thus, BT can be used as a method for *audience segmentation*.

Several ways of doing BT have been proposed, and we review some of them in the next section. Since in many cases advertisers measure the value of a segment by how its users respond to the ads (measured most often by click-through rate), most BT/segmentation efforts focus at least in part in choosing an audience likely to click on topic-specific ads (so-called "click models"). We use a probabilistic model, described in Section 3.1, combining features of the user and of the advertisement to

ADKDD '11, August 21, 2011, San Diego, California, USA.

compute the probability of the user clicking on the advertisement. Once such a model has been built, we can use it to evaluate click probabilities for users on a set of pre-classified advertisements, believed to be representative of a specific topic, and then threshold the click probabilities according to desired size and specificity of the resulting user segment. A question that arises when building such models is whether one should build separate models for each individual topic, or build a single model that simultaneously predicts membership in all or a subset of the topics for which segments are being built. Arguments can be made both ways, as we discuss in section 3.2, and we explore both alternatives empirically using user search and browse data collected from behaviours observed on Bing and on the Microsoft display network.

2. BEHAVIOURAL TARGETING

The idea of dividing a marketing audience into segments with the aim of achieving a better response to an advertising campaign is older than online advertising, and corresponding statistical methods have been developed, e.g. for targeted catalogue mailings [1]. However, with the WWW, online advertising and the continued exponential growth in computing power, the possibilities, but also needs, for performing audience segmentation have changed dramatically. It is therefore not surprising that a number of commercial systems (see [14]) have been developed for this purpose. In the academic community, results from research into behavioural targeting for online advertising has only recently appeared. Yan et al. [14] evaluated two clustering methods to identify user segments (clusters) in log data from a commercial search engine. Every entry in this log data corresponded to a click made by a user following a query. Users were represented either by the collection of clicked URLs (page-views) or the collection of query terms they used. Regardless of which user representation was used, and whether a log entry corresponded to a click on an advertisement or on a link to a document returned by the search engine, no information as to whether the item clicked had any relevance to a specific topic or brand was used, so for the purpose of audience segmentation, this corresponds to learning segments in an unsupervised fashion. However, Yan et al. demonstrated that some of the identified segments had a much higher than average click through rate (CTR) on specific ads. Wu et al [13] took a similar approach, but used a more sophisticated clustering approach, focussing on the query representation of users, found more promising also in [14].

A different approach was taken by Provost et al. [11], who used anonymized data from a social network site to build up a graph over users, where a link is created between every pair of users who had visited the same (social network) web page; a link was created for every distinct page, and weights on the links were used to indicate the frequency of visits. User segments were then extracted by picking neighbours to *seed users* (nodes) in the graph believed to have a high affinity with a brand or topic; this belief draws on observations of seed users taking "brand actions," e.g. clicking on corresponding advertisements. Thus, this is more akin to supervised learning of a non-parametric model for predicting brand actions. A different supervised approach is presented by Liu et al. [8], who construct a model for ranking users in terms of their affinity to defining characteristics of segment, such as a topic or a brand, by essentially learning a click prediction model. Chen et al also build a click prediction model [4], but they take a probabilistic approach, where click and impression counts are modelled under a Poisson observation model.

3. BT USING CLICK PREDICTION

Our approach to BT assumes that the action of clicking on an advertisement is an indicator that the user is interested in the topic(s) associated with the advertisement. Under this assumption, the probability of clicking on an advertisement will be directly linked to the probability of the user being interested the topic of the advertisement. Click prediction is a key functionality in any online advertising system and much time and effort has been devoted to this task [7].

While one can argue over whether clicking on an advertisement is a good single proxy for indicating an interest in the topic of the advertisement, it is certainly one such indicator. Moreover, it is a quantity that can be easily observed, and one that carries great weight in advertisers' assessments of advertising campaigns. Predicting the probability of clicking then becomes a natural methodological approach, since with this probability at hand, we have all the information we need to trade off the reach (size) of an audience segment against its expected CTR (precision). Even if predicting clicks is not the ultimate goal of the BT exercise, we can argue that as a sufficient but not necessary component of affinity modelling, it will be a component in any robust BT system.

3.1 Matchbox

Our models are based on Matchbox [12], a probabilistic Bayesian model for matching users and items. Matchbox is an approach to collaborative filtering which incorporates user and item metadata (features). In traditional collaborative filtering, users and items are represented in terms of how they match, i.e., a user is represented in terms of which items he (dis)likes whereas items are represented in terms of which users (dis)like them. This representation allows a model to identify groups of users with similar preferences for items and similarly identify collections of items that are all rated similarly by the same users. By using metadata, Matchbox can make predictions for items with few or no ratings. In the context of this paper, the items are advertisements and users show their preference for these by clicking or not clicking on them.

A Matchbox model is a combination of two sub-models: a linear model, usually referred to as the context model which models bias effects such as how often a given advertisement is clicked in general, and a bi-linear model where features representing users and advertisements are mapped to a latent trait space where the affinity between items and users is measured by the inner product. It is this bi-linear part that will allow Matchbox to model the interaction between advertisements and users and hence, which advertisements are more likely to be clicked on by which particular users. Indeed, one question that we wanted to answer in our experiments was whether Matchbox would reveal latent traits that could be related to topics or brands and thus be the basis of a modelling approach that directly captured user-topic affinity.

Matchbox can operate with a range of feedback (observation) models, corresponding to different encodings of the users' preferences for items, but in this paper we only consider a binary feedback model, corresponding to the click or no-click feedback obtained from the user. As impression and click counts were aggregated on a daily basis, we considered the use of a binomial feedback model, but preliminary experiments suggested that this approach was computationally more demanding than simply replicating the clicked and non-clicked impressions according to the corresponding counts, probably due to the fact that these counts generally were low, only rarely exceeding a few tens of impressions. Another possibility, which we have not yet

investigated, would be to use a Poisson model [4] to model the observed impression and click counts.

Matchbox allows all user and advertisement features, including IDs, to be used in both the linear and bi-linear model. Determining which features to use where, as well as the dimensionality of the latent trait space, must be done through experimentation, as described in Section 5. Given a chosen set of features and a data set from which these features can be extracted, the construction of a Matchbox model requires inferring posterior distributions over all parameters in the model. This is achieved using an approximate, deterministic inference algorithm. The updates for this algorithm were computed using the Infet.NET library [10]. This inference step is commonly referred to as model *training* and the data used is consequently referred to as *training data*, in contrast to *test data* used to score the trained models.

3.2 Single- vs Multi-Topic Models

One of the key issues that we wanted to address in this work was whether we should build individual models using selected data corresponding to a single topic, or build models from data covering multiple topics. There are reasons for and against both of these alternatives. When we build a single-topic model, the model will be solely devoted to make optimum prediction for the chosen topic. A model built to make predictions for multiple topics may have to trade off the performance on one topic against the performance on another topic. On the other hand, if the probability of a click is topic-independent with respect to some of the features, a model built using multi-topic data should be able to better capture the relationship between these features and the probability of click, and thus provide a more accurate model. Moreover, so called multi-task learning [3], where multiple related tasks are learnt simultaneously, has proven useful for determining suitable data representations—in terms of our models this would correspond to selecting which features to include—resulting in improved generalization.

If there were to be strong topic specific dependencies on some features, we would expected that Matchbox models trained on corresponding multi-topic data would discover and exploit such dependencies, provided it had a sufficient number of latent traits at its disposal. However, in situations where such topic specific dependencies are only weakly reflected in the data, a model attempting to exploit these may end up overfitting [2] to the training data.

4. DATA

We used log data from Bing and the Microsoft display network, containing demographic and behavioural data about users (Table 1) descriptive data of advertisements (Table 3) and records of impressions of advertisements to users and whether the user clicked on the advertisement or not.

The data set contained 15 days' worth of data; impression data was aggregated on a daily basis, so that for any particular user and any particular advertisement, we counted the number of impressions and clicks that occurred each day. In total, 284M impressions with 606K clicks, involving 1.8M users and 3270 advertisements, where recorded during the 15-days period. However, we restricted our attention to impressions where both the advertisement and the user met certain conditions. A large proportion of the total number of impressions corresponded to a fairly small number of advertisements that appeared in a number of sizes and formats, even though the content was always the same; for our experiments, we excluded those impressions, since we believed they could bias the results an unrepresentative way. Of the users, we selected only those from the US, UK or Canada, aged between 12 and 75, inclusive, that had seen a minimum of 10 impressions, including at least 2 distinct advertisements, and had clicked on at least 1. In the end, we had a data set comprising 78M impressions with 174K clicks, involving 127K users and 2793 advertisements.

Table 1. User features used in our models.

Name	Type
ID	Unique numerical ID
AgeBand	Multinomial (<12, [12,18], [18,25], [25,35], [35,45], [45, 61], [61, 75], >75)
Country	Multinomial (US, UK, CA)
Gender	Binomial (Male, Female)
QueryExplicitCategoryCount, QueryAncestorCategoryCount, PageViewExplicitCategoryCount, PageViewAncestorCategoryCount	Sparse, real-valued vector features (see text)

Table 3. Advertisement features used in our models.

Name	Type
ID	Unique numerical ID
Type	Multinomial indicating shape and content type (e.g. animated), 17 values
Industry	Multinomial indicating industry of advertiser, 19 values
Size (width and height)	Integers

The behavioural data that we used (page-views and queries) was represented using categories. Every page-view and every query is automatically assigned to one category. Categories are related to one another in a forest-like structure, where individual trees consist of increasingly specific categories; examples of these categories can be seen in Table 2. We tried using counts of only explicitly assigned categories as well as also counting all the parent categories. We then took logarithm of the counts as we

Table 2. Examples of categories used to represent page-views and queries. "/" separate categories in the same category tree. The first two rows provide two examples from the same category tree.

Vehicles_&_Transportation/Automobiles/By_Make_&_Model/BMW/7-Series/750
Vehicles_&_Transportation/Automobiles/By_Make_&_Model/Volvo/XC/XC60
Health_&_Wellness/Mental_Health/Anxiety_Disorders/phobias
Society_&_Culture/Law_&_Legal_Services/Lawyers_&_Legal_Information/Estate_Planning_&_Administration

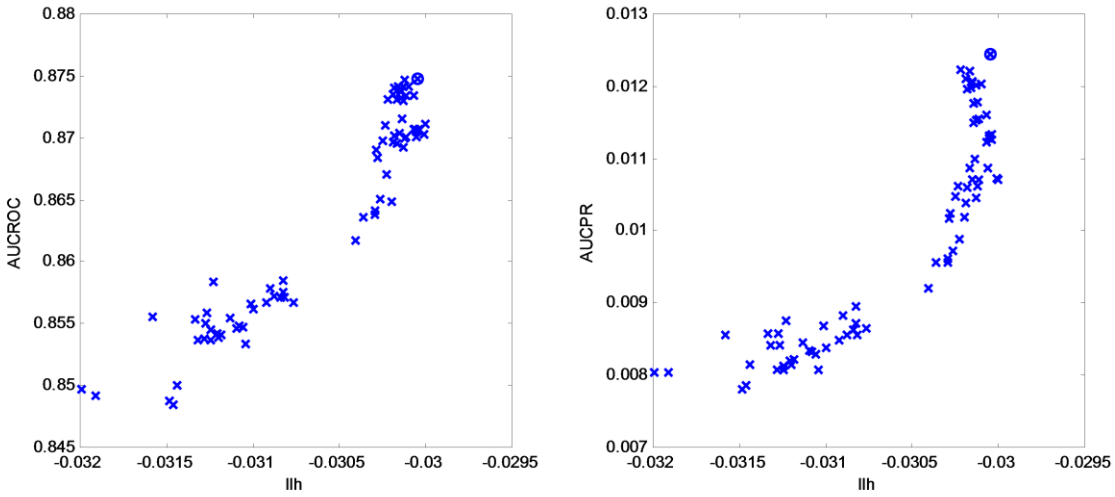

Figure 1. AUCROC (left) and AUCPR (right) plotted llh for different feature combinations and number of latent traits.

believe it is the order of magnitude that is important rather than precise counts. During the 15-day period, the users in the selected data set viewed 4.6 billion pages and issued half a billion queries.

5. EXPERIMENTS

In order to determine the best combination of user, advertisement and context features (see Section 3.1), as well as the most appropriate dimensionality of the latent trait space, we used a compute cluster to run a large number of experiments, fitting different models and scoring them, using performance metrics that will be discussed in Section 5.1. We then moved on to construct models from impression data corresponding to either just a single topic or multiple topics, and compared the performance of these models on test data from the chosen topic. Finally, we made exploratory experiments to investigate the viability of models that are in some sense half-way between single- and multi-topics models.

Our general experimental setup was similar to that of Provost et al. [11], in that we used data from the first M days to build the model (training data), whereas data from (some of) the remaining $15 - M$ days were used to score the models. We tried different values for M, ranging from 9 to 14. The Matchbox models were all trained using assumed density filtering (ADF) [9], with a single pass through the entire data set. We also tried Expectation-Propagation [9] with multiple passes through the training data during preliminary experiments, but we never observed any significant performance gain on test data compared to models trained using ADF.

Before proceeding to the results, we briefly describe the performance metrics used.

5.1 Performance Metrics

To compare different models, we used three different measures: area under the receiver operator characteristic curve (AUCROC), the area under the precision-recall curve (AUCPR) [5] and marginal log-likelihood (llh), defined as

$$\text{llh} = \frac{1}{N}\sum_{n=1}^{N} \ln p(c_n),$$

where $p(c_n)$ denotes the probability of the observation c_n (click or no click) for the nth impression under the model, and N is the number of observations in the test data. The log-likelihood serves as an approximation to the log-evidence of the test data, which is the obvious score to use for Bayesian model selection. These quantities are of course related, as can be seen in Figure 1.

As a further check that our models indeed are learning to model the quantity we are interested in, we plotted *calibration curves* for the best scoring models. We construct this curve by binning the range of probabilities of clicks output by the model, aggregating the impressions in these bins, according to the probability of click predicted by the model for each impression and compute the empirical probability of click within each bin. We can then plot these empirical click probabilities against the corresponding bin centres.

6. RESULTS

We start by looking at the problem of determining an appropriate model. This involves determining which features to use in the context model, which user and advertisement features to use in the bi-linear model and the number of latent traits in the bi-linear model. Figure 1 show AUCROC and AUCPR plotted against llh, showing scores obtained from 75 models with a range of different feature combinations and different number of latent traits. Scores for 97 other models with llh scores below -0.032 were excluded from this plot in the interest of clarity. This plot shows that these three quantities are closely correlated, as one would expect, but also that they still rank the same models slightly differently.

Figure 2 show mean, and standard deviations of the three test scores, along with minimum and maximum values, plotted against the number of latent traits; models with zero latent traits are models with only context features, i.e. Bayesian logistic regression models, and these are the ones that generally do best. Since the three test scores are not in complete agreement about which model does best, we simply choose a model that scores well on all three of them. This model includes all advertisement features and all demographic user features (rows 1–4 in Table 1) as well as ancestor category counts for both page-view and queries; these are all being used as context features. The corresponding model scores are indicated with a ○ in Figure 1. The calibration curve for this model is shown in Figure 3.

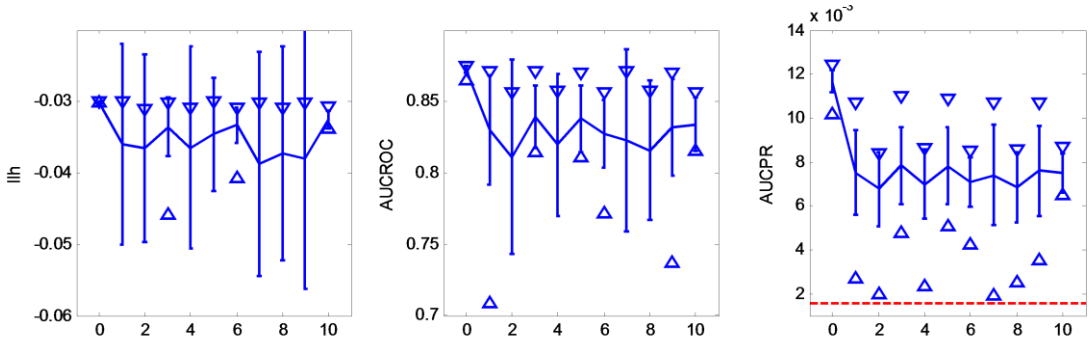

Figure 2. Test scores plotted against number of latent traits. The solid line show average scores (mean) along with error bars (1 standard deviation), minimum (△, some out of range) and maximum (▽) values. The dashed horizontal line in the AUCPR plot corresponds to the baseline CTR in the test data.

6.1 Single- vs Multi-Topic Models

To investigate the relative merits of single- and multi-topic trained models we made use of manually chosen topics applied to the advertisements in the data; in total, 128 topics were present in in data, organised in a similar fashion to the categories used for the behavioural data; however, it should be noted that the topics are quite distinct from these categories.

The available topics were screened for the total number of impressions and the total number of clicks observed during the period of data collection. We picked five topics that were reasonably diverse, had at least 800K impression and CTR that exceeded a minimum threshold of 0.005:

- Financial Services/Financial Planning & Management /College Financing (Financial)
- Health/Drugs (Health)
- Information Technology/Telecommunications /Mobile Phones/Smart Phones (IT)
- Shopping/Clothing & Accessories (Shopping)
- Vehicles/Automobiles (Vehicles)

"/" separates topics in the same 'topic tree' and in parentheses are the labels used in subsequent figures.

For each topic, we built Matchbox models that were trained on impressions from either just the chosen topic or all five topics, during the training period, and then scored the resulting model only on impressions from the chosen topic during the test period. We also looked at using the entire data set for training, including impressions also from other topics, but this did not lead to improved results. For the majority of these experiments, we used a Bayesian logistic regression model (Matchbox with zero latent traits) with the feature configuration that had been found to produce the best results across the larger data set. We also ran experiments to select the feature configuration and number of latent traits using only data from the five selected topics, but the performance of the best resulting model was worse than that reported below.

Figures 4 and 5 show the scores obtained on test data with single- and multi-topic trained models for the five topics; Figure 5 also show the per-topic average CTR. From these figures, it seems that the single topic models do better and thus that the advantages gained by focusing on a single topic outweigh any potential gains from learning topic independent feature dependencies across multiple topics. While this result is perfectly reasonable, the question still remains, whether there are features whose effects on the predicted click probabilities are topic independent and, if so,

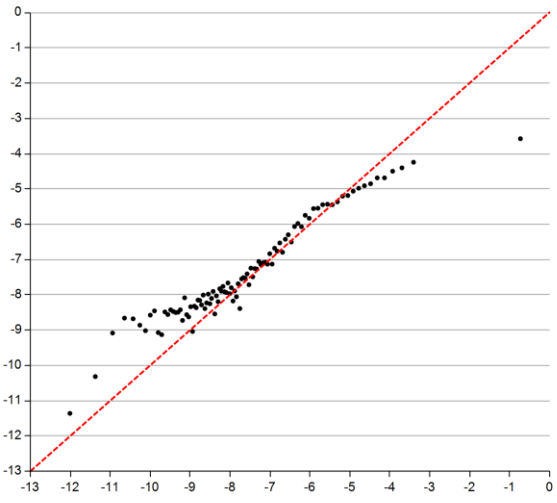

Figure 3. Calibration curve for the best scoring model, showing the empirical log-CTR on y-axis plotted against the predicted log-CTR on the x-axis.

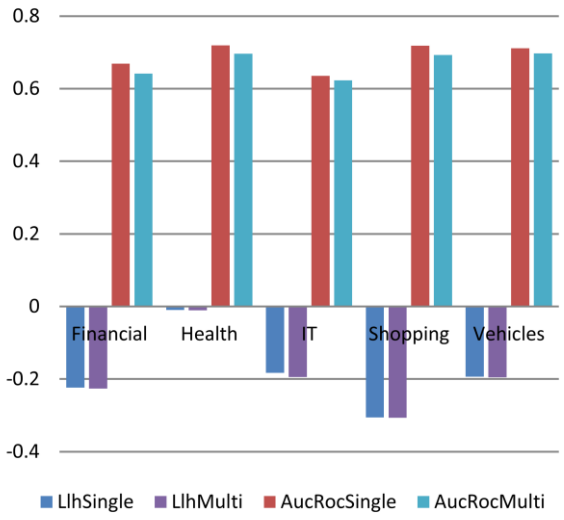

Figure 4. llh and AUCROC topic test scores for single- and multi-topic trained models.

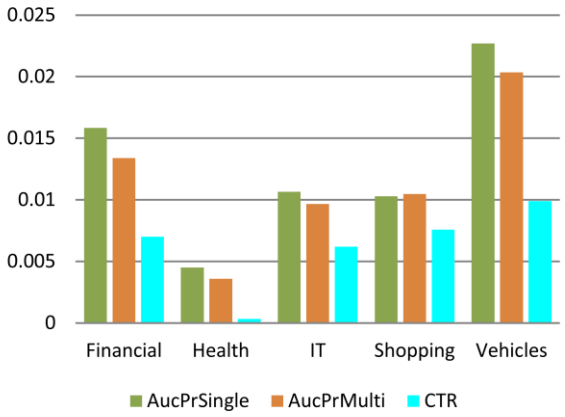

Figure 5. AUCPR topic test scores for singe- and multi-topic trained models, and the per-topic, average, test data CTR.

how we could capture this in our models.

Here we propose to derive new, topic dependent features by forming Cartesian products of topics and existing features of moderate dimensionality.

For example, we can form the Cartesian product of user gender and topics to obtain a 10-nomial feature, covering all possible combinations of gender and our five topics. With such topic dependent features, multi-topic models could be built that would be able to exploit topic specific dependencies of some features while treating other features, for which no such dependencies exist, in a uniform manner across all topics. An alternative to forming the Cartesian product of a feature and all topics is to create a 'topical' feature, which has one representation of a feature for a specific topic and another representation used for all other topics; in other words, we form the Cartesian product of the chosen feature and only two topics: ChosenTopic and NotChosenTopic.

We investigated the use of topic and topical features for the user features Gender and AgeBand and the advertisement features Type and Industry. Generally, the performance obtained with these new features was comparable with that obtained with the multi-topic models trained on the features used in previous experiments. However, results varied across topics and features. Figure 6 compares topic and topical features for user AgeBand; for the topic Shopping/Clothing & Accessories; we observed a small improvement in AUCPR, and the same treatment of the advertisement Industry feature gave a similar result on this topic.

7. DISCUSSION

To summarise our results, we found that of the several Matchbox models we tried, a Bayesian logistic regression type model gave the best performance. Furthermore, the strategy of using training data corresponding to a single topic (audience segment) gave better performance on test data than the alternative strategy of using more training data including multiple topics. The differences in performance between single- and multi-topic models are not very large, but if they are genuine, they can still correspond to significant differences in revenue.

A question raised by this result is why the signals in the data exploited by the single topic models could not be exploited by Matchbox multi-topic models with one or more latent traits? Our first set of experiments suggests this is not the case, but these experiments were carried out on a data set with potentially many more topics represented. This data set was also larger than in the experiments comparing single- vs multi-topic models, but it may not have been large enough to compensate for the increased diversity. The single-topic strategy *forces* the models trained to focus on just a single topic; giving a model 'the option' to

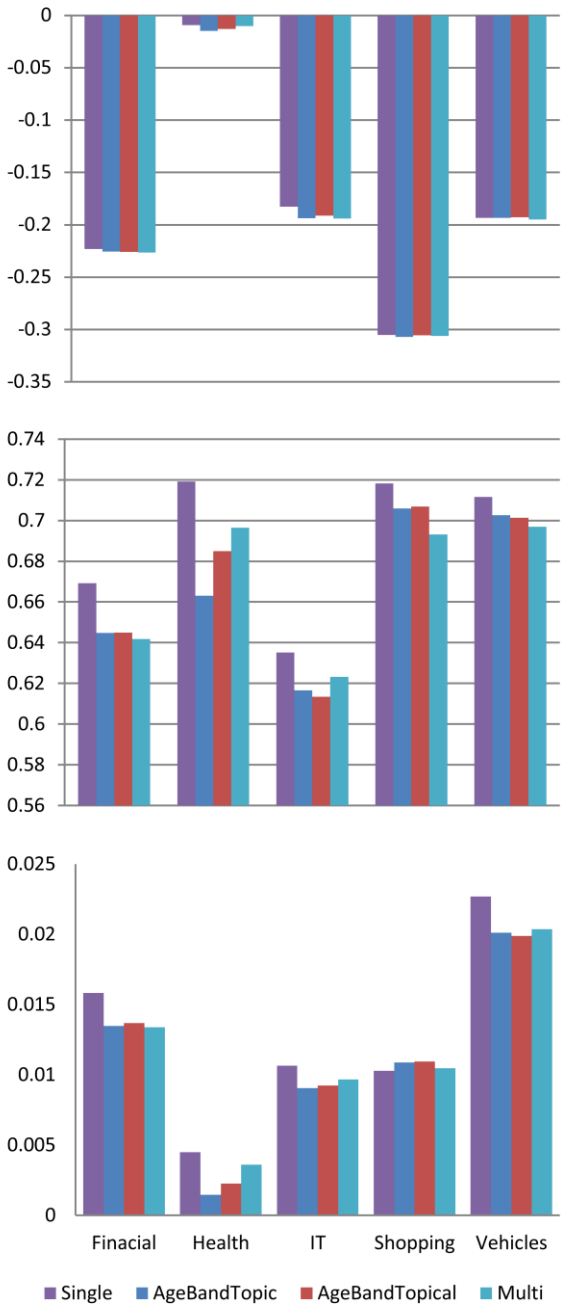

Figure 6. llh (top), AUCROC (middle) and AUCPR (bottom) test scores obtained with single- and multi-topic trained models, as well as multi-topic models trained with AgeBand-topic specific features and AgeBand-topical features (see main text). The model colour coding and topic stack order is the same in all three panels.

represent weak trends in the data may not have the same effect. While multi-topic models in theory should have the potential to perform as well as or better than single topic models, this might be difficult to realize in practice.

Our final set of experiments indicates that topic-specific features may allow for further improvements in performance. Strategies need to be devised for how best identify feature-topic combinations that are likely to give the biggest increases in prediction performance.

In this paper, we have used users' clicks on advertisements as indications of interest in topics that are associated with the advertisement. However, we would like to be able to identify additional, complementary indicators that can help us in more reliably gauging users' interests in different topics. Data that come out of social networking sites, as well as new data from users operating via Smartphones and lightweight mobile computers, such as the iPad, maybe useful for this aim, provided privacy issues that come with these data can be addressed. What seems clear is that behavioural targeting will become an increasingly important tool for companies selling advertising opportunities on the WWW.

8. REFERENCES

[1] C Apte et al., "Segmentation-Based Modeling for Advanced Targeted Marketing," in *SIGKDD*, San Fransisco, 2001.

[2] Christopher M Bishop, *Pattern Recognition and Machine Learning*.: Springer, 2006.

[3] Rich Caruana, "Multitask Learning," *Machine Learning*, vol. 28, pp. 41-75, 1997.

[4] Ye Chen, Dmitry Pavlov, and John F Canny, "Large-Scale Behavioral Targeting," in *Proceedings SIGKDD*, Paris, France, 2009, pp. 209-217.

[5] Jesse Davis and Mark Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, Pittsburg, USA, 2006.

[6] eMarketer. (2011, April) US Digital Ad Spending: Online, Mobile, Social. [Online]. http://www.emarketer.com/Reports/All/Emarketer_2000794.aspx

[7] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borschert, and Ralf Herbrich, "Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine," in *International Conference on Machine Learning*, Haifa, Israel, 2010.

[8] Ning Liu et al., "Learning to Rank Audience for Behavioral Targeting," in *Proceedings SIGIR*, Geneva, Switzerland, 2010, pp. 719-720.

[9] Thomas P Minka, "A family of algorithms for approximate Bayesian inference," MIT, Boston, USA, PhD Thesis 2001.

[10] T Minka, J Winn, J Guiver, and D Knowles, Infer.NET 2.4, 2010, http://research.microsoft.com/infernet.

[11] Foster Provost, Brian Dalessandro, Rod Hook, Xiaohan Zhang, and Allan Murray, "Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting," in *Proceedings SIGKDD*, Paris, 2009.

[12] David Stern, Ralf Herbrich, and Thore Graepel, "Matchbox: Large Scale Online Bayesian Recommendations," in *Proceeding of WWW'09*, Madrid, Spain, 2009.

[13] Xiaohui Wu et al., "Probabilistic Latent Semantic User Segmentation for Behavioral Targeted Advertising," in *Proceedings ADKDD (CD)*, Paris, 2009, pp. 10-17.

[14] Jun Yan et al., "How much can Behavioral Targeting Help Online Advertising?," in *18th International World Wide Web Conference (WWW2009)*, Madrid, 2009, pp. 261-270.