Online Multimedia Advertising:

Techniques and Technologies

Xian-Sheng Hua
Microsoft Research Asia, China

Tao Mei
Microsoft Research Asia, China

Alan Hanjalic
Delft University of Technology, The Netherlands

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Senior Editorial Director: Kristin Klinger
Director of Book Publications: Julia Mosemann
Editorial Director: Lindsay Johnston
Acquisitions Editor: Erika Carter
Typesetters: Michael Brehm, and Milan Vracarich, Jr.
Production Coordinator: Jamie Snavely
Cover Design: Nick Newcomer

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

Library of Congress Cataloging-in-Publication Data

Online multimedia advertising : techniques and technologies / Xian-Sheng Hua, Tao Mei and Alan Hanjalic, editors.
p. cm.
Includes bibliographical references and index.
Summary: “This book unites recent research efforts in online multimedia advertising and includes introductions to basic concepts and fundamental technologies for online advertising, basic multimedia technologies for online multimedia advertising, and modern multimedia advertising schemes, theories and technologies”--Provided by publisher.
ISBN 978-1-60960-189-8 (hbk.) -- ISBN 978-1-60960-191-1 (ebook) 1. Internet advertising. 2. Internet marketing. 3. Multimedia systems. I. Hua, Xian-Sheng, 1973- II. Mei, Tao, 1978- III. Hanjalic, A.
HF6146.I58O56 2011
659.14’4--dc22
2010051809

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 9
Adapting Online Advertising Techniques to Television

Sundar Dorai-Raj
Google Inc., USA

Yannet Interian
Google Inc., USA

Igor Naverniouk
Google Inc., USA

Dan Zigmond
Google Inc., USA

ABSTRACT

The availability of precise data on TV ad consumption fundamentally changes this advertising medium, and allows many techniques developed for analyzing online ads to be adapted for TV. This chapter looks in particular at how results from the emerging field of online ad quality analysis can now be applied to TV.

INTRODUCTION

As online advertising has exploded in the past decade, it has often been contrasted with traditional media such as television, print, and radio. The inherently connected nature of online content has enabled unprecedented tracking and analysis of online advertising, and the resulting explosion of data has allowed Internet companies to develop ever more sophisticated algorithms for allocating and pricing advertising inventory. Using user-initiated signals (like "clickthrough"), these companies can indirectly measure the relevance of ads in specific contexts, and build models to predict which ads will most interest future users and maximize revenues for online publishers (Richardson et al., 2007).

In contrast, traditional television measurement has typically relied on relatively small panels of pre-selected households to report their viewing

DOI: 10.4018/978-1-60960-189-8.ch009

behavior. This approach often has meant that reliable measurements took days or weeks to produce, and could not be produced at all for niche programming that appealed to very small audiences. These methods also did not generate enough data to build determine which TV ads were most appealing to viewers, nor to predict which ads would be most appealing in the future.

However, a new source of TV-related data has emerged in recent years, one that is closer to Internet scale. The set-top boxes (STBs) used by most cable and satellite TV subscribers are often capable of collecting data on viewing behavior. These data can then be stripped of personally-identifiable information and anonymously aggregated, allowing for very detailed measurement of television viewing behavior. While previous panels collected data from thousands or perhaps tens of thousands of households, set-top box data are available from many million US households and similar numbers in other countries.

With these data in hand, it is now possible for television to adopt many of the analytical techniques pioneered in online advertising. In particular, ads can be scored for relevance or quality based on statistical models that predict how viewers are likely to respond. Ad inventory can then be allocated so as to maximize relevance, and to compensate publishers for any loss of audience due to ads.

This chapter will explore this emerging discipline. We will introduce the basic statistical techniques underlying the approach, and give examples for how such models can be implemented in software. We will also discuss applications of these models to television advertising, and some of the issues raised in the application of these techniques.

TV Audience Measurement

Panel-based audience measurement has a long history. In the US, Arthur Nielsen began measuring television audiences in 1950 based on a nationwide sample of 300 households (Nielsen, 2009a). Because there were only 48 commercial television stations in the US at the time and no more than a handful of viewing choices in any one local area, the audience of any given station could be adequately estimated with such a small sample. The Nielsen Company continues measuring TV audiences today, now with a sample of over 9,000 households (Nielsen, 2009b). Understanding and using these audience ratings has evolved into a discipline unto itself (see, for example, Webster et al., 2006).

Even the modern, expanded panel, however, is at times unable to measure the increasing fragmented TV audience flocking to niche programming (Bachman, 2009). For television ads (as opposed to programs), this bias is further compounded: attempts to judge the reaction of TV audiences to ads have focused on only the most popular programming. For the 2009 Super Bowl, for example, Nielsen published a likeability score and a recall score for the top ads [1]. The scores were computed using 11,466 surveys, and Nielsen reported only on the top 5 best-liked ads and most-recalled ads.

Several companies have started using data from STBs to measure TV audiences. In addition to Google, TNS, CANOE, Retrak, Tivo, and The Nielsen Company itself are using STB data (Mandese, 2009). Several of these companies will make such data available to media researchers on a subscription basis.

Measuring Ad Quality

In the world of online advertising, the term "ad quality" has taken on a very specific meaning. Roughly speaking, an ad quality is "a measure of how relevant an ad is" (Yahoo, 2009); in the context of Internet keyword search ads, ad quality scores "measure how relevant [the] keyword is to [the] ad text and to a user's search query" (Google, 2009). In other words, ad quality represents a judgment on an ad primarily from a user's

or viewer's perspective. It asks, to what extent is the ad the viewer would most like to see.

Ad quality is not, in other words, a measure of cost effectiveness from an advertiser's perspective. It is easy to imagine ads that viewers consider highly enjoyable and relevant, but which are ineffective at meeting the goals of the advertisers themselves. Similarly, it is not hard to find examples of ads that viewers may dislike, but nevertheless seem to be cost-effective. (Much "junk mail" and the online equivalent – email spam – might fall into this second category.)

However, there are reasons to predict a synergy between ad quality (as defined above) and ad effectiveness, at least in the long run. If viewers become accustomed to seeing relevant ads, they may pay more attention and thus increase the effectiveness of ads generally. Furthermore, TV programmers have a vested interest in keeping viewers from changing the channel and so may reap an economic benefit from increasing viewer-perceived ad quality.

SOLUTIONS AND RECOMMENDATIONS

Precise television usage data is now available from several sources. Google aggregates data collected and anonymized by DISH Network L.L.C., describing the precise second-by-second tuning behavior for several million of US television set-top boxes, covering millions of US households, for thousands TV ad airings every day. From this raw material, we have developed several measures that can be used to gauge how appealing and relevant commercials are to TV viewers. One such measure is the percentage initial audience retained (IAR): how much of the audience, tuned in to an ad when it began airing, remained tuned to the same channel when the ad completes.

In many respects, IAR is the inverse of online measures like click-through rate (CTR). For online ads, CTR is a positive action; advertisers want users to click through. This is somewhat reversed in television advertising, in which the primary action a user can take is a negative one: to change the channel. However, we see broad similarities in the propensity of users to take action in response to both types of advertising. Figure 1 shows tune-away rates (the additive inverse of IAR) for 182,801 TV ads aired in January 2009. This plot looks similar to the distribution of CTRs for paid search ads also ran that month. Although the actions being taken are quite different in the two media, the two measures show a comparable range and variance.

A significant challenge in interpreting TV audience data like this is that many factors appear to impact STB tuning during ads, making it difficult to isolate the effect of the specific ad itself on the probability that a STB will tune away. Rather than using raw measures of tune-away directly, we have developed a "retention score" that attempts to capture the creative effect itself.

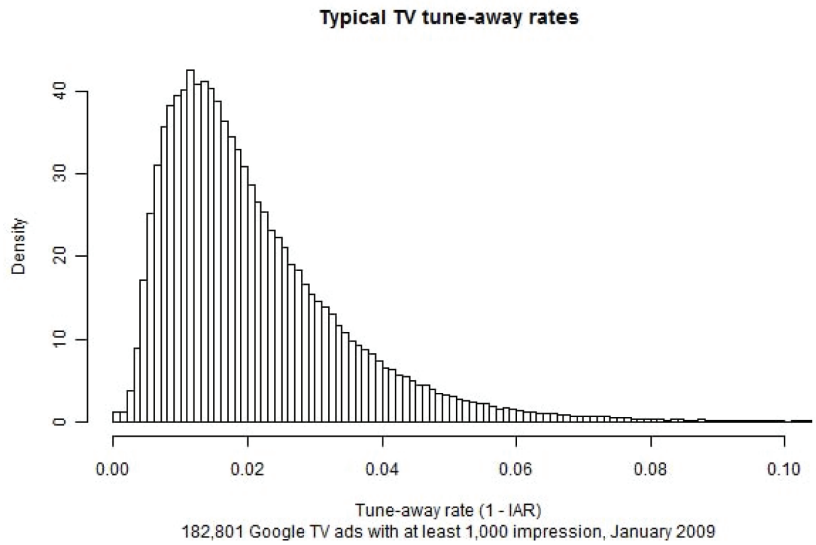
Definition

We calculate per airing the fraction of initial audience retained (IAR) during a commercial. This is calculated by taking the number of TVs tuned to an ad when it began which then remained tuned throughout the ad airing as shown in equation (1).

$$\mathrm{IAR} = \frac{\text{Audience that viewed whole ad}}{\text{Audience at beginning of the ad}} \quad (1)$$

The hypothesis behind this measure is that when an ad does not appeal to a certain audience, they will vote against it by changing the channel. By including only those viewers who were present when the commercial started, we hope to exclude some who may be channel surfing. However, even these initial viewers may tune away for other reasons. For example, a viewer may be finished watching the current program on one channel and begin looking for something else to watch.

Figure 1. Density of tune away rate for TV ads, defined by the percentage of watchers who click away from an ad



We can interpret IAR as the probability of tuning away from an ad. In order to isolate extraneous factors like the network, day part, and day of the week from the effect of the creative, we define the *Expected IAR* of an airing as

$$\hat{\text{IAR}}=\text{E}(\text{IAR}|\mathbf{x}) \tag{2}$$

where **x** is a vector of features extracted from an airing, which exclude any features that identify the creative itself; for example, hour of the day and TV channel, but not the specific campaign or advertiser. Then we define the *IAR Residual* as in equation 3 to be a measure of the creative effect.

$$\text{IAR Residual} = \text{IAR} - \hat{\text{IAR}} \tag{3}$$

There are a number of ways to estimate $\hat{\text{IAR}}$, which we will discuss in this chapter. Using equation 3 we can define *underperforming airings* as the airings with IAR residual below the median. We can then formally define the *retention score* (RS) for each creative as one minus the fraction of airings that are underperforming.

$$\text{RS}=1-\frac{\text{Number of underperforming airings}}{\text{Total number of Airings}} \tag{4}$$

The remainder of this chapter focuses on methods for obtaining accurate retention scores. We explore four different statistical models based on three different algorithms. We will rank each method by metrics we have constructed to measure its usefulness as a measure of ad quality. We have also conducted extensive experiments to validate the usefulness of retention scores as an ad quality signal. These include comparing retention scores to subjected human evaluations for ads, and the extent to which past retention scores predict future audience retention (Zigmond et al., 2009).

MODEL ESTIMATION

In this section we introduce three estimation algorithms to obtain predictions for IAR. All three approaches are based on logistic regression with

IAR as our response. To motivate the algorithm descriptions, consider the following two events:

$$C_0 = \text{STB tuned to the beginning of an ad leaves before the ad ends}$$
$$C_1 = \text{STB tuned to the beginning of an ad remains through the entire ad,} \quad (5)$$

where $\Pr(C_1) = 1 - \Pr(C_0)$. Then, given a vector of features $\mathbf{x}_i$ for observation i, we model the log-odds of C_1 as

$$\log\left(\frac{\Pr(C_1|\mathbf{x}_i)}{\Pr(C_0|\mathbf{x}_i)}\right) = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j. \quad (6)$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is a length-k vector of unknown coefficients. We then obtain estimates of $(\beta_0, \boldsymbol{\beta})$ by maximizing the binomial log-likelihood of $\Pr(C_1 \mid \mathbf{x}_i)$ given the observed IAR:

$$\max_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{k+1}}\left[\frac{1}{N}\sum_{i=1}^{N} n_i\left\{\text{IAR}_i \log(\Pr(C_1|\mathbf{x}_i)) + (1-\text{IAR}_i)\log(\Pr(C_0|\mathbf{x}_i))\right\}\right], \quad (7)$$

where n_i is the number of viewers at the beginning of the ad for observation i and $\Pr(C_1 \mid \mathbf{x}_i)$ is given by

$$\Pr(C_1|\mathbf{x}_i) = \frac{1}{1+\exp\{-\beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j\}}. \quad (8)$$

For this discussion, an "observation" depends on the feature list. At its most basic level, an observation is an individual STB. However, to improve computational efficiency, our algorithms aggregate over STBs with identical feature sets, where each group of STBs is thought of as a single observation. The one exception is the last algorithm, which is efficient enough to handle each STB as a separate observation.

Glmnet Logistic Regression

The Glmnet algorithm controls overfitting and any possible correlations by applying an L1 penalty on the coefficients during the estimation. In essence, we are still maximizing the log-likelihood given by (7) with an additional penalty with the form

$$\max_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{k+1}}\left[\frac{1}{N}\sum_{i=1}^{N} n_i\left\{\text{IAR}_i \log(\Pr(C_1|\mathbf{x}_i)) + (1-\text{IAR}_i)\log(\Pr(C_0|\mathbf{x}_i))\right\} - \lambda\sum_{j=1}^{k}|\beta_j|\right], \quad (9)$$

where λ is a regularization parameter. Software for obtaining coefficient estimates using Glmnet is available through the R package glmnet (Friedman et al., 2009). R is an open source scripting language primarily used for statistical data analysis and visualization (R Development Core Team, 2010).

Choice of the regularization parameter λ affects the amount of shrinkage applied to each β_j. The larger λ is, the smaller the resulting β_j's. Some β_j's will be shrunk to zero, implying this coefficient has no impact on IAR or that it is correlated with another feature in the models.

Principal Components Logistic Regression

The next algorithm we tried is based on principal components logistic regression (Aguilera et al., 2006). The algorithm is as follows:

1. Build a model matrix $\mathbf{X}$ of size n x p, where n is the number of rows and p is the number of parameters (columns).
2. Assuming $\mathbf{X}$ has an intercept, drop the first column of $\mathbf{X}$ and center and scale the remaining columns. Call the result $\mathbf{X}^*$.
3. Transform $\mathbf{X}^*$ using singular value decomposition (SVD) into matrix components $\mathbf{U}_{nxk}$, $\mathbf{d}_{(kxk)}$, and $\mathbf{V}_{(kxk)}$, where

$$\mathbf{X}^* = \mathbf{UdV} \quad (10)$$

and $k = p - 1$.

Define $\mathbf{W}$ as

$$\mathbf{W} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{V} \end{bmatrix}, \tag{11}$$

where $\mathbf{0}$ is a column vector of length k containing all zeros.

Define $\mathbf{Z}$ as

$$\mathbf{Z} = \mathbf{XW}, \tag{12}$$

which is n x p.

6. Keep only the first m columns of $\mathbf{Z}$. There are many published ways for choosing m. See Aguilera et al. (2006), for example.

   Maximize the log-likelihood binomial with response IAR against the first m columns of $\mathbf{Z}$:

$$\max_{\boldsymbol{\beta}^* \in \mathbb{R}^m} \left[\frac{1}{N} \sum_{i=1}^{N} n_i \left\{ \mathrm{IAR}_i \log(\Pr(C_1 | \mathbf{z}_i)) + (1 - \mathrm{IAR}_i) \log(\Pr(C_0 | \mathbf{z}_i)) \right\} \right], \tag{13}$$

where

$$\Pr(C_1 | \mathbf{z}_i) = \frac{1}{1 + \exp\{-\sum_{j=1}^{m} z_{ij} \beta_j^*\}}, \tag{14}$$

and z_{ij} is the ij^{th} element of $\mathbf{Z}$, and β_j^* is the j^{th} regression coefficient based on the j^{th} principal component of $\mathbf{Z}$. Note that (13) and (14) are fundamentally the same to (7) and (8), respectively, except that m is typically much smaller than k.

By maximizing (13) with respect to the β_j^* s and given our observed IAR, we obtain estimates $\hat{\beta}_j^*$. Since $\hat{\beta}_j^*$ is estimated in the transformed space, we convert back to the original space spanned by $\mathbf{X}$ by performing a matrix multiplication given by:

$$\hat{\boldsymbol{\beta}} = \mathbf{W}_m \hat{\boldsymbol{\beta}}^*, \tag{15}$$

where $\mathbf{W}_m$ contains the first m columns of $\mathbf{W}$ and $\hat{\boldsymbol{\beta}}^*$ a column vector containing the $\hat{\beta}_j^*$ s. The resulting vector $\hat{\boldsymbol{\beta}}$ contains coefficients for all features in the model. As with the Glmnet estimators, the estimated coefficients will be shrunk towards zero as more correlations are present in $\mathbf{X}$ and the larger m is.

The latter algorithm is used in the demographics model because of the strong relationship between the makeup of a household and which networks viewers in that household watch.

A Proprietary Logistic Regression Implementation

The last algorithm we tried is based on proprietary logistic regression implementation designed by Google to handle very large data sets. The basic algorithm optimizes over coefficients β_j with respect to the log-likelihood function in (7). However, we apply additional regularization techniques to shrink unimportant or highly correlated coefficients while also merging similar coefficients.

The main advantage of this method is efficiency. The algorithm design allows for significant parallelization, which means we can model more features on a greater number of STBs. In fact, unlike the Glmnet and principal component algorithms, with our proprietary algorithm we model each STB as an individual observation rather than rely on grouping of STBs by feature. Given that we have data for several million STBs, that level of prediction is not possible with most statistical software packages, such as R.

RETENTION SCORE MODELS

We have devised several models for computing retention scores, which help us rank ads based on their quality. In this chapter, we will discuss our findings in the areas of household demographics and user behavior. In addition, we have developed metrics to rank our models based accuracy of predictions and their ability to discriminate ads consistently. In this section we introduce the basic model first developed by Google, along with 3 competing models later designed as possible improvements.

All model comparisons are based on training and test data from October 2009. The primary tool for analysis is the R language, which suffers from fairly severe memory constraints. For this reason, we limited our analysis to the 25 top-viewed networks to improve memory efficiency for the Glmnet and principal components algorithms. For the machine learning algorithm, we used an internal software tool that is much more scalable. However, for comparison purposes, we still limited the analysis to the same networks.

The Basic Model

The first model relates the observed IAR to the Day Part, Weekday, Ad Duration, and Network. Each feature is described in more detail below. All time-based features are EST/EDT.

1. Day Part – a categorical variable with the following levels:
2. Weekday – a Boolean variable determining whether the ad was placed on the weekday (TRUE) versus weekend (FALSE).
3. Ad Duration – the duration of the ad in seconds. Most ads are 15 or 30 seconds, but 45, 60, and 120-second ads are also shown. In the basic model, we treat Ad Duration as a numeric variable and not categorical. Doing so assumes a linear relationship between IAR and Ad Duration.
4. Network – a categorical feature of network id. For the Dish population, this variable has roughly 100 levels.

Each observation used in the Basic Model represents a single airing. We typically use three weeks of data to train a model and estimate coefficients, and predict for only a single week.

Demographics Model

The demographics model is motivated by the fact that households of certain demographic composition tend to watch the same networks. To observe this behavior we first conducted a simple principal components analysis to determine the reduction in dimensionality achieved by modeling variations in household makeup rather than network viewership. In essence, we attempted to determine whether demographics were a proxy for network viewership. This would imply we could subsequently remove or diminish network from our Basic Model as a feature for predicting IAR in return for including certain household conditions.

First let us discuss our findings of the principal components analysis. For the month of October, the percentage of time a given STB was tuned to a particular network was recorded, provided the STB was tuned to that network for at least 60 seconds and under two hours. As mentioned above, only the 25 top-viewed networks were included in the analysis. In addition, the actual names of the networks have been obfuscated. Table 2 contains the demographics we considered in our study.

Figure 2 shows strong correlations between 113 household demographics and viewership for 25 networks. To explain this relationship in further detail, we performed a principal components analysis (PCA) on these percentages. PCA is a dimension reduction technique that allows view the highly correlated data in Figure 2 with just a few uncorrelated variables, or principal components (Shaw, 2003).

Table 2. Table of demographics we considered as a partial proxy for network. Note that Single is not the opposite of married, as there exists households containing unmarried couples. STBs labeled as unknown were removed from the principal components analysis, but are included in the retention score model

Gender	Kids	Married	Single	Age
Male	Yes	Yes	Yes	18-24
Female	No	No	No	25-34
Both	Unknown	Unknown		35-44
Unknown				45-54
				55-64
				65-74
				75 plus+

Figure 3 shows the cumulative percentage of variance for each principal component. As we can see, over 95% of the total variation is explained in the first three principal components. Figures 4 and 5 explore these three dimensions even further. Figures 4 plots each principal component versus the age group and split by presence of children and gender. From the first principal component we see clear variations due to age. However, there also exists a separation due to the presence of children. The second principal component also shows differences due to age, while the third component in particular shows the greatest amount of variation is due to gender. This is seen by the top two dashed lines, which correspond to households that contain either a single female or female head of household.

Figure 5 shows biplots of the first three principal components. Biplots are useful for overlaying the principal components (shown in black) with the rotated data (shown in gray) (Gabriel, 1971). To interpret a biplot, we focus on areas where the data and the principal components are somewhat aligned. For the biplot containing the first and second principal component, we see as strong correlation between older adults without children and viewership of cable news networks such as "News Channel 2" and "News Channel 1". In addition, we also see that younger adults with children tend to watch "Cartoon Network 2" and "Cartoon Channel 3". The latter observations match very well with our comments about the first and second principal components in Figure 4.

For the biplot containing the second and third principal components shows a relationship of gender and age to network viewership. Note that most of the data for females between the ages of 45 and 64 are in the upper left corner of the plot, relating to networks such as "Women's Network 1", "Women's Network 2", and "DIY Channel 2". In addition, households with males or mixed genders along the bottom of the figure show a strong correlation to the networks "Documentary Channel 2", "Sports Channel 1", and "Mixed Programming Channel 4".

For our demographics model we simply added the 113 different combinations from Table 2 along with the same list of features in the Basic model, which lead to 143 coefficients in the model. Applying the algorithm for principal components regression described above, we reduced the dimensionality to 125 coefficients, or 90% of the total variation. This is considerably more dimensions than the demographics-to-network PCA discussed in this section. However the PCA study described the viewership of networks, while our demographics model describes STB tune-out at ads.

Figure 2. This figure shows the percentage of time over the course of a month that a STB was tuned to a particular network (vertical axis) versus the demographic makeup of the household (horizontal axis). The darker area, the more time a particular household was tuned to that network. For example, older people without kids (upper right) tend to watch "News Channel 2" the most, while younger people with kids (upper right) tend to watch more "Cartoon Channel 2" and "Cartoon Channel 3". This plot shows 25 networks and 113 demographic groups

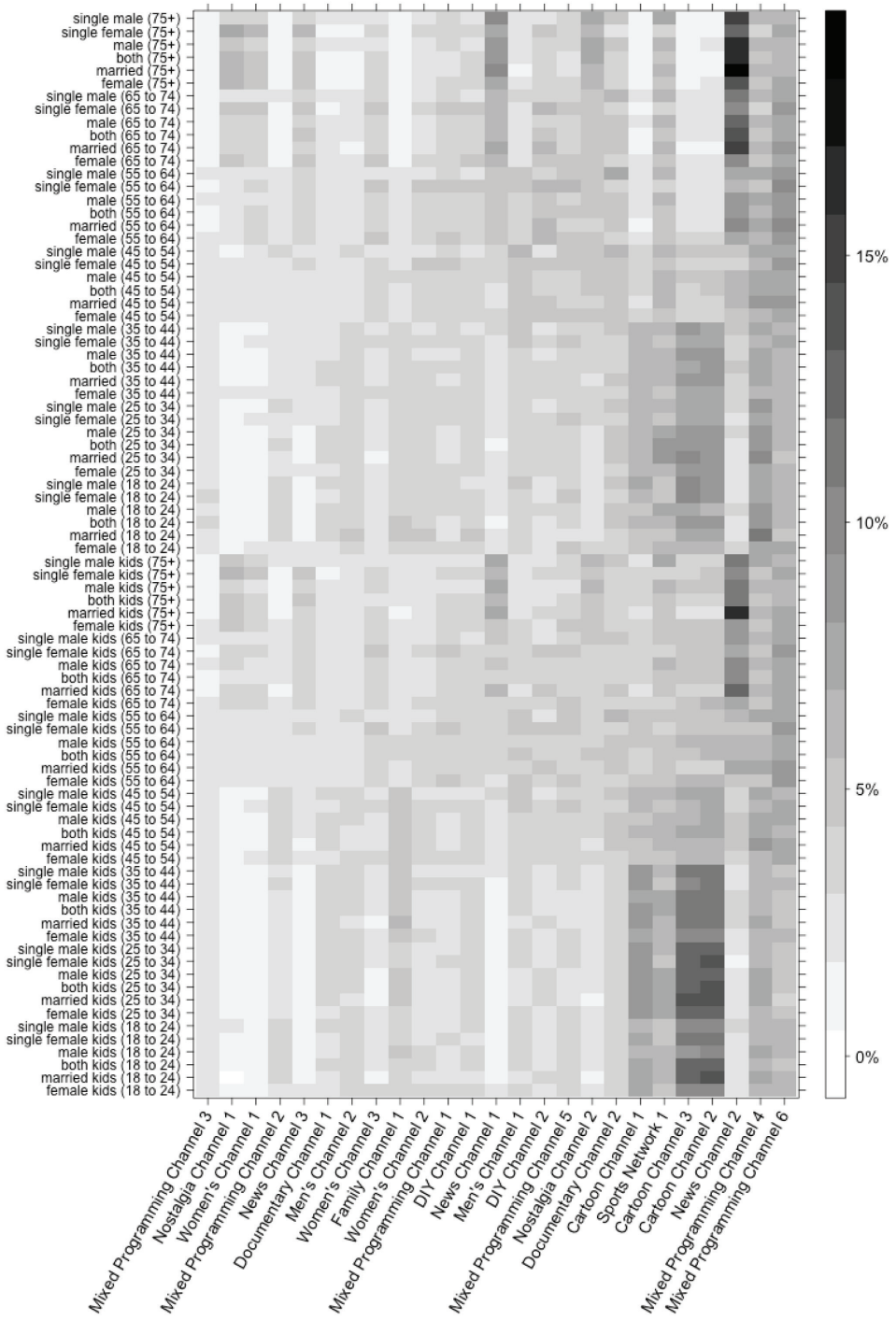
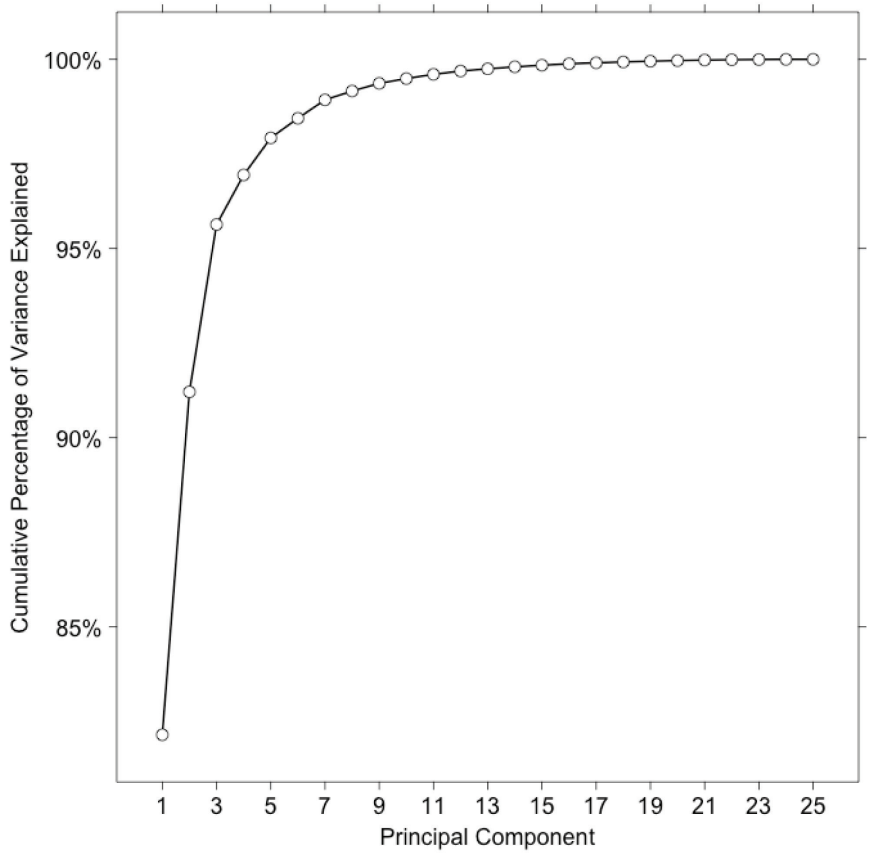


Figure 3. This plot shows the cumulative percentage of variance explained by 25 principal components. The first three principal components explain more than 95% of the total variation



Local User Behavior Model

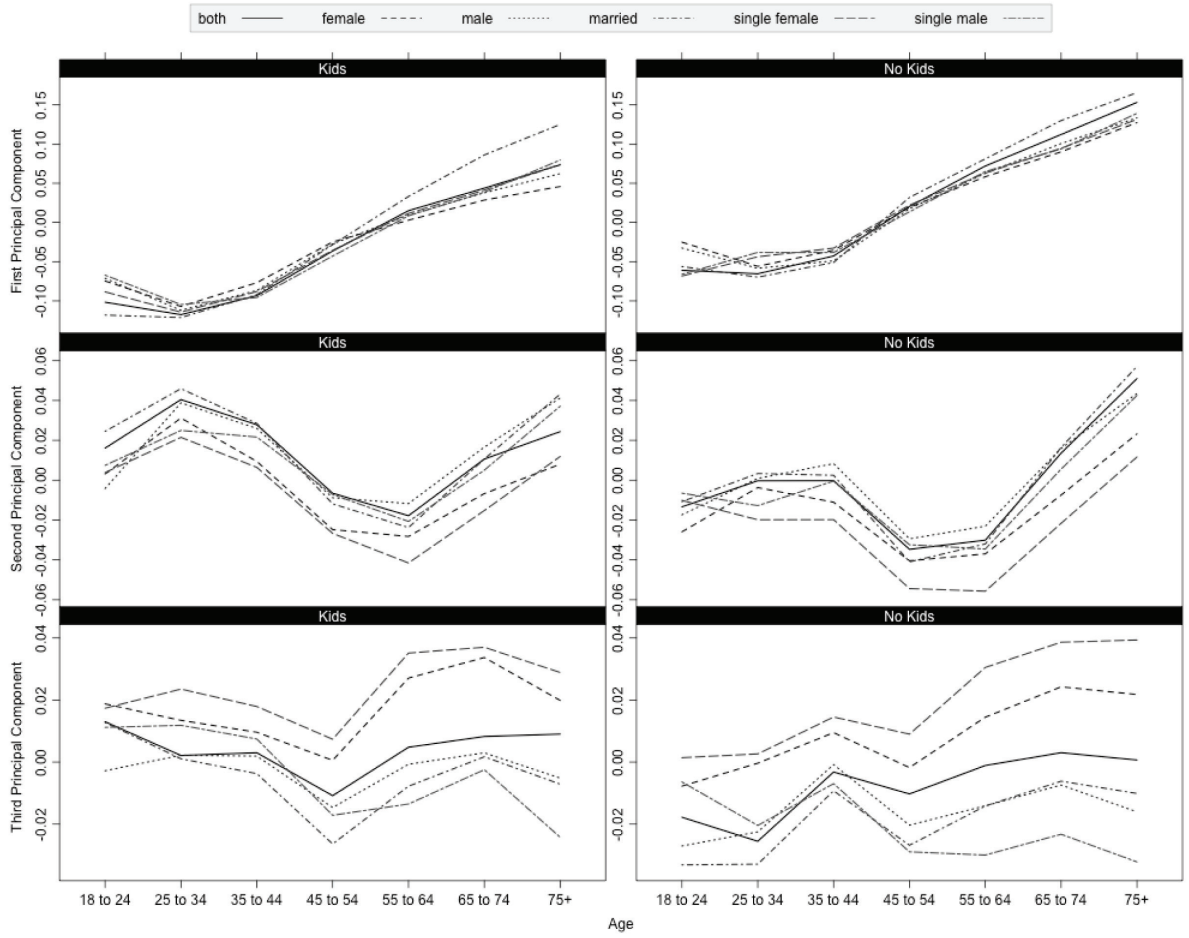
In this model we monitor STB behavior up to the point of a particular ad. Our hypothesis is that users how have a channel change closer to an ad insertion are more likely to tune away from the ad than users have not had a recent channel change. In essence we are trying to separate "active" users from "passive" users based on recent events.

For this model we define the following features:

1. LastEvent – whether the last event occurred in the previous minute, 10 minutes, 30 minutes, 60 minutes, or greater than 60 minutes.
2. NumberEvents – the number of channel changes in the 60 minutes prior to tuning away from an ad. STBs with 5 or more events are grouped in the same category.
3. MF – Male, Female, Both, or Other.
4. AGE – Over 65, Under 65, or Unknown
5. Network – Limited to the top 25 viewed networks.
6. Ad duration (in seconds).

The first two features separate the active viewers from the passive ones, which helps us predict who is more likely to tune away from an ad. Figure 7 shows the distribution of IAR for the viewers who had an event within the last hour ("Active")

Figure 4. Plots of first 3 principal components vs. age group, split by gender (lines) and presence of children (panels). The first principal component (top row) varies mostly by age differences and presence of children. The second principal component (middle row) reveals additional variation in age. The third principal component (bottom row) shows that women tend to watch different shows from other gender groups



versus the IAR for viewers who had no events in the past hour. Figure 8 is a similar plot for the NumberEvents feature. These two plots clearly show a dependence on how much a viewer is actively watching TV prior to an ad's airing.

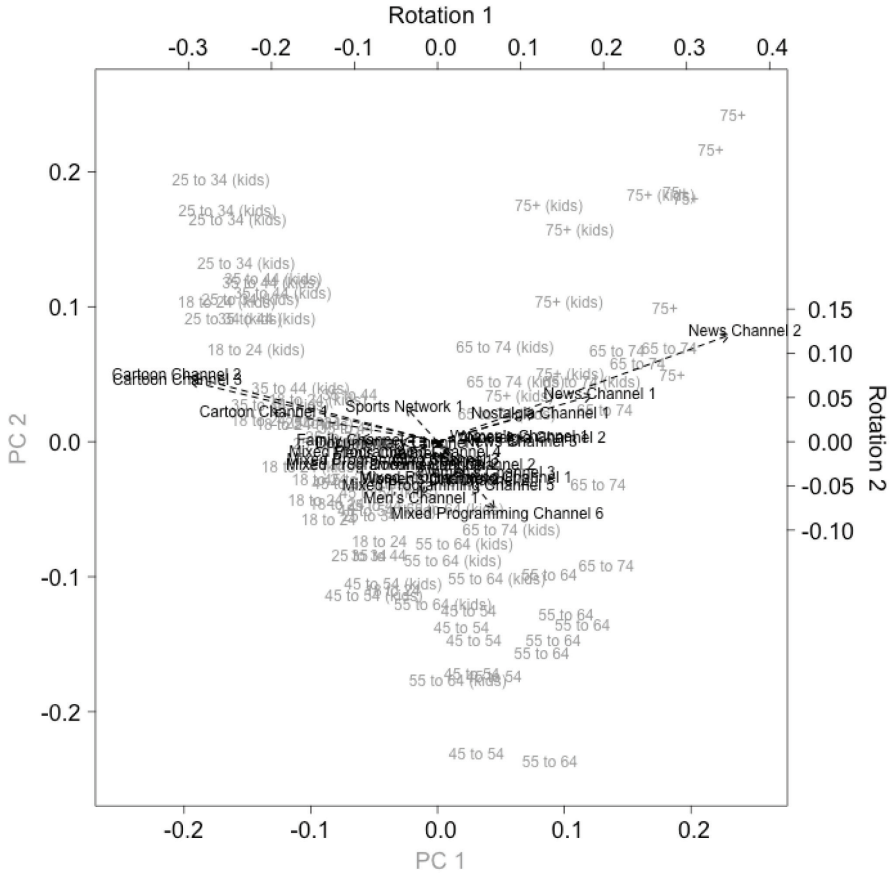
We also added two demographic features to help predict IAR: MF and AGE. We chose these two because of the principal components analysis discussed earlier in the chapter. However, due to the memory constraints of the R language we limited the levels of each demographic.

Figure 9 shows the distribution of IAR for the two demographic groups in our model. For the gender demographic we see that men are less tolerant of ads and tend to have a lower IAR than women. Similarly, older adults tend to watch more ads than younger adults.

A Machine Learning Approach

The machine learning approach relies on our proprietary logistic regression implementation. This

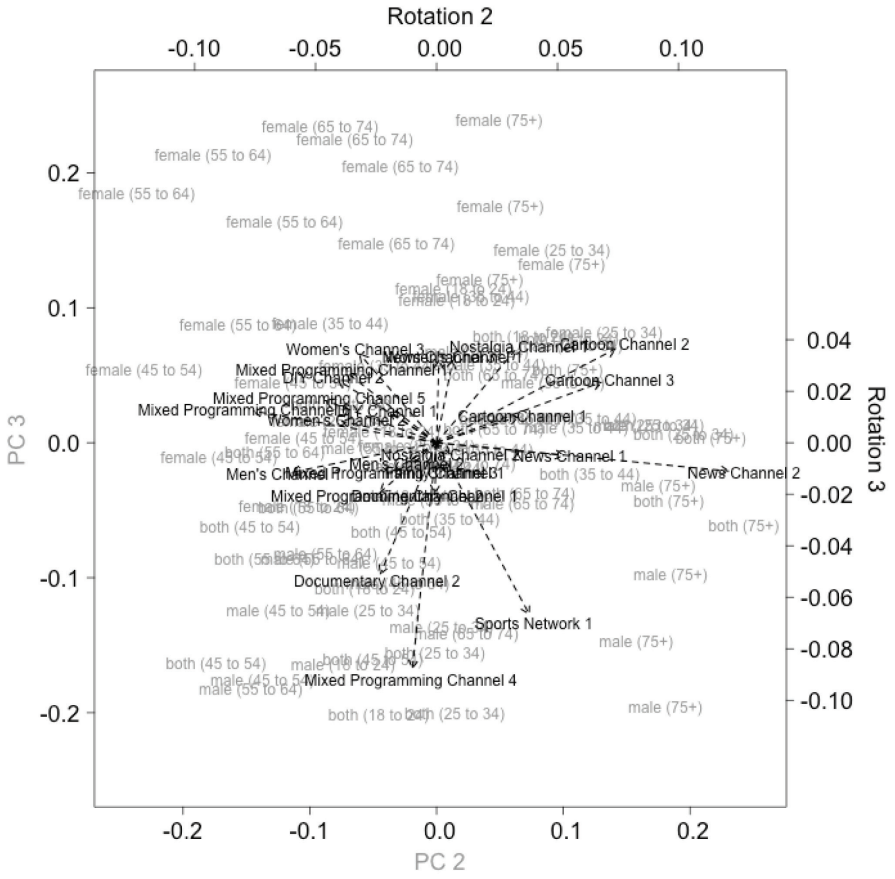
Figure 5. This figure shows biplot of principal components 2 vs. 1. From this figure we see that older adults with no children present tend to watch "News Channel 2", "News Channel 1", and "Nostalgia Chanel 1", while younger adults with children tend to watch more "Cartoon Channel 2" and "Cartoon Channel 3"



model greatly expands upon the previous models but does not suffer from the memory limitations of R. The main features included in this model are the following:

1. LN(Time to last event) – the natural log of the time to the last event. This is a continuous version of the "LastEvent" in the previous model.
2. DurationSec and sqrt(DurationSec) – the ad duration and its square root. The square root transformation controls for longer ads, which do not have the same effect on IAR as shorter ads.
3. The genre of the show where the ad was placed (e.g. adventure, comedy, etc.)
4. Five minutes from the end or beginning of show, which models the behavior that viewers tend to have higher tune-away rates at the beginning or ending of a show.
5. The show name (e.g. Phineas and Ferb, Without a Trace, etc.).
6. Network – Not limited to the top 25 networks as in the previous models. However, for comparison purposes we filtered the results

Figure 6. This figure shows biplot of principal components 3 vs. 2. From this figure we see women from aged 45-74 tend to watch "Women's Channel 3", "DIY Channel 2", and "Mixed Programming Channel 5"



to the same airings as the data used in the first three models.

7. Household demographics, including those in Table 2, as well as ethnicity and occupation.
8. STB time zone and geographic location.
9. Day part (see Table 1).
10. Weekday vs. weekend.
11. Stickiness, which is defined as the total percentage of times a STB tuned away from an ad over a month of time.

All these features are used in the model. Unlike the previous algorithms, this implementation looks at more than just the main effects. Because this approach is extremely memory efficient, we also investigated higher order interactions.

The previous algorithms have no mechanism for updating coefficients that may become stale over time. However, the machine-learning approach automatically recalculates the model coefficients with the introduction of new data, while down weighting the contribution of older data to the estimation.

PREDICTIVE POWER

To compare the models introduced in the chapter, we need a metric that demonstrates their ability

Figure 7. This figure shows the density of IAR for active viewers versus passive viewers. Active viewers changed the channel in the hour prior to an ad, while passive viewers did not

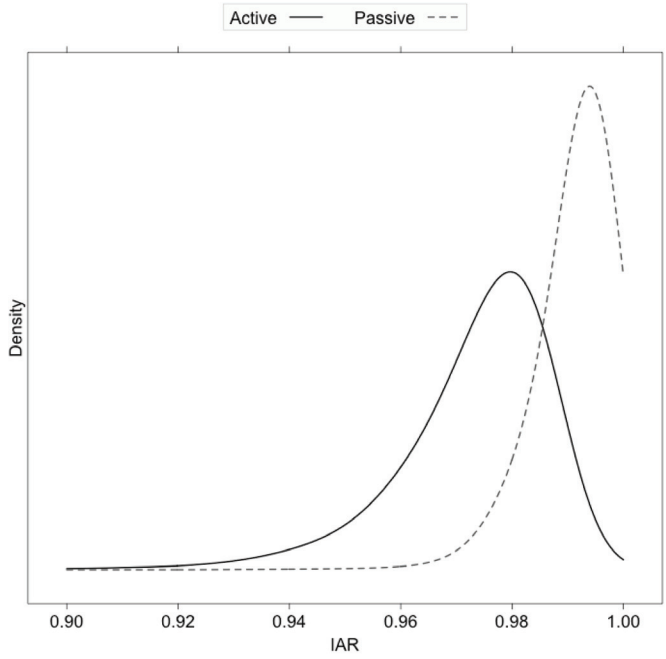


Figure 8. This figure shows the IAR for viewers as number of channel changes increased within the past hour. The plot is truncated for to 5 or more events

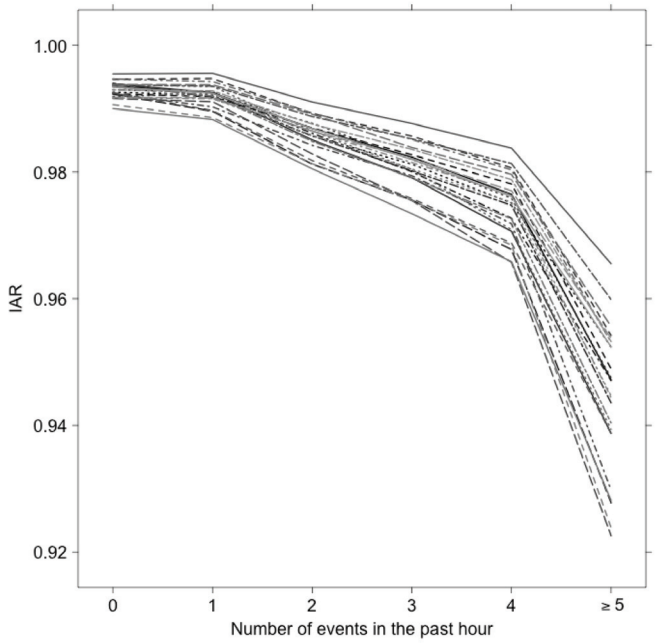


Figure 9. Density plots of IAR by gender and age. Men tend to be tune away from ads more than women. We also see that adults under 65 tend to tune away from ads than older adults

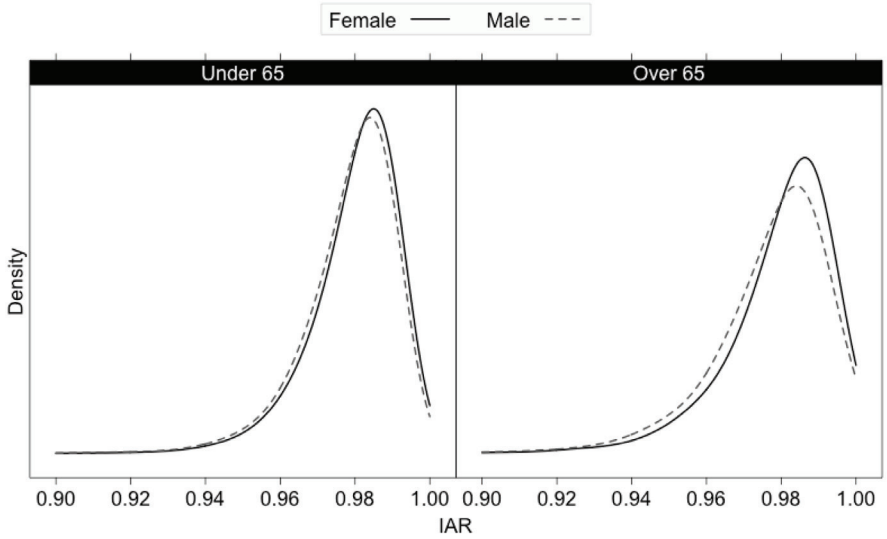


Table 1. Category definition for Day Part. Most networks start and end their broadcasting day at 5 am Eastern time

Day Part	Time
Morning	5am to 10am
Daytime	10am to 2pm
Late Afternoon	2pm to 5pm
Evening	5pm to 8pm
Prime	8pm to 12pm
Overnight	12pm to 5am

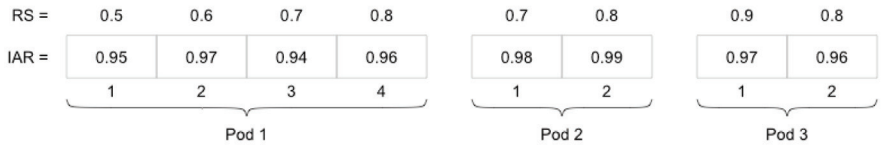
to produce accurate retention scores. In this section we discuss a metric that measures a models capability to predict retention scores. For all the algorithms we trained our models on the first 75% of the airings in October 2009 and predicted retention scores the remaining 25% of the airings.

Algorithm

The main metric we use to compare retention score models is called *predictive power*. A model with high predictive power accurately predicts ad rankings based on current retention scores. We interpret this metric as the total percentage of ads correctly sorted by our retention score algorithm. The following provides the algorithm for our metric:

1. Using a training dataset, build a model to predict IAR.
2. With the fitted model, obtain predicted IARs on a test dataset.
3. Aggregate the observed and predicted IARs to the airing level (i.e. each row in the test dataset represents a single ad placement).
4. For each airing, use (3) to compute a residual.
5. For each creative, obtain a retention score (RS) based on (4).
6. For each airing, determine all other airings within the same commercial break, or *pod*.
7. For each pod, compute the pairwise differences of observed IAR as well as the pairwise differences for the predicted retention scores for each airing.
8. For all ad pairs whose difference in RS is in the interval $(\Delta, \Delta + 0.01)$, where $0 < \Delta <$

Figure 10. Illustration of how we estimate predictive power. Comparisons of ads are only made within a commercial break, or pod, since all factors are essentially the same at this level



1, determine the proportions of those pairs whose corresponding observed IAR agrees with their respective RS.

9. Define the *predictive power* as the weighted mean of all proportions determined in the previous step, with weights determined from the total number of ad pairs in each interval.

Illustration Of Predictive Power

Figure 10 illustrates the algorithm for computing predictive power. In this example, we have eight airings distributed among three pods. The numbers in each box are the observed IAR for each airing, while the numbers above the box are the retentions scores (RS) of the creative. The RS values are determined from one of the models. In Pod 1, with four airings, we make six comparisons, of which three of the comparisons have the same sign while three have opposite sign. All comparisons for each pod are shown in Table 3.

Model Comparison

The predictive power for the four models discussed in this paper is shown in Table 4. Their relative performances can be seen more easily in Figure 11. The model that faired the best is the "Local Events + Demographics" model with 76.5% of all ad pair correctly sorted. However, this model is not as scalable as the "Machine Learning" model and thus loses some of its attractiveness. The Machine Learning model is the first of its kind that we tried and is easily expandable to include other features.

Table 3. Table of IAR and RS comparisons for the example shown in Figure 10. The predictive power is the percentage of comparisons where IAR agrees with RS. Comparisons agree when ΔRS and ΔIAR have the same sign. For this example, there are five comparisons that agree (Y) and three comparisons that do not agree (N). The predictive power for this example is 5/8 = 62.5%.

	Pod 1			Pod 2			Pod 3		
	ΔRS	**ΔIAR**	**Agrees**	**ΔRS**	**ΔIAR**	**Agrees**	**ΔRS**	**ΔIAR**	**Agrees**
2-1	0.1	0.02	Y	0.1	0.01	Y	-0.1	-0.01	Y
3-2	0.1	-0.03	N						
3-1	0.2	-0.01	N						
4-3	0.1	0.02	Y						
4-2	0.2	-0.01	N						
4-1	0.3	0.01	Y						

Table 4. Predictive power for each of the models. The metric is interpreted as the percentage of all ad pairs within a pod that are correctly sorted by their respective retention scores. The best model is the "Local Events + Demographics" model followed by the "Machine Learning" algorithm

Model	Predictive Power
Basic	72.8%
Demographics	65.8%
Local Events + Demographics	76.5%
Machine Learning	73.8%

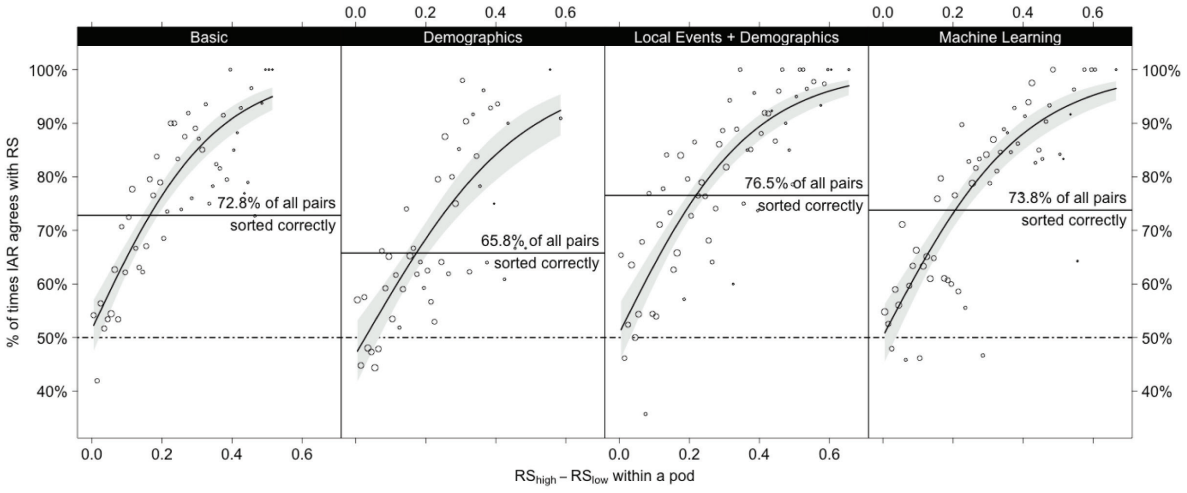
CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this chapter we have introduced three models for predicting ad retention on TV. However, we still have room for improvement, of which the machine-learning algorithm has the greatest potential because of its scalability. This is an area ripe for further innovation. Many new (and perhaps yet-to-be-developed) machine learning algorithms could be applied to this problem, and ever-greater quantities of data can be fed to such models to produce more accurate predictions of future audience behavior.

In the first decade of the twentieth century, "offline" media such as TV, radio, and print were thought to be in conflict with emerging online opportunities, such as Web-based display advertising and paid search ads. In the coming decade, however, the distinction between online and offline is likely to blur considerably. Content will often be available in many forms: traditionally offline media such as newspapers will be read on the Web, while online videos from sites like YouTube will be downloaded to be watched later on possibly disconnected devices.

The online-offline division will soon be replaced by a new distinction, between measured and unmeasured. Data from set-top boxes and similar sources will provide a degree of measurability and accountability to TV and other "offline" advertising that had previously only been available online, and allow the traditional advertising world

Figure 11. Comparison of the four retention score models in terms of predictive power. The curves are logit trend lines with 95% confidence bands. The size of each point is proportional to the number of ad pairs in the denominator of the percentage



to adopt the quantitative techniques pioneered in online settings. We have described one such application in this chapter: creating ad quality scores for TV ads similar to those first developed for paid search advertising.

Because what can be measured can also be analyzed and optimized, well-measured media will develop a natural efficiency advantage over unmeasured. Advertising budgets will naturally flow to those media that are best able to generate and capitalize on data. Successfully applying the quantitative lessons from online advertising will become essential for the survival of all advertising media.

REFERENCES

Ana, M. A., Manuel, E., & Mariano, J. V. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, (50): 1905–1924.

Bachman, K. (2009). Cracking the set-top box code. *AdWeek*, August 17, 2009. Retrieved December 22, 2009, from http://www.adweek.com/aw/content_display/news/e3i8fb28a31928f66a5893aa9825dee83f2.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.1-3. http://www-stat.stanford.edu/~hastie/Papers/glmnet.pdf

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrics*, (58): 453–467.

Google. (2009). What is 'Quality Score' and how is it calculated? Retrieved December 23, 2009, from http://adwords.google.com/support/aw/bin/answer.py?hl=en&answer=10215.

Mandese, J. (2009). Research Rivals Nielsen, comScore, Rentrak, TiVo, TNS Agree to Pool TV Set-Top Data. Retrieved January 14, 2009, from http://www.mediapost.com/publications/?fa=Articles.showArticle&art_aid=105217

R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting Clicks: Estimating The Click-Through Rate For New Ads. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, (pp. 521–530). New York, NY, USA.

Shaw, P. (2003). *Multivariate statistics for the Environmental Sciences*. Hodder-Arnold.

The Nielsen Company. (2009a). The Evolution – and Revolution – of Meters. Retrieved December 22, 2009, from http://www.nielsenmedia.com/lpm/history/History.html

The Nielsen Company. (2009b). Nielsen TV Audience Measurement. Retrieved December 22, 2009, from http://en-us.nielsen.com/tab/product_families/nielsen_tv_audience

Webster, J. G., Phalen, P. F., & Lichty, L. W. (2006). *Ratings Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Yahoo. (2009). Writing ads: Ad Quality and Quality Index. Retrieved on December 23, 2009, from http://help.yahoo.com/l/us/yahoo/ysm/sps/articles/writing_ads4.html.

Zigmond, D., Dorai-Raj, S., Interian, Y., & Naverniouk, I. (2009). Measuring Advertising Quality on Television: Deriving Meaningful Metrics from Audience Retention Data. *Journal of Advertising Research*, *49*(4), 419–428. doi:10.2501/S0021849909091090