Advertising on YouTube and TV: A Meta-analysis of Optimal Media-mix Planning

Georg M. Goerg, Christoph Best, Sheethal Shobowale, Nicolas Remy, Jim Koehler

Google Inc.

December 3, 2015

Abstract

In this work we investigate under what circumstances a TV campaign should be complemented with online advertising to increase combined reach. First, we use probabilistic models to derive necessary and sufficient conditions. We then test these optimality conditions on empirical findings of a large collection of TV campaigns to answer two important questions: i) which characteristics of a TV campaign make it favorable to shift part of its budget to online advertising?; and ii) if it should shift, how much cost savings and additional reach can advertisers expect? First, we use classification methods such as linear discriminant analysis, logistic regression, and decision trees to decide whether a TV campaign should add online advertising; secondly, we train linear and support vector regression models to predict optimal budget allocation, cost savings, or additional reach. To train these models we use optimization results on roughly $26,000$ campaigns. We do not only achieve excellent out-of-sample predictive power, but also obtain simple, interpretable, and actionable rules that improve the understanding of media mix advertising.

1 Introduction

Online media has become increasingly important for advertisers to complement their television (TV) ads. In order to make best use of their advertising budget, advertisers need to figure out the optimal budget allocation between TV and online media. Jin et al. (2013) introduce probabilistic models to estimate combined TV & online ads effectiveness and use these models to find the optimal budget split which maximizes reach. This optimization yields two important results: the optimal *shift* of budget to online media, as well as the expected additional reach after adding online media – also referred to as *extra reach*.

In this work we apply these optimization algorithms to a large collection of campaigns with a variety of target demographics, budgets, maximal TV reach, etc. We then train predictive models to find those characteristics of a campaign that make it more likely to benefit from shifting advertising budget; and if it should shift, we predict the attained extra reach and cost savings. Such prediction models can be used by advertisers for simple, yet predictively powerful rules to find TV campaigns that would most likely benefit from adding online advertising.

In Section 2 we study bivariate reach surfaces and show that marginal cost per reach is the single most important metric that determines the optimal budget allocation. Section 3 summarizes the TV data and the optimization results. In Section 4.2 we train classification models to decide whether a campaign should shift budget to online advertising. Section 4.3 presents several regression models with excellent predictions of optimal shift, extra reach, and cost savings. Finally, Section 5 summarizes the main findings and discusses future work. Detailed analytical derivations and proofs can be found in Appendix B.

2 Methodology

Before going into detailed theoretical analysis of the budget allocation problem, we present terminology and notation used throughout this work.

2.1 Notation and terminology

Table 1 lists the most important abbreviations; notation for derivatives can be found in Appendix A.

Let an advertiser have a total budget B (or, equivalently, cost C) and let them buy $I \geq 0$ impressions of advertising content. Rather than on the absolute impressions level, we use

the industry standard of gross ratings points (GRPs) to measure the size of a campaign. GRPs are impressions I normalized by the total population P times 100; $G = \frac{I}{P} \cdot 100$. For example, a campaign size of 200 GRPs means that - on average - two impressions per person are shown. In order to evaluate the economical efficiency of a campaign, it is common to consider *cost per point (cpp)*, which is the average cost per GRP, $cpp = \frac{C}{G}$. Since advertisers usually buy a set of GRPs for certain price, *cpp* is constant as a function of GRPs. We thus often use budget and GRPs interchangeably to refer to the size of a campaign.

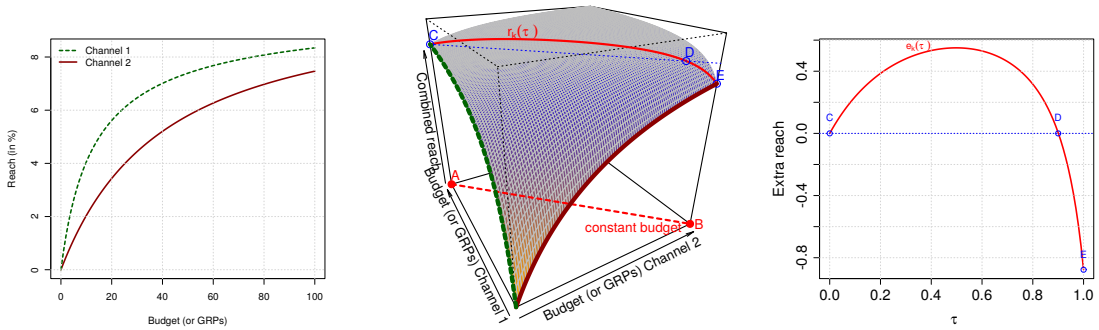
Advertisers want to know how many *different* people they can reach with a given number of impressions (budget). Let $R_k \leq P$ be the total number of different people reached at least k times.[1] Again, we usually use relative reach, $r_k = \frac{R}{P}$, to make campaigns for different target audiences comparable. We typically view $k+$ reach as a function of $GRPs$ or cost, $r_k = r_k(g) = r_k(c)$. For example, Figure 1a shows two single-channel reach curves, with different shape and slope of each curve. In particular, note that channel 2 (dark red, solid) has lower reach than channel 1 (green, dashed) at the same GRP level. However, at the last observed data point (here: $GRP = 100$), channel 2 has a higher *marginal* reach than channel 1, i.e., the curve has a larger slope at $GRP = 100$. This is important for advertisers as it means that it is more reach-efficient to show the 101^{st} GRP on channel 2.

In the theoretical analysis below the *cost per effective reach point* (Rossiter and Danaher, 1998), $cperp = \frac{C}{R_k}$, will play the principal role in determining the optimal shift. Contrary to cpp, cperp is increasing with the size of a campaign since the reach curve is concave as a

[1]The specific choice of k depends on the interest of advertisers. For example, the industry standard in the United States is 3+ reach; in Germany, it is 1+ reach.

Symbol	Description	Variable type	Computation
$B = C$	total budget = cost of an advertising campaign	currency	
I	content impressions (ad, video, ...)	count	
P	total (target) population	count	
R_k	$k+$ absolute reach, i.e., the absolute number of different people who have seen the content	count	
G	GRPs = gross rating points	%	$\frac{I}{P} \times 100$
r_k	$k+$ relative reach, i.e., percentage of people who are reached at least k times	%	$\frac{R_k}{P} \times 100$
frequency	average number of times user sees an impression (among those who have seen it at least k times)		$\frac{GRPs}{r_k}$
cpp	cost per point	currency	$\frac{C}{GRPs}$
cperp	cost per effective reach point	currency	$\frac{C}{r_k}$

Table 1: Abbreviations and notation



(a) Reach curve for single channel **(b)** Combined reach surface **(c)** Extra reach as a function of budget share τ

Figure 1: Reach in the single channel and the two-channel scenario.

function of GRPs. It is exactly this non-linear increase that determines the optimal budget allocation.

2.2 Bivariate probability surface

For a campaign that uses multiple advertising channels combined reach is a function of the multidimensional budget vector $(B_1, \ldots, B_N)$. In this work we consider the $N = 2$ channel scenario (e.g., TV and online media), where combined $k+$ reach, $r_k(B_1, B_2)$, can be represented as a surface along the two channel dimensions (Fig. 1b). Here each point on the surface represents the proportion of the target audience that has been reached on channel 1 *or* channel 2 as a function of budget on each channel. At the boundary of $B_1 = 0$ or $B_2 = 0$ it reduces to two single-channel reach curves in Figure 1a.

2.2.1 Modeling reach as a probability

Like Jin et al. (2013), we model relative $k+$ reach as the probability that a randomly drawn person u sees at least k impressions, i.e.,

$$r_k = \mathbb{P}\left(I_u \geq k\right), \tag{1}$$

where I_u are the number of impressions of person u. Such a probabilistic view allows us to use parametric probability models to compute entire reach curves (see e.g., Jin et al., 2012; Goerg, 2014; Cannon et al., 2002).

2.3 Reach optimization at fixed budget

Jin et al. (2013) consider two optimization scenarios: i) maximize combined reach, at constant budget; ii) minimize budget, at constant reach. By default, the constant budget (reach) is the historically attained budget (reach) on channel 1. For analytic derivations we restrict ourselves to the "maximize reach, constant budget" case; similar derivations can be obtained for "minimize budget, constant reach". In the applications section we again consider both scenarios and provide classification and regression models for each.

In Figure 1b the fixed budget constraint is shown as the dashed, red line in the (B_1, B_2) plane. At constant budget combined reach reduces to a one-dimensional curve along the surface (red, solid). It can thus be parametrized by the one-dimensional variable τ, which represents the budget share of channel 2: let $B_1(\tau) = (1-\tau)B$ and $B_2(\tau) = \tau B$. For τ moving from 0 to 1, budget allocation moves from point A to B, and combined reach

$$[0,1] \ni r_k(\tau) = r_k^{1\&2}\left((1-\tau)B, \tau B\right) \tag{2}$$

moves from C to E. The additional $k+$ reach of a media mix compared to the channel 1-only campaign ($\tau = 0$) equals

$$e_k(\tau) = r_k(\tau) - r_k(0) \in [-1, 1], \tag{3}$$

where e_k stands for the *extra* $k+$ reach (Fig. 1c).

In the example from Fig. 1, 100 GRPs on channel 1 yield higher reach than on channel 2. However, as the red $r_k(\tau)$ curve along the surface shows, combined reach achieves its maximum at $\tau^* \approx 0.5$ (see also Fig. 1c). This means that moving 50% of advertising budget from channel 1 to channel 2 would increase the combined campaign reach compared to single-channel advertising.

2.4 Optimality conditions for maximizing combined reach

Remark For better readability, we drop the subscript k in r_k for the remainder of Section 2.4. This will avoid confusion with derivative subscripts $r_x := \frac{\partial}{\partial x} r(x, y)$ (see also Appendix A).

The optimal budget allocation τ^* occurs either at the single-channel boundary, $\tau^* = 0$ or

$\tau^* = 1$, or where

$$\frac{\partial}{\partial \tau} r(\tau) = r'(\tau) = 0. \tag{4}$$

Lemma 2.1 *A two-channel campaign achieves maximum combined reach at constant budget when*

$$\frac{\partial}{\partial x} r(x(\tau^*), y(\tau^*)) = \frac{\partial}{\partial y} r(x(\tau^*), y(\tau^*)), \tag{5}$$

or at the boundary, $\tau^* \in \{0, 1\}$.

Proof In Appendix B.

Lemma 2.1 formally shows that budget should be shifted from channel 1 to channel 2 as long as the marginal increase reach on channel 2 (y) is greater than on channel 1 (x).

Without any modeling assumptions about the reach curves and surfaces, (19) can not be simplified any further. However, for the single-channel case ($\tau = 0$) we obtain a simpler condition.

Corollary 2.2 *A single-channel campaign should add another channel if*

$$r_y(B, 0) > r_x(B, 0), \tag{6}$$

or equivalently

$$C_y(r, 0) < C_x(r, 0), \tag{7}$$

where $C_y(r, 0) = \frac{1}{r_y(B,0)}$ *is marginal cost per reach of channel 2 (y) at maximum reach (analogous for* $C_x(r, 0)$*).*

Lemma 2.1 and Corollary 2.2 show that – in theory – the sole predictor of shift versus no shift is the difference between the marginal cost per reach of channels 1 and 2. Figure 1a illustrates this condition (6): if the campaign on channel 1 has already reached the flat part of the curve for large budget ($r_x(B, 0) \approx 0$), then it is more likely to be a good candidate for shifting (since $0 \approx r_x < r_y$).

In general, TV-only advertisers do not (yet) have information about reach and reach efficiency for the online channel. Thus for the empirical analysis and predictive modeling in Section 4.2 and 4.3 we only use data from the one-dimensional TV reach curve, $r^{TV}(g)$, $g \in [0, G]$.

2.5 Estimating marginal cost per reach

So far we have considered reach as a function of GRPs and cost. It is useful to consider the inverse relation, $C(r_k)$, cost as a function of reach. As shown above, marginal cost per reach is the sole indicator whether a campaign should shift or not.[2] Advertisers, however, often do not know their *marginal* cost, but only their *average* (or total) cost. Goerg (2014) presents methodology to estimate the entire reach curve using only total GRPs and reach. The functional form of this reach GRP curve is

$$r_k(g) = \frac{G^{total} \cdot r_k^{total} \cdot g}{(G^{total} - g) \cdot \frac{1}{\iota_k} \cdot r_k^{total} + g \cdot G^{total}}, \tag{8}$$

where ι_k is a nuisance parameter that represents the expected number of total impressions for the first person to see k impression.[3] Trivially, $\iota_1 = 1$ since the first impression must go to the first person. For $k > 1$, they approximate it with $\iota_k \approx k + \log_2 k$.[4]

The marginal reach per GRP at $g = G^{total}$ equals

$$r_k'(g = G^{total}) = \frac{1}{\iota_k} \left(\frac{r_k^{total}}{G^{total}} \right)^2. \tag{9}$$

And consequently the marginal cost per reach at total GRP and reach has a surprisingly simple form of

$$C'(r_k) = \iota_k \cdot cperp^{total} \cdot \text{frequency}^{total}, \tag{10}$$

where $cperp^{total}$ is cost per effective reach point at the total size of the campaign.

The regression and classification models in Section 4.2 and 4.3 show that (10) does indeed provide excellent predictions.

Number of campaigns (ignoring demo targeting)	2,914
Number of demographic groups	9
Number of all analyzed campaigns (by demo)	26,222
Start date	2015-01-01
End date	2015-09-30
Effective frequency (k+ reach)	3
YouTube Watchpage cost per mille (cpm) (in USD)	20
Maximum possible shift	100%

Table 2: Control settings for optimization.

3 Data summary

For the remainder of this work we investigate the two channel scenario for TV (channel 1) and YouTube (channel 2). The analysis is based on 2,914 quarterly TV campaigns in the US from 2015-01-01 to 2015-09-30. Each campaign was optimized for 9 different target demographics split by age and gender. We further restricted the campaigns to only those that had at least 100 GRPs per quarter for all demographics. This yields a total of 26,222 analyzed campaigns for this meta study. The TV campaign data used in the optimization results and this meta study is based on Nielsen's Cross Media Panel (Nielsen Solutions, 2013) and Nielsen's Monitor-Plus.

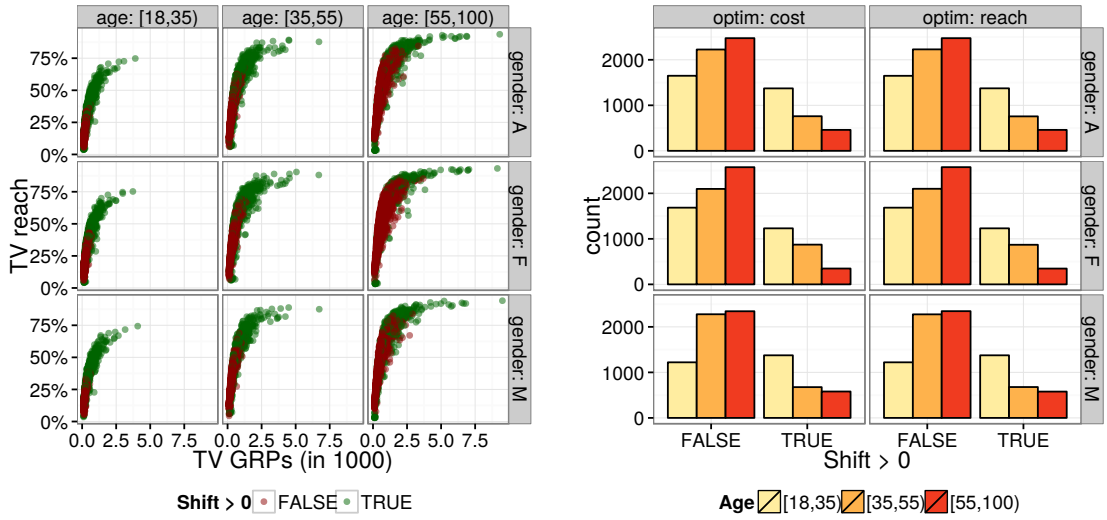
The optimization results for the optimal media mix between TV and YouTube depend on several control settings (Table 2). Changes in these controls will, in general, give different results.

Computations and figures were done in R (R Core Team, 2014); tables were generated with the *stargazer* R package (Hlavac, 2014).

[2]The derivations above assumed 1+ reach. For k+ reach with $k > 1$ derivations become a bit more cumbersome. While the relationship between shift and marginal cost is not as direct, we still use marginal cost as a proxy that determines shift and extra reach.

[3]In the rare (empirical) case that $G^{total} < \frac{1}{\iota}\frac{r_k^{total}}{1-r_k^{total}}$ the reach curve estimate in (8) must be replaced with $r_k(g) = \frac{g \cdot r_k^{total}}{G^{total}+(g-G^{total})\cdot r_k^{total}}$. See Goerg (2014) for details.

[4]Note that this is an approximation and obtaining exact values for ι_k for $k > 1$ remains a task for future work.



(a) GRP reach curve: every point represents the total GRP and reach of a historical TV-only campaign.

(b) Distribution of positive shift versus no-shift campaigns.

Figure 2: TV plans in the TV-only plan and their likelihood to shift (green).

3.1 EDA

In most scatterplots each point represents one campaign, and many figures are split by age group and/or gender. Interpreting the demographic groups is straightforward, e.g., A[18,35) refers to adults from 18 to 34 years old; F[35,55) refers to females from 35 to 54 years old; M[55,100) refers to males from 55 to 99 years old.

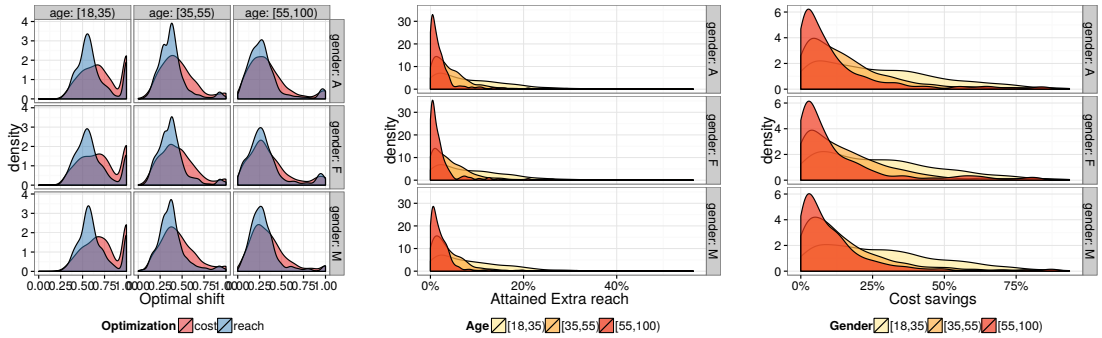
Figure 2a shows TV reach as a function of GRPs for the TV-only plan and Fig. 2b shows the frequency of positive shift campaigns split by gender and optimization method. They show two expected patterns: first, larger campaigns are more likely to shift; secondly, adding online advertising is more beneficial for campaigns with younger (and male) audiences.

Figure 3 and Table 3 summarize the three optimized metrics (shift, reach, savings).[5] Overall about 29% of TV campaigns would benefit from online advertising. This proportion varies across demographic targets with the lowest proportion (12%) for F[55,100), and highest (53%) for M[18,35) (see also Figure 2b). Of those campaigns that do shift, the average shift (for reach optimization) is 50% with an average attained extra reach of 7 percentage points

[5]Recall that the shareshift methodology (Jin et al., 2013) allows to maximize reach (at constant cost) or minimize cost (at constant reach). In this manuscript these two scenarios are usually displayed separately; if such a distinction is missing in figures or tables, then *reach* optimization results are shown by default.

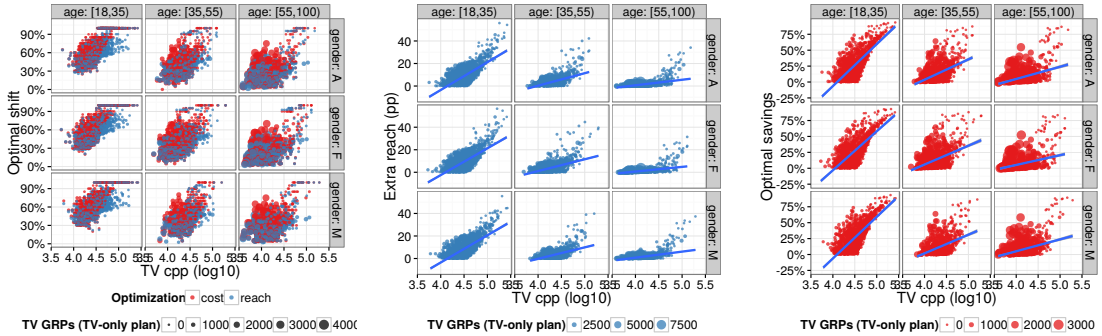
Table 3: Averages for optimization results grouped by demo and optimization scenario. Averages are reported in %; for 'Avg. extra reach' in percentage points (pp).

Demo	Optimization	P(shift >0)	Avg. shift	Avg. extra reach	Avg. cost savings
A[18,35)	cost	45.4	70.7	0	27.5
A[35,55)	cost	25.5	44.7	0	16.4
A[55,100)	cost	15.7	34.5	0	12.7
F[18,35)	cost	42.2	69.5	0	27.7
F[35,55)	cost	29.4	44.6	0	16.6
F[55,100)	cost	11.9	35.7	0	13.1
M[18,35)	cost	53.0	72.6	0	28.3
M[35,55)	cost	22.9	44.9	0	15.5
M[55,100)	cost	19.8	34.6	0	12.3
A[18,35)	reach	45.4	62.3	9.2	0
A[35,55)	reach	25.4	38.5	4.7	0
A[55,100)	reach	15.7	30.3	2.7	0
F[18,35)	reach	42.2	61.2	9.5	0
F[35,55)	reach	29.3	38.2	4.8	0
F[55,100)	reach	11.9	31.8	2.8	0
M[18,35)	reach	53.0	63.5	9.3	0
M[35,55)	reach	23.0	38.5	4.3	0
M[55,100)	reach	19.8	29.8	2.8	0



(a) Optimal shift **(b)** Attained extra reach **(c)** Optimal cost savings

Figure 3: Overview of optimization results (only when shift is beneficial).



(a) Optimal shift **(b)** Extra reach **(c)** Cost savings

Figure 4: Predictive relationship between cost per GRP (cpp) on TV and optimal shift, extra reach, and cost savings.

(pp). In the cost savings scenario, the average shift (of those that do shift) lies at 57% with an average cost savings of 22%.

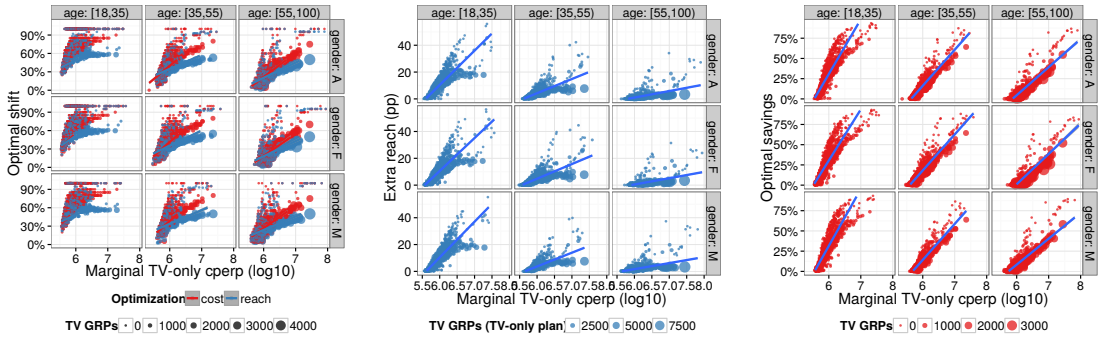
Figure 4 and 5 display a) optimal shift from cost per TV GRP (cpp), b) attained extra reach, and c) optimal cost savings, respectively, broken down by age and gender.

Figure 4a and 4b) show that it is difficult to predict optimal shift, while it does better at predicting extra reach. The main reason for this lies in the flatness of the extra reach curve (recall Figure 1c in Section 2), which makes the optimal shift (x-axis) very sensitive to noise, whereas the attained optimum (y-axis) is relatively stable. Table 4 shows the results of performing both an ordinary least squares (OLS) as well as a robust linear regression[6]

[6]We use the R function `rlm`. It performs linear regression, but instead of minimizing the sum of squared residuals, it minimizes a Huber-type loss of residuals, which is more robust to outliers. See Huber (1981) for an overview.

Table 4: Linear regression estimates for 'reach' optimization. ρ^2 is the squared correlation between data and fit (on original scale).

	Dependent variable: Extra reach (pp)		
	normal	*robust linear*	
	all	all	$shift > 0$ only
	(1)	(2)	(3)
Constant	-0.83^{***} (0.01)	-0.54^{***} (0.02)	-0.98^{***} (0.02)
log10.orig.tv.cpp	0.20^{***} (0.001)	0.13^{***} (0.004)	0.24^{***} (0.01)
age.group[35,55)	0.52^{***} (0.01)	0.46^{***} (0.02)	0.55^{***} (0.03)
age.group[55,100)	0.70^{***} (0.01)	0.51^{***} (0.02)	0.71^{***} (0.03)
genderF	0.004^{***} (0.001)	0.001^{***} (0.0002)	0.005^{***} (0.001)
genderM	-0.003^{***} (0.001)	-0.001^{***} (0.0002)	-0.004^{***} (0.001)
log10.orig.tv.cpp:age.group[35,55)	-0.12^{***} (0.002)	-0.11^{***} (0.004)	-0.13^{***} (0.01)
log10.orig.tv.cpp:age.group[55,100)	-0.17^{***} (0.002)	-0.12^{***} (0.004)	-0.17^{***} (0.01)
ρ^2	0.56	0.53	0.56
Observations	26,222	26,222	7,669
Note:			*p<0.1; **p<0.05; ***p<0.01



(a) Optimal shift **(b)** Extra reach **(c)** Cost savings

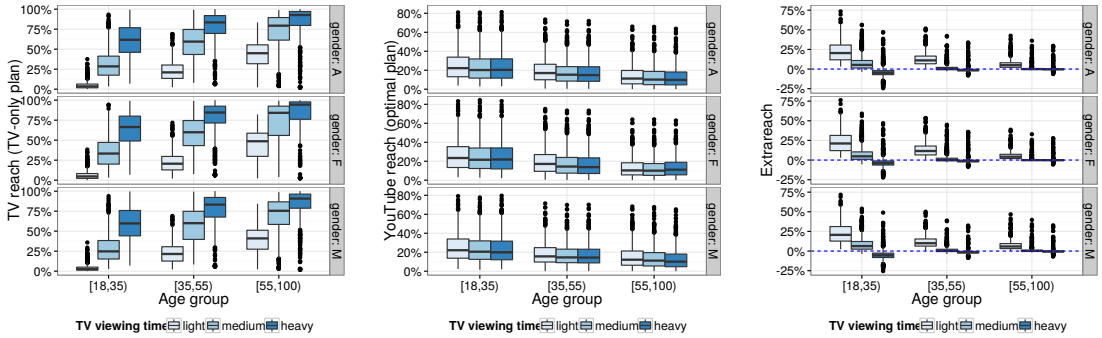
Figure 5: Marginal cost per effective reach point (cperp) on TV as a predictor of optimal shift, extra reach, and cost savings.

of extra reach on $\log_{10}(cpp)$ and demographic information. For example, the slope estimate for the (omitted) youngest age group for $\log_{10}(cpp)$ of $\widehat{\beta}_j = 0.2$ (see Table 4): this means that – all others equal – a campaign with a 10% higher TV cpp can expect additional reach of $\log_{10}(1.10) \cdot \widehat{\beta}_j \cdot 100 = 0.83$ percentage points (pp) in a media mix scenario; for the older [55,100) target demographic the increase is only 0.14 pp. The other coefficient estimates also confirm the findings from Figure 4 that differences are more pronounced across age groups than across gender. If any, then a campaign with a male target demo can expect slightly lower extra reach. Figure 5 confirms the theoretical findings in Section 2.4 that marginal cost per effective reach point (cperp) (see e.g., Rossiter and Danaher, 1998) is an even better predictor of optimal shift, extra reach, and optimal cost savings (see Section 4.3 for details).

3.2 TV viewing buckets

In order to better understand why online advertising can be more efficient at reaching new audiences, it is useful to consider TV viewing buckets. Here the population is split in three equally sized buckets, with different levels of TV viewing consumption. More precisely, we first computed 33.3% quantiles of TV viewing time per day, and then put each member of the population in its corresponding bucket: light, medium, and heavy.

Figure 6 compares TV, YouTube, and extra reach across the TV viewing buckets. By construction, TV has much higher reach for heavy TV viewers. YouTube, on the other hand, has a quite balanced reach across TV viewing buckets. As a result (Fig. 6c) extra reach is mostly due to large reach increases for light TV viewers, whereas medium and heavy TV viewers do not get as much additional reach (some campaigns even *decrease* their

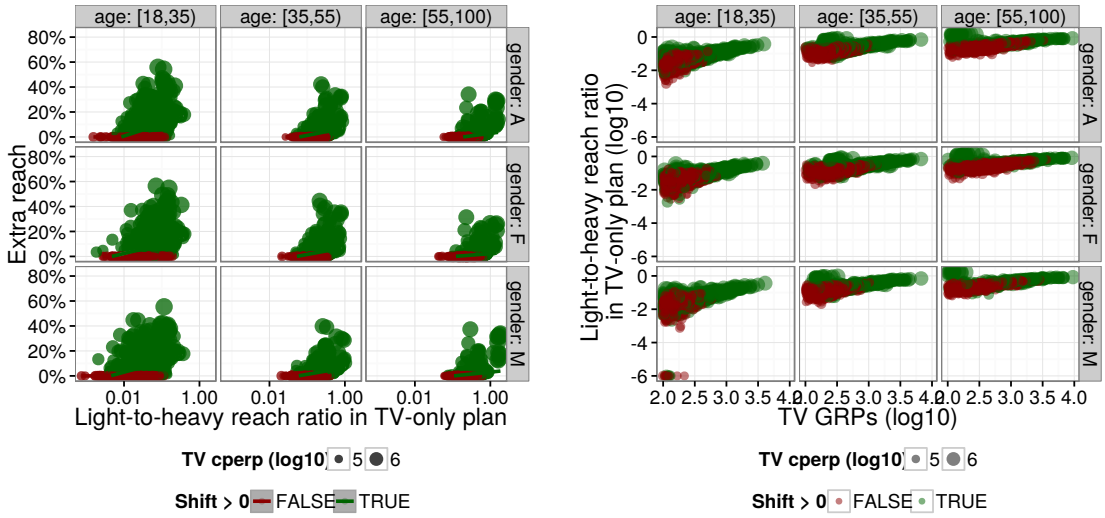


(a) TV reach of TV-only plan **(b)** YT reach of optimal plan **(c)** Extra reach of optimal plan

Figure 6: Where does YouTube gain new audiences compared to TV?: TV, YouTube, and extra reach split by TV viewing time of the population.

combined reach in those buckets).

The distribution across light to heavy viewers can also be useful to explain when a campaign is more likely to shift. As a univariate quantity to summarize the distribution over buckets consider the light-to-heavy ratio of TV reach and GRPs, which describes the (in)equality across buckets. Figure 7 shows that a campaign with a high light-to-heavy reach ratio will more likely benefit from adding online advertising. Furthermore, the patterns in size of the points (proportional to $\log_{10}$(TV cperp)) suggest that adding average TV cperp can further improve predictions.



(a) Light-to-heavy reach ratio of TV-only plan and its effect on the attainable extra reach.

(b) Light-to-heavy reach ratio in TV-only plan, size of the campaign, and how it affects whether a campaign shifts or not (green vs. red).

Figure 7: Light-to-heavy viewer TV reach ratios of TV-only plan and their predictability of extra reach and likelihood to shift.

4 Predicting The Optimal Media-Mix

In Section 4.2 and 4.3 we use common classification and regressions techniques from statistics and machine learning. We will not discuss mathematical or statistical details of each method but refer to standard text books in statistical modeling and machine learning such as Bishop (2006); Hastie et al. (2001).

4.1 Notation

Before presenting the results we review notation used for several classification and regression models – using the classic linear model as a baseline example, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$.[7]

For the predicted variable $\mathbf{y}$ we either use: a) $\mathbf{y} = \mathbb{1}\,(\text{shift} > 0)$ (for classification), or b) $\mathbf{y}$ as some continuous variable like optimal shift or cost savings (for regression). In some cases we use link functions directly on the response variable or in generalized linear models.

For the predictor matrix $\mathbf{X}$ we use all the information we have about a campaign from TV data, e.g., GRPs, reach, frequency, cpp, cperp, demographics, etc. In many cases we also use variables on log-scale to account for multiplicative effects (which is especially important for socio-economic quantities such as budget and population sizes). As we are interested in a prediction tool for TV-only campaigns, we restrict the analysis to TV data only; no online media information is used as a predictor.

Dropping marginal cperp from generalized linear models Lemma 2.1 shows that – in theory – marginal cperp is the single most important variable to for predicting likelihood of shifting. If we could only choose one variable to use as a predictor marginal cperp is a better option than average cpp, average cperp, or average frequency. The prediction models we use below, however, are mostly multivariate and many of them are generalized linear models.

Since we compute marginal cperp as a constant times average cperp times frequency, it is clear that on log-scale these three variables are perfectly collinear

$$\begin{aligned}\log(\text{marginal cperp}) &= \log(\text{const}) + \log(\text{average cperp}) - \log(\text{frequency}) && (11)\\ &= \log(\text{const}) + \log(\text{average cperp}) - (\log GRPs - \log reach)\,. && (12)\end{aligned}$$

Since we are mostly interested in good predictions, and not in inference about a coefficient β_j for marginal cperp, we will neither use (logarithm of) marginal cperp nor (logarithm

[7] We do use more advanced methods than linear regression, but the predictor and predicted variables remain the same throughout.

of) frequency as part of the $\mathbf{X}$ matrix (if GRPs, average cpp, and reach are included), but allow the model to determine the best combination of logarithmic average cperp, logarithmic GRPs, and logarithmic reach to give the best fit.

4.2 Classify which campaigns should shift

Out of 26,222 campaigns 29.25% would benefit to shift part of their TV budget to YouTube, and thus increase their combined reach at constant budget. While these recommendations were obtained through a combination of several layers of probabilistic model estimates, it would be useful to have a good rule of thumb to say whether a campaign is likely to shift or not. A trivial baseline model assigns each campaign the label with highest frequency ("majority vote"); for this dataset, the majority label is 'no shift', yielding an overall classification error of 29.25%.

In this Section we use linear discriminant analysis (LDA) (Section 4.2.1), logistic regression (Section 4.2.2), Support Vector Machines (SVM) (Section 4.2.3), and decision trees (Section 4.2.4) to classify campaign in 'shift' versus 'no shift'. Decision trees in particular have very good prediction accuracy and yield interpretable rules.

4.2.1 Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) aims to find a linear combination of of variables, $\mathbf{z} = \boldsymbol{\beta}'\mathbf{X}$, so that a simple threshold rule on $\mathbf{z}$ can discrimate well between classes in $\mathbf{y}$. In two dimensions this corresponds to a rotation of coordinates such that a horizontal (or vertical) lines can separate the classes to the top and bottom (or left and right).

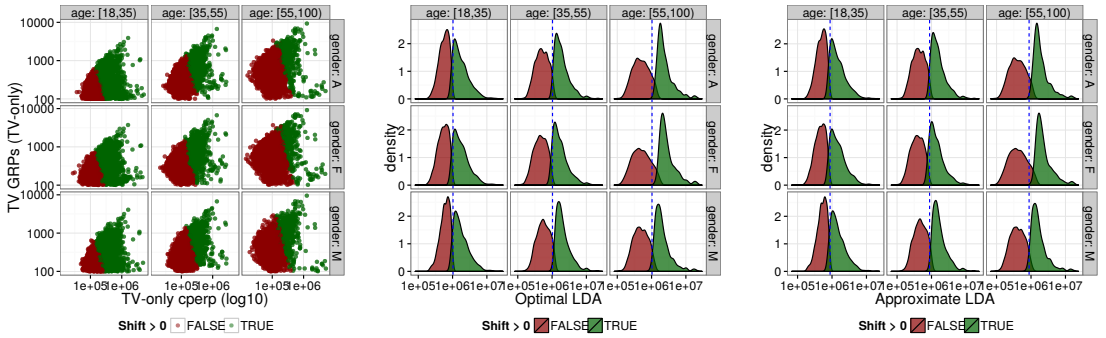
Figure 8a suggests to use a LDA for the logarithm ($\log_{10}$) TV GRPs and TV cperp

$$\mathbf{z} = \beta_1 \log_{10} cperp + \beta_2 \log_{10} GRPs \lessgtr c, \tag{13}$$

where β_i parametrize the coordinate rotation, and c is the optimal threshold for classification. Without loss of generality assume that $\beta_1 = 1$.[8] The estimated optimal classifier, $\widehat{\beta} = (1, 0.28)$ and $10^{\widehat{c}} = 1.03 \times 10^6$, has a 5.76% training error (CV: 5.83%). Since (13) is on $\log_{10}$ scale, this is equivalent to using the transformed variable

$$10^{\mathbf{z}} = \text{cperp} \cdot \text{GRP}^{0.28} \lessgtr 10^c. \tag{14}$$

[8] One can always divide (13) by β_1 and thus make $\tilde{\beta}_1 = 1$, $\tilde{\beta}_2 = \beta_2/\beta_1$, and $\tilde{c} = c/\beta_1$.



(a) Cost per effective reach point (cperp) and GRPs determine whether a campaign shifts or not.

(b) Optimal LDA estimate $10^{\mathbf{z}}$

(c) Approximate LDA estimate $10^{\tilde{\mathbf{z}}}$

Figure 8: Linear discriminant analysis (LDA) on cperp and GRPs: two-dimensional scatterplot of the original data including class assignments (by color) and the resulting densities of the LDA estimate. Optimal threshold represented by dashed, blue line.

The density estimates in Fig. 8b show how $10^{\mathbf{z}}$ can clearly separate shift vs. no shift campaigns.

As a more interpretable proxy, one can also use $\tilde{\mathbf{z}} = \text{cperp} \cdot GRP^{\frac{1}{4}}$ with $10^{\widehat{\tilde{c}}} = 8.79 \times 10^5$ in (14) (shown in Fig. 8c), which has a 5.52% training error (CV: 5.55%).

4.2.2 Logistic regression

Here we interpret optimal shift as a probability. An advertiser should shift budget with probability p, and this probability depends on the characteristics of a campaign. Logistic regression tries to model this probability as a generalized linear function of the predictor variables $\mathbf{X}$, $p = \text{logit}^{-1}(\mathbf{X}\boldsymbol{\beta})$, where logit^{-1} is the inverse of the $\text{logit}(p) = \log(p/(1-p))$ link function.

Table 5 summarizes the results of a logistic regression for $\mathbb{P}\left(I(\text{shift} > 0) \mid X\right)$, where the model matrix X contains previously described metrics of the TV campaign (and others). The CV error for logistic regression with LASSO (Tibshirani, 1994) lies at 3.06%. This means that logistic regression achieves a 90% error reduction compared to the 29.25% baseline.

Table 5: Logistic regression estimates for $\mathbb{P}(\text{shift} > 0 \mid \mathbf{X})$. The left GLM is a baseline model with only few predictors for better interpretation and statistical inference; the right is a GLM with a large variety of available predictors – mainly used for prediction.

	shift.greater.zero		
	GLM	GLM	LASSO (CV)
	(1)	(2)	(3)
Constant	−223.0*** (5.1)	−244.0*** (8.6)	−148.0
log10.lhr.orig.tv.reach		0.5* (0.3)	0.0
age.group[35,55)	−3.1*** (0.1)	19.0** (7.9)	−1.5
age.group[55,100)	−7.8*** (0.2)	33.0*** (7.7)	0.0
genderF	−0.02 (0.1)	−0.005 (0.1)	0.0
genderM	0.2 (0.1)	0.1 (0.1)	0.0
log10.orig.tv.grps	13.0*** (0.3)	9.5*** (0.5)	6.0
log10.orig.tv.reach		8.1*** (0.7)	4.1
log10.lhr.orig.tv.reach:age.group[35,55)		−1.5** (0.7)	0.8
log10.lhr.orig.tv.reach:age.group[55,100)		−5.8*** (0.8)	0.0
age.group[35,55):log10.orig.tv.cperp		−4.6*** (1.5)	−0.001
age.group[55,100):log10.orig.tv.cperp		−8.4*** (1.5)	−1.0
log10.orig.tv.cperp	36.0*** (0.8)	43.0*** (1.5)	26.0
Class. error	0.03	0.03	0.031
Observations	26,222	25,620	25,620
Note:			*p<0.1; **p<0.05; ***p<0.01

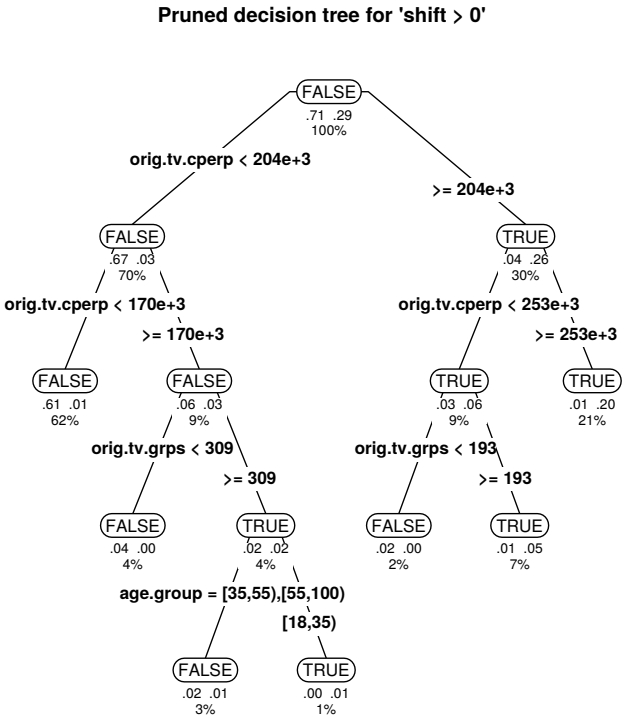


Figure 9: Pruned decision tree fit for $I(\text{shift} > 0)$ with a cross validation error of 5.56%. Every node represents the predicted label; every branch a decision rule. The two numbers below each node are the proportion of true labels; the percentage below each node refers to the percentage of observations in each node (thus the two proportions add up to the percentage).

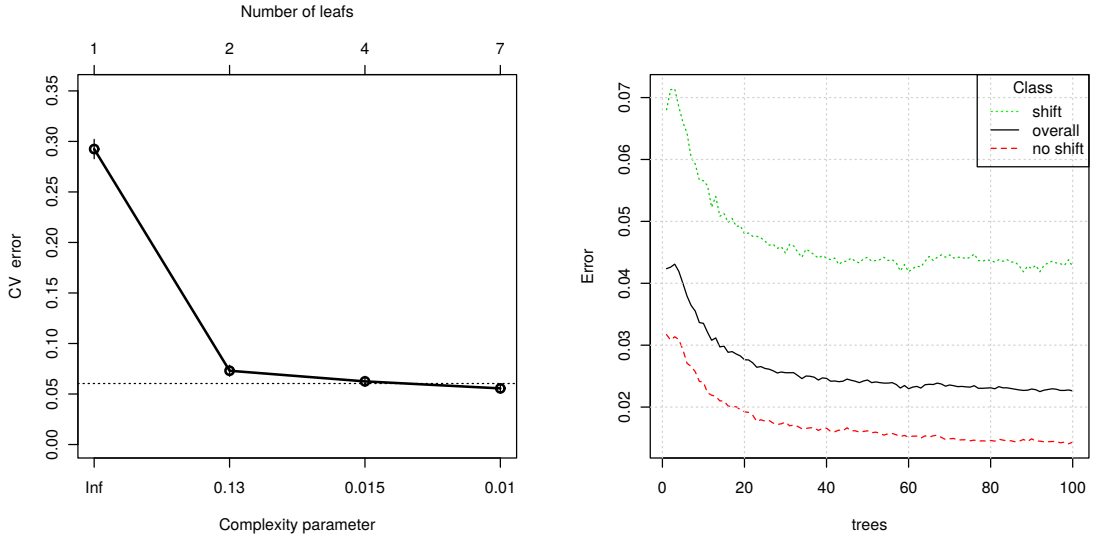
4.2.3 Support Vector Machines (SVM)

Support vector machines (SVMs) (Burges, 1998) are a popular machine learning classification method. An SVM tries to find a hyperplane through a set of points, such that it divides the space into a subspace with points from (mostly) one class and the complementary subspace with points (mostly) from the other class. Here we use *linear* as well as a non-linear extension, a *radial* SVMs. For details see the References in the `e1071` R package (Dimitriadou et al., 2010).

A linear (radial) SVM to predict positive shift has a training error of 2.97% (radial: 1.85%). The 10-fold cross validation error is 3.03% (radial: 1.93%).

	0	1	total
0	0.68	0.03	0.71
1	0.02	0.26	0.29
total	0.71	0.29	1.00

Table 6: Normalized cross tabulation of predictions (row) versus data (columns) of pruned decision tree (total: 26,222 observations).



(a) CV error by size of tree, i.e., the total number of tests (nodes) in the decision process.

(b) Error as function of number of trees in a random forest.

Figure 10: Classification tree and random forest cross-validation (CV) error for predicting $I(\text{shift} > 0)$.

Table 7: Importance of each variable in random forest (ordered by mean decrease in accuracy error when adding the variable).

	FALSE	TRUE	MeanDecreaseAccuracy	MeanDecreaseGini
orig.tv.cperp	0.160	0.430	0.240	5,180
orig.tv.grps	0.068	0.100	0.078	1,082
orig.tv.cpp	0.053	0.120	0.074	2,391
age.group	0.036	0.088	0.051	507
orig.tv.reach	0.035	0.047	0.039	647
lhr.orig.tv.reach	0.027	0.032	0.028	511
lhr.orig.tv.grps	0.021	0.039	0.026	349
gender	0.004	0.010	0.006	150

4.2.4 Decision trees and random forests

Decision trees are a powerful non-linear classification technique, with a straightforward interpretation. A decision tree looks at one variable at a time and tries to find the best threshold to split the data; it then splits the data into two subgroups and starts this best variable and threshold selection again. This iterative process results in a tree, where each node is a rule and a data point is classified in each bin based on whether it satisfies the rule or not ('yes' or 'no'). Every level further down the tree gives more fine grained (but eventually overfitting) predictions.

Figure 9 shows the (regularized / pruned) decision tree, which achieves 5.13% training error rate (5.56% for CV). Table 6 shows a cross tabulation of predictions versus observed data. The label in the rounded box at each node represents the class label and the proportion below the box indicates the classification error at this node.

A Random Forest (Breiman, 2001; Therneau et al., 2013) improves the error to 2.26% (Figure 10). It also allows us to rank variables by importance, i.e., their ability to decrease classification error (Table 7). As above, TV-only cperp is the most important variable to predict shift versus no shift.

4.3 Predicting optimal shift and extra reach

In the previous section we presented a collection of classification models to tell advertisers if a specific TV campaign is likely to benefit from online advertising. Once advertisers determine if a TV campaign is a good candidate for online advertising, the next questions are: how large should the online media portion be and how much extra reach can be expected.

In this section we thus build models to predict optimal shift and extra reach from TV campaign characteristics, such as target demographics, total budget, GRPs, and total reach. We make the assumption that the classification models above can successfully separate between shift and no-shift campaigns. For training the prediction models we thus restrict the data to only those campaigns that had a positive shift.

4.3.1 Linear regression (OLS, robust)

Table 8 displays parameter estimates for multivariate regression predicting optimal shift, for a generalized linear model with a logarithmic link function and a robust linear fit (no link function).

Table 8: Linear regression estimates for 'reach' optimization results and only campaigns with shift > 0. ρ^2 is the squared correlation between data and fit (on original scale).

	Dependent variable: optimal shift (logit)		
	glm: gaussian *link = logit* Subset of variables	*robust* *linear* Subset of variables	*glm: gaussian* *link = logit* All LASSO
	(1)	(2)	(3)
Constant	3.60*** (0.19)	2.90*** (0.40)	−3.50
lhr.orig.tv.reach	4.00*** (0.46)	1.20 (0.98)	0.01
lhr.orig.tv.grps			−0.15
log10.orig.tv.cperp			0.87
log10.orig.tv.cpp			−0.22
log10.orig.tv.reach			−0.49
age.group[35,55)	−0.56*** (0.05)	−0.68*** (0.12)	−0.06
age.group[55,100)	−0.84*** (0.08)	−1.20*** (0.17)	−0.11
orig.tv.cperp			−0.0000
orig.tv.cpp			0.0000
orig.tv.reach			0.00
orig.tv.grps			0.0000
genderF	0.01 (0.02)	−0.01 (0.02)	0.01
genderM	0.04** (0.02)	0.07*** (0.02)	−0.003
log10.orig.tv.grps	−3.30*** (0.11)	−2.70*** (0.09)	0.00
log10.lhr.orig.tv.reach	−1.50*** (0.19)	−4.50*** (0.39)	−0.05
log10.orig.tv.freq	4.80*** (0.10)	5.50*** (0.17)	
lhr.orig.tv.reach:age.group[35,55)	−2.40*** (0.32)	−2.20*** (0.68)	
lhr.orig.tv.reach:age.group[55,100)	−3.20*** (0.35)	−2.40*** (0.78)	
log10.orig.tv.grps:log10.lhr.orig.tv.reach	0.66*** (0.09)	2.20*** (0.17)	
log10.lhr.orig.tv.grps			−0.01
ρ^2	0.73	0.7	0.89
Observations	7,669	7,669	7,669
Note:			*p<0.1; **p<0.05; ***p<0.01

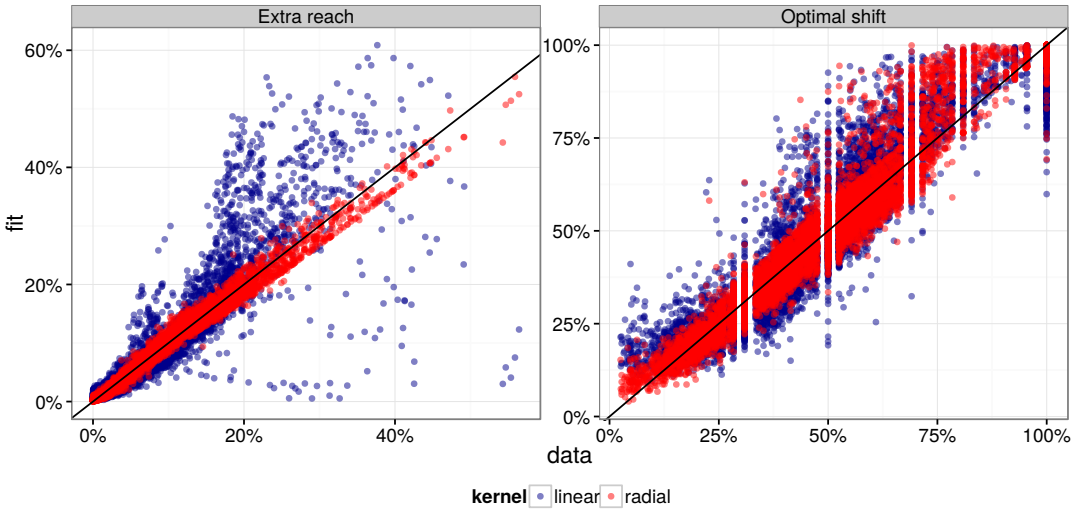


Figure 11: SVR model check: data versus fit. Solid, black line represents perfect prediction

4.3.2 Support vector regression (SVR)

Support vector regression (SVR) is an extension of SVMs: while in SVMs the hyperplane is the means to separating data points into classes, in SVRs the hyperplane is the actual regression function that should be estimated, and the sign of the residuals represents the two classes (see Schölkopf and Smola, 2002, for an overview). Similarly to SVMs, SVRs also benefit from its ability to easily model non-linear dependencies.

For extra reach and optimal shift predictions SVRs have much better predictive power than (generalized) linear models. For predicting optimal shift (on logit scale) the cross validated squared correlation between data and fit, ρ^2, equals 64.97% for the linear SVR; the radial SVR achieves 83.62%. Similarly for predicting (the logit of) extra reach: 86.84% for linear SVR (radial: 94.85%).

Figure 11 compares data versus fit for both kernels. The linear SVR deviates from the 45° line and slightly underestimates large reach and overestimates small reach. The radial SVR, on the other hand, adapts to different dependencies for small and large campaigns and thus can accurately predict reach for a wide range of TV campaigns.

5 Discussion

In this meta study we predict optimal budget allocation between YouTube and TV from TV-only campaigns. We train classification and regression models on TV-only advertising

data to decide whether a campaign should shift budget to YouTube and to predict how much shift and extra reach advertisers can expect.

We find that the most critical variable for predicting shift is cost per effective reach point (cperp) on TV, and – to lesser extent – the size of the campaign, measured by GRPs. A linear discriminant analysis (LDA) on cperp and GRP yields a decision rule with a very low 5.8% error rate. It is a simple threshold rule, based on well known metrics in TV advertising, with a clear interpretation: a campaign benefits from adding online advertising if TV is too costly, and the cost threshold gets smaller for larger campaigns. Using more advanced classification methods we can reduce the misclassification rate below 3%.

Similarly, regression models give good predictions for optimal shift, optimal savings, and extra reach. Linear regression models have good predictive power (squared correlation coefficient $\rho^2 \approx 93\%$), but they are outperformed by non-linear methods such as kernel support vector regression (SVR) ($\rho^2 \approx 99\%$); however, the latter lose the interpretability of linear regression.

Overall, our works provides recommendation for advertisers, who can use these models to set expectations about how a particular campaign might fare with online media in their advertising mix.

Acknowledgments

We would like to thank Tony Fagan, Penny Chu, Raimundo Mirisola, Oli Gaymond, Andras Orban, Mikko Sysikaski, Yuxue Jin, Vanessa Bohn, Elissa Lee, Daniel Meyer, Yunting Sun, and Xiaojing Wang for providing tools to obtain the data, their insightful discussion, and constructive feedback.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

Cannon, H. M., Leckenby, J. D., and Abernethy, A. (2002). Beyond effective frequency: Evaluating media schedules using frequency value planning. *Journal of Advertising Research*, 42(6).

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2010). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.5-24.

Goerg, G. M. (2014). Estimating the reach curve from only one data point. Technical report, Google, Inc.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hlavac, M. (2014). *stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables*. Harvard University, Cambridge, USA. R package version 5.0.

Huber, P. (1981). *Robust Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley.

Jin, Y., Koehler, J., Goerg, G. M., and Remy, N. (2013). The Optimal Mix of TV and Online Ads to Maximize Reach. Technical report, Google, Inc. `http://research.google.com/pubs/pub41669.html`.

Jin, Y., Shobowale, S., Koehler, J., and Case, H. (2012). The Incremental Reach and Cost Efficiency of Online Video Ads over TV Ads. Technical report, Google Inc.

Nielsen Solutions (2013). Nielsen Presents: Cross-Platform Home Panels. `http://www.nielsen.com/us/en/solutions/audience-measurement.html`.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rossiter, J. and Danaher, P. (1998). *Advanced Media Planning*. Advanced Media Planning. Springer US.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Therneau, T., Atkinson, B., and Ripley, B. (2013). *rpart: Recursive Partitioning*. R package version 4.1-3.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

A Notation for derivatives

Let $s : \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto s(x, y)$ be a differentiable function. We use common notation, s_x, to denote the partial derivative of s with respect to x, $\frac{\partial}{\partial x} s(x, y)$ (and s_y for $\frac{\partial}{\partial y} s(x, y)$).

Let $h : \mathbb{R} \to \mathbb{R}, \tau \mapsto h(\tau)$ be one-dimensional differentiable functions of τ. The derivative of h with respect to τ is denoted as $h'(\tau)$.

Let $(x(\tau), y(\tau))$ be a generic one-dimensional curve in $\mathbb{R}^2$ parametrized by $\tau \in [0, 1]$. We denote $h(\tau) := h(x(\tau), y(\tau))$ as the mapping of the curve from $\mathbb{R}$ to $\mathbb{R}$ via $h(x, y)$. The derivative of h with respect to τ can be computed with the total derivative

$$\dot{h}(\tau) = \frac{\partial}{\partial \tau} h(\tau) = h_x(x(\tau), y(\tau)) \cdot x'(\tau) + h_y(x(\tau), y(\tau)) \cdot y'(\tau). \tag{15}$$

If we view $h(\tau)$ merely as a one-dimensional function of τ, rather than a one-dimensional curve in a higher-dimensional space, we use $h'(\tau)$ to denote its derivative.

B Analytical derivations and proofs

Proof of Lemma 2.1 The derivative of $r_k(\tau) = r_k(x(\tau), y(\tau))$ with respect to τ equals (dropping the k subscript to avoid confusion with partial derivative notation)

$$\begin{aligned} \dot{r}(\tau) &= r_x(x(\tau), y(\tau)) \cdot x'(\tau) + r_y(x(\tau), y(\tau)) \cdot y'(\tau) && (16) \\ &= r_x(x, y) \cdot (-B) + r_y(x, y) \cdot B && (17) \\ &= B \cdot [r_y(x, y) - r_x(x, y)] && (18) \end{aligned}$$

Thus combined reach is increasing at $\tau \in [0, 1]$ if (dividing by $B > 0$)

$$r_y(x(\tau), y(\tau)) > r_x(x(\tau), y(\tau)). \tag{19}$$

The optimal budget allocation τ^* is achieved when (19) holds with equality or at the boundary $\tau^* \in \{0, 1\}$.

Proof of Corollary 2.2 Follows from Lemma 2.1 and since single-channel campaign has $\tau = 0$, and thus $x(0) = B$ and $y(0) = 0$.