Ehrenberg-Bass Institute **Working Paper**:

Finding Creative Drivers of Advertising Effectiveness with Modern Data Analysis

---

*This working paper, dated 12 September 2022, is forthcoming in the **International Journal of Market Research.***

Authors:

Dr John Williams - University of Otago
Dr Nicole Hartnett - Ehrenberg-Bass Institute
A/Prof. Giang Trinh - Ehrenberg-Bass Institute

University of South Australia | Ehrenberg-Bass Institute for Marketing Science

Finding Creative Drivers of Advertising Effectiveness with Modern Data Analysis

The Modern Data Analysis paradigm (Williams 2021) advocates using multiple methods to address the same research question, which is rarely done in studies of advertising creative effects. In this paper, we apply the MDA paradigm to data from Hartnett *et al.* (2016a), which coded 158 creative variables for 312 television advertisements with commercially validated short-term sales effectiveness outcomes. We found that many models give higher classification accuracy than the ordinal regression model previously applied, some significantly higher. Importantly, by applying many alternative but equally plausible analytical methods, we can identify creative variables associated with commercial success and have evidence-based confidence that these creative variables are artefacts of the data, and not artefacts of any particular analytical method and its associated assumptions. The findings reveal several alternative creative variables that are consistently associated with sales success across methods, which relate to the timing aspects of visual branding.

Keywords: advertising, creative, sales effectiveness, Modern Data Analysis

Introduction

Successful advertising, with respect to growing brands, is largely a result of effective creative execution. What creative execution or messaging elements of video advertising more effectively drive people to buy? Researchers have spent decades gathering evidence to answer this complicated question. Is it a humorous tone? Is it a character-driven narrative? Is it communicating product benefits?

Studies examining many creative variables have consistently found that the explanatory power of any *single* creative variable against advertising outcomes is small (Hartnett *et al.* 2016a; Stewart & Furse 1986; Stewart & Koslow 1989). One interpretation for this finding is that advertising effectiveness is a result of the combinations of creative variables used and given the many creative variables available to advertisers, there are hundreds of possible combinations across advertisements. The likely non-linear interactions between creative variables (Carlson 2011) make it difficult to isolate and interpret effects. Another consideration is context-based contingencies for an advertiser, such as product type or brand history, where creative variables are differentially effective across conditions (Armstrong 2011a, b). To address these quandaries and draw robust insights about which creative variables matter more requires extensive experimental research and/or very large data sets with the application of advanced analysis techniques.

We see two issues with prior studies of advertising's creative effects. Firstly, historically, a lot of the empirical work in this space has applied ordinary least squares (OLS) or logistic regression models to determine relationships between creative variables and various advertising outcomes (e.g., Bellman *et al.* 2012; Bellman *et al.* 2017; Chandy *et al.* 2001; Hartnett *et al.* 2016a; Kim *et al.* 2013; Stewart & Furse 1986; Stewart & Koslow 1989). These models, by default, assume the effect of explanatory variables is independent of the effect of other variables. Sometimes analysts add two- or three-way interaction effects,

however, these approaches are rare and still limited by *a priori* theory to identify a set of creative variables that can work together to produce superior advertising outcomes. Secondly, we observe that most studies lack replication of analysis, in that they apply a single analytical method and do not use a hold-out or alternative data set for validation. There is no guarantee that significant relationships identified by a single analytical model will hold if another method or model (algorithm) is used. Without replicating results, either through applying different models to the same data set, or the same model across different data sets, we must proceed cautiously and with limited confidence that those results will lead to useful explanations of more effective advertising creative.

In this research, we apply the MDA paradigm to a data set originally reported by Hartnett *et al.* (2016a), which examined creative variables from 312 television advertisements against their short-term sales effects using (only) an ordinal logistic regression model. We re-analyse the same advertisements, using a host of traditional and AI and ML models (manual and automated) with two objectives in mind. The first was to compare the accuracy of the predictions produced across traditional and newer models. The second was to identify which creative variables contributed the most to sales performance across models, or practically, which creative variables more consistently explain commercial success. By applying many models to the same data set of advertisements to determine what more can be learned, we illustrate the issue of single method bias. The findings allow meaningful advertising theory development, as well as giving advertising managers guidance on how they might produce more sales effective advertisements.

Background

Modern Data Analysis

The MDA paradigm is not a method but a set of principles that encapsulate the approach to analytical work implicitly employed by data scientists (Williams 2021). The MDA paradigm is expressed in a core set of principles: (a) a de-emphasis of the frequentist null-hypothesis significance testing paradigm; (b) triangulation over data by bootstrapping and cross-validation; and (c) triangulation over methods by applying all valid methods to a problem and seeing what works best. Implied in "modern" is using state-of-the-art methods in (c), especially methods that can *automatically* deal with non-linearity and multi-collinearity.

Modern methods of predictive analytics that are often understood to be examples of artificial intelligence (AI) or machine learning (ML) hold promise for improving our knowledge about marketing effectiveness generally (Chintalapati & Pandey 2022; Mustak *et al.* 2021; Vlačić *et al.* 2021), and advertising design and effectiveness specifically. The promise of AI and ML methods is that they often provide superior performance than currently used methods, where "performance" is what statisticians call "goodness of fit" and data scientists call "predictive accuracy". Meanwhile, the perils of AI and ML in the context of predictive analytics are manifold, but for the most part reduce to two issues: a little knowledge is a dangerous thing (i.e., a powerful and complex tool in the hands of an inexperienced or unskilled user is likely to produce undesirable outcomes); and prediction and explanation appear to be a trade-off in many empirical applications. Prediction may be enough in some managerial contexts, notably the real-time bidding algorithms used in programmatic advertising and recommendation algorithms, but in other contexts, explanation is arguably more important. That is if an algorithm predicts that action *X* will be successful, it

is helpful to know *why* the algorithm made that prediction, specifically which variables contributed most to the prediction.

In the context of advertising creative effects research, comprehensively quantifying the creative variables present in advertisements produces many variables for analysis, and we argue that traditional statistical methods are limited for analysing such data. Newer ML methods are better able to accommodate large numbers of creative variables. Some ML methods, notably Artificial Neural Networks (ANNs) and random forests, do not suffer from numerical problems due to multicollinearity and interactions that are not explicitly (manually) modeled. If ML and AI methods predict an advertisement execution will succeed commercially, we would like to know which aspects of that video are associated with commercial success, so we can choose to include them in future campaigns.

Creative advertising effects

Studies looking to understand what aspects of advertising content affect consumer responses have a long history. We limit our discussion to studies that investigate multiple creative variables for video advertisements. Even then, there exists a substantive body of literature with studies proving quite different in their execution. For example, studies differ markedly in their data collection methods, with varied advertising outcomes such as intermediate memory and/or evaluation survey measures (e.g., McEwen & Leavitt 1976; Walker & von Gonten 1989; Yelkur *et al.* 2013), observed viewing time or avoidance measures (e.g., Becker *et al.* 2022; Olney *et al.* 1991), or in-market sales/demand response (e.g., Chandy *et al.* 2001; Dall'Olio & Vakratsas 2022; Tellis 2004), sometimes focusing on a particular subset of creative variables, such as branding tactics (e.g., Bruce *et al.* 2020; Romaniuk 2009; Teixeira *et al.* 2010). These differences can make it difficult to reconcile inconsistent findings with respect to which creative tactics matter more for improving advertising performance.

Stewart & Furse (1986) is considered the seminal work in this space. They developed an extensive, reliable creative codebook and examined the relationships between 160 creative variables and three advertising measures (related recall, message comprehension, and persuasion) from lab-based pre-testing conducted on more than 1,000 advertisements. Several years later the study was replicated with a new data set of 1,000 advertisements that had undergone the same pre-testing (Stewart & Koslow 1989). There were some consistent findings, namely that a brand differentiating message had the strongest, positive affect on all three outcome measures. Others have since applied the codebook, in part or in full, to explore different advertising conditions with the same pre-testing measures (Laskey *et al.* 1994; Phillips & Stanton 2004; Stanton & Burke 1998), or different outcome measures, such as biometric responses (Bellman *et al.* 2019), requests for information (Bellman *et al.* 2012), and brand sales (Hartnett *et al.* 2016a; Guitart & Stremersch 2021).

Hartnett *et al.* (2016a) is a notable study because it was the first study to link advertising execution to sales effects using single source data. Unlike econometric studies that examine gross rating points (GRPs) supporting campaigns against aggregated weekly sales, single source data directly links receiving a specific advertisement with subsequent purchases for individuals. As such, single source data is better able to isolate and measure the short-term creative effect of an advertisement on brand buying (Jones 2007), which is the outcome marketers are generally most interested in. Also, unlike other studies, Hartnett *et al.* (2016a) applied Stewart & Furse's (1986) codebook in full, making it the first full replication and extension since Stewart & Koslow (1989). Being a full replication is important to comprehensively identify in/consistencies across those creative variables available to advertisers.

Hartnett *et al.* (2016a) used ordinal logistic regression, and found the following creative factors[1] significantly ($p<.10$) increased odds of more sales effective advertisements:

- Negatively framed appeals (with celebrities and psychological benefits)
- Humour (with an on-camera spokesperson)
- Enjoyment appeals (including information about a sensory experience)
- Animated characters (that are brand created or continuing characters)

Conversely, creative factors that significantly ($p<.10$) decreased the odds of more sales effective advertisements:

- Mood music (includes dominant music driving mood or action)
- Puffery (makes baseless superiority claims)
- New ending information (tagging the advertisement with information not related to the body, e.g., introducing special offers or extended product range)

Drawing comparisons to Stewart & Furse (1986), the positive effect of humour was consistent (specific to related recall and message comprehension), but other findings were either not directly comparable due to the principal components analysis producing divergent creative factors, or contradictory in effects (e.g., Stewart & Furse (1986) found that mood music had a positive effect on related recall and message comprehension). The limited consistency in findings was not unexpected considering the incidence of creative tactics between these data sets varied greatly and the outcome variables captured were also different (i.e., survey measures from pre-testing vs. observed buying behaviours). Perhaps another contributing factor, not previously discussed, is using different models to analyse the data. This issue is another reason to follow the MDA paradigm, which explicitly addresses the

[1] A principal components analysis was conducted to reduce the large number of codes into a more manageable set of 28 creative factors.

question “to what degree are our findings an artefact of the data (i.e., the empirical world, albeit imperfectly measured) as opposed to method we have chosen to analyse it?”

Modern Data Analysis applied to creative advertising effects research

The studies discussed above, along with most others that examine creative effects, only apply a single analytic model. It is therefore unclear whether the creative variables identified are the “true” drivers of short-term sales (or other advertising outcomes) or identified as such due to the assumptions and specifications of the specific method used. Similarly, most studies do not test or do not report model fit or prediction with a holdout sample (with exceptions, e.g., Guitart & Stremersch 2021).

Hartnett *et al.* (2016a) tested how an ordinal regression model could fit, or “reverse” predict, the outcomes for the same advertisements from which the model was developed. This approach is not as compelling as testing on a holdout sample, but the sample size of advertisements was insufficiently large to justify a holdout. The outcome variable was measured as a three-level ordinal variable of *above-average*, *average,* and *below-average* performance relative to product category and country norms. The model classified the correct outcome well for the *above-average* and *below-average* categories (68% and 73% respectively) but had a 100% failure rate for the *average* category. Seemingly *average* advertisements presented combinations of creative variables that made them impossible for the model to classify. Consequently, the results do not inspire a great deal of confidence to use the model as a basis for creative decision-making and justify the application of alternative models to improve prediction.

Given the most recent formalisation of MDA, we felt is pertinent to employ this approach to the pre-existing dataset, where the application to sales effects makes this data and new findings very relevant for marketers.

Methods

Television advertisements and advertising effectiveness

Our data set was provided by a global owner and advertiser of packaged goods products, originally reported on by Hartnett *et al.* (2016a). The data set included 312 television advertisements that were aired and measured in five developed markets, for more than 60 brands of different sizes. Most of the advertisements were for the manufacturer's own brands. The advertisements were accompanied by a commercially validated measure of sales effectiveness using single-source data, which is a rarity in the advertising literature and very relevant to the objectives of advertisers and their decision-making. The significance of analysing this type of data cannot be overstated, making a direct link to financial outcomes.

Single-source data combines television advertising delivery logs recorded passively with set-top boxes and scanned grocery purchases using store cards or devices operated in-home (hand-held scanners). These data are collected for the same households; that is, two observational data sets originating from a "single-source". Such disaggregated data allows for a quasi-experimental design to be extracted from "as it lies" real-world data, namely by separating households that received advertising versus those that did not. Households were recruited and remunerated for their participation by large market research companies, which operate panels of many thousands of households across countries. These data were sold to the advertiser, which analysed the data in-house to determine the sales effect of advertisements aired in a specified timeframe.

The advertiser's analysis approach resembled Jones' (1995) short-term advertising strength (STAS) system. Category purchases made in a four-week period were linked to advertising opportunities-to-see (OTS), for the target brand advertisement, in the four-week period immediately prior. The measure of short-term sales effectiveness is an index that uses a proprietary Bayesian technique comparing brand-specific purchases among households *with*

advertising OTS against the same brand's sales for households *without* advertising OTS. Regarding which households received advertising, this occurred naturally, and without controls forced by the advertiser or panel when the advertisements aired. Socio-demographics of the exposed and unexposed households were checked for biases and if the groups were markedly different, the advertisement was not measured (i.e., not included in our data set). The industry partner further used contingency tables to control for other potential impacting variables (Roberts 1996), such as price promotions and exposure frequency. These quality controls (among others) meant that the advertiser, and these authors, were confident of the validity of the advertising measurement.

The advertiser provided the authors with video clips of the ads and as mentioned previously, the corresponding effectiveness outcomes reported as a three-level ordinal variable as *above average*, *average*, or *below average* sales effectiveness. The raw indices (which were not shared with the authors) were converted to an ordinal variable because the data set spanned multiple product categories, and some product categories are more responsive to advertising than others. This adjustment for category norms meant outcomes were directly comparable across these conditions. These outcomes were present in roughly equal thirds of advertisements across the data set.

Coding procedure for creative variables

The creative codes form the independent variables (IVs) for our study. Hartnett *et al.* (2016a) used the Stewart & Furse (1986) codebook to identify the creative devices present in the advertisements. It remains one of the most exhaustive codebooks, spanning 160 strategies and tactics, including things such as whether advertisements present a slice of life, the principal character(s) are male or female, make direct comparisons with competitors, etc. Most codes are binary (the creative variable is present or absent, e.g., a humor appeal is used or not), but some codes are categorical (e.g., the audio portion of the message is delivered by a voice over

only, or a combination of voice over and character on screen, or by a character on screen only) while others are continuous (e.g., counting the number of seconds a brand name or logo is shown throughout the advertisement).

Five judges independently coded each advertisement for 158 creative variables (i.e., two codes were omitted). All judges were extensively trained by one of the authors familiar with the codebook, with codes illustrated using example advertisements from outside the data set, such that all codes were consistently understood and recognised in context. Intercoder reliability was assessed based on the average pairwise percent agreement (APPA) across coders. Pairwise agreement examines the percentage of decisions each combination of the five coders agreed on the presence, categorisation, or timing of a creative variable (i.e., coder 1 vs. coder 2, coder 1 vs. coder 3, etc., agrees on whether ad 1, ad 2, ad 3, etc., "includes sensory information", which is a binary code). The average percent agreement for the 10 coder pairings was calculated for each creative variable using ReCal, a publicly available online software (Freelon 2010) (e.g., coders agreed that sensory information was present or absent 73% of the time, on average). Most binary codes achieved more than 80% APPA, which is well above chance agreement. Only codes above 60% APPA were included in our analysis, which exceeds the recommended benchmark (Rust & Cooil 1994). Coding discrepancies (for nonbinary variables that failed to achieve a 'majority rule' among judges) were adjudicated by one of the authors. The coding resulted in almost 50,000 observations of creative variables across the 312 advertisements.

It is worth noting that some continuous timing variables remained as absolute counts (i.e., number of seconds) while others were converted to proportions (i.e., percent of total time). For example, the amount of time a brand logo is displayed on screen may be best measured in absolute terms (because it takes an absolute amount of time for the human brain to process information) or proportional terms (to standardise across advertisements of

different lengths). Including both measures in the same model is conceptually appealing but numerically disastrous, so a choice was made. The decision of whether to use the absolute or proportional measure was made separately for each variable *a priori* and then checked empirically.

Analysis

We applied 317 statistical models, using over 100 separate methods (algorithms) to these data using binary, nominal, and ordinal scaled measures of the DV, with both raw variables and factor scores of the IVs as covariates. This was done to apply the largest set of analytical methods possible to the data. These data options give a 3×2 matrix of variable codings. Within each cell of this matrix, all methods of the ***R*** package caret (Kuhn 2017; Kuhn & Johnson 2013) and the *h2o* package (LeDell *et al.* 2018), that were suitable for this combination of DVs and IVs were applied.

The company *h2o* provides open-source software for automatic ML and AI, which was used for some of the analyses. A total of 115 models ran successfully for the binary data, 64 for nominal and seven for ordinal. The complete set of results can be found in the appendix, which reports the fit indices of all the models.

To select "good" models, we used Accuracy, Area Under the Curve (AUC) and F_1 for model evaluation of training data. In addition to these measures, we use Recall and Balanced Accuracy for model evaluation of test data. For training data, we use the conventional cut-off of 80% for AUC (Hosmer & Lemeshow 2000). For test data, it is more managerially useful to know what makes an *above-average* advertisement than an *average* or *below-average* one. Therefore, we used Recall as the most important criterion to maximize. A good model will also predict all classes reasonably well, so we used Balanced Accuracy as the second most important criterion. For both Recall and Balanced Accuracy, a cut-off of at least 50% was applied. AUC is well developed for binary outcomes; however, multi-class AUC does not

have a single universally accepted method of calculation and is also not without interpretational problems, especially as a basis for model comparison. For these reasons the F_1 measure is recommended if one must rely on a single metric of multi-class classification model performance. This measure considers both false positives and false negatives. However, it is entirely possible for a model to have high values of AUC, F_1 and mean Recall, but still have very poor performance on one category of the DV.

The MDA paradigm mandates *benchmarking* interpretable models against the *best possible* model. In our context, we would like to find interpretable models of the creative variables associated with commercial success in the same ballpark of accuracy as the best possible model (of all the models applied).

Results

This section is lengthier than most reports of data analysis because we are using the MDA paradigm. The point of MDA is to ensure reliable results by considering many plausible models and evaluating whether the same variables are important in the best performing models. If we simply pick the "best" by fit alone, and the variables that are most important in that model are not important in any others, then the results are almost certainly an artefact of the model, not the data. The analysis is lengthy to illustrate this principle.

Firstly, we summarise the best models in terms of fit metrics, showing the range of performance over models with the dependent and independent variables constructed in several ways. This achieves the objective of comparing a single-method analysis as reported by Hartnett *et al.* (2016a) with the best possible model in terms of fit. It should be noted that in classification tasks there are several metrics associated with model performance, and opinion is divided over which is best, or even whether such a concept applies. Hence, we have tabulated several performance metrics. The point of benchmarking is to gain confidence

that a single traditional method is as good, or almost as good, as the best possible model, or less so.

Secondly, we examine the variables that are most influential among the best performing models. Crucially, the MDA paradigm eschews the notion of relying on a single best model if the influential variables of that model differ dramatically from models of similar performance. It is essential to examine the influential variables from several equivalent (in terms of fit) models to ascertain to what degree the results of each model are an artefact of the model or the data.

Thirdly, we conclude by examining the implications of the best interpretable model and the best "uninterpretable" model, using recently developed tools to "open the black box" of ANNs. The utility of these tools is illustrated by examining the best and worst performing advertisements from the best ANN model to ascertain what led to the model classifying them as such. This illustrates the use of these tools for creative development and theory generation. But crucially, the interpretation of the best available methods of identifying the reasons *why* ANNs made classifications are still radically different than the interpretability of classical models. This final section illustrates that point.

Accuracy of models

Table 1 shows mean and maximum fit indices for all combinations of variables across models. Many methods provided higher fit indices than the SPSS PLUM model previously applied to these data. Most also have the advantage that they predict the *average* sales effectiveness category well, in contrast to PLUM. The best model for the ordinal DV (see appendix) has mean accuracies of 55% for the test data and 53% for the training data, a small but probably commercially relevant improvement over the PLUM analysis previously reported. Running that model with the entire data set (rather than split into test and training data, and cross-validating) gives balanced accuracy of 68%, 50% and 72% for the *below-*

average, *average* and *above-average* categories, and Sensitivity of 62%, 20% and 69%, compared with the PLUM Sensitivity results: 73%, 0% and 68%. The best models for the binary DV have accuracies of 100% for the test data and 100% for the training data. However, these models are all based on random forests, applied to a small sample, so caution is in order.

Table 1. Summary statistics for all models (test data)

DV	IVs	Mean Recall	Max Recall	Mean Balanced accuracy	Max Balanced accuracy	Mean Accuracy	Max Accuracy	Mean AUC	Max AUC[1]	Mean F1	Max F1
Ordinal	Raw	0.53	0.56	0.59	0.71	0.44	0.57	0.61	0.75	0.49	0.61
Nominal	Factors	0.45	0.96	0.54	0.64	0.38	0.47	0.56	0.7	0.41	0.53
Binary	Factors	0.26	0.75	0.53	0.75	0.6	0.75	0.56	0.71	0.71	0.84
Binary	Raw	0.33	0.71	0.53	0.7	0.58	0.7	0.55	0.73	0.68	0.81
Nominal	Raw	0.34	0.56	0.51	0.63	0.34	0.44	0.56	0.66	0.34	0.51
Ordinal[2]	Factors				0.43[3]		0.50		0.65		

1. The conventional interpretation of value ranges of AUC, due to Hosmer and Lemeshow (2000), is that AUC < 0.7 is unacceptable. 0.7 < AUC < 0.8 acceptable; 0.8 < AUC < 0.9 excellent and AUC > 0.9 outstanding.
2. The last column of Table 1 contains the PLUM results that were previously published. The PLUM results are not directly comparable because they were derived from the full data set (i.e., not split into test and training samples) and were not cross validated.
3. Balanced accuracy and F_1 are technically undefined when recall is undefined, and recall is undefined when true positives + false negatives = 0. Hence the balanced accuracy reported here is the mean of the balanced accuracies in each category, i.e. 0.6376, 0 and 0.6756, just to give an approximation.

Table 2 displays models with *both* Recall and Balanced Accuracy ≥ 0.6 (a criterion chosen to result in a small, but not too small, set of models to consider) and the subset of those models that are not of the binary DV, to allow comparison of the multi-category DV models with the binary DV models.

Table 2. Best Models (Recall & Balanced Accuracy >= 0.6)

			Test					Training		
DV	**IVs**	**Method**	**Recall**	**Bacc**	**Acc**	**AUC**	**F_1**	**Acc**	**AUC**	**F_1**
Binary	Factors	aml_deeplearning[1]	0.75	0.75	0.66	0.75	0.84	1	1	1
Binary	Raw	vglmAdjCat[2]	0.71	0.7	0.73	0.7	0.74	0.86	0.79	0.84
Binary	Raw	gamLoess[3]	0.68	0.61	0.64	0.59	0.63	0.96	0.88	0.91
Nominal	Factors	wsrf[4]	0.64	0.62	0.63	0.45	0.52	0.87	1	1
Nominal	Factors	vglmContRatio[5]	0.6	0.6	0.63	0.47	0.5	0.68	0.52	0.58

1. h2o AutoML Deep Learning
2. Adjacent Category ordinal regression
3. Generalized Additive Model using Loess smoothing
4. Weighted Subspace Random Forest (Zhao *et al.* 2017)
5. Continuation Ratio ordinal regression

Important variables common to several methods

Table 3 shows the top 10 important creative variables which are present in more than one model. The most consistently important variables across the nominal *and* binary DV models are advertisement length (in seconds), time until product or package is shown (in seconds) and time the brand name or logo is shown on screen (as a proportion of advertisement length). These variables occur 75% in the binary DV models and 28% to 39% in the nominal DV models. Hence, the visual branding elements relating to product and brand name exposure (introduction and duration respectively) are common both within and between the models of the raw variables.

Table 3. Variables that Occur in the Top 10 Most Important Variables for 50% or More of Binary and Nominal Models[1]

Binary DV			Nominal DV		
IVs	***n***	**%**	**IVs**	***n***	**%**
Ad Length (X02)	3	75	Mood Music (FAC20)	9	50
Time Until Product or Package Shown in Seconds (X06)	3	75	Enjoyment (FAC17)	7	39
Time Brand Name or Logo is Shown as a Proportion of Ad Length (X10PROP)	3	75	**Time Until Product or Package Shown in Seconds (X06)**	7	39
Results of Using (D019)	2	50	Negative Appeals (FAC12)	6	33
New Product or New Features (D024)	2	50	Puffery (FAC28)	6	33
Visual Pace (D107)	2	50	**Time Brand Name or Logo is Shown as a Proportion of Ad Length (X10PROP)**	6	33
Time Until Product Category Identified in Seconds (X04)	2	50	Ending Information (FAC11)	5	28
Brand Name First Identified in the Last Third (X05PROP_2Last 3rd)	2	50	Humor with Speaking Characters (FAC2)	5	28
Brand Name First Identified in the Middle Third (X05PROP_2Middle 3rd)	2	50	Nutrition (FAC22)	5	28
Category: Pet Food (X14)	2	50	**Ad Length (X02)**	5	28

1. Bolded items are common to both binary and nominal models

The best models

The best interpretable model

Table 4 shows the results of the best interpretable model, an Adjacent Category ordinal logistic model fitted to the binary DV, using the VGAM package (Yee 2015). The column labeled **d** is Cohen's *D*, an effect size transformation (Chen *et al.* 2010) of the odds ratio (OR). Table 4 reports estimates significant at $p < 0.05$ and $D > 0.2$, which is the conventional cut off between a small and medium effect size (0.8 is the cut off for medium and large).

Table 4. Effect sizes for the best interpretable model (ordinal regression)

IVs	Estimate	Std. Error	z value	Pr(>|z|)	OR	d
Brand Name First Identified in the Last Third (X05PROP_2Last 3rd)	4	1.8	2.2	0.026	55	2.2
Time Brand Name or Logo is Shown as a Proportion of Ad Length (X10PROP)	2.4	1.1	2.2	0.027	11	1.3
Ad Length (X02)	2.1	0.97	2.1	0.034	7.8	1.1
Time Until Product or Package Shown in Seconds (X06)	-1.4	0.66	-2.1	0.034	0.25	-0.77
Results of Using (D019)	1.3	0.46	2.8	0.0044	3.7	0.73
Number of On-Screen Characters (X13)	-1.3	0.44	-3	0.0028	0.27	-0.72
Spoken Tagline (D047)	-1.1	0.33	-3.5	0.00056	0.32	-0.62
New Product or New Features (D024)	0.88	0.42	2.1	0.038	2.4	0.48
Rational and/or Emotional Appeal (D138)	0.87	0.4	2.2	0.03	2.4	0.48
Voice Over and/or Spokesperson On-Camera (D134)	0.77	0.36	2.1	0.033	2.2	0.42
Surrealistic Visuals (D041)	0.73	0.31	2.3	0.02	2.1	0.4
Packaging (D010)	0.72	0.35	2.1	0.04	2.1	0.4
Music (D108)	-0.7	0.29	-2.4	0.015	0.5	-0.38
Background Cast (D127)	0.69	0.28	2.4	0.015	2	0.38
Front-End Impact (D081)	0.58	0.27	2.1	0.032	1.8	0.32
Slice of Life (D090)	-0.58	0.29	-2	0.044	0.56	-0.32
Visual Tagline (D043)	0.5	0.24	2.1	0.037	1.7	0.28

The best performing model with the raw variables as IVs corroborates the random forest and neural network results to a degree. Table 5 shows the effects of variables from the best interpretable model that are also common across all binary and nominal models. From this we can see that variables related to timing dominate the important effects.

Table 5. Effects of Variables from the Best Interpretable Model Common to Best Uninterpretable Model

IVs	Label	Pr(>|z|)	OR	d
X05PROP_2LastThird	Brand name is introduced in the last third of the advertisement	0.026	55	2.2
X10PROP	Time (as a proportion of ad length) the brand name or logo is shown	0.027	11	1.3
X06	Time (in seconds) until product or packaging is shown	0.034	0.25	-0.8
X13	Number of characters on screen	0.003	0.27	-0.7

The best uninterpretable model

To generate the results for the AutoML Deep Learning model, we used the ***R*** package iml (Interpretable Machine Learning) (Molnar & Schratz 2020). We use *Accumulated Local Effects* (ALE) (Apley & Zhu 2016) as the indicator of variable importance. Figure 1 shows the most influential variables from this model. Of the eight most important variables from the Deep Learning model shown in Figure 1, half are common to the adjacent categories Ordinal Logistic model, which includes brand name introduced in the first, middle or last third (X05PROP_2), time until product/packaging is shown (X06), duration of brand name presence (X10) and number of characters (X13), indicating a reasonable degree of corroboration.

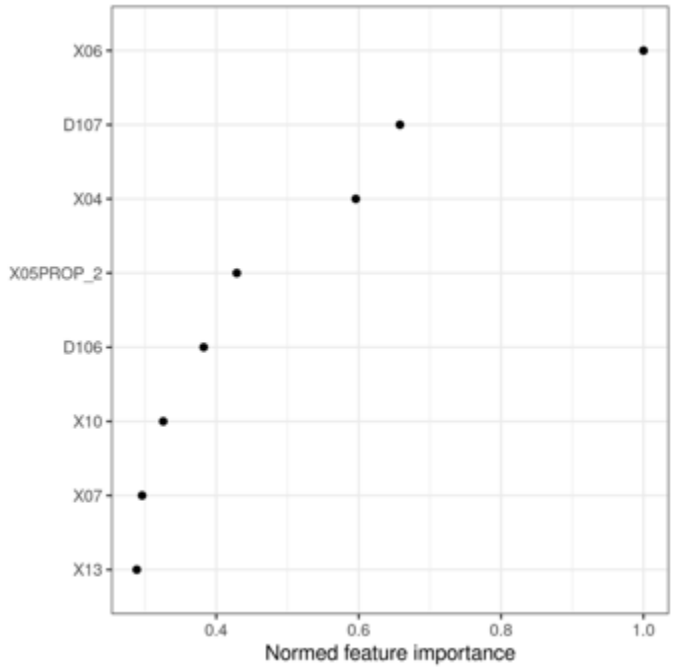

Figure 1. Normed Feature Importance (ALE) from the Best Uninterpretable Model

We use *Individual Conditional Expectations* (ICE) (Goldstein *et al.* 2015) to show the nature of the influence of each IV. Figures 2 to 5 show an ALE plot on the left and ICE plot on the right for each of the four variables of interest. The ICE plots show the ICE of each individual advertisement, and the average of those. The software we used automatically allocates the average/below average level of the DV as the 'positive' class, so increasing values on the *y* axis indicate increased probability of a poorly performing ad.

Figure 2 indicates that if the *brand name* is introduced in the second or last third of the advertisement, the probability of it being classed as average or below average is lower.

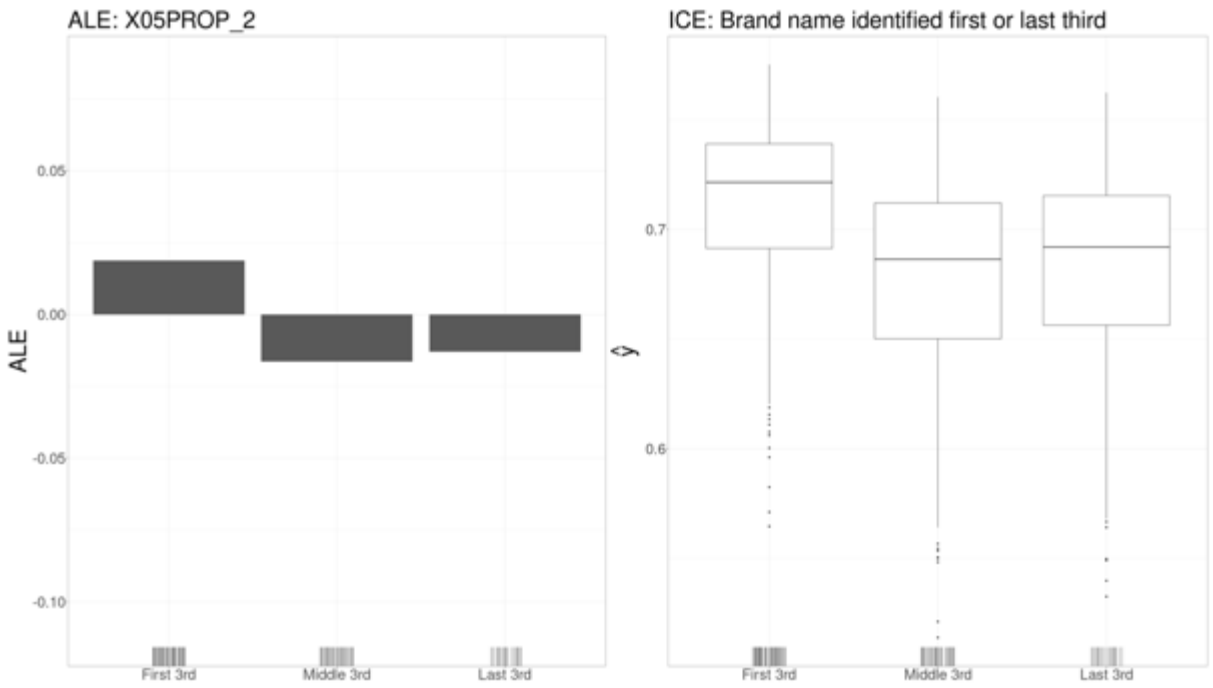

Figure 2. ALE and ICE Plots for Brand Name Introduction Timing

Figure 3 shows that the two methods (ALE and ICE) are in broad agreement, in that if the *product or packaging* is shown immediately, the probability of being less than good is low but climbs steeply and abruptly as time goes on. The ICE plots on the right show that most advertisements follow this pattern with some outliers.

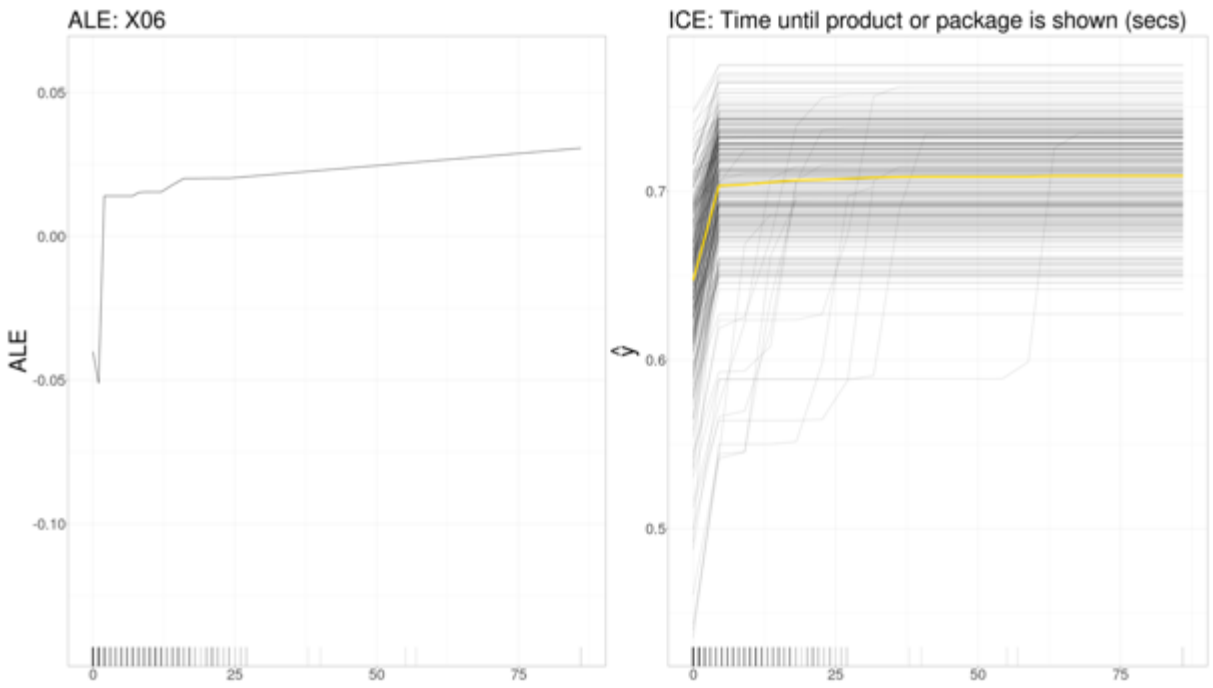

Figure 3. ALE and ICE Plots for Time Until Product of Package is Shown

Figure 4 shows an additional story with respect to timing, referring to the brand name being shown for a shorter or longer amount of time. Shorter duration is associated with better performance. The variation in the individual ICE curves is much larger than those in Figure 2. The outliers can be seen more clearly here also.

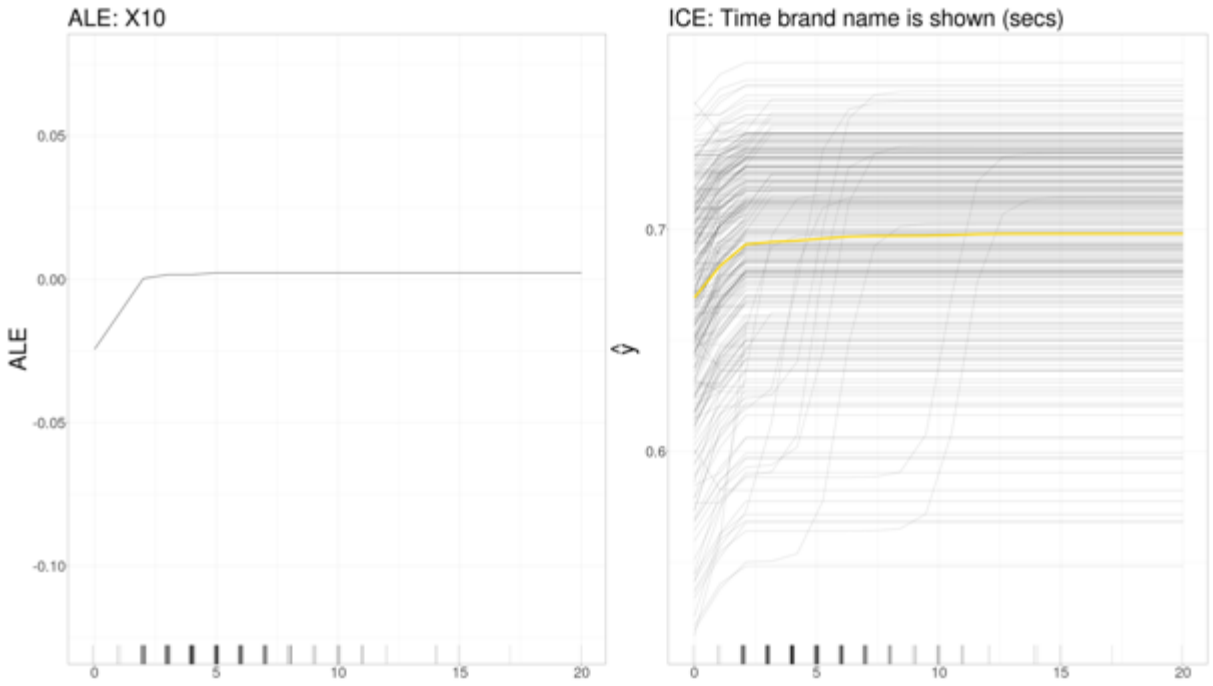

Figure 4. ALE and ICE Plots for Time Brand Name is Shown

Figure 5 shows the pattern in terms of a low probability increasing, and the heterogeneity among the ICE curves. There are fewer outlying cases here than in the previous plots. The plots indicate that a smaller number of characters in an advertisement is preferable.

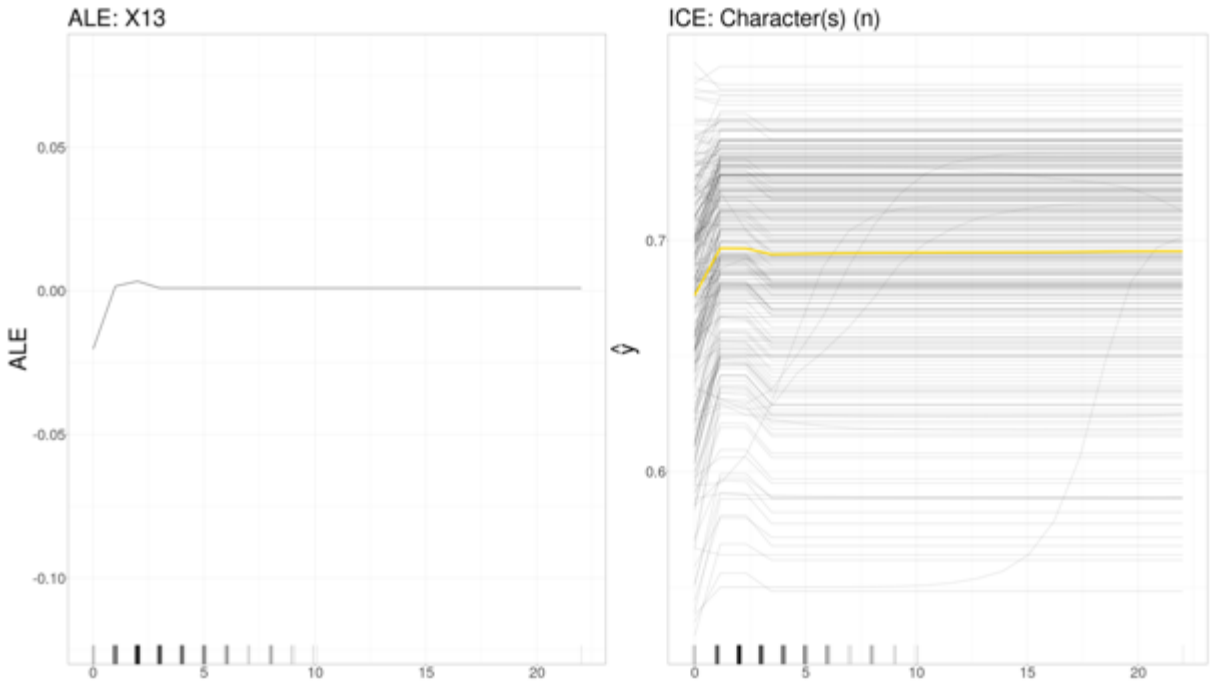

Figure 5. ALE and ICE plots for number of characters on screen

We use Shapley values (Lundberg & Lee 2017; Shapley 1953; Štrumbelj & Kononenko 2014) for the best advertisement. The results are shown in Figure 6. The estimation software automatically assigns the 'positive' category, and in this case, it is the worst category. The figure shows that what makes the best advertisement least bad is that the product or package is shown immediately, and that it has no characters. Interestingly, although the product or package is shown immediately, the brand name is not identified until the middle third of the advertisement. We can see three of the four variables common to the best interpretable model here: time until product/packaging is shown (X06), number of characters (X13) and brand name introduced in the first, middle or last third (X05PROP_2).

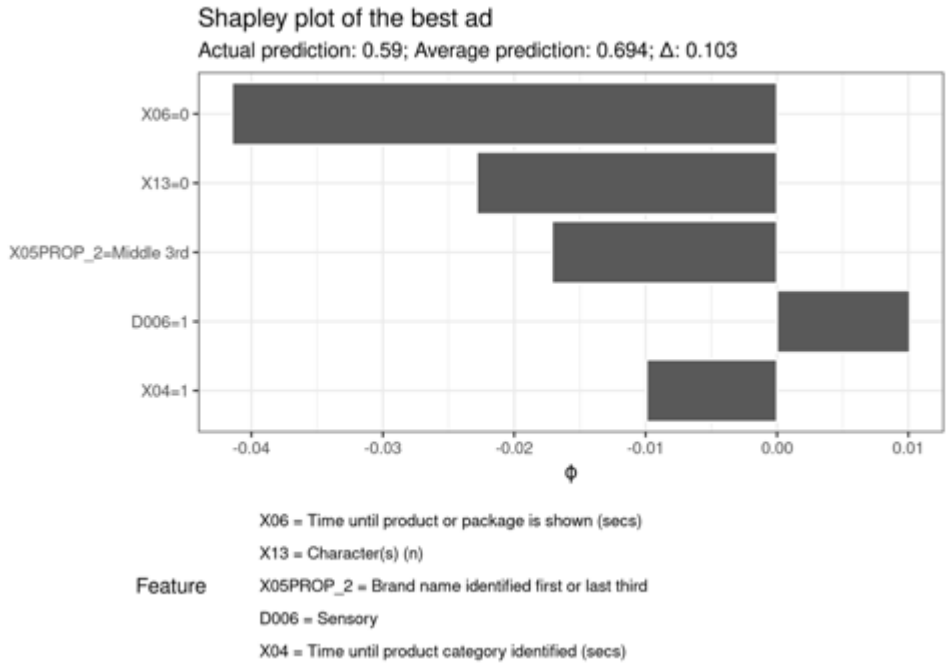

Figure 6. Shapley plot for the best advertisement

At the other end of the spectrum, Figure 7 shows that the worst advertisement is characterised by not having the product or package shown until four seconds have elapsed, and the brand name is introduced in the first third of the advertisement. These are the only two variables in the plot that are common to the best interpretable model. Shapley values explain what contributes to the case class probability being different from the average class probability, so these values should not be over-interpreted, and are presented here simply as an example of the information that is available to a marketing executive with extensive domain knowledge, who can look at the advertisement, interpret the Shapley plot and judge whether the Shapley plot makes sense.

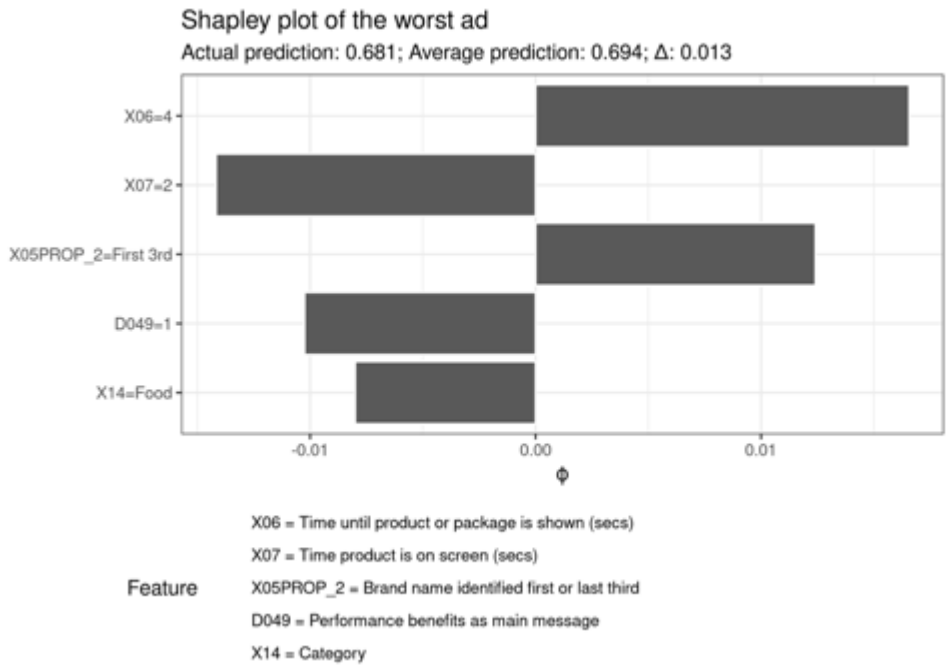

Figure 7. Shapley plot for the worst advertisement

In summary, we have seen broad agreement between the interpretable logistic regression model and the (until-recently regarded as) uninterpretable Deep Learning model.

Discussion

This research shines a light on the single method bias inherent to many studies of creative effects, addressing a critical issue that is relevant to marketing research most broadly. Our method-agnostic approach demonstrates the multitude of analytic methods that can help solve advertising problems, and that using as many of them as possible enables avoidance of single-method bias and enhanced confidence in the robustness of the results.

Implications for academic research

Previous research on advertising effectiveness has almost exclusively used a single analytical method to evaluate the effects of creative variables. In this paper, we show that each method has advantages and disadvantages and indicates different creative variables as important, therefore, highlighting the risk of producing conflicting results when using a single method.

We further show that ML and AI methods generally give better fit and prediction than the classical method used previously to analyse these data.

Researchers should consider the advantages of applying multiple ML and AI methods to their data rather than defaulting to only the classical methods. We do not claim that every AI or ML method will be superior to all traditional methods, in every circumstance; rather we are advocating being method-agnostic. This advice may seem daunting to users of traditional software like SPSS, but modern software such as R and Python have packages that automate much of the work involved.

In this analysis, several important creative variables generalised across a wide range of models to predict high performing advertisements. These creative variables were not shown as statistically significant by the previously reported classical ordinal regression model (Hartnett *et al.* 2016a), which identified creative devices such as humour (or music) as increasing (or decreasing) the odds of sales effectiveness for example. The newly revealed important creative variables from the best ML and AI models are predominantly related to the timing of visual branding elements, including *when* the brand name and product are introduced, and *how long* the brand name is present.

Our novel findings highlight branding tactics require careful consideration and execution by advertisers. In practice, advertising must strike a balance between branding and creative (non-branding) elements. Creative tactics, such as humour or character-driven narrative, are important for drawing attention to advertising. Executing branding effectively, however, such that the brand is noticed among those creative tactics, is necessary if advertisements are to build useful memory associations anchored to the *brand*, to then affect the probability of the brand to come to mind at the next purchase occasion.

We found that *early* product introduction was linked to improved sales. This is a reasonably well-supported tactic in the literature, linked primarily to improved related recall

(Stewart & Furse 1986; Stewart & Koslow 1989), and also persuasion (Stanton & Burke 1998). We further found that *later* brand name introduction was also linked to improved sales. It may seem that *early* product introduction and *later* brand name introduction run counter to one another because the brand is (often) an inherent part of the product or packaging. However, it is possible to show the actual product and its use without the brand name overtly present (e.g., a bowl of cat food).

On later brand name introduction, the literature presents varied findings. Numerous studies link earlier branding with improved related recall (Baker *et al.* 2004; Romaniuk 2009; Stewart & Furse 1986; Stewart & Koslow 1989; Walker & von Gonten 1989), however, others have found nil or negative effects using the same measure (Phillips & Stanton 2004; Stanton & Burke 1998). A most recent study analysed a series of controlled lab experiments and found that shorter advertisements with early branding had significantly lower related recall than longer advertisements with late branding, concluding advertisement length is more important than branding for memory effects (Varan *et al.* 2020). For the reasons outlined in the introduction, we again acknowledge this is a complex research area, spanning different advertising outcomes that are themselves not strongly correlated (e.g., related recall and sales as reported by Lodish *et al.* 1995).

We also found that *shorter* branding duration (i.e., the total amount of time the brand name is shown) positively affects sales. Shorter duration goes together with later introduction because there is less scope for showing the brand name for longer. There is also limited convergence across the literature regarding branding duration effects. Several past studies have failed to observe a relationship between branding duration and memory measures (Romaniuk 2009; Stewart & Furse 1986; Stewart & Koslow 1989). One experiment found *static* longer branding duration triggered viewers to avoid advertising, but this could be subverted if executed as a series of shorter pulses (Teixeira *et al.* 2010). Another study

reported that branding duration, among branding cues, has the largest effect on long-term advertising elasticity, where small brands had the greatest scope for improved effectiveness (Bruce *et al.* 2020). Perhaps time spent showing the brand does not always compensate for *how* the brand is integrated as part of the creative narrative or is more important for brands that lack awareness and familiarity.

Based on our findings, we propose an empirically derived theory that showing the product early serves as a contextually important cue to set up introducing the brand. Presenting relevant information in this manner arguably works by strengthening the memory association between the product category, brand name, and response to the advertising content (Baker *et al.* 2004). We do suggest more studies, ideally applying the MDA paradigm, are needed to provide greater clarity around conditions when specific branding tactics are more or less important for supporting superior *sales*. These studies should also consider elements that consumers uniquely associate with brands, such as symbols, characters, or taglines, which can be leveraged to indirectly signal the brand if/when downplaying the brand name (Hartnett *et al.* 2016b; Romaniuk 2018).

Implications for advertising practice

Marketing managers are already aware of the importance of being data-driven but often struggle to implement data-driven strategies due to perceived high cost of hardware and software, training, and most importantly hiring and retaining skilled employees. However, the barriers are falling; for moderate sample sizes (e.g., a few 100,000 cases and less than 10,000 variables), state of the art models can run on commodity hardware using open-source software by non-experts.

The implication of this research for advertisers is that they can likely generate new insights by applying multiple models to their advertising data, increasing the breadth of creative variables at their disposal to improve creative effectiveness. In a creative industry

such as advertising production, having more options is favourable to avoid being overly constricted in the execution of creative ideas.

Importantly, we do not wish to imply that these creative variables guarantee sales effective advertisements; not all advertisements that use the creative variables highlighted here will be successful. There is no foolproof formula for how to make a great, sales driving advertisement. The findings do, however, highlight a handful of creative tactics (e.g., early product introduction, later branding) that proved more consistently effective than other tactics when subjected to varied analytic testing. From this most comprehensive analysis, we encourage marketers to open a discussion with their advertising agencies about these creative variables, which present greater odds for success, and check their creatives for how the product and brand are incorporated.

Limitations

We were limited by a relatively small but costly sample of advertisements linked to individual-level sales effects, which is among the largest and highest quality data in this area of research. We applied many algorithms that were designed for large samples, and a paradigm designed for large samples (splitting the sample into test and training sets), which means that the results from these methods may be not as good as if a larger sample were available. However, methods designed for *any* sample size were also used. Some ML and AI models can give similar or better fit than the classical logistic regression when applying to small samples of training data, while they are superior in terms of prediction of test data. We do not claim that the consistency of accuracy measures across ML and AI and traditional methods for small samples is a general finding, but it has been shown in the literature, applied to even smaller data sets than ours (Panesar *et al.* 2019).

We used human coders to observe and record the creative variables in advertisements, which could be perceived as contradictory to the research intent; *modern* approaches and

application of newer ML and AI models. Relying on human coders is slower and prone to human error compared to computer programs, which can nowadays extract a variety of video characteristics from advertisements on a frame-by-frame basis, such as colorfulness, scene cuts, and the presence of faces and facial expression (Schwenzow *et al.* 2021). However, many of the creative variables captured in this study cannot be sufficiently recognised by machines, such as the use of humour, an unusual setting for product usage, or mood creating music. Perhaps one day it will be possible for machines to code these creative variables, but for now, recruiting human coders was necessary for our purposes.

References

Apley, D.W. and Zhu, J. (2016) Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. Available at: https://arxiv.org/abs/1612.08468.

Armstrong, J.S. (2011a) Evidence-Based Advertising: An Application to Persuasion. *International Journal of Advertising*, **20**, 5, pp.743-67.

Armstrong, J.S. (2011b) Reply to the Comments on 'Evidence-Based Advertising: An Application to Persuasion'. *International Journal of Advertising*, **30**, 5, pp.790-94.

Baker, W.E., Honea, H., and Russel, C.A. (2004) Do Not Wait to Reveal the Brand Name: The Effect of Brand-Name Placement on Television Advertising Effectiveness. *Journal of Advertising*, **33**, 3, pp.77-85.

Becker, M., Scholdra, T.P., Berkmann, M., and Reinartz, W.J. (2022) The Effect of Content on Zapping in Tv Advertising. *Journal of Marketing*, **ahead-of-print**, pp.1-105.

Bellman, S., Nenycz-Thiel, M., Kennedy, R., Hartnett, N., and Varan, D. (2019) Best Measures of Attention to Creative Tactics in Tv Advertising: When Do Attention-Getting Devices Capture or Reduce Attention? *Journal of Advertising Research*, **58**, 4, pp.295-311.

Bellman, S., Nenycz-Thiel, M., Kennedy, R., Larguinat, L., McColl, B., and Varan, D. (2017) What Makes a Television Commercial Sell? Using Biometrics to Identify Successful Ads. *Journal of Advertising Research*, **56**, 4, pp.1-14.

Bellman, S., Schweda, A., and Varan, D. (2012) Interactive Tv Advertising: Itv Ad Executional Factors. *Journal of Business Research*, **65**, 6, pp.831-39.

Bruce, N.I., Becker, M., and Reinartz, W. (2020) Communicating Brands in Television Advertising. *Journal of Marketing Research*, **57**, 2, pp.236-56.

Carlson, L. (2011) Comments on Scott Armstrong's 'Evidence-Based Advertising: An Application to Persuasion' - Tastes Great but Less Filling (Than It Could Have Been). *International Journal of Advertising*, **30**, 5, pp.769-74.

Chandy, R.K., Tellis, G.J., MacInnis, D.J., and Thaivanich, P. (2001) What to Say When: Advertising Appeals in Evolving Markets. *Journal of Marketing Research*, **38**, 4, pp.399-414.

Chen, H., Cohen, P., and Chen, S. (2010) How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics—Simulation and Computation*, **39**, 4, pp.860-64.

Chintalapati, S. and Pandey, S.K. (2022) Artificial Intelligence in Marketing: A Systematic Literature Review. *International Journal of Market Research*, **64**, 1, pp.38-68.

Dall'Olio, F. and Vakratsas, D. (2022) The Impact of Advertising Creative Strategy on Advertising Elasticity. *Journal of Marketing*, **ahead-of-print**, pp.1-19.

Freelon, D.G. (2010) Recal: Intercoder Reliability Calculation as a Web Service. *International Journal of Internet Science*, **5**, 1, pp.20-33.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015) Peeking inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, **24**, 1, pp.44-65.

Guitart, I.A. and Stremersch, S. (2021) The Impact of Informational and Emotional Television Ad Content on Online Search and Sales. *Journal of Marketing Research*, **58**, 2, pp.299-320.

Hartnett, N., Kennedy, R., Sharp, B., and Greenacre, L. (2016a) Creative That Sells: How Advertising Execution Affects Sales. *Journal of Advertising*, **45**, 1, pp.102-12.

Hartnett, N., Romaniuk, J., and Kennedy, R. (2016b) Comparing Direct and Indirect Branding in Advertising. *Australasian Marketing Journal*, **24**, 1, pp.20-28.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression.* New York, NY, United States: Wiley Series in Probability and Statistics.

Jones, J.P. (1995) Single-Source Research Begins to Fulfill Its Promise. *Journal of Advertising Research*, **35**, 3, pp.9-16.

Jones, J.P. (2007) *When Ads Work: New Proof That Advertising Triggers Sales.* New York: M.E. Sharpe, Inc.

Kim, J., Freling, T.H., and Grisaffe, D.B. (2013) The Secret Sauce for Super Bowl Advertising: What Makes Marketing Work in the World's Most Watched Event. *Journal of Advertising Research*, **53**, 2, pp.134-49.

Kuhn, M. (2017) Caret: Classification and Regression Training. Available at: https://CRAN.R-project.org/package=caret.

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling.* New York: Springer.

Laskey, H.A., Fox, R.J., and Crask, M.R. (1994) Investigating the Impact of Executional Style on Television Commercial Effectiveness. *Journal of Advertising Research*, **34**, 6, pp.9-16.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., and Kraljevic, T. (2018) H2o: R Interface for 'H2o'. Available at: https://github.com/h2oai/h2o-3.

Lodish, L.M., Abraham, M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., and Stevens, M.E. (1995) How Tv Advertising Works: A Meta-Analysis of 389 Real World Split Cable Tv Advertising Experiments. *Journal of Marketing Research*, **32**, 2, pp.125-39.

Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. In *31st Conference on Neural Information Processing Systems*, Long Beach, California, United States, pp.4765-74.

McEwen, W.J. and Leavitt, C. (1976) A Way to Describe Tv Commercials. *Journal of Advertising Research*, **16**, 6, pp.35-39.

Molnar, C. and Schratz, P. (2020) Iml: Interpretable Machine Learning. Available at: https://christophm.github.io/interpretable-ml-book/.

Mustak, M., Salminen, J., Plé, L., and Wirtz, J. (2021) Artificial Intelligence in Marketing: Topic Modeling, Scientometric Analysis, and Research Agenda. *Journal of Business Research*, **124**, pp.389-404.

Olney, T.J., Holbrook, M.B., and Batra, R. (1991) Consumer Responses to Advertising: The Effects of Ad Content, Emotions, and Attitude toward the Ad on Viewing Time. *Journal of Consumer Research*, **17**, 4, pp.440-53.

Panesar, S.S., D'Souza, R.N., Yeh, F.-C., and Fernandez-Miranda, J.C. (2019) Machine Learning Versus Logistic Regression Methods for 2-Year Mortality Prognostication in a Small, Heterogeneous Glioma Database. *World Neurosurgery: X*, **2**, pp.1-12.

Phillips, D.M. and Stanton, J.L. (2004) Age-Related Differences in Advertising: Recall and Persuasion. *Journal of Targeting, Measurement and Analysis for Marketing*, **13**, 1, pp.7-20.

Roberts, A. (1996) What Do We Know About Advertising's Short-Term Effects? *Admap*, February, pp.42-45.

Romaniuk, J. (2009) The Efficacy of Brand-Execution Tactics in Tv Advertising, Brand Placements and Internet Advertising. *Journal of Advertising Research*, **49**, 2, pp.143-50.

Romaniuk, J. (2018) *Building Distinctive Brand Assets.* South Melbourne, Victoria: Oxford University Press.

Rust, R.T. and Cooil, B. (1994) Reliability Measures for Qualitative Data: Theory and Implications. *Journal of Marketing Research*, **31**, 1, pp.1-14.

Schwenzow, J., Hartmann, J., Schikowsky, A., and Heitmann, M. (2021) Understanding Videos at Scale: How to Extract Insights for Business Research. *Journal of Business Research*, **123**, pp.367-79.

Shapley, L.S. (1953) A Value for N-Person Games: Contributions to the Theory of Games. *Annals of Mathematic Studies*, **2**, 28, pp.307-17.

Stanton, J.L. and Burke, J. (1998) Comparative Effectiveness of Executional Elements in Tv Advertising 15 Versus 30 Second Commercials. *Journal of Advertising Research*, **38**, 6, pp.7-13.

Stewart, D.W. and Furse, D.H. (1986) *Effective Television Advertising: A Study of 1000 Commercials.* Lexington, MA: Lexington Books.

Stewart, D.W. and Koslow, S. (1989) Executional Factors and Advertising Effectiveness: A Replication. *Journal of Advertising*, **18**, 3, pp.21-32.

Štrumbelj, E. and Kononenko, I. (2014) Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems*, **41**, 3, pp.647-65.

Teixeira, T., Wedel, M., and Pieters, R. (2010) Moment-to-Moment Optimal Branding in Tv Commercials: Preventing Avoidance by Pulsing. *Marketing Science*, **29**, 5, pp.783-804.

Tellis, G.J. (2004) *Effective Advertising: Understanding When, How and Why Advertising Works.* SAGE.

Varan, D., Nenycz-Thiel, M., Kennedy, R., and Bellman, S. (2020) The Effects of Commercial Length on Advertising Impact: What Short Advertisements Can and Cannot Deliver. *Journal of Advertising Research*, **60**, 1, pp.54-70.

Vlačić, B., Corbo, L., e Silva, S.C., and Dabić, M. (2021) The Evolving Role of Artificial Intelligence in Marketing: A Review and Research Agenda. *Journal of Business Research*, **128**, pp.187-203.

Walker, D. and von Gonten, M.F. (1989) Explaining Related Recall Outcomes: New Answers from a Better Model. *Journal of Advertising Research*, **29**, 3, pp.11-21.

Williams, J. (2021) Modern Data Analysis-a Paradigm for Robustness: Lessons for Marketing Researchers from the Machine Learning Literature. In R. Nunkoo, V. Teeroovengadum and C.M. Ringle (eds)*Handbook of Research Methods for Marketing Management*, Cheltenham, UK: Edward Elgar Publishing, pp. 91-127.

Yee, T.W. (2015) *Vector Generalized Linear and Additive Models: With an Implementation in R.* New York, United States: Springer.

Yelkur, R., Tomkovick, C., Hofer, A., and Rozumalski, D. (2013) Super Bowl Ad Likeability: Enduring and Emerging Predictors. *Journal of Marketing Communications*, **19**, 1, pp.58-80.

Zhao, H., Williams, G.J., and Huang, J.Z. (2017) Wsrf: An R Package for Classification with Scalable Weighted Subspace Random Forests. *Journal of Statistical Software*, **77**, 3, pp.1-30.

Appendix: Quality Indices for All Models

dv	iv	method	recall_te	bacc_te	auc_te	acc_te	f1_te	auc_tr	acc_tr	f1_tr
Binary	Factors	aml_deeplearning	0.75	0.75	0.66	0.75	0.84	1	1	1
Binary	Factors	rlda	0.61	0.56	0.66	0.55	0.6	0.71	0.62	0.64
Binary	Factors	Linda	0.57	0.58	0.58	0.58	0.64	0.68	0.62	0.69
Binary	Factors	pcaNNet	0.54	0.57	0.48	0.58	0.64	1	1	1
Binary	Factors	dda	0.5	0.56	0.59	0.58	0.65	0.73	0.66	0.7
Binary	Factors	kknn	0.5	0.57	0.57	0.59	0.67	1	1	1
Binary	Factors	mda	0.5	0.5	0.51	0.5	0.56	0.93	0.86	0.89
Binary	Factors	nnet	0.5	0.54	0.6	0.55	0.62	1	1	1
Binary	Factors	mlp	0.46	0.54	0.57	0.57	0.65	0.93	0.94	0.96
Binary	Factors	rpart1SE	0.46	0.59	0.59	0.62	0.7	0.89	0.84	0.88
Binary	Factors	LogitBoost	0.43	0.51	0.43	0.53	0.61	0.78	0.75	0.81
Binary	Factors	naive_bayes	0.43	0.58	0.58	0.62	0.71	0.75	0.73	0.8
Binary	Factors	vglmCumulative	0.43	0.53	0.52	0.55	0.64	0.75	0.7	0.78
Binary	Factors	QdaCov	0.39	0.55	0.56	0.59	0.69	0.8	0.8	0.84
Binary	Factors	svmRadial	0.36	0.61	0.69	0.67	0.77	0.94	0.88	0.91
Binary	Factors	gamSpline	0.32	0.56	0.51	0.62	0.72	0.76	0.7	0.78
Binary	Factors	gpls	0.32	0.56	0.62	0.62	0.72	0.73	0.69	0.78
Binary	Factors	mlpML	0.32	0.47	0.53	0.51	0.62	0.96	0.97	0.98
Binary	Factors	multinom	0.32	0.58	0.57	0.64	0.75	0.72	0.69	0.78
Binary	Factors	pls	0.32	0.55	0.57	0.61	0.71	0.73	0.67	0.77
Binary	Factors	rotationForest	0.32	0.51	0.52	0.57	0.67	0.99	0.97	0.98
Binary	Factors	svmRadialCost	0.32	0.62	0.66	0.7	0.79	0.95	0.76	0.84
Binary	Factors	vglmAdjCat	0.32	0.53	0.51	0.58	0.69	0.76	0.7	0.78
Binary	Factors	bayesglm	0.29	0.55	0.65	0.62	0.73	0.74	0.7	0.79
Binary	Factors	C5.0	0.29	0.5	0.51	0.55	0.67	1	0.99	0.99
Binary	Factors	glm	0.29	0.52	0.59	0.58	0.69	0.74	0.7	0.79
Binary	Factors	treebag	0.29	0.53	0.57	0.59	0.7	1	1	1
Binary	Factors	vglmContRatio	0.29	0.51	0.48	0.57	0.68	0.75	0.73	0.81
Binary	Factors	wsrf	0.29	0.58	0.61	0.66	0.76	1	1	1
Binary	Factors	nb	0.25	0.45	0.52	0.5	0.62	0.74	0.69	0.76
Binary	Factors	pda	0.25	0.56	0.56	0.64	0.76	0.73	0.7	0.79
Binary	Factors	rda	0.25	0.5	0.56	0.57	0.69	0.96	0.91	0.93
Binary	Factors	regLogistic	0.25	0.52	0.49	0.59	0.71	0.76	0.73	0.8
Binary	Factors	gamLoess	0.21	0.47	0.51	0.54	0.67	0.99	0.96	0.97
Binary	Factors	simpls	0.21	0.49	0.53	0.57	0.69	0.74	0.72	0.8
Binary	Factors	svmLinearWeights	0.21	0.59	0.63	0.68	0.79	0.68	0.69	0.79
Binary	Factors	hda	0.18	0.45	0.59	0.53	0.66	0.6	0.63	0.76
Binary	Factors	kernelpls	0.18	0.54	0.49	0.63	0.75	0.73	0.72	0.8
Binary	Factors	RRFglobal	0.18	0.53	0.53	0.62	0.74	1	1	1
Binary	Factors	svmLinear2	0.18	0.56	0.71	0.66	0.78	0.69	0.69	0.79
Binary	Factors	svmPoly	0.18	0.52	0.57	0.61	0.73	1	0.98	0.98
Binary	Factors	parRF	0.14	0.49	0.49	0.58	0.71	1	1	1

dv	iv	method	recall_te	bacc_te	auc_te	acc_te	f1_te	auc_tr	acc_tr	f1_tr
Binary	Factors	rf	0.14	0.54	0.52	0.64	0.77	1	1	1
Binary	Factors	C5.0Rules	0.11	0.5	0.5	0.61	0.74	0.35	0.73	0.82
Binary	Factors	evtree	0.11	0.5	0.5	0.61	0.74	0.42	0.68	0.79
Binary	Factors	gamboost	0.11	0.53	0.65	0.64	0.77	0.78	0.73	0.82
Binary	Factors	rbfDDA	0.07	0.53	0.51	0.64	0.78	0.78	0.83	0.88
Binary	Factors	rotationForestCp	0.07	0.49	0.47	0.61	0.75	0.35	0.67	0.79
Binary	Factors	slda	0.07	0.49	0.55	0.61	0.75	0.63	0.69	0.79
Binary	Factors	svmRadialSigma	0.04	0.52	0.66	0.64	0.78	0.97	0.69	0.8
Binary	Factors	C5.0Tree	0	0.5	0.5	0.63	0.77	0.5	0.63	0.77
Binary	Factors	mlpWeightDecay	0	0.5	0.52	0.63	0.77	0.6	0.63	0.77
Binary	Factors	mlpWeightDecayML	0	0.5	0.64	0.63	0.77	0.6	0.63	0.77
Binary	Factors	pda2	0	0.5	0.66	0.63	0.77	0.7	0.63	0.77
Binary	Factors	rpart	0	0.5	0.5	0.63	0.77	0.5	0.63	0.77
Binary	Factors	rpart2	0	0.47	0.53	0.59	0.74	0.42	0.69	0.8
Binary	Factors	svmLinear	0	0.5	0.56	0.63	0.77	0.79	0.63	0.77
Binary	Factors	svmRadialWeights	0	0.5	0.59	0.63	0.77	1	0.69	0.8
Binary	Raw	vglmAdjCat	0.71	0.7	0.73	0.7	0.74	0.86	0.79	0.84
Binary	Raw	gamLoess	0.68	0.61	0.64	0.59	0.63	0.96	0.88	0.91
Binary	Raw	multinom	0.57	0.61	0.61	0.62	0.68	0.86	0.78	0.82
Binary	Raw	gamSpline	0.54	0.52	0.51	0.51	0.56	0.88	0.82	0.86
Binary	Raw	nnet	0.54	0.54	0.55	0.54	0.6	0.94	0.88	0.9
Binary	Raw	vglmContRatio	0.54	0.55	0.56	0.55	0.61	0.9	0.85	0.88
Binary	Raw	kknn	0.5	0.57	0.57	0.59	0.67	1	1	1
Binary	Raw	mda	0.46	0.51	0.49	0.53	0.6	0.97	0.92	0.93
Binary	Raw	C5.0Tree	0.43	0.47	0.56	0.49	0.56	0.98	0.95	0.96
Binary	Raw	gam	0.43	0.46	0.52	0.47	0.55	1	1	1
Binary	Raw	glm	0.43	0.52	0.55	0.54	0.62	0.89	0.82	0.86
Binary	Raw	pda	0.43	0.53	0.59	0.55	0.64	0.87	0.79	0.84
Binary	Raw	RRFglobal	0.43	0.63	0.6	0.68	0.77	1	1	1
Binary	Raw	vglmCumulative	0.43	0.53	0.58	0.55	0.64	0.94	0.97	0.97
Binary	Raw	C5.0Rules	0.39	0.57	0.44	0.62	0.71	0.85	0.86	0.89
Binary	Raw	pcaNNet	0.36	0.43	0.54	0.45	0.53	1	1	1
Binary	Raw	aml_deeplearning	0.33	0.5	0.51	0.69	0.81	1	1	1
Binary	Raw	bayesglm	0.32	0.42	0.54	0.45	0.54	0.85	0.79	0.84
Binary	Raw	slda	0.32	0.56	0.51	0.62	0.72	0.73	0.69	0.78
Binary	Raw	C5.0	0.29	0.56	0.47	0.63	0.74	1	1	1
Binary	Raw	treebag	0.29	0.55	0.59	0.62	0.73	1	0.99	0.99
Binary	Raw	rpart1SE	0.21	0.47	0.55	0.54	0.67	0.79	0.78	0.84
Binary	Raw	evtree	0.14	0.5	0.5	0.59	0.73	0.42	0.67	0.78
Binary	Raw	wsrf	0.14	0.52	0.51	0.62	0.75	1	1	1
Binary	Raw	rpart2	0.11	0.52	0.48	0.63	0.76	0.45	0.66	0.78
Binary	Raw	rf	0.07	0.51	0.61	0.63	0.77	1	1	1
Binary	Raw	nb	0.04	0.51	0.49	0.63	0.77	0.81	0.63	0.77

dv	iv	method	recall_te	bacc_te	auc_te	acc_te	f1_te	auc_tr	acc_tr	f1_tr
Binary	Raw	parRF	0.04	0.47	0.54	0.58	0.73	1	1	1
Binary	Raw	naive_bayes	0	0.5	0.57	0.63	0.77	0.82	0.63	0.77
Binary	Raw	pda2	0	0.5	0.53	0.63	0.77	0.74	0.63	0.77
Binary	Raw	rpart	0	0.5	0.5	0.63	0.77	0.5	0.63	0.77
Nominal	Factors	vbmpRadial	0.96	0.49	0.49	0.33	0.49	0.84	1	1
Nominal	Factors	C5.0	0.8	0.46	0.54	0.31	0.45	0.43	0.44	0.54
Nominal	Factors	wsrf	0.64	0.62	0.63	0.45	0.52	0.87	1	1
Nominal	Factors	mda	0.6	0.53	0.55	0.33	0.45	0.78	0.74	0.76
Nominal	Factors	svmRadialWeights	0.6	0.59	0.61	0.41	0.49	0.84	0.73	0.74
Nominal	Factors	vglmContRatio	0.6	0.6	0.63	0.47	0.5	0.68	0.52	0.58
Nominal	Factors	C5.0Rules	0.56	0.63	0.61	0.47	0.52	0.77	0.82	0.86
Nominal	Factors	nnet	0.56	0.58	0.54	0.43	0.47	0.71	0.55	0.69
Nominal	Factors	pda	0.56	0.64	0.6	0.44	0.53	0.67	0.57	0.57
Nominal	Factors	RRFglobal	0.56	0.59	0.55	0.37	0.48	0.85	1	1
Nominal	Factors	vglmAdjCat	0.56	0.56	0.63	0.44	0.46	0.69	0.48	0.56
Nominal	Factors	hda	0.52	0.54	0.48	0.33	0.43	0.66	0.49	0.56
Nominal	Factors	multinom	0.52	0.62	0.61	0.43	0.5	0.68	0.54	0.52
Nominal	Factors	rpart2	0.52	0.53	0.55	0.37	0.43	0.64	0.56	0.61
Nominal	Factors	svmLinear2	0.52	0.61	0.58	0.44	0.49	0.69	0.6	0.61
Nominal	Factors	svmRadialSigma	0.52	0.54	0.55	0.37	0.43	0.77	0.67	0.69
Nominal	Factors	vglmCumulative	0.52	0.53	0.56	0.39	0.43	0.67	0.44	0.49
Nominal	Factors	dda	0.48	0.6	0.56	0.41	0.47	0.65	0.54	0.52
Nominal	Factors	kernelpls	0.48	0.48	0.52	0.33	0.38	0.73	0.53	0.62
Nominal	Factors	polr	0.48	0.51	0.52	0.29	0.4	0.71	0.51	0.61
Nominal	Factors	svmRadial	0.48	0.58	0.61	0.4	0.45	0.78	0.74	0.77
Nominal	Factors	svmRadialCost	0.48	0.52	0.56	0.37	0.41	0.8	0.61	0.76
Nominal	Factors	C5.0Tree	0.44	0.52	0.56	0.39	0.39	0.88	0.94	0.93
Nominal	Factors	mlp	0.44	0.62	0.58	0.43	0.48	0.81	0.84	0.81
Nominal	Factors	pcaNNet	0.44	0.6	0.62	0.37	0.46	0.86	0.87	0.88
Nominal	Factors	pls	0.44	0.5	0.54	0.35	0.38	0.68	0.57	0.6
Nominal	Factors	regLogistic	0.44	0.61	0.61	0.39	0.47	0.69	0.57	0.55
Nominal	Factors	svmLinear	0.44	0.6	0.58	0.45	0.46	0.7	0.5	0.5
Nominal	Factors	kknn	0.4	0.55	0.6	0.37	0.4	0.85	0.92	0.92
Nominal	Factors	Linda	0.4	0.52	0.51	0.39	0.38	0.63	0.52	0.5
Nominal	Factors	mlpWeightDecay	0.4	0.56	0.59	0.37	0.41	0.71	0.7	0.71
Nominal	Factors	mlpWeightDecayML	0.4	0.55	0.59	0.37	0.4	0.77	0.74	0.76
Nominal	Factors	rda	0.4	0.55	0.48	0.39	0.4	0.67	0.56	0.61
Nominal	Factors	aml_deeplearning	0.37	0.53	0.56	0.37	0.48	0.87	1	1
Nominal	Factors	evtree	0.36	0.51	0.51	0.36	0.35	0.75	0.73	0.73
Nominal	Factors	rlda	0.36	0.51	0.53	0.32	0.35	0.68	0.53	0.58
Nominal	Factors	simpls	0.36	0.52	0.55	0.37	0.36	0.7	0.56	0.59
Nominal	Factors	svmPoly	0.36	0.55	0.7	0.39	0.38	0.89	1	1
Nominal	Factors	nb	0.32	0.49	0.54	0.32	0.32	0.8	0.83	0.82

dv	iv	method	recall_te	bacc_te	auc_te	acc_te	f1_te	auc_tr	acc_tr	f1_tr
Nominal	Factors	rpart	0.32	0.5	0.49	0.36	0.33	0.67	0.53	0.57
Nominal	Factors	rpart1SE	0.32	0.49	0.58	0.39	0.32	0.74	0.72	0.73
Nominal	Factors	LogitBoost	0.29	0.49	0.63	0.38	0.3	0.76	0.8	0.83
Nominal	Factors	mlpML	0.28	0.48	0.55	0.31	0.29	0.74	0.87	0.87
Nominal	Factors	naive_bayes	0.28	0.51	0.55	0.36	0.31	0.65	0.56	0.51
Nominal	Factors	parRF	0.28	0.46	0.57	0.32	0.28	0.84	1	1
Nominal	Factors	pda2	0.28	0.51	0.55	0.39	0.31	0.69	0.52	0.55
Nominal	Factors	rf	0.28	0.49	0.5	0.35	0.3	0.84	1	1
Nominal	Factors	slda	0.28	0.44	0.53	0.32	0.27	0.62	0.51	0.48
Nominal	Factors	treebag	0.28	0.52	0.5	0.32	0.32	0.81	1	1
Nominal	Raw	parRF	0.56	0.59	0.61	0.44	0.48	0.85	1	1
Nominal	Raw	wsrf	0.56	0.48	0.6	0.32	0.41	0.85	1	1
Nominal	Raw	kknn	0.52	0.63	0.55	0.33	0.51	0.83	1	1
Nominal	Raw	slda	0.52	0.6	0.6	0.4	0.48	0.64	0.52	0.53
Nominal	Raw	vglmCumulative	0.52	0.54	0.58	0.36	0.43	0.78	0.55	0.65
Nominal	Raw	nnet	0.48	0.55	0.55	0.43	0.43	0.87	0.73	0.73
Nominal	Raw	RRFglobal	0.48	0.56	0.63	0.37	0.44	0.85	1	1
Nominal	Raw	rf	0.44	0.56	0.58	0.37	0.42	0.86	1	1
Nominal	Raw	rpart2	0.44	0.44	0.55	0.29	0.34	0.7	0.53	0.55
Nominal	Raw	pcaNNet	0.4	0.53	0.47	0.33	0.38	0.78	0.77	0.83
Nominal	Raw	rpart1SE	0.4	0.52	0.47	0.36	0.38	0.7	0.64	0.69
Nominal	Raw	treebag	0.4	0.51	0.48	0.36	0.37	0.87	1	1
Nominal	Raw	vglmAdjCat	0.4	0.57	0.59	0.36	0.42	0.8	0.58	0.66
Nominal	Raw	C5.0Tree	0.36	0.55	0.53	0.41	0.38	0.81	0.92	0.93
Nominal	Raw	pda	0.32	0.46	0.55	0.32	0.3	0.77	0.73	0.77
Nominal	Raw	vglmContRatio	0.32	0.5	0.62	0.33	0.33	0.84	0.86	0.82
Nominal	Raw	aml_deeplearning	0.31	0.49	0.52	0.31	0.5	0.88	1	1
Nominal	Raw	evtree	0.24	0.53	0.56	0.4	0.3	0.72	0.62	0.6
Nominal	Raw	multinom	0.2	0.48	0.49	0.31	0.24	0.79	0.81	0.83
Nominal	Raw	nb	0.2	0.39	0.66	0.19	0.2	0.77	0.6	0.65
Nominal	Raw	C5.0	0.16	0.42	0.52	0.29	0.18	0.71	0.7	0.71
Nominal	Raw	C5.0Rules	0.12	0.47	0.52	0.31	0.16	0.74	0.83	0.85
Nominal	Raw	mda	0.12	0.45	0.58	0.25	0.15	0.83	0.9	0.89
Nominal	Raw	naive_bayes	0.12	0.49	0.62	0.24	0.17	0.66	0.5	0.46
Nominal	Raw	rpart	0.12	0.5	0.51	0.36	0.18	0.45	0.46	0.3
Nominal	Raw	pda2	0	0.5	0.6	0.37	NA	0.7	0.37	0.03
Ordinal	Raw	custom	0.56	0.71	0.75	0.57	0.61	0.74	0.56	0.56
Ordinal	Raw	vglmAdjCat	0.56	0.54	0.55	0.37	0.44	0.78	0.54	0.65
Ordinal	Raw	vglmContRatio	0.56	0.58	0.55	0.37	0.47	0.79	0.55	0.62
Ordinal	Raw	vglmCumulative	0.44	0.55	0.59	0.43	0.42	0.8	0.83	0.77