Journal of Business Research 151 (2022) 324–338

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

The usefulness of socio-demographic variables in predicting purchase decisions: Evidence from machine learning procedures

Towhidul Islam [a,*], Nigel Meade [b], Richard T. Carson [c], Jordan J. Louviere [d], Juan Wang [e]

[a] Department of Marketing and Consumer Studies, University of Guelph, 50 Stone Road East, Guelph, Ontario N1G 2W1, Canada
[b] Imperial College Business School, 53 Princes Gate, South Kensington Campus, London SW7 2AZ, UK
[c] University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0021, United States
[d] University of South Australia, 101 Currie St, Adelaide, SA 5001, Australia
[e] Department of Marketing and Consumer Studies, University of Guelph, 50 Stone Road East, Guelph, Ontario N1G 2W1, Canada

A R T I C L E I N F O

Keywords:
Buy–no buy
Consumer targeting
Machine learning
Propensity score
Socio-demographic variables

A B S T R A C T

Research has long debated the effectiveness of socio-demographics in understanding purchase behavior, with mixed conclusions. The appeal of socio-demographic data for customer relationship marketing is based on its low acquisition cost and the growing array of variables on which marketers can condition messages and offers. We reinvestigate the value of socio-demographic variables, focusing on the potential of machine learning procedures (MLPs) to extract a stronger and reliable signal than the standard linear-in-parameters (logistic) regression models. We explore how predictive power can be increased through the nonlinearities and interactions identified with MLPs; our experimental set ranges from well-established procedures to newer entrants in this space. We also examine causality vis-à-vis predictability using a propensity scoring approach. Empirics are based on six grocery product categories and more than 7,000 panelists. We find that, relative to logistic regression models, MLPs using demographic variables yield a 20% to 33% improvement in out-of-sample predictive accuracy.

1. Introduction

Decades of conventional wisdom in market research suggest that demographic variables are of marginal value to most businesses in predicting consumer demand. In this paper, we revisit this issue and argue that answering the question of usefulness is much more nuanced than previously realized. We consider the following three factors underlying this conventional wisdom:

1. A common empirical finding is that when researchers possess information on prior purchase behavior, having individuals' demographic characteristics adds little predictive power (e.g., Ferber, 1962; Frank, Massy, & Boyd, 1967; Gupta & Chintagunta, 1994; Twedt, 1964). However, the argument that demographic variables are not useful to a business does not logically follow from these results. An obvious example is the case of completely new products, for which information on past purchases in the category is unavailable; similar logic holds for new entrants in an existing category, to the extent that they offer a differentiated alternative.
2. Previous research often fails to recognize the underlying behavioral sequence. Demographic variables predict past purchase behavior, but, conditional on observing past purchase behavior, these variables do not further improve predictability of current purchase behavior. Businesses should focus on the following questions: What is the difference in the predictive power of a model built only on demographic variables versus a model based on past purchase behavior, and what are the costs of obtaining these two types of variables? Low-cost demographic variables for households across many countries are now available, but this was far from the case during much of the period over which conventional wisdom on this topic developed. The variety of socio-demographic variables has also expanded dramatically and introduced new considerations. For example, liking a barbecue restaurant's page on a social media site does not reflect past purchase of barbecue sauce in a grocery store, but it may be a good targetable predictor of such behavior.
3. The notion that demographic variables are inadequate for forecasting purchase behavior comes from papers using linear-in-parameters models (e.g., variants of ordinary least squares [OLS] and logistic regression) and written when issues regarding how to

* Corresponding author.
E-mail addresses: islam@uoguelph.ca (T. Islam), n.meade@imperial.ac.uk (N. Meade), rcarson@ucsd.edu (R.T. Carson).

https://doi.org/10.1016/j.jbusres.2022.07.004
Received 18 March 2021; Received in revised form 27 June 2022; Accepted 1 July 2022

select predictive variables from a broad set and avoid overfitting were still largely unresolved (Diamantopoulos, Schlegelmilch, Sinkovics, & Bohlen, 2003; Laukkanen, 2016; Wheatley, Chiu, & Stevens, 1980; Zwick, 1957). The literature is full of examples in which researchers have found deviations from linear-in-parameters models to be important in specific situations, such as an inverted U-shaped relationship (e.g., Mihalopoulos & Demoussis, 2001; Ricciuto, Tarasuk, & Yatchew, 2006), and via interactions with other demographics (Laukkanen, 2016; Verbeke, 2005). Modern machine learning (ML) techniques were designed to avoid rigid specifications and provide ways to handle an abundance of possible predictors while avoiding overfitting, which can lead to poor-quality out-of-sample predictions.

Because most previous research on this topic comes from studies involving products sold in grocery stores, it is logical to use this context to reexamine the usefulness of traditional socio-demographic variables in the age of ML. Our results include a range of common grocery store products that can provide a useful benchmark for subsequent investigations involving substantively different product categories (e.g., new cars, mobile service providers) and novel covariates like media attention variables and categorizations of small spatial areas. To enhance the credibility of our results, we use purchase data of six product categories from more than 7000 panelists of Information Resources, Inc. (IRI), one of the best-known commercial sources for the type of data we use and an established data source for many academic papers.

Our null hypothesis is sharply focused: Do standard linear-in-parameters models extract all the useful information relevant to predicting the purchase behavior of interest (as many have long contended; Dawes & Corrigan, 1974) relative to currently available ML approaches? We examine a set of ML techniques because the well-known “no free lunch” theorems for optimization (Wolpert & Macready, 1997) suggest that different ML algorithms likely have varying strengths and weaknesses that are context-specific in the finite samples that characterize empirical research.

The set of ML techniques we consider includes a mix of established procedures known to perform well in many contexts, in addition to some newer procedures that appear promising for standard marketing research applications like ours. Our central focus is the question of whether ML procedures can squeeze more information out of standard socio-demographic covariates that would be useful for predicting purchases. Comparing and contrasting the relative performance of specific ML procedures with respect to individual products is often a useful adjunct to that objective.

We find that, in general, ML procedures do extract substantially more information from standard socio-demographic covariates in many instances. This result should lead marketing researchers to give their use a fresh look and prompt practitioners to think about the cost of acquisition compared with value in targeting. Our experience, and that of many top data scientists, indicates that the relative performance of different ML procedures across different products displays a fair amount of variability. Conventional wisdom in the data science community suggests that random forests, while not always the best performer, is consistently among the best (e.g., Athey & Imbens, 2019; Choudhury, Allen, & Endres, 2021). Top analysts likely would have used one or more of these ML procedures after seeing our results, and we chose procedures typical of what a new MBA student who took one semester of a good data science course should know. Thus, we understate the value of standard socio-demographic covariates to organizations with top analytical modeling teams.

There also are differences across products in terms of how well our set of covariates predicts purchase behavior relative to the predictive power of a simple linear model. These differences are related to the importance of deviations from linearity, such as including interactions with other variables. The practical importance of ML techniques lies in whether identified deviations from linearity are targetable in some way that advances one or more business objectives. While we cannot answer this question in the general sense, we do provide estimates for our set of products and covariates from IRI, suggesting that, in many instances, the answer is likely to be yes. We further investigate whether the nonlinear relationships are causal, in a narrow context based on a widely used propensity score–weighting approach, and we find that many are. In summary, we have three main objectives:

(1) to reassess the predictive power of socio-demographic characteristics on purchase decision using a suite of MLPs,
(2) to explore the nature of some nonlinear relationships between socio-demographic variables uncovered by MLPs, and.
(3) to examine whether socio-demographic variables have a narrowly defined demand-side causal effect on the purchase decision.

The paper is organized as follows. First, we briefly review relevant literature streams in Section 2. In Section 3, we identify and position our theoretical contribution. Section 4 presents our methodological framework and a detailed description of the data obtained from IRI used in our analysis. The modeling results are presented in Section 5. This is followed by a discussion of the implications of our findings in Section 6 and concluding remarks in Section 7.

2. Literature review

The first stage of customer relationship management is the identification of consumers and their socio-demographic characteristics. The importance of focusing marketing activity on consumers with higher purchase probabilities is self-evident. It is critical for marketing practitioners to accurately identify which consumers to target and the attribute combinations of a product that consumers are most likely to buy. The major challenge retailers face is determining how to better identify their potential customers and whether easily accessible socio-demographic information can be effectively used to predict consumer purchase decisions (Hood, Urquhart, Newing, & Heppenstall, 2020).

The benefits of targeted marketing based on socio-demographic variables are well documented. For example, in an early study, Zwick (1957) found that variations in short-run price elasticities for meat, fish, and poultry purchases were related to both income and age. In a coupon promotion study, Bawa and Shoemaker (1989) found that targeting large and educated households greatly increased sales. More recently, Dong, Manchanda, and Chintagunta (2009) reported that individual-level targeting increased profits by 14% to 23%.

Prior work exploring the relationship between socio-demographics and behavior has mainly focused on two problems: (1) identifying the specific buying behavior of interest, and (2) uncovering socio-demographic variables that help profile consumers and predict future buying behavior. For example, Verbeke (2005) examined the joint impact of socio-demographic, cognitive, and attitudinal factors on consumer acceptance of functional foods, where the presence of ill family members was identified as an important driver; meanwhile, socio-demographics such as age did not independently impact consumer acceptance but interacted with consumer knowledge to generate an effect. Myers, Stanton, and Haug (1971) found that income was more informative than social class in explaining purchases of low-cost packaged goods.

Consumer socio-demographic profiles have been constructed for many products. Among food products, the effect of socio-demographic variables on consumer decisions has been investigated for fruit, vegetable, and meat consumption (Ricciuto et al., 2006; Vaughan, Collins, Ghosh-Dastidar, Beckman, & Dubowitz, 2017), fast-food choices (Akbay, Tiryaki, & Gul, 2007), and purchase of organic foods (Massey, O’Cass, & Otahal, 2018; Thompson & Kidwell, 1998). In these studies, household income, education level, and number of adults and children in a household were consistently reported as strong predictors. Beyond food purchase, the impacts of socio-demographics such as gender, income, and age were used to predict e-commerce behavior (Hood et al., 2020; Soopramanien & Robertson, 2007; Weltevreden, 2007); gender, age, and ethnicity were used to predict hedonic consumptions, such as choice of movies (Palomba, 2020); occupation, income, and education level were used to predict concert attendance (White & Tong, 2019); and age and education level predicted use of renewable-energy technology (Diamantopoulos et al., 2003; Sardianou & Genoudi, 2013).

As we have noted, the usefulness of socio-demographic variables in predicting consumer decisions has been questioned (e.g., Sheth, 1977). Some early work claimed that demographics were poor predictors of purchasing decisions for grocery products (Ferber, 1962; Frank et al., 1967; Koponen, 1960; Twedt, 1964). For instance, Rossi, McCulloch, and Allenby's (1996) pioneering study of the value of past purchase behavior found that only 7% of the variability in price sensitivity was explained by socio-demographics. However, the low predictability of socio-demographic variables on purchase decision does not hold across product categories or contexts (Ricciuto et al., 2006). For example, using choice experiments and real purchase data, Feit, Beltramo, and Feinberg (2010) found that the inclusion of demographic variables greatly improved prediction accuracy for minivan sales. Sun and Morwitz (2010) found socio-demographic variables such as the occupation of the head of household, income, and type of residence useful for reconciling differences between purchase intentions and final purchase decisions for automobiles and personal computers. In addition to predicting purchase behavior, socio-demographics have been shown to predict anti-consumption, such as consumer alienation (Lambert, 1981). Although some researchers have found that socio-demographics underperformed against other factors, such as psychographic and purchase history variables (Rossi et al., 1996; Verbeke, 2005), we argue that these comparisons overlooked the acquisition costs and access feasibility of different data (McDonald & Dunbar, 1998; Sheth, 1977; Wheatley et al., 1980). In general, socio-demographic data are more accessible and can be acquired at lower cost.

Finally, and perhaps more relevant to this article, most previous studies reporting low predictability of socio-demographic variables relied on conventional regression-based econometric and statistical models (Diamantopoulos et al., 2003; Laukkanen, 2016; Ricciuto et al., 2006; Vaughan et al., 2017; Wheatley et al., 1980; Zwick, 1957). For example, when focused on discrete choices such as buy–no buy decisions or product selection among a set of competitors, the prediction models used were often binary and multinomial logit models (Greene, 2018). For example, Gupta and Chintagunta (1994) used a multinomial logit model to profile segments for a set of competing products using demographic variables. A weakness of these models is that they typically capture only linear or linear-in-logit relationships between socio-demographics and consumer decisions. If nonlinear relationships among variables were captured, the predictive accuracy of the model might be enhanced. A nonlinear impact (e.g., inverted U-shape, a pattern that increases at a decreasing rate) of income has been observed for food product purchases (Ricciuto et al., 2006), and household size was found to influence fast-food consumption in an inverted U-shaped pattern (Mihalopoulos & Demoussis, 2001). In addition to nonlinear relationships, different socio-demographics may interact to affect purchase decisions. Examples include interactions between occupation and income, which affect the purchase intentions for household products (Namias, 1960); gender and age, which affect the adoption of mobile banking services (Laukkanen, 2016); age and consumer knowledge, which affect functional food acceptance (Verbeke, 2005); and income and customer satisfaction, which affect loyalty within financial services (Cooil, Keiningham, Aksoy, & Hsu, 2007). To capture these interaction effects, consumer behavior researchers usually rely on theories to make predictions and create interaction terms in regression models for further tests; however, some insights may not be foreseen by existing theories and must first be captured empirically by advanced methods.

3. Artificial intelligence and machine learning in theory building

Artificial intelligence (AI) represents the broad concept of machines carrying out tasks intelligently. Originally dominated by rule-based deductive reasoning, over time AI has moved in the direction of inductive reasoning using ML techniques. This change has been driven by the technological revolution of the internet digitizing social, economic, political, and cultural activities across the world and generating a rich repository of digital data as a by-product. Corresponding developments in ML to help understand and interpret this data have come from computer science, computational neurosciences, econometrics, and statistics, leading to the emerging field of computational social science (Lazer et al., 2020).

The marketing datasets in these applications are typically numeric, but the use of nonnumeric datasets is increasing (Sheth & Kellstadt, 2021; for a review, see Ma & Sun, 2020). Within ML, there are various learning types. Supervised learning is focused on how to predict a particular variable of interest given a set of potential predictor variables (analogous to stepwise regression), and unsupervised learning focused on finding novel patterns (analogous to principal component analysis). For a review of AI applications in marketing, see Verma, Sharma, Deb, and Maitra (2021).

The dominant use of MLPs in marketing, and the one we explore in a specific context in this article, has been to build supervised learning models with flexible, interdependent, and nonlinear relationships not specified a priori (Choudhury et al., 2021; Dzyabura & Yoganarasimhan, 2018). Examples of MLP use in various marketing contexts (e.g., Cui & Curry, 2005; Lemmens & Croux, 2006; Schaeffer & Sanchez, 2020) lead us to expect that the MLPs will improve the predictive power of socio-demographics on purchase decisions. Davenport, Guha, Grewal, and Bressgott (2020), for instance, note an approximately 20% increase in sales by targeting non-purchasers with MLPs.

Among others, Hofman et al. (2021) call for an integration of prediction and explanation into a data-driven computational social science. In information management, Kar & Dwivedi (2020) and Dwivedi et al. (2019) argue that the use of data-driven research should be expanded from pattern recognition and prediction to theory building. Shrestha et al. (2021) advocate that MLPs can be used to develop theory inductively or abductively, as they make fewer a priori assumptions about the functional form of the underlying model that best represents the data. More simply, researchers can use MLPs to explore novel and robust patterns that may lead to theory building. Despite a rich literature on data-driven theory development making the case for using MLPs (Eisenhardt, 1989, 2021), in the social sciences the adoption of data-driven theory development has been comparatively slow.

Lehmann (2020) notes that a core objective of marketing theory is to couple explanation with storytelling. Achieving this objective may precede or follow a data-driven approach to analysis. Previous marketing literature (Bass, 1995; Ehrenberg, 1995) has defined science as a process of interaction between empirical generalization and theory, leading to more advanced theories. However, insistence on theory first effectively rules out the deep insights generated by ML coupled with big data. We hope that the present research can help change that mindset (Breiman, 2001a).

In this study, we adopt an integrative modeling framework that blends data and theory by combining prediction and explanation (see Hofman et al., 2021; Kar, & Dwivedi, 2020; Lehmann, 2020). We implement seven MLP algorithms to predict consumer purchasing decisions. In so doing, we demonstrate the application of MLPs in three modes: pattern discovery, predictive performance, and causal inference. Partial dependence plots (PDPs) are an example of pattern visualization that can reveal (possibly unexpected) nonlinear and interdependent relationships between socio-demographic variables. We use cross-validation to measure the accuracy of each MLP's prediction of the outcome variable. Causal inference from observational data is central to many conceptions of theory (Gregor, 2006). As a guide to causality, PDPs can be used to visualize the marginal effects of one or more variables on the outcome variable (see Zhao, & Hastie, 2021). Here, we focus on estimating causal effects between socio-demographics (cause) and purchasing decisions (effect) using a counterfactual approach. Effectively, our approach is to rule out alternative causal explanations that are due to confounding variables in our observational data. This approach defines a causal effect as the difference in an outcome variable (here, to buy or not), where the same unit of observation experiences different levels of the causal variable (here, socio-demographic variables).

Our findings reveal nonlinear patterns such as inverted U-shapes and interactions of socio-demographics in purchasing decisions that can be further extended for theory building. For example, an inverted U-shape may arise as a result of an aversion to extremes; countervailing (often latent) forces, one positive and one negative (Haans, Pieters, & He, 2016); or a combination of direct and indirect effects driving the underlying process (Islam, Meade, & Sood, 2022).

4. Methodological framework, data, and research objectives

In this section, we describe the linear binary logistic regression model that we use as the baseline statistical and econometric classification model representing standard practice. The baseline model employs either an OLS linear probability model or a probit regression that produces similar results. Then, we describe seven MLPs: five major MLPs used in management research (see Choudhury et al., 2021) and two from recent advances in bio-inspired algorithms (Ab Wahab, Nefti-Meziani, & Atyabi, 2015; Kar, 2016), genetic algorithms (GAs), and particle swarm optimization (PSO). Bio-inspired algorithms are mainly used to solve combinatorial and continuous-parameter optimization problems, with only a few applications for solving classification problems (Sachdeva, Kumar, Gupta, Khandelwal, & Ahuja, 2013). We used multiple MLPs because different ML algorithms are likely to have different strengths and weaknesses that are context-specific in finite samples (Wolpert & Macready, 1997).

We see traditional statistical models and MLPs as complements and use the strengths of each approach to investigate our research questions. Although it is possible to use binary logistic models to test nonlinearities such as quadratic or interaction effects, they need to be hypothesized a priori. Yet, nonlinearities often are unknown beforehand, and the number of possible nonlinearities increases dramatically with the number of variables. The strength of traditional approaches lies in interpretable coefficients that can be used to test hypotheses. MLPs build models with flexible, interdependent, and nonlinear relationships that maximize the models' predictive performance. We exploit this strength to reveal hitherto unspecified complex relationships between X and Y. We test causal relationships between purchase decisions and demographics using inputs from MLPs. Finally, we discuss how the results from the estimated models can be used to address the objectives of this study.

4.1. Statistical and machine learning procedures

Our objective is to model and predict a household's product purchase/nonpurchase decisions based on a series of household socio-demographic variables. Thus, the outcome Y is categorical and binary ($Y = 1$ or $Y = 0$), and the m predictor variables $X_1, X_2, X_3, \ldots, X_m$ are continuous or categorical. We use conventional binary logistic regression as the benchmark model for estimation and compare its results with the performance of five frequently used ML binary classifiers and two bio-inspired algorithms. The standard statistics-oriented references for the MLPs we use are Hastie, Tibshirani and Friedman (2009) and James, Witten, Hastie, and Tibshirani (2013). For each classifier, there are one or more hyper-parameters that are optimized using a training sample.

Binary logistic. We use binary logistic regression to model the relationship between a categorical variable with two possible outcomes and one or more categorical or continuous predictor variables, analogously to linear regression (Greene, 2018; Hosmer, Lemeshow, & Sturdivant, 2013). The binary logistic model can be expressed as:

$$\Pr(Y = 1|X) = \frac{1}{1 + exp\left(- \left(\beta_0 + \sum_{i=1}^{m}\beta_i X_i\right)\right)} \tag{1}$$

If we denote $\Pr(Y = 1|X)$ as π, then the log of the odds ratio for a purchase is:

$\text{logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{m}\beta_i X_i$. (2).

The relationship between Y and X is specified in terms of the conditional distribution of $Y|X$. The effect of each predictor is summarized by a coefficient that captures the marginal effect of a change in that predictor, assuming all other variables are held constant. In addition to its restricted functional form, there are several other limitations of this model. A common issue is that with multiple categorical predictors, cell counts may be too sparse for reliable parameter estimation, making it difficult to include high-order interaction effects.

Random forests. Random forests is an ensemble technique based on the use of a set of small classification and regression trees (CARTs) (Breiman, Friedman, Stone, & Olshen, 1984). Algorithms using ensembles of trees, such as random forests, can approximate functions more smoothly by averaging over the step-functions of single trees, which can help capture nonlinearities and complex interactions (Strobl, Malley, & Tutz, 2009).

Fig. 1 presents a simple single classification tree. The socio-demographic variables of a household are used sequentially to determine whether a consumer is a purchaser. The classification is nonlinear because, at each node, the sample data are divided into subsets that are subsequently treated differently. A random forest effectively averages across many bootstrapped classification trees. To classify a household, the variables are fed into each tree in the forest, resulting in each tree "voting" for a classification (i.e., purchase or nonpurchase). The classification with the most votes is the choice of the random forest.

If there are n households in the sample used for testing and m characteristics describing the households, and a number m_0 (much smaller than m) is specified, then each tree in the forest is generated as follows:

n households are sampled with replacement.

At each node in the tree, m_0 characteristics are chosen at random.

Each tree is grown as far as possible without pruning.

Breiman (2001b) shows that the error rate of the random forest ensemble depends on the correlation between trees and the error rate of individual trees. This correlation decreases as m_0 increases, and the algorithm optimizes the value of m_0.

For further details concerning random forests, see Lin and Jeon (2006), Biau, Devroye, and Lugosi (2008), and Strobl et al. (2009). In marketing, Lemmens and Croux (2006) used random forests to model churn and found greater classification accuracy compared with conventional binary or multinomial models. Schaeffer and Sanchez (2020)

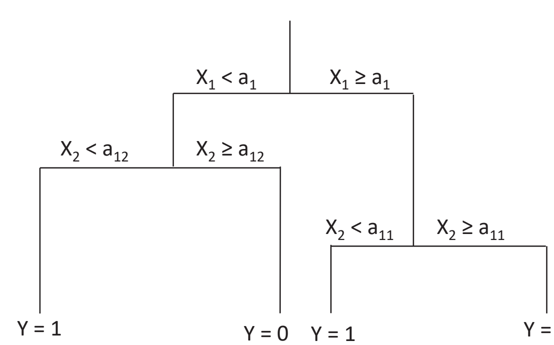

Fig. 1. An example of a simple classification tree.

used random forests, a support vector machine (SVM), and k-nearest neighbors to forecast client retention in prepaid services. A key strength of random forests is their very effective performance "out of the box"—that is, they require relatively little tuning compared with other, more complex, methods (Athey & Imbens, 2019). A weakness is their relatively low interpretability relative to traditional binary logistic models; however, random forests have been successfully extended to investigate causality (Wager & Athey, 2017).

Gradient boosting. Gradient boosting is another widely used ensemble approach that relies on combining many relatively weak simple models to obtain a stronger ensemble prediction (Schapire, 2003). Boosting is a sequential forward-stagewise procedure. Models (e.g., classification trees) are fitted iteratively to training data, gradually increasing emphasis on observations that are poorly modeled by the existing collection of trees. The nonlinearity in this procedure is adaptive reweighting, which gives previously misclassified samples an increased weight in the next iteration but reduces weight to samples correctly classified in the previous iteration. The final classification is based on a weighted majority vote of the sequence of classifiers. A gradient-descent-based formulation of boosting methods has been derived by Friedman, Hastie, and Tibshirani (2000) and Friedman and Meulman (2003). To achieve better predictive accuracy both in- and out-of-sample, we use a nonparametric classification approach (Friedman, 2002), also known as extreme gradient boosting. Gradient boosting's strength lies in its origin in learning theory (Valiant 1984), where weak learners are "boosted" to produce a strong learner. The algorithm generally generates a globally optimal solution but may find only a local optimum. Another weakness is that, unlike random forests, boosted trees have a correlated structure, making the relative importance of variables more difficult to discern (Kuhn & Johnson, 2013).

Support vector machine. An SVM classifies an observation based on where it lies on each side of a hyperplane. The choice of hyperplane depends on the accuracy required and the tolerance for misclassification. SVMs build optimal separating boundaries by solving a constrained quadratic optimization problem, the solution of the following optimization problem (James et al., 2013, p. 346):

Maximize D.

$$\text{Subject to } \sum_{j=1}^{m} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}) \geq M(1 - \epsilon_i)$$

$$\text{where } \epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C. \tag{3}$$

D is the margin, the distance between the observations and the hyperplane; m is the number of predictors X (the vector of socio-demographic variables); n is the training sample size; Y is the outcome variable (purchase, nonpurchase); ϵ are slack variables that allow individual observations to be on the wrong side of the hyperplane; C is the tuning parameter that controls for bias–variance trade-off; and βs are the weights associated with predictors X.

The SVM is effective in high-dimensional spaces and in cases where the number of dimensions, m, is greater than the number of samples, n. The SVM uses memory efficiently because only a subset of training points (the support vectors) is used in the decision function. Functions of the distance between predictors are employed to allow the SVM to use nonlinear decision boundaries. In particular, two kernels are common in empirical work:

$$\text{Polynomial Kernel}: \ \mathrm{K}(X_i, X_i') = \left(1 + \sum_{j=1}^{m} x_{ij}, x_{ij}'\right)^d$$

$$\text{Radial Basis Kernel}: \ \mathrm{K}(X_i, X_i') = \exp\left(-\gamma \sum_{j=1}^{m} \left(x_{ij} - x_{ij}'\right)^2\right)$$

where d is the degree of polynomial and is a positive constant. SVM combines the strength of conventional theory-driven statistical methods and that of data-driven, distribution-free, and robust ML methods (Ravisankar, Ravi, Rao, & Bose, 2011). Cui and Curry (2005) noted SVM's strength in solving nonlinear problems using a linear framework with kernel transformations but highlighted SVM's weakness in not providing probability estimates or predictive/posterior bounds.

K-nearest neighbor. The principle behind any nearest-neighbor method is first to use a predefined distance metric to identify several observations (e.g., households) in the training sample that are closest to the new observation. The values of the outcome variable of the identified observations are then averaged to predict the new observation that falls into the neighborhood. The number of observations required for the prediction can be a user-defined constant (k-nearest neighbor learning) or can vary depending on the local density of points (radius-based neighbor learning). For each new observation, the objective is to find K nearest neighbors, where the distance metric is chosen in the context of the problem; it can be Euclidean for continuous data or a measure of similarity for categorical data. Finally, the new observation is classified using a majority vote among the K neighbors (Hastie et al., 2009, p. 463). Viaene, Derrig, Baesens, and Dedene (2002) provide details for a business application of K-nearest neighbor. The strength of k-nearest neighbors lies in its simple implementation, requiring only two tuning parameters, k and the choice of distance measure. However, unlike random forests, k-NN requires feature scaling and is sensitive to missing observations (Hastie, Tibshirani, & Friedman, 2009).

Neural network. Neural networks encompass a large class of models. A network without a hidden layer is identical to a logistic regression model if the sigmoid activation function is used. Here, as an exemplar, we use a standard single-layer perceptron, which consists of an input layer with a node for each input variable (household characteristic), one hidden layer, and an output layer with two nodes in our case (purchase and nonpurchase). The one hidden layer we used is generally sufficient to classify most datasets (Dreiseitl & Ohno-Machado, 2002). This neural network can approximate a wide range of input–output maps. A single hidden layer's vector of outputs (called hidden units), Z, is created from linear combinations of the m inputs, X:

$$\boldsymbol{Z} = \sigma\left(\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}^T \boldsymbol{X}\right)$$

The vector of outcomes, Y, is a function of linear combinations of the values of Z:

$$\boldsymbol{Y} = g\left(\boldsymbol{\beta}_0 + \boldsymbol{\beta}^T \boldsymbol{Z}\right) \tag{4}$$

The vectors α_0 and β_0 are bias vectors. The function is usually a sigmoid, linear, or tanh function. The output function, $g(\bullet)$, allows a final transformation using the softmax (normalized exponential) function, for an element T_k of vector T,

$$g_k(\boldsymbol{T}) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}$$

In a classification context, this means the decision boundary can be nonlinear, making the model more flexible than logistic regression. For further technical details, see Hastie et al. (2009, p. 392); for examples of business applications of neural networks, see Ravisankar et al. (2011). The primary strength of neural networks is their effectiveness in complex settings with large numbers of features such as image classification, text mining, and other large-scale numeric data sets. Neural networks' weakness lies in the domain knowledge and substantial amount of tuning necessary in a given application compared with the other MLPs used here (Athey & Imbens, 2019).

Genetic algorithm. Inspired by Darwinian evolution, GAs employ the concepts of inheritance, mutation, and natural selection to iteratively improve a solution to an optimization problem. A GA "simulates the process observed in a natural system where the strong tend to adapt and survive while the weak tend to perish" (Ab Wahab et al., 2015, p. 2/36). GAs explore a far greater range of potential solutions than do conventional algorithms (Holland, 1992). For a nonsmooth solution surface, they are designed not to be trapped by local optima, where a gradient-based method could mistakenly identify a local optimum as a global optimum (Meade & Islam, 2006). A GA maintains a population of possible solutions that are processed simultaneously and iteratively modified by crossover (merging different solutions) or by mutation (changing elements of a single solution). The fitness (a model selection criterion) of each population member is evaluated, and a survivor selection process favoring fitter solutions forms a new generation of the population. The process converges when the fittest solution ceases to improve. Grubinger, Zeileis, and Pfeifer (2014) apply a GA to the classification problem. Like random forests, their GA builds on Breiman et al.'s (1984) CART, where rules recursively partition the data into groups using a forward-stepwise procedure. The GA uses a population of sets of values for the parameters of the rules (like a_{ij} in Fig. 1). The population is iterated through many generations until no further improvements in fitness (classification accuracy) can be found. Kar (2016) notes that the strength of a GA is its capability to solve a variety of single or multi-objective combinatorial and nondeterministic problems. Its main weakness is slow convergence toward an optimal solution due to the randomness of crossovers and mutations (Ab Wahab et al. 2015).

Particle swarm. Swarm optimization is inspired by the ways in which insects (for example, ants or bees) communicate the location of a food source. A particle swarm optimization (PSO) algorithm uses a swarm (set) of particles (candidate solutions). The positions of the particles in the search-space are adjusted by following some simple rules. These rules are driven by the current best position of the particle and the overall best position in the swarm. The adaptive Michigan particle swarm optimizer (AMPSO) is specially designed to address the classification problem (for details, see Cervantes, Galván, & Isasi, (2009)). This algorithm combines elements of k-NN and PSO. Each particle acts as a classifier, a vector of attribute values, and a classification—in our case, a vector of socio-demographics and either purchase or nonpurchase. At each iteration, the sample data are assigned to their nearest particle (using Euclidean distance), and the accuracy of the classification of each particle is measured. The position of each particle is adjusted by taking its previous best position and the previous positions of its nearest neighbors. PSO shares several strengths with other MLPs: it can be used on noisy and irregular problems; it has few tuning parameters and is insensitive to the scaling of design variables; and, because it is reliant on social interaction, PSO does not require differentiation, unlike classical optimization algorithms (Ab Wahab et al., 2015). Weaknesses of PSO include a tendency to result in a fast and premature convergence to a suboptimal point, exhibiting a slow convergence in a refined search area (Poli, Kennedy, & Blackwell, 2007).

4.2. Data description

We accessed purchase data for six grocery product categories from IRI, a main source for marketing intelligence data. This dataset was obtained as part of a large grant focused on combining stated and revealed preference data using discrete choice experiments and volumetric choice experiments. For each product category, independent random samples of IRI panelists were selected from both the purchaser and non-purchaser consumer pools. Consumers were classified into these pools depending on whether they had purchased from the particular product category at least once in the previous year. The product categories are baby formula, single-serving coffee, snacks, detergent powder, toothpaste, and canned tuna. These categories differ in idiosyncratic ways and span a range of standard shelf-based products. Major *baby formula* brands are Similac, Enfamil, Gerber, and private labels. Products vary by size, price, packaging, flavor, sugar content, additives, and type of formula. Major *single-serving coffee* brands include Starbucks, Donut House, Folgers, Gevalia, and Green Mountain. Products vary by packaging, flavor, amount of caffeine, and organic information, among other attributes. Major *snack* brands are Atkins, General Mills, Clif, Nature Valley, and Kellogg's. Products vary by size, price, packaging, flavor, sugar content, type of coating, fat content, among other attributes. Major *detergent* brands are Tide, All Mighty, Purex, and Arm & Hammer. Products vary by size, price, wash load, packaging, concentration level, strength, and additives. Major *toothpaste* brands are Colgate, Crest, Aquafresh, Sensodyne, and private labels. Products vary by price, flavor, quantity, form, and package size, among other features. *Canned tuna* brands are Starkist, Bumble Bee, Chicken of the Sea, and private labels. Products vary by form, whether the tuna is packed in water or oil, color, regular type or albacore, and size, among other attributes.

We explore a range of grocery items, but refrigerated products such as beer, milk, and fresh fruits and vegetables may show slightly different purchase patterns. We chose our product categories because the related purchase decisions are believed to be mainly influenced by consumer needs in an ordinary consumable sense. The categories do not represent all consumer products, but we believe they are sufficiently diverse to allow us to draw conclusions about the usefulness of socio-demographics in predicting consumer choice.

The socio-demographic variables are standard ones that IRI routinely collects. Table 1 lists the sample size for each product category and provides a summary of the socio-demographic variables of buyers and nonbuyers.

The proportion of buyers in each product category is greater than 60% in all cases, except for canned tuna. The distribution of socio-demographic variables exhibits variation across product categories. For instance, due to its idiosyncratic nature, baby formula shows a different distribution pattern on some attributes. Take the household-size variable as an example: only 46% of one- and two-member households buy baby formula, while more than 70% of purchases for the other product categories come from this small household-size group. The full dataset is in supplementary appendix A.

4.3. Research objectives

We expand on the three main objectives of our research.

Reassessment of the predictive power of socio-demographic characteristics on the purchase decision using a suite of MLPs. In contrast with most of the marketing literature, which has emphasized significance testing of coefficients more than assessing predictive ability of models, we focus on out-of-sample predictive accuracy in distinguishing between purchasers and nonpurchasers using socio-demographic variables. We used the standard binary logistic model as a benchmark to assess the predictive power of the seven MLPs. The key question is whether the assessment of the usefulness of socio-demographic variables differs substantively depending on the MLP used. We focus on the six product categories, which span a range of grocery items.

Exploration of the nature of some nonlinear relationships between socio-demographic variables that are uncovered by MLPs. In a binary logistic model (and its close analogues), the effects of independent variables on the dependent variable are usually condensed into a single linear index function where all observations are used for the estimation of coefficients, and all observations are treated alike. MLPs, in contrast, depart from linearity in various ways. Tree-based approaches repeatedly split datasets into subsets that are analyzed separately. Gradient boosting identifies poorly fitted observations and applies extra focus to them. Nearest-neighbor procedures isolate a subset of data as the base for predictions, while ignoring the remainder of the dataset. Neural networks produce a potentially nonlinear mapping to capture complex interactions between demographic variables. Thus, the potentially superior predictive accuracy of MLPs over the conventional models is

Table 1
Summary of demographics (%).

Demographics		Baby Formula	Canned Tuna	Deter-gent Powder	Single-Serving Coffee	Snacks	Tooth-paste
Sample Size		1225	837	1154	1328	1258	1218
Purchase	Yes	64.5	50.8	63.5	68.8	67.5	66.9
	No	35.5	49.2	36.5	31.2	32.5	33.1
Household Size	1 to 2 members	46.0	69.5	73.2	75.8	75.0	77.8
	3 members	19.3	13.4	12.5	10.1	12.1	10.8
	4 or more members	34.7	17.1	14.3	14.2	12.9	11.4
Race	White	76.9	84.1	83.4	86.1	86.8	84.2
	Other	23.1	15.9	16.6	13.9	13.2	15.8
Ethnicity	Hispanic	9.6	6.1	5.1	4.4	4.1	4.5
	Non-Hispanic	90.4	93.9	94.9	95.6	95.9	95.5
Affluence	Getting By	53.1	53.4	52.0	55.5	51.3	47.6
	Doing Well	18.9	20.7	21.1	17.2	19.3	24.2
	Comfortable	27.9	25.9	26.9	27.3	29.4	28.2
Household Income (USD)	<25, 000	12.5	14.9	15.2	12.0	15.2	18.1
	25,000 to 49,000	26.7	27.7	28.3	26.9	29.2	31.4
	50,000 to 69,000	18.0	18.0	19.9	19.9	17.6	18.0
	70,000 to 99,000	21.5	22.6	20.7	22.5	19.3	17.6
	100,000 or more	21.3	16.7	15.9	18.8	18.8	15.0
Presence of Child < 18 Y	Yes	45.1	19.4	16.6	13.6	13.4	12.9
	No	54.9	80.6	83.4	86.4	86.6	87.1
Age of Head Household	<35 Years	23.8	8.5	5.7	4.7	4.9	4.5
	35 to 44 Years	20.8	11.7	9.0	9.2	10.2	8.7
	45 to 49 Years	8.8	8.0	8.8	7.1	8.2	7.0
	50 to 54 Years	9.2	13.6	12.3	9.7	11.6	10.3
	55 to 64 Years	23.6	29.3	35.1	36.8	33.1	33.5
	65 Years or more	13.8	28.9	29.0	32.5	32.0	36.0
Education of Female Household	Some or Graduated High School	24.0	27.2	28.5	26.1	26.5	30.9
	Some College	28.7	29.5	33.0	31.4	32.4	29.2
	Graduated College	33.0	29.9	26.8	28.9	26.1	26.5
	Post College	14.3	13.4	11.7	13.6	15.1	13.4
Census (Sampling Regions)	Central	27.5	25.4	25.9	26.6	27.0	26.4
	East	16.8	19.1	16.2	19.6	17.2	19.4
	South	39.4	36.7	40.9	36.9	33.6	35.9
	West	16.2	18.8	17.0	16.9	22.1	18.4

due in part to their ability to identify one or more nonlinearities among variables to improve predictions for at least some subsets of the observations.

Examination of whether socio-demographic variables have a demand-side causal effect on the purchase decision. In observational studies, an association between a set of contemporaneously recorded covariates and the outcome variable of interest should not be taken as proof of a causal relationship. For instance, the value differences in other covariates between the treatment and control groups may lead to biased estimates of treatment effects (D'Agostino, 1998). To address this issue, we use propensity scoring to reweight observations to ensure that the distribution of measured baseline covariates is similar between the treatment and control groups. Our focus on socio-demographic variables, clearly exogenous from the perspective of the decision of interest, also reduces the risk of uncovering noncausal relationships.

5. Results of a comparison between conventional and machine learning methods

We present our analysis in three subsections addressing predictive ability, nonlinear patterns found by MLPs, and determination of causal inference. To measure and compare predictive ability, we divided the dataset into a 70% training sample and a 30% test sample, fit models to the training sample, and made predictions in- and out-of-sample. Our criterion for prediction accuracy was sensitivity, defined as the proportion of actual purchasers correctly identified. We demonstrated how MLPs capture nonlinearities by comparing the models estimated by random forests with conventional binary logistic regression.

We focus much of our discussion on the results for random forests because this technique has become the standard workhorse in the ML world for our sort of problem, akin to OLS (e.g., Athey & Imbens, 2019; Choudhury et al., 2021). It is less of a "black box" than many ML techniques, and random forests uses the concept of a CART (Breiman et al., 1984), which consists of simple binary splits. Its ensemble notion of different trees voting to determine the best predictor makes the technique (relatively) easy to explain to managers. We used a propensity score adjustment to demonstrate how causal relationships can be robustly assessed from observational analyses using MLPs.

5.1. Model estimation and predictive performance

For each product category, we split the dataset randomly into a 70% training sample and a 30% testing sample. Each model (logistic regression, random forests, gradient boosting, SVM, k-nearest neighbor, neural network, genetic algorithm, and PSO) was first fit to the training sample and then used to make purchase predictions both in- and out-of-sample. We estimated the binary logistic and first five MLPs using the Scikit-learn library in Python in two stages. First, we tuned hyperparameters using cross-validation with GridsearchCV in Scikit-learn. Second, we optimized model parameters. The GA solution used the evtree package in R (Grubinger et al., 2014). The particle swarm classifier was programmed in Python for this analysis.

We display predictive performance using a confusion matrix (see

Table 2
Confusion matrix for evaluation of predictive performance.

		Predicted Class	
		Purchase	No Purchase
Actual Class	Purchase	True Positive (TP)	False Negative (FN)
	No Purchase	False Positive (FP)	True Negative (TN)

Accuracy = (TP + TN)/(TP + FP + FN + TN); Sensitivity = TP/(TP + FN).

Table 2). True positives and true negatives are observations that were correctly predicted. Classification procedures are designed to minimize false positives and negatives. With balanced data, where purchasers and nonpurchasers share equal proportions, predictive performance is usually measured by *accuracy*, the ratio of correctly predicted observations to total observations. In our data, where there is an imbalance between the classes, accuracy is misleading (Choudhury et al., 2021; Saito & Rehmsmeier, 2015). Gu, Zhu, and Cai (2009) showed superiority of sensitivity over accuracy for imbalanced data like ours. Thus, we use *sensitivity* to evaluate the predictive performance of different models. Here, "sensitivity" refers to the correctly identified fraction of all the panelists who purchased the product.

Table 3 shows in-sample and out-of-sample sensitivities of the predictions by the eight classification procedures for each product category. The MLPs clearly outperformed parametric logistic regression. Across products, the median improvement in sensitivity by each MLP varies from 16% to 42% in-sample and from 20% to 33% out-of-sample. Random forests performed well both in- and out-of-sample, achieving the highest sensitivity for four product categories and second highest for the other two categories. The predictive performance of the SVM is comparable to the random forests procedure. Gradient boosting is the third most effective model of these procedures. The bio-inspired algorithms both performed similarly to gradient boosting. K-nearest neighbor is the least effective of the MLPs, but it still convincingly outperforms logistic regression. The neural network is the only model with a noticeable deterioration between in-sample and out-of-sample performance.

Results from the MLPs suggest that socio-demographic variables exert a nonlinear influence on purchase decisions that is not fully captured by the traditional logistic regression model. Furthermore, MLPs such as random forests, SVM, and gradient boosting identify interactions between socio-demographics by selecting appropriate subsets of the data for estimation. K-nearest neighbor has a similar objective, but it is less flexible than the aforementioned MLPs. The algorithms with higher model capacity and flexibility, such as neural networks, require more observations and expertise to prevent overfitting. The neural network was fitted with relatively few observations in our sample and tended to overfit the data, leading to a decrease in out-of-sample predictive accuracy.

5.2. Nonlinear relationships in the data: A comparison of variable effect estimation between the random forest model and binary logistic regression

We use a simple example to illustrate the differences between a conventional linear regression model and an MLP. While the binary logistic regression predicts product purchases by using *all* the values of *all* the socio-demographic variables in an *additive linear* logit function, the random forests procedure makes the purchase prediction of baby formula with a subset of four variables and includes them stepwise in the prediction model (see Fig. 2).

We demonstrate how the random forests model works with an example tree for baby formula. The tree first separates the data according to the variable of *presence of a child* (two categories: yes vs. no). After the first data split, both subsets are further separated using the *age of the head of household*, but with different levels determining the split, and neither age split being monotone. After the second split, households with heads in specific age groups go directly to the output node, and others are further split based on *income* and then *census area*. The final row of output nodes indicates the probability of purchasing baby formula by a household classified into that node. For example, the output node furthest to the right shows that households with a child and a head not aged 45–49 years have a purchase probability of 0.84. In summary, the data were divided into subsets with subsequent treatment conditional on the values of other socio-demographic variables. Interactions between two, three, and four variables are considered in the random forests procedure. When classifying an out-of-sample household as a potential purchaser or nonpurchaser, we assign each tree one vote, and the final classification is decided by a majority vote.

To compare the effectiveness of the conventional binary logistic regression and the random forests model in predicting household purchase of baby formula, we focused on two variables, *household income* and *education level of the female head of household*, to illustrate use of PDPs. Compared with traditional statistical models, MLPs have low interpretability; however, the use of partial effects is common in the social sciences when model parameters are not immediately interpretable (King, Tomz, & Wittenberg, 2000). We use a PDP proposed by Friedman (2002) as a visualization tool to contrast nonlinearity and/or interactions with the binary logistic. The PDP reveals the marginal effect of one or two characteristics on the predicted outcome of an MLP. The plot shows whether the relationship between the target and a characteristic is linear, monotonic, or more complex. When applied to a linear model (such as the binary logistic), PDPs indicate a linear relationship. In Fig. 3, we show the PDP of two demographics and their interactions for baby formula.

Comparing the two left-hand plots for the conventional approach and the MLP, we see that the linear dependence of the logistic regression approximates the nonlinear effects of household income and education level of a female head of household from the random forests. The nonlinear effects of the two characteristics, income and education, contrast strongly with linear contours from logistic regression. The latter imply that, for a given education level, the probability of purchase decreases as income increases, whereas the random forests plot shows an increase in probability of purchase with increasing income until a sharp drop at the highest income level.

5.3. Causal inference using a propensity score approach

In observational studies, observed associations between treatments (explanatory variables) and outcome variables do not adequately prove a causal relationship. Thus, in our case a correlation cannot be used to infer causal links between socio-demographic characteristics and the purchase decision. There are several threats to any set of linkages translating correlations into causation, with a major one being the confounding that comes from the differences in observed variables between treatment and control groups. These differences can lead to biased estimates of treatment effects (D'Agostino, 1998).[11] Consider the following example. Does *the presence of a child* (i.e., treatment group) causally predict the purchase of a product? For an experimental study, the treatment condition is the *presence of a child* and the control condition is *no presence of a child*. However, the observed differences in purchase probability could also be attributed to differences in the other socio-demographics, such as income and household size, instead of the presence of the child itself (see Fig. 4).

To illustrate how a causal inference can be better established with observational studies, we chose to study the effect of *presence of a child* (yes or no) on purchase/nonpurchase decisions. In an ideally controlled experiment, households with and without a child would be matched to be statistically equivalent on all other potentially relevant socio-demographics. Otherwise, any observed effects of the focal variable, *presence of a child*, could be confounded with other socio-demographic differences between the two groups (D'Agostino, 1998; Greenland & Morgenstern, 2001). However, this confounding issue has largely been ignored in prior investigations. For example, in our datasets for baby formula, single-serving coffee, and canned tuna, the imbalance is evident for two socio-demographic variables, *household size* and *age of the head of household* (see Table 4). This imbalance is mainly due to the correlations between some socio-demographic characteristics. For

[1] There are, of course, many other threats to identification of causal relationships, particularly in revealed preference data, that require a panel structure to sort out (Greene, 2018).

Table 3
Comparison of sensitivity for in- and out-of-sample predictive performance.

	Classifiers	Baby Formula	Canned Tuna	Deter-gent Powder	Single- Serving Coffee	Snacks	Tooth-paste	Median % Improve-ment over Binary Logistic	Average Rank
In-sample sensitivity	Binary Logistic	0.785	0.577	0.814	0.797	0.537	0.579		7.8
	Random Forest	0.959	0.790	0.971	1.000	0.990	0.948	31.2	1.7
	Gradient Boosting	0.938	0.733	0.944	0.976	0.957	0.919	24.7	3.8
	SVM	0.926	0.790	0.952	1.000	0.969	0.888	31.2	3.2
	k-NN	0.887	0.577	0.927	0.950	0.957	0.912	16.5	5.5
	Neural Network	0.947	0.937	0.973	0.970	0.964	0.966	42.0	2.0
	Genetic Algorithm	0.892	0.946	0.928	0.904	0.941	0.711	21.0	5.7
	Particle Swarm	0.858	0.935	0.921	0.955	0.916	0.790	27.1	5.3
Out-of-sample sensitivity	Binary Logistic	0.740	0.560	0.816	0.705	0.533	0.567		7.7
	Random Forest	0.892	0.704	0.958	0.992	0.950	0.957	33.2	1.5
	Gradient Boosting	0.865	0.640	0.915	0.953	0.913	0.901	26.0	3.8
	SVM	0.879	0.680	0.948	0.995	0.888	0.906	31.3	3.0
	k-NN	0.825	0.520	0.840	0.907	0.883	0.927	20.1	5.7
	Neural Network	0.749	0.728	0.811	0.833	0.796	0.871	24.1	6.3
	Genetic Algorithm	0.865	0.945	0.849	0.898	0.925	0.693	28.9	4.0
	Particle Swarm	0.843	0.910	0.891	0.953	0.879	0.784	34.5	3.8

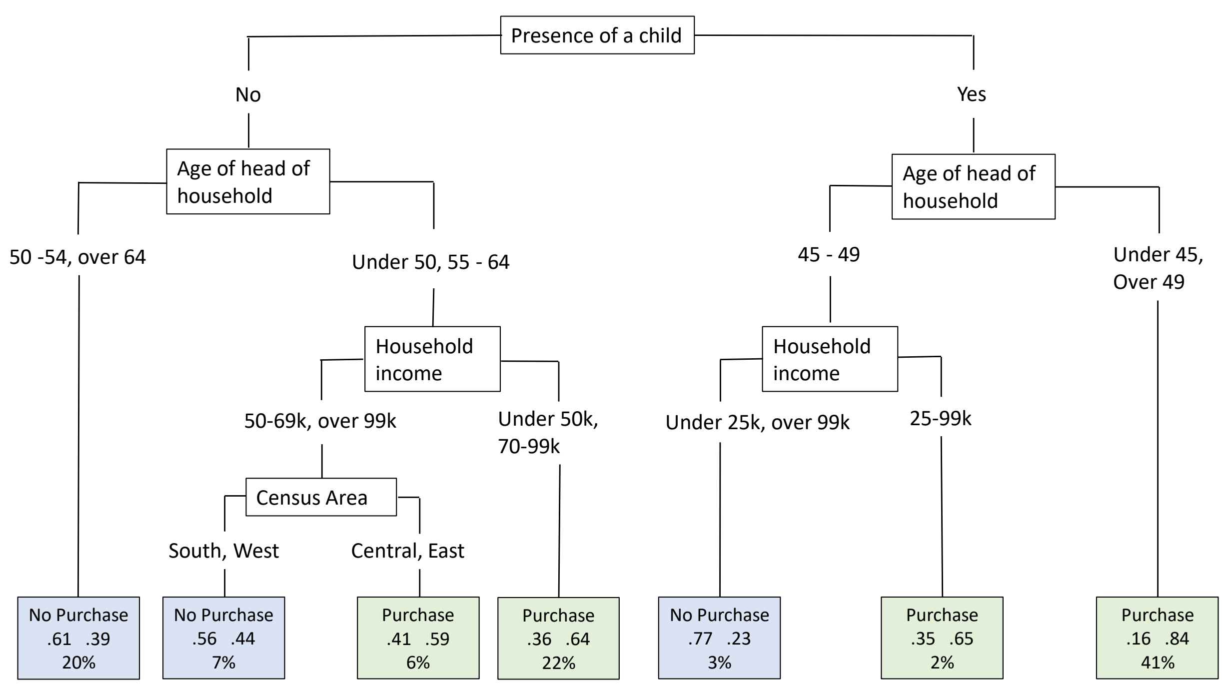

Fig. 2. Example tree from random forests procedure predicting purchase of baby formula Note: For output nodes, the most likely predicted outcome is given (no purchase/purchase), the split [Pr(no purchase), Pr(purchase)] of each type within the node, and percentage of observations in the sample classified as being in this output node.

instance, the presence of each child increases the size of the household, and parents with only one child tend to be younger than 35 years old. Although the rationale for the imbalance is explicable, obtaining unbiased estimates of the effects of socio-demographic characteristics may require appropriate conditioning on this imbalance in the statistical estimation procedure.

Propensity scores enable us to address the confounding concerns explained previously. To help overcome the confounding problem and to facilitate the inference of causal relationships in observational studies, Rosenbaum and Rubin (1983) defined propensity score as $\Pr(Z = 1 \mid X)$, the probability of a treatment assignment (treatment, $Z = 1$; control, $Z = 0$) given an observed covariate, X. Austin (2011) noted that propensity is a balancing score and, once made conditional on this score, the distribution of observed baseline variables will be similar between treatment and control groups (e.g., households with or without a child), thus allowing for the inference of causal effect. Under the assumption of no unmeasured confounding, Rosenbaum and Rubin showed that the propensity score approach yields unbiased estimates in the presence of confounding. Hence, for our dataset, comparing households with identical propensity scores but different realized purchasing decisions is

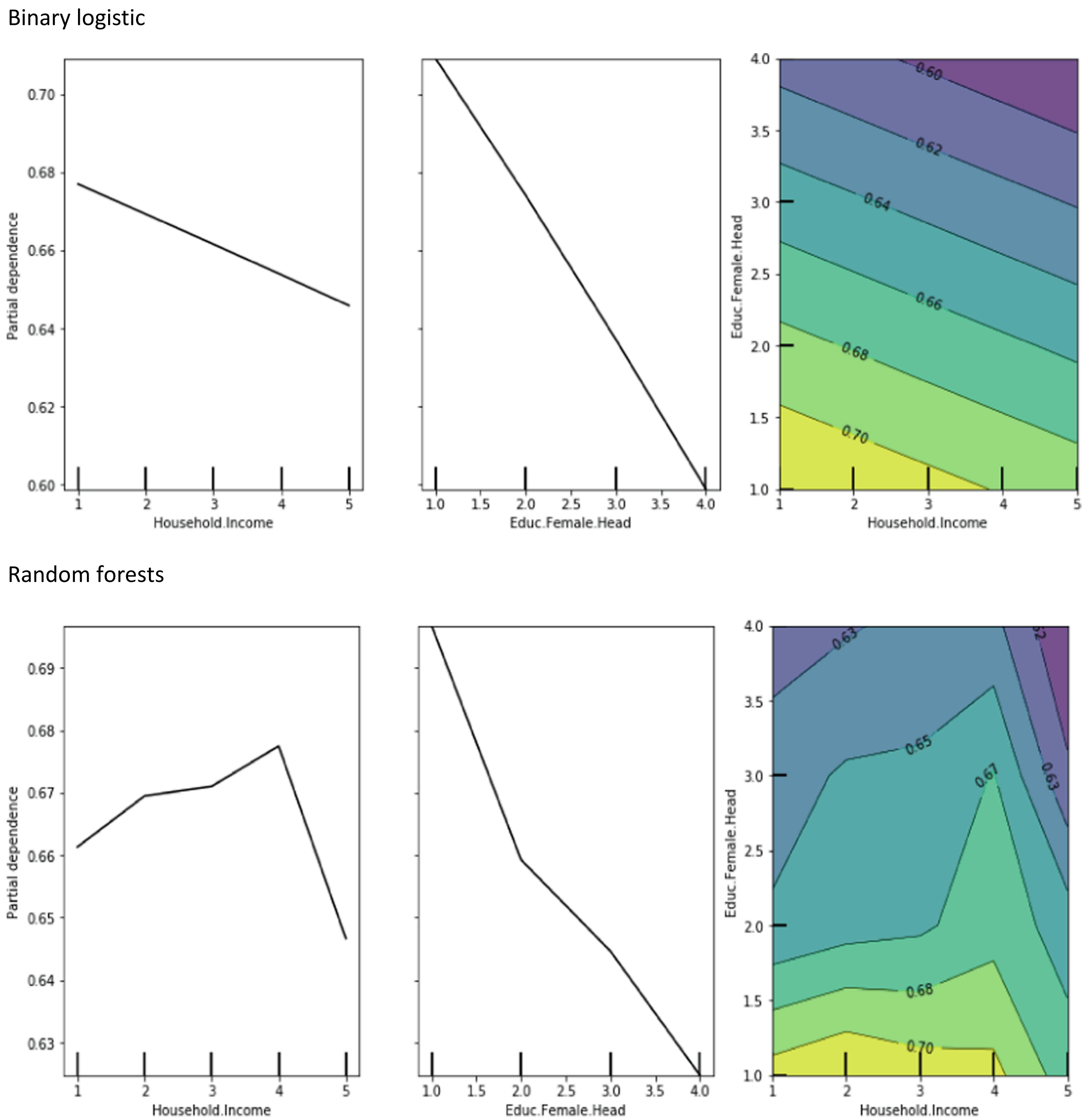

Fig. 3. Partial dependence plots of *household income* and *education level of female head* for logistic regression and random forests for baby formula. The vertical axis of the two left-hand plots is the predicted probability of purchase given the value of the characteristic. The contour lines in the right-hand plot represent the predicted probability of purchase given the values of both characteristics.

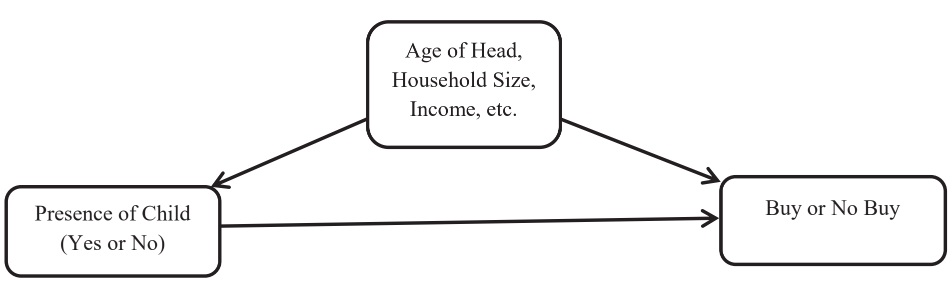

Fig. 4. Potential confounders to the relationship between *presence of child* (yes/no) and purchase decision.

analogous to conducting a randomized experiment, and the results obtained provide a valid basis for the inference of causal relationship (Bingenheimer, Brennan, & Earls, 2005).

Propensity score adjustment is a two-stage method. In the first stage, confounding variables are balanced across groups. The propensity score–adjusted effects are then estimated in the second stage. Logistic

Table 4
For households with a child (yes/no), the percentages for other characteristic levels reveal the extent of the data imbalance.

Demographics	Households with Child	Baby Formula yes	Baby Formula no	Single-Serving Coffee yes	Single-Serving Coffee no	Canned Tuna yes	Canned Tuna no
	Category Levels						
Household Size	1 to 2 members	2.1	79.9	5.0	88.4	7.4	86.0
	3 members	33.0	13.1	24.1	7.6	32.0	9.2
	4 or more members	64.9	7.0	70.9	4.1	60.6	4.8
Household Income (USD)	<25,000	13.4	13.5	9.1	12.3	11.3	14.7
	25,000 to 49,000	28.4	28.8	15.0	24.0	25.1	30.1
	50,000 to 69,000	20.0	17.3	10.8	17.1	16.9	17.3
	70,000 to 99,000	20.3	21.9	22.4	17.3	28.1	21.8
	100,000 or more	17.9	18.5	21.6	16.0	18.6	16.2
Age of Head Household	<35 Years	53.3	14.5	12.7	2.5	19.9	3.9
	35 to 44 Years	30.0	10.0	38.0	4.4	41.1	3.6
	45 to 49 Years	5.5	8.0	20.8	5.4	14.3	4.9
	50 to 54 Years	4.3	12.6	11.9	10.5	16.0	12.3
	55 to 64 Years	5.5	35.6	14.4	39.4	6.1	36.5
	65 Years or more	1.3	19.3	2.2	37.8	2.6	38.8
Education of Female Head of Household	Some or Graduated High School	17.8	28.3	18.3	26.0	17.7	26.9
	Some College	30.8	26.2	30.5	29.4	28.1	25.3
	Graduated College	35.1	28.4	34.9	23.2	39.4	25.5
	Post College	14.1	11.7	13.0	11.4	12.6	11.4

regression is commonly used to estimate the propensity score in the first stage. However, if logistic regression's underlying assumptions of linear functional form and specification of interactions are violated, then covariate balance will not be achieved, and effect estimates will be biased (Drake, 1993; Williamson, Morley, Lucas, & Carpenter, 2012). In contrast, regardless of sample size and the extent of nonadditivity or nonlinearity, the propensity scores of MLPs such as random forests provide excellent covariate balance and effect estimation (Lee, Lessler, & Stuart, 2010; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008).

To provide a benchmark, we use logistic regression to estimate characteristic effects using the unweighted data. Then, to turn our observational study into a pseudo-randomized study, in the first stage we derive MLP propensity scores to achieve covariate balance. In the second stage we use logistic regression to test causal effects by reweighting samples by propensity scores. We follow the research agenda suggested by Wedel and Kannan (2016) for estimation of causal effects, using the strengths of both traditional procedures and MLPs. Specifically, we use the R package "twang" (McCaffrey, Ridgeway, & Morral, 2004) to get propensity scores using boosted trees for binary treatments (e.g., presence of a child: yes vs. no) and multinomial treatments (e.g., family income groups, education levels). Boosted trees

Table 5
Estimates of a sample of socio-demographic variables using unweighted data and data reweighted by propensity scores. Agreement on variable significance is shown in the right-hand column (Y indicates significance at 5%, N otherwise).

			Binary Logistic – Unweighted Data (BL)		Propensity score adjusted (PA)		Agreement /Disagreement
			Estimate(β)	**p-val**	**Estimate(β)**	**p-val**	**(BL/PA)**
Presence of Child(age < 18 yrs.)	Baby Formula	Yes	0.564	0.021	0.975	0.000	YY
	Canned Tuna	Yes	0.066	0.815	−0.652	0.007	**NY**
	Detergent Powder	Yes	−0.092	0.716	−0.382	0.328	NN
	Coffee	Yes	0.481	0.069	−0.199	0.583	NN
	Snacks	Yes	−0.200	0.447	−0.204	0.640	NN
	Toothpaste	Yes	0.564	0.021	0.975	0.000	YY
Household Income Groups	Baby Formula	25–49 K	0.062	0.821	0.212	0.336	NN
		50–69 K	0.071	0.839	0.180	0.443	NN
		70–99 K	0.265	0.517	0.348	0.130	NN
		100 K or more	−0.031	0.940	−0.014	0.949	NN
	Canned Tuna	25–49 K	0.485	0.074	0.284	0.207	NN
		50–69 K	0.275	0.447	0.038	0.877	NN
		70–99 K	0.154	0.707	−0.082	0.726	NN
		100 K or more	−0.145	0.731	−0.375	0.134	NN
	Detergent Powder	25–49 K	0.372	0.167	−0.037	0.866	NN
		50–69 K	0.168	0.616	−0.284	0.215	NN
		70–99 K	−0.057	0.880	−0.505	0.026	**NY**
		100 K or more	−0.012	0.976	−0.435	0.068	NN
	Coffee	25–49 K	0.186	0.437	0.406	0.045	**NY**
		50–69 K	0.106	0.731	0.533	0.014	**NY**
		70–99 K	−0.359	0.322	0.351	0.092	NN
		100 K or more	−0.247	0.506	0.394	0.069	NN
	Snacks	25–49 K	−0.017	0.942	0.084	0.662	NN
		50–69 K	−0.053	0.864	0.306	0.154	NN
		70–99 K	−0.352	0.333	0.263	0.208	NN
		100 K or more	0.033	0.928	0.571	0.009	**NY**
	Toothpaste	25–49 K	−0.339	0.372	−0.015	0.946	NN
		50–69 K	0.314	0.298	0.465	0.015	**NY**
		70–99 K	0.088	0.733	0.214	0.306	NN
		100 K or more	0.489	0.040	0.351	0.100	**YN**

are a hybrid of random forest and gradient-boosting procedures. We compare the estimates of the characteristics using the benchmark (with unweighted data) with estimates using data reweighted by propensity scores (see Table 5).

In Table 5, for parsimony we present sample output for two demographics: household income (a multinomial treatment) and the presence of a child (a binary treatment). For baby formula, there is no disagreement about the significance of a linear effect of the two characteristics; for the other five products there is a disagreement, and in most cases a characteristic is found significant using the propensity score–reweighted data. In Table 5, there is disagreement for 6 out of 30 effects (20% disagreement). For all the estimated effects for all the demographics, we find 25% disagreement. These results demonstrate that socio-demographics may have a causal influence on purchase decisions that is overlooked in logistic regression models due to inadequate methodology.

6. Discussion

In the prediction of a consumer purchase decision, a major attraction of the use of socio-demographic variables, relative to psychographic variables, is their low acquisition cost. However, the usefulness of such variables in identifying purchasers has long been debated. Early work (e.g., Koponen, 1960; Rossi, McCulloch, & Allenby, 1996; Twedt, 1964) found that socio-demographic variables had little value in predicting consumer decisions, yet later work (e.g., Ricciuto et al., 2006; Feit, Beltramo, & Feinberg, 2010) argued that such variables did have predictive value. Though related, the question we have sought to address does not pertain to whether one set of variables is better or whether socio-demographics add explanatory power when attitudinal variables and past purchase behavior are available. Rather, our question is: Can modern ML techniques extract more actionable information from low-cost socio-economic variables than traditional tools?

We investigated the predictive accuracy of seven MLPs for six distinct product categories using large industry-standard datasets from IRI and obtained a roughly 20%–30% improvement in out-of-sample predictive performance with the different MLPs across the six product categories. There are two reasons for the enhanced predictive value of socio-demographics. First, the MLPs consistently construct (in different ways) simple nonlinear response functions from the socio-demographic variables, in addition to the linear effects detected by more conventional methods. Second, the MLPs detect important interactions between the covariates while using different methods to avoid the overfitting that would occur from the inclusion of all such interactions. Visualizing the patterns of nonlinear relationships and interactions or converting them to marginal quantities such as elasticities can enhance understanding. In addition, one of the key threats to a causal interpretation of the use of socio-demographic variables is selection bias. We show how propensity score matching can help address this issue. The low cost of acquiring the socio-demographic variables and the convenience of implementing MLPs would allow for widespread uptake of this practice among firms and could aid to further theory building.

6.1. Theoretical contributions

In our analysis, we make theoretical contributions under two of Whetten's (1989) framework items: *how* variables in the model are related conceptually and *what* key drivers should be considered for a conceptual model. Regarding *how*, we establish the superiority of MLPs relative to the commonly used generalized linear models. MLPs capture and exploit important nonlinear effects in socio-demographic variables. For example, we identify the nonlinear effects of income and education and show the strong contrast with the linear effect found by logistic regression. Without prior guidance, MLPs identify interactions between these variables, resulting in better predictive performance. For example, we show that the random forests procedure considered interactions between two, three, and four variables when modeling the purchase decision.

Regarding *what*, we demonstrated that socio-demographic variables can be causally linked to purchase decisions, extending the prior correlational relationships identified with observational data. This approach should be particularly useful for the exploration of how basic demographics such as gender, education level, and stage in a family's life cycle influence an individual's value system and/or psychological needs and, thus, their purchase and consumption of goods and services (Tharp, 2001; White & Tong, 2019).

6.2. Implications for practice

The widespread availability of MLPs can fundamentally change how researchers approach the search for nonlinearities when developing predictive models. By construction, a simple linear (logistic) model does not extract all the information from a set of predictors unless that model's specification matches the true data-generating process. This rarely is the case with standard socio-demographic variables because, often in complex ways, they underlie attitudes and past purchase behaviors that are the more proximate causal variables. MLPs are designed to help capture this underlying structure in a way that avoids the confirmatory bias that can arise from fitting a large number of opportunistic models. A perceived disadvantage of MLPs is their "black box" nature; these procedures do not deliver the estimated coefficients, standard errors, and p-values of conventional statistical models. However, although not demonstrated here, MLPs can produce marginal odds ratios or marginals useful for identifying key socio-demographic characteristics that influence product purchase probabilities, as well as standard measures such as price elasticities or gains in lift in response to promotions.

Socio-demographic variables are widely available and relatively inexpensive to acquire. These variables, coupled with large datasets, form the ideal environment for marketing-related analysis. We have demonstrated MLPs' ability to deliver improved predictions of the purchase decisions compared with conventional statistical methods. Further, we have provided insights about the relative strengths of the seven MLPs considered. Of the more established MLPs, we found random forests and gradient boosting to offer greater improvements in predictive accuracy over binary logistic regression. We also demonstrated that neural networks tended to promise more in-sample than was delivered out-of-sample. Of the newer procedures we implemented, swarm optimization shows great promise. In summary, the use of easily available socio-demographic data for segmentation and targeting is attractive to marketers, and we show MLPs to be an effective analytical tool for targeting potential purchasers and increasing sales.

6.3. Limitations and future research directions

We note three key limitations of our work. First, although the six product categories we examined represent a reasonably diverse selection of grocery products, the transferability of our results to contexts other than packaged goods is an open question. In particular, we would expect MLPs to extract less useful information from socio-demographic variables in situations where, for instance, product purchase is mainly driven by other marketing communication factors, such as the use of celebrity endorsement, instead of basic consumption needs. In addition to generalizability, we acknowledge the challenge of comparing algorithms by their performance on relatively small samples of data, as noted by Wolpert and Macready (1997).

Second, the socio-demographic variables that we used are standard and have been widely used in marketing. The range of socio-demographic variables now available from database vendors is dramatically larger at both the individual and small-geographic-unit levels.

Third, establishing causality is a difficult and multifaceted enterprise. We took one step in that direction by using the propensity score approach to transform an observational study into a pseudo-randomized trial study (Rosenbaum & Rubin, 1983), which yields unbiased estimates in the presence of measured confounding variables (Robins & Greenland, 1992). This is an important point to note, as it is not always possible to collect all socio-demographic variables in practice. We recommend sensitivity analyses to assess the extent to which unmeasured confounding variables might change the estimated effects, as suggested by Ding and VanderWeele (2016).

Future research directions should include further probing into the nonlinear nature of the relationship between the consumer purchase decision and socio-demographic data. The increasing size of databases in terms of volume and availability of more variables is likely to make further exploration rewarding. Exploration using discrete choice experiments (Louviere, Hensher, & Swait, 2000) with formal randomization is likely to further illuminate issues of the causality underlying the consumer purchase decision.

7. Concluding remarks

After investigating the predictive accuracy of seven MLPs for six distinct product categories, we found that all the MLPs exhibited greater predictive accuracy both in- and out-of-sample than logistic regression. We demonstrated that this superior accuracy was due to the MLPs' ability to better capture the nonlinearities present in the underlying data-generating process of purchaser behavior compared with the linear framework of logistic regression. It would be interesting to revisit the issue of past purchase behavior in terms of how much more these variable(s) add after an appropriate ML approach extracts the signal from the socio-demographic variables.

Conventional statistical procedures such as logistic regression allow hypothesis testing to be used for causal inference. We used propensity scoring in a two-stage procedure in this context and demonstrated that applying an MLP to pre-weight the data improved inference by reversing approximately 25% of the conclusions from the unweighted analysis. Given the low acquisition cost of socio-demographic variables and the ease of using MLPs, our findings suggest that major benefits can be derived from their use in predicting purchase/nonpurchase decisions.

CRediT authorship contribution statement

Towhidul Islam: Writing – original draft, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Nigel Meade:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Conceptualization. **Richard T. Carson:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Jordan J. Louviere:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Juan Wang:** Writing – review & editing, Visualization, Resources, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada, Grant No. 430199.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbusres.2022.07.004.

References

Ab Wahab, M. N., Nefti-Meziani, S., & Atyabi, A. (2015). A comprehensive review of swarm optimization algorithms. *PloS One, 10*(5), Article e0122827.
Akbay, C., Tiryaki, G. Y., & Gul, A. (2007). Consumer characteristics influencing fast food consumption in Turkey. *Food Control, 18*(8), 904–913.
Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685–725.
Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399–424.
Bass, F. M. (1995). Empirical generalizations and marketing science: A personal view. *Marketing Science, 14*(3_supplement), G6–G19.
Bawa, K., & Shoemaker, R. W. (1989). Analyzing incremental sales from a direct mail coupon promotion. *Journal of Marketing, 53*(3), 66–78.
Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research, 9*(June), 2015–2033.
Bingenheimer, J. B., Brennan, R. T., & Earls, F. J. (2005). Firearm violence exposure and serious violent behavior. *Science, 308*(5726), 1323–1326.
Breiman, L. (2001a). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.
Breiman, L. (2001b). Random forests. *Machine Learning, 45*(1), 5–32.
Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman and Hall.
Cervantes, A., Galván, I. M., & Isasi, P. (2009). AMPSO: A new particle swarm method for nearest neighborhood classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39*(5), 1082–1091.
Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal, 42*(1), 30–57.
Cooil, B., Keiningham, T. L., Aksoy, L., & Hsu, M. (2007). A longitudinal analysis of customer satisfaction and share of wallet: Investigating the moderating effect of customer characteristics. *Journal of Marketing, 71*(1), 67–83.
Cui, D., & Curry, D. (2005). Prediction in marketing using the support vector machine. *Marketing Science, 24*(4), 595–615.
D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*(19), 2265–2281.
Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*(1), 24–42.
Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95–106.
Diamantopoulos, A., Schlegelmilch, B. B., Sinkovics, R. R., & Bohlen, G. M. (2003). Can socio-demographics still play a role in profiling green consumers? A review of the evidence and an empirical investigation. *Journal of Business Research, 56*(6), 465–480.
Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology, 27*(3), 368–377.
Dong, X., Manchanda, P., & Chintagunta, P. K. (2009). Quantifying the benefits of individual-level targeting in the presence of firm strategic behavior. *Journal of Marketing Research, 46*(2), 207–221.
Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49*, 1231–1236.
Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*(5–6), 352–359.
Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2019). Re-examining the unified theory of acceptance and use of technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers, 21*(3), 719–734.
Dzyabura, D., & Yoganarasimhan, H. (2018). Machine learning and marketing. In *Handbook of marketing analytics* (pp. 255–279). Northampton: Edward Elgar Publishing.
Ehrenberg, A. S. (1995). Empirical generalisations, theory, and method. *Marketing Science, 14*(3_supplement), G20–G28.
Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532–550.
Eisenhardt, K. M. (2021). What is the Eisenhardt Method, really? *Strategic Organization, 19*(1), 147–160.
Feit, E. M., Beltramo, M. A., & Feinberg, F. M. (2010). Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science, 56*(5), 785–800.
Ferber, R. (1962). Research on household behavior. *American Economic Review, 52*(1), 19–63.
Frank, R. E., Massy, W. F., & Boyd, H. W. (1967). Correlates of grocery product consumption rates. *Journal of Marketing Research, 4*(2), 184–190.
Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378.
Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics, 28*(2), 337–374.
Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine, 22*(9), 1365–1381.
Greene, W. H. (2018). *Econometric analysis, 8th ed.* Upper Saddle River, NJ: Pearson.
Greenland, S., & Morgenstern, H. (2001). Confounding in health research. *Annual Review of Public Health, 22*(1), 189–212.
Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611–642.

Grubinger, T., Zeileis, A., & Pfeifer, K.-P. (2014). evtree: Evolutionary learning of globally optimal classification and regression trees in R. *Journal of Statistical Software, 61*(1), 1–29.
Gu, Q., Zhu, L., & Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In *International symposium on intelligence computation and applications* (pp. 461–471). Berlin: Springer.
Gupta, S., & Chintagunta, P. K. (1994). On using demographic variables to determine segment membership in logit mixture models. *Journal of Marketing Research, 31*(1), 128–136.
Haans, R. F., Pieters, C., & He, Z. L. (2016). Thinking about U: Theorizing and testing U-and inverted U-shaped relationships in strategy research. *Strategic Management Journal, 37*(7), 1177–1195.
Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference and prediction, 2nd ed.* New York: Springer.
Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., … Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature, 595*(7866), 181–188.
Holland, J. H. (1992). Genetic algorithms. *Scientific American, 267*(1), 66–73.
Hood, N., Urquhart, R., Newing, A., & Heppenstall, A. (2020). Sociodemographic and spatial disaggregation of e-commerce channel use in the grocery market in Great Britain. *Journal of Retailing and Consumer Services, 55*(July), Article 102076.
Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression, 3rd ed.* New York: Wiley.
Islam, T., Meade, N., & Sood, A. (2022). Timing market entry: The mediation effect of market potential. *Journal of International Marketing, 30*(1), 40–54.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
Kar, A. K. (2016). Bio inspired computing: A review of algorithms and scope of applications. *Expert Systems with Applications, 59*, 20–32.
Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research—Moving away from the "What" towards the "Why". *International Journal of Information Management, 54*, Article 102205.
King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science, 44*(2), 347–361.
Koponen, A. (1960). Personality characteristics of purchasers. *Journal of Advertising Research, 1*(September), 6–12.
Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
Lambert, Z. V. (1981). Profiling demographic characteristics of alienated consumers. *Journal of Business Research, 9*(1), 65–86.
Laukkanen, T. (2016). Consumer adoption versus rejection decisions in seemingly similar service innovations: The case of the Internet and mobile banking. *Journal of Business Research, 69*(7), 2432–2439.
Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., … Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science, 369*(6507), 1060–1062.
Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337–346.
Lehmann, D. R. (2020). The evolving world of research in marketing and the blending of theory and data. *International Journal of Research in Marketing, 37*(1), 27–42.
Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research, 43*(2), 276–286.
Lin, Y., & Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association, 101*(474), 578–590.
Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: Analysis and applications*. Cambridge: Cambridge University Press.
Ma, L., & Sun, B. (2020). Machine learning and AI in marketing: Connecting computing power to human insights. *International Journal of Research in Marketing, 37*(3), 481–504.
Massey, M., O'Cass, A., & Otahal, P. (2018). A meta-analytic study of the factors driving the purchase of organic food. *Appetite, 125*(1), 418–427.
McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observation studies. *Psychological Methods, 9*(4), 403–425.
McDonald, M., & Dunbar, I. (1998). *Market segmentation—how to do it, how to profit from it. 2nd ed.* London: Macmillan.
Meade, N., & Islam, T. (2006). Modelling and forecasting the diffusion of innovation: A 25-year review. *International Journal of Forecasting, 22*(3), 519–545.
Mihalopoulos, V. G., & Demoussis, M. P. (2001). Greek household consumption of food away from home: A micro econometric approach. *European Review of Agricultural Economics, 28*(4), 421–432.
Myers, J. H., Stanton, R. R., & Haug, A. F. (1971). Correlates of buying behavior: Social class vs. income. *Journal of Marketing, 35*(4), 8–15.
Namias, J. (1960). Intentions to purchase related to consumer characteristics. *Journal of Marketing, 25*(1), 32–36.
Palomba, A. (2020). Consumer personality and lifestyles at the box office and beyond: How demographics, lifestyles and personalities predict movie consumption. *Journal of Retailing and Consumer Services, 55*(July), Article 102083.
Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization. *Swarm intelligence*, *1*(1), 33-57.
Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50*(2), 491–500.
Ricciuto, L., Tarasuk, V., & Yatchew, A. (2006). Socio-demographic influences on food purchasing among Canadian households. *European Journal of Clinical Nutrition, 60*(6), 778–790.
Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology, 3*(2), 143–155.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.
Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science, 15*(4), 321–340.
Sachdeva, J., Kumar, V., Gupta, I., Khandelwal, N., & Ahuja, C. K. (2013). Segmentation, feature extraction, and multiclass brain tumor classification. *Journal of Digital Imaging, 26*(6), 1141–1150.
Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One, 10*(3), Article e0118432.
Sardianou, E., & Genoudi, P. (2013). Which factors affect the willingness of consumers to adopt renewable energies? *Renewable Energy, 57*(September), 1–4.
Schaeffer, S. E., & Sanchez, S. V. R. (2020). Forecasting client retention: A machine-learning approach. *Journal of Retailing and Consumer Services, 52*(January), Article 101918.
Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification* (pp. 149–171). New York: Springer.
Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*(6), 546–555.
Sheth, J. N. (1977). Demographics in consumer behavior. *Journal of Business Research, 5*(2), 129–138.
Sheth, J., & Kellstadt, C. H. (2021). Next frontiers of research in data driven marketing: Will techniques keep up with data tsunami? *Journal of Business Research, 125*, 780–784.
Shrestha, Y. R., He, V. F., Puranam, P., & von Krogh, G. (2021). Algorithm supported induction for building theory: How can we use prediction models to theorize? *Organization Science, 32*(3), 856–880.
Soopramanien, D. G., & Robertson, A. (2007). Adoption and usage of online shopping: An empirical analysis of the characteristics of "buyers", "browsers", and "non-internet shoppers". *Journal of Retailing and Consumer Services, 14*(1), 73–82.
Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348.
Sun, B., & Morwitz, V. G. (2010). Stated intentions and purchase behavior: A unified model. *International Journal of Research in Marketing, 27*(4), 356–366.
Tharp, M. C. (2001). *Marketing and consumer identity in multicultural America*. Thousand Oaks, CA: SAGE Publications.
Thompson, G. D., & Kidwell, J. (1998). Explaining the choice of organic produce: Cosmetic defects, prices, and consumer preferences. *American Journal of Agricultural Economics, 80*(2), 277–287.
Twedt, D. W. (1964). How important to marketing strategy is the "heavy user"? *Journal of Marketing, 28*(1), 71–72.
Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.
Vaughan, C. A., Collins, R., Ghosh-Dastidar, M., Beckman, R., & Dubowitz, T. (2017). Does where you shop and who you are predict what you eat? The role of stores and individual characteristics in dietary intake. *Preventative Medicine, 100*, 10–16.
Verbeke, W. (2005). Consumer acceptance of functional foods: Socio-demographic, cognitive and attitudinal determinants. *Food Quality and Preference, 16*(1), 45–57.
Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights, 1*(1), Article 100002.
Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance, 69*(3), 373–421.
Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing, 80*(6), 97–121.
Weltevreden, J. W. (2007). Substitution or complementarity? How the Internet changes city centre shopping. *Journal of Retailing and Consumer Services, 14*(3), 192–207.
Wheatley, J. J., Chiu, J. S., & Stevens, A. C. (1980). Demographics to predict consumption. *Journal of Advertising Research, 20*(6), 31–38.
Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review, 14*(4), 490–495.
White, C. J., & Tong, E. (2019). On linking socioeconomic status to consumer loyalty behaviour. *Journal of Retailing and Consumer Services, 50*(September), 60–65.
Williamson, E., Morley, R., Lucas, A., & Carpenter, J. (2012). Propensity scores: From naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research, 21*(3), 273–293.
Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.
Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics, 39*(1), 272–281.
Zwick, C. (1957). Demographic variation: Its impact on consumer behavior. *Review of Economics and Statistics, 39*(4), 451–456.

Dr. Towhid Islam is a Professor of Marketing and Consumer Studies and University Research Leadership Chair at the University of Guelph, Canada. His PhD is in Management Science is from Imperial College Business School, London, UK. His work has appeared in leading journals including Management Science, Journal of Consumer Research, Journal of Consumer Psychology, European Journal of Operational Research, International Journal of Research in Marketing, Journal of Business Research, and the International Journal of Forecasting among other outlets. The originality and novelty of his research approach has directly contributed to being awarded three Social Sciences and Humanities Research

Council of Canada (SSHRC) grants plus major funding from the Australian Research Council (ARC).

Dr. Nigel Meade is Emeritus Professor of Quantitative Finance at Imperial College London Business School. He has a BSc in Mathematics and Statistics and an MSc in Statistics from the University of Sheffield and a PhD in Management Science from Imperial College. His academic interests can be broadly summarized as statistical model building. He is associate editor of the International Journal of Forecasting.

Dr. Richard T. Carson is a distinguished professor in and former chair of the Department of Economics at the University of California San Diego Division of Social Sciences. Carson served as president of the Association of Environmental and Resource Economists (AERE). He is an elected Fellow of AERE and of the Agricultural and Applied Economics Association. Carson received a Ph.D. in resource economics and M.A. in statistics from the University of California, Berkeley and an M.A. in international relations from George Washington University. He has extensive experience in assessing the benefits and costs of environmental policies. His specialty is valuing non-marketed goods and new commercial products using a wide array of techniques. Carson's recent work on forecasting Chinese carbon dioxide emissions received international attention. His econometric research has focused on experimental and sampling designs, computer intensive and robust statistical techniques, discrete choice models and forecasting.

Dr. Jordan Louviere is Emeritus Professor of Marketing at the University of South Australia. He is an internationally recognized expert in conjoint analysis and consumer choice modelling, the inventor of best-worst scaling, and a researcher who comfortably bridges the gap between academia and industry, having consulted with many private and public sector organizations throughout Australasia, Europe and the Americas. He has authored or edited four books, published over 160 refereed journal articles and contributed more than 50 chapters to edited books. He has been the recipient of many grants from universities, industry bodies and government departments to conduct his research, and is currently working on a Social Sciences and Humanities Research Council of Canada (SSHRC) funded project to compare different designs for choice experiments and test their external validity against real consumer purchases and an US NSF funded project to understand and model how NIH grant panels make funding decisions.

Dr. Juan Wang is an Assistant Professor at the department of Marketing and Consumer Studies, University of Guelph. She obtained her Ph.D. from the Ivey Business School, University of Western Ontario (Canada). Her research stream focuses on how consumers learn and evaluate new products and new experiences, as well as consumer information processing in relation to consumer-brand relationship. Juan's research has been published in Journal of the Academy of Marketing Science (JAMS), Journal of Consumer Psychology (JCP), and Review of Marketing Research among other outlets.