The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Using Digital Purchasing Data to Generate Public Health Evidence: Learning Unhealthy Beverage Demand from Grocery Transaction Data

Hiroshi Mamiya,[1] Xing Han Lu,[1] Yu Ma,[2] David L. Buckeridge[1]

[1]McGill Clinical and Health Informatics, McGill University
[2]Desautels Faculty of Management, McGill University
hiroshi.mamiya@mail.mcgill.ca, xinghanlu@mail.mcgill.ca, yu.ma@mcgill.ca, david.buckerdige@mcgill.ca

Abstract

Unhealthy diet plays a major role in driving chronic disease incidence and prevalence. Taxation of unhealthy food has been proposed to improve population-level dietary patterns, and its effectiveness can be estimated by the prediction of the change in unhealthy food purchasing upon increase of food price. Recent availability of grocery transaction data from scanner technologies enables an accurate prediction of food sales. However, the very large number of product attributes in these data prohibits the application of conventional statistical learning algorithms. In this study, we explored the predictive performance of learning algorithms adapted for high-dimensional data, namely the Least Absolute Shrinkage and Selection Operator (LASSO) and Decision Tree Regressor with Adaptive Boosting (DTR-AdaBoost), in comparison with a conventional statistical learning based on Ordinary Least Square (OLS). LASSO demonstrated superior predictive accuracy to OLS, possibly due to its ability to reduce overfitting and collinearity across predictive features of food sales. DTR-AdaBoost showed the best predictive accuracy, suggesting the presence of extensive non-linearity between the predictive features in the transaction data and sales.

Introduction

Unhealthy diet is the leading preventable cause of death and disability globally, responsible for the epidemics of obesity and major chronic diseases including cardiovascular diseases, various forms of cancers and type-2 diabetes mellitus. Nutrition-related illness was responsible for 11 million deaths and a loss of 241 million disability adjusted life years in 2012 (Forouzanfar et al. 2015). Prevention of these chronic illness thus requires effective public health policies aimed at promoting selection of healthy food and discouraging energy-dense and nutrition-poor food items (i.e. junk food) at population scale.

Taxation of unhealthy food has received heightened interest as a viable public health intervention to reduce the population-level consumption of unhealthy food (Anne Marie Thow, Downs, and Jan 2014), most notably carbonated soft drinks (i.e. soda), which are a primary source of artificially added sugar intake and have an established epidemiologic association with obesity/overweight and other chronic illness (Cabrera Escobar et al. 2013; Hu 2013). Predictions of the change in soda sales upon an increase in their price can provide critical evidence for the effectiveness of price-based public health interventions. The validity of the prediction relies largely on the data, which must contain accurate sales records and measurement of factors affecting sales (e.g. price, promotion), but collection of such data for many items in a large number of stores has been prohibitively expensive until recently.

Ubiquitous scanner technologies now allow collection and centralization of electronic transaction data generated at the time of product purchase. These transaction data contain product attributes including price, promotion status and quantities sold on a weekly basis for tens of thousands of food items sold at multiple retail chain and store types (e.g. supermarket, pharmacy, convenience stores, supercenters, etc.) sampled from a wide geographic region. The data have been a valuable asset in guiding retail planning and practice (e.g. inventory management and price optimization) and providing market researcher insights in customer behaviors (Ma and Fildes 2017).

From the perspective of public health agencies and researchers, these data offer a tremendous potential for exploring the determinants of healthy and unhealthy food purchasing and for predicting community and consumer response (purchase) under the influence of policy interventions, such as taxation. Although limited in scale, applied public health

researcher occasionally used these data to evaluate potential effectiveness of intervention such as taxation (Zhen, Brissette, and Ruff 2014) on food purchase at population level.

Conceptually, prediction of food sales in a store can be generalized by the following linear time-series demand function;

$$Y_{it} \sim \beta_{io} + \beta_x W_{it} + \beta_A A_t + \beta_z Z_{jt} + \varepsilon_{it} \qquad (1)$$

Where Y_{it} represents the sales or market share of product *i* at time *t*, and W_{it} represents a matrix whose columns consist of product *i*'s own feature at time *t*, such as its price, promotion (e.g. price discounting, flyers, display at special location) and product group (e.g. soda, fruit juice, chocolates), with the corresponding vector of regression parameters presented as β_x. A_t represents a matrix of temporal features affecting food sales, typically including consumer price index, temperature, and week and/or month indicators depending on the temporal unit of analysis. Residuals for product *i* at time *t* are denoted by ε_{pt}.

Importantly, sales of item *i* are not only influenced by its own product attributes (features in the W_{it} matrix), but are also affected by the price, promotion and other attributes of competing items in the same store (e.g., sales of Coca-Cola maybe affected by price discounting of Pepsi). Attributes of such competing products *j* at time *t* are represented by the matrix Z_{jt}. Because a typical supermarket contains a large assortment of products, the number of distinct item, *Nj*, can exceed 2,000; thus, a sales prediction algorithm accounting for the influence price and promotion of *Nj* items on the sales of item *i* forms a high-dimensional matrix, with *Nj* columns multiplied by the number of product attributes per item.

In a large store, the dimension of the competing products in model matrix can exceed the number of transaction instances, with many product attributes linearly dependent on each other. In such a setting, traditional statistical estimators (e.g. Ordinary Least Square and Maximum Likelihood) may overfit the data or fail to find optimal parameter values for a prediction model, thereby requiring dimensionality reduction prior to developing a predictive function of sales.

Traditionally, reducing the dimension of competing products was addressed by an ad-hoc dimensionality reduction driven by convenience, such as limiting the competing products to a small number of high-market share brands, or aggregating their feature into broader product category (e.g. mean price of items belonging to the category of fruits juice) (Ma and Fildes 2017; Bajari et al. 2015). Without statistical rationale, these approaches discard the richness of product information, and thus fail to account for a complex pattern of competition among food items, resulting in reduced accuracy of prediction.

Accurate prediction of food sales is essential for a sound simulation of public health interventions targeting retail food markets. Using carbonated soft drinks (hereafter called Soda) as an example, we developed a framework for predicting food sales using the product features of all *Nj* competing beverage products for soda items in a store. Our study advances current methods by demonstrating how to effectively use high-dimensional retail purchase data for public health nutrition research.

Data and Methods

Transaction data

The point-of-purchase (i.e. store-level) grocery transaction data were generated by a product scanner at the time of purchase and were collected from a random sample of retail food outlets in the Province of Quebec, Canada by the Nielsen company. The data consist of food product information in each week at each sampled store, where product is uniquely defined by Universal Product Code (UPC). We extracted the transaction records for year 2013 from five stores belonging to a medium-sized supermarket chain located in the province of Quebec, Canada.

Although our transaction data was available at a weekly resolution, we aggregated the data to monthly by averaging the weekly sales and product attributes. Weekly data may reveal greater detail in the time-varying status of predictive features such as product price, but they also exhibit greater variance; therefore, monthly data often provide an equivalent prediction accuracy, in addition to the benefit of a reduced number of transaction instances (Volpe and Li 2012).

The response (output) variable was the log-transformed quantity of soda item *i* at month *t*, standardized to a single serving size (240ml). As in equation (1) predictor (input) variables consist of the own product attributes of *i* at month *t*, temporal factors, and the product attributes of *Nj* competing items in the same store at month *t*. These features are presented in Table 1. Because the number of competing items vary by store even within a single chain due to varying store size, local demand and inventory, the dimension of the competing product matrix varied by store. We therefore fit separate models for each of the five stores.

Demand prediction of soda

We aimed to predict sales of soda item *i* at month *t* using the following approaches; Ordinary Least Square (OLS), Least Absolute Shrinkage and Selection Operator (LASSO), and Decision Tree Regression with Adaptive Boosting (DTR-AdaBoost). Although detailed discussion of these algorithms is provided by others (Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome 2009), we will provide a brief description of their properties in the context of predicting sales from high-dimensional transaction data.

	Predictive features
Own product features	Monthly price
	Monthly discounting percent
	Beverage product category
	Frequency of display promotion
	Frequency of flyer promotion
Temporal features	Month indictor
	Monthly Consumer Price Index in Quebec
	Mean daytime temperature
	Mean monthly precipitation
	Number of statutory holiday
Competing product features	Monthly price
	Monthly discounting percent

Table 1: Predictive Feature of Soda Sales.

Excluding time index t for simplicity, we define the data with *N* instances of transaction (row of model matrix) each containing sales y_i for product *i* as well as its *p*-dimensional predictive features $x_i = (x_{i1} \dots x_{ip})$. A simple functional form for such data is specified as;

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad (2)$$

β_0 (an intercept for regression) and β_j represent regression parameters, and ε_i represents residual. The optimal value of β can be derived using the following objective function of OLS;

$$\min_{\beta_0,\beta_j} \sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad (3)$$

OLS is a widely used linear demand model. However, overfitting can occur when *p* is large. Moreover, when the number of columns exceeds the number of transaction instances ($p > N$, typically in a large store where the number of competing product, or Z matrix is large) or when the predictive features are linearly correlated with each other, a unique solution for β is not available. Thus, when training a model using such rank-deficient data, analysts are required to subset the predictive features including only the features that are uncorrelated with each other and are the most predictive of y_i. However, identifying such features can require extensive modeling effort.

LASSO combines the least square optimization (equation 3 above) and an ℓ_1 norm that imposes constraints to the value of parameters to find a sparse solution for the parameter vector β. Estimation of the LASSO parameters in a form of penalized residual sum of square is as follows;

$$\min_{\beta_0,\beta_j} = \sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j \quad (4)$$

LASSO allows a data-driven feature selection whose degree of sparseness (the number of β set to zero) can be controlled by the penalization term λ, whose value can be empirically determined by cross-validation.

LASSO and OLS impose a linearity assumption between predictor variables and the response (output) variable, which is often violated in the data generated from consumer behaviors. In such case, explicit specification of functional form (e.g. introduction of polynomial features) based on the theory in marketing science or empirical observation is necessary. However, searching correct non-linear specification for hundreds or thousands of predictor variables is not feasible. We thus implemented a non-parametric learning algorithm (DTR in this study) which is free from explicit specification of non-linear functional form.

DTR is a heuristic learning algorithm based on recursive binary splitting of predictive features, where the cut-point for each feature represents a decision boundary minimizing prediction error of the response variable y_i. As with LASSO and OLS, prediction error can be represented as the sum of squared error. The partitioned (branched) data can be further split based on the cut-point of another predictive feature. The partitioning can be repeated until certain criteria are met, for example reaching the maximum number of branches or minimum number of data points at each terminal node (leaf) to prevent overfitting of the tree to training data. Because a learned tree structure and the resulting predictive accuracy can be unstable i.e. predictive accuracy radically changes with a slight perturbation of training data, we implemented the algorithm in an ensemble (boosting) framework.

Boosting creates an accurate prediction algorithm from weighted combination of prediction generated by multiple algorithms. In Adaptive boosting, series of weak learning algorithms are trained on successively weighted data, where weight is specific to each transaction instance and proportional to the magnitude of prediction error by a previous algorithm.

Let the number of learning iteration as *M* and the *m*th learner as h_m, fitted to training data consisting of *N* transaction instance denoted as $x_i = (x_1 \dots x_N)$. For each instance of training data, weight w_i^m is applied. For the initial iteration, the sample weights are initialized uniformly as $w_i^1 = 1/\mathrm{N}$ for all instance *i*. Given a loss function such as Absolute Relative Prediction Error; $ARPE_i^m = |(h_m(x_i) - y_i)/y_i|$, weight-adjusted error can be determined as;

$$error_m = \sum_{i=1}^{N} ARPE_i^m \; w_i^m \quad (5)$$

which yields the following confidence of learner; $\alpha_m = error_m/(1 - error_m)$. Weights for the subsequent iteration of learning are updated as;

$$w_i^{m+1} = w_i^m \, \alpha_m^{1-ARPE_i^m} \quad (6)$$

The ensemble prediction of output for *i*th instance, $h_{final}(x_i)$, is a median of weighted output from each learner, where the weights are $\log(1/\alpha_m)$.

Tuning of learning algorithms

The transaction records from the sample of five stores consist of 5999 (4512 features), 5175 (3999 features), 4012 (2916 features), 5843 (4377 features), and 3737 (2676 features) instances. We randomly selected 30 percent of transaction instances to learn the optimum value of the penalization term λ that minimizes Mean-Squared prediction Error (MSE) for LASSO using 5-fold cross-validation. Using the best value of penalization, we learned LASSO regression parameters (β) to create a prediction function and measured its test prediction error from the rest of the 70% of transaction data. We again used 5-fold cross-validation, where training data was used to learn the parameters, whereas the test data was used to measure MSE from the resulting prediction function.

For DTR-AdaBoost, we learned the optimum tree depth and the number of DTR estimators to be combined in boosting using a grid search on random 30% of data, again with the selection of optimum value based on the minimization of MSE. Using these tuned values, we again developed (booted, or ensemble) tree-based prediction function and measured test prediction error using 5-fold cross-validation on the rest of the 70% of transaction instances.

For OLS, we did not perform dimensionality reduction, as the number of transaction record was slightly larger than the number of parameter ($p < N$). We estimated (trained) regression parameters and measured prediction error (MSE) of resulting function, again using 5-fold cross-validation using the random 70% of data as in the above two algorithms. The rest of the 30% of data that was used for the tuning of LASSO and DTR-AdaBoost was unused for OLS.

Results

The test errors from each model in each store are presented in Table 2. For all stores, LASSO demonstrated substantially smaller prediction error than OLS. The finding suggests extensive correlation across the predictors, which resulted in inaccurate parameter estimation by OLS. Such multi-collinearity is highly plausible, since price and promotion patterns of certain beverage items are likely to fluctuate together, especially when the items belong to the same brand. As an example, 2.0 L bottle and a package of 350ml cans of Coca-Cola could be discounted within the same time period.

Store ID	Model	MSE
707	OLS	1.091
	LASSO	0.703
	AdaBoost + DTR	0.428
1183	OLS	1.289
	LASSO	0.936
	AdaBoost + DTR	0.537
1613	OLS	1.877
	LASSO	1.028
	AdaBoost + DTR	0.758
1627	OLS	1.485
	LASSO	0.932
	AdaBoost + DTR	0.505
39699	OLS	1.498
	LASSO	0.953
	AdaBoost + DTR	0.637

Table 2: Mean Squared Error (MSE) of Predicted Log Soda Sales by Algorithms for Five Retail Supermarkets.

DTR-AdaBoost demonstrated substantially lower prediction error than LASSO for all stores. It is possible that the relationship between the predictors and sales are highly non-linear in our data. Given a large feature space in our transaction data, it was not feasible to introduce best-fit polynomial features for each predictor or combination of predictors for OLS and LASSO. Imposing the linearity constraint to these features led to the underfit prediction models generated by OLS and LASSO. On the other hand, non-parametric learning algorithms such as DTR accommodate any functional form in the presence of extensive non-linearity.

Importantly, we observed a noticeable difference in prediction error across stores, even though these stores are from the same supermarket chain. The difference was particularity large for OLS. The variation likely reflects unobserved attributes of store practice (e.g. product assortment, price and promotion) and customer attributes.

Discussion

Accurate prediction of sales of unhealthy and healthy food is critical to assess the likely effects on food consumption of retail-based public health interventions, such as taxation. Grocery transaction data provide detailed product information from a large number food items sold in numerous

retail food outlets. These data allow the development of accurate prediction function for food sales. However, because sales of a one item are influenced by a large number of competing items in the same store, prediction algorithms capable of incorporating complex pattern of competition must be used to make use of the rich product information contained in transaction data. We explored the accuracy of parametric (OLS and LASSO) and non-parametric learning algorithms (DTR-AdaBoost) for predicting soda sales in a sample of retail supermarkets.

Although OLS is regarded as the best unbiased linear estimator with widespread application, its utility for building prediction functions from high-dimensional feature spaces (in our study, incorporating many competing products into prediction equation) may be limited, especially when extensive overfitting and correlation among predictive features is suspected. In addition, although we selected medium-sized supermarkets whose product assortment (and thus dimensionality) is smaller than transaction instances ($p < N$) in this study, many stores have lager products space, resulting in unidentifiable parameter values using OLS.

It is possible to improve the performance of OLS by removing correlated predictors by existing feature selection method, such as stepwise selection. However, such modeling may take additional computational efforts when applied to many of stores (our original transaction data was generated by greater than 400 stores), whereas LASSO has a built-in feature selection method whose degree of penalization, or removal of unwanted features, can be readily controlled by a single penalization parameter. Thus, researchers analyzing scanner data may benefit from using non-traditional econometric modelling methods.

Given its the lowest prediction error, DTR-AdaBoost is our algorithm of choice based on the data from the five stores we selected in this study. This finding is consistent with a similar study, which found ensemble tree-based algorithm to have superior performance to other learning methods (Bajari et al. 2015). Although parametric methods, such as OLS and LASSO are computationally efficient and require smaller amounts of training data, their predictive performance can suffer when the data violate the assumptions of parametric models, such as a linear association between the predictor and response variables. We thus recommended researchers and analyst to investigate the performance of both approaches.

A main limitation of our study is a failure to include non-beverage food item as competing product of soda. It is possible that pricing and promotion of salty snacks or solid energy-dense hood (e.g., chocolates), also influence purchase of soda, by either competing for consumers' budget or by complementing each other (e.g., soda and snacks tend to be bought together). Although inclusion of price and promotion status from tens of thousands of solid food items may improve prediction accuracy, the large increase in dimension and resulting computational overhead prohibited us in this study from creating an even larger-scale prediction model that would take advantage of the full richness of the product space in grocery transaction data.

Future work includes the prediction of multiple food products. Although the current study focused on soda as an initial example, our methodological framework can be readily extended to other food types of public health interest, including sales of fruits, vegetable, and sports and energy drinks (significant and rising sources of artificially added sugar among adolescents). Similarly, the current study focused on five superstores as an initial exploration, we plan to estimate the predictive accuracy for all 427 stores that provided grocery transaction data. Doing so may allow us to investigate spatial patterning of prediction accuracy as a step to searching for factor(s) driving the variation of prediction accuracy across stores.

References

Anne Marie Thow, Shauna Downs, and Stephen Jan. 2014. "A Systematic Review of the Effectiveness of Food Taxes and Subsidies to Improve Diets: Understanding the Recent Evidence." *Nutrition Reviews* 72 (9):551–65.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. "Demand Estimation with Machine Learning and Model Combination." Working Paper 20955. National Bureau of Economic Research.

Cabrera Escobar, Maria A., J. Lennert Veerman, Stephen M. Tollman, Melanie Y. Bertram, and Karen J. Hofman. 2013. "Evidence That a Tax on Sugar Sweetened Beverages Reduces the Obesity Rate: A Meta-Analysis." *BMC Public Health* 13:1072.

Forouzanfar, Mohammad H, Lily Alexander, H Ross Anderson, Victoria F Bachman, Stan Biryukov, Michael Brauer, Richard Burnett, et al. 2015. "Global, Regional, and National Comparative Risk Assessment of 79 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks in 188 Countries, 1990–2013: A Systematic Analysis for the Global Burden of Disease Study 2013." *The Lancet* 386 (10010):2287–2323.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. 2009. *The Elements of Statistical Learning - Data Mining, Inference, Second Edition*. Second edition.

Hu, F. B. 2013. "Resolved: There Is Sufficient Scientific Evidence That Decreasing Sugar-Sweetened Beverage Consumption Will Reduce the Prevalence of Obesity and Obesity-Related Diseases." *Obesity Reviews: An Official Journal of the International Association for the Study of Obesity* 14 (8):606–19.

Ma, Shaohui, and Robert Fildes. 2017. "A Retail Store SKU Promotions Optimization Model for Category Multi-Period Profit Maximization." *European Journal of Operational Research* 260 (2):680–92.

Volpe, Richard J., and Chenguang Li. 2012. "On the Frequency, Depth, and Duration of Sales at High–Low Pricing Supermarkets." *Agribusiness* 28 (2).

Zhen, Chen, Ian F. Brissette, and Ryan Richard Ruff. 2014. "By Ounce or by Calorie: The Differential Effects of Alternative Sugar-Sweetened Beverage Tax Strategies." *American Journal of Agricultural Economics*, June.