MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190221

Application of Machine Learning and Grocery Transaction Data to Forecast Effectiveness of Beverage Taxation

Xing Han Lu[a,b*], Hiroshi Mamiya[a*], Joseph Vybihal[b], Yu Ma[c], David L. Buckeridge[a]

[a] *Surveillance Lab, McGill Clinical and Health Informatics, McGill University, Montreal, Quebec, Canada*
[b] *School of Computer Science, McGill University, Montreal, Quebec, Canada*
[c] *Desautels Faculty of Management, McGill University, Montreal, Quebec, Canada*

Abstract

Sugar Sweetened Beverages (SSB) are the primary source of artificially added sugar and have a casual association with chronic diseases. Taxation of SSB has been proposed, but limited evidence exists to guide this public health policy. Grocery transaction data, with price, discounting and other information for beverage products, present an opportunity to evaluate the likely effects of taxation policy. Sales are often non-linearly associated with price and are affected by the prices of multiple competing brands. We evaluated the predictive performance of Boosted Decision Tree Regression (B-DTR) and Deep Neural Networks (DNN) that account for the non-linearity and competition across brands, and compared their performance to a benchmark regression, the Least Absolute Shrinkage and Selection Operator (LASSO). B-DTR and DNN showed a lower Mean Squared Error (MSE) of prediction in the sales of most major SSB brands in comparison to LASSO, indicating a superior accuracy in predicting the effectiveness of SSB taxation. We demonstrated the application of machine learning methods and large transactional data from grocery stores to forecast the effectiveness food taxation.

Keywords:

Beverages; Machine learning; Public policy

Introduction

Unhealthy diet is the leading preventable cause of global death and disability, claiming 11 million lives and 241 million disability adjusted lost life years in 2012 [1]. Diet-related chronic diseases, such as obesity, cardiovascular diseases, cancers, and type-2 diabetes mellitus impose a considerable burden on society and individuals. Taxation has been proposed as a public health policy to discourage the purchasing of unhealthy foods [2], most notably Sugar Sweetened Beverages (SSB). The primary source of artificially added sugar with a recognized association with obesity and major chronic diseases [3,4], SSB consists of beverages such as soda (carbonated soft drinks), fruits drinks, sports and energy drinks each containing many product brands (e.g. Coca-Cola™ and Pepsi™ in the category of soda). The expected effectiveness of taxation is determined by the magnitude of reduction in SSB purchasing that is likely to occur with an increase in the price of SSB. Formally, this value is called the *price elasticity of demand* and is quantified as the percent reduction in product purchased in response to a one percent increase in price.

Grocery transaction data are generated by scanning products at the time of purchasing from a large number of retail food outlets and can be used to estimate changes in SSB purchasing from price fluctuations. The data include the quantity of food products sold and ourchase details, such as the price and promotions (e.g. discounting, in-store display, and flyer). Although rarely used for public health research, grocery transaction data can be used to predict SSB sales conditional on pricing, promotions, consumer demographic and economic attributes of the store neighborhood (e.g. income and family size). Because sales of a product are influenced by its features (*focal features*), but also by the features of competing products in the same store (*competing features*), the prediction of beverage purchasing must take into account the influence of numerous competing brands.

Due to correlations in price and promotion across many food products, feature selection is critical. Researchers previously performed ad-hoc dimensionality reduction, such as aggregating product sales and features into broader SSB categories or modeling only a small number of brands [5]. These approaches masks the complex patterns of competition among individual food products, potentially resulting in biased estimate of price elasticity due to aggregation bias or omitted confounders. Therefore, prediction at the level of individual food items or brands is critical.

More importantly, associations between product features and sales are non-linear (i.e. deal-effect curve), and multiple product features can jointly affect sales through statistical interactions, or non-additive effects due to competitive interference and synergistic effect of promotions. The shape of the relationship between sales and product features is determined by a mix of factors including: 1) a threshold effect, where the quantity of sales remains constant until the price or promotion reaches a threshold value; 2) a saturation effect, where the growth of sales plateaus as consumers are de-sensitized by price reduction or promotions above certain levels; 3) a cross-item deal effect, where the sales of brands are affected by the price of a large competing brands; 4) an interaction between features of multiple brands e.g., price reduction of a brand fails to increase sales due to promotional activities of a larger competing brand; and 5) an interaction between features of the same brand e.g. the effect of discounting can be enhanced by promotions of the same brand [6].

While parametric estimators (e.g. linear regression) are traditionally used to model product demand, manual specification of non-linear functions and interactions is not feasible with dozens or hundreds of competing product features. In contrast, non-parametric algorithms, such as decision trees and artificial neural networks, naturally incorporate non-linear associations and interactions. Although typically used for classification, these algorithms can be used in a regression context where the output variable is numeric (i.e. sales).

Table 1 - Description of Brand-level, Temporal, and Store Neighborhood Predictive Features of Sugar Sweetened Beverage (SSB) sales. Each transaction record consists of sales (target variable) of specific SSB brands and these features at a given week and store. The third column describes either the number of categories (if categorical), or the range of values (if numerical).

Feature description	Type	Category count or numeric range
Brand-level features		
Chain code where product was sold	Categorical	3
Percent price discount (%)	Numeric	0 - 98.225
Prices in Canadian cents	Numeric	0.001 - 1399
Display advertisement frequency	Numeric	0 – 1
Flyer advertisement frequency	Numeric	0 – 1
Brand name	Categorical	154
Store code where product was sold	Categorical	74
Temporal features		
Month of Sale	Categorical	12
Week of Sale	Categorical	53
Store neighborhood features		
Proportion of post-secondary certification	Numerical	0.435 - 0.891
Average family size	Numerical	2 – 3
Proportion of family with child	Numerical	0.207 - 0.530
Proportion of single parent family	Numerical	0.103 - 0.274
Median family income ($/family)	Numerical	38046 - 108735
Proportion of immigrants	Numerical	0.005 - 0.606
Number of dwellings (families)	Numerical	2885 - 58710
Total population (inhabitants)	Numerical	7460 - 133570
Dwelling density (families/km2)	Numerical	1.145 - 6289.960
Total		
Log of Weekly Sales of brand	Numerical	-1.403 - 11.761

To date, SSB taxation is rarely implemented in developed nations, and the price elasciticy of demand is for the most part estimated by observational data using parametric demand estimation models that tend to suffer the limitations mentioned above [7]. Traditional econometric approaches (e.g. linear regression) to estimate the effect of price on SSB (price elasticity of SSB demand) is infeasible when modling the interaction of a large number of beverage products. We thus aim to evaluate the accuracy of non-parametric learning algorithms for predicting the sales of SSB from scanner grocery transaction data.

Data

We obtained weekly transaction records of food products purchased from 44 stores sampled to be geographically representative of three large retail grocery chains in the province of Quebec, Canada between 2008 and 2013. The data were indexed by time (week), store identification code, product name, price, and three promotional activities: discounting, in-store display (placement of a product in a prominent location) and flyer advertising. Price indicates the dollar amount paid by the consumer at the time of purchase (i.e. net price), while discounting is the percent reduction of the purchase price from the regular price calculated as the maximum purchase price in a three month trailing window [8].

Among many types of beverages, we were interested in predicting the sales of SSB, or artificially sweetened beverages which included soda, energy and sports drink, carbonated fruit beverage, frozen fruit juice, fruit drink, and carbonated soda water. There were 2,608 distinct SSB products defined by brand, flavor, and package type. Because products in the same brand tended to exhibit similar pricing and promotional patterns, we aggregated the value of sales, pricing and promotion into a smaller set of 154 distinct SSB brands, such as Coca-Cola™ and Pepsi™. Brand-level predictive features (i.e. price, discounting, display, and flyer advertisement) were calculated as the mean (price and discounting) and proportion promotion (display and flyer) across the products belonging to the brand.

Let $t := week, i := brand, j := store$. There were 1,509,280 weekly transaction records for the 154 SSB brands across all stores, with each record representing the brand-specific sales denoted as Y_{ijt}, which was the target variable and defined as the natural-log of the sales of brand i in store j at week t. The sales quantity was standardized to the U.S Food and Drug Administration serving size of 240 milliliters. Although the log transformation was relevant to parametric regression modeling only [9], we applied this transformation in accordance with existing practice in demand modeling.

The vector of brand-level focal features was denoted as X_{ijt} (Table 1, Brand-level features). We let S_j be the features of the store where the products were sold (categorical indicator of chain and store identification code) and store neighborhood socio-economic and demographic features that may influence food purchasing were obtained from the 2011 Canadian census (Table 1, Store neighborhood features). We let M_t and W_t represent categorical features indicating the month and week for each record to account for temporal fluctuations in purchasing (e.g. increases on holidays). As noted above, sales of a brand depend on the pricing and promotion of that brand (*focal brand features*) and on the features of popular competing brands (*competing brand features*). Because a few brands accounted for most of the market share in each SSB category (e.g. Coca Cola™ and Pepsi™ have nearly 70% of share in the soda category), their brand features have a strong influence on the sales of other brands. Thus, we extracted price and promotions of twenty brands with the highest market

Table 2 - Mean Squared Error of most popular brands of Sugar Sweetened Beverages

	Pepsi	Coca Cola	Seven Up	Crush	Sprite	Canada Dry	Nestle
LASSO	0.51	0.44	0.46	0.35	0.45	0.38	0.41
B-DTR	**0.17**	**0.16**	0.22	0.28	**0.22**	**0.24**	0.41
DNN	0.19	0.23	**0.21**	**0.23**	0.23	0.27	**0.31**

share among SSB that are denoted as C_{kjt}. The dimension of each feature vector was: (X_{ijt}, 245), (C_{kjt}, 80), (S_j, 9), (M_t, 12), and (W_t, 53).

Methods

We used two non-parametric methods: an ensemble of Decision Trees with Adaptive Boosting (B-DTR) and a fully-connected Deep Neural Network (DNN). The baseline model was a regularized linear parametric model (LASSO, or Least Absolute Shrinkage Selection Operator). The DNN was implemented in Keras [10], and the other models were implemented in Scikit-Learn [11]. Unless otherwise noted, normalization was done using standard mean shifting and variance scaling.

LASSO regression identifies a sparse set of features through shrinkage via L_1 regularization [12] and has been previously used for demand forecasting in high-dimensional feature space [13], even though explicit specification of non-linear features (e.g. spline) becomes unrealistic when modeling the sales of a large number of brands. We selected the regularization parameter λ by iterating over a range of values and selecting the one with lowest average mean squared error (MSE) through three-fold Cross Validation.

Decision Tree Regression is a rule-based learning algorithm that identifies a binary segmentation of predictive features, where the cut-point for each feature represents a decision boundary that minimizes the prediction loss (e.g sum of squared errors) for a target vector Y_{ijt}. The partitioning ends when pre-specified criteria, such as a maximum number of branches or a minimum number of observations at each terminal node, are met. The decision trees implement the Classification and Regression Trees algorithm due to its ability to predict numerical values [14].We used Drucker's improved Adaptive Boosting [15] meta-estimator to form an ensemble of 100 weak learners. The weight of each learner was determined by a linear loss. Each learner was a Decision Tree with varying depths, set to a maximum depth of 30 nodes. The value of each node was determined by the partition that best minimized the MSE.

The DNN model with the best results had four fully connected layers. Adam optimization was used to enable convergence with large data and noisy gradients [16]. The optimum values of exponential decay rates and fuzzy factors were selected based on training stability and the ability to converge. The network weight parameters were initialized using Normalized Initialization [17] to accelerate convergence. We trained the model using mini-batches of 128 samples to leverage the richness of the data and to provide inherent regularization [18], while maintaining a stable training process. We chose the activation function to be a Rectified Linear Unit due to its biological properties and strong experimental results on high-dimension datasets [19], due in part to its non-linearity, which allows the DNN to learn complex relationships between features.

The DNN had an input layer dimension of 389, and fed a 400-dimension vector to the first hidden layer. The first hidden layer output a 100-dimension vector to the next layer with a L_1 regularization and ReLU activation. The last hidden layer output was a 25-dimension vector to the output layer. The final layer outputs a single numerical value corresponding to the predicted log of sales, using a linear activation function to take into account negative target values (brands with extremely low sales has negative log values).

We extracted the first five years (2008-2012) of the transaction data for training and validation. We randomly sampled 90% of these data as the training set for learning algorithm parameters, leaving the remaining 10% as the validation set for evaluating the prediction accuracy of the algorithms. The final year (2013) of data was reserved to estimate prediction accuracy, measured as Mean Squared Error (MSE). Data were managed using Numpy, Pandas and PostgreSQL.

Results

The MSE for the prediction of all SSB brands in the 2013 transaction data was 0.67, 0.72, and 0.91 for DNN, B-DTR, and LASSO, respectively.

At the individual brand level, DNN, B-DTR, and LASSO showed best predictive performance for 80, 31, and 21 brands, respectively. Prediction error of seven most popular SSB brands driving overall sales of SSB is presented in Table 2. The DNN and B-DTR had comparable prediction accuracy for these brands, while LASSO showed the lowest accuracy except for the Nestle brand. Category-level prediction is of public health interest, since each beverage category can have different health effects. We thus calculated a category-level MSE (Table 3). The DNN had the highest prediction accuracy for the soda category, the most important source of added sugar consumption. LASSO showed the lowest prediction accuracy for all the categories.

Table 3 – Mean Squared Error for each Beverage Category, Across all Stores

	Soda	Soda Water	Energy Drinks	Frozen Juice	Frozen Drinks
B-DTR	0.650	0.743	0.956	**0.600**	**0.631**
DNN	**0.609**	**0.647**	**0.691**	0.660	0.640
LASSO	0.756	0.852	0.819	1.228	0.887

Using the most accurate predictive algorithm (DNN), we generated price elasticity, which is the predicted percent reduction in SSB sales conditional on increased beverage prices in reference to the predicted SSB sales due to observed SSB price. The elasticity across all stores was -1.74 (95% Confidence Interval [CI]: -0.89, -2.69), implying that the increase of SSB price by 1% results in the expected reduction of SSB purchasing by 1.74%.

Discussion

The superior prediction accuracy demonstrated by B-DTR and DNN over LASSO is likely due to their ability to model non-linear relationships and interactions across predictive features of the 154 brands. Conventional approach to model low-dimentional data, such as Ordinary Least Square and its extension adapted for a high-dimentional varaible space (LASSO) may be a suboptimal approach in predicting the sale of SSB in a competitive retail environment due to its linear constraint Although it is theoretically possible to manually specify appropriate non-linear functional forms guided by model-fit criteria (e.g. Akaike's Information Criterion) in LASSO, this approach is not feasible when the number of competing brands grows large.

The estimated own-price elasticity was slightly higher than the mean estimate from a systematic review of previously reported observatioinal studies in the United States (-1.21, range: -0.71 to -3.87) [7] but was lower than the elasticity (-3.1) estimated by the exsiting natural experiment in Berkley, California [20].

Non-parametric learning algorithms, in particular DNN as demonstrated by the superior prediction accuracy in our study, are free from the linearity assumptions. They are a promising alternative to predict the sales of a large number of brands, thus providing a better quality of public health evidence to guide policy on SSB taxation.

Future work includes in-depth investigation of store-level difference in the estimated effectiveness of taxation, or price elasticity. Identification of store-level features (e.g. promotion and the number of competing items) and neighborhood features driving differential store-level elasticity is a critical public health interest, since the analysis allows the characterization of communities that are less likely to benefit from taxation and consequently in need of community-specific interventions addressing local obstacles of healthy eating. As well, variation of prediction error (MSE) across stores and beverage categories warrants thorough investigation to improve the performance of the models.

A limitation of our study, as in the majority of research estimating price elasticity using observational transaction data, is the lack of the sales data generated by the simultaneous increase of SSB price across all beverages due to taxation, which may induce purchasing patterns different from the observed price fluctuation without the taxation. Thus, our ability to formerly validate the predictive performance of these algorithms under SSB taxation is limited. Future research therefore includes the evaluation of these algorithms in a (rare) setting where the taxation is applied. Nevertheless, our study adds an important contribution to the existing methodological research in estimating the price elasticity of demand using the transaction of a large number of food products.

From a public health perspective, unique aspects of our study include the evaluation of the effectiveness of health policy using a large amount of transactional data, which have become recently available to public health researchers. More importantly, analytical strategies for learning food demand from high-dimensional data were also lacking. Our novel approach to predicting SSB sales using machine learning methods should improve the accuracy of estimating the effectiveness of SSB taxation, while demonstrating the effective use of scanner transaction data for public health research. In addition, our analytical framework offers a scalability in terms of geographic coverage (e.g. expansion to national-level analysis) and food categories that are not modelled in this study but are often proposed as part of SSB including 100% fruit juice and sweetened milk, albeit having much lower market share than soda.

Conclusions

Overall, our study demonstrated the utility of non-parametric machine learning algorithms in predicting unhealthy beverage sales in the presence of complex interactions across a large number of SSB brands and non-linear effect of sales and product attributes. The higher accuracy of the non-parametric learning algorithms, in particular DNN, over parametric algorithm highlights its utility in predicting the potential effect of SSB taxation.

References

[1] M.H. Forouzanfar, L. Alexander, H.R. Anderson, V.F. et al., 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** (2015), 2287–2323. doi:10.1016/S0140-6736(15)00128-2.

[2] A.M. Thow, S. Downs, and S. Jan, A systematic review of the effectiveness of food taxes and subsidies to improve diets: Understanding the recent evidence, *Nutr Rev* **72** (2014), 551–565. doi:10.1111/nure.12123.

[3] M.A.C. Escobar, J.L. Veerman, S.M. Tollman, M.Y. Bertram, and K.J. Hofman, Evidence that a tax on sugar sweetened beverages reduces the obesity rate: a meta-analysis, *BMC Public Health* **13** (2013), 1072.

[4] F.B. Hu, Resolved: there is sufficient scientific evidence that decreasing sugar-sweetened beverage consumption will reduce the prevalence of obesity and obesity-related diseases, *Obes Rev Off J Int Assoc Study Obes* **14** (2013), 606–619. doi:10.1111/obr.12040.

[5] P. Bajari, D. Nekipelov, S.P. Ryan, and M. Yang, Demand estimation with machine learning and model combination, National Bureau of Economic Research, 2015. http://www.nber.org/papers/w20955 [accessed April 22, 2016].

[6] H.J. Van Heerde, P.S. Leeflang, and D.R. Wittink, Semiparametric analysis to estimate the deal effect curve, *J Mark Res* **38** (2001), 197–215.

[7] L.M. Powell, J.F. Chriqui, T. Khan, R. Wada, and F.J. Chaloupka, Assessing the potential effectiveness of food and beverage taxes and subsidies for improving public health: a systematic review of prices, demand and body weight outcomes, *Obes Rev* **14** (2013), 110–128. doi:10.1111/obr.12002.

[8] J.S. Raju, The Effect of Price Promotions on Variability in Product Category Sales, *Mark Sci* **11** (1992) 207–220.

[9] P. Leeflang, J.E. Wieringa, T.H.A. Bijmolt, and K.H. Pauwels, Modeling Markets: Analyzing Marketing Phenomena and Improving Marketing Decision Making, Springer-Verlag, New York, 2015. //www.springer.com/gp/book/9781493920853 (accessed April 26, 2018).

[10] F. Chollet, and others, Keras, 2015.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, Scikit-learn: Machine learning in Python, *J Mach Learn Res* **12** (2011), 2825–2830.

[12] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat Soc Ser B Methodol* (1996) 267–288.

[13] S. Ma, and R. Fildes, A retail store SKU promotions optimization model for category multi-period profit maximization, *Eur J Oper Res* **260** (2017), 680–692. doi:10.1016/j.ejor.2016.12.032.

[14] L. Breiman, Classification and regression trees, Routledge, 2017.

[15] H. Drucker, Improving Regressors Using boosting techniques, *ICML* **97** (1997), 107–115..

[16] D.P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *CoRR* **abs/1412.6980** (2014). http://arxiv.org/abs/1412.6980 (accessed April 1, 2019).

[17] X. Glorot, and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, *Proc Thirteen Int Conf Artif Intell Stat* (2010),249–256.

[18] Y. Bengio, Practical recommendations for gradient-based training of deep architectures, in: Montavon, G, Orr, G.B, Muller, K-R. (eds.) *Neural Netw. Tricks Trade* (2nd ed.) (pp. 437–478), Springer, Berlin,2012..

[19] X. Glorot, A. Bordes, and Y. Bengio, Deep sparse rectifier neural networks, in: Proc. Fourteenth Int. Conf. Artif. Intell. Stat., 2011: pp. 315–323.

[20] J. Falbe, H.R. Thompson, C.M. Becker, N. Rojas, C.E. McCulloch, and K.A. Madsen, Impact of the Berkeley excise tax on sugar-wweetened beverage consumption, *Am J Public Health* **106** (2016) 1865–1871. doi:10.2105/AJPH.2016.303362.

Address for correspondence

* These authors contributed equally to this work.

Corresponding author name: Hiroshi Mamiya

Address: 1140 Ave Pine, Montreal, Quebec, Canada, H1A 1A3

E-mail: hiroshi.mamiya[at]mail.mcgill.ca