Determining the Importance of an Attribute in a Demand System

Structural versus Machine Learning Approach

Syed Badruddoza, Washington State University, s.badruddoza@wsu.edu
Modhurima Dey Amin, Washington State University, modhurima.amin@wsu.edu

Selected Poster prepared for presentation at the 2019 Agricultural & Applied Economics Association Annual Meeting, Atlanta, GA, July 21-23

Determining the Importance of an Attribute in a Demand System
Structural versus Machine Learning Approach

Presented in 2019 conference of AAEA Agricultural & Applied Economics Association

Syed Badruddoza and Modhurima Dey Amin
School of Economic Sciences, Washington State University

Washington State University

- Products are just collection of attributes. For example, a sedan car has attributes like miles per gallon (MPG), rating, weight etc.
- Wouldn't it be great if you could figure out which attributes are the most valuable to your customers?
- What if you did that without constructing complex structural models and finding instrumental variables?
- Our simulations show that a machine learning model might help you get a rough idea on how important the attributes of your product are to the customers!

Structural Approach: BLP Model

Berry-Levinsohn-Pakes (BLP 1995) is a common structural approach in demand analysis, where the indirect utility of consumer i with income y from brand j in market m is,

$$u_{ijm} = \alpha_i(y_i - p_{jm}) + \mathrm{x}_{jm}\beta_i + \xi_{jm} + \epsilon_{ijm}$$

where, x_{jm} and ξ_{jm} are vectors of observed and unobserved product attributes, respectively; p is price, ϵ is mean-zero stochastic term with distribution G, and $\{\alpha, \beta\}$ is a set of parameters to be estimated.

For instance, if j is sedan car brands, $x1$ is miles per gallon (MPG), higher β_{x1} indicates higher utility from additional MPG. Some parameters are heterogeneous across consumers, called random coefficient, e.g., $\beta_{x1} \sim \mathrm{Normal}(\bar{\beta}_{x1}, \sigma^2_{\beta_{x1}})$.

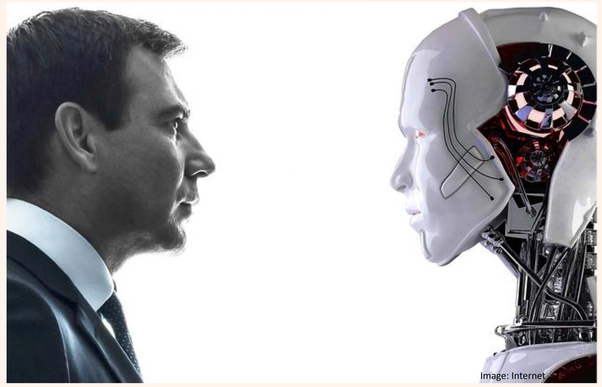
The probability of one brand j will be preferred to another brand k is

$$P(u_{ijt} > u_{ikt}) = P(\epsilon_{ikt} - \epsilon_{ijt} < \alpha_i p_{jm} - \alpha_i p_{km} + \mathrm{x}_{km}\beta_i - \mathrm{x}_{jm}\beta_i + \xi_{km} - \xi_{jm})$$
$$= \int_\epsilon I(\alpha_i p_{jm} - \alpha_i p_{km} + \mathrm{x}_{km}\beta_i - \mathrm{x}_{jm}\beta_i + \xi_{km} - \xi_{jm})\, dG(\epsilon)$$

where I is an indicator function representing the inequality. BLP assumes that ϵ has Type I extreme value distribution, so the difference has a Logit form. Brand j's share in market m is aggregation of individual preferences,

$$s_{jm} = \frac{\exp(-\alpha_i p_{jm} + \mathrm{x}_{jm}\beta_i + \xi_{jm})}{1 + \sum_{k \neq j} \exp(-\alpha_i p_{km} + \mathrm{x}_{km}\beta_i + \xi_{km})}$$
$$\Rightarrow \ln(s_{jm}) \approx \ln(s_{0m}) - \alpha_i p_{jm} + \mathrm{x}_{jm}\beta_i + \xi_{jm}$$

where, 0 represents brands that are missing in the data. Price is endogenous to market share. So BLP estimation requires instrumental variables in a Generalized Method of Moments framework.



Machine Learning Approach: Random Forests (RF)

RF is a supervised machine learning algorithm widely used for classification, regression, and prediction. The model randomly constructs a multitude of decision trees and aggregates their predictions (Breiman 2001).

Assume q attributes $p, \mathrm{x} \in \chi \subset \mathbb{R}^q$ are observed, and the goal is to predict log of observed market share vector $\ln(s)$. Define the nonparametric regression function as $\mathcal{S}(\chi) = \mathbb{E}[\ln(s)|\chi]$ and sample $\mathcal{D}_n$.

A random forest is a predictor consisting of a collection of T randomized regression trees. The t-th tree estimation takes the form

$$\mathcal{S}_n(\chi; \lambda_t, \mathcal{D}_n) = \sum_{\mathcal{D}^*_{t,n}} \frac{I_{\chi \in A(\chi;\lambda_t,\mathcal{D}_n)} \ln(s)}{\sum^n I_{\chi \in A(\chi;\lambda_t,\mathcal{D}_n)}}$$

where, random variable λ is used to resample and successive directions for splitting, $\mathcal{D}^*_{t,n}$ is observations resampled with replacement, $A(\chi; \lambda_t, \mathcal{D}_n)$ is the cell containing χ. Trees are aggregated by

$$\mathcal{S}_{T,n}(\chi; \lambda, \mathcal{D}_n) = T^{-1} \sum_1^T \mathcal{S}_n(\chi; \lambda_t, \mathcal{D}_n)$$

Importance of an attribute q is the increase in Mean-Square-Error of prediction when the attribute is randomly removed from the model. The importance is,

$$T^{-1} \sum^T [R_n[\mathcal{S}_n(.\,, \mathcal{D}^q_{t,n})] - R_n[\mathcal{S}_n(.\,, \mathcal{D}_{t,n})]]$$

where, the prediction error is defined by $R_n[\mathcal{S}_n(.\,, \mathcal{D})] = |\mathcal{D}|^{-1} \sum_{\mathcal{D}} (\ln(s) - \mathcal{S}_n(\chi; \lambda_t))^2$.

Sample Simulation Result

BLP Marginal Utility Estimates vs. RF Importance of Product Attributes

Attributes	Type	True value	BLP Estimate	BLP p-values	RF Importance
x6	Ex	3	3.019	0.000	4.707
x7	Ex	-2.2	-2.117	0.000	2.575
x5	Ex	2.2	2.250	0.000	1.918
price	En	-0.2	-0.195	0.000	0.726
x1	RC	0	0.311	0.674	0.451
x4	Ex	1	1.081	0.000	0.302
x2	RC	-2	-2.763	0.183	0.129
x10	Ex	0.7	0.689	0.000	0.076
x3	Ex	-0.32	-0.317	0.000	0.034
x8	Ex	0	0.001	0.984	0.008

Note: Ordered by RF Importance. Ex=Exogenous variable, En=Endogenous variable, RC=Exogenous variable with random coefficient. Results are based on 50 markets and 40 brands.

This is what we did

1. Assume consumers' marginal utility of attributes (true values of β).
2. Construct a data set with assumed number of brands and markets.
3. Use BLP model to derive mean marginal utilities from product attributes.
4. Run RF regression of $\ln(s)$ on attributes p, x only (WITHOUT any instrumental variables) and derive the importance of each product attribute in predicting $\ln(s)$.
5. Compare the importance with initially assumed valuation; does higher marginal utility of an attribute mean higher importance in RF?
6. Repeat the process thousand times with different number of brands and markets.

This is what we found

1. If consumers do not get utility from an attribute, its importance predicted by RF is almost always the lowest. If consumers are sensitive to an attribute (absolute value of parameter is high), its importance is high.
2. There is a positive association (71%) between RF importance and absolute value of BLP estimates. Above 90% if attributes are exogenous only. The closer the absolute marginal utility of two attributes, the closer their importance.
3. The order or marginal utility may not coincide with the order of importance in case of endogenous and random coefficient attributes. Endogenous attributes appear to be more important than they actually are in consumers' utility. Attributes with random coefficients appear to be of random importance due to variation.

So what? Who cares?

1. Machine Learning model does not replace structural model.
2. But gives a sensible idea about the importance of product attributes, especially under time, calculation, and instrumental variable constraints.
3. Firms can focus on product attributes that are more important to consumers, thus saving valuable resources and receiving better appreciation of their products.
4. The costs of construction, computation, and communication of a model should be compared to the added-value the complexity offers to the researcher. Our approach offers a balanced solution to the trade-off between data dimension and model rigor. The model is particularly helpful when number of attributes are high but the data set is small.

Syed Badruddoza and Modhurima Dey Amin are Doctoral Students at Washington State University. Correspondence may be sent to s.badruddoza@wsu.edu
Citation: Badruddoza, S. and Amin, M.D. 2019. *Determining the Importance of an Attribute in a Demand System: Structural versus Machine Learning Approach*. Presented in AAEA Annual Meeting, July 2019. (Working Paper).

Works cited

- Berry, S., Levinsohn, J. and Pakes, A., 1995. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pp.841-890.
- Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.