22) Review: Machine Learning Methods

Vitor Kamada

April 2019

Reference

Tables, Graphics, and Figures from:

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). **Machine learning methods for demand estimation.** American Economic Review, 105(5), 481-85.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). **Demand estimation with machine learning and model combination.** No. w20955. National Bureau of Economic Research.

Demand Equation

Product j have demand in market m at time t :

$$\ln Q_{jhmt} = f(p_{mt}, a_{mt}, X_{mt}, D_{mt}, \epsilon_{jmt}; \theta)$$

h : nests of products allow the substitution patterns

α : advertising and promotional measures

$$\ln Q_{jhmt} = \alpha' p_{mt} + \beta_1' X_{mt} + \beta_2' D_{mt} + \gamma' a_{mt}$$

$$+ \lambda' \mathcal{I}(X_{mt}, D_{mt}, p_{mt}, a_{mt}) + \zeta_{hm} + \eta_{mt} + \epsilon_{jmt}$$

$\mathcal{I}$: Interactions Operator

IRI Marketing Research Data

Information Resources, Incorporated (IRI)

- Scanner panel data from grocery stores within one grocery store chain for six years
- Sales data on salty snacks based on UPC (Universal Product Code)

Sparse data: most of the elements are zeros

$$q_{jmt} = 0\text{: No sale or out-of-stock}$$

Bag of Words

Unstructured data: text description of a bag of chips and the image of the bag

x_{jw}: # of times word w appears in bag of chips j

- Unsupervised learning (Clustering)
- Add prediction power

Gentzkow & Shapiro (2010). **What drives media slant? evidence from U.S. daily newspapers**. Econometrica 78 (1).

Summary Statistics

Variable	Mean	Median
Price	2.12	1.99
Quantity	15.80	6.00
Dollars	28.11	12.19
#Stores	1,560	
#Weeks	313	
#UPC	3,337	
#Obs	3,045,513	

Variable	#Levels	Three Most Frequent Values		
Brand	237	Pringles	Utz	Lays
Product Type	4	Potato Chip	Potato Crisp	Potato Chip and Dip
Packaging	20	Bag	Canister	Plastic Wrapped Cardboard
Flavor	207	Original	Sour Cream & Onion	BBQ
Fat Content	16	Missing	Low Fat	Fat Free
Cooking Method	47	Missing	Kettle Cooked	Old Fashion Kettle Cooked
Salt Content	14	Missing	Lightly Salted	Sea Salt
Cutting Type	32	Flat	Missing	Ripple

Linear Regression

Log Quantity	Estimate	Std. Error	t value
Log Price	-0.639	0.055	-11.708
Promotion	0.466	0.039	11.926
Feature: None	-0.630	0.067	-9.334
Display:			
Minor	0.708	0.049	14.341
Major	0.637	0.049	13.119
Brand:			
Herrs	-0.351	0.156	-2.253
Jays	-1.101	0.244	-4.516
Kettle Chips	-0.995	0.236	-4.217
Lays	-0.337	0.159	-2.124
Lays Bistro Gourmet	-0.656	0.188	-3.480
Lays Natural	-1.662	0.327	-5.079
Lays Stax	-1.481	0.183	-8.104
Lays Wow	-0.485	0.204	-2.379
Michael Seasons	-1.655	0.239	-6.921
Pringles	-0.794	0.156	-5.090
Pringles Cheezums	-0.644	0.211	-3.055
Pringles Fat Free	-0.624	0.189	-3.308
Pringles Prints	-1.876	0.314	-5.982
Pringles Right Crisps	-0.881	0.128	-6.892
Ruffles Natural	-1.379	0.389	-3.549
Ruffles Snack Kit	-1.555	0.307	-5.061
Utz	-0.543	0.149	-3.635
Wise	-0.505	0.165	-3.062
Wise Ridgies	-0.984	0.167	-5.888
Volume	0.469	0.113	4.142
Package:			
Canister	0.437	0.091	4.800
Canister In Box	0.453	0.130	3.494
Flavor:			
BBQ	0.167	0.066	2.534
Cheddar	0.241	0.080	3.026
Cheese	-0.443	0.205	-2.164
Ketchup	-0.680	0.244	-2.787
Onion	0.339	0.066	5.107
Original	0.704	0.061	11.588
Spicy	-0.211	0.105	-2.005
Salt: No Salt	-0.446	0.212	-2.099
Type of Cut: Flat	0.308	0.070	4.411
Store Fixed Effects	Yes		
Week Fixed Effects	Yes		
Adjusted R-squared	0.884		

Logit with Regression Selection

Log Share	Estimate	Std. Error	t value	$\Pr(>\|t\|)$	
Log Price	0.296	0.113	2.624	0.009	**
Promotion	-0.441	5.192	-0.085	0.932	
Feature: None	0.263	0.151	1.745	0.081	.
Display:					
Minor	-0.215	0.104	-2.080	0.038	*
Major	-0.338	0.113	-3.000	0.003	**
Store Fixed Effects	No				
Week Fixed Effects	No				
AIC	6884.4				
Significance	0 ***	0.001 **	0.01 *	0.05 .	

Variance Inflation Factors: $VIF_j = \frac{1}{1-R^2_{-j}}$

	VIFs after Selection	
Variable	**Linear**	**LASSO**
Product Type - Potato Chip And Dip	$+\infty$	3.5084
Brand - Ruffles Snack Kit	$+\infty$	3.4729
Logprice	4.1750	3.2319
Volumn	3.9775	3.1541
Cooking - Missing	$+\infty$	3.1100
Cooking - Kettle	$+\infty$	2.6495
Package - Canister	$+\infty$	1.8047
Fat - Regular	76.6610	1.5930
Brand - Lays	104.5904	1.5187
Promotion	1.4806	1.4388
Feature - None	2.3398	1.3369
Brand - Kettle Chips	27.3608	1.3222
Flavor - Original	2.8610	1.2875
Brand - Ruffles	50.1427	1.2802
Salt - Regular	3.0660	1.2732
...		

Random Forest Variable Importance

Log Quantity	%Increase in Mean Squared Error	Increase in Node Purity
Log Price	74.83	1196.68
Volume	56.81	855.79
Display: Minor	49.79	455.98
Promotion	43.76	519.72
Display: Major	43.29	267.43
Feature: None	42.05	592.37
Brand: Lays	39.82	367.29
Brand: Ruffles	33.21	76.97
Brand: Wavy Lays	32.95	143.46
Flavor: Classic	32.31	219.00
Flavor: Sour Cream & Onion	30.26	62.28
R-Squared	0.42	

L_2 Boost Coefficients

Log Quantity	Coefficient
Log Price	-19.57
Promotion	18.24
Feature: Medium Ad	4.79
Feature: None	-19.85
Display: Minor	12.78
Display: Major	18.88
Brand: Kettle Chips	-3.41
Brand: Lance Thunder	-0.48
Brand: Lays	16.50
Brand: Lays Stax	-2.30
Brand: Ruffles	6.26
Brand: Wavy Lays	10.06
Flavor: Classic	11.30
Flavor: Sea Salt & Vinegar	-0.45
Type: Potato Chip and Dip	-0.49
Type: Potato Crisp	-1.10
Package: Canister in Box	-4.08
...	

Only 36 out of 2243 variables have non-zero coefficients

Bates & Granger (1969): Linear Model Combination

$$Y = \gamma_1 \hat{Y}_{ols} + \gamma_2 \hat{Y}_{\sqrt{Lasso}} + \gamma_3 \hat{Y}_{SVM} + \gamma_4 \hat{Y}_{Boosting} + \gamma_5 \hat{Y}_{Logit} + \epsilon$$

$$\Sigma \gamma_i = 1$$

	Train		Validation		Test		
	RMSE	Std. Err.	RMSE	Std. Err.	RMSE	Std. Err.	Weight
Linear	0.766	0.010	0.994	0.017	1.010	0.015	10.41%
Sqrt Lasso	0.977	0.007	0.984	0.013	0.995	0.009	1.71%
Support Vector Machine	0.543	0.007	0.889	0.018	0.900	0.012	87.57%
L2 Boosting	1.053	0.004	1.016	0.013	1.028	0.012	0.00%
Logit	3.282	0.170	3.509	0.340	3.915	0.263	0.32%
Linearly Combined			0.887		0.898		100.00%
# of Obs	1,827,308		456,827		761,379		
Total Obs	3,045,513						
% of Total	60%		15%		25%		

Model Comparison: Prediction Error

	Train		Validation		Test		
	RMSE	Std. Err.	RMSE	Std. Err.	RMSE	Std. Err.	Weight
Linear	0.766	0.010	0.994	0.017	1.010	0.015	17.73%
Stepwise	0.930	0.008	0.969	0.017	0.980	0.014	0.00%
Forward Stagewise	0.977	0.007	0.985	0.015	0.995	0.013	0.00%
Sqrt Lasso	0.977	0.007	0.984	0.013	0.995	0.009	0.00%
Random Forest	0.927	0.007	0.914	0.017	0.916	0.013	37.46%
Support Vector Machine	0.543	0.007	0.889	0.018	0.900	0.012	44.79%
L2 Boosting	1.053	0.004	1.016	0.013	1.028	0.012	0.00%
Logit	3.282	0.170	3.509	0.340	3.915	0.263	0.02%
Linearly Combined			0.879		0.887		100.00%
# of Obs	1,827,308		456,827		761,378		
Total Obs	3,045,513						
% of Total	60%		15%		25%		

Combining Models in Random Forest

RF: robust to missing values

	Train		Validation		Test		
	RMSE	Std. Err.	RMSE	Std. Err.	RMSE	Std. Err.	Var. Imp.
Linear	0.766	0.010	0.994	0.017	1.010	0.015	32.435
Stepwise	0.930	0.008	0.969	0.017	0.980	0.014	32.647
Forward Stagewise	0.977	0.007	0.985	0.015	0.995	0.013	24.607
Sqrt Lasso	0.977	0.007	0.984	0.013	0.995	0.009	23.135
Random Forest	0.927	0.007	0.914	0.017	0.916	0.013	34.845
Support Vector Machine	0.543	0.007	0.889	0.018	0.900	0.012	52.972
L2 Boosting	1.053	0.004	1.016	0.013	1.028	0.012	23.977
Logit	3.282	0.170	3.509	0.340	3.915	0.263	0.932
Combined by Random Forest			0.920		0.902		
# of Obs	1,827,308		456,827		761,378		
Total Obs	3,045,513						
% of Total	60%		15%		25%		

Top 20 Products vs Other Products

Training set: Top 20

	Top 20 Products		Other Products		
	RMSE	Std. Err.	RMSE	Std. Err.	Weight
Linear	0.397	0.034	2.037	0.037	35.37%
Stepwise	0.768	0.023	1.437	0.024	0.00%
Forward Stagewise	0.882	0.017	1.371	0.018	0.00%
Sqrt Lasso	0.935	0.015	1.374	0.017	0.00%
Random Forest	0.759	0.018	1.530	0.017	0.00%
Support Vector Machine	0.318	0.042	1.537	0.020	64.63%
L2 Boosting	0.920	0.021	1.378	0.019	0.00%
Logit	1.331	0.124	2.685	0.134	0.00%
Linearly Combined	0.277		1.433		100.00%
# of Obs	504,337		2,541,176		
Total Obs	3,045,513				
% of Total	16.56%		83.44%		

Promotion Variables

Variable	Value	Frequency(%)
Promotion	Price Reduction $< 5\%$	74.11
	Price Reduction $>5\%$	25.89
Feature	Large Ad	5.35
	Medium Ad	4.98
	Small Ad	0.33
	None	89.33
Display	None	81.16
	Minor	10.66
	Major	8.18
Total Obs	3,045,513	

Model Comparison: Promotional Lift

$$Mean = E(Y_{j,actual}|T = 1) - E(\hat{Y}_{j,predicted}|T = 0)$$

	Mean	t	95% Conf. Int.		Weight
Linear	9.646	8.171	7.332	11.960	23.37%
Stepwise	20.124	19.516	18.103	22.145	7.01%
Stagewise	22.458	21.018	20.363	24.552	0.00%
Sqrt Lasso	22.440	21.006	20.346	24.534	0.00%
Random Forest	18.276	17.705	16.253	20.299	68.00%
Support Vector Machine	25.920	23.428	23.752	28.089	0.00%
L2 Boost	22.995	21.386	20.887	25.102	0.00%
Logit	22.671	20.474	20.500	24.841	1.61%
Linear Combination	19.017	18.456	16.998	21.037	100.00%

FE model has the wrong sign on promotional lift