Machine Learning Methods for Demand Estimation

By Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang

Aydın Bayraktar / EC48W 3rd Assignment

ML METHODS & DEMAND ESTIMATION

Modeling consumer behavior with computer science and statistics

Large data sets

To improve business decisions

price, quantity, price discrimination, marketing strategies, relative prices of competitors

Commonly used by firms in the retail, health care, and internet industries

Table 2 Product Categories

Category	Dollars per 1000 HH ($)	Percent of HH's buying (%)	Purchase cycle, (days)	Perishability	Stockpilability	Percent of volume on any deal (%)	Average percent off price reduction (%)
Beer/ale/alcoholic cider	21,503	29.9	67	m	m	31.0	13.4
Carbonated beverages	76,567	91.9	40	l	m	58.2	23.6
Coffee	17,026	57.3	65	l	h	40.8	26.2
Cold cereal	46,555	87.2	48	m	m	43.4	30.7
Deodorant	6,020	53.4	94	l	h	35.5	28.0
Diapers	12,021	14.7	55	l	l	35.0	16.2
Facial tissue	8,611	59.9	70	l	h	38.9	25.1
Photography supplies	2,911	18.4	104	l	l	34.2	29.2
Frankfurters	9,896	65.8	82	h	l	46.9	31.1
Frozen dinners/entrees	51,552	80.3	51	l	l	40.7	25.7
Frozen pizza	19,087	63.4	64	l	l	50.2	26.3
Household cleaner	10,397	70.0	82	l	h	22.7	25.0
Mustard & ketchup	4,647	71.2	91	l	h	32.4	28.3
Mayonnaise	6,652	72.8	95	m	h	41.1	29.1
Laundry detergent	18,294	68.0	80	l	h	46.2	26.2
Margarine/spreads/butter blends	8,994	74.1	65	m	m	29.3	27.0
Milk	61,588	93.4	29	h	l	22.4	22.5
Paper towels	11,809	64.6	78	l	l	45.0	24.4
Peanut butter	6,311	61.0	82	m	h	32.9	25.3
Razors	1,258	9.2	87	l	h	34.0	20.6
Blades	4,448	28.5	106	l	h	20.3	21.5
Salty snacks	44,234	93.3	41	h	l	40.4	25.4
Shampoo	6,302	55.2	87	l	h	35.1	22.5
Soup	27,418	90.3	45	l	h	38.5	29.0
Spaghetti/Italian sauce	8,908	67.6	72	l	h	42.5	27.0
Sugar substitutes	2,731	21.7	82	l	h	14.4	23.2
Toilet tissue	24,189	75.3	67	l	l	45.4	23.9
Toothbrush	6,862	49.3	87	l	h	33.1	27.1
Toothpaste	7,997	62.8	89	l	h	40.1	25.8
Yogurt	23,556	71.8	50	h	l	34.7	24.3

Notes. Total U.S.—Grocery, drug, and mass excluding Wal-Mart. For 52 weeks, ending 6/25/2006. l = low, m = medium, h = high.
Source: IRI Builders Suite.

PROBLEM & DATA

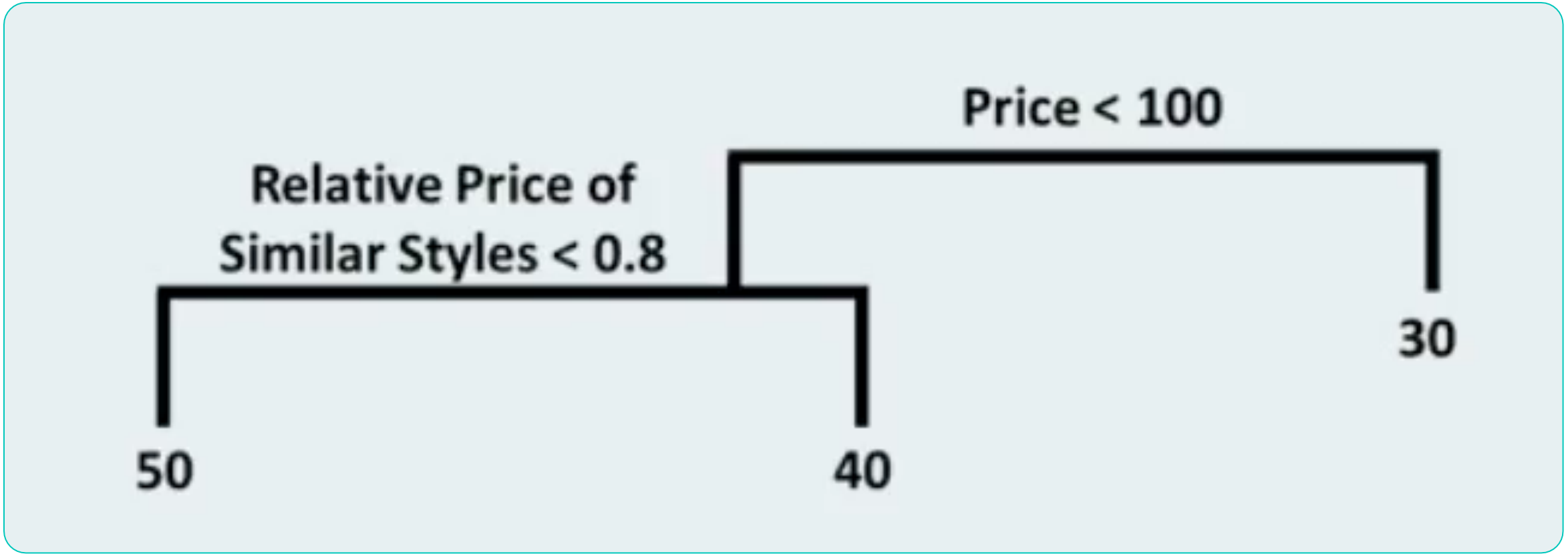
- A canonical demand estimation problem
- IRI Marketing Research (Bronnenberg, Kruger and Mela, 2008)
- Scanner panel data from grocery stores within one grocery store chain for six years.
- Number of observations: 837,460, which includes 3,149 unique products.

METHODS

Linear regression	LASSO	Stepwise regression	Bagging
Conditional logit	Forward stage wise regression	Support vector machines	Randomforest

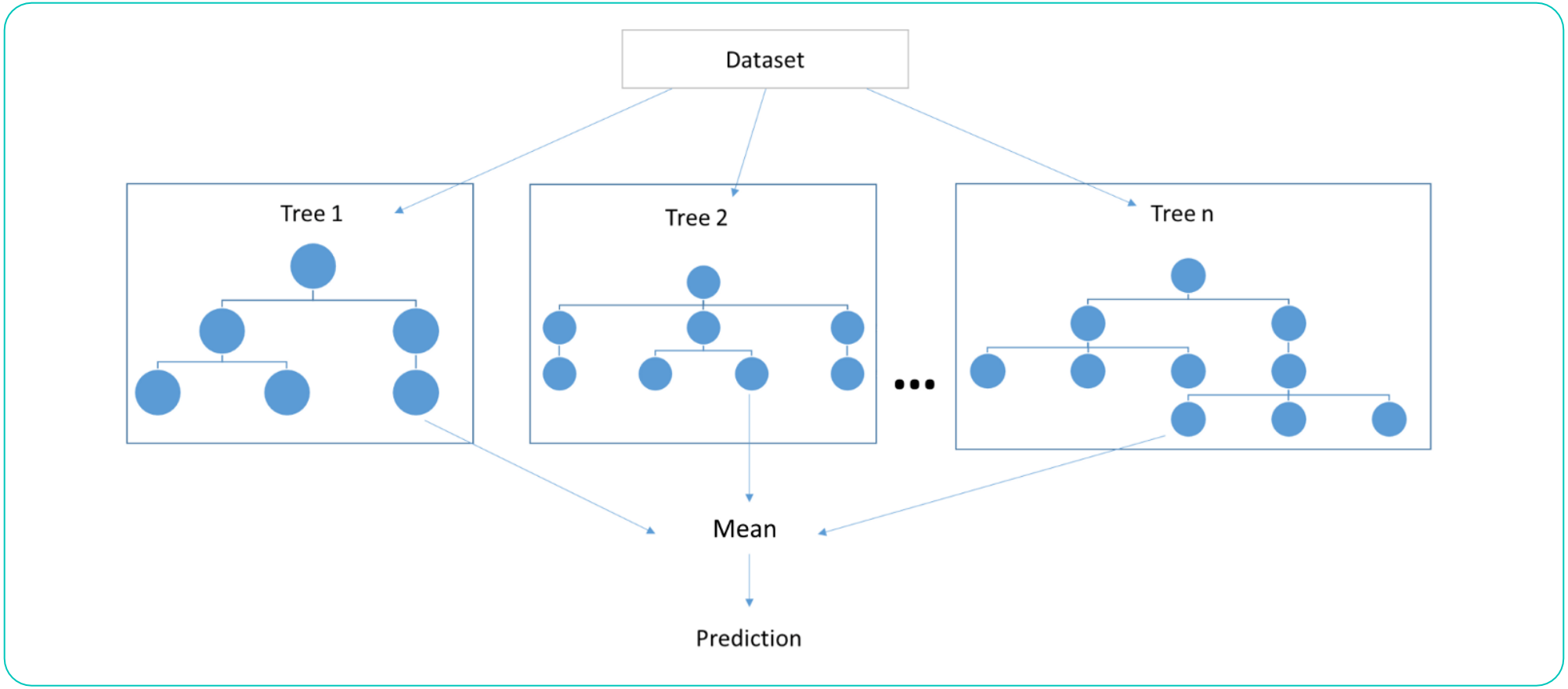
TABLE 1—MODEL COMPARISON: PREDICTION ERROR

	Validation		Out-of-Sample		Percent
	RMSE	Std. Err.	RMSE	Std. Err.	Weight
Linear	1.169	0.022	1.193	0.020	6.62
Stepwise	0.983	0.012	1.004	0.011	12.13
Forward Stagewise	0.988	0.013	1.003	0.012	0.00
LASSO	1.178	0.017	1.222	0.012	0.00
Random Forest	0.943	0.017	0.965	0.015	65.56
SVM	1.046	0.024	1.068	0.018	15.69
Bagging	1.355	0.030	1.321	0.025	0.00
Logit	1.190	0.020	1.234	0.018	0.00
Combined	0.924		0.946		100.00



REGRESSION TREES

- Use features to partition styles sold in past, and only use relevant styles of predict demand
- Allow for non-monotonic price/demand relationship



RANDOM FOREST

CONCLUSION

- ML methods can produce superior predictive accuracy as compared to a standard linear regression or logit model when estimating demand.