Interpretable machine learning for demand modeling with high dimensional data using Gradient Boosting Machines and Shapley values

Evgeny A. Antipov, Elena B. Pokryshevskaya

Affiliation (for both authors):
National Research University Higher School of Economics

Kantemirovskaya St. 3, Saint-Petersburg, Russia, 194100

Corresponding author:
Evgeny A. Antipov eantipov@hse.ru

Aknowledgements

The research was supported by the Russian Science Foundation (project № 18-71-00119)

Interpretable machine learning for demand modeling with high dimensional data using Gradient Boosting Machines and Shapley values

Evgeny A. Antipov, Elena B. Pokryshevskaya

Affiliation (for both authors):
National Research University Higher School of Economics

Kantemirovskaya St. 3, Saint-Petersburg, Russia, 194100

Corresponding author:
Evgeny A. Antipov eantipov@hse.ru

Evgeny A. Antipov is Associate Professor in the National Research University Higher School of Economics (Russia). He is also the President of *StatAdvice.com*, a statistical and business research consulting company. His research interests include empirical analysis of demand, consumer preferences and pricing.

Elena B. Pokryshevskaya is Associate Professor in the National Research University Higher School of Economics (Russia). Her research interests are concentrated in applied statistics and econometrics.

Keywords
sales forecasting; Shapley value; interpretable machine learning; Random Forest; Gradient Boosting Machines; Elastic net

Aknowledgements
The research was supported by the Russian Science Foundation (project № 18-71-00119)

Abstract

Forecasting demand and understanding sales drivers are one of the most important tasks in retail analytics. However, traditionally, linear models and/or models with a small number of predictors have been predominantly used in sales modeling. Taking into account that real-world demand is naturally determined by complex substitution and complementation patterns among a large number of interrelated SKUs, nonlinear effects of prices, promotions, seasonality, as well as many other factors, their lagged values and interactions, a realistic model has to be able to account for all that. We propose a conceptual model for sales modeling based on standard POS data available to any retailer and generate almost 500 potentially useful predictors of a focal SKU's sales accordingly. In our comparison of three classes of models, Gradient Boosting Machines outperformed Random Forests and Elastic nets. By using interpretable machine learning methods, we came up with actionable insights related to the importance of various groups of predictors from the conceptual model, as well as demonstrated how helpful it can be for marketing managers to decompose predictions into the effects of individual regressors by using an approximation of Shapley values for feature attribution.

1. Introduction

Quantifying price sensitivity of sales, as well as true effectiveness of promotions remain challenging problems due to the complexity of relationships that determine sales. For example, non-promoted products may be cannibalized by promoted ones, the effectiveness of a promotion is likely to depend on promotional activity in previous weeks, joint promotions of certain products may have a synergy effect on some other product's sales, etc. As a result, either none or only a portion of the sales lift generated by a promotion is incremental for retailers (Gedenk, 2018). More specifically, it has been shown that more than 50% of all promotions were not profitable for the retailer (Ailawadi et al., 2007). Even if a promotion is beneficial for a manufacturer, it is likely to cause a negative impact for a retailer because switching happens from non-promoted items to promoted ones that typically have lower margin. To manage promotions effectively, one needs to evaluate and quantify the effects that promotions have on sales, which is why a good sales response model is needed.

One of the classical models used in the literature on mathematical models in marketing for sales forecasting as well as price and promotion elasticity of demand estimation is the SCAN*PRO model and its variations that are estimated using both ordinary least squares and more advanced econometric methods. The model was proposed in 1988 (Wittink et al., 1988) and since then its variations have been discussed and compared in the literature (Andrews et al., 2008; Van Heerde et al., 2002). The SCAN*PRO model and its extensions decompose sales for a brand into own- and cross-brand effects of price, feature advertising, aisle displays, week effects, and store effects. Despite the fact that it accounts for the impact of the current week's prices and promotions of some of the competitors, being based on traditional regression methods, the model accounts for intertemporal (pre- and post-promotion dips) and cross-category effects only to a very limited extent given the large assortment of a typical store. Most other studies on demand modeling also used parsimonious models a'la SCAN*PRO with up to 20 features (Haupt et al., 2014). Even though the power of big data and machine learning algorithms has been demonstrated in some sales forecasting papers of the last decade, the number of features has typically been not overwhelming (Ali et al., 2009; Ferreira et al., 2015; Sun et al., 2008; Yang and

Zhang, 2018).

One of a few papers that has ever emphasized the importance of dealing with high dimensional data for demand forecasting was the work of Ma et al. (2016) where a 4-step approach was suggested for feature selection and the importance of accounting for cross-product and cross- period influences has been emphasized. They note that most studies do not account for inter- category, as well as lagged variables and fill this gap. Their approach was further combined with an optimization algorithm to come up with an optimization model for category multi-period profit maximization (Ma and Fildes, 2017). Their method involves a subjective stage of determining relevant product categories. Even though this step allows making the model computationally more efficient, it may still substantially restrict the patterns of substitution and complementation. Even though it is reasonable to expect substitution and complementation patterns to be stronger for SKUs from the same or somewhat related categories, we will allow more flexibility by assuming that any SKU's price and/or promotion features can impact sales of any other SKU if the SKUs are from at least somewhat related categories. For example, if a person purchases a lot of discounted ice-cream, they may be less likely to buy chocolate because they feel it would be too much non-healthy food for this week or because they have little money left after purchasing ice-cream. Even though this unrestrictive assumption may seem to be bad from the computational efficiency perspective, today's big data technologies allow making such assumptions and including a large number of predictors in our models. Taking into account that nowadays big data technologies already allow working with petabytes of retail data (Bradlow et al., 2017), we tend to worry less and less about the number of potential predictors involved. Therefore, algorithms that handle hundreds and even thousands of features effectively, giving each feature an opportunity to be used in modeling, such as Random Forests and Gradient Boosting Machines, seem to be more promising than methods that require getting rid of some features completely.

While subjectively limiting the number of potential predictors can often be helpful, a more important limitation of the approach proposed in Ma et al. (2016) is that they restrict the flexibility of the functional form by essentially using linear autoregressive distributed lags models. Therefore, despite all the modifications aimed at working with multidimensional data, Ma's and Fildes' basic econometric specification was still the traditional ADL model, the parsimony of which was achieved through a multi-stage feature selection procedure (Ma and Fildes, 2017).

Indeed, economists and marketing scientists have still preferred traditional econometric models allowing for an easy interpretation of parameter estimates. At the same time, it is reasonable to assume that the large assortment of supermarkets inevitably leads to increasingly complex interactions and nonlinearities in the data-generating process, which cannot be accounted for using parametric regression models when the number of relevant features is large. For example, due to stockpiling, for most products a promotion in, say, period t-2 will weaken the effect of a promotion in period t, but the effect of that lagged promotion is likely to be moderated by whether there were promotions and/or temporary price reductions in periods t-3 and t-1.Traditional linear models are not flexible enough to account for all such interactions and nonlinearities, especially when the number of features exceeds several dozens. Even though parametric regressions are still widely used, "big data tricks for econometrics" have been found to be applicable for many problems in economics and business (Bajari et al., 2015; Einav and Levin, 2014; Varian, 2014), but interpreting black box model still remained a challenge until the field of interpretable machine learning has started to emerge recently.

In this study we present a conceptual model and a machine learning approach to the problem of forecasting sales and interpreting the impact of several groups of features on sales. More specifically, besides demonstrating the superiority of flexible nonlinear tree-based models and conducting feature importance analysis, we propose how the problem of explaining a machine learning sales model's predictions can be fully overcome by introducing to the demand and sales analytics domain a recently developed unified approach to interpreting model predictions called SHAP (Shapley Additive Explanations). SHAP is a recent data science development, allowing to interpret any black box model as easily as a traditional regression model, which is the only possible consistent and locally accurate additive feature attribution method based on expectations (Lundberg and Lee, 2017). Due to a large number of features we use one of its approximate implementations (Molnar, 2018; Štrumbelj and Kononenko, 2014). We have not found any published applications of this approach to sales analytics yet, while it is one the fields where it is crucially important for analysts to decompose the sales bump and be able to explain why certain results were predicted and/or achieved for the sake of marketing planning and evaluation (Bohanec et al., 2017).

2. Data

We use a dataset of weekly sales made available by Dunnhumby − a retail analytics company − for academic purposes and known as "Breakfast at the Frat"[1]. It contains sales and promotion information on the top five products from each of the top three brands within four categories: mouthwash, pretzels, frozen pizza and boxed cereal, gathered from a sample of stores over 156 weeks. Despite being somewhat limited in terms of the number of SKUs and product categories included into it, this dataset is still unique because it contains real-world, high-quality and publicly available data from a leading provider of retail data. We will concentrate on analyzing this dataset from a boxed cereal manufacturer's perspective. More specifically, we will assume that our aim is to do pricing and promotional analytics for GM HONEY NUT CHEERIOS 12.25 OZ, a popular SKU from the category of cold cereals and the subcategory of all family cereals.

According to the analysis of promotional interactions at category level conducted by Ma et al. (2016), cold cereal purchases are influenced by salty snacks and frozen pizza promotions, which allows accounting for realistic patterns of complementation and substitution using the available dataset. However, the contribution of inter-category effects is not expected to be very high: Ma et al. (2016) reported that of the improvement achieved (compared to the model with only focal product's own price and promotion predictors) 95 percent comes from the intra-category information, and only 5 percent from the inter-category information. The list of 13 UPCs used in our analysis is presented in Table 1.

Table 1. List of UPCs used in the analysis

UPC	DESCRIPTION	MANUFACTURER	CATEGORY	SUB_CATEGORY	PRODUCT_SIZE	UPC_SHORT
1111085345	PL RAISIN BRAN	PRIVATE LABEL	COLD CEREAL	ADULT CEREAL	20 OZ	AC1
1111085350	PL BT SZ FRSTD SHRD WHT	PRIVATE LABEL	COLD CEREAL	ALL FAMILY CEREAL	18 OZ	AFC2
1600027527	GM HONEY NUT CHEERIOS	GENERAL MI	COLD CEREAL	ALL FAMILY CEREAL	12.25 OZ	AFC3
1600027528	GM CHEERIOS	GENERAL MI	COLD CEREAL	ALL FAMILY CEREAL	18 OZ	AFC4

[1] https://www.dunnhumby.com/careers/engineering/sourcefiles

1600027564	GM CHEERIOS	GENERAL MI	COLD CEREAL	ALL FAMILY CEREAL	12 OZ	AFC5
3800031829	KELL BITE SIZE MINI WHEAT	KELLOGG	COLD CEREAL	ALL FAMILY CEREAL	18 OZ	AFC7
3800031838	KELL FROSTED FLAKES	KELLOGG	COLD CEREAL	KIDS CEREAL	15 OZ	KC3
7192100337	DIGRN SUPREME PIZZA	TOMBSTONE	FROZEN PIZZA	PIZZA/PREMIUM	32.7 OZ	PI5
7192100339	DIGRN PEPP PIZZA	TOMBSTONE	FROZEN PIZZA	PIZZA/PREMIUM	28.3 OZ	PI6
1111009477	PL MINI TWIST PRETZELS	PRIVATE LABEL	BAG SNACKS	PRETZELS	15 OZ	PR1
1111009497	PL PRETZEL STICKS	PRIVATE LABEL	BAG SNACKS	PRETZELS	15 OZ	PR2
1111009507	PL TWIST PRETZELS	PRIVATE LABEL	BAG SNACKS	PRETZELS	15 OZ	PR3
2840004768	RLDGLD TINY TWISTS PRTZL	FRITO LAY	BAG SNACKS	PRETZELS	16 OZ	PR5

Ailawadi et al. (2006) found out that cross-store variation accounts only for about 2% of total variation in net unit impact of a sales promotion for a retailer, making it reasonable to assume that pooled POS data is sufficient to provide reliable results. However, we pooled data from 12 stores while including a set of corresponding dummy indicators that allow accounting for individual store differences. Our dataset was converted to "wide" format so that each row of our dataset is identified by WEEK_END_DATE and STORE_NUM.

Predictors were generated from the original dataset based on the proposed conceptual model for a focal SKU's sales forecasting (Figure 1), which assumes the accessibility of only standard POS data commonly available for most retailers when enriching such data with external data (on advertising shocks, news and other influential factors) is not feasible. Promotional variables were assumed known to the retailer at *t+6* in our model, as they usually form part of a promotional plan agreed with suppliers. The lag length was chosen so as to account for a sufficient number of past periods, while not decreasing the number of degrees of freedom too much. A detailed justification of the inclusion of 6 lags to a related model was given by Van Heerde et al. (2004) with references to results of other researchers.

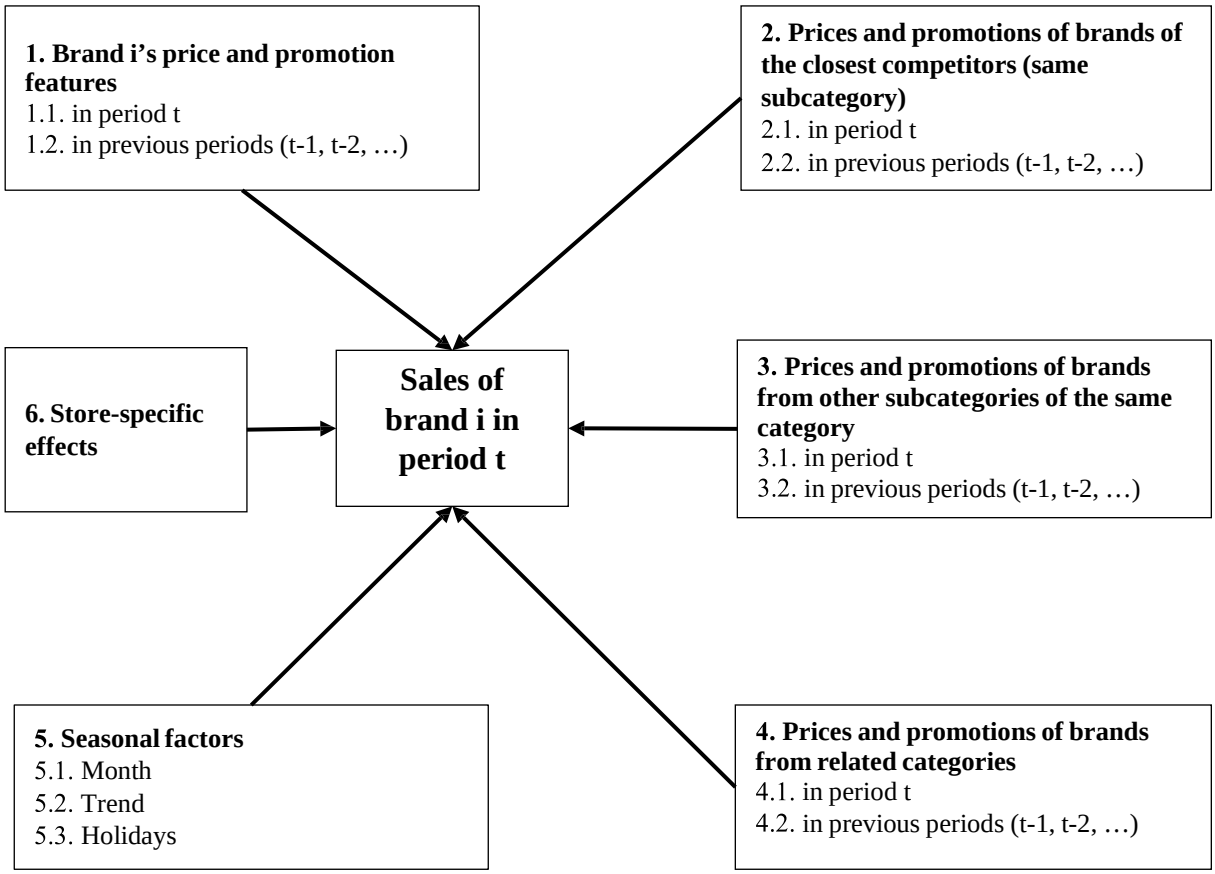

Figure 1. Conceptual model for SKU-level sales forecasting using POS data

The list of predictors contains 499 variables. Variable names for groups 1.1 – 4.2 of the conceptual model are constructed as follow: PREFIX.[LAG (IF ANY)].[UPC_SHORT]. For example, PRICE.AFC3 stands for the price of product AFC3 in the current week, while DISPLAY.2.AFC3 is whether there was a display promotion for product AFC3 two weeks ago. The following 7 prefixes were used: UNITS stands for unit sales, PRICE is the actual price charged, FEATURE stands for feature promotion (1 – used, 0 – not used), DISPLAY – display promotion (1 – used, 0 – not used), TRP_ONLY – temporary price reduction only (1 – yes, 0 – no), DISCOUNT is the size of discount in %. Seasonal variables include 12 month dummies (month_1 − month 12), trend (week_num) and 8 holiday dummies (NewYearsDay, MemorialDay, IndependenceDay, LaborDay, ThanksgivingDay, ChristmasDay, GoodFriday, Halloween), indicating whether the corresponding holidays were on a given week or not. Individual effects of stores were accounted for using 23 dummy variables.

3. Methods

Three classes of models that are capable of handling a large number of features without requiring any subjective preliminary dimensionality reduction procedures were used in our benchmarking: Elastic net models, Random Forests (RF) models and Gradient Boosting Machines (GBM). In this benchmarking Elastic Net models are in the spirit of traditional linear functional forms used in sales forecasting literature previously while GBM and RF are nonlinear techniques which have been used for retail sales modeling in the presence of almost 500 features for the first time. All algorithms were implemented in R language

using H2O – a fast Java-based framework for machine learning. A short description of each method is presented below.

- Gradient Boosting Machines (Friedman, 2001) is a forward learning ensemble method. The guiding heuristic is that good predictive results can be obtained through increasingly refined approximations. H2O's GBM sequentially builds regression trees on all the features of the dataset in a fully distributed way - each tree is built in parallel.
- A Random Forest is an ensemble of regression trees, built using the CART algorithm (Breiman, 1984). All the trees of the ensemble are built independently according to the following algorithm (Breiman, 2001). Let N be the size of the learning sample and M – the total number of predictors. A subset of m<M randomly chosen predictors is used to grow each tree on a bootstrap sample of the training data. The size of the bootstrap sample is n<N. For each of the bootstrap samples, an unpruned regression tree is grown, with the following modification: at each node, rather than choosing the best split among all predictors, the best split among a random sample of m regressors is made. After a large number of trees are generated, predictions are averaged over the different trees.
- Elastic Net is a regularized generalized linear modeling (GLM) algorithm used to attempt to solve problems with overfitting that can occur in traditional GLM (Friedman et al., 2010). Penalties can be introduced to the model building process to avoid overfitting, to reduce variance of the prediction error, and to handle correlated predictors. The two most common penalized models are ridge regression and LASSO (least absolute shrinkage and selection operator). The elastic net combines both penalties using both the alpha and lambda options (i.e., values are allowed to be greater than 0 for both).

Random discrete grid search was used to find the best set of hyperparameters for each algorithm according to Table 2. Mean Absolute Error (MAE), a measure robust to outliers and often used as a relevant performance measure for retail sales forecasting (Ali et al., 2009; Ma et al., 2016), was used as the stopping metric with stopping tolerance of 0.001 (i.e. a 0.1% improvement is considered to be substantial to continue the search process), 15 stopping rounds (the search stops if no improvement is achieved in 15 models in a row) and a maximum runtime of 7200 seconds. The dataset was split as follows: weeks 7-120 – training sample, 121-138 – validation sample (used for early stopping of the random grid search and ranking models), 139-156 – leaderboard (holdout) sample. Before assessing each model's performance on the holdout sample they were re-estimated using all weeks from 7 to 138.

Table 2. List of hyperparameters for model tuning and random grid search results

Algorithm	Hyperparameter	Range of hyperparameter values for the grid search	Value for the best model	MAE of the best model on the validation sample	MAE of the best model on the hold- out sample after re-estimating using the rest of the data
Elastic net	alpha: regularization distribution between L1 and L2	from 0 to 1, by 0.05	1	33.7	44.9
	Lambda: regularization strength	from 0 to 10, by 0.1	3.4		
GBM	max_depth: depth of each tree	from 1 to 5, by 1	5	21.5	38.4
	min_rows: fewest observations allowed in a terminal node	1, 5 and 10	1		
	learn_rate: rate to descend the loss function gradient	0.01, 0.05 and 0.1	0.05		
	learn_rate_annealing: allows you to have high initial learn_rate, then gradually reduce as trees are added (speeds up training)	0.99 and 1	0.99		
	sample_rate: row sample rate per tree	0.5, 0.75, 1	0.5		
	col_sample_rate: column sample rate per tree	0.8, 0.9 and 1	0.8		
RF	ntrees: number of trees	300, 500 and 700	500	20.5	43.7
	mtries: number of columns to randomly select at each level	From 50 to 400, by 10	380		

The following methods were used to explain the best model:

- **Variable Importance Analysis.** When calculating variable importances, the difference in the squared error before and after the split using a particular variable is considered the improvement (Rifkin and Klautau, 2004). Each feature's improvement is then summed up at the end to get its total feature importance (and then scaled between 0-1). Then two measures of importance were calculated:

relative importance is the importance of the i^{th} variable relative to the predictor with a maximum decrease in the squared error and the percentage contribution is the i^{th} variable importance relative to the sum of the changes in the squared error for all predictors.

- **Shapley additive explanations of model predictions.** Feature contributions for individual predictions were computed using the Shapley value approach from cooperative game theory. According to the approach, the feature values cooperate to achieve the prediction for a particular instance. In order to explain individual predictions we used a method from coalitional game theory that produces Shapley values (Lundberg and Lee, 2017), one of the key concepts of modern coalitional game theory. The Shapley value (ϕ) represents the contribution of each respective variable towards the predicted value compared to the average prediction for the data set. The idea behind Shapley values is to assess every combination of predictors to determine each predictor's impact. Focusing on feature x_i, the approach will test the accuracy of every combination of features not including x_i and then test how adding x_i to each combination improves the accuracy. As a result, the Shapley value fairly distributes the difference of the instance's prediction and the dataset's average prediction among the features. The following approximate Shapley estimation algorithm implemented in R's package iml was used (Molnar, 2018):

ob = single observation of interest
1: for variables j in {1,...,p} do
| m = random sample from data set
| t = data frame obtained by adding ob to m (horizontally)
| f(all) = compute predictions for t
| f(!j) = compute predictions for t with feature j values randomized
| diff = sum(f(all) - f(!j))
| phi = mean(diff)
end
2. sort phi in decreasing order

4. Results

4.1. Model comparison

Random grid search settings ensured a computer-intensive, but also a very comprehensive search of optimal hyperparameters: 2121 Elastic nets, 47 GBM and 37 RF models were estimated until the improvement of MAE fell under the specified tolerance. The parameters and mean absolute errors of the best model in each class are presented in the last three columns of Table 2. The rankings of models on the validation and the holdout samples were consistent: tree-based ensembling techniques (RF and GBM) outperformed the regularized linear regression approach, yet with a smaller variability of MAE among algorithms in the hold-out sample compared to the validation sample. Our experiments with creating stacked ensembles of the best RF base learners did not lead to any improvement on the hold-out sample compared to the best base learner (RF). The increase of the error in the hold-out sample compared to the validation sample is likely to be a result of the higher variance of sales in the hold-out sample (SD=106.8) compared to the validation sample (SD=65.5), whereas in the training sample the heterogeneity of sales was similar to that of the holdout sample (SD=106.9). This is not surprising due to the time-ordered nature of sales data for which unequal heterogeneity is common. It is worth mentioning that, even though our study is not directly comparable to Ma et al. (2016) MAE we achieved compares favorably to that found for the

cold cereals in the above-mentioned study category if we normalize it relative to the mean number of units sold (95.0, 88.9 and 121.0 for the training, validation and the hold-out samples, respectively). It is especially inspiring since unlike Ma et al. (2016) we do not use the focal product's first lag of sales as a predictor, because it would prevent us from planning for a few weeks ahead. We find this ability to be somewhat more actionable for pricing and promotional planning than ultra-short-term one-step ahead forecasting.

It is worth mentioning that optimal hyperparameters of the GBM model turned out to be robust to changes in the size of the training and validation samples. We omit the results of this sensitivity check for brevity. Before proceeding to model interpretation (section 4.2), the best model was applied to the whole dataset to use all the available information (Molnar, 2018), like we would do before making forecasts for new, unseen, data.

4.2 Variable importance analysis

We obtained individual importance scores for each variable and added up percentage contributions so as to evaluate the contribution of each group of predictors (Figure 2). It is worth mentioning that groups of variables contain different number of features, making it easier for groups with a greater number of features to be considered important. That is why we find it useful to complement the analysis by reporting the position of each group's most important variable in the feature importance ranking.

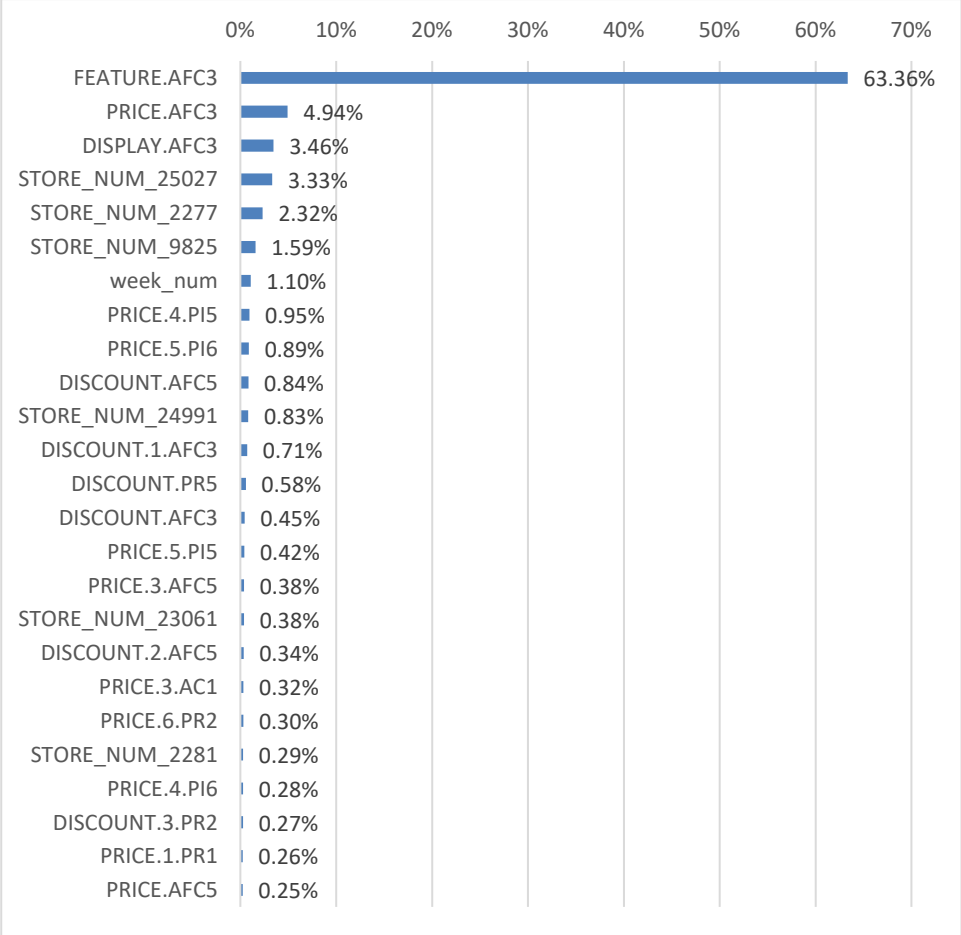

Figure 2. Top-25 features by importance
(% contribution to decrease of the sum of squared residuals)

While each individual feature's contribution may seem to be negligibly small (with a few exceptions), this is mainly because of the large number of features involved in the analysis. The analysis of aggregate percentage contributions of groups of predictors (Table 3) revealed that prices and promotions of a focal brand in week t are by far the most predictive (72.2%), while the corresponding lagged values explain only a small portion of the focal SKU's sales (1.7%). Some groups of features from our conceptual model contribute less than 1% to the explanatory power of the model, but it is important to keep in mind that these contributions may vary from one product category to another and, in addition, conclusions can be made at a different level of aggregation. For instance, various information about prices and promotions in related subcategories and categories (groups 2.1-4.2) contribute 15.7% to the improvement of predictive accuracy, while store-specific effects are accountable for 9.3% of the improvement. Therefore, the importance of competitive and store-specific effects is actually rather high, especially taking into account that our estimate is likely to be close to the lower bound of the true contribution of competitive information, because data on only a limited set of potential competitors was available to us. The total contribution of all lagged variables was 14.5%, which is why the inclusion of various lagged effects can generally be recommended. The analysis has shown that, among other things, prices of some related products 3-4 weeks ago and current discounts on them are the most influential in corresponding groups of features (2.1-4.2), while ChristmasDay is the most influential holiday.

Table 3. Aggregate percentage contributions for groups of predictors

Group of features	% contribution	Minimum rank	Maximum rank	The group's most important feature
1.1. Brand i's price and promotion features in period t	72.21	1.00	310.00	FEATURE.AFC3
1.2. Brand i's price and promotion features in previous periods (t-1, t-2, …)	1.68	12.00	322.00	DISCOUNT.1.AFC3
2.1. Prices and promotions of brands of the closest competitors (same subcategory) in period t	1.30	10.00	354.00	DISCOUNT.AFC5
2.2. Prices and promotions of brands of the closest competitors (same subcategory) in previous periods (t-1, t-2, …)	3.65	16.00	372.00	PRICE.3.AFC5
3.1. Prices and promotions of brands from other subcategories of the same category in period t	0.23	33.00	374.00	DISCOUNT.AC1
3.2. Prices and promotions of brands from other subcategories of the same category in previous periods (t-1, t-2, …)	1.73	19.00	388.00	PRICE.3.AC1
4.1.Prices and promotions of brands from related categories in period t	1.36	13.00	458.00	DISCOUNT.PR5
4.2. Prices and promotions of brands from related categories in previous periods (t-1, t-2, …)	7.42	8.00	475.00	PRICE.4.PI5
5.1. Seasonal factors: Month	0.02	181.00	492.00	month_3
5.2. Seasonal factors: Trend	1.10	7.00	7.00	week_num
5.3. Seasonal Factors: Holidays	0.03	148.00	499.00	ChristmasDay
6. Individual store effects	9.29	4.00	483.00	STORE_NUM_25027

4.3. Shapley additive explanations of model predictions

This subsection illustrates how the model came up with the highest and the lowest predicted values of sales by considering the corresponding cases and computing approximate Shapley values to come up with a local explanation and shed light on factors leading to high (Table 4) and low (Table 5) sales according to the estimated model. The highest prediction was observed in a situation

when there was a feature and a display promotion in a popular store (STORE_NUM=25027), the price was almost 50% lower than the mean price for our focal product. Unusually low prices for product AFC5 of the same brand from the same subcategory on the current week as well as one and two weeks ago also contributed positively to sales of the focal SKU in the current week. A possible reason is that this created some sort of income effect – buyers of AFC5 had more money left at their disposal and decided to spend them on a similar product (AFC3). Even though not all effects are easy to explain, the largest of them are insightful, while less interpretable effects are usually small and are associated with control variables.

Table 4. Shapley value decomposition. Case 1:
Highest predicted sales (predicted value=892.69, average prediction=92.22, top-15 features)

Feature	Phi	Feature value	Mean feature value
FEATURE.AFC3	287.41	1	0.11
STORE_NUM_25027	88.80	1	0.04
DISCOUNT.AFC5	52.11	-47.34	-10.92
PRICE.AFC3	40.85	1.64	2.81
DISCOUNT.1.AFC3	38.94	-49.84	-4.84
DISPLAY.AFC3	33.92	1	0.09
DISCOUNT.2.AFC5	22.48	-64.58	-11.62
week_num	20.89	7	81.50
DISCOUNT.PR5	18.43	-13.84	-2.00
DISCOUNT.6.PR2	14.11	-8.00	-1.80
DISCOUNT.1.AFC5	12.87	-68.34	-11.29
PRICE.1.AFC5	11.73	1.01	2.73
DISCOUNT.1.PR1	8.40	-1.34	-1.78
TPR_ONLY.4.PR5	8.37	1	0.04
PRICE.4.AFC4	7.84	4.54	4.34

Table 5 presents top-15 factors that contributed to the lowest predicted sales. Among these most important features, only the top-4 contributed more than two units to the difference between the predicted value and the average prediction. Explanations suggested by the algorithm include no feature promotion for the focal product, the fact that the store is neither store 24991, nor store 9825, but store 23061 and possible stockpiling created by a large discount on a similar product (AFC4) 4 weeks ago.

Table 5. Shapley value decomposition. Case 2:
Lowest predicted sales (predicted value=47.48, average prediction=92.22, top-15 features)

Feature	Phi	Feature value	Mean feature value
FEATURE.AFC3	-20.40	0	0.11
STORE_NUM_23061	-6.17	1	0.04
PRICE.4.AFC4	-2.83	2.72	4.34
STORE_NUM_24991	-2.21	0	0.04
FEATURE.4.AFC4	-1.71	1	0.08
PRICE.AFC3	-1.27	2.51	2.81
DISCOUNT.4.AFC4	-1.13	-40.74	-4.06
DISCOUNT.PI5	-0.93	-36.05	-9.13
STORE_NUM_9825	-0.74	0	0.04
PRICE.5.AFC2	-0.72	2.35	2.14
week_num	-0.66	62	81.50
DISCOUNT.5.PR2	-0.53	1.36	-1.783
DISCOUNT.AFC3	-0.52	-12.24	-4.73
PRICE.PR1	-0.48	1.46	1.38
PRICE.6.AFC2	-0.44	2.35	2.14

5. Conclusion

The fact that Gradient Boosting Machines outperformed other algorithms with another tree-based technique (Random Forests) being a close second emphasizes the importance of accounting for nonlinearities and interactions and proves that such tree-based ensembles are very competitive universal out-of-box tools which should be used in sales response modeling more widely. This result agrees with Ali et al. (2009) who came up with a conclusion that for periods with promotions, regression trees improve accuracy substantially compared to exponential smoothing and linear regression methods. The limited applicability of regularized GLM approaches that tend to assign zero weights to many variables agrees with the claim given in Ma et al. (2016) that such methods may select some unimportant predictors which are highly correlated with the important predictors but fail to select the truly important predictors.

We have demonstrated the usefulness of model-agnostic interpretable machine learning techniques for sales response modeling, where interpretable, yet less flexible, traditional econometric models have predominantly been used. The estimation of variable importance allows us to recommend accounting for competition effects as they have been shown to contribute more than 15% to the overall increase in the sum of squared residuals. There is some evidence suggesting that categories related to the focal product's category (such as frozen pizza/salty snacks in the case of all-family cold cereals) are not less important than related subcategories within the same category (such as kid's cereals in the case of all-family cold cereals). Shapley additive explanation of individual predictions turned out to be very insightful and allowed uncovering the effects of prices and promotions, as well as some non-trivial ideas related to stockpiling and cross-product effects.

One of the main problems of sales forecasting is that due to the observational nature of data we have to assume that the demand is always met and, therefore, sales reflect the true demand. Even though this assumption is rather realistic taking into account that in our study we are considering products with a long shelf life, the quantity demanded can be larger than the number of units sold, especially in the case of perishable products (Ozhegov and Teterina, 2018). Therefore, we recommend that retailers account for whether sales reflect censored demand or not whenever possible (at least, by keeping track of whether and how long the SKU was out of stock for some time during each week).

One more limitation of our study is that approximate Shapley values do not exactly add up to the difference between the prediction and mean prediction (however in the two cases we considered this efficiency property was almost perfectly met), while exact Shapley Values are rarely computationally feasible when the number of features is large. Even though it is often sufficient for applied analysis, a potential alternative is to reduce the number of features after variable importance analysis, rerun the model using only the main 50-80 features and estimate exact Shapley Values afterwards, i.e. to conduct exact Shapley value analysis based on a model built using a set of the most important features. Additional feature selection may be useful to account for the non-additivity in explaining machine-learning models: when using complex models it is important to consider not only features independently but also in sets of potential interactions, which is fully feasible only when the number of predictors is small.

6. References

1. Ailawadi, K.L., Harlam, B.A., Cesar, J., Trounce, D., 2006. Promotion profitability for a retailer: the

role of promotion, brand, category, and store characteristics. J. Mark. Res. 43, 518–535.
2. Ailawadi, K.L., Harlam, B.A., César, J., Trounce, D., 2007. Practice prize Report— Quantifying and improving promotion effectiveness at CVS. Mark. Sci. 26, 566–575.
3. Ali, Ö.G., Sayin, S., Van Woensel, T., Fransoo, J., 2009. SKU demand forecasting in the presence of promotions. Expert Syst. Appl. 36, 12340–12348.
4. Andrews, R.L., Currim, I.S., Leeflang, P., Lim, J., 2008. Estimating the SCAN∗ PRO model of store sales: HB, FM or just OLS? Int. J. Res. Mark. 25, 22–33.
5. Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M., 2015. Machine learning methods for demand estimation. Am. Econ. Rev. 105, 481–485.
6. Bohanec, M., Borštnar, M.K., Robnik-Šikonja, M., 2017. Explaining machine learning models in sales predictions. Expert Syst. Appl. 71, 416–428.
7. Bradlow, E.T., Gangwar, M., Kopalle, P., Voleti, S., 2017. The role of big data and predictive analytics in retailing. J. Retail. 93, 79–95.
8. Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
9. Breiman, L., 1984. Classification and regression trees. Chapman & Hall/CRC.
10. Einav, L., Levin, J., 2014. Economics in the age of big data. Science (80-.). 346, 1243089.
11. Ferreira, K.J., Lee, B.H.A., Simchi-Levi, D., 2015. Analytics for an online retailer: Demand forecasting and price optimization. Manuf. Serv. Oper. Manag. 18, 69–88.
12. Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1.
13. Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
14. Gedenk, K., 2018. Retailer promotions, in: Handbook of Research on Retailing. EdwardElgar Publishing.
15. Haupt, H., Kagerer, K., Steiner, W.J., 2014. Smooth Quantile-Based Modeling Of Brand Sales, Price And Promotional Effects From Retail Scanner Panels. J. Appl. Econom. 29, 1007–1028.
16. Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems. pp. 4765–4774.
17. Ma, S., Fildes, R., 2017. A retail store SKU promotions optimization model for category multi-period profit maximization. Eur. J. Oper. Res. 260, 680–692.
18. Ma, S., Fildes, R., Huang, T., 2016. Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. Eur. J. Oper. Res. 249, 245–257.
19. Molnar, C., 2018. Interpretable machine learning: A guide for making black box models explainable. Leanpub.
20. Ozhegov, E., Teterina, D., 2018. The Ensemble Method For Censored Demand Prediction. High. Sch. Econ. Res. Pap. No. WP BRP 200.
21. Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Mach. Learn. Res. 5, 101–141.
22. Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41, 647–665.
23. Sun, Z.-L., Choi, T.-M., Au, K.-F., Yu, Y., 2008. Sales forecasting using extreme learning machine with applications in fashion retailing. Decis. Support Syst. 46, 411–419.
24. Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R., 2004. Decomposing the sales promotion bump with store data. Mark. Sci. 23, 317–334.
25. Van Heerde, H.J., Leeflang, P.S.H., Wittink, D.R., 2002. How promotions work: SCAN* PRO-based evolutionary model building. Schmalenbach Bus. Rev. 54, 198–220.
26. Varian, H.R., 2014. Big data: New tricks for econometrics. J. Econ. Perspect. 28, 3–27.
27. Wittink, D.R., Addona, M.J., Hawkes, W.J., Porter, J.C., 1988. SCAN*PRO: The estimation, validation and use of promotional effects based on scanner data. Intern. Pap. Cornell Univ.
28. Yang, D., Zhang, A.N., 2018. Forecast UPC-Level FMCG Demand, Part IV: Statistical Ensemble, in: 2018 IEEE International Conference on Big Data (Big Data). pp. 3180–3185