MIS Quarterly

RESEARCH ARTICLE

USING FORUM AND SEARCH DATA FOR SALES PREDICTION OF HIGH-INVOLVEMENT PROJECTS

Tomer Geva and Gal Oestreicher-Singer
The Coller School of Management, Tel Aviv University, P.O. Box 39040,
Tel Aviv 6997801, ISRAEL
{tgeva@tau.ac.il} {galos@tau.ac.il}

Niv Efron and Yair Shimshoni
Google, Tel Aviv, ISRAEL
{niv@google.com} {shimsh@google.com}

Appendix A

List of Main Robustness Checks

Robustness Check	Appendix
Predictive capacity using NN algorithm	C
Predictive capacity using Expanding Window approach	D
Predictive capacity according to "premium" and "value" car brand characteristics based on price, perceived quality, and willingness to recommend metrics. Using one or two lags of data.	E
Car model-level Analysis	G
Predictive capacity using extended keyword selection	H
Predictive capacity using MSE criteria	H

Appendix B

List of Brands and Grouping by Price, Perceived Quality, and Willingness to Recommend

Table B1 provides details about the brands included in this study. The list of brands includes all car brands with average U.S. sales exceeding 5,000 cars per month during the period 2007–2010 (source: Automotive News).[1] Quality and willingness to recommend rankings are based on the YouGov BrandIndex product. (See additional details in Appendix F.)

[1]Combined search volume of multiple keywords can be obtained from Google Trends by utilizing the "+" sign between different terms.

Table B1. List of Brands and Grouping by Price, Perceived Quality, and Willingness to Recommend

Brand	Keyword(s)	Price	Average Quality Ranking	Average Recommend Ranking
Acura	acura	high price	high	high
Audi	audi	high price	high	low
BMW	bmw	high price	high	high
Buick	buick	high price	low	low
Cadillac	cadillac	high price	high	low
Chevrolet	chevrolet, chevy	low price	low	high
Chrysler	chrysler	low price	low	low
Dodge	dodge	low price	low	low
Ford	for	low price	low	high
GMC	gmc	low price	low	low
Honda	honda	low price	high	high
Hyundai	hyundai	low price	low	low
Infiniti	infiniti	high price	high	low
Jeep	jeep	low price	low	low
Kia	kia	low price	low	low
Lexus	lexus	high price	high	high
Lincoln	lincoln	high price	low	low
Mazda	mazda	low price	low	low
Mercedes Benz	mercedes	high price	high	high
Nissan	nissan	low price	high	high
Subaru	subaru	low price	low	high
Toyota	toyota	low price	high	high
Volkswagen	volkswagen	low price	high	high

Appendix C

Analysis Using the NN Algorithm

In addition to evaluating the performance of models based on the linear regression algorithm (LR), for robustness we repeat the analysis using the back-propagation neural network (NN) algorithm. This is a nonlinear method that is estimated by the backprop algorithm (Werbos 1974). One of the strongest properties of the NN algorithm is that it inherently accounts for nonlinear relationships and complex interactions between variables (Bishop 1995). (See Appendix I for more details about the NN algorithm.) Such a method may have the capacity to capture (potentially complex, or unexpected) interactions and relations that underlie the real-life data generating process, without the need for the researcher to formally specify (or even be aware of) all existing relations. Thus, the NN approach can potentially "extract" more predictive power out of the data compared with linear methods, as it is not constrained by linearity and prespecified interactions. To implement the NN algorithm we used the "nnet" package in R software. This implementation involves one layer of hidden nodes, and the minimization of a sum-of-square-errors criterion. In specifying the network architecture, one must choose the number of nodes in the hidden layer. While the literature does not offer clear rules about the optimal complexity of the network in terms of hidden nodes, it proposes general guidelines (Zhang et al. 1998); for instance, the number of hidden nodes should be proportional to the number of inputs. Following this guideline, we employed an NN model architecture in which the number of hidden nodes was equal to the number of inputs multiplied by 0.5.[2] Finally, while NNs are well-known for their ability to "learn" complex relations, in practice NN results may sometimes produce unstable predictions, overfit the data, or converge to a local optimum. As a safeguard against these problems, our specific implementation utilized the median prediction of an ensemble of 100 NNs, each using a different random seed.

[2]If the number of inputs is an odd number, we round up the number of hidden nodes.

Figure C1 displays the results obtained with NN using the different data representations (i.e., the different model types defined in Table 1, in the main body of the paper). Table C1 presents the differences in MAPE values between models utilizing different sets of data and the corresponding significance values, using a bootstrap confidence interval. In sum, we reach findings that are similar to those reported in the main body of the paper regarding the relative performance of the different prediction models using search trend data and forum data (as defined in Table 1).

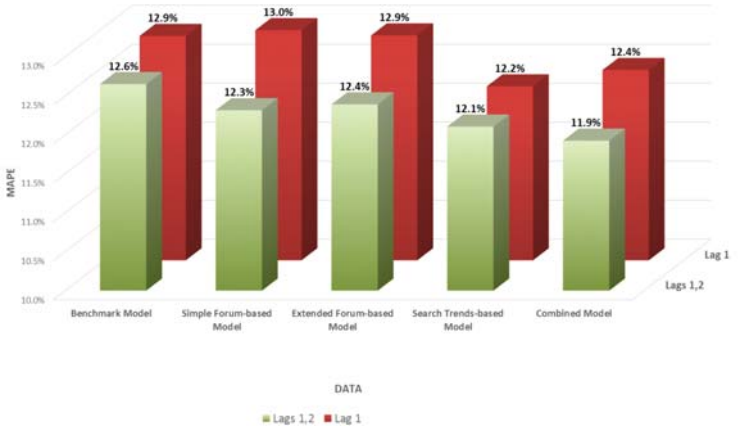


Figure C1. Prediction Results Using the NN Algorithm

Table C1. MAPE Differences and One-Sided Confidence Intervals for the Difference in MAPE Values Using the NN algorithm for Each Model

Model A	Model B	NN – Lag 1	NN – Lag 1,2
Forum-Based Model	Benchmark Model	-0.07%	0.34%**
Extended Forum-Based Model	Benchmark Model	-0.01%	0.26%
Search Trends-Based Model	Benchmark Model	0.65%***	0.55%***
Combined Model	Benchmark Model	0.44%**	0.73%***
Search Trends-Based Model	Forum-Based Model	0.72%***	0.21%*
Search Trends-Based Model	Extended Forum-Based Model	0.65%***	0.28%*
Combined Model	Forum-Based Model	0.51%**	0.39%***
Combined Model	Extended Forum-Based Model	0.44%**	0.47%***
Combined Model	Search Trends-Based Model	-0.21%	0.18%

Table C1 reports the difference in MAPE using two models (*Model A* and *Model B* - each based on different data inputs) while considering 1 or 2 lags with the NN algorithm. Specifically, the table reports the difference: diff = MAPE(Model B) - MAPE(Model A). Therefore, a positive value associated with the comparison between Model A and Model B indicates better predictive accuracy of Model A over Model B. Lower confidence interval bounds for *diff* were calculated using 2000 iterations of the BCA bootstrapping confidence interval calculation method implemented in R software. A lower confidence interval bound for *diff*, with a positive value, provides confidence that *MAPE(Model A)* is indeed better than *MAPE(Model B)*.

We report the following lower confidence bounds:

* 0.9 lower confidence bound for *diff* is positive

** 0.95 lower confidence bound for *diff* is positive

*** 0.99 lower confidence bound for *diff* is positive

Appendix D

Analysis Using Expanding Window Approach

In addition to using the moving window validation methodology, for robustness we also evaluated an expanding window approach using 24 months of expanding training data. Implementing this method, we followed common practice and used (at least) two-thirds of our data as training set and one-third of our data as validation. We therefore report performance based on the entire out-of-sample validation period (months $t = 25, \ldots, 36$). That is, for each validation month t we measure performance while applying the model trained during the preceding months (months 1 to $t - 1$). We note that month $t = 1$ is January 2008 and month $t = 25$ is January 2010.[3]

Figure D1 displays the results obtained with LR using the different data representations. Table D1 presents the differences in MAPE values (performance differences) between models utilizing different sets of data and the corresponding significance values using a bootstrap confidence interval. Overall, the findings obtained using the "expanding window" approach are similar to those obtained using the "moving window" approach.

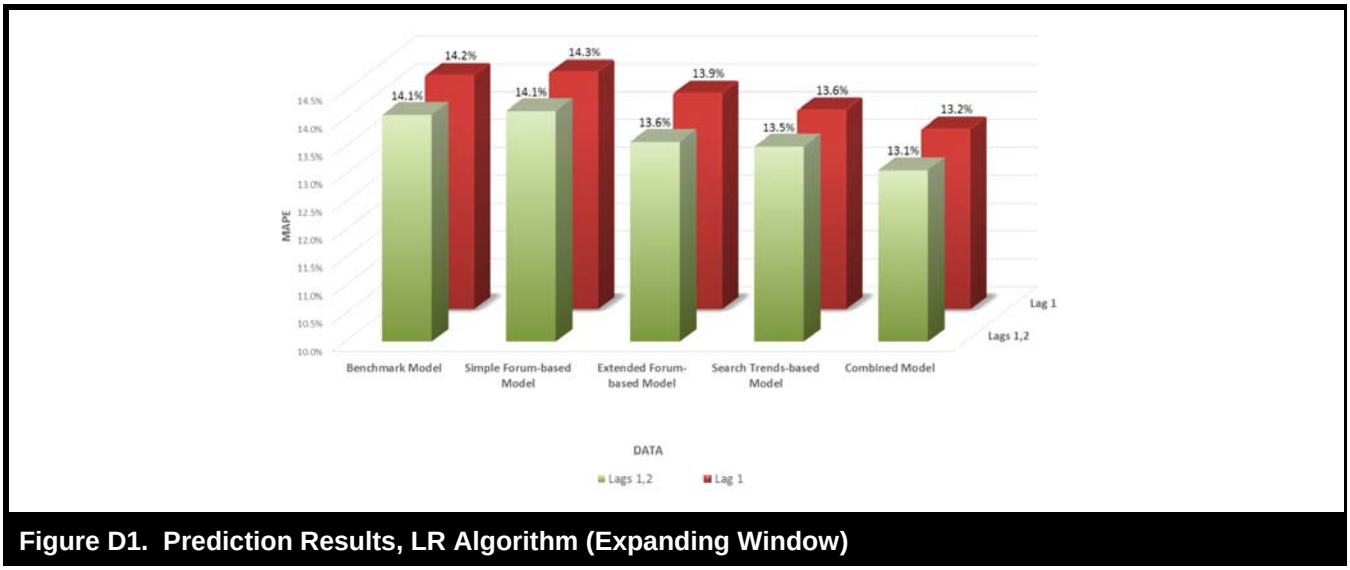


Figure D1. Prediction Results, LR Algorithm (Expanding Window)

[3]Our data are for January 2007–December 2010. We "lose" 12 months' worth of data when accounting for seasonality.

Table D1. MAPE Differences and One-Sided Confidence Intervals for the Difference in MAPE Values Using the LR algorithm for Each Model (Using Expanding Window Validation)

Model A	Model B	LR – Lag 1	LR – Lag 1,2
Forum-Based Model	Benchmark Model	-0.06%	-0.07%
Extended Forum-Based Model	Benchmark Model	0.32%*	0.48%**
Search Trends-Based Model	Benchmark Model	0.62%***	0.57%***
Combined Model	Benchmark Model	0.96%***	0.99%***
Search Trends-Based Model	Forum-Based Model	0.68%***	0.64%***
Search Trends-Based Model	Extended Forum-Based Model	0.30%	0.09%
Combined Model	Forum-Based Model	1.02%***	1.06%***
Combined Model	Extended Forum-Based Model	0.65%***	0.51%***
Combined Model	Search Trends-Based Model	0.34%*	0.42%*

Table D1 reports the difference in MAPE using two models (*Model A* and *Model B* - each based on different data inputs) while considering 1 or 2 lags with the LR algorithm. Specifically, the table reports the difference: diff = MAPE(Model B) - MAPE(Model A). Therefore, a positive value associated with the comparison between Model A and Model B indicates better predictive accuracy of Model A over Model B. Lower confidence interval bounds for *diff* were calculated using 2000 iterations of the BCA bootstrapping confidence interval calculation method implemented in R software. A lower confidence interval bound for *diff*, with a positive value, provides confidence that *MAPE(Model A)* is indeed better than *MAPE(Model B)*. We report the following lower confidence bounds:

* 0.9 lower confidence bound for *diff* is positive
** 0.95 lower confidence bound for *diff* is positive
*** 0.99 lower confidence bound for *diff* is positive

Appendix E

Analysis According to Car Brand Characteristics

In this appendix, we provide additional results and robustness checks regarding the predictive accuracy of the extended forum-based model and the combined model. Specifically, Figures E1 and E2 provide an additional graphic illustration of different models' predictive accuracy in the cases of premium and value brands, where a brand's affiliation with either category is based on its perceived quality (Figure E1) and on the extent to which it is associated with willingness to recommend (Figure E2). This analysis was carried out using one lag of data. For robustness we repeated the comparisons between value and premium brands (Figures E3–E5) using two lags of data. In sum, we observe similar findings to those reported in the main body of the paper. Specifically, for value brands, the combined model significantly outperforms the extended forum-based model, whereas for premium brands differences are smaller and not significant. (In Appendix F we provide more details regarding the YouGov data that were used for the perceived quality and willingness to recommend measurements.)

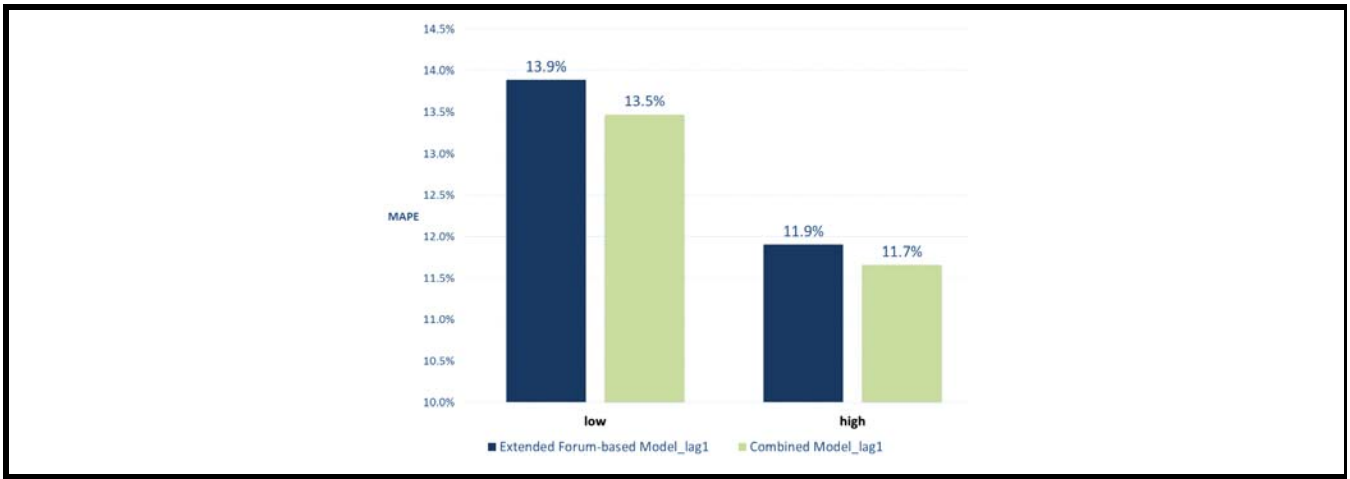


Figure E1. Prediction Results for Low- Versus High-Perceived Quality Brands Using LR and One Lag of Data

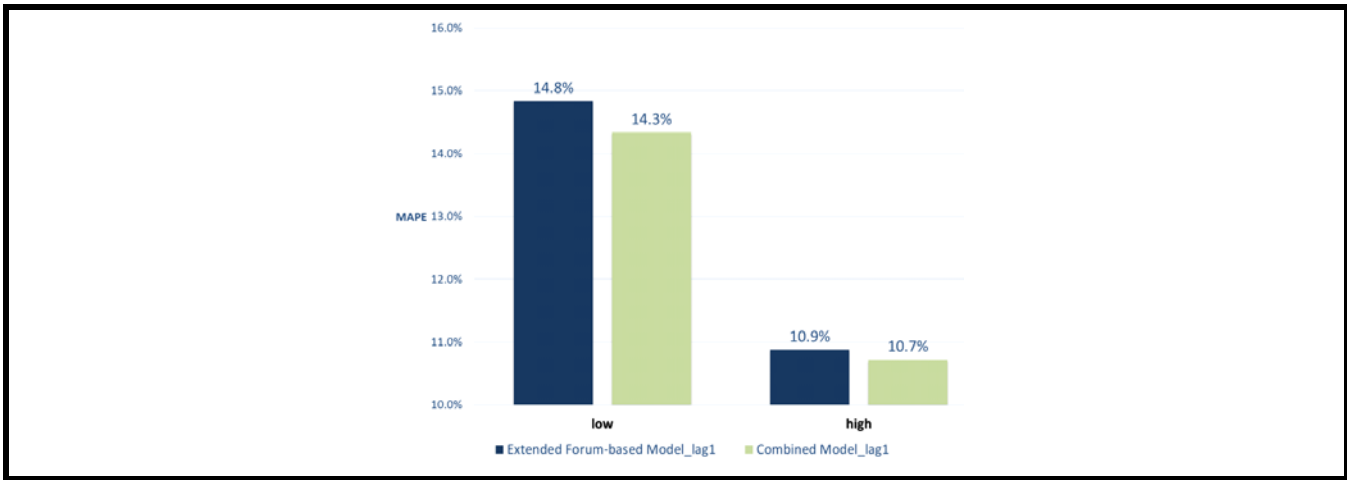


Figure E2. Prediction Results for Low- Versus High-Willingness-to-Recommend Brands Using LR and One Lag of Data

Figure E3. Prediction Results for Low- Versus High-Price Brands Using LR and One Lag of Data

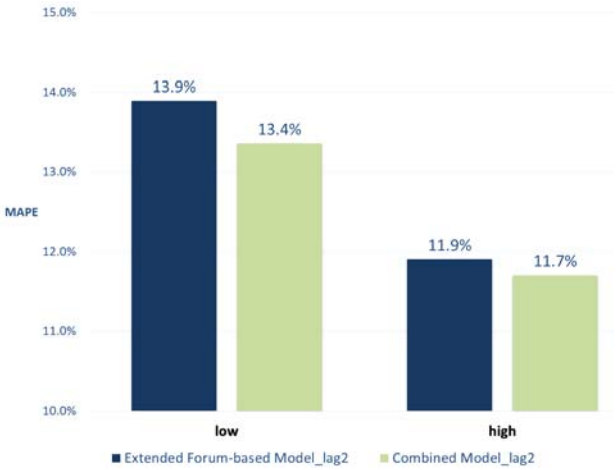


Figure E4. Prediction Results for Low- Versus High-Perceived Quality Brands Using LR and Two Lags of Data

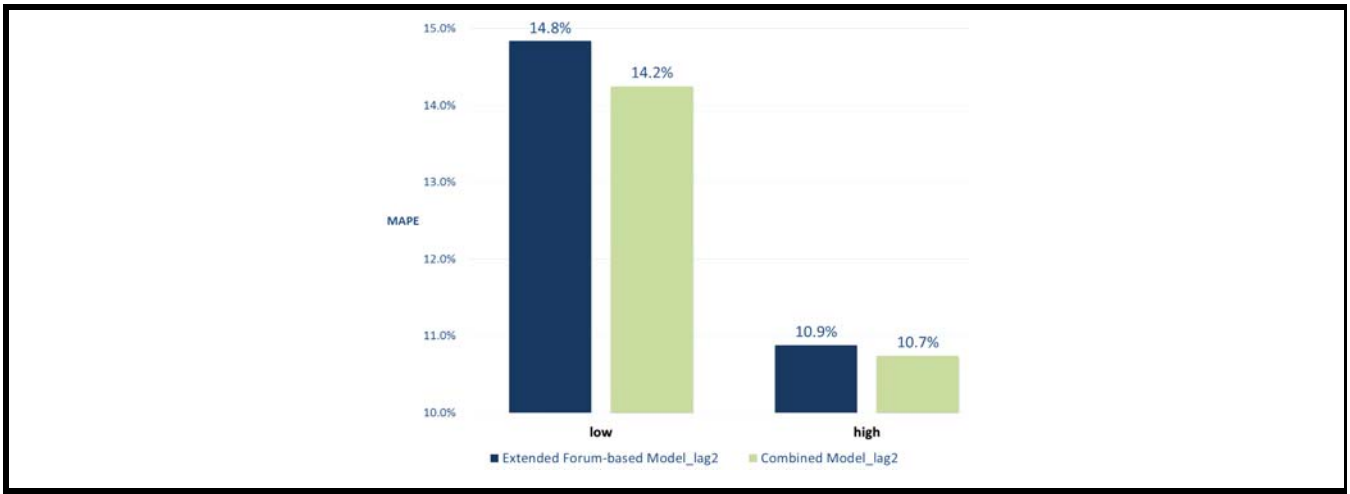


Figure E5. Prediction Results for Low- Versus High-Willingness-to-Recommend Brands Using LR and Two Lags of Data

Appendix F

YouGov Survey Data

In this work we utilized customer perception data obtained from YouGov plc. YouGov monitored a panel of 5,000 people in the United States, on a daily basis, and reported on brand-related perceptions. Our data regarding perceived quality and willingness to recommend were based on survey participants' average ratings for the period 2008–2010.[4] Regarding product quality, panel participants were presented with the following questions on a daily basis:

- Which of the following brands do you think represents good quality?
- Now which of the following do you think represents poor quality?

Daily quality scores were calculated according to the following formula:

$$Perceived\ Quality = \frac{PositiveCount - NegativeCount}{PositiveCount + NegativeCount + NeutralCount}$$

Regarding willingness to recommend, panel participants were presented the following questions on a daily basis:

- Which of the following brands would you recommend to a friend or colleague?
- Which of the following brands would tell a friend or colleague to avoid?

Daily willingness-to-recommend scores were calculated in a similar manner to the quality scores, according to the following formula:

$$Willingness\ to\ Recommend = \frac{PositiveCount - NegativeCount}{PositiveCount + NegativeCount + NeutralCount}$$

To avoid question biases, YouGov utilizes different respondents for each question. During the relevant time period, the average number of daily respondents for the quality question was 130 (s.d. 25.2). The average number of daily respondents for the willingness-to-recommend questions was 126 (s.d. 23.7).[5]

[4]For the Hyundai and Kia brands, data are available only for 2010.

[5]Source: YouGov BrandIndex product documentation and data.

Appendix G

Car Model-Level Analysis

The main body of the paper reports on predictions of car brand sales. In this appendix we present a robustness check that involves prediction of car model sales. To evaluate predictive accuracy at the car model level, we created a data set consisting of all the car models whose annual sales exceeded 10,000 units in the United States, and that were sold continuously (and not replaced by a new model with the same name) during the years 2007–2010 (a total of 78 car models).

Main Analysis

Figure G1 displays the car model-level results obtained with LR using the different data representations. Table G1 presents the differences in MAPE values (performance differences) between models utilizing different sets of data and the corresponding significance values using a bootstrap confidence interval. The results reported include models with one lag of data and with two lags of data. Notably, although we tested prediction models using up to five lags of data, we found that adding data from lag 3 or higher actually degraded predictive accuracy for all the models. (In effect, this degradation actually begins at lag 2 for all models except for the forum-based models, which display a minor improvement at lag 2.) Overall, the car-model-level results are in line with the car-brand-level results reported in the main body of the paper: Specifically, the combined model significantly improves predictive accuracy as compared with models based on forum data alone, and models based on search trend data alone obtain comparable (or superior) results to those of forum-based and extended forum-based models.

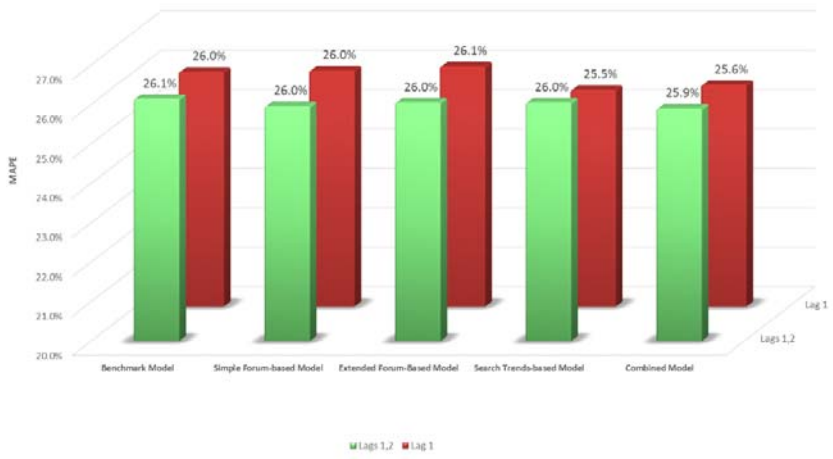


Figure G1. Prediction Results, Car Model Level

Table G1. MAPE Differences and One-Sided Confidence Intervals for the Difference in MAPE Values

Model A	Model B	LR – Lag 1	LR – Lag 1,2
Forum-Based Model	Benchmark Model	-0.04%	0.19%***
Extended Forum-Based Model	Benchmark Model	-0.14%	0.09%
Search Trends-Based Model	Benchmark Model	0.46%***	0.09%
Combined Model	Benchmark Model	0.33%***	0.25%**
Search Trends-Based Model	Forum-Based Model	0.5%***	-0.10%
Search Trends-Based Model	Extended Forum-Based Model	0.6%***	0.00%
Combined Model	Forum-Based Model	0.37%***	0.06%
Combined Model	Extended Forum-Based Model	0.47%***	0.15%*
Combined Model	Search Trends-Based Model	-0.13%	0.15%*

Table G1 reports the difference in MAPE using two models (*Model A* and *Model B* - each based on different data inputs) while considering 1 or 2 lags with the LR algorithm. Specifically, the table reports the difference: diff = MAPE(Model B) - MAPE(Model A). Therefore, a positive value associated with the comparison between Model A and Model B indicates better predictive accuracy of Model A over Model B. Lower confidence interval bounds for *diff* were calculated using 2000 iterations of the BCA bootstrapping confidence interval calculation method implemented in R software. A lower confidence interval bound for *diff*, with a positive value, provides confidence that *MAPE(Model A)* is indeed better than *MAPE(Model B)*. We report the following lower confidence bounds:

* 0.9 lower confidence bound for *diff* is positive
** 0.95 lower confidence bound for *diff* is positive
*** 0.99 lower confidence bound for *diff* is positive

Premium Versus Value Models

For robustness we compared the performance of the combined model with that of the extended forum-based model for different car characteristics at the car model level. Figure G2 presents the prediction results for high price (premium) versus low price (value) car models (defined, as in the main body of the paper, using a price threshold of $20,000, and one lag of data). Results are consistent with the brand-level results in that the difference in predictive accuracy between the combined model (adding search trends over forums data) and the extended forum-based model is larger for low-price car models.

Figure G2. Prediction Results for High-Price Versus Low-Price Car Models

Appendix H

Additional Robustness Checks Based on Extended Keyword Selection and MSE Criteria

Extended Keyword Selection

To test the robustness of our results to the keyword selection process, we repeated the analysis using a different keyword selection method. Specifically, we added predictors (explanatory variables) to the models described in the main body of the paper, based on search and forum data derived from additional keywords: specifically, keywords associated with the top selling car model for each brand. Figure H1 displays the results obtained with LR models. Overall, the findings obtained using this keyword approach are similar to those obtained using the keyword approach reported in the main body of the paper.

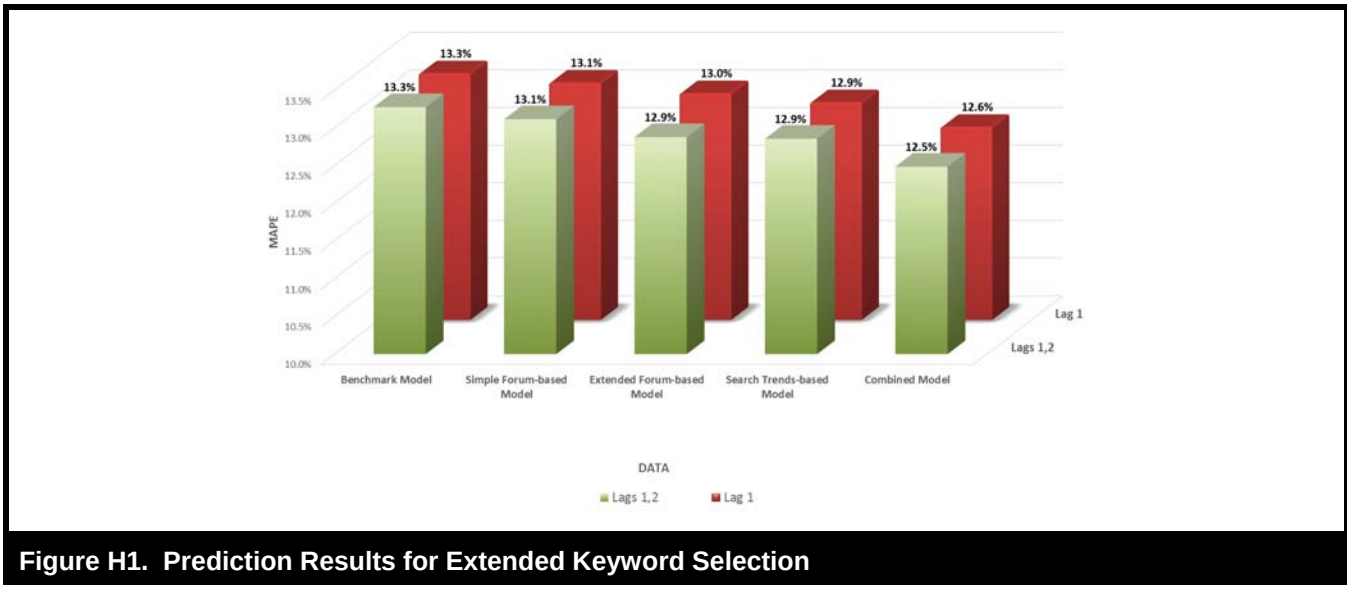


Figure H1. Prediction Results for Extended Keyword Selection

MSE Results

In this section we provide results using Mean Squared Error (MSE) criteria rather than MAPE, which was used in the main body of the paper. Figure H2 displays the results obtained with LR using the different data representations. Analyses carried out using the MSE criteria yielded similar results to those reported in the main body of the paper using the MAPE criteria.

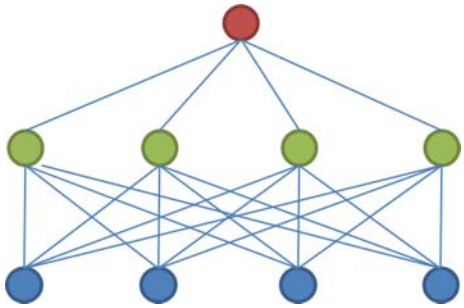
Figure H2. Prediction Results (MSE)

Appendix I

Additional Information on Neural Networks

The neural network is a biologically inspired model that attempts to learn patterns from data directly (Rumelhart et al. 1986). The NN is represented by a weighted directed graph containing three types of nodes (or neurons) organized in layers: the input nodes, the hidden nodes and the output node(s). A neuron receives input signals from a previous layer, aggregates those signals based on an input function, and generates an output signal based on an output (or transfer) function. The output signal is then routed to the other nodes in the network according to the network configuration. Each link connecting any two nodes is characterized by a weight. These weights are determined through a training process in which the NN repeatedly receives examples of past data instances for which the actual output is known, thereby allowing the system to adjust the weights.

Finding the values of these weights requires solving an optimization problem. Perhaps the most common method is the backpropagation algorithm (see, for instance, Bishop 1995). This method consists of two phases: feed-forward and backward-propagation. In the feed-forward phase, outputs are generated for each node on the basis of the current weights and are propagated to the output nodes to generate predictions. Then, in the backward-propagation phase, "prediction errors" are propagated back, layer by layer, and the weights of the connections between the nodes are adjusted for error minimization. The feed-forward and backward-propagation phases are executed iteratively, until convergence.



An illustration of an NN with 4 input nodes (blue), one layer with 4 hidden nodes (green), and a single output node (red).

Figure I1. An Illustration of a Simple NN

Appendix J

Forum Data Characteristics

In order to represent forum data we used Google's vast scan of the Internet. To the best of our knowledge, this is the most comprehensive scan of forum data that has been made available for any academic research. Figure J1 provides a distribution of the relative volume of forum mentions per brand. Figures J2 and J3 provide distributions of the volume of forum mentions with positive or negative sentiment, respectively.

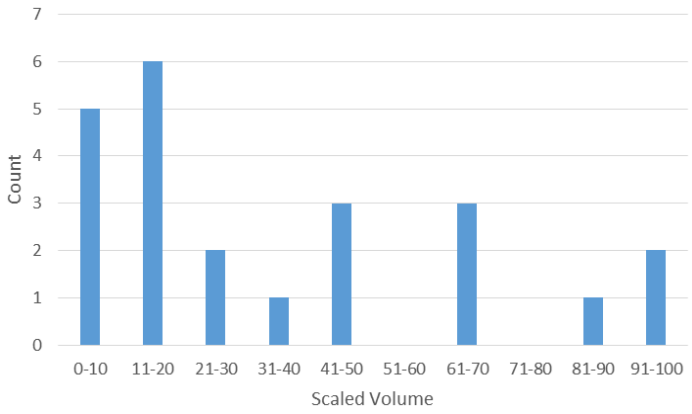


Figure J1. Brand Forum Mention Average Volume (by Decile)

Figure J1 presents a distribution of scaled forum mention volume for the 23 different brands. To preserve data confidentiality, the volume of forum mentions for each brand is scaled using the following method: (brand mention volume) / [(volume for the most mentioned brand) – (volume for the least mentioned brand)]

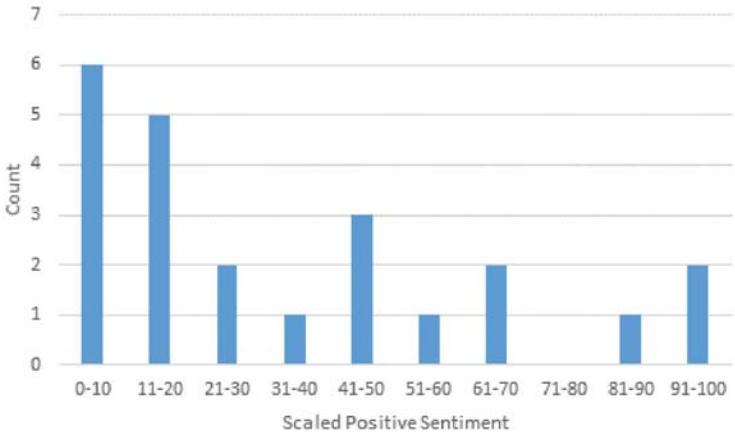


Figure J2. Brand Forum Positive Sentiment Average Volume (by Decile)

Figure J2 presents a distribution of scaled positive sentiment mentions for the 23 different brands. To preserve data confidentiality, the volume of positive mentions for each brand is scaled using the following method: (brand positive mentions volume)/[(positive mentions for the brand with the highest volume of positive mentions) – (positive mentions for the brand with the lowest volume of positive mentions)]

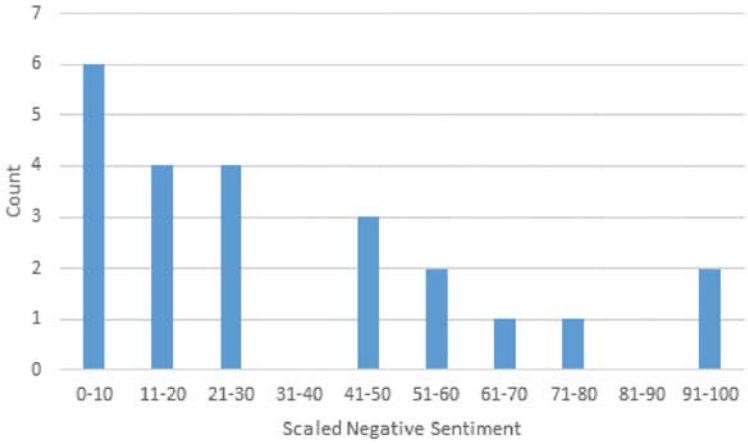


Figure J3. Brand Forum Negative Sentiment Average Volume (by Decile)

Figure J3 presents a distribution of scaled negative sentiment mentions for the 23 different brands. To preserve data confidentiality, the volume of negative mentions for each brand is scaled using the following method: (brand negative mentions volume)/[(negative mentions for the brand with the highest volume of negative mentions) – (negative mentions for the brand with the lowest volume of negative mentions)]

References

Bishop, C. . 1995. *Neural Networks for Pattern Recognition*, Oxford, UK: Clarendon Press.

Rumelhart, D. E., McClelland, J. L., and Williams, R. J. 1986. "Learning Internal Representation by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, MA: MIT Press, pp. 318-362.

Werbos, P. J. 1974. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences," unpublished Ph.D. dissertation, Harvard University.

Zhang, G., Patuwo, B. E., and Hu, M. Y. 1998. "Forecasting with Artificial Neural Networks: The State of the Art," *International Journal of Forecasting* (14:1), pp. 35-62.