Paper 1185-2017

Visualizing Market Structure Using Brand Sentiments

Praveen Kumar Kotekal, MS in Business Analytics, Oklahoma State University
Dr. Goutam Chakraborty, Oklahoma State University, Dr. Amit Ghosh, Cleveland State University

Abstract

Increasingly, customers are using Social media and other Internet-based applications such as review sites and discussion boards to voice their opinions and express their sentiments about brands. Such spontaneous and unsolicited customer feedback can provide brand managers with valuable insights about competing brands. There is a general consensus that listening and reacting to the "voice of the customer" is a vital component of brand management. However, the unstructured, qualitative, and textual nature of customer data that is obtained from customer's poses significant challenges for data scientists and business analysts.

In this paper we propose a methodology that can help brand managers visualize the competitive structure of a market based on an analysis of customer perceptions and sentiments that are obtained from blogs, discussion boards, review sites, and other similar sources. The brand map is designed to graphically represent the association of product features with brands, thus helping brand managers assess a brand's "true" strengths and weaknesses based on the voice of customers. Our multi-stage methodology uses the principles of Topic Modelling and Sentiment Analysis in text mining. The results of text mining is analyzed to represent the differentiating attributes of each brand. We empirically demonstrate the utility of our methodology by using data collected from Edmunds.com – a popular review site for car buyers.

Introduction

Brand management is turning out to be very essential with growing competition, and reviews given by customers are turning out to become first impressions of a brand for several prospects. Keeping the growing importance of brand management and capacity of customer reviews, we chose this topic to elucidate an example on how reviews from various blogs and new sites can be used to know about the brand perception. The dataset compiled for this project serves as a foundation for additional research.

Data Extraction

The Edmunds.com web site allows consumers to post reviews on different car models. Consumers can rate their vehicle's model, list pros and cons, and write their own review. In this paper we examine consumer reviews and ratings on four popular car brands named as Subaru, Chevrolet, Toyota and Honda. We extracted the data from Edmunds by accessing the Dealers API of Edmunds using the JSON package in Python. Using this API, we were able to fetch reviews and ratings provided by users on various car models based on brand, model name and year. To extract data, we built a tool that automatically iterates through all the model's available for a car brand in a year. Using this tool, we scrapped 2176 text reviews along with numerical ratings for car models released in 2012 till 2016.

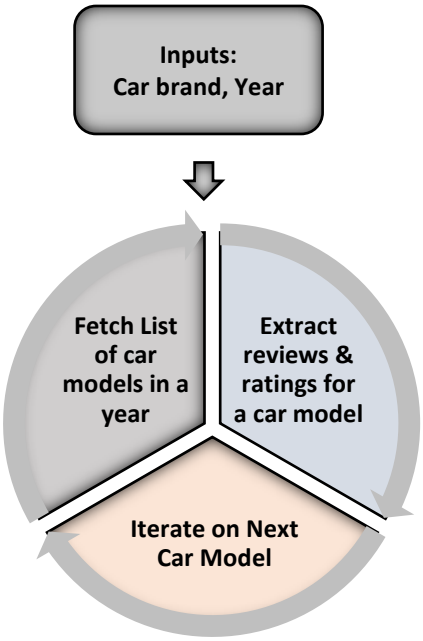


Fig1. Data Extraction Flow Chart

Data Cleaning & preparation

The raw extract data has reviews having stale information, we went through the data and filtered out such reviews. After initial text parsing, we came across terms which doesn't give any insights, we removed such reviews and we were finally left with a dataset of 2176 review.

Methodology

The process that we followed to come up with the analysis further described is shown in the flow chart below.

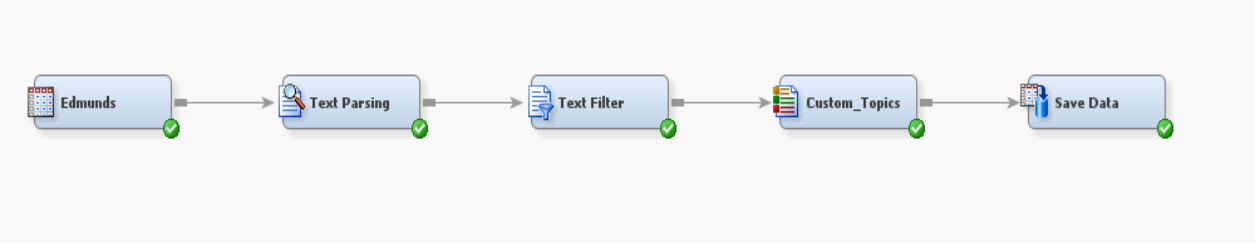
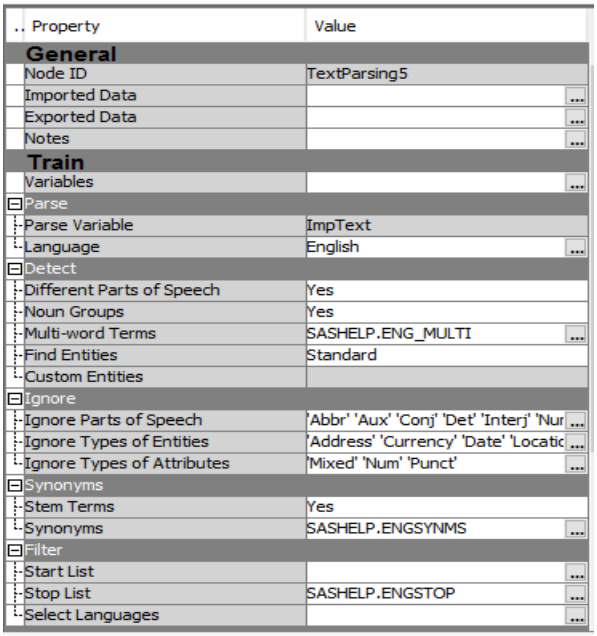


Fig2. Process Flow Chart

Text Parsing & Text Filtering

We used Text Parsing node in SAS® Enterprise Miner to parse and understand the reviews. Through text parsing, we wanted to eliminate terms related to articles and connectors which have high frequency, and also terms that do not actually describe any feature in particular. For example, the stop list can include the names of different car models, so that topic extraction results disregard language use patterns that

revolve around specific car models and focus on high-level concepts. We created a list of such stop words and used it while parsing to exclude them from further analysis. Besides stop words, we used a custom dictionary to detect synonyms and a dictionary to detect multi-word terms. To suit to our analysis goals, we have ignored parts of speech like 'Abbr', 'Aux', 'Conj', 'Det', 'Interj', 'Num', 'Part', 'Pref', 'Prep', 'Pron', 'Prop', entities like 'Address','Currency','Date','Location','Measure','Percent','Phone','Timeperiod' and attributes like 'Punct', 'Mixed', 'Num'. The configurations of text parsing node are shown below.



Property	Value
General	
Node ID	TextParsing5
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Parse	
Parse Variable	ImpText
Language	English
Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	Standard
Custom Entities	
Ignore	
Ignore Parts of Speech	'Abbr' 'Aux' 'Conj' 'Det' 'Interj' 'Nur
Ignore Types of Entities	'Address' 'Currency' 'Date' 'Locatic
Ignore Types of Attributes	'Mixed' 'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	
Stop List	SASHELP.ENGSTOP
Select Languages	

Fig3. Text Parsing Node Configurations

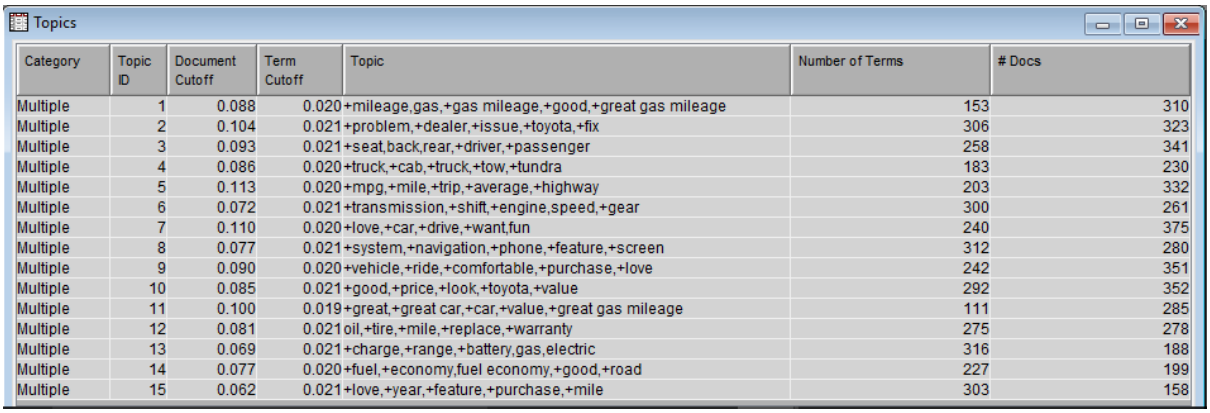
Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
+ be	Verb	Alpha	8083	1756	N	+	25091	1
+ have	Verb	Alpha	4391	1484	N	+	24980	2
+ car	Noun	Alpha	4053	1355	Y	+	9743	3
not	Adv	Alpha	3481	1334	N		25000	4
+ do	Verb	Alpha	2020	983	N	+	25143	5
+ get	Verb	Alpha	1833	976	N	+	24926	6
+ drive	Verb	Alpha	1252	779	Y	+	15206	7
s	Noun	Alpha	1362	727	N		24923	8
very	Adv	Alpha	1204	724	N		24957	9
+ mile	Noun	Alpha	1170	692	Y	+	15281	10
+ good	Adj	Alpha	1010	678	Y	+	1424	11
+ buy	Verb	Alpha	865	634	Y	+	749	12
+ much	Adj	Alpha	856	547	N	+	25159	13
no	Adv	Alpha	838	546	N		25172	14
+ go	Verb	Alpha	761	545	N	+	24946	15
zero maintenanc...	Noun Group	Alpha	1	1	Y		13035	6918
zero mechanical...	Noun Group	Alpha	1	1	Y		4258	6918
+ zero mechanic...	Noun Group	Alpha	1	1	Y	+	2423	6918
+ zero other com...	Noun Group	Alpha	1	1	Y	+	1176	6918
+ zero unexpect...	Noun Group	Alpha	1	1	Y	+	290	6918
zero warranty	Noun Group	Alpha	1	1	Y		1057	6918
zip	Noun	Alpha	1	1	Y		17262	6918
zippin	Noun	Alpha	1	1	Y		6812	6918
zippy little car	Noun Group	Alpha	1	1	Y		1883	6918
+ zombie	Noun	Alpha	1	1	Y	+	7060	6918
+ zone	Verb	Alpha	1	1	Y	+	12699	6918
zone climate co...	Noun Group	Alpha	1	1	Y		11229	6918
zoom	Noun	Alpha	1	1	Y		4699	6918
zx	Noun	Alpha	1	1	Y		11958	6918
+ zx car	Noun Group	Alpha	1	1	Y	+	17363	6918

Fig4. Terms Table

The terms table output from Text Parsing node shows the frequency of occurrence of terms. The output of text parsing node above shows the term by frequency output, it is evident from the output that highly frequent terms again turned out to be articles and other words which doesn't explain any special feature. Again terms with low frequency also don't give much information as there are not many occurrences. So we used a Text Filter node with term weight as 'Inverse Document Frequency' instead of frequency to assign weights to the terms, and set minimum number of documents to 4, so that terms that appear in less than four documents in the collection are filtered out. The default setting in the text filter node filters out all the words with low weight. We have also enabled spell check property to accommodate any typo errors, we used custom English dictionary to accommodate spell checks. The text filter node generates term by document matrix, viewing results in interactive filter mode will enable us to explore further using concept links.

Text Topic

The Text Topic node enables you to create topics of interest from a list of terms. Text Topic node extracts topics from text documents. A document can have multiple topics. Parameters of the text topic node can be manually set based on the size of the data and the frequency of unique terms. For this project the numbers of terms in a topic are limited to 15. Below term topic matrix shows us the distribution of various terms together as topics.


Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.088	0.020	+mileage,gas,+gas mileage,+good,+great gas mileage	153	310
Multiple	2	0.104	0.021	+problem,+dealer,+issue,+toyota,+fix	306	323
Multiple	3	0.093	0.021	+seat,back,rear,+driver,+passenger	258	341
Multiple	4	0.086	0.020	+truck,+cab,+truck,+tow,+tundra	183	230
Multiple	5	0.113	0.020	+mpg,+mile,+trip,+average,+highway	203	332
Multiple	6	0.072	0.021	+transmission,+shift,+engine,speed,+gear	300	261
Multiple	7	0.110	0.020	+love,+car,+drive,+want,fun	240	375
Multiple	8	0.077	0.021	+system,+navigation,+phone,+feature,+screen	312	280
Multiple	9	0.090	0.020	+vehicle,+ride,+comfortable,+purchase,+love	242	351
Multiple	10	0.085	0.021	+good,+price,+look,+toyota,+value	292	352
Multiple	11	0.100	0.019	+great,+great car,+car,+value,+great gas mileage	111	285
Multiple	12	0.081	0.021	oil,+tire,+mile,+replace,+warranty	275	278
Multiple	13	0.069	0.021	+charge,+range,+battery,gas,electric	316	188
Multiple	14	0.077	0.020	+fuel,+economy,fuel economy,+good,+road	227	199
Multiple	15	0.062	0.021	+love,+year,+feature,+purchase,+mile	303	158

Fig5. Text Topic output

From the default text topics that were formed, we see that customers are talking about various aspects of a car from seat comfort to transmission, warranty and mileage. We further created our own custom topics by grouping relevant terms based on the text topic node output and concept links from text filter node. The custom topic that are created describes every aspect of a car from comfort to safety and technology. Below table provides information about how many words exists in each user defined topic and number of times these topics are identified in text reviews.

Topic	No. of Terms	No. of Documents
Comfort	37	756
Interior	23	685
Performance	55	1138
Safety	29	483
Technology	24	554

Fig6. Custom Topics

Comfort
Adjustable heat
Adequate leg room
auto reverse mirror
cruise control
keyless ignition

Interior
seat height
passenger cabin
passenger side mirror
power window
bucket seat

Performance
engine acceleration
engine noise problem
turbo engine
fuel consumption
engine wear

Safety
bullet proof
airbag system
warning system
blind spot detection
traction control

Technology
radar cruise control
gps touch screen
sound insulation
stereo module
bluetooth system

Fig7. Text Topics – Top 5 Terms

Custom Topic Analysis

We further analyzed each of these topics by brand to find the proportion of customers talking about these attributes by brand. The bar chart below shows the percentage of customers talking about each of the topics by brand.

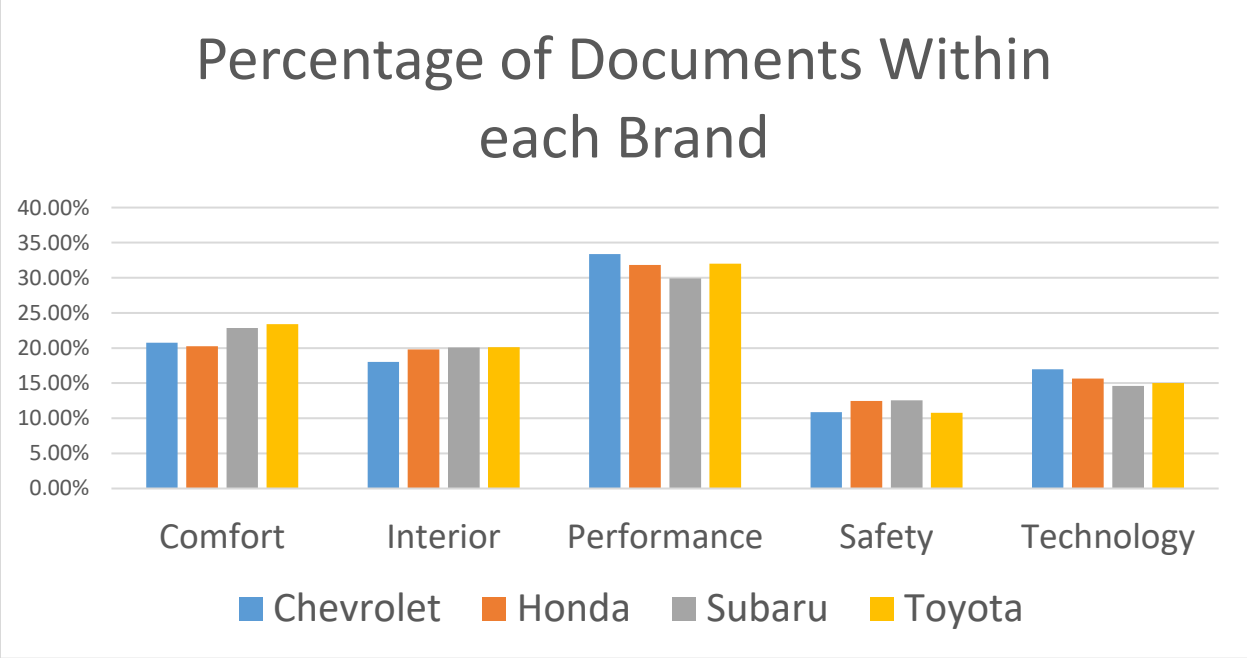


Fig8. Importance of Text Topics

Brand wise regression analysis

We further analyzed to see the sentiment of customers towards the brands in each of these areas. Consumers have provided ratings on a scale of 1-5 on categories such as Performance, Comfort, Technology, Interior, and Safety. Basically, numerical ratings provided by consumers are having maximum 25 percent missing values. By understanding the data, we have considered median value within each brand to impute missing values for a brand.

We have categorized all the `reviews which contains any of the terms mentioned under each of those topics and with the corresponding topic rating less than three as negative and all the reviews that contain any of the terms related to the custom topics discussed above an d with the respective topic rating greater than three as positive. We performed a regression analysis to confirm the same.

Rating	Binary Rating
1	Negative
2	
3	
4	Positive
5	

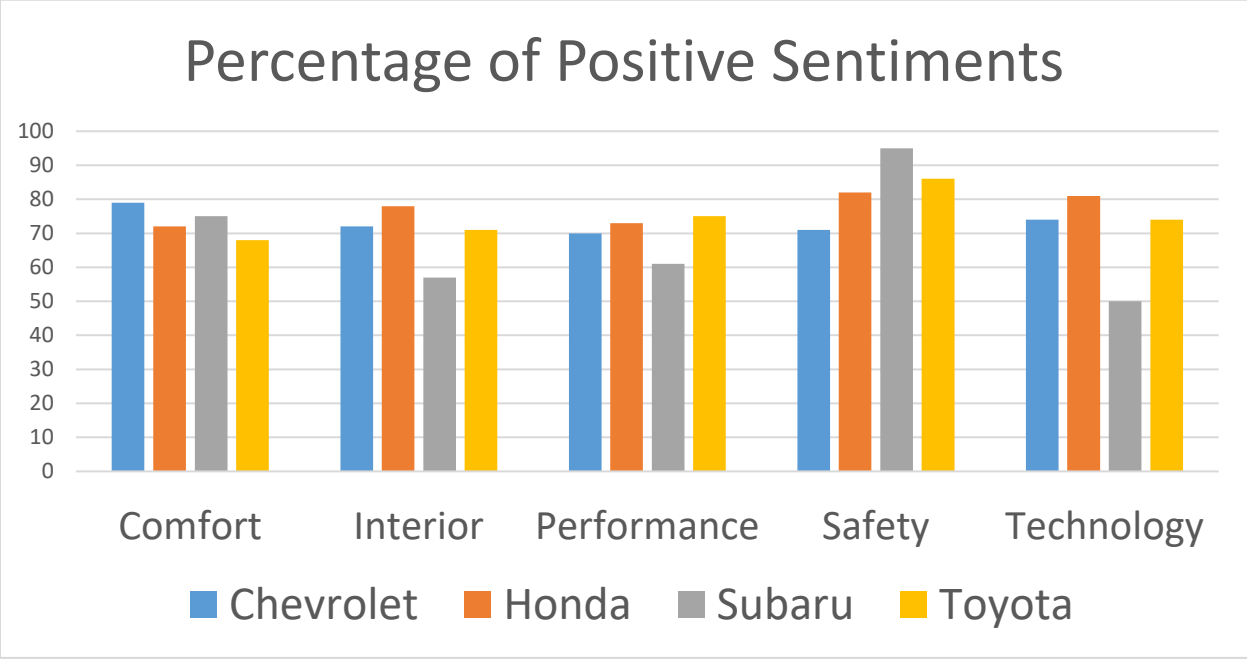
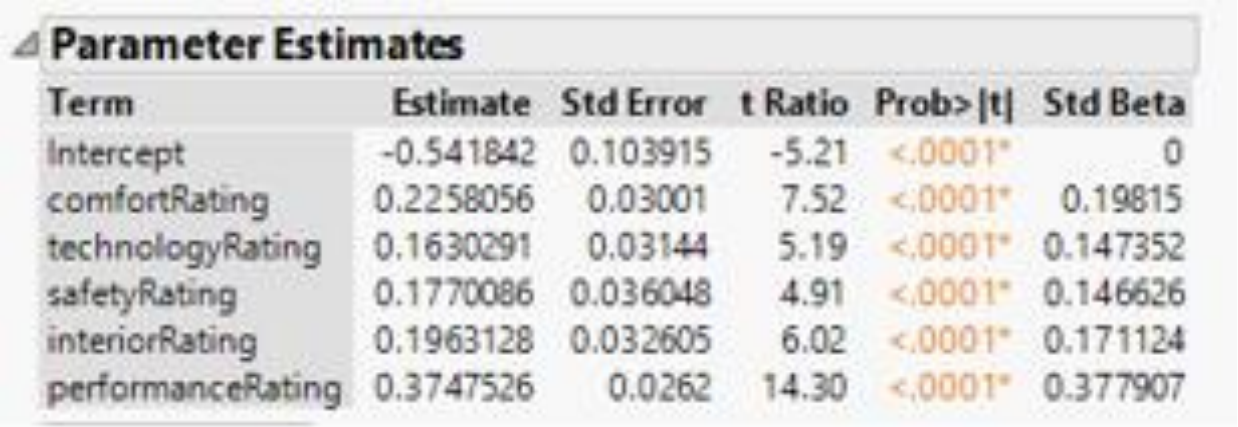


Fig9. Brand Sentiments

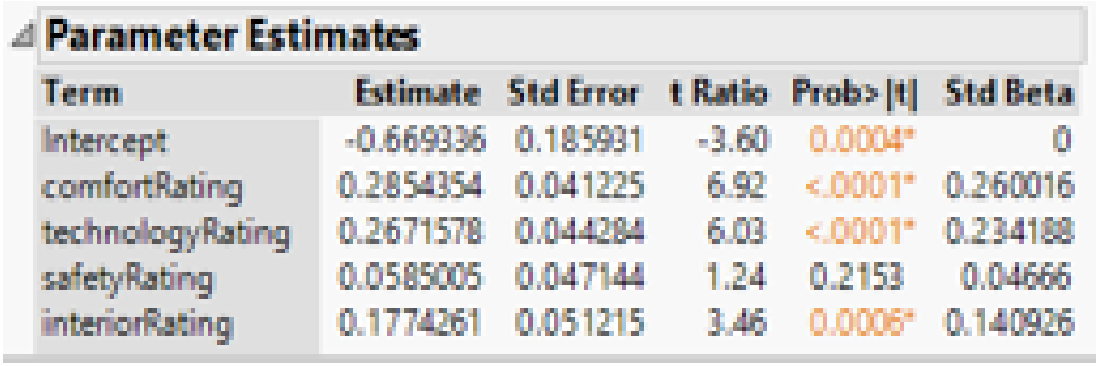


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	-0.541842	0.103915	-5.21	<.0001*	0
comfortRating	0.2258056	0.03001	7.52	<.0001*	0.19815
technologyRating	0.1630291	0.03144	5.19	<.0001*	0.147352
safetyRating	0.1770086	0.036048	4.91	<.0001*	0.146626
interiorRating	0.1963128	0.032605	6.02	<.0001*	0.171124
performanceRating	0.3747526	0.0262	14.30	<.0001*	0.377907

Fig10. Regression results for Chevrolet brand

The top two important predictor variables based on LogWorth value for this brand are Performance Rating and Comfort Rating. About 77% of variance is explained by this model.

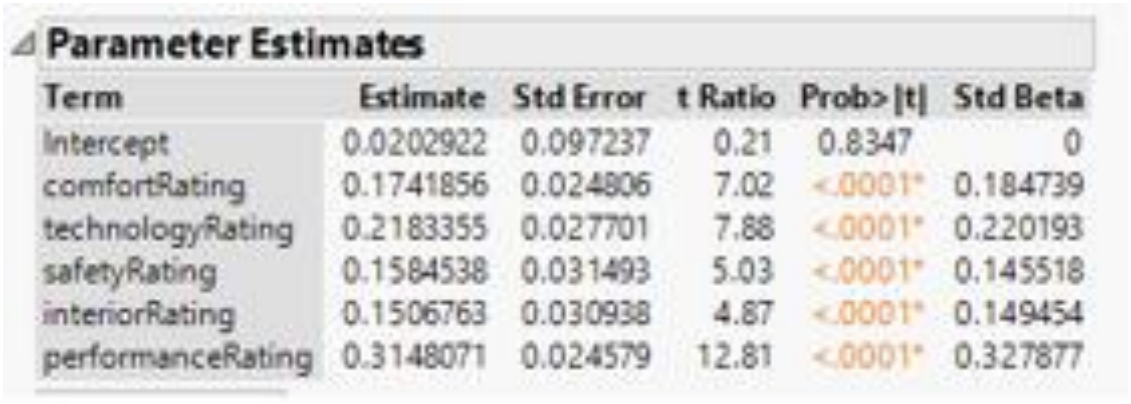


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	-0.669336	0.185931	-3.60	0.0004*	0
comfortRating	0.2854354	0.041225	6.92	<.0001*	0.260016
technologyRating	0.2671578	0.044284	6.03	<.0001*	0.234188
safetyRating	0.0585005	0.047144	1.24	0.2153	0.04666
interiorRating	0.1774261	0.051215	3.46	0.0006*	0.140926

Fig11. Regression results for Honda brand

The top two important predictor variables based on LogWorth value for this brand are Comfort Rating and Technology Rating. About 65% of variance is explained by this model.

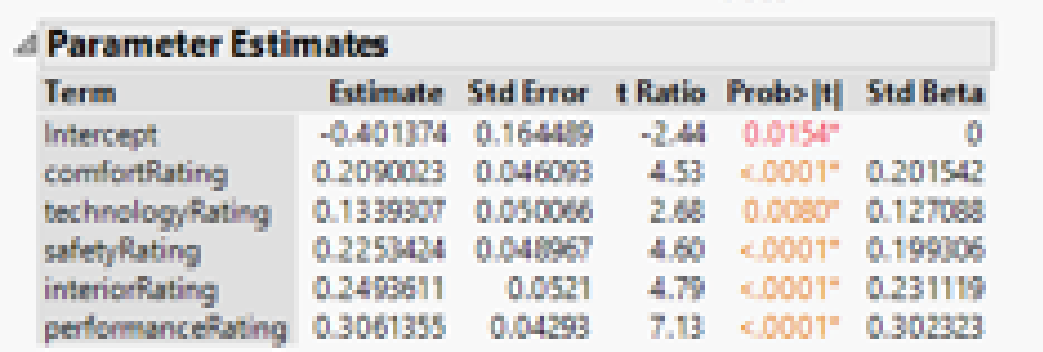


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	0.0202922	0.097237	0.21	0.8347	0
comfortRating	0.1741856	0.024806	7.02	<.0001*	0.184739
technologyRating	0.2183355	0.027701	7.88	<.0001*	0.220193
safetyRating	0.1584538	0.031493	5.03	<.0001*	0.145518
interiorRating	0.1506763	0.030938	4.87	<.0001*	0.149454
performanceRating	0.3148071	0.024579	12.81	<.0001*	0.327877

Fig12. Regression results for Toyota brand

The top two important predictor variables based on LogWorth value for this brand are Performance Rating and Technology Rating. About 72% of variance is explained by this model.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	-0.401374	0.164489	-2.44	0.0154*	0
comfortRating	0.2090023	0.046093	4.53	<.0001*	0.201542
technologyRating	0.1339307	0.050066	2.68	0.0080*	0.127088
safetyRating	0.2253424	0.048967	4.60	<.0001*	0.199306
interiorRating	0.2493611	0.0521	4.79	<.0001*	0.231119
performanceRating	0.3061355	0.04293	7.13	<.0001*	0.302323

Fig13. Regression results for Subaru brand

The top two important predictor variables based on LogWorth value for this brand are Performance Rating and Interior Rating. About 76% of variance is explained by this model.

Conclusion

We would like to conclude by summarizing the results obtained above and insights derived from the same. A methodology to facilitate brand management using textual data on brand sentiments is developed. This exercise reveals the strengths and weaknesses of brands and provides valuable diagnostic information. Subaru has the fewest reviews. Highest dissonance levels in terms of interior, performance and technology categories. Extreme positive consensus about safety. Niche marketing is recommended. Chevrolet has the highest reviews, moderate dissonance on all factors. Mass marketing is seen.

Limitations and Future Research

In this paper we only focused on four car brands and we identified brand sentiments using customer ratings on various aspects. In our future work we are planning to do sentiment mining on the text reviews to compare and contrast the results obtained from numeric ratings. We also would like to continue our research to see if brands can be positioned based on brand sentiment.

References

[1] Chakraborty, G., M. Pagolu and S. Garla. 2013. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®. Cary, NC: SAS® Institute.

[2] Kelley Blue Book Co. 2016. “Kelley Blue Book.” Accessed March 18, 2016. http://www.kbb.com/.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Praveen Kumar Kotekal
Email: praveen.kotekal@okstate.edu
MS in Business Analytics Program
Oklahoma State University, Stillwater, OK

Dr. Amit Ghosh, Cleveland State University
Email: A.GHOSH@csuohio.edu
Amit K. Ghosh is Chair & Professor of Marketing Department.

Dr. Goutam Chakraborty, Oklahoma State University, Stillwater OK
Email: goutam.chakraborty@okstate.edu
Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University.