Predict Automobile Sales Based on Consumer Online Search Behavior

Ruimei Zhang

School of Economics and Management, Xidian University, Xi'an 710071, China

15129011877@163.com

Abstract

Accurate forecasting of sales is crucial to marketing. In this paper, we collected sales data and online search dada of 46 vehicle types in the automobile market of China to investigate the relationship between online search and offline purchase. The regression results show that online search index is a good indicator for predicting sales. And adding the online search index can improve the goodness of fit of prediction model. The value of online information declines sharply over time.

Keywords

Online search, Sales prediction, Automobile.

1. Introduction

For a few billions of people around the world, the Internet has become an important part of their daily life. From the 1990s, Internet technology has profoundly influenced the way people access information, communications, cultural exchanges and transactions. In particular, the search engines have changed the way business operation by providing information and distribution channels for both retailers and consumers. For consumers, they can easily access product information without visiting a firm. For retailers, they can view who is visiting their website in real time and analyze the clickstream data in detail to get more information in advance.

According to the report published by China Internet Network Information Center (CNNIC), 77% of Internet users choose to search relevant information in advance when they purchase products. When consumers search online, they are actually sending signals about themselves: what they are searching for, what pages they visited before and after the search.... Every time a consumer searches for product information through the Internet, it reflects the consumer's willingness to buy. Therefore, consumer online search has become a valuable resource for understanding market demand.

Accurate forecasting of sales is crucial to marketing, because this information can be used to adjust the distribution of market budget and overall marketing strategy. If the online search data can be used to predict accurately, this method will be more effective than traditional methods such as questionnaire survey and social media monitoring. Because this method can not only get real-time data, but also can achieve zero cost almost. This paper studies the relationship between online search behavior and offline transactions on the basis of real data generated when consumers search information and make purchase decisions. Automobile requires large capital investment, consumers usually evaluate the overall characteristics when they decide to buy automobiles, so we choose the automobile as the research object. We collected sales data and online search dada of 46 vehicle types in the automobile market of China from January 2015 to June 2017 to investigate the relationship between automobiles sales and online search data.

2. Related Literature

Consumer information search is defined as a process in which consumers actively acquire and integrate information from various internal and external sources before making purchase decisions. When internal information search fails to meet demand, consumers will conduct external information search. Many aspects of consumer information search, such as information sources, influencing factors of search behavior, are the focus of research.

Online search data reflects consumers' focus and attention to the product. In recent years, with the opening of search engine data, the research based on search data prediction has gradually appeared. Ginsberg et al(2009) analyzed Google search request data to study the development path of influenza disease firstly, this method can predict the occurrence of influenza 1-2 weeks in advance compared to the government monitoring report. Kulkarni(2012) found that the network search before the release of the film followed a certain pattern, and the search model containing the online search data can significantly improve the prediction accuracy of the box office. Wu and Brynjolfsson (2014) found that using search data generated by search engines such as Google to predict is one of the ways to accurately predict future business activity, collecting search indices for the "real estate" category in Google Trends, The study found that the housing search index could significantly improve the accuracy of price and sales forecasts in the real estate market. Liu et al. (2016) found that using data on platforms such as Google Trends, Wikipedia and Twitter could improve the predictive effectiveness of TV ratings.

3. Empirical Analysis

In this section, we will empirically analyze the relationship between the online search data and offline sales of automobile. We first discuss our data sets and variable definitions, and then build the econometric model. Finally, we will show which variable are statistically significant.

3.1 Data and Variable

Our data mainly include the following aspects:

(1)Sales data. The monthly sales data of automobile come from Sohu automobile. Cars are divided into eight classes, namely mini car, small car, compact car, medium vehicles, large vehicles, luxury vehicles, MPV, and SUV. We selected 10 vehicle types from each class as candidate samples, and then the vehicle types that were missing more than 12 months between January 2015 to June 2017 were excluded. Finally, 46 vehicle types were collected in the automobile market of China. It is important to note that the purpose of this paper is to study how online search reflects current sales and predicts future trends. Therefore, when selecting samples, we should avoid those vehicles which are prone to ambiguity. For example, Haima automobile has a vehicle type called "Knight", considering that consumers enter "Knight" for search does not necessarily target this vehicle type, but may search for basketball information, this vehicle type is avoided in the selection of samples. In addition, when selecting vehicle types, it includes both high-sale types (such as HAVAL H6) and low-sale types to demonstrate the universality of this study.

(2)Search data. The search data comes from Baidu Index. Baidu is the largest Chinese search engine around the world. Baidu Index is a data sharing platform based on the real data of Baidu's massive Internet users' search behavior, which can reflect the attention of keywords in the past period .There are several different ways of keyword selection when determining search keywords in related research. Seebach (2011) and Geva (2013) use brand names as keywords to collect data, such as BMW, Volkswagen to collect the data of the subordinate types of the brand. However, using brand name as a key word may not reflect the real purchase intention of consumers, because search requests containing brand names may include non-purchasing information needs, such as understanding the brand's history, management or recruitment information. Obviously, these search requests have nothing to do with buying a car. The report of search market in automobile industry (2016) showed that brand words and product words are the most core words of consumer online search. Therefore, this paper selects "brand + model" as the search keywords, such as "Audi A4".

(3)The data of control variables. Since automobile sales are also influenced by other factors, referring to previous studies, we also collects four kinds of control variables, including automobile performance indicators, price information, policy variables and other variables. The description and data source of the control variables are detailed in Table 1.

Table 1 The Variable Definition and Data Source of Vehicle Sales Model

Variables	Definition	Data Source
Sale	The monthly sales data of the specific type	Sohu automobile
Search	The monthly search data of the specific type	Baidu Index
lSearch	the search index of the previous period	Baidu Index
gas	Average fuel consumption per 100 km	Auto home
Dummy size	Dummy variable indicating the class of the vehicle type	Auto home
Price	The minimum price of the specific type	Auto home
Oilprice	The oilprice	The national development and reform commission
tax	The purchase tax on cars	State Administration of Taxation
Dummy country	Dummy variable indicating the country of origin	Auto home
Dummy season	Dummy variable indicating the season	

Table 2 presents the descriptive statistics of our data after processing.

Tabel 2 Descriptive Statistics

Variables	Sample size	Mean	Median	Max	Min	SD
Sale	1380	11064.34	8600	80500	28	10215.9
Search	1380	180214.2	111795	1099229	2970	194338.2
Searcht-1	1380	180157.6	110865	1099229	2970	194993
Price	1380	153719.7	144800	449800	27900	102716.2
Oilprice	1380	7626.367	7548.5	8326	7026	376.7086
Gas	1380	6.481818	6.25	11	5	1.294932
tax	1380	7.786364	7.5	10	5	2.256362

3.2 Econometric Model

Firstly, we study the main effect, that is, the influence of online search index on automobile sales. In this model, the current search index and the previous search index are taken as independent variables to test the impact of the current month's online search and the last month's online search on automobile sales. The benchmark model is shown in Formula 1:

$$Sale_{it} = \alpha + \sum_{j=1}^{J} \beta_j Search_{it-j} + \mu_i + \varepsilon_{it} \quad (1)$$

And then, in order to show that the influence of online search index on automobile sales is independent of control variables, we build the control model (see Formula 2) and combination model (see Formula 3) respectively. In the control model, the control variables proposed above are used as independent variables, and the online search index is added to the combined model on the basis of the control model.

$$Sale_{it} = \alpha_2 + \sum_{k=1}^{K} \beta_k Controls_{it} + \mu_i + \varepsilon_{it} \quad (2)$$

$$Sale_{it} = \alpha_3 + \sum_{j=1}^{J} \beta_j Search_{it-j} + \sum_{k=1}^{K} \beta_k Controls_{it} + \mu_i + \varepsilon_{it} \quad (3)$$

3.3 Regression Results

We use Generalized Least Squares (GLS) method estimate the above three models. From the regression results shown in Table 3, we can see that Model 1 can explain the variance of 35.9% sales. *ln (search)* and *ln(*search $_{t-1}$*)* are significantly positively correlated with automobile sales (for *ln (search)*, β= 0.405, P < 0.001, for *ln(search* $_{t-1}$*)*, β= 0.124, P < 0.05). This shows that the online search index is a good indicator for predicting sales.

The estimated results of the combined model further show that when the network search index is added to the control model, Adj-R2 increases from 52.6% to 65.4%, which shows that adding the online search index can improve the goodness of fit of the model.

For the control variables, the regression results of model 2 show that Tax has a significant negative effect on sales (β=-0.013, P < 0.05). This shows that the purchase tax has a significant impact on automobile sales. With the increase of the purchase tax on automobiles, automobile sales will decline. Gas also has a significant negative effect on sales (β=-0.296, P < 0.001). This shows that the average fuel consumption per 100 km as an important performance index of automobiles may be one of the basis for consumers to make purchasing decisions. The higher the fuel consumption, the higher the price of the fuel required, so that consumers are more likely to buy cars with low fuel consumption under the same conditions.

In addition, it should also be noted that in the sales model of this paper, the lag period of the online search index is not arbitrarily determined, but after trying different lag periods. According to the regression results, the impact of two or more lag periods of network search index on automobile sales is not significant (P > 0.05). The possible explanation is that consumers have several months of information search and evaluation process before making purchase decisions, but the search intensity is small. Until one month before purchase, the number of searches will reach a peak, so the online search only shows a significant impact on sales one month before purchase. This result shows that the search for the target product on the online platform is very short before consumers make purchase decisions. The value of online information declines sharply over time.

Table 3 Regression Results

	Model 1	Model 2	Model 3
Control variables			
ln（price）		-0.232***	-0.018
Gas		-0.296***	-0.319***
ln（oilprice）		0.177	-0.414**
Tax		-0.013*	-0.011
Dummycountry1		0.397***	0.270***
Dummycountry2		-0.199***	-0.163**
Main Effect			
ln(search)	0.405***		0.381***
ln(searcht-1)	0.124*		0.138*
Adj-R2	0.359	0.526	0.654

*P<0.05；**P<0.01；***P<0.001.

4. Summary

With the development of search engine, the role of online search data has begun to generate the growing practice of analytics and extensive research in academia. Based on the previous research, this paper studies the relationship between online search behavior and offline transactions on the basis of real data generated when consumers search information and make purchase decisions. We collected monthly sales data and online search dada of 46 vehicle types in the automobile market of China from January 2015 to June 2017. The regression results show that online search index is a good indicator for predicting sales. And adding the online search index can improve the goodness of fit of prediction model. The value of online information declines sharply over time.

References

[1] Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data.[J]. Nature, 2009, 457(7232):1012.

[2] Kulkarni G, Kannan P K, Moe W. Using online search data to forecast new product sales[J]. Decision Support Systems, 2012, 52(3): 604-611

[3] Wu L, Brynjolfsson E. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities[C]// International Conference on Information Systems, Icis 2009
[4] Liu X, Singh P V, Srinivasan K. A structured analysis of unstructured big data by leveraging cloud computing[J]. Marketing Science, 2016, 35(3): 363-388.
[5] Erickson G M, Johansson J K, Chao P. Image Variables in Multi-Attribute Product Evaluations: Country-of-Origin Effects[J]. Journal of Consumer Research, 1984, 11(2):694-699.