Predictive Analytics of E-Commerce Search Behavior for Conversion

Full Paper

Xi Niu
University of North Carolina at Charlotte
xniu2@uncc.edu

Chuqin Li
University of North Carolina at Charlotte
cli30@uncc.edu

Xing Yu
Indiana University, Indianapolis
yu64@umail.iu.edu

Abstract

This study explores online customer search behavior on a large e-commerce website—Walmart.com. In order to more accurately predict customer purchase conversion based on their search behavior, we adopt a modern machine-learning technique, random forest, as well as logistic regression to develop two computational models. We also integrate information retrieval literature to propose metrics to quantify online consumers' search behavior. Results show that the random forest model performs better with a very high accuracy rate (76%) in predicting customers who will purchase the item they browsed. Among all the predictors, page and session dwell time, user type, click entropy, and click position are the strongest influential factors for the conversion behavior. The findings suggest that, with the enhanced metrics and modeling approaches, search behavior could offer strong cues about customers' purchasing decision. Additionally, the findings also suggest operational implications about how to accommodate and induce the desired search behavior with the e-commerce website.

Keywords

E-Commerce, Search Behavior, Conversion Rate, Computational Models, Random Forest

Introduction

Despite the fast growth of e-commerce sales, the online conversion rates across industries are usually very low, rarely exceeding 5% (eMarketer, 2014). Conversion rate is defined as the percentage of visits that result in purchases. Low conversion rates imply that most e-commerce websites traffic only represents casual visitors as opposed to serious buyers. E-commerce business stakeholders are struggling with marketing that attempts to recover an abandoned shopping cart through various marketing methods in order to boost the conversion percentage (eMarketer, 2014). One of the primary benefits for doing business online is that several aspects of customers' behavior can be tracked with assistance of modern technology. Due to vast amounts of behavioral data available online, extensive literature has explored the possibility to use those data to understand and predict their conversion decisions (da Silva, 2014; Fernandes, 2015).

One important aspect of such recorded behavior is consumers' search behavior on e-commerce websites. A customer who has an item in mind needs to engage a series of searches across the web to find a product of satisfactory quality and price. The literature in the information retrial (IR) field has provided many commonly used as well as advanced measures to quantify search behavior in various context (e.g., Spink, Jansen, & Ozmultu, 2000; Belkin, 2003; Rieh & Xie, 2006). However, those measures have not been fully leveraged into the information systems (IS) and marketing research. In addition, prior IS and marketing studies mainly rely on traditional econometrics models to explore online consumer behavior. Despite their historical and conceptual importance, linear regression models often perform poorly relative to newer predictive modeling approaches from the machine-learning literature (Pearson, 2016). "Large dataset may allow for more flexible relationships than simple linear models" (Varian, 2014, pp. 3). Therefore, machine-learning techniques are recommended to reveal and investigate complex relationships.

To fill those gaps, in this study we borrowed metrics from the IR field and used a modern machine-learning technique, random forest, to construct a predictive model for online purchasing conversion rates. We also constructed a logistic regression model as a baseline model. As the result, random forest outperforms logistic regression in both accuracy and false negative rate. In addition, random forest reveals more nuanced and complicated impacts of predictors that are beyond linear and one-way relationship. The models developed in this paper have implications for online business stakeholders who need to deploy operational tactics to increase customer engagement, thus boosting revenue. The stakeholders include both the e-commerce managers and the external parties such as vendors and suppliers who are competing for customer attention online.

Literature Review

In recent decades, there has been a series of research addressing consumers' search depth and dynamics using the clickstream data from the ComScore database. For example, Johnson et al. (2004) use the data to characterize the search behavior at three levels: depth of search, dynamics of search, and activity of search. There are different opinions on the impact of modern search engines on consumers' search depth. Peterson and Merino (2003) believe that the availability of search tools will increase the amount of information and therefore increase the search depth. However, Holland and Mandry (2013) analyze a large amount of the Internet panel data from various e-commerce websites and conclude that the search depth in all sectors is significantly shallower than expected. As for search dynamics, empirical research on the temporal characteristics of search is sparse, perhaps due to the partial observation data of consumer search. Notable exceptions are De Los Santos et al. (2013) and Koulayev (2014) who explain that consumers revisit items that were searched previously because of learning or non-stationary search costs. This rich body of literature goes deeper into search behavior compared to those investigating online search benefits. However, depth and dynamics are just two high-level aspects of search behavior. More behavioral details at the action level, such as query formulation, scanning and viewing items, clickthrough, and dwell time, are expected in the research community. Missing those pieces might fail to capture the potential cues to infer conversion behavior.

In recent two years, there is a series of studies in business analytics that incorporated machine-learning models to predict customers' purchase intentions. Examples are da Silva's (2014) dissertation research that constructed four machine-learning models on a clickstream dataset; Fernandes' (2015) dissertation research that built a sequence model to predict real-time purchase likelihood. In addition, Vieira (2016)'s used deep learning algorithms to analyze purchase behavior. These studies have marked milestones in applying machine-learning techniques into the business analytics field. They have also collectively weaved a story about customers' interactions with an e-commerce website. However, due to data restrictions, the features or indicators used by them were limited and at an aggregate level, such as "number of page views for a customer during the last week". The derived insights were not detailed enough or easily translated into actions that could be done for the e-commerce websites. Our study complements those insights by working on the original customers' behavioral data at a single action level harvested by Walmart.com, the second largest e-commerce website in the U.S. The data is information rich and at a sufficiently detailed level to understand an individual customer's action sequence. The model will generate more actionable insights for those online business stakeholders.

Since the early 2000s, a series of IR studies has examined all kinds of aspects of search behavior metrics recorded in search logs (e.g., Belkin, 2003; Rieh & Xie, 2006). Table 1 below summarizes the widely used ones at different levels and categories in the IR research community. In this study, we aim to borrow these metrics and apply them into our model to predict the online conversion behavior.

Data Collection

The data used in this study is the weblogs from Walmart.com, a complex enterprise platform that consists of a series of components, like search/browse, catalog, store finder, gift registry, customers' account management, shopping cart, financial management, etc. The data in this study is from the search/browse component that tracks the customers' searching and browsing behavior prior to an item being added to a shopping cart. A brief introduction of the dataset is in the following section.

	Measure	Description	Example Studies
Query Actions	Query Length	The number of words in a query	Spink, Jansen, & Ozmultu, 2000; Belkin, 2003
	Number of Queries in a Task	The total number of issued queries for a search task	Spink, Jansen, & Ozmultu, 2000; Jansen &Spink, 2006
	Query Reformulation	The action of re-issuing another query based on an original query	Rieh & Xie, 2006; Jansen, Booth & Spink, 2009
	Query Abandonment	The action of giving up a query without any click	Das Sarma, Gollapudi, & Leong, 2008; Li, Huffman, & Tokuda, 2009
	Query Pagination	The action of clicking next page	Spink, Jansen, & Ozmultu, 2000; Wu, Kelly, & Sud, 2014
Click Actions	Click Depth	The deepest rank clicked on for a query	Wu, Kelly, & Sud, 2014; Niu, 2012
	Click Entropy	The diversity and messiness of clicks for a particular query	Mei & Church, 2008; Deng, King, & Lyu, 2009
	Number of Clickthrough in a Query	The total number of clicks for a query	Jansen &Spink, 2006; Deng, King, & Lyu, 2009
Time Engagement	Dwell Time	Time spent on viewing a result document, scanning a search results page, working on a query, or working on a whole search session.	Agichtein, Brill, & Dumais, 2006; Liu, White, & Dumais, 2010; Wu, Kelly, & Sud, 2014

Table 1. Metrics for online search

Data Processing

In August 2012, Walmart announced the Polaris search engine for its e-commerce website. Powered with a built-in data collection application, Walmart.com is able to periodically dump its server logs that represent customers' requests to the server and interactions with its site. The dumped data is in JSON (Javascript Object Notation) format. The dataset used in this study consists of 6,944,274 lines of records, each representing a search request sent to the U.S. based servers from 0:00 to 24:00 PDT on June 2, 2014. After the data preprocessing, we used Python scripts to extract the direct variables as well as to calculate those advanced variables, such as click entropy and session dwell time. Finally, we used R to perform model constructions and evaluations.

Variables

Based on the variables in Table 1 and considering availability and relevance to this study, we have extracted variables in multi-level (3 levels) and multi-category (4 categories), as summarized in Table 2. A search session is defined as a certain period of time (maximum of 30 minutes in this study according to the common practice of log analysis (Niu & Hemminger, 2015)) with the same cookie information during which a customer performs a series of actions.

	Query behavior	Click behavior	Time engagement	Search context
Click level		ClickPosition	PageDwellTime	
Query level	QueryLength	AvgClickPosition NumSearchResults NumClickQuery ClickEntropyQuery		HourOfDay
Session level	CurrentQueryPosition NumQuery	NumClickSession AvgClickEntropySession	SessionDwellTime	UserType Device

Note:

- **QueryLength**: number of words in a query
- **CurrentQueryPosition**: the position of the current query in the current session. For example, the current query might be the second query issued by a searcher in the same session
- **NumQuery**: total number of queries in the current session
- **ClickPosition**: the rank of the current click in the result list
- **AvgClickPosition**: the average rank of all the clicks in the result list under the current query
- **NumSearchResults**: total number of search results returned by the current query
- **NumClickQuery**: total number of clicks in the current query
- **NumClickSession**: total number of clicks in the current session

- **ClickEntropyQuery**: the click entropy for the current query
- **AvgClickEntropySession**: the average click entropy in the current session
- **PageDwellTime**: how much time spent on viewing an item page
- **SessionDwellTime:** how much time spent on the whole search session
- **UserType**: whether the customer was registered or not with Walmart.com
- **Device**: the device the customer was using to access the website. It has three values: desktop, tablet, and phone
- **HourOfDay**: the local hour of the day of accessing the website. It ranges from 0 to 23.

Table 2. Variables at three levels and in four categories

Most of these variables are very straightforward to understand. It is worth some explanation for *ClickEntropyQuery* and *AvgClickEntropySession*, which represent search ambiguity. After years of analyzing search logs, we have observed that queries are usually at different levels of ambiguity. For instance, "single-lens camera" is more ambiguous than "Canon 60D", and "camera" is more ambiguous than "single-lens camera". Search ambiguity may reflect consumers' specificity of search motivation. For example, casual visitors are more likely to search ambiguous terms than directed buyers. Queries that are more ambiguous may also require higher information processing efforts.

To capture query ambiguity, we apply a measure called click entropy, developed by Dou, Song, and Wen (2007), which calculates the variability in clicked results across individuals. The formula is:

$$ClickEntropy(q) = -\sum_{C \in URL(q)} p(c \mid q)\log_2(p(c \mid q))$$

Here *URL(q)* is the set of the URLs clicked for a query *q*, and $p(c \mid q)$ is the probability that URL *c* is clicked under a query *q*. For example, if 3 out of 8 clicked URLs under a query *q* are *c*, $p(c \mid q)$ will be 3/8 = 0.375. A high click entropy score means users click many different results under the same query, suggesting that the query is very ambiguous. On the other hand, a zero click entropy implies all the users click the same item under a query, suggesting that this query is very specific. In this study, we use *ClickEntropyQuery* to measure the click entropy for a particular query, and *AvgClickEntropySession* to measure the average click entropy for all the queries in a search session. From the initial scanning of the logs, we have been under the impression that the specific queries such as "Canon 60D" are more likely to lead to a purchase. Having the entropy variables included in the model, we want to use data analytics to back up this anecdotal observation.

Model Development

From our previous studies on search engines (e.g., hidden for the review purpose), search behaviors have profound impact on the information gathered. For the same token, customers' search behaviors may have complex relationships with their purchasing behavior. In machine-learning literature, tree-based models are usually better at predicting complex relationships because they are insensitive to the interactions of indicators and the distribution of indicators. Random forest is a common and popular tree-based model because it ensembles a large number of decision trees and takes the majority vote, thus largely reducing the error rate. We will introduce random forest and its implementation in R in the next section. Meanwhile, a logistic regression model will also be constructed as the baseline model. The model's target is to predict customers' conversion, treated as a binary decision where 1 means buying and 0 means not buying.

Random Forest

Random forest (RF, Breiman, 2001; Breiman et. al., 1984) consists of a sequence of classification trees grown by randomly selecting several variables from the variable list at each node to split. The splitting criterion is the maximum heterogeneity reduction in the target variable. This procedure is used together with bagging (Breiman, 1996), which is the random selection of a subsample from the original training set at each tree. The RF prediction is the majority vote of the tree predictions for the target variable, computed by passing down each tree only the observations that did not contribute to the model construction (out-of-bag predictions). The RF model also returns variable importance metric (VIMs) that is used to identify the most influential predictors (Breiman, 2002). In this study, we used the *R* package *snowfall* (Knaus, 2010) to implement the procedures mentioned above to evaluate variable importance.

The performance of the model on our target variable can be described by the "confusion matrix", a squared contingency table with m rows (the categories predicted by the model) and m columns (the true categories observed in the sample). Many measures have been proposed based on this confusion matrix. Three common measures are the accuracy, sensitivity (true positive recognition rate), and specificity (true negative recognition rate). Having this background knowledge, we will apply random forest to our dataset, and evaluate the model performance against the logistic regression technique.

Results

Of the 1,530,738 transaction records, only 71,159 (4.9%) are the conversion cases and the remaining are the non-conversion records. Both the descriptive statistics and the model construction results will be reported in below.

Descriptive statistics of customers shopping with Walmart.com

On average, shoppers formulated 8.3 queries and stayed around 3 minutes (189.3 seconds) on an item page. Although the number of the clicks under a particular query is 1.92, the average click rank is 5.36, deeper than the depths reported with general search engines (Song, Ma, Wang, & Wang 2013) and therefore suggesting a higher search depth. The average click entropy is 3.45 for a query and 3.55 for a search session respectively, meaning there are 8 ($-\log_2 1/8$ =3) to 16 ($-\log_2 1/16$ = 4) distinct clicks under the same query. The click diversity speaks to the fact that most queries have a moderate degree of ambiguity.

Random forest model results

As common practice in machine-learning, we randomly sampled 80% of the dataset as the training set and the remaining 20% as the test set. The unbalanced distribution (95.1% vs. 4.9%) of the independent variable is a potential problem for training the model, since most common classification algorithms would minimize the overall error rate rather than paying special attention to the minority class (Chawla, 2005). In this study, we adopted the method of under-sampling the majority class (He & Garcia, 2009) to make both types of cases roughly balanced. Then we used the balanced dataset as the training dataset. The test set remains untouched.

During the RF construction process, we included the 15 predictor variables in Table 2 and 1 target binary variable with two levels (buying and not buying). We performed the tuning and training of the RF model using the *R* package *caret*. The only parameter that needs turning is *mtry*, the number of variables randomly selected at each split. In this study, *mtry* was set to vary from 1 to 4 ($\approx \sqrt{15}$). To provide more accurate prediction, we employed a 10-folds cross-validation with 15 repetitions. Table 3 summarizes the modeling results for all the values of *mtry*. The accuracy reaches the peak (0.7008) when *mtry* was 3, suggesting the model with 3 random variables selected at each split performed the best.

Mtry	Accuracy	Accuracy SD
1	0.6805	0.0050
2	0.6998	0.0049
3	**0.7008**	**0.0049**
4	0.6998	0.0048

Table 3. Random forest result

The high accuracy rate (0.7008) confirms our assumption that using customers' behavioral variables and constructing a tree-based machine-learning model is able to predict buying behavior at a highly accurate level. However, the model accuracy has provided us a "black-box" view of the model performance. To penetrate into details, we need to evaluate the influence of the 15 variables in term of their predicting power for the conversion rate. Therefore, we applied the random forest algorithm with 3 variables randomly selected at each split (as picked from Table 3) and adopted the heuristic correction strategy for 100 iterations (Sandri & Zuccolotto, 2010). Gini VIM distributions over the 100 iterations for each variable were examined. The higher the Gini VIM is, the more influential the variable is in predicting customers' buying behavior. The top five influential variables are *PageDwellTime, UserType, AvgClickEntropySession, ClickEntropyQuery*, and *AvgClickPosition*. This finding implies that the amount of time a customer spends on an item page, as well as how diversified the user clicks are, play an important role in predicting customers' conversion. On the other hand, the five least important variables are *Device, QueryLength, HourOfDay, CurrentQueryPosition, and NumQuery*, which suggests query behavior and the search context in general do not factor much into the final purchase behavior of the customer. To our surprise, the device variable carried the least Gini VIM. This might be due to the highly skewed distribution of devices toward desktop computers (88%) in the sampled data. Since the Gini VIM threshold for keeping a variable into the predictive model is 0.5 (Sandri and Zuccolotto, 2010), all 15 variables are valid to be included in the model construction.

Logistic regression model results

The second model was built using the logistic regression technique. Since the 15 predictors are potentially correlated, both the VIF (Variance Inflation Factor) and the LASSO method were used to select variables to minimize the issue of variable multicollinearity. As the result, four variables (*AvgClickPosition, NumQuery, NumClickSession, SessionDwellTime*) were removed. The logistic regression model was constructed on the remaining 11 variables. The coefficients, the p-values, and the odds ratios are presented in Table 4. Although most predictors are significant at the .05 level, most of the odds ratios are close to 1, indicating small effect size. One variable with large odds ratio is *UserType*. Compared to unregistered users, registered users were 2.28 times more likely to purchase things they were looking at.

Model Evaluations

Each of the two models was run on the test set to evaluate their performance. The accuracy, sensitivity, and specificity are listed in Table 5. As the result, the random forest model outperforms the logistic regression in terms of all the metrics. Since the accuracy of the RF model is 0.76, we conclude that, using search behavioral as cues, we are able to predict the customers' conversion decisions at a very high accuracy level. If the cases are broken down into positive and negative ones, the confusion matrices for both models are listed in Table 6. Since in the e-commerce context, false negative, mistaking the buyers for non-buyers, is more undesired than the false positive, a better model should also maintain a low false negative rate. From Table 6, the false negative rates for the random forest model and the logistic regression model are 18.2% and 39.7% respectively. From this aspect, random forest is also better compared to the logistic regression model.

Variable	Coefficient	p value	Odds Ratio
QueryLength	-0.053***	<.0001	0.95
CurrentQueryPosition	0.000	.4340	1.00
ClickPosition	0.017***	<.0001	1.02
NumSearchResults	0.000**	.0296	1.00
NumClickQuery	-0.008**	.0291	0.99
ClickEntropyQuery	-0.099***	<.0001	0.91
AvgClickEntropySession	-0.082***	<.0001	0.92
PageDwellTime	0.000***	<.0001	1.00
HourOfDay	-0.022***	<.0001	0.98
UserType (Registered)	0.822***	<.0001	2.28
Device(Cellphone)	-0.448***	<.0001	0.64
Device (Tablet)	-0.141***	<.0001	0.87

Significance levels: ** < 0.05, *** < 0.001

Table 4. Logistic regression result

Follow-Up on the Random Forest Model - Partial Dependencies

From the above section, we chose the RF model as our final predictive model for customers' conversion. An objection frequently leveled at these newer model types is difficulty of interpretation relative to linear regression models (Pearson, 2016). In linear regression, we can gain considerable insight into the structure and interpretation of the model by examining its coefficients. In random forest, there is no comparably simple parametric description available, making the interpretation of these models more difficult. To address this difficulty, Friedman (2001) proposed the use of partial dependence plots, to investigate the marginal effect of one variable on the outcome while we hold other variables constant. In this study, we implemented the partial dependence plots using a series of R commands. By plotting the partial dependence for the 15 variables, we have an understanding of the direct impact of each variable on the conversion decision.

Classifiers	Accuracy	Sensitivity	Specificity
Random Forest	0.76	0.73	0.82
Logistic Regression	0.61	0.61	0.60

Table 5. Performance metrics of the models

	Random Forest		Logistic Regression	
Prediction	Actual (Reference)		Actual (Reference)	
	True (buy)	False (not buy)	True (buy)	False (not buy)
True (buy)	135,621	50,614	93,402	59,990
False (not buy)	24,373	109,397	65,805	100,014

Table 6. Confusion matrices for the two models

As shown in Figure 1, for most variables, the conversion probability curve fluctuates when the variable takes different values. The fluctuation implies their relationship to conversion is more complex than a linear and one-way correlation. In most curves, we find one or several spikes at particular values. For example, the conversion probability increases dramatically when *PageDwellTime* increases from 0 and reaches the highest when *PageDwellTime* is around 50 seconds. The probability remains much lower and consistent after that and drops gradually after 150 seconds (2.5 minutes). A stay characterized as too short or too long on an item page reduces the purchase probability. When the click entropy at both the query and the session levels remain around 3, the conversion rate is the highest. Queries that are too specific or too broad both lower this rate by a significant amount. The purchase likelihood reaches a maximum when the customers have clicked 3 items and the average rank of the 3rd clicked item is 5. This is consistent with previous studies on the search depth (for example, Holland & Mandry, 2013) that number of suppliers in consideration prior to purchase is about 2 or 3.

Discussion

This research applies metrics in the IR field to study search behavior on Walmart.com. Generally speaking, shoppers' search engagement level is higher than that in a general search engine, such as Google, for which the querying behavior, clicking behavior, and time engagement are well documented by a rich body of literature (Spink, Jansen, & Ozmultu, 2000; Jansen & Spink, 2006; Jansen, Booth, & Spink, 2009). Shoppers on Walmart.com issue more queries, scanned deeper into the result rank, click on more items, and spend more time on a search than those using a general search engine. For an e-commerce website, most of the visitors search for items with the intention of buying. By nature they are deeper searchers than those who access Google for a quick look-up of a piece of information. In addition, an e-commerce website is product-based and serves as a product catalog for its customers. Customers need to navigate to the item pages to access the product descriptions, compare prices, and read reviews prior to the purchase decision. The process is more pipelined than most searches typically conducted on search engines, where information need can sometimes be satisfied by a snippet information on a search results page without landing onto that page. Although occasionally clicking on a lower ranked item on a search a depth that a customer can reach, a point beyond the first page is very rare in reality. In our study, the average click rank is 5.36—a point on the first results page. For e-commerce website evaluations, instead of comparing the ranking algorithms along a result list, it would make more sense for customers to just compare those top results. In most situations, suppliers on the search results page are just one or several clicks away. These few clicks might make a huge difference in competing for the customers' attention. For suppliers, being listed on the second results page implies being in an unreachable place by customers.

Page dwell time is found to have the strongest associations with the final buying decisions. The purchase probability reaches the highest when a person spends around 50 seconds on the item page. This could be served as a practical guideline for the placement of information on an item page. A worth of elements that needs around 1 minute to digest might be most appropriate for the conversion rate. Either too little or too much information may lose customers' attention. Too little information might be insufficient for customers to make a purchase decision whereas too much information may overload them. Other than the page dwell time, user type has been found to be strongly correlated with conversion. Walmart.com offers

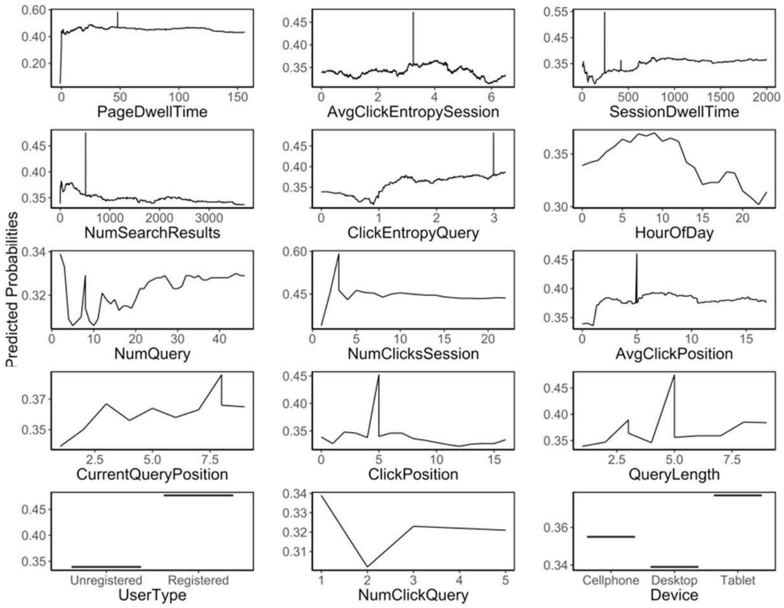

Figure 1. Partial dependence (marginal effect) of the 15 variables

the flexibility that a customer does not have to register an account prior to a purchase. However, it turns out that the purchase rate is much higher for the registered customers than those unregistered ones. An implication for the e-commerce websites is that certain marketing and promotion strategies should be taken to encourage all the customers to register before exploring the website. Click entropy has been found to be an important variable too. The purchase probability reaches the maximum when a query has a click entropy around 3, meaning 8 different clicks under this query. We may leverage some query assistance technique, such as query/term suggestions, query auto-completion, to assist people in broadening too specific or narrowing too broad queries into a desired level of specificity for the purpose of a higher probability of purchase. The next influential variable is the click position. The buying probability is the highest when a customer has clicked 3 items, and it drops dramatically after the 5th click. The finding supports the findings of previous studies (e.g., However, Holland, & Mandry, 2013) on the depth of consideration set prior to a purchase. It also agrees with the commonly accepted "three-click rule" (Zeldman, 2001) for more general website design, which suggests that a user of a website should be able to find any information with no more than three mouse clicks. Otherwise, they would become frustrated and leave the site.

Our study also identified those variables that do not carry much importance in predicting customers' conversion, such as the user device. One likely reason that the device being used by the customer was not found to be significant is due to the highly skewed data distribution across the three types of devices. The high skewness suggests that the majority of users still preferred traditional computers over mobile devices when visiting e-commerce websites to search for products, at least for the time point of 2014 when the data was collected. Other than the device, the length of query is not an important predictor because there is not much variation in this variable as evidenced by the small standard deviation (1.35) of the variable. Averagely speaking, customers type 2- or 3-word query into the search box. To our surprise, although we saw a peak volume of site visits after dinner time (7pm), the peak volume of purchase did not necessarily happen at that time period. The hour of the day does not impact the conversion rate much. In addition, the current query position in a search session and the total number of the queries issued by the customer for this search have a small correlation with the conversion rate too. This suggests that queries were "created equal" in a search session. Their order and the total number of them have roughly the same chances of leading to a purchase.

In this study, the RF model achieves a very high accuracy (76%). The high accuracy in this study implies that search behavior could be used as indicators to understand and predict customers' purchase decisions. This is a very significant finding in today's digital world where human behaviors are traceable with the assistance of various sophisticated technologies. The RF models are especially powerful when a predictive model needs to be constructed with a number of variables whose respective importance with regards to predicting the target variable is unknown. By taking advantage of a large number of "random sampling" and "repetition", random forest treats the modeling process as a black box and could achieve incredible prediction accuracy.

Conclusions

Despite the large amount of existing literature on various aspects of search behavior in the e-commerce context, few studies have successfully linked search behavior with conversion decisions. In this paper, we used 15 search behavioral variables to construct a random forest predictive model. The model achieved the prediction accuracy of 76% for online conversion. In addition, the RF model also provided a rank of variables according to their importance to the model construction. As the result, page and session dwell time, the user type, click entropy, and click position are the most important variables while devices, query length, hour of the day, and query position do not contribute much in the predictive model. In the follow-up analysis, we examined the marginal effect of each of the 15 variables on the conversion probability. For most variables, their relationship with the conversion probability is non-linear and more complicated than a one-way correlation. The findings suggest certain strategies for e-commerce stakeholders such as optimizing the amount of information provided on item pages, encouraging users to register, offering query formulation/reformulation assistance, and placing the desired items within the top 3 or 5 in the result list. Those strategies would potentially lead to a higher chance of purchase, and therefore the higher conversion rates for customers from casual site visitors to serious buyers.

There are several limitations in the work. One major limitation lies in the nature of server log analysis. The logged data does not capture the whole picture of users' behavior because it misses the requests cached on the local machine or proxy servers. In addition, the session-level analysis depends on identification of the session boundaries, which is impossible to be precise without applications to track when sessions begin and end. Our identification in this study is a best possible estimate based on literature and experience. The analysis based on logs also misses much context information to understand customers' motivations, feelings, and particular reasons for some unusual behavior. In the future, we need to combine server log analysis with more client-side data analysis and user studies to obtain more comprehensive insights about customers' search behavior and purchase decisions.

References

Agichtein, E., Brill, E., and Dumais, S. 2006. "Improving web search ranking by incorporating user behavior information," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19–26.

Belkin, N. J., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., Yuan, X.-J., and Cool, C. 2003. "Query length in interactive information retrieval," in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 205–212.

Breiman, L. 1996. "Bagging predictors," Machine learning, (24:2), Springer, pp. 123–140.

Breiman, L. 2002. "Manual on setting up, using, and understanding random forests v3. 1," Statistics Department University of California Berkeley, CA, USA, (1).

Breiman, L. 2001. "Random forests," Machine learning, (45:1), Springer, pp. 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. "Classification and regression trees. Wadsworth & Brooks," Monterey, CA.

Brynjolfsson, E., and Saunders, A. 2009. Wired for innovation: how information technology is reshaping the economy, MIT Press.

Chawla, N. V. 2005. "Data mining for imbalanced datasets: An overview," in Data mining and knowledge discovery handbook, Springer, pp. 853–867.

Das Sarma, A., Gollapudi, S., and Ieong, S. 2008. "Bypass rates: reducing query abandonment using negative inferences," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 177–185.

da Silva, J. M. M. 2014. "The Road to Enlightenment: Generating Insight and Predicting Consumer Actions in Digital Markets," Dissertation Research Submitted to University of Porto.

Deng, H., King, I., and Lyu, M. R. 2009. "Entropy-biased models for query representation on the click graph," in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 339–346.

Dou, Z., Song, R., and Wen, J.-R. 2007. “A large-scale evaluation and analysis of personalized search strategies,” in Proceedings of the 16th international conference on World Wide Web, pp. 581–590.

E-Marketer. 2014. “Global B2C Ecommerce Sales to Hit $1.5 Trillion This Year Driven by Growth in Emerging Markets - eMarketer.,”

Fernandes, R. F., & Teixeira, C. M. 2015. “Using clickstream data to analyze online purchase intentions”, Dissertation Research Submitted to University of Porto.

Friedman, J. H. 2001. “Greedy function approximation: a gradient boosting machine,” Annals of statistics, JSTOR, pp. 1189–1232.

He, H., and Garcia, E. A. 2009. “Learning from imbalanced data,” IEEE Transactions on knowledge and data engineering, (21:9), IEEE, pp. 1263–1284.

Holland, C. P., and Mandry, G. D. 2013. “Online search and buying behaviour in consumer markets,” in 46th Hawaii International Conference on System Sciences (HICSS), 2013, pp. 2918–2927.

Jansen, B. J., Booth, D., and Spink, A. 2009. “Predicting query reformulation during web searching,” in CHI’09 Extended Abstracts on Human Factors in Computing Systems, pp. 3907–3912.

Jansen, B. J., and Spink, A. 2006. “How are we searching the World Wide Web? A comparison of nine search engine transaction logs,” Information processing & management, (42:1), Elsevier, pp. 248–263.

Johnson, E. J., Moe, W. W., Fader, P. S., Bellman, S., and Lohse, G. L. 2004. “On the depth and dynamics of online search behavior,” Management science, (50:3), INFORMS, pp. 299–308.

Kamvar, M., Kellar, M., Patel, R., and Xu, Y. 2009. “Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices,” in Proceedings of the 18th international conference on World wide web, pp. 801–810.

Knaus, J. 2010. “snowfall: Easier cluster computing (based on snow),” R package version, (1).

Li, J., Huffman, S., and Tokuda, A. 2009. “Good abandonment in mobile and PC internet search,” in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 43–50.

los Santos, B., Hortaçsu, A., and Wildenbeest, M. R. 2015. “Search with learning for differentiated products: Evidence from e-commerce,” Journal of Business & Economic Statistics, (just-accepted), Taylor & Francis, p. 0.

McKinsey&Company. 2011. “Measuring the value of search.,”

Mei, Q., and Church, K. 2008. “Entropy of search logs: how hard is search? with personalization? with backoff?,” in Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 45–54.

Niu, X., & Hemminger, B. 2015. “Analyzing the interaction patterns in a faceted search interface”, Journal of the Association for Information Science and Technology, 66(5), 1030-1047.

Pearson, R. 2016. “Interpreting Predictive Models Using Partial Dependence Plots.,” (available at https://cran.r-project.org/web/packages/datarobot/vignettes/PartialDependence.html).

Peterson, R. A., and Merino, M. C. 2003. “Consumer information search behavior and the Internet,” Psychology & Marketing, (20:2), Wiley Online Library, pp. 99–121.

Rieh, S. Y., and others. 2006. “Analysis of multiple query reformulations on the web: The interactive information retrieval context,” Information Processing & Management, (42:3), Elsevier, pp. 751–768.

Sandri, M., and Zuccolotto, P. 2010. “Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms,” Statistics and Computing, (20:4), Springer, pp. 393–407.

Song, Y., Ma, H., Wang, H., and Wang, K. 2013. “Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance,” in Proceedings of the 22nd international conference on World Wide Web, pp. 1201–1212.

Spink, A., Jansen, B. J., and Cenk Ozmultu, H. 2000. “Use of query reformulation and relevance feedback by Excite users,” Internet research, (10:4), MCB UP Ltd, pp. 317–328.

Varian, H. R. 2014. “Big data: New tricks for econometrics,” The Journal of Economic Perspectives, (28:2), American Economic Association, pp. 3–27.

Wu, W.-C., Kelly, D., and Sud, A. 2014. “Using information scent and need for cognition to understand online search behavior,” in Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 557–566.

Zedlman, J. 2001. “Taking your talent to the web: making the transition from graphic design to web design,” Indianapolis Indiana: New Riders.