Did We Get It Right? Predicting Query Performance in E-commerce Search

Rohan Kumar
Flipkart
rohankumar@flipkart.com

Mohit Kumar
Flipkart
k.mohit@flipkart.com

Neil Shah*
Carnegie Mellon University
neilshah@cs.cmu.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

ABSTRACT

In this paper, we address the problem of evaluating whether results served by an e-commerce search engine for a query are good or not. This is a critical question in evaluating any e-commerce search engine. While this question is traditionally answered using simple metrics like query click-through rate (CTR), we observe that in e-commerce search, such metrics can be misleading. Upon inspection, we find cases where CTR is high but the results are poor and vice versa. Similar cases exist for other metrics like time to click which are often also used for evaluating search engines.

We aim to learn the quality of the results served by the search engine based on *users' interactions with the results*. Although this problem has been studied in the web search context, this is the first study for e-commerce search, to the best of our knowledge. Despite certain commonalities with evaluating web search engines, there are several major differences such as underlying reasons for search failure, and availability of rich user interaction data with products (e.g. adding a product to the cart). We study *large-scale* user interaction logs from Flipkart's[1] search engine, analyze behavioral patterns and build models to classify queries based on user behavior signals. We demonstrate the feasibility and efficacy of such models in accurately predicting query performance. Our classifier is able to achieve an average AUC of 0.75 on a held-out test set.

KEYWORDS

Information Retrieval, Evaluation, Query Performance, e-commerce, mobile search behavior, implicit feedback

ACM Reference format:
Rohan Kumar, Mohit Kumar, Neil Shah[2], and Christos Faloutsos. 2018. Did We Get It Right? Predicting Query Performance in E-commerce Search. In *Proceedings of ACM SIGIR Workshop on eCommerce, Ann Arbor, Michigan, USA, July 2018 (SIGIR 2018 eCom)*, 7 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

[1]Flipkart is the largest e-commerce platform in India.
[2]Dr. Shah is now at Snap Inc.

SIGIR 2018 eCom, Ann Arbor, Michigan, USA
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

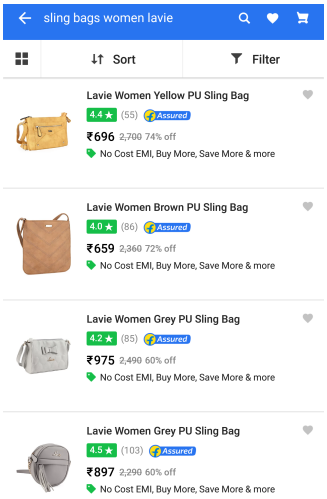


Figure 1: Mobile app e-commerce results page for the query "sling bags women lavie", showing relevant products.

1 INTRODUCTION

Search engines are a fundamental component of most modern Internet applications, and evaluating their performance on a query is not only needed for evaluating their overall performance, but is also critical in the iterative process of improving the algorithms that power them. This is important since bad performance of a search engine leads to customer attrition as described in White and Dumais [21]. Traditionally, the performance of a search engine on a query is measured using metrics derived from ordinal ratings of the search results given by human experts [4, 13, 23]. However, obtaining such manual judgments is prohibitive for the large document collections and high number of unique queries commonly encountered in most modern Internet applications. While one could solicit explicit feedback on the quality of search results from the users of a search engine, this may be detrimental to their experience of the application.

More recent work [8] has focused on automating the evaluation of search engine performance by using implicit feedback on the quality of search results derived from various user activity signals generated by the interactions between users and the results presented to them. Most of this work has been done for Internet search engines while in this paper, we focus on e-commerce search engines. The users of e-commerce applications tend to look for

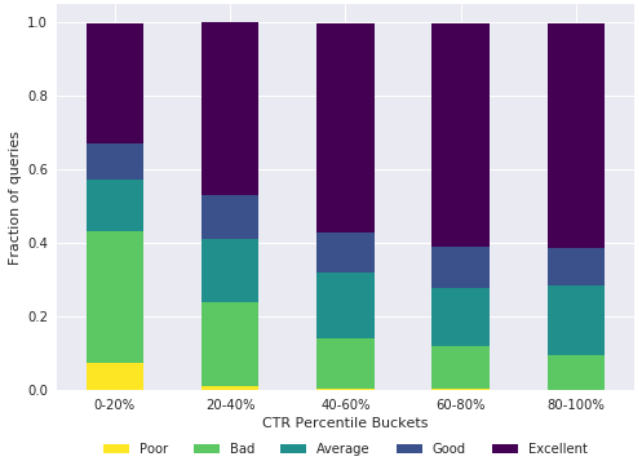


Figure 2: Distributions of search result ratings across percentile based CTR buckets.

products and services, and thus the queries typically encountered by e-commerce search engines are fundamentally different from the informational and navigational queries typically encountered by Internet search engines.

The most popular user activity signal in the aforementioned work is clicks and it is used to define the Click-Through Rate (CTR) metric. The CTR of a query is often used as a proxy for the performance of the search engine on that query, and this approximation is based on the assumption that clicks on search results are a reliable indicator of performance. However, Hassan et al. [10] points out that while clicks are a useful indicator of performance, they can nevertheless be quite noisy.

We validate this observation for e-commerce search by studying the distributions of the ordinal ratings of search results given by human experts to queries having a wide range of CTR values randomly sampled from Flipkart search query-logs. We discretized the CTR values into 5 buckets with the bucket boundaries at the 20th, 40th, 60th, and 80th percentiles of the CTR values of our sampled queries. The distributions of search result ratings across these percentile-based CTR buckets is shown in Figure 2. The details of how queries are sampled from our query-logs and how the associated search results are rated by human experts are given in Section 3.

From Figure 2 it is evident that while the fraction of queries whose results are rated as poor decreases as we go from the lowest CTR bucket to the highest CTR bucket, a significant fraction of queries whose results are rated as bad still exists even in the highest CTR bucket. Figure 1 shows an example of a search engine results page (SERP) that appears in Flipkart's mobile app for the query "sling bags women lavie". The query has good results even though it belongs to the 0-20% CTR bucket from Figure 2. This highlights the need for a richer set of user activity signals beyond click behavior. Guo et al. [8] made use of such signals, but their focus was on Internet search where the set of user activity signals available is limited in comparison to e-commerce search, where we have additional signals available such as the time taken to click an add-to-cart or buy-now button. Using a richer set of such user activity signals, we build a classification model to predict whether the results for any query from our query-logs would be rated as bad or good by human experts and thus automate the evaluation of our search engine performance. Such a system also serves as a first step towards building a system to predict user satisfaction at the level of individual user activity sessions as studied in [6, 9, 18].

Our classifier is able to achieve an average AUC of 0.75 on a held-out test set. On certain product categories like Mobile Phones, we achieve an average AUC of 0.88 on the held-out test set.

Summarily, the primary contributions of our work are:

(1) We identify a rich set of user activity signals that help predict whether the results for any search query would be rated as bad or good by human experts.
(2) We demonstrate that it is possible to use user activity signals to automate the evaluation of search engine performance for e-commerce applications.
(3) We analyze the performance of our classifier and derive insights into the effectiveness of automated systems for evaluating search engine performance that are of particular interest to e-commerce applications.

2 RELATED WORK

2.1 Query Performance

Evaluating search engine performance has been well-studied in the domain of web search. Topical relevance based metrics like nDCG [13], expected reciprocal rank [4] and weighted information gain [23] require explicit human labeled relevance judgments for query-document pairs which are prohibitively costly to calculate at scale for real-world web scale evaluation.

Several methods were proposed to automatically measure various characteristics of the documents retrieved for a query, which can then be used for measuring overall system performance. Clarity score [5] evaluates query performance by measuring the relative entropy between query language model and corresponding collection language model. The Robustness score [22] exploits the fact that query-level ranking robustness is correlated with retrieval performance. It is measured as the expected value of Spearman's rho between ranked lists from original collection and a corrupted collection. Carmel et al. [1] find Jensen-Shannon divergence between queries, relevant documents and the entire collection to be an indicator of query performance. However, [23] experimentally show the ineffectiveness of these metrics in measuring search performance on web-scale engines.

User click behavior has been used as an alternative to expert judgments for automatically tuning retrieval algorithms (predicting document relevance) as well as estimating IR evaluation metrics [3, 7, 8, 15]. Kim et al. [16] show that only analysing user clicks naively may not indicate satisfaction, but rather using dwell time per click appropriately indicates query level satisfaction in a better manner. Guo et al. [8] also make use of interaction features and engine switches as signals to predict DCG@3.

2.2 Search Session Performance

There has been considerable work in the area of analyzing user satisfaction at a session level rather than at an individual query

level. Fox et al. [6] conducted one of the first studies that found association between explicit ratings and implicit measures of user interest, concluding that user satisfaction can be predicted using such implicit signals. Hassan et al. [9] show empirically that user behavior alone can give an accurate picture of the success of the user's web search goals, without considering the relevance of the documents displayed. There have been studies focusing on graded satisfaction [14] as well as specific user behaviors like query reformulation [10, 18] and interaction sequences [17] for understanding satisfaction.

2.3 E-commerce Search Performance

Most studies have been geared towards web search where user search goals are different from those in product/e-commerce search. However, there has been some work recently in the context of product search. Singh et al. [19] study the user behavior in the e-commerce search context in a specific scenario when the search engine doesn't retrieve any results. [20] is a recent study that addresses the user's session satisfaction in product search. They approach the problem by firstly identifying a taxonomy of user intents while interacting with product search, and then analyze the user's behavior in the context of the defined taxonomy. They predict user session satisfaction by utilizing the interaction behavior, where they build separate models for different intents with the demonstration that user behavior is different under different intents. Our work, while building upon the learnings from these studies, differs in that we are interested in measuring only the aggregate query performance instead of more user-centric task of session satisfaction. The example mentioned by Su et al. [20] where the results expected by two different users for the same query *iphone* may be different and thus they may be individually dissatisfied even though the results shown are "relevant." We aim to address the simpler, albeit more business-critical problem of understanding a query's result relevance in a user-agnostic fashion. The underlying reason(s) for a search engine's poor query performance is due to factors like incorrect spell error handling, vocabulary gap [2], selection gap (when the e-commerce platform does not sell a particular item – e.g. *chocolate* when packaged food items are not sold), and more. Thus understanding and measuring the user-agnostic query performance can help improve the core relevance algorithm of the search engine.

3 QUERY PERFORMANCE JUDGEMENTS

At Flipkart, regular search quality analysis is done for a random sample of queries (stratified on query volume segment) from search logs by a team of quality experts. They are requested to rate queries on a five-point scale (PBAGE: Poor-1, Bad-2, Average-3, Good-4, Excellent-5) based on result relevance. To ensure the consistency of labeling across experts, inter-rater agreement is continuously monitored. In this work, we make use of the expert editorial judgments for the month of January 2018.

We selected 18,613 queries from this randomized set of expert-labeled queries which occurred more than 100 times in a week in order to ensure reasonable user activity data. This set of queries corresponded to 127M query impressions, 149M clicks and 14M other interactions (e.g. filters application, sort application) from activity by 21M users collectively spending almost 4M hours on the platform. The data is collected from Flipkart's mobile app, significantly reducing the chances of bot traffic. All user behavior data is captured for the same week in which the query was labeled by an expert. We assume the search system and hence user activity remain constant throughout the week as there are no manual or algorithmic fixes applied during the week.

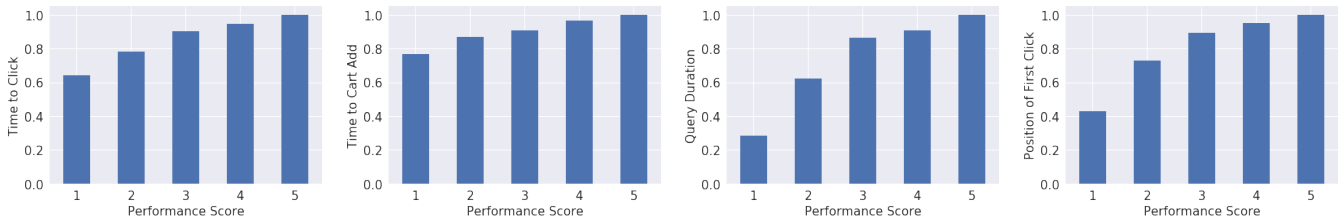
Table 1: Features used to distinguish between good and poorly performing search queries

Activity time	
timeToFirstClick	Time taken to click first product
timeToFirstCart	Time taken to add a product to the cart
queryDuration	Total dwell time of the query
Positional	
posFirstClick	Position of first product clicked
Activity aggregates	
numClicks	Number of clicks
numSwipes	Number of swipes
numCarts	Number of cart adds
numFilters	Number of times a filter was applied
numSorts	Number of times user changed sorting
numImpressions	Number of product impressions in the viewport
clickSuccess	Any product clicked for query
cartSuccess	Any product added to cart for query
Query text characteristics	
charQueryLen	Length of the query in characters
wordQueryLen	Length of the query in words
LMScore	Query language model perplexity score
querySim	Similarity to next query
containsSP	Query contains specifiers
containsMT	Query contains modifiers
containsRS	Query contains range specifiers
containsUnits	Query contains units like liters
Meta aspects	
queryCat	Category (mobile phones, books etc.) of the query based on taxonomy
queryType	Type of the query (specific product, broad category etc.)
queryCount	Frequency of the query
isAutoSuggestUsed	Auto-completed query or not
isGoodNetwork	Network type is WiFi or 4G
numProductsFound	Number of products matching the query

4 SIGNALS OF USER BEHAVIOR

Table 1 lists the metrics along with their descriptions that we extracted for every query instance. We characterize the user behavior metrics as *Activity time*, *Positional* and *Activity aggregates*. We characterize the non-user metrics as *Query text characteristics* and *Meta aspects*.

Activity time features capture the time taken by the user for various activities. timeToFirstClick is the time taken by the user to click a product after the results are displayes. timeToFirstCart is similar



(a) Time to first click **(b) Time to first cart** **(c) Query duration** **(d) First Click Position**

Figure 3: Normalized Distribution of Activity Time and Positional feature values with respect to Performance Score

to timeToFirstClick except it captures time taken to add a product to the cart. queryDuration is the total time spent in interacting with the query results including all interactions with product pages, cart etc. Figures 3a-3c show the distribution of *Activity time* features with respect to query performance. We observe interestingly that time taken for first click increases with the query performance. This is counter-intuitive in that when the query performance is good, it still takes users longer to click. This is however potentially explained with Figure 4a, which shows the distribution of the number of clicks against query performance. We observe that when the query performance is low the total number of clicks is lower and it increases with query performance. Intuitively, the users usually don't click any products when the query performance is poor but when they click products for a poorly performing query they do it faster. Similar pattern is observed for the add-to-cart behavior, in Figures 3b and 4c.

Positional features correspond to the position of result interaction. posFirstClick captures which position the user clicked first. A lower position value indicates that the results were shown near the top of the page. We observe that the average position of the first result click increases with improving query performance. This is correlated with the previous observation where time to first click of poorly performing queries is lower and correspondingly the user is clicking the results in lower positions (faster). The total number of clicks is low when query performance is low. Similar to Activity Time features, users usually don't click products when the query performance is poor but then they click products for a poorly performing query they do it at lower positions.

Activity aggregates features capture the aggregated summary of user's actions for a query. We observe that all the activity aggregates are positively correlated with the query performance – i.e. increasing user activity indicates better query performance. Number of product clicks (numClicks: Figure 4a), product swipes (numSwipes: Figure 4b), cart additions (numCarts: Figure 4c), filters applied (numFilters: Figure 4d), sort applied (numSorts: Figure 4e), product impressions per query (numImpressions: Figure 4f), query successful click through rate (clickSuccess: Figure 4g), query successful cart conversion rate (cartSuccess: Figure 4h) are all positively correlated with query performance.

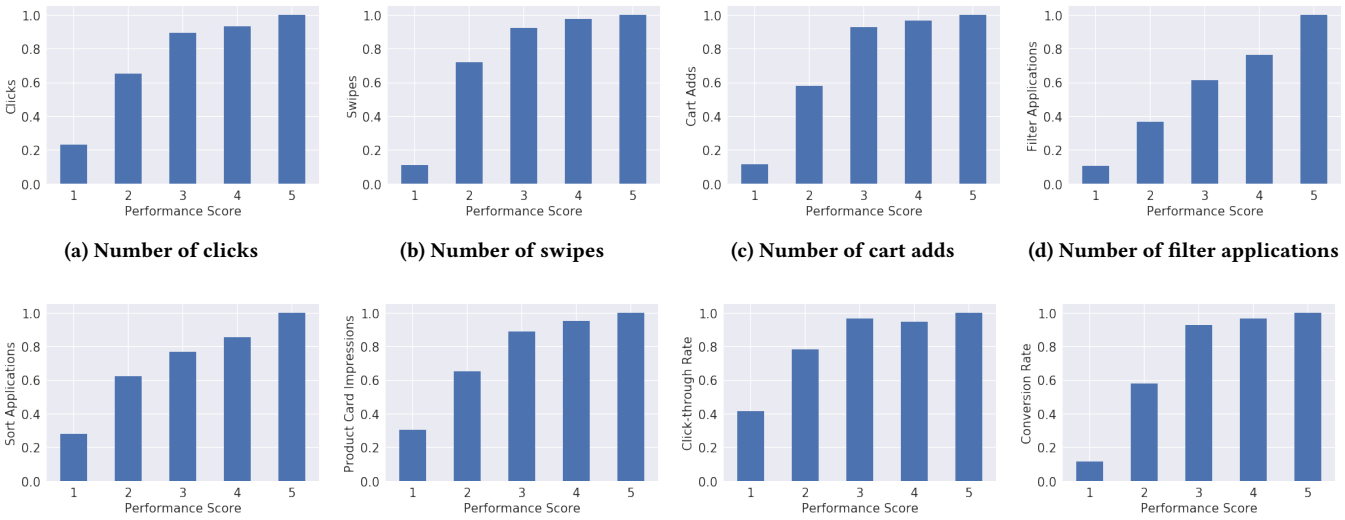
Query text characteristics features capture the textual properties of the query. charQueryLen and wordQueryLen are length of query in characters and words respectively. LMScore is the perplexity score of the query based on a language model[12] trained on the query logs. querySim is the text similarity between the current query and the following query defined by the measure described in Hassan et al. [11]. We also make use of certain domain-dependent text features indicating if the query contains specifiers (e.g. "greater than"), modifier phrases (e.g. "least expensive"), range specifiers (e.g. "between") or units (e.g. "liters", "gb"). The intuition here is that search engines may face difficulty in product retrieval when queries contain such phrases which require semantic understanding.

Meta aspect features include additional information about the query. queryCat indicates the e-commerce product category. These are broad lines of business, namely Mobile Phones, Books, Electronics, Lifestyle, and Home and Furniture. Each query is assumed to belong to one of these categories. The intuition for using this feature is that the query performance and user behavior might be dependent on the specific categories. queryType indicates the type of query which is classified amongst three kinds, namely "Product", "FacetCategory' and "Category." Queries in which the exact product that the user is looking for is mentioned are called "Product" queries (e.g. *iPhone X*). Queries which refer to a broad group of products are called "Category" queries (e.g. *shoes*). "FacetCategory" queries typically contain one or more attributes followed by a category (e.g. *red Nike shoes*). For both queryCat and queryType, we make use of modules which are able to assign appropriate values for a given query (details of these modules is outside the scope of this paper). queryCount is the total number of times the query was issued by users in the past week. isAutoSuggestUsed indicates whether the user selected the query from the suggested queries (auto-suggest). The intuition is that the queries suggested by the search engine typically perform better than query issued by user. isGoodNetwork indicates whether the user has a good Internet connection (defined as WiFi or LTE) while issuing the query. This is important, as the user experience and behavior might be altered if he/she doesn't have a good Internet connection leading to bad experience independent of the search engine's performance. numProductsFound indicates the total number of products found in the search index for the query. The intuition here is that the number of products found in conjunction with the type of the query may indicate if the search engine is not able to retrieve relevant results.

5 EXPERIMENTS

5.1 Experimental setup

In this work, we formulate the problem of query performance prediction as a binary classification task, as is done in [20]. As described in Section 3, we obtained expert judgments for 18,613 queries across a 5-point scale. Similar to [20], we label "poor," "bad" and "average" queries as DSAT and "good" and "excellent" as SAT.



(a) Number of clicks **(b) Number of swipes** **(c) Number of cart adds** **(d) Number of filter applications**

(e) Number of sort applications **(f) Number of product card impressions** **(g) Click-through rate** **(h) Conversion rate**

Figure 4: Normalized Distribution of Activity Aggregates with respect to Performance Score

This results in 6,949 DSAT and 11,664 SAT queries. We treat DSAT as the positive class (classifier target) as the interventions in future based on the model's prediction will be for this class.

We aggregate the metrics, described in previous section, across all the instances of the query in the week to obtain aggregate user behavior corresponding to the query. For metrics which may not have values for all query instances (e.g. timeToFirstClick), we only include instances for which values are present, in the aggregate calculation. These aggregate metrics are used as features for the classification model. We experiment with various descriptive statistics for the features, namely, average, median, standard deviation, inter-quartile range[5]. We bin each numeric feature into 10 percentile buckets and convert them to one-hot encoded features. We also defined certain interaction features such as clickSuccess × queryCount.

We split the labeled data into 80% training and 20% test set. During training, we performed feature selection using recursive feature elimination along with model hyper-parameter tuning. The hyper-parameter tuning is done using five-fold cross validation with class-stratification and optimized for area under the ROC curve (AUC).

5.2 Results

We analyze the results of our model along the following aspects: performance of learnt classifier, feature importance, performance across e-commerce categories, performance across query types and performance across query volume. We use AUC to evaluate the prediction performance.

5.2.1 Performance of classifier. We train a binary random forest model based on the methodology described earlier in section 5.1. Figure 5 shows the AUC curve and Figure 6 shows the PR curve. The overall test AUC obtained is 0.75. We observe that the classifier is able to achieve a reasonably good performance, thus establishing that it is feasible to predict query performance based on user interaction signals.

One application of this predictive model is to enable automated interventions for unsatisfactory queries i.e. when the classifier is confident that the results are poor, we can enable certain interventions like triggering an interactive intent solicitation module. Towards that end, we need a reasonably high precision operating point. Based on discussion with business/product team, the operating point that can be used is 85% precision where we will be able to achieve 20% recall with the current model.

5.2.2 Feature importance. Given below is the list of top-10 most important features based on Gini index:

(1) numSwipes
(2) clickSuccess
(3) queryType
(4) wordQueryLen
(5) numProductsFound
(6) cartSuccess
(7) numFilters
(8) numClicks
(9) numSorts
(10) queryCount

We observe a mix of features from various groups in the top features. It is interesting to see the number of page-to-page swipes as a very indicative feature of query performance. We conjecture that the users tend to click and swipe more in exploratory searches when they are satisfied with the initial results and want to continue exploring in the same set without reformulation. As expected, clickSuccess, cartSuccess and numClicks are indicative of query performance. queryType in conjunction with numProductsFound is a good indicator where we expect a small number of products

[5]In all the figures above, we show qualitative analysis of the features with only the "averaged" metric which sufficiently indicates the patterns.

Table 2: Prediction performance for different product categories

Product Category	AUC
Books	0.70
Electronics	0.74
Home And Furniture	0.72
Lifestyle	0.70
Mobile Phones	0.90

Table 3: Prediction performance query types

Query Type	AUC
Category	0.74
Facet Category	0.72
Product	0.87

for "Product" queries and larger number of products for "Category" queries. Interestingly numFilters and numSorts which indicate further refinement of results are also indicative of query performance, where based on Figures 4d and 4e we observe positive correlation with query performance. One surprising observation is that none of the *Activity Time* features are amongst the top 10 features; even though they are indicative, they are less indicative than other structured features like filters and sorts applied.

5.2.3 Performance across categories. Table 2 shows the performance of the model across the e-commerce categories (described in Section 4). We observe that the model is able to predict the query performance in "Mobile" categories considerably better than all other categories. We conjecture this is due to model's performance across query types (detailed below in section 5.2.4). The "Mobile" category has 7x more "Product" queries compared to the "Lifestyle" category. Additionally, "Mobile" category has 3x less "Facet Category" queries. The model is able to perform much better for 'Mobile' category due to the underlying query type distribution which is biased towards "Product" queries. This is fairly important from a business perspective as the "Mobile" category contributes to a significant portion of overall sales.

5.2.4 Performance across query types. Table 3 shows the results across query types. There are three query types, namely, "Product," "Facet Category" and "Category" as discussed in Section 4.

We observe that performance of "Product" queries, where the user's intent and language is very specific, is significantly better than other query types. We conjecture that indicators like numProductsFound and numClicks are particularly indicative of the query performance for "Product" queries.

5.2.5 Performance across query volume segments. Queries are categorized into three segments based on weekly volume: Head, TorsoHigh and TorsoLow. Table 4 shows that classifier performance improves as the volume increases. The average queryCount for queries belonging to the Head segment is about 34x that of queries belonging to TorsoBottom segment. Despite the huge difference in amount of data available per query, the classifier is able to predict performance for queries in all three segments reasonably well.

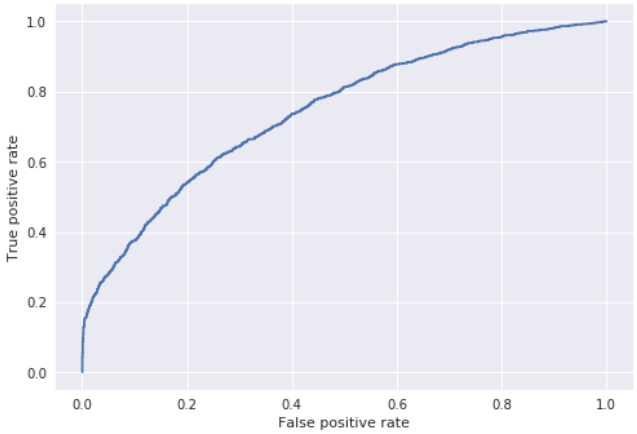


Figure 5: Receiver Operating Characteristic Curve for Binary Classification of Query Performance

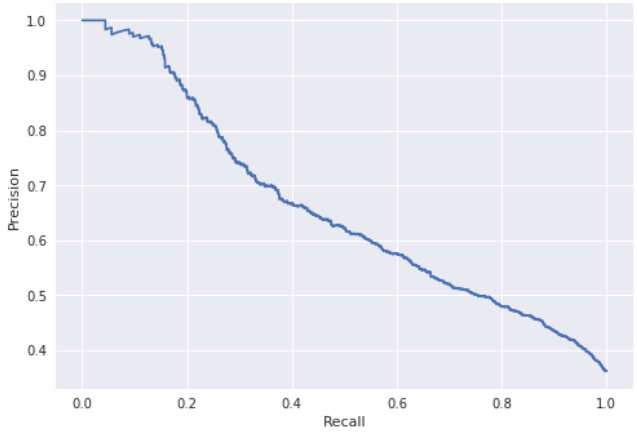


Figure 6: Precision Recall Curve for Binary Classification of Query Performance

Table 4: Prediction performance for different query volume segments

Volume Segment	AUC
Head	0.76
TorsoHigh	0.75
TorsoBottom	0.72

6 CONCLUSION AND FUTURE WORK

Measuring search engine performance is essential to building and improving retrieval algorithms. Query performance evaluation allows for a fine-grained measure of search performance. CTR can be a noisy metric in that high CTR queries may still have poor performance, and vice versa. A more sophisticated analysis of search behavior is needed to distinguish poor and well performing queries. In this work, we successfully demonstrate that query performance can be predicted based on user's interaction with the result set. This

is the first study to our knowledge that has collectively defined these signals in the context of query performance prediction for e-commerce search. Specifically, we propose and use several user interaction signals that help characterize query performance and enabled us to achieve good classification performance using these signals. Notably, our model achieved an overall AUC of 0.75 in the binary SAT/DSAT prediction task. We have analyzed the results across various factors like category of the query, query type and query volume. Key takeaways from the performance analysis are (a) We achieve significantly higher AUC of 0.90 on certain categories like "Mobile" making the result very promising from business impact perspective, (b) Classifier performance varies across query types ("Product", "Facet Category" and "Category") and is best for "Product" queries, and (c) Classifier performance improves with engagement volume, and is better for Head queries than TorsoBottom queries.

Future Work The study can be extended to have a finer prediction target of issue type like spell error, vocabulary gap, selection gap etc. which would make the classifier prediction more easily actionable by giving finer details on the query. Even richer signals of user activities can be used for prediction. For example, the notion of good dwell time (healthy engagement such as reading or voting on reviews) and bad dwell time (unhealthy engagement such as changing seller) might be used. Reducing the number of observations required (currently set to 100) for robustly predicting query performance would be another avenue of future work. This would allow the classifier to scale an even larger number of queries which do not have many instances in a fixed time period.

7 ACKNOWLEDGEMENTS

We thank Mr. Priyank Patel and Mr. Subhadeep Maji for their helpful comments.

REFERENCES

[1] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. 2006. What makes a query difficult?. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 390–397.

[2] Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)* 44, 1 (2012), 1.

[3] Ben Carterette and Rosie Jones. 2008. Evaluating search engines by modeling the relationship between relevance and clicks. In *Advances in Neural Information Processing Systems*. 217–224.

[4] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 621–630.

[5] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 299–306.

[6] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.

[7] Fan Guo and Chao Liu. 2009. Statistical Models for Web Search Click Log Analysis. In *Tutorial at the 19Th ACM International Conference on Information & Knowledge Management (CIKM '09)*. ACM.

[8] Qi Guo, Ryen W White, Susan T Dumais, Jue Wang, and Blake Anderson. 2010. Predicting query performance using query, result, and user interaction features. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 198–201.

[9] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: User Behavior As a Predictor of a Successful Search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/1718487.1718515

[10] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond Clicks: Query Reformulation As a Predictor of Search Satisfaction. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 2019–2028. https://doi.org/10.1145/2505515.2505682

[11] Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. 2014. Struggling or exploring?: disambiguating long search sessions. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 53–62.

[12] Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 187–197.

[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[14] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W. White. 2015. Understanding and Predicting Graded Search Satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. ACM, New York, NY, USA, 57–66. https://doi.org/10.1145/2684822.2685319

[15] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 154–161.

[16] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Modeling Dwell Time to Predict Click-level Satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 193–202. https://doi.org/10.1145/2556195.2556220

[17] Rishabh Mehrotra, Imed Zitouni, Ahmed Hassan Awadallah, Ahmed El Kholy, and Madian Khabsa. 2017. User Interaction Sequences for Search Satisfaction Prediction. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 165–174. https://doi.org/10.1145/3077136.3080833

[18] Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. 2015. Struggling and Success in Web Search. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 1551–1560. https://doi.org/10.1145/2806416.2806488

[19] Gyanit Singh, Nish Parikh, and Neel Sundaresn. 2011. User Behavior in Zero-recall Ecommerce Queries. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. ACM, New York, NY, USA, 75–84. https://doi.org/10.1145/2009916.2009930

[20] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 547–555. https://doi.org/10.1145/3159652.3159714

[21] Ryen W White and Susan T Dumais. 2009. Characterizing and predicting search engine switching behavior. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 87–96.

[22] Yun Zhou and W Bruce Croft. 2006. Ranking robustness: a novel framework to predict query performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, 567–574.

[23] Yun Zhou and W Bruce Croft. 2007. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 543–550.