ID Cover Page

Summary of WP Student Team

Drivers and prediction of organic search engine CTR - The effects of SERP features and their business implications

Group constitution:

Student Name	Program	Individual Title
Niclas Groll	Business Analytics	Drivers and prediction of organic search engine CTR - The effects of SERP features and their business implications

Work project carried out under the supervision of:

Advisor: Qiwei Han

Note: this template is a support document for the jury members before the defense, only.

A Work Project, presented as part of the requirements for the Award of a Master's degree in

Business Analytics from the Nova School of Business and Economics.

Drivers and prediction of organic search engine CTR

- The effects of SERP features and their business implications (48699)

NICLAS GROLL (48699)

Work project carried out under the supervision of:

Qiwei Han, PhD

25-01-2023

Abstract

33% of web traffic in the $5.7 trillion e-commerce industry originates from organic search engine results. Thus, website providers benefit from a holistic understanding of the drivers of click-through rates (CTR) on organic searches to increase traffic. However, providers face a knowledge gap, as existing literature focuses on position as the primary CTR influence, disregarding other result page characteristics. To solve this problem, we use an extensive dataset comprising organic Google result page information. We conduct an elaborate data analysis highlighting the impact of four categories of result page characteristics before determining suitable CTR prediction modeling techniques. We discover novel patterns impacting CTR for each category and find tree-based models to outperform state-of-the-art deep-learning models.

Furthermore, we reveal particular SERP feature effects on CTR and highlight their possible business implications.

Keywords

organic click-through rate, CTR prediction, e-commerce, SERP features

Acknowledgements

We are very grateful for the continued support of our supervisors, Qiwei Han and Maximilian Kaiser. Additionally, we highly appreciate Grips enabling us to conduct this work by providing the extensive data set used.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Introduction

Over the past years, the e-commerce business has seen extraordinary growth, with the Covid-19 pandemic additionally boosting its expansion. In 2022, Statista estimates global e-commerce sales at 5.7 trillion US-Dollars (Chevalier 2022a) which exceeds the 2021 GDP of the world's third-largest economy, Japan (O'Neill 2022). With this immense potential, online businesses are flocking the market, trying to gain a share of the steadily growing pie of e-commerce revenue. This revenue is created through sales on a website leading to the ultimate challenge of generating as much traffic on a website as possible. A prominent way to accomplish this is search engine optimization (SEO), an online marketing strategy with the goal of achieving the highest possible traffic for a website (Ledford 2009).

For e-commerce, 33% of the total web traffic is estimated to originate from organic traffic (Chevalier 2022b). Organic traffic refers to cases in which a user searches a keyword in a search engine and arrives at a website by clicking on one of the non-ad results. Because of the high relevance of this organic traffic, it is critical for website providers to improve their chances of generating clicks through search engines. In general, the clicks to a website on a search result page for a specific keyword can be formulated as

$$(1)\quad Clicks = Impressions \; * \; CTR$$

where CTR refers to the click-through-rate of a result (Google 2022a). Accordingly, to achieve higher web traffic, SEO practitioners want to increase at least one of both factors. While the impressions for any keyword can be approximated with the search volume available (Semrush 2022a), the CTR is only available to a website owner for keywords they already rank for. Thus, website providers can easily understand which other keywords are most promising to rank for but usually have very little insight into how much of the volume will convert into actual traffic to their website. Providing an understanding of what influences CTR can overcome this hurdle and enable companies to receive more website traffic by achieving a higher CTR.

To increase CTR, SEO efforts tend to focus on the position of a website on a result page, as top-ranked websites receive more user attention and clicks (Lewandowski, Sünkler, and Yagci 2020). To rank high on a result page, a website needs to be considered as relevant for the particular keyword. To achieve this, SEO practitioners concentrate their efforts on matching a website's content and metadata to the keyword. (Cui and Hu 2011; Shih et al. 2013; Bala and Verma 2018; Ziakis et al. 2019; Das 2021; Olson et al. 2021)
This approach of focusing only on the ranking of a website on a result page to increase CTR appears one-dimensional, as it ignores how other characteristics of a search page directly influence user behavior leading to the desired clicks.
These characteristics can be grouped into four categories potentially influencing CTR:

1. The *position* of a result on a particular result page.
2. K*eyword* characteristics: These characteristics can relate to a keyword itself (e.g. length or content), the intent of the search, or the accompanying metadata about the result page, such as the search volume or difficulty of ranking high for a keyword.
3. Characteristics of a *result:* Such include URL, title, or description of a result.
4. *SERP features:* These are visual elements containing information from organic search results and aim at making Google result pages more engaging (Google 2022b). These increasingly prominent elements of a search engine result page (SERP) can, for example, take the form of an instant answer to a user's question, a knowledge panel on the right of the result page, or an image next to a result. Appendix I shows an illustration of SERP features.

The categories demonstrate that apart from the position, a website provider can influence various variables to gain users' attention. In order to do so and accurately predict CTR, however, one first needs to understand the characteristics' effects and dynamics. This opens two research questions:

RQ1: How do these characteristics influence CTR, and how do they impact the importance of a website's position?

RQ2: Which result page characteristics and machine learning models can be used to accurately predict the organic CTR of a website on a Google result page?

In this work, we extend the existing analysis of position as the main direct influence on CTR to additional keyword, result, and SERP feature characteristics summarized in a unique dataset. An analysis including these characteristics provides a novelty in CTR research. In addition to that, we will identify practical CTR prediction model techniques to apply to this data. This enables ex-ante evaluation of the effectiveness of SEO efforts by estimating expected website traffic through predicted CTR and commonly known impressions.

In the following, the second chapter will provide an introductory analysis of related publications on CTR influences and CTR prediction models. The third chapter introduces the used dataset and applied methodology in detail. After presenting the findings of an exploratory data analysis in chapter 4, chapter 5 examines which CTR drivers have the most predictive power and what models perform best for the prediction of CTR.

2. Related works

When raising the question of how to predict CTR and which features impact it, one is mainly looking at two related research fields. First, current research into drivers of organic CTR, and second, recent research on models and data used to predict CTR. Therefore, this section will provide a high-level literature analysis of the former.

2.1 Research into drivers of organic CTR

Search engine optimization (SEO) is arguably one of the most wide-ranging online advertising strategies today (Olson et al. 2021; Kingsnorth 2022). According to the Oxford dictionary, search engine optimization is defined as ‘the process of maximizing the number of visitors to a particular website by ensuring that the site appears high on the list of results returned by a search engine’ (Stevenson 2010). The definition stresses the direct relationship between traffic and a website's position on result pages. This importance of position is also expressed by the vast majority of research on SEO marketing. (Cui and Hu 2011; Shih et al. 2013; Bala and Verma 2018; Ziakis et al. 2019; Das 2021; Olson et al. 2021). SEO achieves higher ranks by precisely tailoring a website’s content to a given keyword, for example, by adding new and qualitative blog posts with the aim to appear more relevant to search engines (Das 2021). Other aspects such as keyword, result, and SERP feature characteristics, if considered, are only analyzed regarding their influence on the position, but not their direct influence on CTR.

To conclude, previous research sees the position as the primary driver of CTR and the most effective way to gain more user attention. We contribute to this perspective by introducing additional characteristics of result, keyword, and SERP features to the analysis of the directly influencing factors of CTR. In the following, we put a special emphasis on SERP features as they are one of the main visual elements that users perceive on a result page, next to organic and paid search results. We will do so by first outlining how they have been described in related works and to what criticism they are linked in public.

While the literature focuses on a website's rank, SERP features are receiving relatively little attention, with little literature researching isolated SERP features. An example is Sam-Martins (2020) analysis of the *featured snippet,* which in turn relies on information found on blogs from influential SEO companies like Semrush. While there is little research on the impact of SERP features, there are a number of influential marketing blogs that advertise the importance of appearing in SERP features (Moz 2022; Semrush 2022b; Wheelhouse 2022). Most blogs argue that SERP features benefit websites featured in them and harm the CTR of websites that appear next to SERP features without being featured (Moz 2022, Wheelhouse 2022). Google itself advertises significant improvements in click-through rate, visits and time spent on websites when chosen to be shown in SERP features (Google 2022b). Critics, however, state that by adding additional features to its result pages' and leveraging information originally published on third-party websites, Google reduces reasons to leave its ecosystem (Tober 2022). This can harm the publishers of the original information that rely on traffic to their websites. Zero-click searches, which are searches where a user does not leave Google's ecosystem, amount to 25.6% of all searches, according to a non-academic study (Tober 2022). This shows that by analyzing the effect of SERP features on CTR, this work has the potential to shed light on the intransparent role of SERP features in Google's search engine ecosystem and provide first empirical research on the SERP features' impact on website performance.

2.2 CTR prediction research

This chapter discusses achievements in previous research on CTR prediction and points out the difference in approaches between previous publications and this work. These differences lie in the nature of the prediction, i.e. whether it is a classification of clicks from individual users or a regression for aggregated CTR, and in the nature of clicks, i.e. whether they are organic or advertisements.

The problem formulation addressed by the vast majority of CTR research is: “Will user X click item Y?”. More technically, the underlying problem is a binary classification problem, i.e. predicting yes or no. The item in question could be an ad on a search engine result page but also an article in an online shop (Yang and Zhai 2022).

This poses a fundamental difference to the problem formulation of this work, as this work deals with a regression problem. It predicts a continuous value based on aggregated data while existing research conducts a classification for a single observation. This difference is also visible in the data used to make predictions. The dataset provided by Grips has high density and low sparsity. Opposingly, the datasets of the existing research, such as the commonly used Criteo dataset, usually contain the columns *user* and *item* with extremely high cardinality for each (Yang and Zhai 2022; Juan et al. 2016; Zhou et al. 2018). To make these columns usable for machine learning models, the values are usually one-hot-encoded, resulting in extremely sparse and high-dimensional feature vectors. Consequently, models are built in such a way that they can incorporate this high dimensionality and sparsity (Zhou et al. 2018).

Recent literature focuses on creating and exploring models suitable for such characteristics. At the beginning of user-item CTR prediction, the research focused on multivariate statistical models like linear regressions enhanced with polynomial features to include feature interaction (Wang, Suphamitmongkol, and Wang 2013; Yan et al. 2014). As these models show difficulties dealing with sparse data, the research progressed to models more suitable for this kind of challenge with a focus on factorization machines and later field-aware factorization machines (Ma et al. 2016; Pan et al. 2018; Yuchin et al. 2016). While these consistently outperform the multivariate models, the literature points out difficulties with generalization (Yang and Zhai 2022). With the rise of convolutional neural networks (CNN), research increasingly attempts to utilize them for CTR prediction (Chen et al. 2016;

Gharibshah et al. 2020). Based on Google's Wide&Deep recommender system architecture (Cheng et al. 2016), the introduction of DeepFM (Guo et al. 2017) marked a new milestone in user-item CTR prediction. DeepFM combines the previously well-working factorization machines with neural networks and lays the foundation for current research. The most recent literature focuses on more accurately modeling feature interactions and produces promising models, such as MaskNet (Wang, She, and Zhang 2021) and DeepLight (Deng et al. 2021), that continuously, though only slightly, outperform DeepFM. Yang and Zhai (2022) provide a detailed overview of the field of CTR prediction in the context of user-item interactions.

Next to the differences between individual click prediction (classification) and aggregated CTR prediction (regression) as described above, a further differentiation lies in the nature of clicks. There are some papers that focus on search engine CTR (Richardson, Dominowska, and Ragno, 2007; Graepel et al. 2010; Zhang et al. 2021). However, only the work of Richardson, Dominowska, and Ragno (2007) focuses on the prediction of aggregated CTR, while the other works predict individual, non-aggregated user interaction. Furthermore, all works focus on paid instead of organic results. This lack of research on organic CTR can be explained by the absence of publicly available data on this topic, as the data is one of the search engines' most significant assets, which therefore has no interest in sharing it.

To summarize, many papers focus on predicting whether an individual user will click on a specific ad and use this information as a recommender system. Additionally, some papers predict aggregated advertisement CTR. However, to the best of our knowledge, there are no publications that predict aggregated organic search engine CTR.

2.3 Implications

The analysis of the research on SEO marketing shows that the position is the most important driver of CTR. Therefore, it promises to play a critical role in predicting organic CTR. In contrast, the effect of SERP features is theorized but barely analyzed. Therefore, it is essential first to analyze and fully understand the drivers of CTR and potential modeling approaches before predicting CTR. Therefore, in chapter 4, an in-depth exploratory data analysis was conducted to provide fundamental insights into the challenge of CTR prediction on organic search engine result data.

Additionally, while there is a large amount of literature on CTR prediction regarding user-item interactions, there is little research on aggregated search engine CTR prediction. Therefore, while it is insightful to apply the previously mentioned models to the data at hand, it is crucial to consider that the nature of the problem and the available data are different. This is likely to require distinct modeling approaches, which are tested in chapter 5.

3. Dataset & Methodology

This section provides an introductory overview of the used data and methodology applied to answer the research question.

3.1 Dataset

The dataset used for the analysis was provided by Grips, formerly Peekd, a German start-up that aims at ‘shining a light on the blind spots [in e-commerce sales] and create the most comprehensive map of online commerce for retailers and brands’ (Grips 2022). The data comprises historic Google search data for a given URL and keyword. One row can, for example, resemble metrics related to the performance of the URL ‘www.novasbe.unl.pt/en/’ for the searched keyword ‘nova university’. As such, the data is provided in a tabular, heterogenous format. It ranges over 43 domains, collected from US-based desktop searches between May 31st, 2022, and August 18th, 2022. In the raw dataset, 70 features are available for 79,853 data points.

Although the dataset was provided by Grips, it was enriched with features from Semrush, a company specializing in search engine marketing, and Google's Keyword Planner (see the source of each feature in appendix II.). The features cover various aspects, ranging from the searched keyword over metadata about the result page to result-specific information and which SERP features are shown. Finally, the data includes the click-through rate for each result which, in line with the research question, is determined as the target value.

The dataset, with its comprehensive set of features consisting of real-world, not publicly accessible data for dozens of e-commerce stores, signifies a novelty in available data.

3.2 Feature subsets

For a better understanding of the effect of position, keyword, result, and SERP feature characteristics on CTR, the variables in the dataset were grouped accordingly into four feature subsets. If applicable, subsets were enriched with self-generated features. The generated feature subsets form the basic structure for the analysis in chapter 4 and 5. A data dictionary describing all features can be found in appendix II. and a list of all features for each subset in appendix III.

Position - The position subset groups features that are exclusively position related. Those are the position of the result during the measurement, the monthly position average, and the difference in positions to the last measurement. While the position subset is technically also handled as a subset in the following analysis, it is important to note that the feature position plays an overarching role in the problem at hand. The literature review and findings of this work highlight that position is the most important feature for predicting CTR. Therefore, in the analysis, the position subset is often used in conjunction with other subsets.

Keyword - The keyword subset consists of the keyword itself and everything directly resulting from the keyword a user types into the search engine. This includes information about the keyword, such as the length of the keyword, information indirectly related to that keyword, like the competition or the search volume, and finally, the search intent. One generally differentiates between four intents: (i) The informational search, where a consumer is looking for information about a subject; (ii) the navigational intent in which a specific website is searched, e.g. a marketplace for a product; (iii) the commercial intent that implies a purchasing decision but it is not yet decided on the final transaction (e.g., “linux alternative”); (iv) the transactional intent where a consumer has an action in mind such as “buy iPhone”.
To quantify the complexity of each keyword, the Flesch reading ease score (Flesch 1948) was computed. Even though newer scores have been developed, Flesch’s reading ease score is used because it is commonly accepted and easily interpretable. It counts the average number of words per sentence and the average number of syllables per word. A high score corresponds to good readability, characterized by short sentences and short words.

SERP features - The SERP features in the dataset are binary features stating for each entry whether each SERP feature is present (1) or not (0). They can be divided into page and positional SERP features. Page SERP features describe whether a certain feature is present somewhere on the page. Positional SERP features describe whether a result is featured in a particular SERP feature. This can take different forms. For example, a *Review* can be shown beneath a URL, or a *Snippet* from the URL can be displayed in the *Knowledge Panel*. Therefore, if a positional SERP feature is present, the equivalent page SERP feature must be present, but not vice versa. The page SERP features also include information about the kind of ads present on a result page. We aggregated this information to a binary feature describing whether ads are displayed. To add a layer of interpretability, the dataset was additionally enhanced with the total count of SERP features both for page and positional features.

Result - The result subset includes all features related to the characteristics of one URL on a result page. While for the three other subsets the features available in the dataset are very comprehensive, available result-related features only cover the URL of the result but not other relevant aspects such as the title or description. To generate informative value out of the URLs, the associated domain is introduced as a one-hot-encoded feature, and two keyword-related features are computed: It is checked whether at least one word of the keyword appears in (1) the domain and (2) the URL.

3.3 Data preparation

Before looking into the data in more detail, it has to be prepared for further analysis. To do so, the Google data preprocessing guidelines were followed (Google 2022c). Duplicate rows and columns as well as columns that had only one unique value or were redundant were dropped. Since $CTR = \frac{clicks}{impressions}$, both 'clicks' and 'impressions' are very closely related to the target variable and are consequently dropped, avoiding collinearity and information spillage. Finally, column data types were changed in order to facilitate effective analysis. To minimize noise through observations of result pages that were shown to too few users, an impression threshold of 20 impressions per result was introduced. 58,898 rows remain for analysis.

For the modeling, all non-binary features are standard-scaled to a mean value of 0 and standard deviation of 1.

4. Drivers of CTR (EDA)

This chapter aims at identifying drivers of CTR through an exploratory data analysis (EDA). We focus on non-obvious and novel insights as well as characteristics that are relevant for CTR prediction models. Thus, the EDA does not only create a basic understanding of the data but also as the foundation for selecting and examining predictive models in chapter 5.

The analysis is structured along the four feature subsets and dedicates a section to each. In addition to that, we analyze feature interactions and present a summary of the main findings and implications for the modeling.

4.1 Position - the most important predictor

There are two main findings regarding the *position* subset: First, the position is more relevant than other subsets in determining CTR. Second, results on position one receive much higher average CTRs than results on any other position.

In line with previous research (see chapter 2 and Iqbal et al., 2022, 4), our data shows that of all features, the position has the strongest impact on CTR. This is based on position and related features having the highest correlations with CTR. The average position per month has a Pearson correlation of 0.5 with CTR. Related features, such as the previous position, have a Pearson correlation of more than 0.4. This is significantly higher than the second most correlated variable with CTR ($\rho = 0.25$)[1].

Our findings show that the CTR on position one is much higher than on other positions, as many search engine users only read and click the first result. The result on the first position receives an average CTR of 0.25. This is 2.5 (3.6) times higher than the average CTR on the second (third) position. Therefore, our data supports the prevailing belief in the academic literature that optimizing the rank is crucial for improving a website's traffic. However, there is a considerable standard deviation of 0.19 for position one. Outliers on the first position reach up to 1.0 CTR, while a CTR of 0.0 is also possible. This can also be seen in the second and third positions, with a CTR range between 0 and 0.8.

This range translates to a significant variance reiterating the need to correctly predict CTRs, as even minor deviations in CTR can have significant click and revenue implications for

[1] ρ denotes the Pearson correlation coefficient

businesses and their SEO efforts. However, the strong variance also underlines the importance of other explanatory variables to explain strong deviations. Lastly, the results highlight position as the single most important feature, which inherits substantial implications for the modeling: We infer that the *position* subset is crucial for accurate predictions and that it can thus serve as baseline subset to use in models to analyze the predictive power of other subsets.

4.2 Keyword characteristics - the role of a keyword's complexity

Among keyword characteristics, the complexity score reveals a particularly noteworthy effect on CTR. For the complexity score introduced in chapter 3, the simplest possible keyword receives a score of 121, with the score decreasing towards a technically infinite negative value for more complex keywords.

Figure 1 shows the average CTR for keywords on position one depending on their complexity. Additionally, it distinguishes between branded[2] and unbranded keywords. The Figure shows that keyword complexity creates mainly two contrasting effects: For more complex keywords consisting of many and/or longer words[3] (score below 80), the average CTR of branded keywords is significantly lower than for unbranded ones. This changes for values with a complexity score of more than 80, where simple branded queries lead to significantly higher average CTRs. For the most simple keywords, branded keyword results, on average, achieve a CTR of more than 50% higher than unbranded keywords. Additionally, a general tendency of decreasing CTR for simpler keywords can be observed for unbranded keywords.

The results imply that optimizing for position one makes sense for branded keywords that are simple and short, where i.e. the brand is likely to be a very prominent component. However, the results also hint that longer and more complex keywords including brand names can even

[2] A keyword, or search query, is branded if it includes a brand name.

[3] To make the score more tangible: A search query consisting of five words with an average number of syllables per word of 1.8 would achieve a reading ease score of 49.5. An example of such a query is "buy cat food online now".

harm the CTR of results on position one. This implies that users are more explorative for such keywords and tend to be more likely to consider results after position one.

The results also highlight feature interactions between the complexity of a keyword and whether it is branded, implying a need for prediction models to be capable of modeling feature interactions.

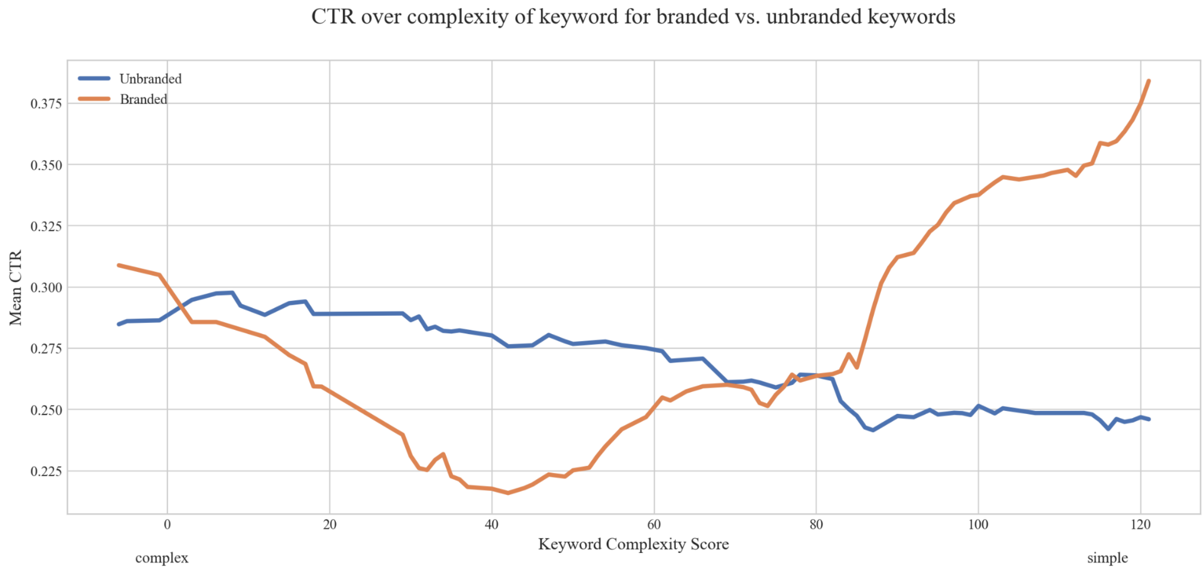


Figure 1: Mean CTR over keyword complexity score for branded and unbranded keywords. Mean CTR is calculated as a centered moving average with a total window size of 30. Keywords are more complex for low values and simpler for high values.

4.3 SERP features - variation in effects over positions

In the following analysis, we examine the impact of SERP features on CTR and outline implications for SEO decision-making concerning SERP features.

As the count of positional SERP features increases, i.e. a single result has more SERP features bound to it, the CTR tends to rise ($\rho = 0.22$), whereas with more page SERP features, there is a less clear trend, with CTR tending to decrease ($\rho = -0.11$).

In chapter 4.1, we have shown the importance of positions. When adding another dimension, namely the presence of a certain SERP feature, noteworthy tendencies arise. While for some SERP features the average CTR per position remains relatively constant regardless of the presence of the SERP feature, for others, this presence has a big influence. Tolerating minor exceptions, there are two main patterns observable: With the presence of a SERP feature, first, the average CTR declines over all positions ('Lower'), and second, the average CTR declines for up to the first three positions but increases for subsequent positions ('Lower → Higher'). See Figure 2 for an example of each pattern category and SERP features it applies to.

Explanations for why these patterns arise for certain SERP features are not clearly identifiable. However, some features are considerably correlated with intents, e.g. *Image* with transactional search intent (ρ = 0.29). Consequently, when interpreting the results, we must keep in mind that effects might also originate from other feature categories, such as the search intents.

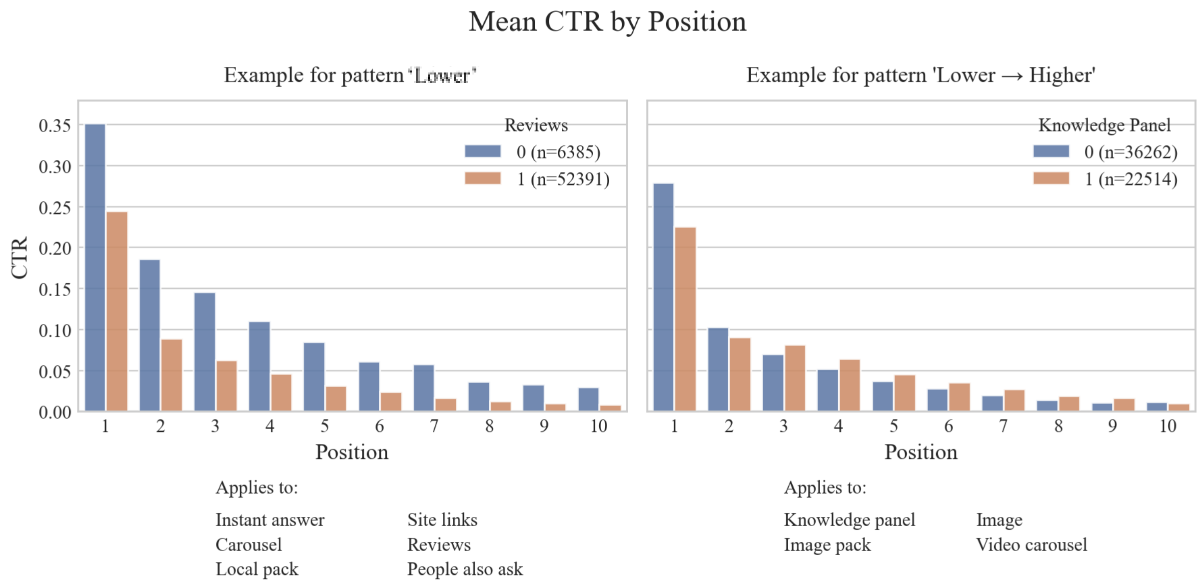


Figure 2: Mean CTR per position if a SERP feature is present (1) compared to when it is not present (0). Left plot shows pattern 'Lower' and right plot pattern 'Lower → Higher'. SERP features without a clear pattern or n < 1000 are not listed.

Although we have previously stated that ranking on a top position is of utmost importance for maximizing CTR, the aforementioned findings on SERP features spur the question of whether, in some cases, it is more important to rank on top positions than in other cases. While above, we see the impacts of SERP features isolatedly, in practice, SERP features appear together and thereby likely influence each other. Thus, in the following, we analyze the importance of positions for each combination of page SERP features. To be able to compare different combinations better, for each combination with more than 200 occurrences, we normalize the CTR to the average value on the first position and then calculate the decay for each subsequent position (Figure 3). As a result, it becomes evident that patterns identified for individual SERP features continue across combinations of them. Among the top three combinations by count, for two, the CTR decays much faster, while for the other one, it decays much slower compared to the entire dataset. For practitioners, this implies that efforts towards ranking on the first position can be of much more value if some SERP features are present, while in the presence of others, a positioning within positions two to five might also suffice. For example, three differing SERP features can already double the maintained CTR on position three (see the difference between the green and blue combination in Figure 3).

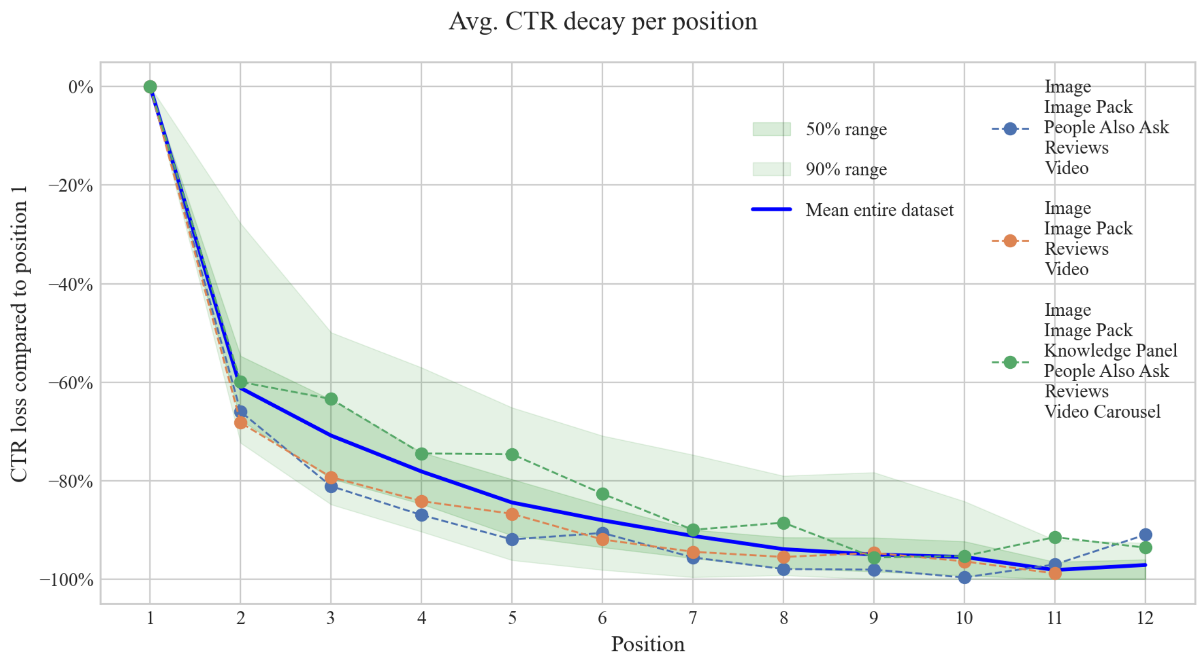


Figure 3: Loss in CTR per position. Values are avg. CTR loss for each position in % of avg. CTR on position 1. Only combinations with > 200 occurrences are considered (n=59). Dashed lines represent the top three most frequent SERP feature combinations. Shaded areas represent the range in which 50% and 90% of all values fall.

Together, the implications for the modeling are twofold. First, we show that impactful patterns can be observed on different levels of aggregation, i.e. when using the count of all SERP features, looking at each feature separately, or looking at combinations of features. Thus, knowledge about SERP features should help models to more accurately predict CTR. Second, we can observe a high degree of interactions. The interactions go beyond the second degree, like the interaction between position and combinations of SERP features. Consequently, to make use of all information inherited in SERP features, models must be able to incorporate feature interactions higher than second degree.

4.4 Result characteristics - how domains & brand perception influence CTR

In this chapter, we focus on the effect of the domains on CTRs. The domains in the analyzed dataset are exclusively from e-commerce websites. They either belong to a manufacturer or a marketplace specialized in product categories like electronics or apparel. Therefore, the domains either represent a brand name or a marketplace and can often be seen as a placeholder for a product category.

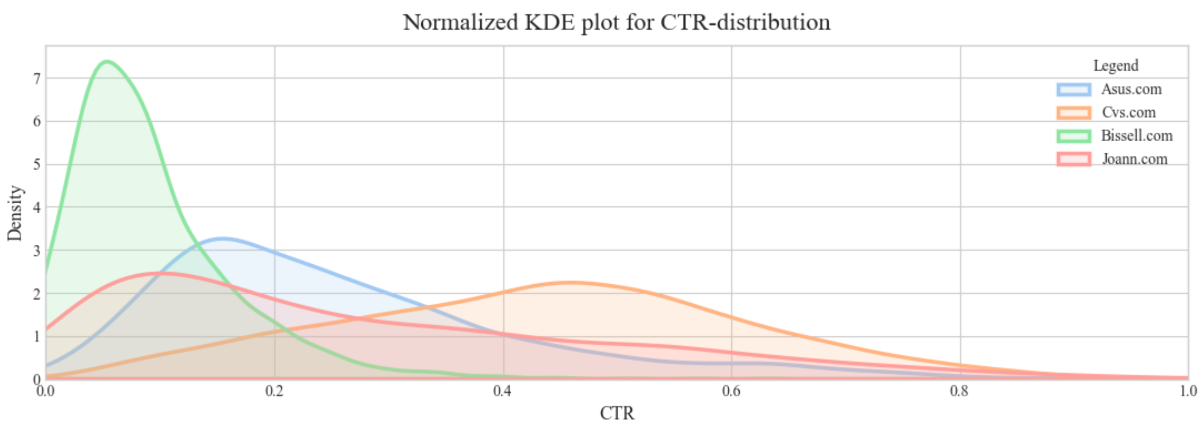


Figure 4: The Kernel Density Estimation (KDE) plot for four of the top five domains by count, normalized to unify the area under the curve.

Figure 4 shows that domains have widely different CTR distributions on position one. CVS, a well-known pharmacy, achieves the highest average CTR among domains that appear at least 1,000 times. While Bissel, a premium vacuum cleaner company, receives the lowest average CTR of these domains. The kernel density estimated distributions of CTR are right-skewed for domains with low average CTR, whereas domains with high average CTR go along with a more uniform distribution. We find that there are no correlations with other variables that could clearly explain these patterns. This indicates that the brand name is still important in determining CTR to a website’s traffic.

This is especially clear when looking at the top and the worst performing domains on position one, CVS and Bissell. Both of the websites are reached from search queries that include their brand name in 87% of cases for Bissel and 77% for CVS. In both cases, 98% of these queries have a transactional (clear purchasing) intent. However, even though both domains appear on position one, in search queries with a purchasing intent for a product of their brand, there is a large difference in CTR. ‘cvs.com’ achieves an average CTR of 26%, compared to 6% for ‘bissel.com’.

A possible explanation could be a certain brand recognition by the consumers. While the pharmacy, CVS, and its online business is a respected marketplace, the premium vacuum cleaner company, Bissel, is known for its product but not as a primary marketplace, as third parties often offer lower prices. When displayed in Google, Bissel's website is in 70% of the cases enriched with the positional SERP feature *Review*. As the *Review* can also show prices, it is likely that third-party retailers show a lower price and drive traffic away. Additionally, searches for Bissel’s products are 55% more likely to include ads than the rest of the data set. This presumably leads to more competition for users’ attention.

Supporting the thesis for the high relevance of a website's brand, we find that the two domains with the highest average CTR (‘hp.com’ & ‘cvs.com’) are the only ones listed in the top 100 most powerful brands index in the US (Tenet Partners 2020), while the others do not appear.

Based on these findings, we conclude that domains are important in determining the CTR of a website and have high relevance for predicting it.

4.5 Feature interactions - SERP features & the brand name in keywords

In our previous analyses, we have uncovered several feature interactions across and within feature categories. In the following, we quantify the strength of feature interactions and reveal the strongest interaction's underlying forces.

Friedman and Popescu (2008) introduce the H-statistic to measure interactions of two or more features in an ML model. The H-statistic quantifies the interaction strength of two features by the difference in the explained variance of a decomposed partial dependence (PD) function and the observed PD function. The decomposed PD function assumes no interactions, whereas the observed PD function measures interactions. Thereby, it calculates the explained variance resulting from interactions.

$$(2)\ H^2_{jk} = \frac{\sum_{i=1}^{N}\left[PD_{jk}(x_{ij}, x_{ik}) - PD_j(x_{ij}) - PD_k(x_{ik})\right]^2}{\sum_{i=1}^{N} PD_{jk}(x_{ij}, x_{ik})^2}$$

If the resulting score amounts to zero, all variance is explained by the individual 'contributions' of the features. Thus, there is no interaction between them. A higher value indicates that more variance is explained by the PD functions with interactions. This suggests that each individual PD function is constant and the effect on the prediction only comes through the interaction (Friedman and Popescu 2008).

The H-statistic has some limitations that need to be addressed when using it. First, it is non-deterministic, i.e. the same input leads to different results. Therefore, ten iterations were completed to see if it delivers stable results. Second, the H-statistic is very computationally expensive when applied to higher-degree interactions. Thus, we only analyze second-degree interactions. Third, the H-statistic only considers the interaction strength. It does not indicate whether the features have a relevant effect on the target variable or the kind of interaction. For these reasons, it is necessary to analyze the feature combinations proposed by the H-statistic, to clearly explain their effects. In the following, we point out the most relevant findings.

As Figure 5 reveals, the strongest second-degree interaction of an H-value of around 0.6 appears for the features branded with the count of page SERP features. These features reveal an interesting interaction pattern, which is displayed in Figure 6. Branded keywords perform nearly three times as well as unbranded ones when few SERP features are displayed by Google. This effect reduces on result pages with more SERP features. When more than seven SERP features are present, the effect switches and unbranded keywords perform better than branded ones.

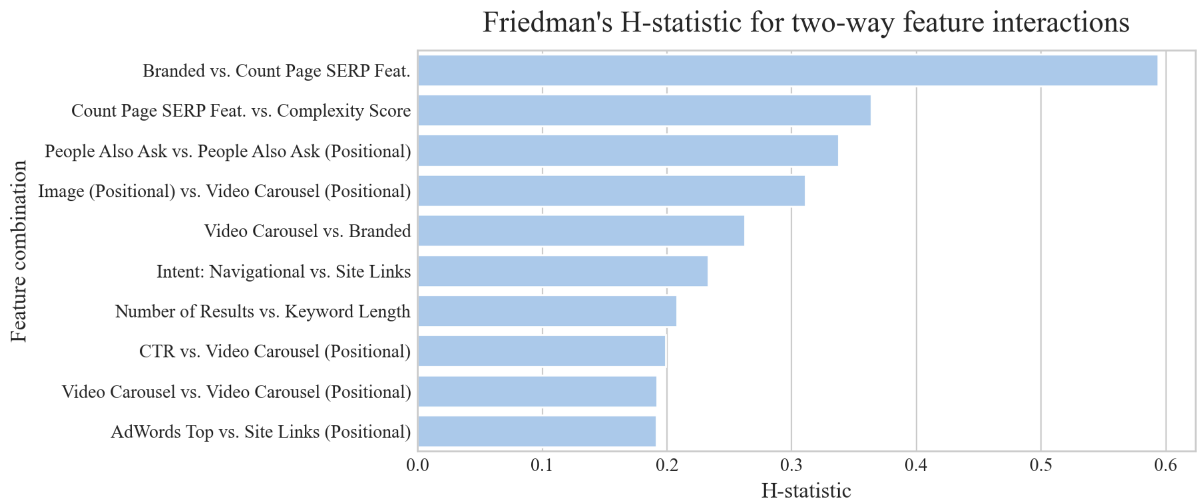


Figure 5: Friedman's H-statistic for two-way feature interactions. Highly correlated features and features with less than 500 occurrences were dropped, as they are likely to bias the H-statistic.

As this effect is not intuitively explainable, it is worth investigating it further. We find that result pages with few SERP features have very different properties, depending on whether the keyword is branded or not. The data also shows that result pages for branded keywords are much more likely to have navigational and transactional intent than those for unbranded ones. In contrast, unbranded result pages with few SERP features are more often informational than their branded counterparts. To summarize, branded result pages with few SERP features have a high CTR, driven by purchase-oriented users who are looking for a specific (branded) website.

The implications are two-fold. First, branded search queries are more likely to generate website traffic if they lead to result pages with few SERP features than to result pages with many SERP features. Second, for unbranded search queries, the number of SERP features is not similarly relevant and a medium CTR can be achieved regardless of them.

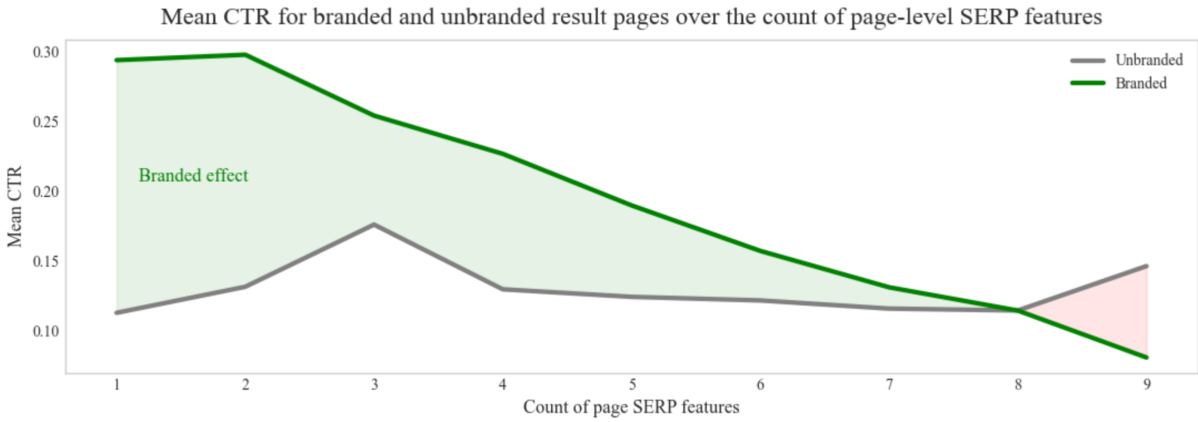


Figure 6: Mean CTR for branded and unbranded result pages over the count of page-level SERP features.

4.6 Findings & implications summary

Next to the detailed analysis of individual drivers of CTR above, there are two main implications based on this data analysis. First, within each subset there are features that have a relevant effect on CTR. Therefore, we will consider all of them as model inputs. Second, we revealed many interactions between features. Most features interact with position, which we either explicitly stated or we analyzed features on only one position, to isolate the feature effects. This shows the high relevance of the position for CTR. Additionally, features within and in between subsets interact with each other. These feature interactions imply that the tested models should either implicitly or explicitly handle feature interactions.

5. Modeling

In the following, different models for the prediction of CTR are tested and the predictive power of the four feature subsets is examined. First, we introduce the tested models. Then, the modeling methodology is outlined. Lastly, we analyze the modeling results.

5.1 Examined models

The different feature subsets are tested over different models. As pointed out in 2.2, the underlying problem is a regression problem. Consequently, the tested models are regression models. Furthermore, these tested models were selected considering their relevance and performance in recent literature as well as their applicability to the dataset and its characteristics as pointed out in the EDA. Additionally, linear regression models are applied as a benchmark due to their simplicity. Three model categories emerge: Linear regression models, neural network regression models, and tree-based regression models.

5.1.1. Linear regression models

Linear regression attempts to model the relationship between two or more variables by fitting a linear equation to observed data (Uyanık and Güler 2013). This allows for fast computation and easy interpretability. However, linear regression models assume the predictor variables to be independent from one another and linearly related with the independent, predicted variable (Su, Yan, and Tsai 2012). Furthermore, linear regression models are prone to outliers and overfitting the training data in the case of a larger number of features (Filzmoser and Nordhausen 2020). As a result, the model can be simplistic and unable to capture real-life complexities in data. Based on the dataset's characteristics outlined in chapter 3, the limitations also apply to the dataset at hand. There are strong feature interactions, i.e., the predictor variables are not independent, and relations with CTR are not always linear.

Regardless of these drawbacks, we will use ordinary least squares regression (OLS), the most popular linear regression model (Su, Yan, and Tsai 2012), as our baseline model. In an OLS model, the squared errors are minimized (De Gryze, Langhans, and Vandebroek 2007). Using this model allows us to compare more elaborate models with a solid, proven statistical framework. Suppose a model performs worse than OLS linear regression, despite the shortcomings and violated assumptions of linear regression described above. In that case, we can assume that it is not suitable for the dataset at hand. We also adapt the OLS model by adding feature interactions as the product of the values of two features (Poly2). As in this way, the number of features grows immensely, we only allow feature interactions with a pearson correlation coefficient > 0.05 in the training data set to be used by the model. Potential overfitting will be addressed using a third variant, a Ridge regression model, which implicitly performs feature selection to avoid overfitting using L2-Norm.

5.1.2. Tree-based regression models

Decision tree models are based on a series of if-then-else decision rules, resulting in tree-like structures, and are already used since the 1960s (Morgan and Sonquist 1963; Messenger and Mandell 1971; Quinlan 1986). Decision trees are enhanced in state-of-the-art models that use ensembles, i.e. a combination of decision trees to make a joint prediction (Omer and Rokach 2018). The considered tree-based models in this work are all ensemble models based on decision trees. The advantages of such models lie in them not requiring specific distributions in the data, being fast without preprocessing, and they are capable of modeling feature interactions (Agarwal et al. 2022). In addition, on medium and small-sized datasets, tree-based ensemble models often outperform more elaborate state-of-the-art models such as neural networks (Grinsztajn et al., 2022; Hancock and Khoshgoftaar, 2020).

Hence, tree-based ensemble models can address both the interactions and limited sample size of the available dataset, making them very promising candidates.

We will test the top three performing tree-based models for tabular data, found in a recent comparative benchmark by Grinsztajn et al. (2022), namely Random Forest (Breiman 2001), Gradient Boosting Decision Trees (GBDT) (Friedman 2001) and XGBoost (Chen and Guestrin 2016), as well as CatBoost, another tree-based model achieving high scores in benchmarks (Prokhorenkova et al. 2018).

Random Forests make use of bagging, an approach where the results of multiple independent models, each trained on a subset of the data, are combined (Breiman 1996). Applied to Random Forests, this means that to get a prediction, the results of several randomized decision trees are aggregated through averaging (Breiman 2001). This model is continuously praised for its versatility, speed, and ease to use (Biau and Scornet 2016).

Contrary to bagging models, boosting models build decision trees sequentially and learn from the residuals of their predecessors (Freund and Schapire 1996). A variant of this are *gradient* boosting models, in which each tree, instead of predicting the target itself, predicts the error of its predecessor. Friedman (2001) laid the foundation for this variant with GBDT. Empirical studies indicate that gradient boosting models perform comparatively well on heterogeneous data, but do poorly on homogenous data (Hancock & Khoshgoftaar 2020).

Based on the concept of gradient boosting, Chen and Guestrin (2016) proposed XGBoost, which makes use of advanced regularization techniques (L1 & L2), improving its generalization capabilities. In addition, the training of XGBoost is parallelizable, making it much faster. XGBoost remains the go-to tool for most practitioners and data science competitions (Kossen et al. 2021).

Compared to the other examined gradient boosting models, CatBoost optimizes decision trees for categorical variables through two ways. First, the use of ordered target statistics allows for efficient handling of categorical features with high cardinality. Second, ordered boosting prevents prediction shift, referring to a phenomenon where what the model learns in the training set is not reflected in the testing set (Prokhorenkova et al. 2018).

5.1.3. Neural network regression models

While Deep Neural Networks (DNN) achieve unprecedented performance on tasks related to homogeneous data (e.g. image, audio, and text data), heterogenous, tabular data is still deemed an “unconquered castle” for DNNs (Borisov et al. 2022; Kadra et al. 2021; Hancock and Khoshgoftaar 2020). However, a recent comparison between the performance of traditional machine learning models and DNNs on tabular data by Borisov et al. (2022), suggests that while for most cases boosting models deliver best performance, on a dataset with more than 10 million samples, DNNs can achieve similar or even better performance. Furthermore, as outlined in chapter 2.2, the best-performing models in recent CTR prediction research incorporate neural networks. Based on these findings, we test a neural network structure especially designed for tabular data, TabNet (Arik & Pfister 2021), as well as Wide&Deep (Cheng et al. 2016) and DeepFM (Guo et al. 2017), neural networks suggested by recent CTR prediction research.

TabNet has been designed for tabular input data and uses the concept of sequential attention to perform row-wise feature selection. This enables it to use its learning capacity only for the most relevant features. Furthermore, TabNet offers a better interpretability than boosting models (Arik and Pfister 2021).

As highlighted in 2.2, CTR prediction research has evolved to currently suggest models that combine a wide and a deep component. These models usually combine the outputs of a deep neural network (deep component) with those of a linear model (wide component) in a common activation function. This model architecture was first proposed by Google as

Wide&Deep (Cheng et al. 2016) and its core strengths lie in its ability to both memorize and generalize (Jais et al. 2019).

State-of-the-art CTR research extends the idea of wide and deep models and adapts them to the specific needs of ad-CTR prediction. Although being outperformed in specific domains, as for example by MaskNet (Wang et al. 2021) on the Criteo dataset, DeepFM (Guo et al. 2017) is still one of the most commonly applied best-performing models in ad-CTR prediction. DeepFM embeds sparse features and interprets the deep component as a neural network while the wide component incorporates a factorization machine. This allows it to capture both two-dimensional feature interactions in the wide component and higher-dimensional interactions in the deep component. As a result, its advantages lie in the flexibility to, in a unified framework, capture low-level feature interactions explicitly and high-level feature interactions implicitly (Guo et al. 2017). However, it also lacks interpretability and imposes high computational complexity (Yang and Zhai 2022).

The implications for organic CTR predictions on dense search engine data are two-fold: On the one hand, DeepFM is capable of capturing feature interactions very well, which is beneficial. On the other hand, it is highly specialized in dealing with sparse feature vectors, which is not necessarily required in the case of result page data and might harm predictions. Furthermore, neural networks often rely on large samples for model training to generate quality predictions (Alwosheel et al. 2018), which is not the case for the available dataset (only medium-sized; ~60k samples).

5.2 Methodology

5.2.1 Subset combinations

To assess each subset's predictive power, different subsets were tested on the models. Testing a model on a subset combination means that all features of the combined subsets are used for model prediction. Based on both the literature review in 2.1 and our analysis highlighting the position as the single most important feature, the *position* subset serves as a baseline subset. It is then combined with the other subsets.

First, the *position* subset is combined solely with each of the other subsets such that the additional predictive power of each subset can be assessed isolatedly. The *result* subset, however, inherits a structural limitation: Because it mainly consists of one-hot-encoded domains, it is very specialized to the underlying dataset. As such, if a prediction for a new domain was to be made in a production environment, the one-hot encoded domains could not contribute to making a more accurate prediction. Thus, the *result* subset has very limited generalizability, and scores need to be interpreted cautiously.

Based on this limitation, we first check the predictive power of all generalizable subsets combined, i.e. *position* + *keyword* + *SERP features*. To get an understanding of the full potential, we lastly also include the *result* subset to combine all available meaningful features. As a result, 6 subset combinations are tested on each model presented in 5.1.

5.2.2 Evaluation metric

In order to compare the effectiveness of different feature subsets and models, we need a common evaluation metric. All models are evaluated using the Root Mean Squared Error (RMSE), defined as

$$(3)\quad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}.$$

This metric allows for the penalization of larger errors which is desirable in the present business case as larger errors can lead to a complete miss investment of SEO efforts. As companies are likely to optimize their SEO marketing for the keyword with the highest expected clicks, it would be a particular financial loss if this estimation is wrong. In addition to that, the RMSE can be used as a quick optimization metric across all model categories, as opposed to, for example, the mean absolute error (MAE).

5.2.3 Hyperparameter tuning

To examine each model's full potential, we conducted hyperparameter tuning where reasonable. The hyperparameters for all linear models were kept constant. Tree-based models were tuned based on the hyperparameter sets suggested by Gorishniy et al. (2021) and Grinsztajn et al. (2022). As the tuning algorithm, Bayesian Optimization (Snoek et al. 2012) with 100 iterations is used since it is reported to perform superior compared to random search (Turner et al. 2021). For the comparison of scores of different hyperparameters, 5-fold cross-validation on the training set was used. Please find the full documentation on tested parameter ranges in appendix IV and the resulting parameters in appendix V.

The training of neural networks was limited to 200 epochs with early stopping after 10 epochs without improvement of the RMSE of the validation set. During the training and evaluation of neural networks, we noticed a large variance in the resulting test scores. To compensate for this, we averaged the scores of 10 different experiments to receive the final test score per neural network based model. However, by doing so, an extremely large number of experiments would need to be performed during hyperparameter tuning, resulting in extraordinarily large computation times[4]. Due to this, we could not conduct in-depth hyperparameter tuning for neural networks. Nevertheless, we performed some manual experiments on hyperparameters based on the domain knowledge gained during the EDA. While for most combinations of features subsets, the altered parameters did not lead to significant improvements, scores for the combination of all subsets were drastically improved

[4] Example: For setup [iterations = 100; No. optimized subsets = 6; CV = 5; experiments per score = 5; training time = 30 sec], the optimization would take 5.2 days per model

(see appendix VI). Thus, for all neural networks, the default hyperparameters were used for all combinations of feature subsets except the combination of all subsets.

5.3 Results & interpretation

In the following, we analyze the model results (Table 1), first concerning how different models perform, and second considering the predictive power each subset incorporates.

5.3.1 Comparison of model performance

Clear patterns arise both between each model and the categories they are grouped in. Apart from summarizing these patterns, we examine whether the assumptions regarding the applicability of models, made in chapter 5.2, hold true.

On average, linear models perform worse than other categories of models, as expected. As described in 5.1.1., we tested the Poly2 model to check the assumption that model performance improves when we allow interactions between features. This assumption holds true as the Poly2 model significantly outperforms the OLS/baseline model. An exception to this is the case when all feature subsets are used, where the error for Poly2 increases drastically. A potential explanation for this is that with more features, the number of interactions grows quadratically, thus leading to so many features that the linear regression model heavily overfits. Ridge only performs marginally better than OLS, indicating that the OLS model works well with the number of features and overfitting is not an issue.

Although mostly outperforming linear models, models based on a neural network architecture cannot keep up with the performance of tree-based models. While scoring very closely to tree-based models for datasets with few features, such as only position, the differences enlarge with the presence of more features and accompanied complexity (see Table 1).

		RMSE of subset combinations					
		Position	**Position + Keyword**	**Position + SERP Feat**	**Position + Result**	**Position + Keyword + SERP Feat**	**Position + Keyword + SERP Feat + Result**
Linear Regression	OLS	0.153	0.145	0.145	0.132	0.137	0.125
	Poly2	0.143	0.129	0.131	0.119	0.122	2.554
	Ridge	0.152	0.144	0.144	0.131	0.137	0.123
Tree Based	Random Forest	0.137	0.108	**0.120**	0.108	0.100	0.088
	GBDT	0.134	0.109	0.122	0.109	0.108	0.091
	XGBoost	**0.133**	0.104	**0.120**	0.106	**0.098**	**0.083**
	CatBoost	0.134	**0.103**	**0.120**	**0.105**	0.099	**0.083**
Neural Network	TabNet	0.137	0.121	0.130	0.111	0.118	0.102
	Wide&Deep	0.137	0.123	0.152	0.110	0.128	0.109
	DeepFM	0.136	0.135	0.137	0.113	0.133	0.114

Table 1: Benchmark results on different feature subsets. The top results for each feature subset are **bold**. We also underline the second-best results.

Furthermore, TabNet, a model resulting from research focusing on using neural networks for inference on tabular data, performs better than models resulting from CTR research (Wide&Deep and DeepFM). Additionally, the strength of CTR research models on sparse data and weakness on tabular data is reflected in the scores. For sparse subset combinations (*position + result*) their error is only 4% above the error of tree-based models, which increases to 19% for dense subset combinations (*position + keyword*).

It is important to note that even without hyperparameter tuning, tree-based models achieved consistently lower errors than the neural networks (see appendix VI). Based on this, we assume that the considerably worse performance of neural networks is mainly rooted in the model's architecture instead of the lack of hyperparameter tuning.

Together, this confirms the previously highlighted inapplicability of models resulting from

CTR prediction research to the dataset due to lack of sparsity. Additionally, the neural networks likely suffer from a relatively small number of observations.

Tree-based models perform clearly the best. It is noticeable that the boosting models mostly perform better than the bagging model. While the random forest performs better than GBDT in half of the feature subsets, it is worse than XGBoost and CatBoost in every subset. XGBoost and CatBoost are the two top-performing models for every feature subset. For two subset combinations, they achieve the same RMSE, and for the remaining subset combinations, one outperforms the other just by 0.001 RMSE. These very similar performances spur the question of which model to use under secondary, non-performance-based factors. XGBoost is better documented and more established than CatBoost, hence facilitating the interpretation, tuning, and understanding of the applied model. For all these reasons, we suggest using XGBoost to predict the organic CTR of websites on Google result pages.

The general optimality of tree-based algorithms is in line with current research. The excellent performance of tree-based models indicates that they are able to capture feature interactions on tabular, heterogeneous data well, while also being able to learn from relatively few samples.

Overall, we observe four main patterns arising across the different models. First, models capturing interactions perform better than those that do not. Second, models designed for the use on tabular data make better predictions than models designed for user-item-specific CTR prediction. Third, the performance differences between models amplify with the presence of more features. Fourth, tree-based models clearly outperform all other model categories.

5.3.2 Explanatory power of feature subsets

By training the models on combinations of feature subsets, it is possible to evaluate the explanatory power of subsets and derive implications on the overall relevance of subsets for CTR.

When comparing the improvement in RMSE for additional feature subsets, we focus on the results of XGBoost. Thereby, the comparison is not diluted by outliers created through badly performing models. Additionally, we validate these comparisons by running paired t-tests on the results of the, on average, five best models[5]. Thus, we can determine if the difference in RMSE of feature subsets is statistically significant.

We find that RMSE scores improve when adding each of the other subsets to the *position* subset. Adding the *keyword* and *result* subsets notably improves the RMSE, reducing it by 22.6% and 21.1%, respectively. These subsets offer more explanatory power than the *SERP features* subset which only improves the RMSE by 9.8%. When adding *SERP features* to *position* + *keyword*, the previously best combination, we only see a slight improvement of 4.9%. The combination of all feature subsets results in the lowest RMSE. It reduces the error of the baseline *position* subset by 37.6% and also improves the error of the *position* + *keyword* + *SERP features* subset by 16.2%. All of the mentioned differences are statistically significant according to the paired t-test.

Together, this implies that all feature subsets are relevant, as they decrease error when being added and improve the baseline model that solely relies on position. Additionally, the larger improvements in RMSE for the *keyword* and *result* subset than the *SERP feature* subset, indicate that they have a larger effect on CTRs. However, this observation is only an

[5] The T-test was conducted with RMSE scores from the following models: CatBoost, XGB, Random Forest, Gradient Boosting and TabNet. The alpha used is 5%.

indication as feature importances are not quantified from a model perspective. Individual part B solves this, by applying interpretation techniques to XGBoost, which analyze feature importances and interactions to understand the relevance of the subsets for CTR.

Increasing the number of combined feature subsets to three and four subsets improves the model performance with each added subset. However, improved results including the *result* subset must be interpreted with caution due to its highlighted lack of generalizability. To increase generalizability individual part C generates a variety of *result* features based on a result's title, that are generally applicable. This helps further understand the predictive power of results.

6. Conclusion

By using a dataset that is novel in its comprehensiveness, this work analyzed drivers of CTR and CTR prediction models. It extended existing SEO research by not only considering the position as a driver of CTR, but also aspects grouped into keyword, result, and SERP feature characteristics. Ultimately, the research questions can be answered as follows:

The analysis confirmed the prevailing SEO literature assumption that position is the most important influence on CTR. However, it was found that also keyword, result, and SERP feature characteristics have a significant impact on CTR. Nonetheless, these impacts are highly dependent on the position of a result and subject to feature interactions.

Furthermore, it was shown that state-of-the-art CTR prediction models are inapplicable for predicting average CTR on Google result pages. Instead, the model analysis revealed that tree-based models, specifically CatBoost and XGBoost, are best suited for the underlying prediction task forecasting CTR. The analysis of model results also revealed that keyword and result characteristics tend to have more predictive power than SERP features.

However, several limitations, which restrict the general applicability of the results, need to be acknowledged. The dataset only covered e-commerce stores in the US, thus being limited to a rather niche segment of the entire Google search spectrum. In addition to that, the analyzed dataset contained only 59k rows. This did not only lead to small sample sizes for some feature combinations so that no statistically significant conclusion could be drawn but also restricted the predictive power of neural networks, which can require large amounts of data to work best. Lastly, limited computation resources did not allow for extensive hyperparameter tuning of neural networks. This potentially prevented those networks from performing better in the model comparison.

Future work could tackle these limitations by applying this methodology to similar data from sectors other than e-commerce. Additionally, building on the analysis above, we recommend future research to attempt confirming the findings on a larger dataset which could also provide a solid base for analyses with neural networks. Furthermore, an enhancement of the data with time series information would allow an analysis of consumer behavior and the dynamics of effects over time. Finally, a similar analysis over different nationalities could provide interesting insights into differences in user behavior and support globally operating website providers.

To extend this analysis the individual parts provide an in-depth study on SERP feature effects on a result level and how these can improve the number of clicks to a website. Additionally, model interpretation techniques are applied to assess the model’s applicability, reveal patterns in the data, and quantify the subsets’ importance. In addition to that, the analysis of results is extended to also incorporate titles of results.

A - Effects of SERP features and their business implications

A.1. Introduction - SERP features as unexplored SEO potential

Currently, search engine optimization (SEO) efforts are strongly focused on improving the rank of a website on a search engine result page (SERP) (Ziakis et al. 2019). However, these result pages changed tremendously in the past years becoming more and more user-centered through the introduction of SERP features (Google 2022d). The above work already touched on the potential these features have for impacting CTR but focused on utilizing this for CTR prediction. However, these SERP features and their effects can also be capitalized on for a provider's CTR improvement when used in the correct context. Thus, the following section shall concentrate on providing this context by working out the concentrated effects of SERP features. Additionally, their implications on the SEO business will be analyzed aiming at providing first empirical research on the theorized impact of SERP features on website performance.

Therefore, first, the effect of a result's presence (and absence) in these SERP features will be analyzed. These effects will then be put into perspective in the context of search intents representing the stages of a user's e-commerce journey. Finally, the results will be compared on a national level in order to potentially detect differences in user behavior between nationalities.

A.2. SERP feature effects on in- and excluded results

SERP features consist of information a search engine takes from specific results. They are shown to provide information faster and make result pages more engaging. It was found before, that they do indeed have an impact on a consumer's behavior on search result pages. However, previously the effect of SERP features was only looked at on a page level. Thereby, page SERP features described the effect of all SERP features on a result page. However, these features only link to selected websites, where they contribute to the result's CTR. Still, other results not shown in the respective SERP feature are indirectly impacted. Consequently, when closely examining the effect of SERP features, a differentiation has to be made between the effect SERP features have on results that they link to and those that they don't. For example, when analyzing the effect of images shown next to a result, it needs to be distinguished between (i) the effect the image has on the result it shows for and (ii) the effect on a result that has no image while others do. In order to facilitate this distinction with the data at hand, a new set of features is introduced. For each SERP feature **present** on a result page, it is distinguished between whether the **result is shown in the SERP feature** (*Result in feature*) or the **result is not shown in the SERP feature** (*Result not in feature*)[6].

We saw before that the more SERP features are shown on a result page, the lower the average CTR. This indicates that these features help consumers to directly find the information they are searching for on the result page, decreasing the need to visit single websites. This negative influence also becomes visible when looking at SERP features separately. As seen in Figure A.1, for most SERP features the correlation with CTR is on average negative.

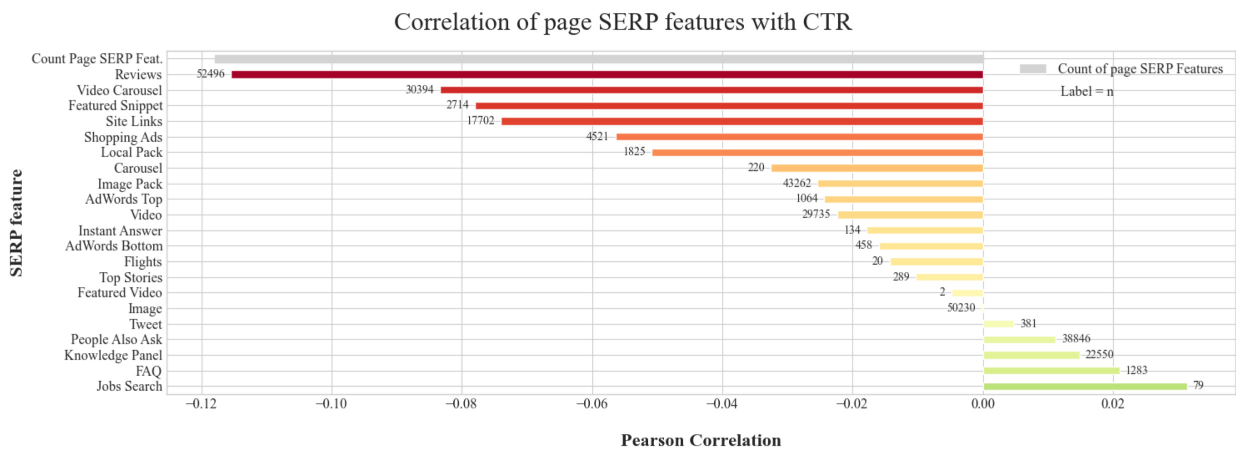


Figure A.1: Correlation of SERP features with CTR when they are generally present on a result page. The color represents the strength of the effect: dark red refers to a strong negative correlation while dark green refers to a strong positive correlation.

---

[6] Referencing the previous example:
An image is present on SERP and for the specific result: *Result in feature = 1, Result not in feature = 0.*
An image is present on SERP but **not** for the specific result: *Result in feature = 0, Result not in feature = 1.*
If no feature is shown for any result on the page, both are 0.

The presence of most SERP features lowers the chance of businesses gaining clicks. Ultimately, this would suggest that businesses should orient their search engine optimization efforts towards keywords with fewer SERP features. However, while this general implication is correct, it is almost impossible to avoid SERP features. In the data set at hand, for example, 99.8% of the result pages listed show at least one SERP feature with the majority showing between four and six features. Consequently, it is crucial for website providers to understand the effects and dynamics of certain SERP features. Especially, whether appearing in specific SERP features can benefit their CTR.

So while the general trend in Figure A.1 is negative, it summarizes the effect of SERP features, regardless of whether the results are featured in them or not. Differentiating these two possibilities paints a different picture.

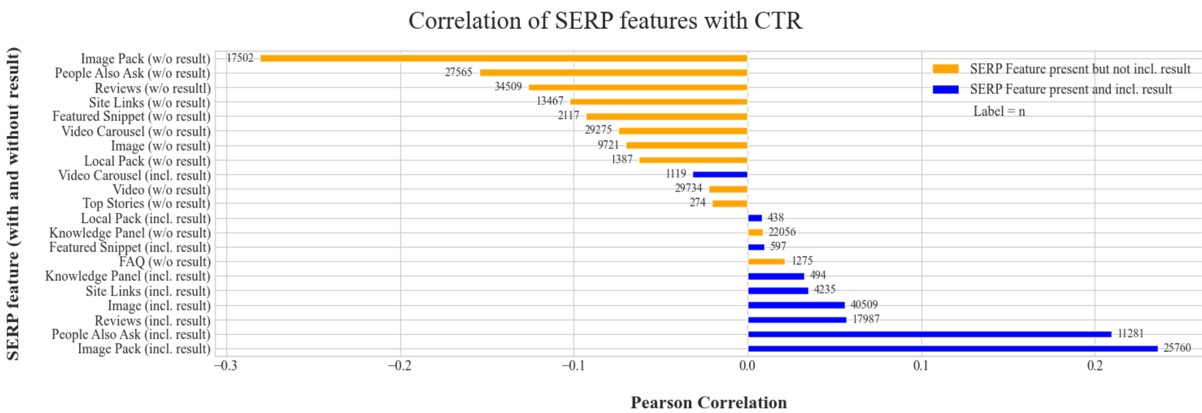


Figure A.2: Correlation of SERP features with CTR differentiating between. For some features, not enough results were included which is why there are more *features w/o result* than *features incl. result*.

Figure A.2 shows that there is indeed a large difference in effect between the two categories. It shows clearly that it has a negative effect to not appear in most SERP features, while it generally increases CTR when the result is in the feature. This raises the question, of what makes SERP features attractive for users and how this can be utilized to benefit providers.

A.2.1 Offering alternatives vs. prominence

Especially remarkable is the effect of the feature *Image Pack*. The feature, which describes a set of images on the result page, has a negative correlation with CTR of almost -0.3 when a result is not shown in it. When a result appears in the *Image Pack*, however, the correlation with CTR is +0.25. This intuitively leads to the assumption that consumers are attracted to visualized content. An argument against this interpretation, however, is the fact that the SERP feature *Image*, describing an image shown next to the result, has a significantly lower impact on CTR whether the result is included or not. This allows a second, more universal explanation. The *Image Pack* consists of several pictures closely related to the keyword shown in a block or row on the search result page. This is both eye-catching and informative at the same time. On the one hand, it creates a large colorful section on a search page which attracts users' attention due to its prominence on the page. On the other hand, it allows a consumer to visualize and select the exact item searched for. In the context of e-commerce, this offers the interpretation that a user is more likely to click on the result when it appears in the *Image Pack* as it is prominent and he or she can be certain, it leads them to the item they are looking for.

A similar picture is seen when looking at the feature *People also ask*. When the result is not present in the SERP feature while the feature is shown, the CTR of the website drops significantly. The opposite is the case when the result is part of the *People also ask* section. This allows the assumption that a consumer finds a question in this section that more accurately answers their query leading to a higher click rate.

A visualization of the effects of *Image Pack* and *People also ask* can be found in appendix A.I. As these two features have the strongest effect both negatively and positively, it suggests two main implications.

One is that consumers are attracted to features offering alternatives to their original query. As it suggests that some search queries are not necessarily perfectly tailored to the actual desired result, it implies a certain trust of consumers in the ability of search engines to nevertheless provide the desired information. This leads to the above-seen higher click-through rates for results appearing in features offering alternatives.

Second, prominent features that break with the continuity of the search result page have a better chance of getting the user's attention, ultimately benefiting the results featured in them.

Now taking another look at Figure A.2, one might argue that other features like the *Knowledge Panel* break with the trend of prominent features benefitting results shown in them. However, upon closer inspection, another interesting explanation presents itself.

A.2.2 SERP feature impact in light of positional effects

The *Knowledge Panel* seems to always be boosting CTR, if only slightly. At first sight, this seems to be doubting the implications from A.2.1 and the findings made in 4.3, where it is found that the average CTR is higher for position one when the *Knowledge Panel* is not present. As the effect of the page feature changes over positions, it raises the question of positional influence on the effect of SERP features and their impacts on results.

Underlining the negative influence of SERP features in general, it was found that especially for position one, almost all SERP features have a negative influence on a results' CTR. Whether results are included in the SERP feature does not play a role yet. This can be intuitively explained through the increased competition for attention with a growing number of SERP features. The positive effect of *Result in feature* seen in Figure A.2 crystallizes itself out from position two onward while the effect of *Result not in feature* stays negative.

As mentioned above, the *Knowledge Panel* - where the result is not in the feature - breaks the negative trend of *Result not in feature*. When inspecting this on the positional level, one finds that like most other features, the *Knowledge Panel* negatively impacts CTR on the first position. Interestingly, its impact stays negative until position three, from where it turns and stays strongly positive. Thus, confirming the findings already made in 4.3, it appears that the *Knowledge Panel* is shifting CTR from higher to lower positions. This strengthens the assumption made before that prominent features steer consumers' attention more actively. It can be suggested that the *Knowledge Panel* leads consumers' attention to the side of the page before returning to the results on the left at a lower point of the page, which explains the positive effect on lower results.

Concluding the above findings, it stands out that, depending on the position of a result, the appearance of SERP features can have varying impacts on the CTR regardless of whether the result is shown in them or not. Especially the results on position one suffer from an increase in SERP features. For results on lower positions, the result’s appearance in SERP features is beneficial, while not showing for SERP features is harmful to a result’s CTR. Overall, as it is impossible to forecast the exact rank for a result, the findings imply that a strategy aiming at generally being shown in SERP features is most promising. Additionally, the findings imply that the presence of a *Knowledge Panel* has the advantage that lower-ranking results receive a larger share of the CTR than when it is not shown. Consequently, providers competing in a more competitive market where the chances of ranking high are low could benefit from keywords with that feature.

A.3. Impact of intents on SERP features and their effects

According to a paper by Schultz (2020), depending on the consumer's search intent - the stage of the e-commerce journey - keyword characteristics like the complexity of a keyword differ. Consequently, different keywords promise higher CTR for providers at different stages of a user's e-commerce journey. Generally, one differentiates between four different intents (see chapter 3.2). The dataset at hand includes queries for all four intents. As one query can also have multiple intents at once, in order to be able to outline clear findings, this study took a differentiated look at intents overlapping and appearing alone. The data at hand never has more than two intents simultaneously. After excluding overlapping sets with a too-low number of appearances, six sets remain[7].

The conducted analysis of the keyword characteristics over these different intents confirms the assumption made by Schultz (2020): Different keyword characteristics benefit different search intents. With a navigational intent, for example, branded keywords outperform unbranded ones regardless of whether a keyword is complex or simple. For the informational and transactional intent, however, this effect is only true for simple keywords. For the latter, unbranded keywords outperform branded ones when the keywords are more complex.

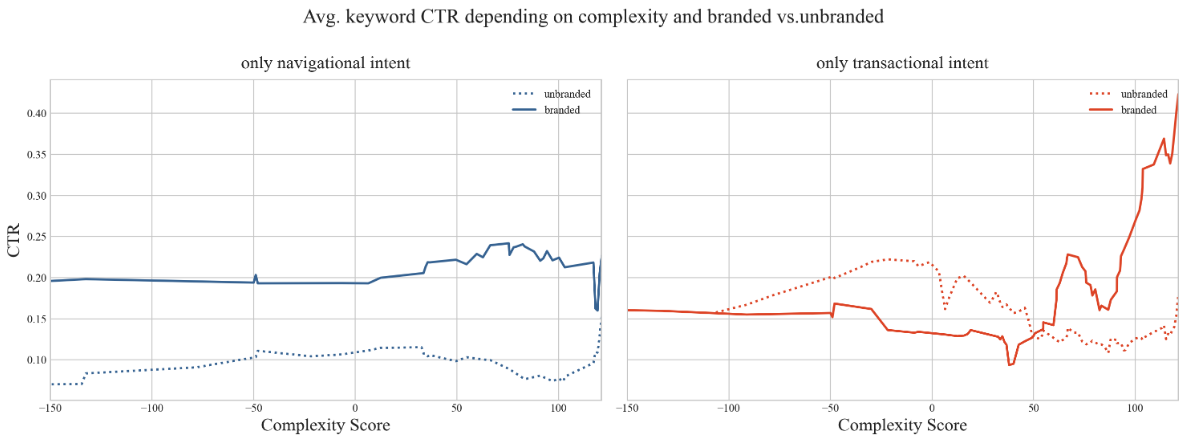


Figure A.3: Performance of keywords depending on their complexity and whether they are branded, for the navigational and transactional intent. A complete set of visualizations for all intents can be found in appendix A.II.

Additionally, the search intent causes a difference in consumer behavior. For example, the transactional and navigational intent combination has a significantly higher average CTR (average 0.28) than all others (average 0.18). This is the result of a clear consumer target, as the information about the location (navigational) together with the transactional purpose of

[7] (i) only transactional, (ii) only commercial, (iii) only navigational, (iv) only informational, (v) transactional with informational, (vi) transactional with navigational

buying a specified product leaves the search engine with relatively little space for interpretation. Thus, the desired results mostly appear on top of the result page leading to higher average CTRs, especially for the first positions. A sole navigational intent shows a similar picture.

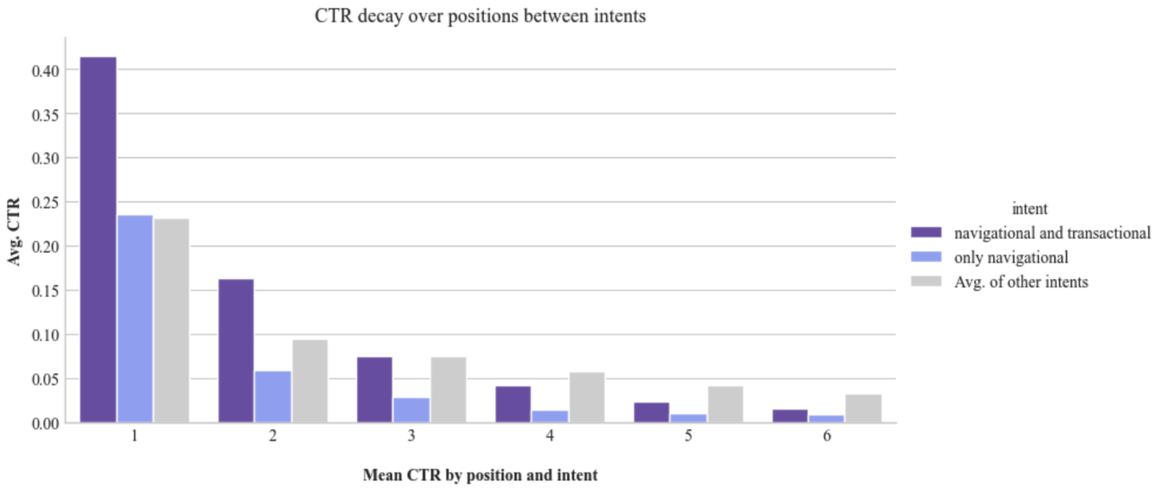


Figure A.4: CTR decay over positions between intents showcasing the significantly stronger decay for navigational intents (appendix A.III.)

Overall, the hypotheses made in the paper by Schultz (2020) can be confirmed as it was also found in the data at hand that the intent has a significant impact on the performance of different keywords. More importantly, however, this allows the question of whether a similar differentiation of behavior can also be detected when SERP feature characteristics and their effects are analyzed in the light of intents.

A.3.1 Special role of navigational intent

A first analysis of the total count of SERP features over intents shows that there are indeed noteworthy differences. While a search with only navigational intent has an average of over 5.5 SERP features per search, a solely informational search only has, on average, 4.5. This raises an interesting question: When more page SERP features harm the average CTR, and the navigational intent has the most SERP features on average, how can it be explained that this intent has the highest CTR on average? The main reason is most likely the strong performance on position one with a higher-than-average CTR. Additionally, we can see an intriguing pattern in the distribution of SERP features. The navigational intent includes more generally positively impacting features like a *Knowledge Panel*. Features that negatively impact CTR, like the *Video Carousel,* barely appear for navigational searches. Together with the strong performance on position one, it causes the higher average CTR. For website providers, this implies that the earlier recommended attention to SERP features does not necessarily apply for navigational intents. Instead, a stronger focus should be laid on ranking.

A.3.2 Intent favoring SERP features

Influential patterns in SERP feature distribution can also be found among other intents. Noteworthy, for example, is the strong presence of visual features in the combination of informational and transactional results. With an average of 5.3 SERP features per search, *Images* appear in almost 100% of the cases, followed by the *Image Pack* in about 81% and *Video Carousel* in around 59% of the cases. For all these features, that is a 20% increase towards all other intents[8].

It was argued above that prominent SERP features can shift CTR between positions. Such seems to also be the case for the SERP feature *Video Carousel.* However, the *Video Carousel*, comprising a set of videos related to the keyword, can logically not withstand the line of reasoning used for the *Knowledge Panel* as it only shows in one row and does not span over larger areas of a result page. When viewing the purpose of a video itself, one would assume that videos are mainly used as a source of information rather than a link to a marketplace. When analyzing this finding on an intent level, it shows that the *Video Carousel* is prominently featured on search queries with informational intent. It appears at least 27% more often for the combination of informational and transactional intent and at least 14% more often for the sole informational intent than for all others. It was found earlier that for those

[8] A visualization of the distribution of SERP features for all intents can be found in appendix A.IV

intents, the distribution of CTR is shown in a more shallow curve spreading more equally over positions leading to higher CTRs on lower positions (see appendix A.III). This suggests that the feature's perceived 'CTR shifting' effect can be directly linked to an intent effect. This reinstates the warning before, that the implications above need to be carefully interpreted, considering all influences simultaneously. Taking this into account, the remaining findings on intent level additionally underline the above made implications for SERP features.

A.4. National comparison of SERP feature effects

Until now, the analysis was conducted on a dataset with around 80,000 data points, including only results from the United States. Over the past years, however, various studies found a difference in consumer behavior over different nationalities (De Mooij 2017).

For SEO providers, this raises two questions: One, do search engines adapt their algorithms to these differences and potentially change the distribution of SERP features? Second, when consumer behavior changes, does this change the effect of different SERP features found before?

In order to research this field, a second dataset combining the same features with data collected in the United Kingdom was analyzed. In the following, the most important similarities and differences found in the study will be highlighted and put into perspective regarding the implications for providers. The new dataset includes, apart from the feature 'branded', the same features used in the US data. Additionally, it features over 800,000 entries allowing to verify the hypotheses made before on a larger sample size. When comparing the findings between the US and UK, generally it can be said that the consumer and search engine behavior appear rather similar in both countries. Results on position one strongly outperform all lower positions, with over two-thirds of the clicks on the top three results. Additionally, the correlation of position with CTR with -0.4 in the UK is similar to the data in the US and suggests that position is the strongest predictor for CTR.

A.4.1 Distribution of SERP features on a country level

The first question raised earlier was whether search engines adapt their algorithms to national differences in consumer behavior and consequently potentially change when SERP features are shown. Interestingly, it was found that the distribution of SERP features on result pages is highly similar between the two countries. This can be confirmed when analyzing the share of SERP features on an intent level. The appearances of the various SERP features are eminently similar in both countries, including the differences between navigational intents and others mentioned in chapter A.3.

Only the share of the features *People also ask* and *Video Carousel* drop significantly in the UK, which explains the overall lower average number of SERP features on a result page in the UK (4.2) opposed to the US (5.1).

Overall, the findings suggest that the search engine does not or only barely adapt its distribution of SERP features for different countries. However, in order to universalize the implications for providers made before, it needs to be analyzed whether the effect of these SERP features is also comparable for the two countries.

A.4.2 Effects of SERP features on a country level

Generally, it can be said that the effect of SERP features on click-through rates and consequently the user behavior appears similar in both countries for most elements, with two anomalies that will be elaborated shortly in the following.

Firstly, the effect of an *Image* shown next to a result in the UK is solely negative, regardless of whether the result itself includes an image or only other results do. This confirms the finding from chapter A.2.1, arguing that it is indeed the prominence of the *Image Pack* gaining the attention of users rather than the visual component.

Secondly, the *People also ask* section has an interesting effect on CTR. We recall that in the US, the feature behaved similarly to the *Image Pack*. It has a highly negative effect when shown but the result not being featured in it, and a strongly positive effect when the result is included. For the UK, on the contrary, the effect is solely and strongly positive in both cases. On the one hand, this strengthens the finding from chapter A.2.1, arguing a positive effect of features providing a user with alternatives to their queries. On the other hand, the effect also being positive when a feature does not appear seems counterintuitive and demands further research.

Overall, it can be summarized that the differences between user behavior in the e-commerce sector are largely similar over nationalities. Consequently, two findings can be concluded: One, that the search engine seems to not or only barely adapt its result pages to different nationalities. Second, that the effect of these features is vastly comparable over the two countries, suggesting a largely similar user behavior.

Ultimately all the above implies an extensive similarity, proposing that the aforementioned implications of SERP features and intents on SEO efforts can be universally adopted. Nevertheless, further research should be undertaken to also confirm these findings, including more cultural diversity.

A.5. Conclusion

This section concentrated on providing a framework to SEO professionals for utilizing the effects SERP features have on a result's CTR in different contexts.

Firstly, it was found that while, on average, more SERP features are decreasing CTR, it has to be distinguished between SERP features where the result is shown and such where it is not.

Ultimately, when optimizing for a keyword, it is found helpful to adapt the website to show for SERP features that are most beneficial for one's CTR. However, the risk that should the latter fail, the CTR can suffer accordingly must be considered. Extensive instructions on how to adapt a website accordingly can be researched on several SEO blogs (Moz 2022).

It additionally became apparent that a consumer's attention can be steered by prominent SERP features. These tend to cause consumers to focus on them, taking attention away from other organic results. Others, like the *Knowledge Panel,* shift user attention to lower parts of result pages impacting CTR distribution.

Secondly, it became eminent that the distribution and impact of SERP features are closely linked to a user's search intent. While some SERP feature effects are reinforced through intent effects, others are negatively impacted. Additionally, carefulness is required when separately interpreting the effects of SERP features with a concentrated distribution on specific intents, as some presumed patterns can result from an underlying intent effect.

Finally, this section proved a large extent of similarity of the aforementioned effects on a national level while highlighting minor differences in user behavior.

As SERP features can not be avoided on current result pages, the findings mentioned above can pose a guideline for SEO practitioners in the e-commerce sector on how to utilize specific SERP feature effects and use them to their advantage.

Cultural similarities and an e-commerce focus are found as the main limitation of this section.

This section suggests that the combination of the US and UK datasets is used for further analysis, as it offers a larger sample size while behaving similarly. Additionally, future research could build on the findings above and compare those to datasets from culturally more diverse nationalities to universally enhance the framework. A continuous repetition of this study is suggested to cover the constant enhancements of the search engine.

References

Agarwal, Abhineet, Yan S. Tan, Omer Ronen, Chandan Singh, and Bin Yu. 2022. "Hierarchical Shrinkage: Improving the Accuracy and Interpretability of Tree-Based Methods." *Proceedings of the 39th International Conference on Machine Learning* 162:111-135.

Alwosheel, Ahmed, Sander van Cranenburgh, and Caspar G. Chorus. 2018. "Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis." *Journal of Choice Modelling* 28:167-182.

Arik, Sercan O., and Tomas Pfister. 2021. "TabNet: Attentive Interpretable Tabular Learning." *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 8 (May): 6679-6687.

Bala, Madhu, and Deepak Verma. 2018. "A Critical Review of Digital Marketing." *International Journal of Management, IT & Engineering* 8, no. 10 (10).

Biau, Gérard, and Erwan Scornet. 2016. "A random forest guided tour." *TEST* 25 (April): 197-227.

Borisov, Vadim, Tobias Leeman, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. "Deep Neural Networks and Tabular Data: A Survey." *arXiv:2110.01889v3*, (June).

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning* 24 (August): 123-140.

Breiman, Leo. 2001. "Random forests." *Machine Learning* 45, no. 1 (October): 5-32.

Chen, Hengrui, Hong Chen, Zhizhen Liu, Xiaoke Sun, and Ruiyu Zhou. 2020. "Analysis of factors affecting the severity of automated vehicle crashes using XGBoost model combining POI data." *Journal of advanced transportation*. https://doi.org/10.1155/2020/8881545.

Chen, Junxuan, Baigui Sun, Hao Li, Hongtao Lu, and Xian-Sheng Hua. 2016. "Deep CTR Prediction in Display Advertising." *In Proceedings of the 24th ACM international conference on Multimedia* 16:811–820. https://doi.org/10.1145/2964284.2964325.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (August), 785-794.

Cheng, Heng-Tze, Levent Koc, Jeremiah Hermsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, et al. 2016. "Wide & Deep Learning for Recommender Systems." *DLRS 2016: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, (September), 7-10.

Chevalier, Stephanie. 2022a. "Global retail e-commerce sales 2026." Statista. Accessed October 26, 2022. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

Chevalier, Stephanie. 2022b. "Global online e-commerce traffic by source and medium." Statista. Accessed October 28, 2022. https://www.statista.com/statistics/820293/online-traffic-source-and-medium-e-commerce-sessions/.

Cui, Meng, and Songyun Hu. 2011. "Search Engine Optimization Research for Website Promotion." *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, (09). 10.1109/ICM.2011.308.

Das, Subhankar. 2021. *Search Engine Optimization and Marketing: A Recipe for Success in Digital Marketing*. N.p.: CRC Press/Taylor and Francis Group.

De Gryze, Steven, Ivan Langhans, and Martina Vandebroek. 2007. "Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression." *Chemometrics and Intelligent Laboratory Systems* 87:147-154.

De Mooij, Marieke. 2017. “Comparing dimensions of national culture for secondary analysis of consumer behavior data of different countries.” *International Marketing Review* 34 (3).

Deng, Wei, Junwei Pan, Tian Zhou, Deguang Kong, Aaron Flores, and Guang Lin. 2021. “DeepLight: Deep Lightweight Feature Interactions for Accelerating CTR Predictions in Ad Serving.” *WSDM '21: Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, (March), 922-930. https://doi.org/10.1145/3437963.3441727.

“The Effect of Paid Search on Organic Traffic.” 2020. Adlucent. Accessed November 27, 2022. https://www.adlucent.com/resources/blog/the-effect-of-paid-search-on-organic-traffic/.

Filzmoser, Peter, and Klaus Nordhausen. 2020. “Robust linear regression for high-dimensional data: An overview.” *WIREs Computational Statistics* 13, no. 4 (July).

Flesch, Rudolph. 1948. “A new readability yardstick.” *Journal of Applied Psychology* 32 (3): 221 - 223.

Freund, Yoav, and Robert E. Schapire. 1996. “Experiments with a New Boosting Algorithm.” *Machine Learning: Proceedings of the Thirteenth International Conference*.

Friedman, Jerome H. 2001. “Greedy function approximation: A gradient boosting machine.” *Annals of Statistics* 29, no. 5 (October): 1189-1232.

Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. N.p.: Springer New York.

Friedman, Jerome H., and Bodgan E. Popescu. 2008. “Predictive Learning via Rule Ensembles.” In *Annals of Applied Statistics*, 916-954. 2nd ed. Vol. 3. http://www.jstor.org/stable/30245114.

Gharibshah, Zhabiz, Xingquan Zhu, Arthur Hainline, and Michael Conway. 2020. "Deep Learning for User Interest and Response Prediction in Online Display Advertising." *Data Science and Engineering* 5:12-26.

Google. 2021. "An update to how we generate web page titles." Google Search Central. Accessed November 6, 2022. https://developers.google.com/search/blog/2021/08/update-to-generating-page-titles.

Google. 2022a. "Click through rate definition." Clickrate (CTR) - Google Ads-Support. Accessed September 20, 2022. https://support.google.com/google-ads/answer/2615875?hl=de.

Google. 2022b. "Enable Search result features for your site | Documentation." Google Developers. Accessed November 14, 2022. https://developers.google.com/search/docs/appearance/search-result-features.

Google. 2022c. "Data preprocessing for ML: options and recommendations | Cloud Architecture Center." Google Cloud. Accessed November 13, 2022. https://cloud.google.com/architecture/data-preprocessing-for-ml-with-tf-transform-pt1?hl=de.

Google. 2022d. "List of Google Search ranking updates." Accessed November 3, 2022. https://developers.google.com/search/updates/ranking.

Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. "Revisiting Deep Learning Models for Tabular Data." *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, (May).

Graepel, T., J. Q. Candela, T. Borchert, and R. Zerbrich. 2010. "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine." *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 13-20.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. 2022. “Why do tree-based models still outperform deep learning on tabular data?” *arXiv:2207.08815*, (July).

Grips. 2022. “Grips - About Us.” Grips - Transaction Intelligence for eCommerce. Accessed November 24, 2022. https://gripsintelligence.com/about.

Guo, Huifeng, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. “DeepFM: A Factorization-Machine based Neural Network for CTR Prediction.” *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 1725–1731.

Hancock, J.T., and T.M. Khoshgoftaar. 2020. “CatBoost for big data: an interdisciplinary review.” *Journal of Big Data* 7 (94): 1-45. https://doi.org/10.1186/s40537-020-00369-8.

Iqbal, Muhammad, Muhammad Noman Khalid, Amir Manzoor, Malik Muneeb Abid, and Nazir Ahmed Shaikh. 2022. “Search Engine Optimization (SEO): A Study of important key factors in achieving a better Search Engine Result Page (SERP) Position.” In *Sukkur IBA Journal of Emerging Technologies*, 1-15. Vol 6 ed. Vol. 1. http://journal.iba-suk.edu.pk:8089/sibajournals/index.php/sjcms/article/view/924/309.

Jais, Imran K., Amelia R. Ismail, and Syed Q. Nisa. 2019. “Adam Optimization Algorithm for Wide and Deep Neural Network.” *Knowledge Engineering and Data Science* 2 (1).

Kadra, Arlind, Marius Lindauer, Frank Hutter, and Josif Grabocka. 2021. “Well-tuned Simple Nets Excel on Tabular Datasets (NeurIPS 2021).” *35th Conference on Neural Information Processing Systems*.

Kingsnorth, Simon. 2022. *Digital Marketing Strategy: An Integrated Approach to Online Marketing*. N.p.: Kogan Page.

Kossen, Jannik, Neil Band, Clare Lyle, Aidan N. Gomez, Tom Rainforth, and Yarin Gal. 2021. “SelfAttention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning.” *arXiv:2106.02584*, (June).

Ledford, Jerri L. 2009. *Search Engine Optimization Bible*. N.p.: Wiley.

Lewandowski, Dirk, Sebastian Sünkler, and Nurce Yagci. 2020. "The influence of search engine optimization on Google's results: Amulti-dimensional approach for detecting SEO." *13th ACM Web Science Conference 2021* 21 (06): 12-20.

Ma, Chao, Yuze Liao, Yuan Wang, and Zhen Xiao. 2016. "F2M: Scalable Field-Aware Factorization Machines." *30th Conference on Neural Information Processing Systems*.

Messenger, Robert, and Lewis Mandell. 1971. "A Modal Search Technique for Predictive Nominal Scale Multivariate Analysis." *Journal of the American Statistical Association* 67, no. 340 (November): 768-772.

Morgan, James N., and John A. Sonquist. 1963. "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association* 58, no. 302 (June): 415-434.

Morningscore. 2022. "What are Title Tags, and do they help with SEO? [+13 writing tips]." Morningscore. Accessed November 26, 2022. https://morningscore.io/what-are-title-tags-in-seo/.

Moz. 2022. "What Is A SERP Feature? Common Types And How To Win Them." Moz. Accessed November 12, 2022. https://moz.com/learn/seo/serp-features.

Olson, Eric M., Kai M. Olson, Andrew J. Czaplewski, and Thomas Martin Key. 2021. "Business strategy and the management of digital marketing." *Business Horizons* 64, no. 2 (04): 285-293.

Omer, Sagi, and Lior Rokach. 2018. "Ensemble learning: A survey." *WIREs Data Mining Knowl Discov* 8 (4).

O'Neill, Aaron. 2022. "Japan GDP 1987-2027." Statista. Accessed November 27, 2022. https://www.statista.com/statistics/263578/gross-domestic-product-gdp-of-japan/.

Pan, Junwei, Jian Xu, Alfonso Lobos Ruiz, Wenliang Zhao, Shengjun Pan, Yu Sun, and Quan Lu. 2018. “Field-weighted Factorization Machines for Click-Through Rate Prediction in Display Advertising.” *In Proceedings of the 2018 World Wide Web Conference* 18.

Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna V. Dorogush, and Andrey Gulin. 2018. “CatBoost: unbiased boosting with categorical features.” *Advances in Neural Information Processing Systems* 31.

Quinlan, John R. 1986. “Induction of Decision Trees.” *Machine Learning* 1 (March): 81-106.

Richardson, M., E. Dominowska, and R. Ragno. 2007. “Predicting clicks: estimating the click-through rate for new ads.” *Proceedings of the 16th international conference on World Wide Web*, 521-530.

Sam-Martin, Kristina. 2020. *Google's featured snippets in the Context of Strategic Content Marketing*. N.p.: E-Book - Kristina SamMartin.

Schultz, Carsten D. 2020. “Informational, transactional, and navigational need of information: relevance of search intention in search engine advertising.” *Information Retrieval Journal* 23:117–135.

Semrush. 2021. “Basic docs.” Semrush Developers. Accessed September 22, 2022. https://developer.semrush.com/api/v3/analytics/basic-docs/#serp-features.

Semrush. 2022a. “A Beginner's Guide to Keyword Search Volume.” Semrush. Accessed November 3, 2022. https://www.semrush.com/blog/keyword-search-volume/.

Semrush. 2022b. “What Are SERP Features? An In-Depth Guide.” SEMrush. Accessed December 1, 2022. https://www.semrush.com/blog/serp-features-guide/#how-to-tell-if-your-competitors%E2%80%99-sites-have-serp-features.

Shih, Bih-Yaw, Chen-Yuan Chen, and Zih-Siang Chen. 2013. “Retracted: An Empirical Study of an Internet Marketing Strategy for Search Engine Optimization.” *Human Factors and Ergonomics in Manufacturing and Service Industries* 23, no. 6 (11).

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. 2012. “Practical Bayesian Optimization of Machine Learning Algorithms.” *Advances in Neural Information Processing Systems 25 (NIPS 2012)*.

Stevenson, Angus. 2010. *Oxford Dictionary of English*. N.p.: OUP Oxford, 2010.

Su, Xiaogang, Xin Yan, and Chih-Ling Tsai. 2012. “Linear regression.” *WIREs Computational Statistics* 4, no. 3 (April): 275-294.

Tenet Partners. 2020. “2020 Top 100 Most Powerful Brands,” The essential brand rises - adapting to core consumer needs. Ranking the brands. Accessed November 28, 2022. https://www.rankingthebrands.com/PDF/Top%20100%20Most%20Powerful%20Brands%202020,%20Tenet.pdf.

Tober, Marcus. 2022. “Zero-clicks Study.” Semrush. Accessed December 6, 2022. https://www.semrush.com/blog/zero-clicks-study/.

Turner, Ryan, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. 2021. “Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020.” *arXiv:2104.10201*, (April).

Uyanık, Gülden K., and Neşe Güler. 2013. “A study on mutliple linear regression analysis.” *Procedia - Social and Behavioral Sciences* 106 (December): 234-240.

Wang, Fang, Warawut Suphamitmongkol, and Bo Wang. 2013. “Advertisement Click-Through Rate Prediction using Multiple Criteria Linear Programming Regression Model.” *Procedia Computer Science* 17:803-811.

Wang, Lih-Werm, Michael J. Miller, Michael R. Schmitt, and Frances K. When. 2013. "Assessing readability formula differences with written health information materials: Application, results, and recommendations." *Research in Social and Administrative Pharmacy* 9 (5): 503-516. https://doi.org/10.1016/j.sapharm.2012.05.009.

Wang, Zhiqiang, Qingyun She, and Junlin Zhang. 2021. "MaskNet: Introducing Feature-Wise Multiplication to CTR Ranking Models by Instance-Guided Mask." *Proceedings of DLP-KDD 2021*.

Wheelhouse. 2022. "How SERP Features Can Negatively Impact Your Business (Updated 2022)." Wheelhouse DMG. Accessed December 9, 2022. https://www.wheelhousedmg.com/blog/important-google-serp-features-and-how-to-optimize-for-them/.

Yan, Ling, Wu-Jun Li, Gui-Rong Xue, and Dingyi Han. 2014. "Coupled Group Lasso for Web-Scale CTR Prediction in Display Advertising." *Proceedings of the 31st International Conference on Machine Learning* 32 (2): 802-810.

Yang, Yanwu, and Panyu Zhai. 2022. "Click-through rate prediction in online advertising: A literature review." *Information Processing & Management* 59, no. 2 (03).

Yuchin, Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. "Field-aware Factorization Machines for CTR Prediction." *Proceedings of the 10th ACM Conference on Recommender Systems* 16 (09): 43-50.

Zhang, Weinan, Jiarui Qin, Wei Guo, Ruiming Tang, and Xiuqiang He. 2021. "Deep learning for click-through rate estimation." *arXiv preprint arXiv:2104.10584*.

Zhou, Guorui, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. "Deep Interest Network for Click-Through Rate Prediction." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 18 (07): 1059-1068.

Ziakis, Christos, Maro Vlachopoulou, Theodosios Kyrkoudis, and Makrina Karagkiozidou. 2019. “Important Factors for Improving Google Search Rank.” *Future Internet* 11, no. 32 (01).

List of Figures

List of Tables

Appendix

I. SERP Features Illustration

Illustration of SERP features. For a detailed description and examples of each SERP feature we also recommend visiting the Semrush SERP feature guide: https://www.semrush.com/blog/serp-features-guide. Image source: Semrush.

II. Data Dictionary

The texts in the description fields with data source “Semrush” are taken from Semrush (2021).

Variable	Description	Data Source	Example
Ads	Any type of ad on the page. This can include ads on top of the search, at bottom, or shopping ads. Describes whether the SERP feature appears at least once on the page.	Self-engineered	1
AdWords Bottom	A series of ads (up to 4) that appear at the bottom of the first search results page. Describes whether the SERP feature appears at least once on the page.	Semrush	1
AdWords Top	A series of ads (up to 4) that appear at the top of the first search results page. Describes whether the SERP feature appears at least once on the page.	Semrush	1
AMP	AMP pages (i.e., results marked with the word "AMP" and a gray lightning bolt) shown in search results on mobile devices.	Semrush	1
Avg. Monthly Volume	Search volume averaged over months	Keywordplanner	1312.65
Branded	If the keyword includes a brand name.	Grips	1
Carousel	A row of horizontally scrollable images displayed at the top of search results. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Clicks	Number of clicks on targeted search result	Grips	25
Competition	Measure of competition on search term with 0.0 being the lowest and 1.0 being the highest.	Keywordplanner	0.22
Count Page SERP Feat.	Count of positive (=1) page SERP features for search result page.	Self-engineered	2
Count Positional SERP Feat.	Count of positive (=1) positional SERP features for search result.	Self-engineered	5

Variable	Description	Data Source	Example
CPC	Cost per click of ad entry.	Keywordplanner	2.34
CTR	Click-through rate (clicks divided by impressions).	Grips	0.25
Domain	Domain of the URL of each search result.	Semrush	'jomashop.com'
FAQ	A list of questions related to a particular search that shows up for a particular organic search result. When clicked on, each of the questions reveals the answer. Describes whether the SERP feature appears at least once on the page.	Semrush	1
FAQ (Positional)	A list of questions related to a particular search that shows up for a particular organic search result. When clicked on, each of the questions reveals the answer. Describes whether the result is featured in the SERP feature.	Semrush	1
Featured Images	A collection of images is usually displayed at the top of the SERP if Google considers visual results to be more relevant than text results. Only for mobile devices. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Featured Snippet	A short answer to a user's search query with a link to the third-party website it is taken from that appears at the top of all organic search results. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Featured Snippet (Positional)	A short answer to a user's search query with a link to the third-party website it is taken from that appears at the top of all organic search results. Describes whether the result is featured in the SERP feature.	Semrush	1
Featured Video	A video result to a search query that is displayed at the top of all organic search results. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Flights	A block that displays flights related to a search query. Flight results include information on flight dates, duration, the number of transfers and prices. Data is taken from Google Flights. Describes whether the SERP feature appears at least once on the page.	Semrush	1

Variable	**Description**	**Data Source**	**Example**
Hotels Pack	A block that displays hotels related to a search query. Hotel results include information on prices and rating, and allows users to check availability for certain dates. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Image	An image result with a thumbnail displayed along with other organic search results. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Image (Positional)	An image result with a thumbnail displayed along with other organic search results. Describes whether the result is featured in the SERP feature.	Semrush	1
Image Pack	A collection of images related to a search query that is usually displayed between organic search results. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Image Pack (Positional)	A collection of images related to a search query that is usually displayed between organic search results. Describes whether the result is featured in the SERP feature.	Semrush	1
Impressions	Impressions of the result, including estimates for results on the second search page..	Grips	100
Instant Answer	A direct answer to a user's search query that is usually displayed at the top of organic search results in the form of a gray-bordered box. Describes whether the SERP feature appears at least once on the page.	Semrush	1
Intent: Commercial	Trying to learn more before making a purchase decision (e.g. “Subaru vs. Nissan”)	Semrush	1
Intent: Informational	Trying to learn more about something (e.g., “What’s a good car?”)	Semrush	1
Intent: Navigational	Trying to find something (e.g., “Subaru website”)	Semrush	1
Intent: Transactional	Trying to complete a specific action (e.g., “buy Subaru Forester”)	Semrush	1

Variable	**Description**	**Data Source**	**Example**
Jobs Search	A number of job listings related to a search query that appear at the top of the search results page. Job listings include the job title, the company offering the job, a site where the listing was posted, etc. Describes whether SERP feature appears at least once on the page.	Semrush	1
Keyword	The searched keyword.	Semrush	'jomashop burberry scarf'
Keyword Complexity Score	Flesch reading ease score (Flesch 1948) of the keyword, with higher values indicating easier-to-read keywords.	Self-engineered	112
Keyword Count	Number of words the search query (keyword) consists of.	Grips	3
Keyword Difficulty	Difficulty to rank for given keyword expressed from 0 to 100	Semrush	87
Keyword Included in Domain	Keyword Included in Domain.	Self-engineered	0
Keyword Included in URL	Keyword Included in URL.	Self-engineered	1
Keyword Length	Number of characters in keyword	Grips	23
Knowledge Panel	Panel on right side of results that often includes images, facts, social media links, and other relevant information to the search query. Describes whether SERP feature appears at least once on the page.	Semrush	1
Knowledge Panel (Positional)	Panel on right side of results that often includes images, facts, social media links, and other relevant information to the search query. Describes whether the result is featured in the SERP feature.	Semrush	1
Local Pack	Embedded Google Maps frame on top of search results. Describes whether SERP feature appears at least once on the page.	Semrush	1

Variable	Description	Data Source	Example
Local Pack (Positional)	Embedded Google Maps frame on top of search results. Describes whether the result is featured in the SERP feature.	Semrush	1
Number of Results	Total number of results for search of keyword	Semrush	456789
People Also Ask	A series of questions that may relate to a search query that appears in an expandable grid box labeled "People also ask" between search results. Describes whether SERP feature appears at least once on the page.	Semrush	1
People Also Ask (Positional)	A series of questions that may relate to a search query that appears in an expandable grid box labeled "People also ask" between search results. Describes whether the result is featured in the SERP feature.	Semrush	1
Position	Position of the result among other search results	Semrush	3
Position (monthly)	Average position in the last month.	Grips	2.45
Position Difference	Position subtracted from previous position, i.e. positive values mean a decrease in position	Semrush	-1
Previous Position	Position last month	Semrush	2
Reviews	Organic search results marked with star ratings and including the number of reviews the star rating is based on. Describes whether SERP feature appears at least once on the page.	Semrush	1
Reviews (Positional)	Organic search results marked with star ratings and including the number of reviews the star rating is based on. Describes whether the result is featured in the SERP feature.	Semrush	1
Search Volume	The search volume of a keyword, cleaned from the original (inflated) value.	Grips	75643

Variable	Description	Data Source	Example
SERP Features by Keyword	SERP features listed by ID, before one-hot-encoded.	Semrush	1
Shopping Ads	A row of horizontally scrollable paid shopping results that appear at the top of a search results page for a brand or product search query, and include the website's name, pricing, and product image. Describes whether SERP feature appears at least once on the page.	Semrush	1
Site Links	A set of links to other pages of a website that is displayed under the main organic search result and for brand-related search queries. Describes whether SERP feature appears at least once on the page.	Semrush	1
Site Links (Positional)	A set of links to other pages of a website that is displayed under the main organic search result and for brand-related search queries. Describes whether the result is featured in the SERP feature.	Semrush	1
Top Stories	A card-style snippet presenting up to three news-related results relevant to user's search query, which is usually displayed between organic search results. Describes whether SERP feature appears at least once on the page.	Semrush	1
Top Stories (Positional)	A card-style snippet presenting up to three news-related results relevant to user's search query, which is usually displayed between organic search results. Describes whether the result is featured in the SERP feature.	Semrush	1
Trends	How much interest web searchers have shown in a given keyword in the last 12 months.	Semrush	1
Tweet	A card-style snippet displaying the most recent tweets related to a search query. Describes whether SERP feature appears at least once on the page.	Semrush	1
URL	The url of a google search result	Semrush	‘https://www.jomashop.com/burberry-4031051.html’

Variable	Description	Data Source	Example
Video	Video results with a thumbnail displayed along with other organic search results. Describes whether SERP feature appears at least once on the page.	Semrush	1
Video (Positional)	Video results with a thumbnail displayed along with other organic search results. Describes whether the result is featured in the SERP feature.	Semrush	1
Video Carousel	A row of horizontally scrollable videos displayed among search results. Describes whether SERP feature appears at least once on the page.	Semrush	1
Video Carousel (Positional)	A row of horizontally scrollable videos displayed among search results. Describes whether the result is featured in the SERP feature.	Semrush	1

III. Feature Subsets

Subset	Features		
Position	Position	Position Difference	Position (monthly)
Keyword	Number of Results	Intent: Transactional	Keyword Length
	Keyword Difficulty	Competition	Keyword Count
	Intent: Commercial	CPC	Branded
	Intent: Informational	Search Volume	Complexity Score
	Intent: Navigational	Avg. Monthly Search Vol.	
SERP Feat.	Instant Answer	Featured Snippet	Site Links (Positional)
	Knowledge Panel	Image	Reviews (Positional)
	Carousel	Jobs Search	Video (Positional)
	Local Pack	Video Carousel	Featured Snippet (Positional)
	Top Stories	People Also Ask	Image (Positional)
	Image Pack	FAQ	Video Carousel (Positional)
	Site Links	Flights	People Also Ask (Positional)
	Reviews	Knowledge Panel (Positional)	FAQ (Positional)
	Tweet	Local Pack (Positional)	Ads
	Video	Top Stories (Positional)	Count Page SERP Feat.
	Featured Video	Image Pack (Positional)	Count Positional SERP Feat.
Result	Keyword in Domain	Keyword in URL	Domain *(One-Hot-Encoded)*

IV. Tested Hyperparameters

Model	Parameter	Values considered for hyperparameter tuning
Random Forest	Max depth	Categorical: [None, 2, 3, 4], p=[0.7, 0.1, 0.1, 0.1]
	Number of estimators	Integer Log Uniform: 10 → 3000
	Criterion	Categorical: [‘squared_error’, ‘absolute_error’]
	Max features	Categorical: ['sqrt', 'log2', None, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
	Min samples split	Categorical: [2, 3], p=[0.95, 0.05]
	Min samples leaf	Integer Log Uniform: 1 → 50
	Bootstrap	Categorical: [True, False]
	Min impurity decrease	Categorical: [0.0, 0.01, 0.02, 0.05], p=[0.85, 0.05, 0.05, 0.05]
GBDT	Loss	Categorical: [‘squared_error’, ‘absolute_error’, ‘huber’]
	Learning rate	Real Log Uniform: 0.01 → 10.0
	Subsample	Real Uniform: 0.5 → 1.0
	Number of estimators	Integer Log Uniform: 10 → 1000
	Criterion	Categorical: [‘friedman_mse’, ‘squared_error’]
	Max depth	Categorical: [None, 2, 3, 4, 5], p=[0.1, 0.1, 0.6, 0.1, 0.1]
	Min samples split	Categorical: [2, 3], p=[0.95, 0.05]
	Min samples leaf	Integer Log Uniform: 1 → 50
	Min impurity decrease	Categorical: [0.0, 0.01, 0.02, 0.05], p=[0.85, 0.05, 0.05, 0.05]
	Max leaf nodes	Categorical: [None, 5, 10, 15], p=[0.85, 0.05, 0.05, 0.05]
XGBoost	Max depth	Integer Uniform: 1 → 11
	Number of estimators	Integer Uniform: 100 → 1000
	Min child weight	Integer Log Uniform: 1 → 100
	Subsample	Real Uniform: 0.5 → 1.0
	Learning rate	Real Log Uniform: 1e-5 → 0.7
	Col sample by level	Real Uniform: 0.5 → 1.0
	Col sample by tree	Real Uniform: 0.5 → 1.0
	Gamma	Real Log Uniform: 1e-8 → 7.0
	Lambda	Real Log Uniform: 1.0 → 4.0
	Alpha	Real Log Uniform: 1e-8 → 100.0

Model	Parameter	Values considered for hyperparameter tuning
CatBoost	Max depth	Integer Uniform: 3 → 10
	Learning rate	Real Log Uniform: 1e-5 → 1.0
	Bagging temperature	Real Uniform: 0.0 → 1.0
	L2 leaf regression	Real Log Uniform: 1.0 → 10.0
	Leaf estimation iterations	Integer Uniform: 1 → 10
TabNet	Number decision steps	Categorical: [3, 5, 10]
	Layer size	Categorical: [8, 16, 64]
	Learning rate	Categorical: [0.01, 0.02]
Wide&Deep	Layer 1 size	Categorical: [64, 128, 256]
	Layer 2 size	Categorical: [64, 128, 256]
	Layer 3 size	Categorical: [None, 64, 128, 256]
	Dropout ratio	Categorical: [0.0, 0.01, 0.05]
	Embedding dimension	Categorical: [4, 16, 32]
DeepFM	Layer 1 size	Categorical: [64, 128, 256]
	Layer 2 size	Categorical: [64, 128, 256]
	Layer 3 size	Categorical: [None, 64, 128, 256]
	Use batch normalization	Categorical: [True, False]
	Dropout ratio	Categorical: [0.0, 0.01, 0.05]
	Embedding dimension	Categorical: [4, 16, 32]

V. Used Hyperparameters

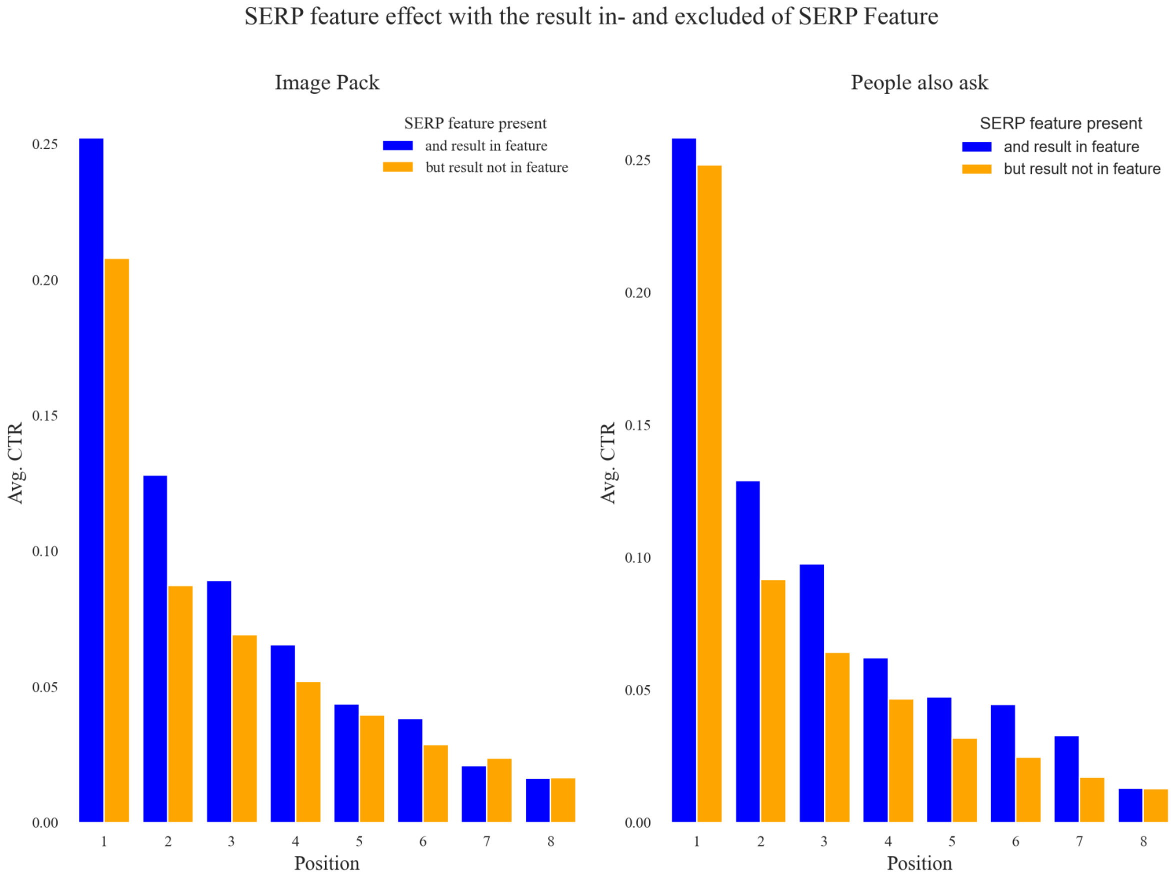
Model	Parameter	Default	Used for Subset: Position	Position + Keyword	Position + SERP Feat	Position + Result	Position + Keyword + SERP Feat	Position + Keyword + SERP Feat + Result
Random Forest	Max depth	None	None	None	None	None	None	None
	Number of estimators	100	897	823	3000	3000	175	516
	Criterion	squared_ error	squared_ error	squared_ error	squared_ error	squarred_ error	squarred__ error	squarred__ error
	Max features	1.0	0.1	sqrt	0.3	0.4	0.3	0.6
	Min samples split	2	3	2	2	3	2	2
	Min samples leaf	1	38	5	8	9	1	2
	Bootstrap	True	True	False	False	True	False	True
	Min impurity decrease	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GBDT	Loss	squarred_ error	squared_ error	squared_ error	squared_ error	squared_ error	squared_ error	huber
	Learning rate	0.10	0.067	0.033	0.030	0.351	0.126	0.767
	Subsample	1.0	0.517	1.0	0.711	0.833	0.634	0.819
	Number of estimators	100	128	497	220	196	487	268
	Criterion	friedman_ mse	friedman_ mse	friedman_ mse	friedman_ mse	friedman_ mse	friedman_ mse	friedman_ mse
	Max depth	3	2	None	None	4	5	2
	Min samples split	2	3	2	2	3	3	2
	Min samples leaf	1	24	50	1	10	24	9
	Min impurity decrease	0.0	0.010	0.050	0.050	0.010	0.010	0.010
	Max leaf nodes	None	15	None	15	5	5	15
XGBoost	Max depth	6	11	11	8	11	11	11
	Number of estimators	100	406	689	603	786	460	714
	Min child weight	1	100	1	1	1	1	1
	Subsample	1.0	1.0	0.993	1.0	1.0	1.0	0.668
	Learning rate	0.30	0.015	0.030	0.035	0.016	0.039	0.031
	Col sample by level	1.0	0.50	0.50	0.50	0.50	0.917	0.50
	Col sample by tree	1.0	1.0	1.0	0.50	0.50	0.708	1.0

Model	Parameter	Default	Used for Subset					
			Position	Position + Keyword	Position + SERP Feat	Position + Result	Position + Keyword + SERP Feat	Position + Keyword + SERP Feat + Result
	Gamma	0.0	1.0e-08	7.7e-05	0.009	1.0e-08	1.0e-08	0.001
	Lambda	1.0	4.0	4.0	2.859	1.0	4.0	1.0
	Alpha	0.0	1.0e-08	0.080	0.113	1.0e-08	0.014	1.0e-08
CatBoost	Max depth	6	5	10	9	9	10	9
	Learning rate	0.030	0.013	0.041	0.033	0.026	0.056	0.054
	Bagging temperature	1.0	0.885	1.0	1.0	0.0	1.0	0.0
	L2 leaf regression	3.0	6.601	10.0	10.0	10.0	10.0	10.0
	Leaf estimation iterations	*Dynamic*	10	10	1	6	10	10
TabNet	Num decision steps	3	3	3	3	3	3	3
	Layer size	8	8	8	8	8	8	64
	Learning rate	0.02	0.02	0.02	0.02	0.02	0.02	0.025
Wide &Deep	Layer 1 size	256	256	256	256	256	256	256
	Layer 2 size	128	128	128	128	128	128	128
	Layer 3 size	64	64	64	64	64	64	64
	Dropout ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.01
	Embedding dimension	4	4	4	4	4	4	16
DeepFM	Layer 1 size	256	256	256	256	256	256	256
	Layer 2 size	128	128	128	128	128	128	128
	Layer 3 size	64	64	64	64	64	64	64
	Use batch normalization	False	False	False	False	False	False	False
	Dropout ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.01
	Embedding dimension	4	4	4	4	4	4	16

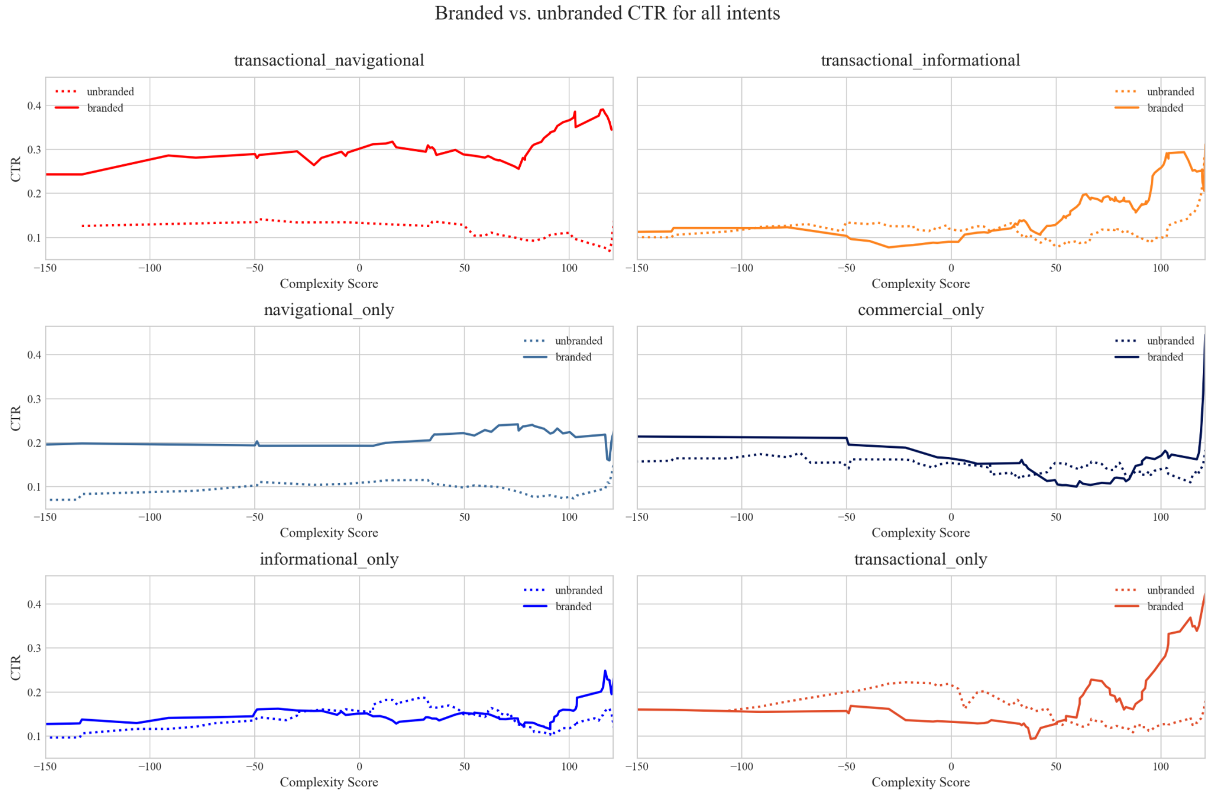
VI. Improvement Through Hyperparameter Tuning

		Improvement over default hyperparameters					
		Position	**Position + Keyword**	**Position + SERP Feat**	**Position + Result**	**Position + Keyword + SERP Feat**	**Position + Keyword + SERP Feat + Result**
Linear Regression	OLS	-	-	-	-	-	-
	Poly2	-	-	-	-	-	-
	Ridge	-	-	-	-	-	-
Tree-Based	Random Forest	0.009	0.003	0.007	0.008	0.003	0.01
	GBDT	0.000	0.009	0.002	0.003	0.007	0.007
	XGBoost	0.001	0.005	0.001	0.001	0.004	0.003
	CatBoost	0.000	0.003	0.000	0.000	0.002	0.001
Neural Network	TabNet	-	-	-	-	-	0.033
	Wide&Deep	-	-	-	-	-	0.010
	DeepFM	-	-	-	-	-	0.013

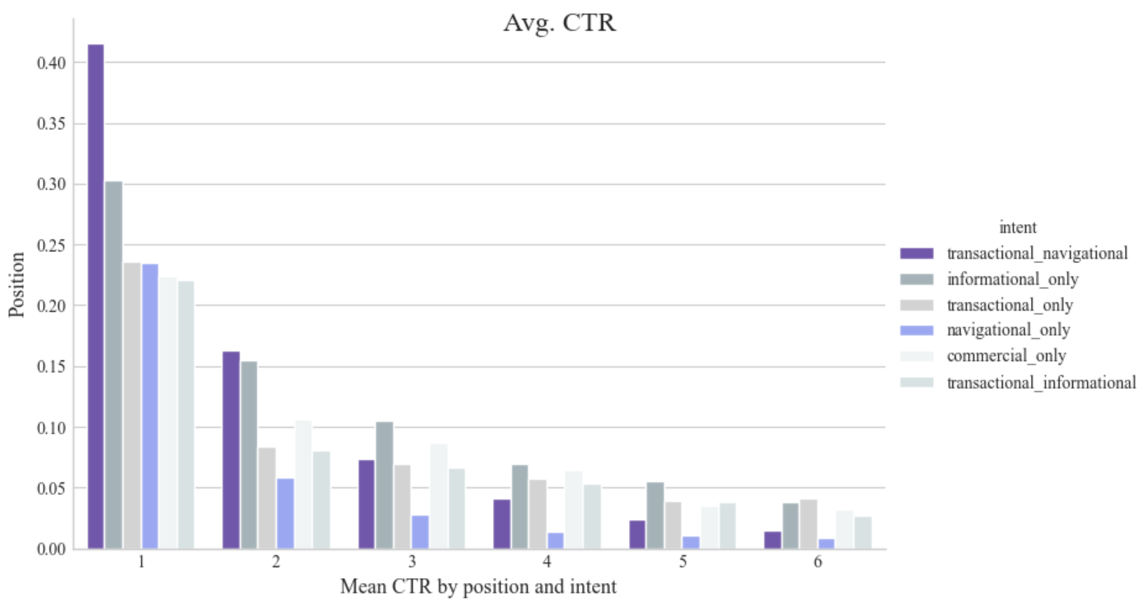
A.I. Effect for Image Pack and People also ask when the result is featured & when the result is not featured



A.II. Complexity vs. branded over intents



A.III. Avg. CTR decay over all intents



A.IV. Distribution of SERP features over different intents