4-2021

The role of comparison sites and image features in consumer search

Peng Yam KOH
Singapore Management University

Citation

KOH, Peng Yam. The role of comparison sites and image features in consumer search. (2021). 1-130.
Available at: https://ink.library.smu.edu.sg/etd_coll/343

THE ROLE OF COMPARISON SITES AND IMAGE FEATURES IN CONSUMER SEARCH

KOH PENG YAM (ALFRED)

SINGAPORE MANAGEMENT UNIVERSITY

2021

The Role of Comparison Sites and Image Features in Consumer Search

Koh Peng Yam (Alfred)

Submitted to Lee Kong Chian School of Business
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy in Marketing

Dissertation Committee:

Ernst C. Osinga (Supervisor / Chair)
Associate Professor of Marketing
Lee Kong Chian School of Business
Singapore Management University

Sandeep R. Chandukala
Associate Professor of Marketing
Lee Kong Chian School of Business
Singapore Management University

Kapil R. Tuli
Lee Kong Chian Professor of Marketing
Lee Kong Chian School of Business
Singapore Management University

Bart B. Bronnenberg
Professor of Marketing
Tilburg School of Economics and Management
Tilburg University

2021

I hereby declare that this PhD dissertation is my original work
and it has been written by me in its entirety.
I have duly acknowledged all the sources of information
which have been used in this dissertation.

This PhD dissertation has also not been submitted for any degree
in any university previously.

Koh Peng Yam (Alfred)
19 April 2021

Acknowledgements

First, I wish to thank my advisor, Prof. Osinga, as he taught me how to write, process data, and estimate models. Without him, I wouldn’t have completed my dissertation. He also took the brunt of my emotional episodes and was always understanding. I still remember being awed by the state-space models which he introduced to the field, and hope to employ it someday.

Second, thanks to Prof. Chandukala who taught me how to do Bayesian estimation which now looks a lot less intimidating than before; I will put your tools to good use in the future.

Third, Prof. Tuli showed me how to theorize from a different perspective as I came from a world which uses a deductive approach. This was really an eye-opener and one which I will remember.

Fourth, Prof. Bronnenberg for being on my committee as your valuable feedback and time helped me to develop myself further as an information search scholar.

Finally, I wish to thank my wife for her patience and understanding as I am full of angst when I work, and for choosing me when she knew I will have diminished earnings for five years. Also, my brother for always being a liquidity providers during the times when I ran dry, and when my stipend ran out.

To my late mother, your belief that education changes and empowers lives has now come true for me and family; I only wish it came in time for me to change your life.

Table of Contents

Introduction

Search is the costly activity of gathering information on an alternative or product (Stigler 1961). For instance, consumers spend on average 6.6 months to search for houses (Chernobai and Hossain 2012) and 15.26 hours to search for cameras (Bronnenberg, Kim and Mela 2016). To reduce search costs, two-sided platforms which match sellers and buyers have emerged, such as comparison sites (Bakos 2001). With the advent of such platforms, new data and new methods are available for researchers to analyze how consumers search. For instance, my first essay uses new forms of data to examine the role of comparison sites in the consumer's search. This new data is the consumers' pre-search characteristics which are surveyed before their online search begins, and which I integrate with the consumers' subsequently observed online search behavior. In my second essay, I utilize state-of-the-art machine learning methods to examine the interactions between a listing's image features which predicts consumer's clicks on the listing, where interactions are important for theory building. Image features are a relatively new form of data available to scholars and have only been recently studied. Below, I motivate the importance of each essay before ending with the overall importance of my dissertation.

In my first essay, I examine how a consumer's initial consideration set and expectations of finding a better deal affect the probability of a comparison site visit, where the initial consideration set is the set of alternatives which the consumer considers at the start of search. Since consumers visit a comparison site to discover new alternatives (Chen and Waldfogel 2005, Moraga-Gonzalez and Wildenbeest 2011), it seems less important for firms to be in the initial consideration set. Specifically, do consumers first search their initial consideration set before looking for new alternatives, or do they override their initial consideration set with information from the comparison sites? Also, if consumers search their initial consideration set first, would they wish to visit a comparison site? To resolve these puzzles, I tap on sequential search to posit a stylized search model relating (i) initial consideration set and expectations of finding a better deal to comparison site visit likelihood (throughout search, at the start of search, and at the end of search) and the (ii) expectations of finding a better deal to the probability of discovering a new alternative that is not in the initial consideration set.

The second essay of my dissertation examines how consumers search for listings on an online property platform, by systematically detecting the interactions between features (i.e., consumer and the listing's characteristics) which predict the target (i.e., clicks on a listing), when researchers do not know a priori which interactions and their functional form are informative of the target. Current management literature focuses on exploring pairwise combinations of features with the highest importance scores as potential interaction candidates (Choudhury et al. 2021), where importance scores measure each feature's relative influence on predicting the target. However, this approach is problematic as it (i) produces many interaction candidates, (ii) assumes that individually important features interact which is not necessarily true, and (iii) does not rank the detected interactions. To resolve these issues, I introduce two machine learning (ML) methods for detecting interactions from the ML literature, i.e., gRITs (Basu et al. 2018) and H-statistic (Friedman and Popescu 2008). I compare gRITs and the H-statistic against literature's current approach and discuss possible differences in their ability to detect interactions, before using simulations to ascertain the best method for correctly detecting, ranking and inferring the directionality of the interactions.

I apply the best method on a large scale consumer search dataset and identify the effect of image features on clicks, by exploiting the characteristic that listings in the same condominium have the same layout to hold their non-image quality constant. This study determines the best-performing ML method for detecting, ranking and visualizing interactions and produces the substantive insights that image features interact to predict a listing's click probability.

Overall, by showing that (i) initial consideration set and expectations of finding a better deal predicts comparison site visit timing, and (ii) interactions between image features are amongst the most important interactions for predicting clicks, I deepen understanding of consumer search. Methodologically, I show the value of combining survey and observational data and demonstrate how novel machine learning techniques can help us develop richer search theories by exploring boundary conditions between features.

Essay 1

On the Interplay between Initial Consideration Sets and Comparison Sites in Consumer Search

Abstract

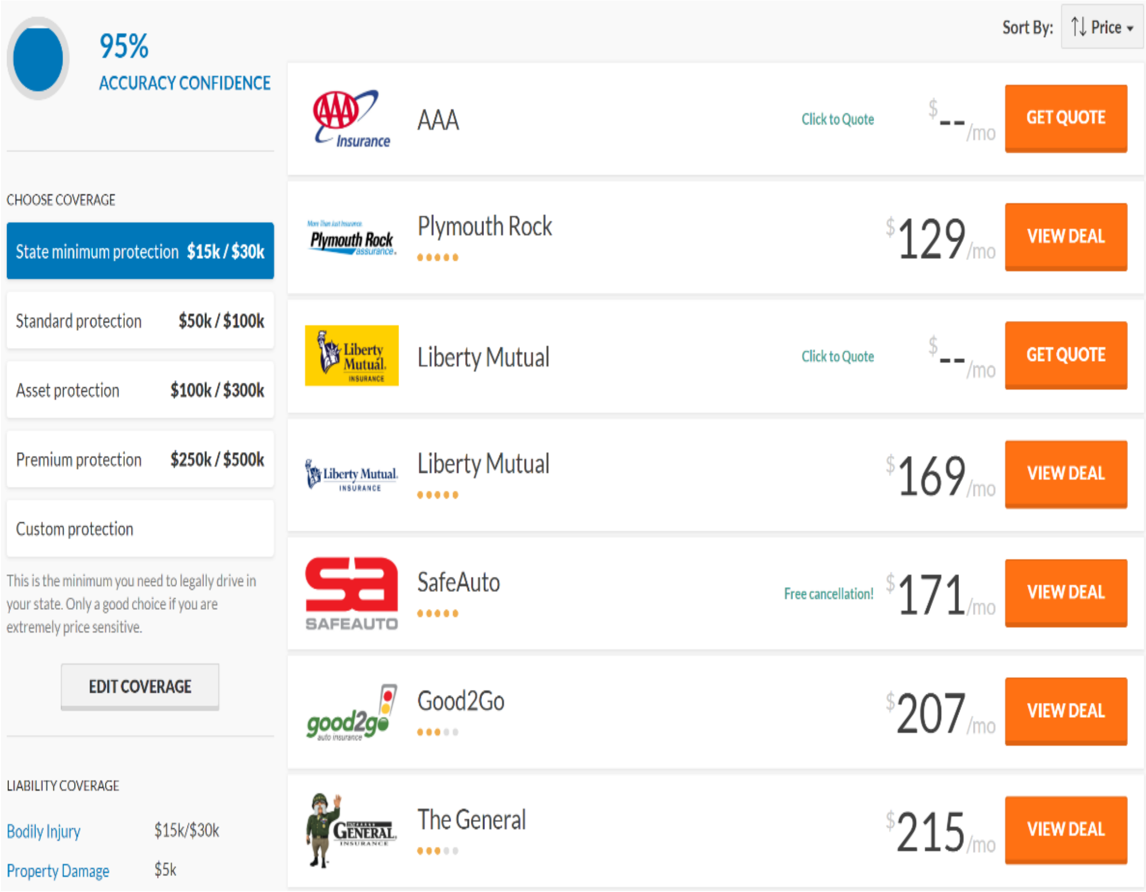
Comparison sites are widely used by consumers. Theory assumes that consumers visit these sites to discover new alternatives, raising questions about the role of the initial consideration set (alternatives considered at the start of search) when comparison sites are available. Will consumers ignore their initial consideration set and directly explore new alternatives? Will consumers with large initial consideration sets avoid comparison sites? Utilizing search and incomplete knowledge theories, the authors intuit that consumers first search their initial consideration set, and visit a comparison site to reduce the search costs of doing so. If a suitable alternative is absent, consumers subsequently visit a comparison site to discover new alternatives. The authors test their expectations on unique data capturing consumers' initial consideration sets and online search and find strong support. Specifically, consumers search a greater proportion of their initial consideration set at the start of search and are more likely to visit a comparison site when their initial consideration set is large. Additionally, consumers are more likely to visit a comparison site when they expect to find a better deal, particularly at the end of search. Finally, only consumers expecting to find a better deal are more likely to explore alternatives not in their initial consideration set.

Keywords: comparison sites, search theory, digital commerce

1. Introduction

Comparison sites allow consumers to compare alternatives based on price and non-price attributes from different sellers (Koçaş and Bohlmann 2008; Natter, Ozimec, and Kim 2015; Ringel and Skiera 2016; Smith 2002). For instance, Insurify offers live quotes from 102 sellers (i.e., insurers) in 48 US states and is the largest online US marketplace for auto insurance (Businesswire 2016). Upon entering information requested by the site such as demographics and the car's age, the site displays quotations from these sellers which are sorted by premiums in ascending order (refer to Figure 1).

Figure 1. Insurify Auto Insurance Results Page



95%
ACCURACY CONFIDENCE

Sort By: Price

CHOOSE COVERAGE

State minimum protection	$15k / $30k
Standard protection	$50k / $100k
Asset protection	$100k / $300k
Premium protection	$250k / $500k
Custom protection	

This is the minimum you need to legally drive in your state. Only a good choice if you are extremely price sensitive.

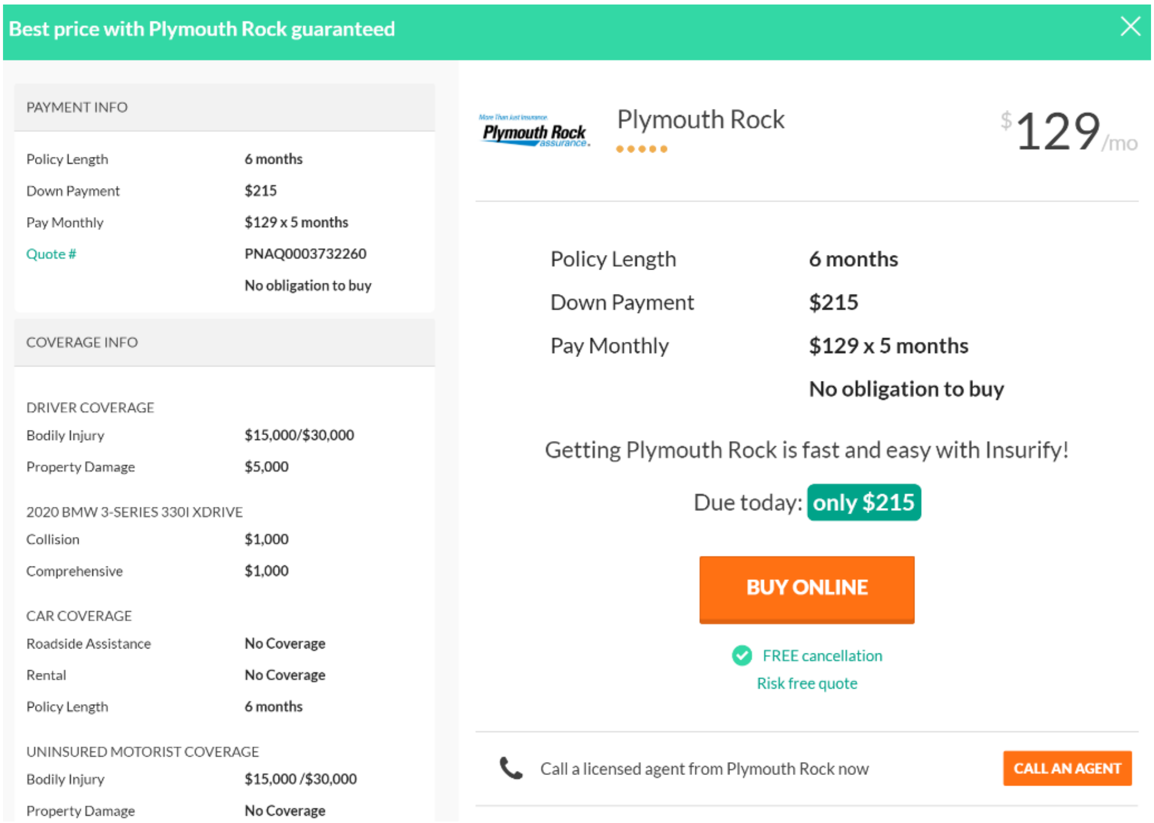
LIABILITY COVERAGE

Bodily Injury	$15k/$30k
Property Damage	$5k

AAA	Click to Quote	$--/mo
Plymouth Rock ●●●●●		$129/mo
Liberty Mutual	Click to Quote	$--/mo
Liberty Mutual ●●●●●		$169/mo
SafeAuto ●●●●●	Free cancellation!	$171/mo
Good2Go ●●●●●		$207/mo
The General ●●●●●		$215/mo

On the comparison site, consumers can obtain additional information such as policy exclusions by clicking on each seller (refer to Figure 2).

Figure 2. Additional Information Shown Upon Clicking a Seller



Best price with Plymouth Rock guaranteed

PAYMENT INFO

Policy Length	6 months
Down Payment	$215
Pay Monthly	$129 x 5 months
Quote #	PNAQ0003732260
	No obligation to buy

COVERAGE INFO

DRIVER COVERAGE	
Bodily Injury	$15,000/$30,000
Property Damage	$5,000
2020 BMW 3-SERIES 330I XDRIVE	
Collision	$1,000
Comprehensive	$1,000
CAR COVERAGE	
Roadside Assistance	No Coverage
Rental	No Coverage
Policy Length	6 months
UNINSURED MOTORIST COVERAGE	
Bodily Injury	$15,000 /$30,000
Property Damage	No Coverage

Plymouth Rock ●●●●● $129/mo

Policy Length	6 months
Down Payment	$215
Pay Monthly	$129 x 5 months
	No obligation to buy

Getting Plymouth Rock is fast and easy with Insurify!

Due today: only $215

BUY ONLINE

✓ FREE cancellation

Risk free quote

☎ Call a licensed agent from Plymouth Rock now CALL AN AGENT

Depending on the specific comparison site, consumers may complete the transaction on the comparison site (Natter, Ozimec, and Kim 2015), or be directed to the seller's site to complete the transaction (Smith 2002). In the UK, about 70% of Internet users visit comparison sites when searching for offers in utilities and financial services (Ronayne 2018) and, in Germany, about 50% of consumers purchase their utilities online primarily from comparison sites (Natter, Ozimec, and Kim 2015). The importance of comparison sites in consumer search is further exemplified by the high fee of approximately $60 to $85 charged by comparison sites in the UK for each consumer switching utilities alternatives[1] (Ronayne 2018).

[1] We use the term 'alternative' to refer to a differentiated product sold by each seller on the comparison site. For instance, insurance products are differentiated along price attributes (premiums) and nonprice attributes (e.g., customer service, coverage). Similarly, utilities providers are differentiated by their prices and nonprice attributes such as service reliability.

Existing theoretical work assumes that consumers visit comparison sites to discover new alternatives (e.g., Baye and Morgan 2001, Moraga-Gonazlez and Wildenbeest 2012), where new alternatives refer to alternatives which the consumer did not consider before visiting the comparison site. Congruent with this assumption, Waldfogel and Chen (2006) find that consumers who visit comparison sites (e.g., Pricescan) reduce their share of visits to well-known retailers (e.g., Amazon) and increase their share of visits to small unknown retailers. Hence, this literature suggests that, in the presence of comparison sites, it is less important for firms to be included in the consumer's initial consideration set, i.e., the set of alternatives which the consumer considers at the start of search. In fact, it makes one wonder about the role of the initial consideration set in consumer search when consumers have access to comparison sites. Do consumers first search their initial consideration set before looking for new alternatives, or do they directly augment or override their initial consideration set with information from comparison sites? Additionally, assuming consumers search their initial consideration set first, do they have a reason to visit a comparison site?

In this paper, we draw on optimal sequential search (Weitzmann 1979) and incomplete knowledge (Fox and Tversky 1995) theories to develop a stylized consumer search model involving initial consideration sets, comparison site visits, and the discovery of new alternatives. We argue that consumers first search alternatives in their initial consideration set and they visit comparison sites at the start of their search to efficiently evaluate alternatives in this set. When this initial search does not give a satisfactory outcome, consumers expecting to find a deal better than their current alternative, then visit comparison sites to explore new alternatives not in their initial consideration set. Our model thus integrates two different uses of comparison sites, (i) evaluating alternatives in the initial consideration set at the start of search and (ii) new alternative discovery at the end of search.

Using this model, we develop six testable expectations involving consumers' initial consideration sets and comparison site use.

Testing these expectations is challenging empirically as one needs to observe a consumer's initial consideration set and online search behavior. For example, Waldfogel and Chen (2006) do not observe the initial consideration set and proxy new alternatives by small unknown retailers. However, these small unknown retailers may be in the consumer's initial consideration set and thus leave unanswered the role of the initial consideration set in the consumer's search when comparison sites are available. We are fortunate to have access to a unique dataset containing consumers' initial consideration sets and their online search obtained from a survey and a browser extension. Thus, we identify truly new alternatives which are discovered. Further, we obviate reverse causality concerns (i.e., the comparison site visit affects the initial consideration set) as we observe the initial consideration set before consumers commence their online search.

We have the following key findings. First, consumers search a larger proportion of alternatives in their initial consideration set at the start than at the end of their search, even if comparison sites are available. As consumers have private information and trust in their own information, they first assess their initial consideration set. Second, initial consideration set size is positively associated with comparison site visit incidence, where the association is relatively stronger at the start of search. Intuitively, comparison sites reduce search costs and this reduction scales with initial consideration set size. Third, consumers with stronger expectations of finding a deal better than their current alternative have a higher comparison site visit incidence, where the relative effect is stronger at the end of search. Finally, we find that consumers with stronger expectations of finding a better deal are more likely to discover new alternatives not in their initial consideration set.

Our empirical findings have important theoretical and managerial implications. Scholars may think that comparison sites' presence renders the initial consideration set insignificant due to the richer set of alternatives displayed on these sites when compared to said set. However, we find that consumers with a larger initial consideration set tend to visit the comparison site with a higher probability, especially at the start of their search. Hence, scholars must recognize the dual use of comparison sites by consumers at different times of their search, i.e., evaluate their initial consideration set at the start and discover new alternatives at the end of search. Managerially, our empirical findings highlight the importance of being in a consumer's initial consideration set because, even in the presence of comparison sites, consumers first search their initial consideration set before exploring new alternatives not in this set.

In the next section, we develop our stylized search model. Thereafter, we detail our unique dataset and variable operationalization and introduce our model specification. We then present our empirical results and end with a discussion of our findings and their implications for scholars, practitioners, and policymakers.

2. Stylized search model

In the following sections, we develop our stylized search model involving the consumer's initial consideration set and expectations of finding a better deal, by integrating search (Weitzmann 1979) and incomplete knowledge (Fox and Tversky 1995) theories with the extant comparison sites literature (e.g., Baye and Morgan 2001, Waldfogel and Chen 2006, Moraga-Gonzalez and Wildenbeest 2012). We derive six expectations from our stylized search model.

2.1 Perspectives on consumer search

Two perspectives of consumer search exist in marketing (Honka and Chintagunta 2017; Honka, Hortaçsu, and Wildenbeest 2019). The first perspective is *simultaneous search* where consumers pre-commit to searching a set of alternatives prior to search (Stigler 1961). The second perspective is *sequential search*, where "consumers determine, after each utility realization, whether to continue searching or to stop" (Honka, Hortaçsu and Wildenbeest 2019). The most general form of sequential search is analyzed by Weitzmann (1979). As explained by Honka, Hortaçsu, and Wildenbeest (2019), consumers calculate their reservation utilities for each alternative, which are the expected utilities net of search costs for the alternative. Next, they search alternatives in order of decreasing reservation utilities (i.e., selection rule). Consumers terminate their search if the maximum of the realized utilities amongst the searched alternatives exceeds the maximum of the reservation utilities from all unsearched alternatives (i.e., stopping rule), and choose the alternative with the highest utility (i.e., choice rule).

Rational consumers prefer sequential search over simultaneous search as they can stop searching and save on search costs if a good offer is found earlier in their search (Honka and Chintagunta 2016). In contrast, simultaneous search requires consumers to continue searching the remaining alternatives that they pre-committed to even when a good offer is found early on in search (Honka, Hortaçsu and Wildenbeest 2019). In some situations, consumers may prefer using simultaneous search (Morgan and Manning 1985). As an example, if search outcomes are observed with a time lag, e.g., a time lag between the moment of requesting and receiving a quotation, and the deadline for choosing alternatives is rapidly approaching, consumers may use simultaneous search to gather information quickly (Morgan and Manning 1985). However, in our study, consumers observe their search outcomes almost instantly which is typical in online consumer search. Thus, we assume that consumers use sequential search as per Weitzmann (1979) as it is more efficient for them than simultaneous search.

2.2 Relating initial consideration set to comparison site visit incidence

In sequential search, consumers rank and search alternatives in declining order of their reservation utilities (Honka, Hortaçsu, and Wildenbeest 2019). However, consumers may have incomplete information about the available alternatives in the market as there are too many alternatives and consumers do not know all available alternatives. Consistent with this intuition, Fox and Tversky (1995) point out that consumers face conditions of "ignorance or ambiguity, where the probabilities of potential outcomes are neither specified in advance nor readily assessed on the basis of the available evidence". Thus, consumers are unable to form alternative-specific expectations for all alternatives and can only form such expectations for alternatives which they do know. Within these alternatives, some alternatives will have higher reservation utilities than others. Thus, they start their search on those alternatives with higher reservation utilities. We term these alternatives which consumers consider at the start of their search as the *initial consideration set*[2].

A critical question that arises is whether consumers immediately augment or override their initial consideration sets in the presence of comparison sites which rank alternatives for easier comparison and evaluation. For instance, consumers can utilize the ranking as a heuristic and only search the top few alternatives (e.g., Masatlioglu, Nakajima, and Ozbay 2012) and ignore their initial consideration set. While attractive, there are two reasons why this intuition is problematic from the perspective of rational consumers. First, consumers possess private information (e.g., Mas-Collel, Whinston and Green 1995) about the attributes which matter most to them. Thus, consumers may place higher weights on some attributes than those assigned by a comparison site. To better reflect their own preferences, consumers prefer to form their initial consideration set before potentially assessing it in light of the

[2] The initial consideration set is conceptually distinct from the awareness set which is the set of alternatives that consumers know of (e.g., Shapiro, MacInnis and Heckler 1997). The awareness set may contain alternatives with low reservation utilities which do not make it into the consumer's initial consideration set.

information on comparison sites. Second, consumers trust their own information above other sources of information (Granovetter 1985) where said information could arise from, for instance, word-of-mouth or the consumer's personal experiences. Hence, consumers utilize their own information to calculate and choose alternatives with higher expected utilities to form their initial consideration set. We thus expect that, even in the presence of comparison sites, consumers search a greater proportion of alternatives in their initial consideration set at the start than at the end of search.

Our stylized model assumes that consumers first form their initial consideration set and then acquire information to reduce their attribute uncertainty for these alternatives. Since comparison sites display information on selected price and non-price attributes of many alternatives (Smith 2002), consumers visit a comparison site to reduce their search costs (Bakos 2001). By visiting a comparison site they incur a single search cost to, in contrast to paying a separate search cost for each alternative's site visited (Baye and Morgan 2001). If the information shown on the comparison site is sufficient for their decision-making, consumers sequentially search each alternative on the comparison site, and stop when a suitable alternative is found. Hence, we propose that consumers visit comparison sites to efficiently evaluate alternatives in their initial consideration set, a use of comparison sites that to our knowledge has not been considered in the literature.

If the information shown on the comparison site is insufficient for their decision-making, consumers can visit alternatives' sites to obtain the requisite information. These consumers still benefit from first visiting comparison sites. Some alternatives in the initial consideration set may score poorly on the attributes displayed on the comparison site and already yield a low utility to the consumer. If more information from the alternatives' sites is insufficient to compensate for this low utility, consumers benefit by not searching these alternatives' sites. As an example, a consumer may learn from a comparison site that an alternative has the highest price. At the same time, the consumer expects this alternative to score well on corporate social responsibility initiatives. To learn about these initiatives, the consumer would need to visit the alternative's site. However, if this positive utility from corporate social responsibility cannot compensate the disutility from high prices, consumers do not visit this alternative's site.

The savings in search costs resulting from a comparison site visit scale with the size of the initial consideration set. Consumers who obtain all required information from a comparison site only visit a single site instead of visiting k sites, where k is the number of alternatives considered. Consumers who require additional information from some alternatives' sites only need to visit p*k sites, where (1-p) is the proportion of alternatives that score poorly on the attributes displayed on the comparison site such that visits to the alternatives' sites becomes redundant. Hence, we expect consumers with a larger initial consideration set size to have a higher probability of visiting a comparison site visit. As argued, consumers search alternatives in their initial consideration set at the start of search. Even within this first stage of search, they likely visit a comparison site early. By visiting early, consumers reduce their search costs as they only need to visit alternatives with high enough expected utilities that sufficiently compensates for their low utility scores based on the comparison site's information. Hence, consumers with a larger initial consideration set size are more likely to visit a comparison site and are more likely to do so at the start of search than at the end of search.

2.3 Relating expectations of finding a better deal to comparison site visit incidence

Since consumers have incomplete information (Tversky and Fox 1995), there is a set of alternatives that consumers are unable to form their alternative-specific reservation utilities on as they have no specific information on these alternatives. In fact, consumers may not even know about their existence. This is the subset of alternatives which consumers may start exploring after *not* finding a suitable alternative in their initial consideration set. Generally, consumers with high *expectations of finding a better deal* are more likely to continue searching as they believe that an additional search is likely to give a better deal than their current alternative.[3] In our stylized model, an alternative provides a better deal to consumers in terms of its product attributes such as price and non-price attributes. This logic is analogous to on-the-job search where employed workers search with higher intensity for other job offers if they expect better offers from other employers (Rogerson, Shimer, and Wright 2005).

Comparison sites provide information on lesser-known alternatives (Waldfogel and Chen 2006), hence, these sites serve as a natural information channel for consumers to learn about these alternatives (Baye and Morgan 2001, Moraga-Gonzalez and Wildenbeest 2012). Moreover, with incomplete information, consumers would need to search new alternatives in an inefficient random order. Since comparison sites rank alternatives based on their value to consumers (Natter, Ozimec, and Kim 2015), consumers can benefit by using this ranking as input for the order in which to search the new alternatives.

Overall, we expect consumers with stronger expectations of finding a better deal to have a higher likelihood of visiting comparison sites, because these sites facilitate the discovery of new alternatives (e.g., Baye and Morgan 2001, Moraga-Gonzalez and Wildenbeest 2012). Indeed, empirical research finds that consumers who visit comparison sites subsequently increase their visits to lesser-known alternatives (Waldfogel and Chen 2006). We further anticipate that consumers who expect to find a better deal are more likely to visit a comparison site at the end of their search as opposed to the start of search, because they first search their initial consideration set before proceeding to discover new alternatives.

[3] We note that consumers do not have alternative-specific expectations but they do have expectations of finding a better deal that is not alternative-specific.

Finally, we expect that consumers expecting to find a better deal are more likely to discover new alternatives not in this set.

2.4 Summary and testable expectations

In summary, our model expects consumers to search a greater proportion of alternatives in their initial consideration set at the start of their search even in the presence of comparison sites. We expect that initial consideration set size is positively associated with comparison site visit incidence, particularly at the start of search. Consumers visit a comparison site to reduce their search costs and efficiently evaluate alternatives in their initial consideration set.

At the end of their search, consumers with stronger expectations of finding a better deal expect unsearched alternatives to yield higher utilities than the realized utilities of alternatives in their initial consideration set. To learn about new alternatives and to decide on the order in which to evaluate these alternatives, consumers may visit a comparison site. We thus anticipate higher expectations of finding a better deal to be positively associated with comparison site visit incidence, particularly at the end of their search. Finally, we expect higher expectations of finding a better deal to be positively associated with new alternative discovery. We summarize our expectations in Table 1.

Table 1. Six Testable Expectations from Stylized Search Model

Expectations
1. A greater proportion of alternatives in the initial consideration set is searched at the start, than at the end of search.
2. The larger the initial consideration set, the more likely a consumer is to visit a comparison site.
3. The effect of initial consideration set size on comparison site visit incidence is more positive at the start of search than at the end of search.
4. The higher the consumer's expectations of finding a better deal, the more likely a consumer is to visit a comparison site.
5. The effect of expectations of finding a better deal on comparison site visit incidence is more positive at the end of search than at the start of search.
6. The higher the consumer's expectations of finding a better deal, the more likely a consumer is to discover a new alternative.

3. Data and empirical setting

3.1. Data

To test our expectations, we need data on consumers' pre-search characteristics (i.e., initial consideration set size and expectations of finding a better deal), and their online search activities including visit to comparison and alternatives' sites. We are fortunate to obtain such unique data from a consumer panel of 2,478 panelists by GfK Netherlands which describes their search for a health insurance alternative. Our data contain three components: (i) pre-search survey data (ii) online browsing data, and (iii) post-purchase survey data.

The pre-search survey was taken by consumers in October 2013 before consumers' commenced search, and we use the data from this survey to operationalize a consumer's initial consideration set size and expectations of finding a better deal elsewhere. Moreover, the pre-search survey queries panelists on whether they have already started their search for a health insurance alternative.[4] As our focal panelists have not started their search when filling out the pre-search survey, we eliminate the possibility that consumers' search behavior drives their answers in the pre-search survey. Moreover, consumers' search from previous years does not affect choices in the current year as alternatives update their price and non-price attributes annually in the Netherlands. Finally, the pre-search survey informs us about panelists' tendency to search for information, as well as other information that we use to operationalize our control variables.

Next, the online browsing data inform about visits to comparison and alternatives' sites and cover the period of October 2013 to January 2014. Online browsing data are collected via a browser extension installed on the panelists' computers. This extension records all web traffic, and we access web traffic data for health insurance.

[4] Our full dataset contains information on 2,649 consumers. 171 consumers indicated to have searched prior to the survey. To avoid endogeneity concerns, we exclude these 171 consumers from our analysis, leaving us with 2,478 consumers.

Finally, we use post-purchase survey data to determine the end of search which is the purchase date. This survey was conducted in January 2014. The earliest purchase was on October 1st, 2013 and the latest occurred on January 12th, 2014.

Our data are single source as each panelist provides information for all three components, i.e., pre-search, browsing history, and post-purchase. Our data are unique in that we not only observe consumers' online search behavior, but also observe their pre-search intentions. This unique feature of the data allows us to empirically test our stylized consumer search model. Below, we first elaborate on the empirical setting of our study before describing our variable operationalization and introducing our empirical model specification.

3.2 Empirical setting

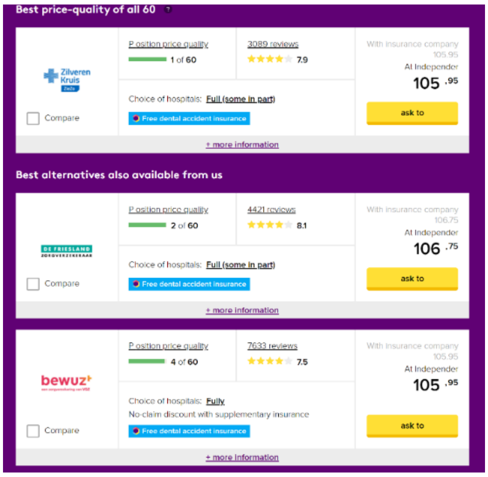
In the Netherlands, health insurance is privatized, and enrolment is compulsory. Every year, consumers can switch to a new alternative. Consumers can choose a new alternative until January 31st, if they have canceled their existing insurance policy by December 31st of the previous year. If they do not cancel before this date, they silently renew with their current health insurance alternative. Each year, the Dutch government announces policy changes related to health insurance on the third Tuesday of September. Thereafter, insurance alternatives announce their new pricing policy for the year. For 2013, alternative DSW made the earliest announcement on September 23rd after the government's policy announcements on September 21st.[5] Thus, no new pricing information was released before September 23rd, 2013, and consumers have little to no reason to start searching for next year's health insurance alternative before this date. Having a fixed start date and fixed deadline provides the unique benefit of being able to time the pre-search survey prior to consumers' search, and observe their search behavior within an exogenously determined time window. Hence, we

[5] https://www.observantonline.nl/English/Home/Articles/articleType/ArticleView/articleId/11274/DSW-sets-the-pace-with-health-premium-price-rise-of-nearly-10

focus on online searches from October 1st, 2013 until the consumer's purchase date. According to Verzekeringen (2013), 80% of Dutch consumers purchased health insurance online in 2012. The 2013 average premium for the basic plan offered by alternatives is 1280 euros annually (Statista 2019) and consumers can add extra coverage if they wish to.

Comparison sites in the Netherlands' health insurance sector include all major health insurance alternatives. Further, comparison sites display the same prices as alternatives' sites. Four basic pieces of information are requested by the comparison site when searching for health insurance which are age, gender, postal code, and whether insurance for family members is required. After entering the basic information, alternatives are shown to the consumer as per the results page in Figure 3. For each alternative, the comparison site shows its ranking, review scores, and monthly premiums. Alternatives are sorted based on the price and quality of the insurance plan as judged by the comparison site. For each alternative, consumers can click on the 'more information' tab to obtain more detailed information from either the comparison site (e.g., independer.nl) or the alternative's site (e.g., zilvenkruis.nl), such as the coverage of physicians and clinics and customer service quality.[6]

Figure 3. Typical Information Provided on a Comparison Site



Best price-quality of all 60

Zilveren Kruis
□ Compare
Position price quality: 1 of 60
3089 reviews ★★★★☆ 7.9
Choice of hospitals: Full (some in part)
Free dental accident insurance
With insurance company 105.95
At Independer 105 ,95
ask to
+ more information

Best alternatives also available from us

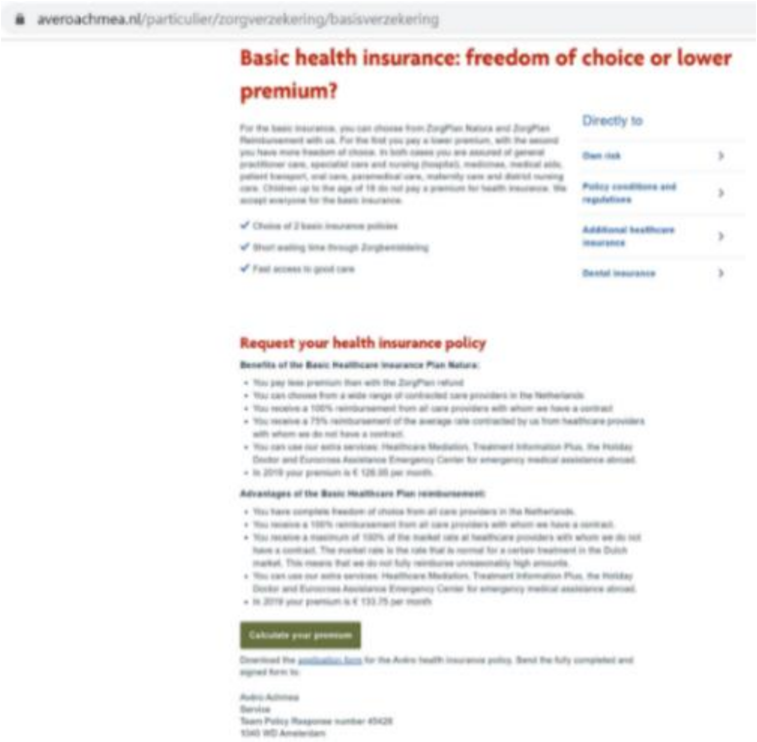
De Friesland Zorgverzekeraar
□ Compare
Position price quality: 2 of 60
4421 reviews ★★★★☆ 8.1
Choice of hospitals: Full (some in part)
Free dental accident insurance
With insurance company 106.75
At Independer 106 ,75
ask to
+ more information

bewuzt
□ Compare
Position price quality: 4 of 60
7633 reviews ★★★★☆ 7.5
Choice of hospitals: Fully
No-claim discount with supplementary insurance
Free dental accident insurance
With insurance company 105.95
At Independer 105 ,95
ask to
+ more information

[6] Out-of-pocket expenses which are a common feature of insurance plans, are standardized into different tiers in the Netherlands, and most alternatives offer all tiers. Thus, out-of-pocket expenses are not a differentiating factor for consumers.

Alternatives' sites contain information which is available from comparison sites as well as information which is not. In Figures 4a and 4b, we present a screenshot of the largest alternative in our empirical setting, which is Avero Achmea. From Figure 4a, we see that the first piece of information provided is the main types of plans available, together with the associated premiums, information that can be obtained from comparison sites as well. By scrolling down the screen, the consumer sees specific information not readily available from comparison sites such as reimbursements and the exclusions in the health insurance policy (Figure 4b). In total, we observe 52 alternatives and 12 comparison sites in our data. We observe 110,983 alternatives' site visits[7] and 1,775 comparison site visits. On average, consumers spend a total of 33.40 minutes on their site visits with the earliest and latest visit on October 1st, 2013 and January 12, 2014 respectively/

Figure 4a. Specific Information for an Insurance Alternative



averoachmea.nl/particulier/zorgverzekering/basisverzekering

Basic health insurance: freedom of choice or lower premium?

For the basic insurance, you can choose from ZorgPlan Natura and ZorgPlan Reimbursement with us. For the first you pay a lower premium, with the second you have more freedom of choice. In both cases you are assured of general practitioner care, specialist care and nursing (hospital), medicines, medical aids, patient transport, oral care, paramedical care, maternity care and district nursing care. Children up to the age of 18 do not pay a premium for health insurance. We accept everyone for the basic insurance.

✓ Choice of 2 basic insurance policies

✓ Short waiting time through Zorgbemiddeling

✓ Fast access to good care

Directly to
Own risk ›
Policy conditions and regulations ›
Additional healthcare insurance ›
Dental insurance ›

Request your health insurance policy

Benefits of the Basic Healthcare Insurance Plan Natura:

- You pay less premium than with the ZorgPlan refund
- You can choose from a wide range of contracted care providers in the Netherlands
- You receive a 100% reimbursement from all care providers with whom we have a contract
- You receive a 75% reimbursement of the average rate contracted by us from healthcare providers with whom we do not have a contract.
- You can use our extra services: Healthcare Mediation, Treatment Information Plus, the Holiday Doctor and Eurocross Assistance Emergency Center for emergency medical assistance abroad.
- In 2019 your premium is € 126.00 per month.

Advantages of the Basic Healthcare Plan reimbursement:

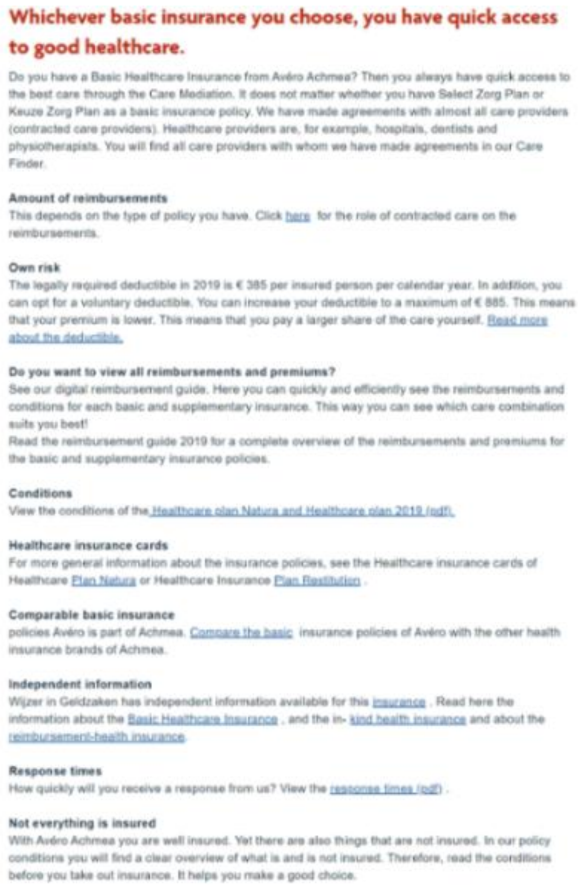
- You have complete freedom of choice from all care providers in the Netherlands.
- You receive a 100% reimbursement from all care providers with whom we have a contract.
- You receive a maximum of 100% of the market rate at healthcare providers with whom we do not have a contract. The market rate is the rate that is normal for a certain treatment in the Dutch market. This means that we do not fully reimburse unreasonably high amounts.
- You can use our extra services: Healthcare Mediation, Treatment Information Plus, the Holiday Doctor and Eurocross Assistance Emergency Center for emergency medical assistance abroad.
- In 2019 your premium is € 133.75 per month.

Download the application form for the Avéro health insurance policy. Send the fully completed and signed form to:

Avéro Achmea
Service
Team Policy Response number 45428
1040 WD Amsterdam

[7] Insurance alternatives may offer other products besides health insurance. We focus exclusively on visits to health insurance sites of these alternatives.

Figure 4b. Information Absent from Comparison Sites



Whichever basic insurance you choose, you have quick access to good healthcare.

Do you have a Basic Healthcare Insurance from Avéro Achmea? Then you always have quick access to the best care through the Care Mediation. It does not matter whether you have Select Zorg Plan or Keuze Zorg Plan as a basic insurance policy. We have made agreements with almost all care providers (contracted care providers). Healthcare providers are, for example, hospitals, dentists and physiotherapists. You will find all care providers with whom we have made agreements in our Care Finder.

Amount of reimbursements
This depends on the type of policy you have. Click here for the role of contracted care on the reimbursements.

Own risk
The legally required deductible in 2019 is € 385 per insured person per calendar year. In addition, you can opt for a voluntary deductible. You can increase your deductible to a maximum of € 885. This means that your premium is lower. This means that you pay a larger share of the care yourself. Read more about the deductible.

Do you want to view all reimbursements and premiums?
See our digital reimbursement guide. Here you can quickly and efficiently see the reimbursements and conditions for each basic and supplementary insurance. This way you can see which care combination suits you best!
Read the reimbursement guide 2019 for a complete overview of the reimbursements and premiums for the basic and supplementary insurance policies.

Conditions
View the conditions of the Healthcare plan Natura and Healthcare plan 2019 (pdf).

Healthcare insurance cards
For more general information about the insurance policies, see the Healthcare insurance cards of Healthcare Plan Natura or Healthcare Insurance Plan Restitution .

Comparable basic insurance
policies Avéro is part of Achmea. Compare the basic insurance policies of Avéro with the other health insurance brands of Achmea.

Independent information
Wijzer in Geldzaken has independent information available for this insurance . Read here the information about the Basic Healthcare Insurance , and the in- kind health insurance and about the reimbursement-health insurance.

Response times
How quickly will you receive a response from us? View the response times (pdf) .

Not everything is insured
With Avéro Achmea you are well insured. Yet there are also things that are not insured. In our policy conditions you will find a clear overview of what is and is not insured. Therefore, read the conditions before you take out insurance. It helps you make a good choice.

4 Empirical tests of expectations

We carry out separate empirical tests for (a) expectation 1, (b) expectations 2 to 5, and (c) expectation 6. For each test, we first introduce our variable operationalization and then explain the relevant statistical model and results.

4.1 Test of expectation 1

Our first expectation is that consumers search a larger proportion of alternatives in their initial consideration set at the start of, than at the end of their search. To test this expectation, we operationalize two variables: (i) the consumer's initial consideration set, and (ii) the alternatives searched at the start and end of search which are in the initial consideration set.

Variable operationalization. We operationalize the initial consideration set via the question in the pre-search survey that asks, "Which alternatives do you consider for your health insurance for 2014?" where 45 alternatives were listed, and an 'Others' option was provided to indicate an alternative that is not in the list of alternatives.

Next, we identify the proportion of alternatives searched by the consumer which are in the initial consideration set at the start of search ($\mathrm{PropCons}_{\mathrm{i,start}}$) and at the end of search ($\mathrm{PropCons}_{\mathrm{i,end}}$) using the consumer's online browsing data.[8] For example, if a consumer has an initial consideration set size of three, searches two alternatives in this set at the start of search and one alternative at the end of search, $\mathrm{PropCons}_{\mathrm{i,start}}$ is $\frac{2}{3}$ and $\mathrm{PropCons}_{\mathrm{i,end}}$ is $\frac{1}{3}$ for this consumer. Next, we define the difference in the proportion of alternatives searched in the initial consideration set at the start and end of search as $\mathrm{PropCons}_{\mathrm{i,diff}} = \mathrm{PropCons}_{\mathrm{i,start}} - \mathrm{PropCons}_{\mathrm{i,end}}$. We define the start and end of search as the first and last 50% of the consumer's site visits, with visits exactly on the 50% mark randomly assigned to either the start or the end of search. Analogous to Bronnenberg, Kim, and Mela (2016), we define a site visit as a visit to an alternative's site or the results page of a comparison site that differs from the preceding site visited. Hence, browsing within a site does not generate additional visits. To illustrate, a consumer that reaches an alternative's site generates a site visit. Once the consumer reaches the results page of a comparison site, s/he generates another site visit. Yet another site visit is generated if the consumer navigates to the previously visited alternative's site, or the results page of another comparison site.

Model specification. To test our first expectation that the proportion of alternatives searched in the initial consideration set is larger at the start than at the end of search, we conduct a paired-samples t-test for the proportion of alternatives searched in the initial consideration set at the start of search, against the end of search, i.e. we test if $\mathrm{PropCons}_{\mathrm{i,diff}}$ is significantly different from zero. This test is appropriate because we measure the same

[8] In our dataset, we do not observe which specific alternatives the consumer focused on comparison sites. Hence, we focus on visits to alternatives' sites.

variable (i.e., proportion of alternatives searched in the initial consideration set) at different points in time, for each consumer. By comparing the difference in the proportion at two moments in time within the same consumer, we control for time-invariant variables such as consumer characteristics.

Results. Our results shown in Table 2 support our first expectation. We find that the difference in proportions of alternatives searched in the initial consideration set between the start and end of search is positive and significant (i.e., $\mathrm{PropCons}_{\mathrm{i,diff}}$ = .14, $p < .001$). Thus, we find empirical support for our first expectation that consumers evaluate a larger proportion of their initial consideration set at the start than at the end of search.

We set out to show that even when comparison sites are available, consumers still prioritize their search on the initial consideration set. However, the results reported above are based on the full sample of consumers, including those that do not visit a comparison site, either because they lack awareness or avoid these sites. Hence, our result that consumers prioritize searching alternatives in their initial consideration set may only hold for those who do not visit a comparison site. To rule out this interpretation, we implement the same t-test on the subset of consumers who have visited at least one comparison site. If our first expectation is correct, we should find a positive and significant difference in proportions (i.e., $\mathrm{PropCons}_{\mathrm{diff}}$) for this subset of consumers as well. Conditioning on at least one comparison site visit, the difference in proportions of alternatives searched is also positive and significant (i.e., $\mathrm{PropCons}_{\mathrm{diff}}$ = .22, $p < .001$). Hence, consumers who have visited a comparison site also search a higher proportion of alternatives in their initial consideration set at the start of search.

Table 2. Results for Differences in Proportions Searched between Start and End of Search

Variable	Unconditional on comparison site visits	Conditional on at least one comparison site visit
$\mathrm{PropCons}_{i,start}$	.21 (.01)	.28 (.01)
$\mathrm{PropCons}_{i,end}$	.07 (.00)	.06 (.00)
$\mathrm{PropCons}_{i,diff}$	.14** (.01)	.22** (.02)

Notes: (i) Numbers (resp. numbers in parentheses) are the means (resp. standard errors). (ii) ** denotes significance based on the 95% confidence interval. (iii) t-test which is unconditional on comparison site visits uses 2,478 observations, and t-test which is conditional on at least one comparison site visit uses 685 observations.

4.2 Test of expectations 2 to 5

In this section, we test our expectations that the consumer's (i) initial consideration set size is positively associated with comparison site probability, where the effect of initial consideration set size on comparison site visit incidence is more positive at the start of search than at the end of search, and (ii) expectations of finding a better deal is positively related with the likelihood of visiting a comparison site, with a more positive effect on comparison site visit incidence at the start than at the end of search. To test these expectations, we need to operationalize three dependent variables (i) comparison site visit incidence, (ii) comparison site visit incidence at the start of search, and (iii) comparison site visit incidence at the end of search. Additionally, we operationalize our two independent variables (i) initial consideration set size, and (ii) expectations of finding a better deal. Finally, we operationalize five control variables (i) price dissatisfaction, (ii) nonprice dissatisfaction, (iii) political discussion, (iv) changes in personal health and personal situation, and (v) search tendency.

Operationalization dependent variables. Comparison site visit incidence (CS_i) takes the value 1 if at least one comparison site was visited by consumer i between Oct 1st 2013, and the end date of search; 0 otherwise. To test our 3rd and 5th expectation, we additionally define comparison site visit incidence at the start the start of search ($\mathrm{CS}_{\mathrm{i,start}}$) as a binary variable which takes the value 1 if consumer i visits a comparison site during the first 50% of visits to comparison or alternatives' sites and 0 otherwise, where site visits are operationalized identically as per expectation 1's empirical test. We define comparison site visit incidence at the end of search ($\mathrm{CS}_{\mathrm{i,end}}$) analogously as a binary indicator with the value of 1 if consumer i visits a comparison site in the last 50% of visits to comparison or alternatives' sites and 0 otherwise.

Operationalization independent variables. We operationalize a consumer's initial consideration set size using the same source question used in the test of expectation 1, and take the natural logarithm of the number of selected alternatives to allow for diminishing returns ($\mathrm{ConsSet}_{\mathrm{i}}$). In operationalizing consumer i's expectations of finding a better deal ($\mathrm{BetterDeal}_{\mathrm{i}}$), we draw on the source question which asks "To what extent do the following aspects play a role in your search for a new health insurance alternative?" We form a multi-item scale based on all items capturing the various aspects by which other alternatives provide a better offer relative to the focal alternative. These items are (i) "I would be able to get better coverage with another health insurance", (ii) "I would be able to pay a lower insurance premium with another health insurance", (iii) "Another alternative offers an interesting promotion (to switch)". We confirm the scale's internal consistency by the Cronbach's alpha value of .80 (Nunally and Bernstein 1994).

Operationalization control variables. We control for five additional reasons underlying consumers' comparison site visits. These are (i) non-price dissatisfaction, (ii) price dissatisfaction, (iii) importance of political discussions, (iv) changes in personal and health situation, and (v) search tendency. We derive constructs (i) to (iv) from the same pre-search survey question used to operationalize our better-deal-elsewhere variable, "To what extent do the following aspects play a role in your search for a new health insurance alternative?" Each item is again scored on a scale of 1 to 4, where 1 indicates no role and 4 indicates a very important role. Construct (v) is derived from a different pre-search survey question that asks consumers to indicate their agreement with the statement "I always shop around if I need a financial product."

Dissatisfaction refers to the unhappiness that results when an existing product or service has not met the expected levels of the consumer (Swan and Combs 1976). When consumers are dissatisfied about the attribute(s) of their current alternative's offering, they gather information on the performance of different alternatives for said attribute(s) (Ratchford et al. 2003). Thus, dissatisfied consumers are more likely to visit comparison sites when their dissatisfaction arises from attributes that are easily summarized on comparison sites. We distinguish two sources of dissatisfaction which are price dissatisfaction and non-price dissatisfaction. Price dissatisfaction ($\mathrm{DissatPrice}_i$) uses all items related to dissatisfaction about the price and the affordability of the current alternative. These items are: (i) "The insurance premium of my current alternative is (too) high", (ii) "The insurance premium of my current health insurance rises too much in 2014", and (iii) "I have troubles to keep paying my current health insurance." Similarly, non-price dissatisfaction ($\mathrm{DissatNonPrice}_i$) uses all items referring to dissatisfaction with non-price related characteristics of the current alternative. These items are: (i) "I am dissatisfied with the coverage of my current health insurance", (ii) "I am dissatisfied about the service offered by my current health insurance alternative", (iii) "I am dissatisfied about my current health insurance alternative no longer contracting (reimbursing) certain hospitals". The Cronbach's alpha for price and non-price dissatisfaction is .77 and .80, respectively, thus establishing convergent validity.

In our empirical setting (i.e., Netherlands), health insurance is fully privatized, but healthcare is subsidized by the government. Thus, the issue of healthcare costs frequently comes up as part of political discussions. Consumers who find these discussions important are likely to have higher involvement. More involved consumers may expend more time and effort to compare information on alternatives using comparison sites. Hence, we control for the consumer's involvement by considering the importance of political discussion to panelists when they search alternatives ($\text{Political}_\text{i}$) which is reflected in the single item "Political discussions" in the same source question used to operationalize our dissatisfaction variables.

Our penultimate control variable accounts for the possibility that panelists' search may be driven by changes in their personal and health situations (Handel and Kolstad 2015). Thus, we construct a multi-item scale using the two items that relate to changes in the panelist's personal or health situations (Changes_i). These items are: (i) "Changes in my personal situation", (ii) "Changes affecting my personal health". The Cronbach's Alpha is .74 and provides support for convergent validity of the scale.

Finally, consumers with a higher search tendency ($\text{SearchTendency}_\text{i}$) are more likely to compare and shop around for alternatives; a task which is facilitated on comparison sites. We proxy for this amount of effort expended by the consumer using $\text{SearchTendency}_\text{i}$. We measure $\text{SearchTendency}_\text{i}$ using the source question "I always shop around if I need a financial product." where the consumer's indicated level of agreement varies from 1 to 5, where '1' is 'very much disagree' and '5' is 'very much agree'.

Discriminant validity. We tested for the discriminant validity of all constructs which are operationalized into independent and control variables using the same source question, using the mono-trait multi-trait (MTMT) method by Henseler, Ringle, and Sarstedt (2015) which computes the ratio of average correlations across and within each pair of constructs. The numerator (denominator) of this ratio is the average correlation between items across (within) constructs. To employ this method, at most one construct can be a single-item construct; thus, the MTMT ratio for two single-item constructs is undefined.

If there is discriminant validity, we expect a lower correlation between items across constructs than between items within constructs. Following Henseler, Ringle, and Sarstedt (2015), discriminant validity is established if the ratio is below .85. All ratios satisfy this condition, and the constructs thus exhibit discriminant validity (Table A1 of Appendix 1). This conclusion is corroborated by the results from a factor analysis with Varimax rotation. Extracting five factors, we obtain high loadings within, and low loadings across, the independent and control variables (Table A2 of Appendix 2). Hence, our results indicate that the independent variables and control variables are sufficiently distinct.

Data description. We provide two tables for easy reference. Table A3 in Appendix 3 briefly describes all variables and the scale used for each variable. Next, in Table 3 we provide descriptive statistics, variance inflation factors (VIFs), and a full correlation matrix for all variables used in this study. From Table 3, we learn that about 35 in 100 consumers visit comparison sites. Better deal elsewhere reflects consumers' expectations of obtaining a better offer than their current alternative and these expectations are relatively high with an average score of 2.54. Although the averages of price and non-price dissatisfaction are similar at 2.23 and 2.03, their correlations are relatively low at .63. Finally, we note that the VIFs for all independent and control variables are sufficiently low.

Table 3. Descriptive Statistics and Correlation Matrix for all Variables

	Descriptive Statistics						Correlation Matrix							
Variables	Mean	Min	Median	Max	Standard Deviation	VIF	1.	2.	3.	4.	5.	6.	7.	8.
1. Comparison site incidence	.35	.00	.00	1.00	.48		1							
2. Consideration set (ln)	.48	.00	.00	3.81	.70	1.12	.13	1						
3. Better deal elsewhere	2.54	1.00	2.67	4.00	.87	2.03	.14	.24	1					
4. Price dissatisfaction	2.23	1.00	2.33	4.00	.87	2.25	.01	.04	.61	1				
5. Non-price dissatisfaction	2.03	1.00	2.00	4.00	.92	1.90	-.02	-.03	.45	.63	1			
6. Political discussion	1.45	1.00	1.00	4.00	.76	1.21	-.01	-.05	.24	.34	.33	1		
7. Changes in personal situation and health	2.04	1.00	2.00	4.00	.97	1.56	.02	.01	.37	.49	.52	.36	1	
8. Tendency to search for information	3.68	1.00	4.00	5.00	.91	1.07	.10	.18	.22	.11	.05	.05	.08	1

Notes: Reported numbers for variables 1-8 are based on 2,478 observations

Model specification for testing expectations 2 and 4. We test expectations 2 (i.e., initial consideration set size is positively associated with comparison site visit incidence) and 4 (i.e., expectations of finding a better deal is positively associated with comparison site visit incidence) by estimating a probit model which explains comparison site visit incidence, $\mathrm{CS_i}$, from the independent and control variables introduced earlier. We define consumer i's utility from a comparison site visit as:

$$\mathrm{U_i^*} = \mathrm{z_i} + \varepsilon_\mathrm{i} \text{ , where } \varepsilon_\mathrm{i} \text{ is a normally distributed error term.} \tag{1}$$

Since utility is a latent variable, we do not observe it directly. Thus, we only observe:

$$\mathrm{CS_i} = \begin{cases} 1, & \text{if } \mathrm{U_i^*} > 0 \\ 0, & \text{if } \mathrm{U_i^*} \leq 0 \end{cases} \tag{2}$$

When $\mathrm{U_i^*} > 0$, we have $\mathrm{z_i} + \varepsilon_\mathrm{i} > 0$ and thus $\varepsilon_\mathrm{i} > -\mathrm{z_i}$. Hence,

$$\mathrm{Prob}\ (\mathrm{U_i^*} > 0) = \mathrm{Prob}\ (\mathrm{CS_i} = 1, \mathrm{z}) = \mathrm{Prob}\ (\varepsilon_\mathrm{i} > -\mathrm{z_i})$$

$$= \mathrm{Prob}\ (\varepsilon_\mathrm{i} < \mathrm{z_i}) = \Phi(\mathrm{z_i}) \tag{3}$$

where Φ is the cumulative distribution function of the standard normal distribution. In the above equations, $\mathrm{z_i}$ is defined as

$$\mathrm{z_i} = \beta_0 + \beta_1\mathrm{ConsSet_i} + \beta_2\mathrm{BetterDeal_i} + \beta_3\mathrm{DissatPrice_i}$$

$$+\beta_4\mathrm{DissatNonPrice_i} + \beta_5\mathrm{Political_i} + \beta_6\mathrm{Changes_i} + \beta_7\mathrm{SearchTendency_i} \tag{4}$$

If expectations 2 and 4 are supported by the data, β_1 and β_2 should be positive and significant.

Model specifications for testing expectations 3 and 5. We model comparison site visit incidence at the start and end of search, $\mathrm{CS_{i,start}}$ and $\mathrm{CS_{i,end}}$, using the same independent and control variables in equation 4. Additionally, to explain $\mathrm{CS_{i,end}}$, we include $\mathrm{CS_{i,start}}$ as another control variable to allow the possibility that the required information is obtained from a comparison site visit at the start of search, which lowers comparison site visit probability at the end of search. We test expectations 3 and 5 by comparing the difference in the effects of our focal independent variables at the start and end of search for each. Because our probit model is nonlinear, we cannot directly compare the coefficients at the start and end and instead compare the uplifts in predicted comparison site visit incidence. Specifically, we first calculate the baseline comparison site visit probability at the start of search when all independent variables are at the mean:

$$\widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1) = \Phi(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{start}}) \tag{5}$$

where Φ is the cumulative normal distribution, $\overline{\mathrm{X}}$ is the vector of independent variables and control variables set at their mean values, and a constant term and where $\widehat{\beta}_{\mathrm{start}}$ is the vector of coefficient estimates. Thereafter, we add a standard deviation to each independent variable and obtain the updated predicted comparison site visit probabilities. For example, the expected comparison site visit probability at the start of search at an increased level of initial consideration set size is given by

$$\widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1) = \Phi(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{start}} + \beta_1\hat{\sigma}_{\mathrm{ConsSet}}) \tag{6}$$

Next, we denote the uplift in the expected comparison site visit probability at the start of search as

$$\Delta_{\mathrm{start,ConsSet}} = \widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1) - \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1) \tag{7}$$

Similarly, we define $\Delta_{\mathrm{start,BetterDeal}}$ as the uplift in the expected comparison site visit probability due to a one standard deviation increase in expectations of finding a better deal. Analogously, we determine $\Delta_{\mathrm{end,ConsSet}}$ and $\Delta_{\mathrm{end,BetterDeal}}$, the uplifts at the end of search. Finally, we calculate the difference in uplifts at the start and end of search:

$$\Delta_{\mathrm{ConsSet}} = \Delta_{\mathrm{start,ConsSet}} - \Delta_{\mathrm{end,ConsSet}} \tag{8}$$

$$\Delta_{\mathrm{BetterDeal}} = \Delta_{\mathrm{start,BetterDeal}} - \Delta_{\mathrm{end,BetterDeal}} \tag{9}$$

Expectations 3 and 5 predict a positive sign for Δ_{ConsSet}, i.e. a larger effect of initial consideration set size at the start than at the end of search, and a negative sign for $\Delta_{\text{BetterDeal}}$, i.e. a larger effect of expectations of a better deal elsewhere at the end than at the start of search. We provide all equations to obtain Δ_{ConsSet} and $\Delta_{\text{BetterDeal}}$ in Appendix 4.

Estimation. We estimate our three models (i.e., the overall, start-of-search and end-of-search models) using maximum likelihood. We bootstrap the standard errors of $\Delta_{\text{ConsSet}}$and $\Delta_{\text{BetterDeal}}$ (1,000 replications) which lets us test expectations 3 and 5.

Results from tests of expectations. In Table 4, we present our estimation results (coefficient estimates and 95% confidence intervals) for our three models.

Table 4. Estimation Results for Basic, Start-of-Search and End-of-Search Models

Variable	CS	CS_{start}	CS_{end}
Constant	-1.34** (-1.61, -1.07)	-1.75** (-2.05, -1.46)	-1.75** (-2.08, -1.41)
CS_{start}			1.13** (1.00, 1.27)
Consideration set	.12** (.04, .20)	.17** (.08, .25)	.03 (-.06, .12)
Better deal elsewhere	.26** (.17, .34)	.19** (.10, .28)	.32** (.22, .42)
Price dissatisfaction	-.09* (-.18, .01)	-.02 (-.12, .07)	-.17** (-.28, -.05)
Non-price dissatisfaction	-.08* (-.16, .01)	-.09** (-.18, -.01)	-.03 (-.13, .06)
Political discussion	-.05 (-.13, .02)	-.04 (-.12, .04)	-.06 (-.16, .03)
Changes in personal situation and health	.02 (-.05, .09)	.02 (-.05, .09)	.05 (-.03, .13)
Tendency to search for information	.11** (.05, .17)	.16** (.09, .23)	.01** (-.07, .08)
Difference in the uplifts due to an increase in initial consideration set size (Δ_{ConsSet})		.02** (.00, .03)	
Difference in uplifts due to an increase of better deal elsewhere ($\Delta_{\text{BetterDeal}}$)		-.03** (-.06, -.01)	

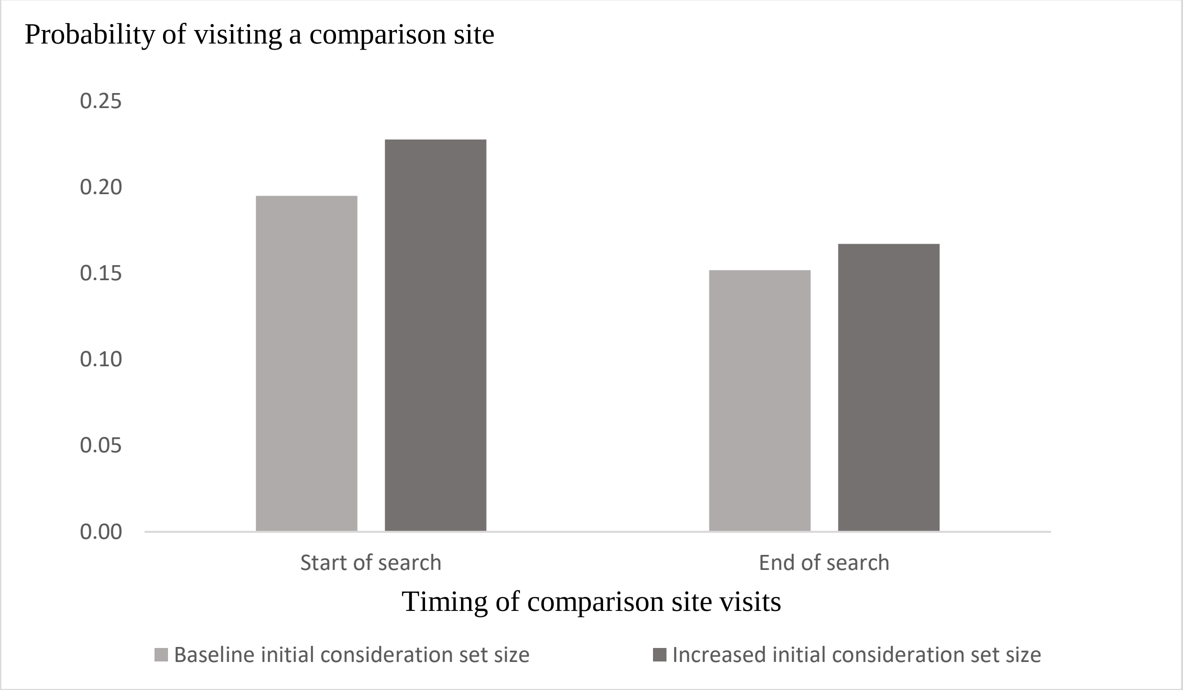
Notes: (i) Numbers are the coefficient estimates (lower and upper limits for the 95% confidence interval). (ii) * and ** denotes significance based on the 90% and the 95% confidence intervals, respectively. (iii) Each model uses 2,478 observations.

Expectation 2 states that consumers with a larger initial consideration set size have a higher probability of visiting comparison sites. We find strong support for this expectation. Initial consideration set size has a positive and significant effect on comparison site visit incidence (β_1 = .12; $p < .01$).

Our stylized search model further intuits that consumers with a larger initial consideration set size have a higher probability of visiting comparison sites at the start of search than at the end of search (expectation 3). We find that initial consideration set size has a significant and positive effect on comparison site visits at the start of search ($\beta_{1,\text{start}}$ = .17; $p < .001$) but an insignificant effect at the end of search ($\beta_{2,\text{end}}$ = .03; $p = .48$). Consistent with expectation 3, the difference in the uplifts in comparison site visit probabilities due to an increase in initial consideration set size is positive and significant, i.e., the mean value of Δ_{ConsSet} .02 and its 95% confidence interval is [.00, .03].

We illustrate this result in Figure 5 by showing how consumers with a baseline initial consideration set size (at the mean) and an increased initial consideration set size (one standard deviation increase to the mean), visit comparison sites with different probabilities at the start and at the end of search; all other variables kept at the mean. The left-hand (right-hand) bars illustrate the probability of a comparison site visit at the start (end) of search, where the light (dark) bars indicate said probability for a consumer with a baseline (increased) initial consideration set size. The difference between dark and light bars thus gives us the uplift in comparison site visit probability due to an increased initial consideration set size set at the start of search ($\Delta_{\text{start,ConsSet}}$) and end of search ($\Delta_{\text{end,ConsSet}}$). Based on our empirical results, we find that $\Delta_{\text{start,ConsSet}}$ (.03) is larger than $\Delta_{\text{end,ConsSet}}$ (.02).

Figure 5. Relative Comparison Site Visit Probabilities for Baseline versus Increased Initial Consideration Set Size



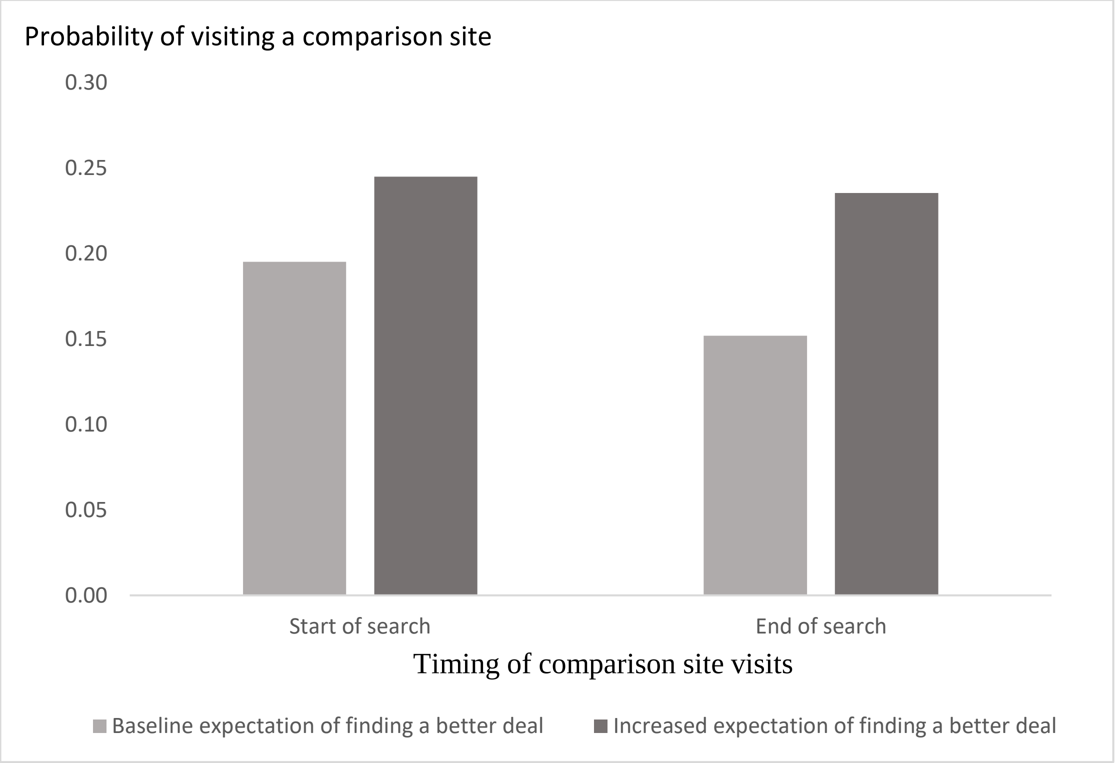
Note: baseline initial consideration set size is at the mean, and increased consideration set size is mean plus one standard deviation.

With regards to expectation 4, that the expectations of obtaining a better deal are positively associated with comparison visit likelihood, we find that higher levels of better deal elsewhere are associated with a higher probability of visiting comparison sites (β_2 = .26; $p < .001$), thereby confirming our fourth expectation. Our finding is in line with existing literature which suggests that comparison sites promote new alternative discovery (e.g., Baye and Morgan 2001, Moraga-Gonzalez and Wildenbeest 2012).

Lastly, we believe that the effect size of expectations of obtaining a better deal elsewhere on comparison site visit incidence is larger at the end of search than at the start of search (expectation 5). Expectations of obtaining a better deal has a significant and positive effect on comparison site visits at the start of search ($\beta_{2,\text{start}}$ = .19; $p < .001$), and at the end of search ($\beta_{2,\text{end}}$ = .32; $p < .001$). Next, we calculate the difference in uplifts in comparison site visit probabilities due to an increase in expectations of finding a better deal elsewhere. We find that the marginal effect at the mean is significantly larger at the end than at the start of search (mean value of $\Delta_{\text{BetterDeal}}$ is -.03; 95% confidence interval is [-.06, -.01]). Thus, our fifth expectation is supported.

In Figure 6, we illustrate how consumers with a baseline level of expectation of obtaining a better deal versus an increased level of expectation of obtaining a better deal, visit comparison sites with different probabilities at the start and end of search, with all other variables kept at mean values. Analogous to Figure 3, the light (dark) bars in Figure 4 depict comparison site visit probability due to a baseline level (an increased level) of expectations of obtaining a better deal, where the left-hand (right-hand) bars denote site visit probabilities at the start (end) of search. The difference between the pairs of light and dark bars shows the uplift in comparison site visit probability at the start and end of search. Specifically, $\Delta_{\text{start,BetterDeal}}$ has a value of .05 and $\Delta_{\text{end,BetterDeal}}$ has a larger value of .08 thus showing the larger uplift in comparison site visit probability at the end of search.

Figure 6. Relative Comparison Site Visit Probabilities for Baseline versus Increased Expectations of Getting a Better Deal



Amongst our control variables, price dissatisfaction is significantly and negatively associated with comparison site visit likelihood in the basic and end-of-search models. One possible interpretation is that these consumers tend to focus on their price dissatisfaction only and thus shun comparison sites which compare alternatives on a wider range of attributes; information that these consumers wish to avoid. As expected, search tendency is positively associated with comparison site visit incidence in all three models; consumers who shop around for alternatives likely find their shopping facilitated by the information provided on comparison sites. Finally, nonprice dissatisfaction is associated with lower comparison site visit incidence in all models. Consumers desiring information on nonprice attributes may prefer the richer information available from alternatives' sites.

Robustness checks. We conduct two robustness checks to rule out competing explanations for our results. First, consumers may visit comparison sites because of targeted emails. If the targeting mechanism is correlated with our independent variables, we may bias our coefficient estimates of the independent and control variables in the comparison site visit incidence model. Since we have access to the URLs prior to any site visit, we observe which consumers reached comparison sites through email clicks. Only two respondents were found to have clicked on an email link. Hence, the influence of emails on comparison site visit incidence is negligible.

Our second check rules out the possibility that consumers' starting date of search is related to their pre-search motivations, thus confounding the interpretation of our findings. We provide the correlations between the starting date of search and independent variables in Appendix 5, Table A5. The highest (absolute) correlation is a mere -.07 for search tendency; hence, our model estimation results are not confounded by the start date of search.

4.2 Test of expectation 6

In this section, we test our final expectation that the higher the consumer's expectations of finding a better deal, the more likely a consumer is to discover a new alternative. Our dependent variable for testing this expectation is new alternative discovery, while we retain the same independent and control variables used in the comparison site visit incidence model (refer to equation 4).

Variable operationalization. We operationalize our dependent variable, new alternative discovery by consumer i ($\mathrm{Discover_i}$) by using the initial consideration set as introduced in our test of expectation 1 and the online browsing data. Specifically, we operationalize $\mathrm{Discover_i}$ as a binary variable taking the value of 1 if consumer i visited the site of an alternative not in the initial consideration set; else this variable takes the value of 0. Our two independent variables (better deal elsewhere and initial consideration set size) and five control variables (non-price dissatisfaction, price dissatisfaction, importance of political discussions, changes in personal and health situations, and search tendency) are operationalized identically as per the variable operationalization in our empirical test for expectations 2 to 5.

Model specification. We test expectation 6 via a probit model with the dependent and independent variables listed in the variable operationalization section. We define consumer i's utility, $\mathrm{W_i^*}$ from new alternative discovery as:

$$\mathrm{W_i^*} = \mathrm{w_i} + \eta_\mathrm{i}, \text{ where } \eta_\mathrm{i} \text{ is a normally distributed error term.} \qquad (10)$$

Next, we specify

$$\mathrm{Discover_i} = \begin{cases} 1, & \text{if } \mathrm{W_i^*} > 0 \\ 0, & \text{if } \mathrm{W_i^*} \leq 0 \end{cases} \qquad (11)$$

and define

$$\mathrm{w_i} = \nu_0 + \nu_1\mathrm{ConsSet_i} + \nu_2\mathrm{BetterDeal_i} + \nu_3\mathrm{DissatPrice_i} + \nu_4\mathrm{DissatNonPrice_i}$$

$$+\nu_5\mathrm{Political_i} + \nu_6\mathrm{Changes_i} + \nu_7\mathrm{SearchTendency_i} \qquad (12)$$

Results. Our results in Table 5 show that the coefficient of better deal elsewhere is positive and significant ($\nu_2 = .12$, $p < .01$). Consistent with our 6th expectation, consumers with higher expectations of finding a better deal than their current alternative have a higher probability of discovering a new alternative. Initial consideration set size is significantly and negatively associated with new alternative discovery ($\nu_1 = -.22$, $p < .001$). This negative association rules out an alternative interpretation of our initial consideration set size measure. One could argue that, instead of measuring the size of the set of alternatives with the highest expected utility, it could measure the consumer's openness to receiving information (Shocker et al. 1991, Swait 1984). However, greater openness to receiving information would increase the probability of new alternative discovery which should lead to a positive association between initial consideration set size and new alternative discovery probability. Thus, we can rule out this alternative interpretation of the initial consideration set size. We provide an interpretation of the significant negative effect of initial consideration set size on new alternative discovery in the discussion section. The possibility remains that consumers with a (very) large initial consideration set have a low probability of discovering new alternatives, simply because there are few alternatives not included in their initial consideration set. To eliminate this interpretation, we rerun our probit model on the subset of consumers with an initial consideration set size of less than 4, and find substantively similar results that a higher level of better deal elsewhere is associated with a higher probability of new alternatives, even for consumers with restricted set sizes. Again, we find that a negative and significant coefficient on initial consideration size thus confirming our reasoning in the previous paragraph.

Table 5. Estimation results for relationship between expectations of finding a better deal and probability of new alternative discovery

Variables	New alternative discovery	New alternative discovery, conditional on initial consideration set size < 4
Constant	.29* (.02, .55)	.34* (.06, .62)
Consideration set	-.22*** (-.30, -.13)	-.42*** (-.57, -.28)
Better deal elsewhere	.12** (.04, .21)	.15** (.06, .25)
Price dissatisfaction	.04 (-.06, .13)	-.00 (-.11, .11)
Non-price dissatisfaction	-.03 (-.11, .06)	-.03 (-.12, .06)
Political discussion	.04 (-.04, .12)	-.01 (-.08, .10)
Changes in personal situation and health	-.05 (-.12, .02)	-.01 (-.09, .07)
Search tendency	.08** (.02, .15)	.08* (.01, .15)

Notes: (i) *, ** and *** denotes $p < .05$, $p < .01$ and $p < .001$ respectively,
(ii) numbers in parentheses denote the lower and upper limits for 95% confidence interval,
(iii) estimated on sample of 2,478 individuals for new discovery, and on sample of 2,144 individuals for new discovery conditional on having initial consideration set size < 4

5 Discussion

In this section, we first summarize our key findings. Next, we highlight the implications of our results for scholars and practitioners before concluding with limitations and future research directions.

5.1 Summary of key findings

Comparison sites are ubiquitous in consumer search. Existing research has mostly focused on the role of comparison sites for new alternative discovery (e.g., Baye and Morgan 2001, Moraga-Gonzalez and Wildenbeest 2001, Waldfogel and Chen 2006). In light of this alternative discovery function of comparison sites, it is unclear whether initial consideration set sizes are still relevant. In this study, we theoretically and empirically address the role of the consumer’s initial consideration set size in consumer search when comparison sites are available to them. We find that consumers search a greater proportion of their initial consideration set at the start than the end of search, also in the presence of comparison sites. Also, in line with a search cost argument, we find that larger consideration set sizes are associated with a higher probability of a comparison site visit, where the effect is more positive at the start of search than at the end of search. Additionally, consumers with higher expectations of finding a better deal are more likely to visit a comparison site, where the relative effect is stronger at the end of search. Finally, expectations of finding a better deal are positively associated with the probability of new alternative discovery, while the initial consideration set size shows a negative association with new alternative discovery.

5.2 Theoretical implications

Our work adds to the comparison site literature and search literature in general. Below, we discuss the two main contributions and their implications for scholars.

First, we complement existing literature that focuses on the alternative discovery role of comparison sites (e.g., Baye and Morgan 2001, Moraga-Gonzalez and Wildenbeest 2012, Waldfogel and Chen 2006) by considering the relationship between initial consideration set size and comparison site visit incidence. A priori, one might expect that a larger initial consideration set size lowers the need for a comparison site visit as the consumer is already considering many alternatives and is less likely to be in need of new alternative discovery. In contrast, we show that a larger initial consideration set size is positively associated with comparison site visit incidence, which we attribute to consumers’ desire to lower their search costs. We thus uncover a use of comparison site visits, evaluation of alternatives in the initial consideration set, not previously considered in the literature. Scholars studying the effects of comparison sites should consider the dual role of comparison sites and allow for heterogeneity in consumers' use of these sites.

Second, we contribute to the comparison site and search literature by showing that the initial consideration set retains relevance when comparison sites are available. Existing literature documents the positive effect of advertising on the probability of entering the consumer's consideration set (Goree 2008, Terui, Ban and Allenby 2011). One might think that the initial consideration set, and thus advertising, loses its importance when alternatives can be easily compared and ranked on comparison sites. We rely on incomplete information arguments to theorize that consumers still start their search on their initial consideration set when comparison sites are available and we find strong empirical support for this. We further find that, while the initial consideration set size is positively associated with comparison site visit incidence, it is negatively related to new alternative discovery. While we did not explicitly hypothesize this result, it is fully in line with our stylized search model where consumers first search their initial consideration set and stop searching once the searched alternatives' realized utilities exceeds the unsearched alternatives' maximum reservation utilities. All things equal, when the initial consideration set is larger, a consumer is more likely to find a suitable alternative in this set and stop searching without considering a new alternative. Scholars should ideally observe consumers' initial consideration sets, as in our study, instead of relying on proxies such whether the retailer is "branded" (cf. Waldfogel and Chen 2006).

5.3 Managerial implications

Our results suggest implications for alternatives and policymakers. Alternatives must ensure that they exist in the consumer's initial consideration set, even when comparison sites are available to consumers, because consumers are more likely to evaluate these alternatives at the start of search. Since advertising increases the odds of entering the consumer's consideration set (Goree 2008, Terui, Ban, and Allenby 2011), firms must not neglect their advertising even when listing on comparison sites. Comparison sites are important for smaller alternatives without the funds to invest in advertising as consumers use these sites for new alternative discovery. Hence, smaller alternatives which may not enter the consumer's initial consideration set must invest in their product's attributes to ensure they are ranked highly on comparison sites. In general, alternatives must calculate the expected benefits and costs of investing in advertising to enter a consumer's initial consideration set and improving its product attributes to rank higher on the comparison site.

Policymakers may believe that comparison sites increase consumers' welfare, as these sites reduce the brand equity of larger firms via greater information provision on lesser-known alternatives (Waldfogel and Chen 2006). Yet, our findings suggest that such a result does not apply to all consumers. Consumers may also visit comparison sites to reduce the costs of searching their initial consideration set, where alternatives with larger brand equity may be more likely to be included in this set. In contrast, consumers who did not find a suitable alternative in their initial consideration set and who visit a comparison site in the later stages of their search, are more likely to explore alternatives not in their initial consideration set. This segment of consumers is more likely to experience increases in their welfare. Hence, policymakers wishing to increase the positive consumer welfare effect from comparison sites should not only target those consumers that do not visit comparison sites, but additionally also target those that visit a comparison site only to evaluate alternatives in their initial consideration set.

5.4 Limitations and future research

We identify three main limitations of our study which suggest future research directions. First, we only focus on the health insurance sector in the Netherlands. Thus, future studies could extend our framework to a multiple-industry and multiple-country setting to uncover possible country-specific and industry-specific effects. Second, we are unable to observe which alternatives and attributes consumers view on the comparison site. Future research could try to obtain this information, e.g., through eye-tracking, and provide a deeper understanding of consumer information acquisition on the comparison site. Third, we rely on observational data for this study. Future research could experimentally manipulate the initial consideration set size and expectations of finding a better deal in a lab (or possibly field) setting to determine their causal impact on comparison site visit incidence.

References

Bakos YJ (2001) The emerging landscape for retail e-commerce. *J. Econom. Perspectives* 15(1):69-80.

Baye MR, Morgan J (2001) Information gatekeepers on the Internet and the competitiveness of homogeneous product markets. *Amer. Econom. Rev*. 91(3):454-474.

Bronnenberg BJ, Kim JB, Mela CF (2016) Zooming in on choice: how do consumers search for cameras online *Marketing Sci.* 35(5):693-712.

Fox RC, Tversky A (1995) Ambiguity aversion and comparative ignorance. *Quart. J. Econom*. 110(3):585-603.

Goeree, MS (2008) Limited information and advertising in the US personal computer industry. *Econometrica*, 76(5):1017-1074.

Granovetter M (1985) Economic action and social structure: The problem of embeddedness. *Amer. J. Sociol.* 91(3):481-510.

Handel BR, Kolstad JT (2015) Health insurance for "humans": Information frictions, plan choice, and consumer welfare. *Amer. Econom. Rev.* 105(8):2449-2500.

Henseler J, Ringle CM, Sarstedt M (2015) A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Marketing Sci.* 43(1):115-35.

Honka E, Ali H, Wildenbeest M (2019) Empirical search and consideration sets. Dubé JP, Rossi PE, eds. *Handbook of the Economics of Marketing* (Elsevier B.V., North-Holland), 193-257.

Koçaş C, Bohlmann JD (2008) Segmented switchers and retailer pricing strategies. *J. Marketing* 72(3):124-142.

Mas-Collel A, Whiston MD, Green JR (1995), *Microeconomic Theory* (Oxford University Press, UK).

Masatlioglu Y, Nakajima D, Ozbay EY (2012) Revealed attention. *Amer. Econom. Rev.* 102(5):2183-2205.

Moraga G, Wildenbeest M (2012) Price comparison websites. Peitz M, Waldfogel J, eds *The Oxford Handbook of the Digital Economy* (Oxford University Press, UK), 224-253.

Honka E, Ali H, Wildenbeest M (2019) Empirical search and consideration sets. Dubé JP, Rossi PE, eds. *Handbook of the Economics of Marketing* (Elsevier B.V., North-Holland), 193-257.

Honka E, Chintagunta P (2017) Simultaneous or sequential? Search strategies in the US Auto Insurance Industry. *Marketing Sci.* 36(1):21-42.

Morgan P, Manning R (1985) Optimal search. *Econometrica* 53(4):923-944.

Natter M, Ozimec AM, Kim JY (2015) Practice prize winner - ECO: Entega's profitable new customer acquisition on online price comparison sites. *Marketing Sci.* 34(6):789-803.

Nunnally JC, Bernstein IH (1994) *Psychometric Theory,* 3rd ed. (McGraw Hill, NY)

Ratchford BT, Lee MS, Talukdar D (2003) The impact of the Internet on information search for automobiles. *J. Marketing Res.* 40(2):193-209.

Ringel DM, Skiera B (2016) Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Sci.* 35(3):511-534.

Rogerson R, Shimer R, Wright R (2005) Search-theoretic models of the labor market: A survey. *J. Econom. Lit.* 43(4):959-988.

Ronayne, D (2018) Price comparison websites. Working Paper, University of Minnesota.

Shapiro S, MacInnis DJ, Heckler SE (1997) The effects of incidental ad exposure on the formation of consideration sets. *J. Consumer Res.* 24(1):94-104.

Shocker DA, Ben-Akiva M, Boccara B, Nedungadi P (1991) Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters* 2(3):181-197.

Smith MD (2002) The impact of shopbots on electronic markets. *J. Acad. Marketing Sci.* 30(4):446-454.

Statista (2019) Average nominal annual premium for basic insurance under Dutch health insurance act (zvw) per person from 2007 to 2019 (in euros). (July 8, 2019) https://www.statista.com/statistics/581710/netherlands-average-nominal-annual-premium-basic-insurance-per-person-under-the-dutch-health-insurance-act-zvw/

Stigler GJ (1961) The economics of information. *J. Political Econom.* 69(3):213-225.

Swait JD (1984) Probabilistic choice set generation in transportation demand models. Doctoral Dissertation, Massachusetts Institute of Technology, Boston.

Swan JE, Combs LJ (1976) Product performance and consumer satisfaction: A new concept. *J. Marketing* 40(2):25-33.

Terui N, Ban M, Allenby G (2011) The effect of media advertising on brand consideration and choice. *Marketing Sci.* 30(1):74-91.

Verzekeringen (2013) Bijna 80% procent consumenten stapt online over van zorgverzekering. (accessed Dec 5, 2017), https://www.verzekeringen.com/nieuws/bijna-80-procent-consumenten-stapt-online-over-van-zorgverzekering

Waldfogel J, Chen L (2006) Does information undermine brand? Information intermediary use and preference for branded web retailers. *J. Indust. Econom*. 54(4): 425-449.

Weitzman ML (1979) Optimal search for the best alternative. *Econometrica* 47(3):641-654.

Appendices

Appendix 1. MTMT Ratios

Table A1. MTMT Ratios

MTMT Ratios	1.	2.	3.	4.	5.
1. Better deal elsewhere	-				
2. Price dissatisfaction	.78	-			
3. Non-price dissatisfaction	.57	.81	-		
4. Political discussion	.33	.28	.48	-	
5. Changes in personal situation and health	.50	.38	.68	.51	-

Appendix 2. Factor Loadings

Table A2. Factor Loadings

Items	1.	2.	3.	4.	5.
Lower price others	**.840**	.115	.087	.241	.027
Better coverage by other alternatives	**.726**	.346	.236	.095	-.082
Promotion offered by other alternatives	**.828**	.060	.089	.108	.195
Premium current too high	.481	.305	.122	**.612**	.058
Premium rises too much	.449	.319	.201	**.560**	.050
Trouble paying current premiums	.101	.240	.201	**.828**	.132
Dissatisfied coverage current	.262	**.783**	.170	.199	.023
Dissatisfied with contracting of current alternative	.133	**.712**	.266	.172	.184
Dissatisfied with service of current alternative	.106	**.819**	.139	.242	.078
Political discussion	.102	.160	.129	.129	**.941**
Changes in personal situation	.169	.189	**.831**	.165	.110
Changes affecting personal health	.138	.246	**.825**	.158	.103

Notes: Loadings larger than .5 are printed in boldface.

Appendix 3. Summary Table for All Variables Used in Study

Table A3. Summary Table for All Variables

Source	Variable	Description	Items	Scale
Browsing data and post-purchase survey	CS_i	Visit to a comparison site by consumer i	Visit to a comparison site by consumer i	Binary variable. 1 for yes; 0 for no
Browsing data and post-purchase survey	CS_{it}	Visit to a comparison site by consumer i at time t, where t is start of search or end of search	Visit to a comparison site by consumer i at the start or end of search	Binary variable. 1 for yes; 0 for no
Pre-search survey	$\mathrm{ConsSet}_i$	Size of initial consideration set	"Which of the following alternatives would you consider for your health insurance?"	Natural logarithm of the number of alternatives considered by consumer
Pre-search survey	$\mathrm{BetterDeal}_i$	Various aspects by which other alternatives have a better deal than current alternative	(i) "I would be able to get better coverage with another health insurance", (ii) "I would be able to pay a lower insurance premium with another health insurance", (iii) "Another alternative offers an interesting promotion (to switch)".	Scale from 1 to 4; where 1 = plays no role in search and 4 = plays a large role in search
Pre-search survey	$\mathrm{DissatPrice}_i$	Dissatisfaction with price related characteristics of current alternative	(i) "The insurance premium of my current alternative is (too) high", (ii) "The insurance premium of my current health insurance rises too much in 2014", and (iii) "I have troubles to keep paying my current health insurance."	Scale from 1 to 4; where 1 = plays no role in search and 4 = plays a large role in search
Pre-search survey	$\mathrm{DissatNonPrice}_i$	Dissatisfaction with non-price related characteristics of current alternative	(i) "I am dissatisfied with the coverage of my current health insurance", (ii) "I am dissatisfied about the service offered by my current health insurance alternative", (iii) "I am dissatisfied about my current health insurance alternative no longer contracting (reimbursing) certain hospitals"	Scale from 1 to 4; where 1 = plays no role in search and 4 = plays a large role in search
Pre-search survey	$\mathrm{Political}_i$	Political discussion	"Political discussion"	Scale from 1 to 4; where 1 = plays no role in search and 4 = plays a large role in search
Pre-search survey	$\mathrm{Changes}_i$	Changes in personal situation and health	(i) "Changes in my personal situation", (ii) "Changes affecting my personal health"	Scale from 1 to 4; where 1 = plays no role in search and 4 = plays a large role in search
Pre-search survey	$\mathrm{SearchTendency}_i$	Tendency to search for information	"I like to shop around when I need a new financial product"	Scale from 1 to 5; where 1 is very much disagree and 5 is very much agree
Post-purchase survey	$\mathrm{Purchase}_i$	Date of insurance purchase	On which date did you sign your health insurance for 2014?	Day, month and year

Appendix 4. Uplifts in Comparison Site Visit Probabilities

We provide the equations necessary to derive the uplifts in comparison site visit probabilities for the initial consideration set size $\Delta_{\mathrm{ConsSet}}$ and the expectations of finding a better deal $\Delta_{\mathrm{BetterDeal}}$. The baseline comparison site visit probability at the start of search when all independent variables are at the mean is:

$$\widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1) = \Phi(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{start}}) \tag{1}$$

where Φ is the cumulative normal distribution, $\overline{\mathrm{X}}$ is the vector of independent variables and control variables set at their mean values and a constant term, and where $\widehat{\beta}_{\mathrm{start}}$ is the vector of coefficient estimates.

Thereafter, we add a standard deviation to each independent variable and obtain the updated predicted comparison site visit probabilities. The expected comparison site visit probability at the start of search at an increased level of (i) initial consideration set size, and (ii) expectations of finding a better deal are:

$$\widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1) = \Phi(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{start}} + \beta_{1,\mathrm{start}}\hat{\sigma}_{\mathrm{ConsSet}}) \tag{2}$$

$$\widehat{\mathrm{P}}_{\mathrm{BetterDeal}}(\mathrm{CS}_{\mathrm{start}} = 1) = \Phi(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{start}} + \beta_{2,\mathrm{start}}\hat{\sigma}_{\mathrm{BetterDeal}}) \tag{3}$$

The baseline comparison site visit probability at the end of search when all independent variables are at the mean is:

$$\begin{aligned}\widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{end}} = 1) = \left(1 - \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1)\right)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}}|\mathrm{CS}_{\mathrm{start}} = 0\right) + \\ \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}}|\mathrm{CS}_{\mathrm{start}} = 1\right)\end{aligned} \tag{4}$$

where Φ is the cumulative normal distribution, $\overline{\mathrm{X}}$ is the vector of independent variables and control variables set at their mean values (with the exception of $\mathrm{CS}_{\mathrm{start}}$ as explained below) and a constant term, $\widehat{\beta}_{\mathrm{end}}$ is the vector of coefficient estimates.

Our end-of-search comparison site visit incidence model uses $\mathrm{CS}_{\mathrm{start}}$ as an independent variable. In equation 4, we thus account for the two possibilities in which a comparison site visit occurs at the end of search. The first possibility is the probability of a comparison site visit at the *end* of search conditional on no visit at the *start* of search which is reflected by the first right hand side term in equation 4, i.e., $\left(1 - \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1)\right)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}}|\mathrm{CS}_{\mathrm{start}} = 0\right)$, where $\overline{\mathrm{X}}$ contains $\mathrm{CS}_{\mathrm{start}}$ which takes the value 0. The second possibility is the probability of a comparison site visit at the *end* of search conditional on a visit at the *start* of search and which is reflected by the second right hand side term in equation 4, i.e., $\widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}}|\mathrm{CS}_{\mathrm{start}} = 1\right)$, where $\overline{\mathrm{X}}$ contains $\mathrm{CS}_{\mathrm{start}}$ which takes the value 1.

Similarly, we add a standard deviation to each focal independent variable and obtain the updated predicted comparison site visit probabilities for initial consideration set size and expectations of finding a better deal:

$$\begin{aligned}&\widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{end}} = 1) \\ &= \left(1 - \widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1)\right)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}} + \beta_{1,\mathrm{end}}\hat{\sigma}_{\mathrm{ConsSet}}|\mathrm{CS}_{\mathrm{start}} = 0\right) + \\ &\widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}} + \beta_{1,\mathrm{end}}\hat{\sigma}_{\mathrm{ConsSet}}|\mathrm{CS}_{\mathrm{start}} = 1\right)\end{aligned} \tag{5}$$

$$\begin{aligned}&\widehat{\mathrm{P}}_{\mathrm{BetterDeal}}(\mathrm{CS}_{\mathrm{end}} = 1) \\ &= \left(1 - \widehat{\mathrm{P}}_{\mathrm{BetterDeal}}(\mathrm{CS}_{\mathrm{start}} = 1)\right)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}} + \beta_{2,\mathrm{end}}\hat{\sigma}_{\mathrm{BetterDeal}}|\mathrm{CS}_{\mathrm{start}} = 0\right) + \\ &\widehat{\mathrm{P}}_{\mathrm{BetterDeal}}(\mathrm{CS}_{\mathrm{start}} = 1)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}} + \beta_{2,\mathrm{end}}\hat{\sigma}_{\mathrm{BetterDeal}}|\mathrm{CS}_{\mathrm{start}} = 1\right)\end{aligned} \tag{6}$$

5 and 6, we apply the same conditions as in equation 4. For example, in equation 5, we condition on the *absence* of a comparison site visit at the start of search, i.e., $\left(1 - \widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1)\right)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}} + \beta_{1,\mathrm{end}}\hat{\sigma}_{\mathrm{ConsSet}}|\mathrm{CS}_{\mathrm{start}} = 0\right)$ and on the *presence* of a comparison site visit at the start of search, i.e.,

$\widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1)\Phi\left(\overline{\mathrm{X}}\widehat{\beta}_{\mathrm{end}} + \beta_{1,\mathrm{end}}\hat{\sigma}_{\mathrm{ConsSet}}|\mathrm{CS}_{\mathrm{start}} = 1\right)$, where $\mathrm{CS}_{\mathrm{start}}$ is part of $\overline{\mathrm{X}}$ and takes the value 0 and 1 in these two terms, respectively. The same applies to equation 6. Equations 5 and 6 thus capture the *direct* effect of an increase in the initial consideration set size and expectations of finding a better deal on the probability of a comparison site visit at the end of search, as well as the *indirect* effect through the probability of a comparison site visit at the start of search (an increase in the initial consideration set size and expectations of a finding a better deal change the probability of a comparison site visit at the start of search, which, in turn, impacts the probability of a comparison site visit at the end of search).

We can then define the uplift in comparison site visit probability at the start of search for each variable as

$$\Delta_{\mathrm{start,ConsSet}} = \widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{start}} = 1) - \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1) \tag{7}$$

$$\Delta_{\mathrm{start,BetterDeal}} = \widehat{\mathrm{P}}_{\mathrm{BetterDeal}}(\mathrm{CS}_{\mathrm{start}} = 1) - \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{start}} = 1) \tag{8}$$

Analogously, the uplift in comparison site visit probability at the end of search for each variable is

$$\Delta_{\mathrm{end,ConsSet}} = \widehat{\mathrm{P}}_{\mathrm{ConsSet}}(\mathrm{CS}_{\mathrm{end}} = 1) - \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{end}} = 1) \tag{9}$$

$$\Delta_{\mathrm{end,BetterDeal}} = \widehat{\mathrm{P}}_{\mathrm{BetterDeal}}(\mathrm{CS}_{\mathrm{end}} = 1) - \widehat{\mathrm{P}}_{\mathrm{baseline}}(\mathrm{CS}_{\mathrm{end}} = 1) \tag{10}$$

The uplift in comparison site visit probability for each variable is then given by the difference in uplifts between the start and end of search:

$$\Delta_{\mathrm{ConsSet}} = \Delta_{\mathrm{start,ConsSet}} - \Delta_{\mathrm{end,ConsSet}} \tag{11}$$

$$\Delta_{\mathrm{BetterDeal}} = \Delta_{\mathrm{start,BetterDeal}} - \Delta_{\mathrm{end,BetterDeal}} \tag{12}$$

Appendix 5. Correlations between Start Date of Search, and the Independent Variables

We provide the correlations between the start date of search and the independent variables. As seen from Table A5, only search tendency is weakly correlated with the start date of search.

Table A5. Correlations between Start Date of Search and Independent Variables

	Start date
Consideration set	-.01
Better deal elsewhere	-.00
Price dissatisfaction	-.01
Non-price dissatisfaction	-.01
Political discussion	.01
Changes in personal situation and health	.01
Search tendency	-.07*

Note: * denotes significance based on α = .05, n = 2,478

Essay 2

Detecting Interactions: A Machine Learning Approach

Abstract

Machine learning (ML) methods are increasingly popular in management and marketing research to discover interaction effects. Existing research focuses on features with high individual importance scores to construct candidates for interactions (HIMPS). However, HIMPS (i) detects many candidates, (ii) assumes that individually important features interact which is not necessarily true, and (iii) does not rank these candidates. To systematically detect interactions, we introduce two ML approaches where the first approach which is generalized Random Intersection Trees (gRITs) detects interactions through generalized Random Intersection Trees (RITs) applied to a random forest (RF) output. Our second approach detects interactions using the H-statistic, which calculates the interaction's share of partial dependence to the sum of each main effect's partial dependence. Comparing HIMPS, gRITs and the H-statistic conceptually and empirically using simulation studies, we find that the H-statistic outperforms HIMPS and gRITs for correctly detecting and ranking pre-specified interactions. We then apply the H-statistic on a real large scale consumer search dataset to obtain substantive insights. Specifically, we find that images with high distortion and high saturation yield lower predicted click probability than images with low distortion and high saturation. We conclude by providing recommendations to scholars and practitioners.

Keywords: interactions, random intersection trees, random forest, machine learning, H-statistic

1. Introduction

There is a growing emphasis in management research on using machine learning (ML) methods to discover new and robust relationships in data (Choudhury et al. 2021) for theory building. Of particular interest to scholars is the discovery of (potentially nonlinear) interactions between features and the target[9] (e.g., Cui and Curry 2005, Dzyabura and Narasimhan 2018, Hagen et al. 2020, Lemmens and Croux 2006, Ma and Sun 2020). Interactions are at the heart of social sciences (Anderson et al. 2014, Cohen et al. 2012) as they explain the boundary conditions of the relationship between constructs and allow managers to adjust their marketing policy levers (MacInnis 2001). Generally, scholars focus on two-way interactions "given the lack of theoretical relevance for higher-order interactions" (Venkatraman 1989). One popular machine learning (ML) method to detect these interactions is the random forest (RF) which produces feature importance scores that measure the relative impact of each feature on predicting the target. Thereafter, researchers focus on features with the highest importance scores (Breiman 2001, Choudhury et al. 2021, Lemmens and Croux 2006) and visualize each feature to surmise its main effect, as well as pairs of features to explore interaction effects. We use the term 'high importance score approach' or HIMPS (for brevity) when referring to the literature's approach of focusing on features with highest importance scores as candidates for interactions.

While HIMPS is popular for discovering interactions (e.g., Chen et al. 2011, Choudhury et al. 2021, Lemmens and Croux 2006, Lunetta et al. 2004, Wei and Li 2007), it is potentially problematic for three reasons. First, feature importance scores indicate each individual feature's relative importance but not their interaction's importance for predicting the target (Darst et al. 2018, Murdoch et al. 2018). Specifically, individually unimportant

[9] Consistent with machine learning taxonomy, we use the terms 'feature' and 'target' to denote an independent variable and dependent variable respectively.

features can interact and influence the target's prediction (Wright et al. 2016). Second, looking at all possible combinations of highest-scoring features for potential interactions is inefficient (Shah and Meinshausen 2014, Shah 2016). For instance, focusing on the 10 highest-scoring features gives us (10*9)/2 = 45 possible two-way interactions to explore. Hence, we need a more systematic manner to discover the most important interaction effects in the data. Third, HIMPS does not rank its detected interactions which raises the interesting question of which interaction is more important, e.g., the interaction between the top-scoring and fourth-scoring feature or the interaction between the second-scoring and the third-scoring feature?

In this study, we respond to the challenges raised in our preceding discussion and make an important contribution towards detecting interaction effects using machine learning methods. We introduce two approaches for detecting interactions from the ML literature. We then discuss the differences between these two approaches and compare them empirically against each other and HIMPS to assess which approach is superior under different contingency scenarios. The first approach is the generalized Random Intersection Trees (gRITs) which detect interactions between binary and non-binary features in predicting the target using a RF to index active features before applying Random Intersection Trees (RITs) on the indexed features (Basu et al. 2018). gRITs are derived from RITs, where RITs detect interactions between binary features on binary targets using the presence of co-occurring features on a decision path from root to leaf nodes across randomly sampled observations, for a given target class (Shah and Meinshausen 2014). Our second approach employs the H-statistic by Friedman and Popescu (2008) which quantifies an interaction's strength using the share of variance explained by the interaction to the sum of each main effect's partial dependence. We discuss the potential differences in correctly detecting, ranking, and inferring the directionality of the interactions between HIMPS, gRITs, and the H-statistic through two contingency scenarios which are likely to occur in empirical settings.

We ascertain which approach performs best amongst HIMPS, gRITs and the H-statistic using three simulation studies with a (i) baseline scenario with equally important features and uncorrelated interacting features, a (ii) contingency scenario 1 where there is a relatively more important non-interacting feature, and a (iii) a contingency scenario 2 with correlated interacting features, respectively. Assessing the effectiveness of these three approaches on correctly detecting, ranking, and inferring the directionality of pre-specified interactions, we find that the H-statistic outperforms both HIMPS and gRITs on all three metrics. This finding suggests that the H-statistic is more effective than HIMPS and gRIT and hence, we adopt H-statistic approach for our subsequent empirical application.

Our empirical application's goal is to determine which features and their interactions predict clicks in the consumer's online search for property listings. For instance, do consumers evaluate each feature individually or jointly as a combination of features for a given listing? Since interactions determine "the boundaries of generalizability, and as such constitute the range of the theory" (Whetten 1989) between a feature and target, interactions allow us to discern between competing theories. Further, marketers can utilize knowledge of boundary conditions to attain their desired marketing outcomes (MacInnis 2001). Our dataset is obtained from the largest property listing platform which recorded all actions taken by a random sample of 6,568 consumers during their browsing, and the image features and non-image features of all condominium rental listings displayed to consumers. As each condominium has a fixed layout for a given number of bedroom and bathrooms, we keep non-image features constant by specifying dummies using a triplet of condominium-bedroom-bathroom and study changes in image features on predicted clicks and its interactions with other features. Since this dataset is heavily imbalanced (i.e., 2% of observations are clicked), the performance of the ML methods for predicting clicks is affected (Hasanin and Khoshgoftaar 2011, Japkowicz and Stephen 2002). Hence, we resampled this dataset by taking all clicked observations, and randomly sampling an equal number of unclicked observations (e.g., King and Zeng 2001). Applying our chosen approach to detecting and quantifying interactions on this resampled dataset, we have the following key findings. First, we find that four of the top five interactions detected by the H-statistic included interactions between image features, and only one interaction did not include image features. In contrast, none of the top six interactions from HIMPS included any image features. By applying HIMPS, we would miss out on interactions involving image features. Second, we detect interactions between image features where jointly higher values of interacting features are associated with a lower predicted click probability than jointly lower feature values. For instance, a listing whose image has a higher distortion and higher saturation is associated with a lower average predicted click probability than a listing whose image has a lower distortion and lower saturation.

Our study contributes to the nascent and growing stream in machine learning methods used in management and marketing. First, we introduce two approaches from the ML literature to efficiently detect interactions which are the gRITs and H-statistic. We compare gRITs and H-statistic to HIMPS and discuss their potential differences before using simulation studies to ascertain which approach performs best under different contingencies for correctly detecting, ranking, and inferring the directionality of the interactions. By doing so, we guide users on the most appropriate approach for their empirical studies. Second, we show substantively that image features interact with each other in predicting clicks on property listings. Since listings with more attractive image features lead to higher demand (e.g., Zhang et al. 2017), marketers can alert users if their listing's current combination of image features yields a lower average predicted click probability than other combinations of image features based on the chosen ML method's predictions. Thereafter, marketers can offer a premium service to users which adjusts the interacting image features jointly using professional photo-editing software to obtain a higher predicted click probability than the current predicted click probability.

The rest of this paper proceeds as follows. We first recap the problems faced by researchers when trying to detect interactions using HIMPS. Thereafter, we introduce and compare two approaches for detecting interactions from the ML literature (i.e., gRITs and the H-statistic) against HIMPS and discuss potential differences in their performance for detecting interactions. Thereafter, we ascertain the most effective approach amongst these three approaches under different conditions using three simulation studies. After determining the most effective approach, we apply this approach on big consumer search data which is obtained from a large property listing platform. We end by summarizing the contributions of our study, and discussing avenues for further development.

2. Methods and visualization

To discover interactions, HIMPS examines interactions between pairs of features with the highest importance scores (e.g., Chen et al. 2011, Choudhury et al. 2021, Lemmens and Croux 2006, Lunetta et al. 2004, Wei and Li 2007). However, feature importance scores cannot detect the interactions between features (Murdoch et al. 2018) as individually important features do not necessarily interact with each other. To systematically detect interactions, a review of the machine learning (ML) literature suggests two general approaches. We explain each approach's strengths to determine the conditions favoring each approach and guide researchers on choosing the appropriate approach.

The first approach is *tree-based* where one collects all features of a randomly chosen observation along its decision path, and gradually removes features by taking pairwise feature intersections with other randomly chosen observations (e.g., Shah and Meinshausen 2014, Thanei et al. 2018). In doing so, feature interactions are detected in a computationally efficient manner. The tree-based approach was pioneered by Shah and Meinshausen (2014) in the context of binary features and targets, and laid the foundations for subsequent studies (e.g., Thanei et al. 2018). To deal with continuous features, gRITs were derived by Basu et al. (2018). We implement gRITs in this study as an example of the tree-based approach as researchers are likely to encounter a mix of binary and continuous features.

The second approach is *target-based* which quantifies the differences in average predicted target value with, and without interactions between features, for combinations of features (e.g., Friedman and Popescu 2008, Greenwell et al. 2018, Tsang et al. 2020). The H-statistic (Friedman and Popescu 2008) is a canonical method used to quantify interaction strength (Oh 2019) in this approach and extended in other studies (e.g., Greenwell et al. 2018, Tsang et al. 2020). Hence, we implement the H-statistic as an example of the target-based approach in this study.

We apply gRITs and H-statistics after fitting an RF to the dataset as gRITs convert continuous to binary features by fitting an RF on the original dataset and H-statistics uses the partial dependences from the fitted RF as an input. To fit an RF, we have two main ways which are (i) fitting a tuned RF (Choudhury et al. 2019) and (ii) fitting iteratively weighted RFs, also known as iRFs (Basu et al. 2018). In (i), researchers use a randomized search to tune the RF's key parameters (i.e., maximum number of features sampled at each node, depth of RF, minimum number of sampled observations in the leaf node, minimum decrease in class impurity at each node before a split can be carried out) (Choudhury et al. 2021). In such a randomized search, many RFs are fit on the training dataset with each RF using a different combination of randomized values for the key parameters. Thereafter, these RFs' performance are quantified using K-fold cross validation which splits the training dataset into

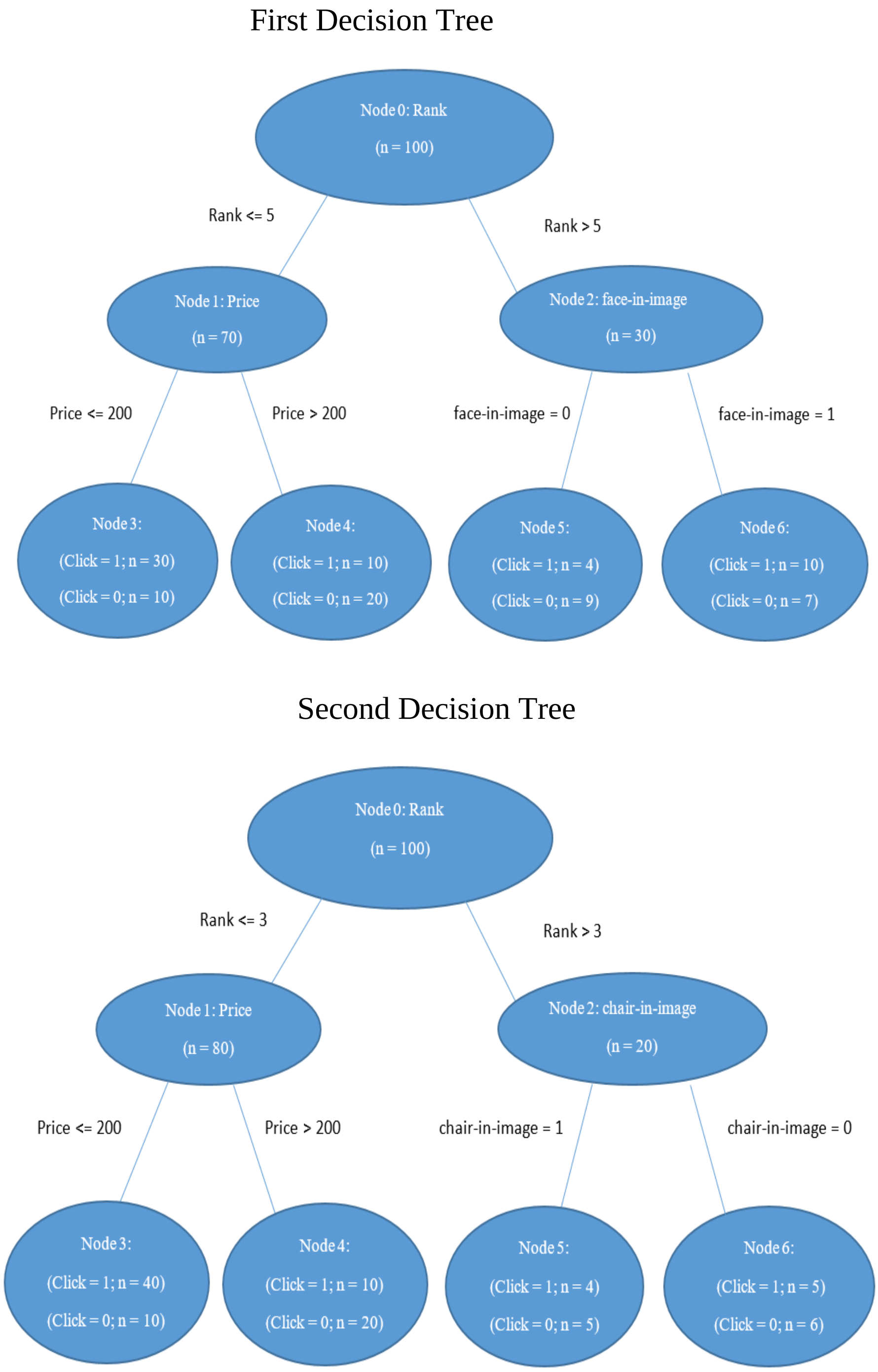
K subsets. Each RF is trained on K – 1 subsets, and its performance (e.g., area-under-curve-precision-recall or AUC-PR) is assessed on the subset that was omitted from its training. Thereafter, researchers choose the best-performing RF and implement it on the test dataset. In (ii), the current RF's feature importances are used as feature sampling probabilities in the RF's next iteration. Hence, features which are more informative in predicting the target are sampled with a higher probability in the iRF than the RF, which aids in the detection of higher-order interactions (Basu et al. 2018). However, the iRF also increases the risks of detecting false positive interactions (Basu et al. 2018, supplementary information 5.1). Since we study two-way interactions, applying the iRF does not compensate us for the increased risks of detecting false positives, for instance, by detecting more true positive two-way interactions. Thus, we assume that researchers use the tuned RF.

2.1 gRITs: Logic and implementation

RITs (Shah and Meinshausen 2014) were created to detect interactions in datasets with binary features and binary targets. In reality, features and targets can be a mix of both binary and non-binary (i.e., continuous) ones. To detect interactions between features, generalized RITs (gRITs) map continuous features into binary features using the rules extracted from the RF's application (Basu et al. 2018). We explain this procedure below in detail.

gRITs procedure. Suppose we have an Airbnb dataset with 100 consumers, where each consumer was randomly exposed to one of ten rental listings. Each listing has binary image features (chair-in-image, face-in-image), continuous non-image features (rank, price, floor area, number of bedrooms), and a binary target click with the value of 1 if the listing was clicked and 0 if not. Chair-in-image and face-in-image are binary features with the value of 1 if a chair and a face is present in the image respectively and price is the listing's daily rental price. Hence, we have a total of 100 observations where 80 observations have click = 1 and 20 observations have click = 0. Following our previous section, we assume that (i) we have obtained a tuned RF and (ii) this tuned RF only has two decision trees. We visualize these two trees in Figure 1 and use these to explain the gRITs procedure, where the goal is to detect interactions which predict a listing's clicks.

Figure 1. First and Second Decision Trees from the Tuned RF.

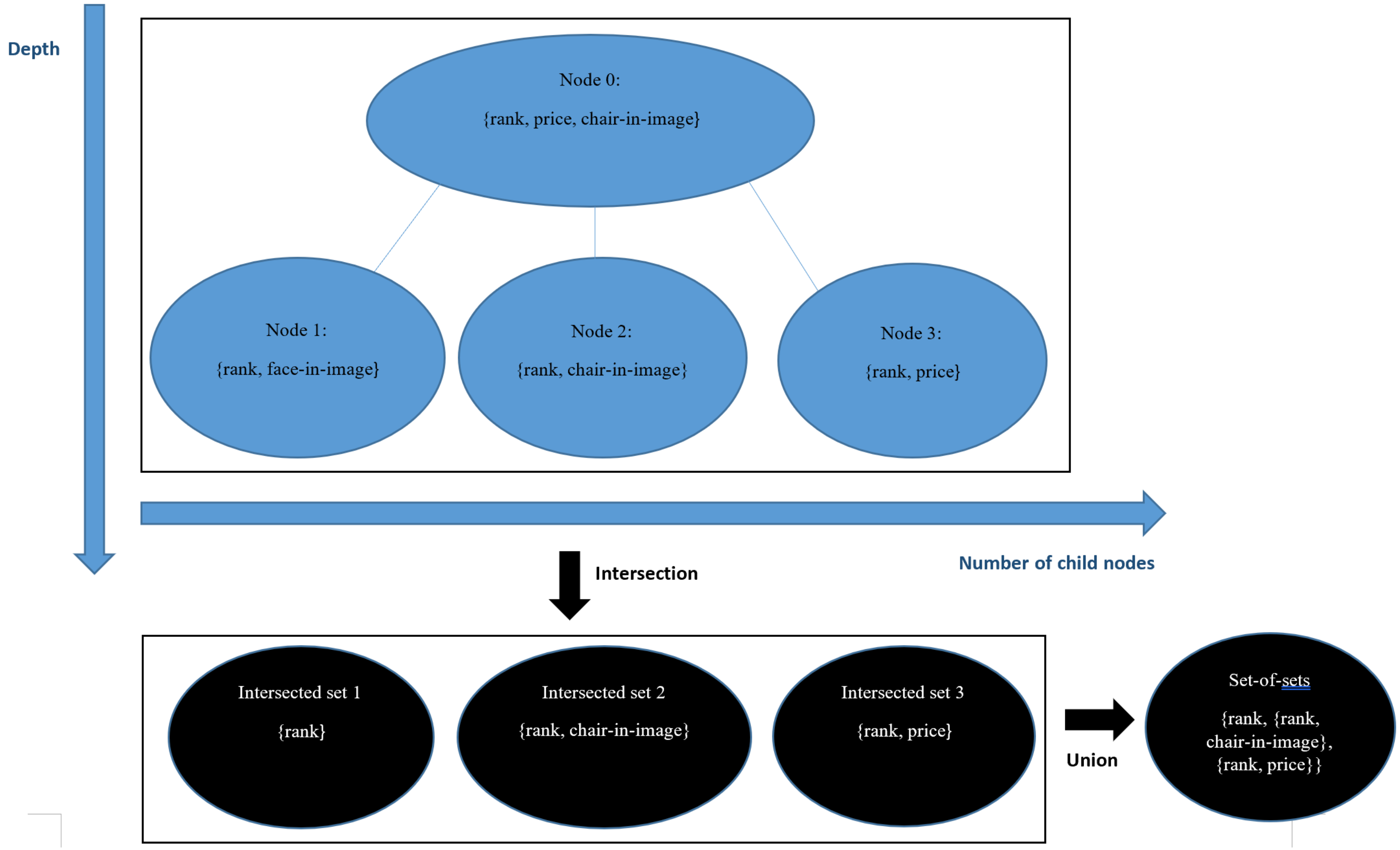


First Decision Tree

Second Decision Tree

In step 1, we randomly sample one observation from the hypothetical dataset which is conditional on a given class (i.e., click = 1). Suppose that this randomly sampled observation is customer 1 and listing 8 with features: (i) chair-in-image = 1, (ii) face-in-image = 1, (iii) rank = 4, (iv) price = $100, (v) floor area = 100 square meters, (vi) number of bedrooms = 3. To determine this observation's active feature sets, we use the two decision trees in Figure 1. From the first tree, we see that this observation is assigned to node 5 (leaf node) as its rank exceeds 3 and its chair-in-image is 1. As rank and chair-in-image are located on the path from the root node to node 5, these two features are the active features for this observation in the first tree. Turning our attention to the second tree in Figure 1, we see that this observation falls in node 3 (leaf node) as its rank is less than 5, and its price is less than $200. Thus, this observation has rank and price as its active features in the second tree as those features are situated on the path from the root node to node 3 (leaf node). Collecting the active features from these two paths from both decision trees, we obtain {rank, price, chair-in-image} as its set of active features. This set gives us the root node in the gRIT depicted in Figure 2.

Figure 2. First Tree in gRITs



In step 2, we grow another layer of nodes in the gRIT which are called child nodes. For each child node, one observation is randomly sampled conditional on the chosen class in step 1 (click = 1 in our example). Further, we assume three child nodes (i.e., nchild = 3) for ease-of-explanation. In each node, the gRIT collects the observation's active features as a set using the same procedure as in step 1. Referring to Figure 2, we can see that these active feature sets are (i) {rank, lights-in-image} for node 1, (ii) {rank, chair-in-image} for node 2, and (iii) {rank, price} for node 3. We keep growing layers of child nodes, where each child node in the current layer has nchild nodes in the subsequent layer, until the pre-determined maximum number of layers is reached. This pre-determined maximum number of layers is also known as the tree depth of gRITs (D), which we set as 1 for expositional purposes in our example, i.e., we grow one layer of child nodes.

In step 3, after creating all layers of child nodes, we intersect the active feature sets on the path from each leaf node to the root node in the gRIT, which gives us as many intersected sets as there are leaf nodes. In Figure 2, we have three leaf nodes which gives us three intersected sets, i.e., {rank}, {rank, chair-in-image}, and {rank, price}. We then take the union of {rank} and {rank, chair-in-image} and obtain a set-of-sets, which is {{rank}, {rank, chair-in-image}}.

In step 4, we repeat steps 1 to 3 M times to obtain M gRITs where each of these M gRITs produces one set-of-sets. We take the union of these set-of-sets to obtain a final set of co-occurring features for these M gRITs, which we denote by U. We summarize these four steps in Table 1 for ease-of-reading.

Table 1. Four Steps in gRITs

gRITs procedure
1. For each tree in each tree in gRITs, starting with the tree's root node, conditional on a class (e.g., class = 1) randomly choose 1 observation and collect its active features as a set in the root node.
2. Continuously grow additional layers of child nodes, where nchild nodes are grown at each layer until the maximum numbers of layers (i.e., tree depth) is reached.
3. For each child node, randomly choose K observations, and for each observation in each child node, collect its active features as a set. Intersect each of these sets with the set in step 2 to obtain an intersected set, then take the union of the intersected sets to obtain a set-of-sets for the given tree.
4. Repeat steps 1 to 3 for all trees, and take the unions of the set-of-sets across all trees to obtain a final set detected by the M gRITs, denoted by U.

Basu et al. (2018) suggest repeating steps 1 to 4 on B bootstrapped datasets which gives us B sets of U, that we denote as U(1), U(2), .., U(B-1), U(B). Using these B sets of U, we quantify the stability score of a set's co-occurring features as the proportion of times that the set's co-occurring features is recovered across B bootstrapped samples. For example, suppose we set B = 100 and recovered the set {rank, price} in ten samples. The stability score for {rank, price} is therefore given by 10/100 = .1. Since Basu et al. (2018) defines a score as a stable score if it is at least .5, {rank, price} is therefore not stable.

To understand why sets of co-occurring features are interpreted as interactions, consider an example where listings only have two binary features pool in image and balcony in image, which take the value of 1 is there is a pool and a balcony in the image respectively. We assume that pool-in-image and balcony-in-image positively interacts such that the probability of click = 1 is .3 when both pool-in-image = 1 and balcony-in-image = 1, and that the probability of click = 1 is only .1 when either pool-in-image = 1, or balcony-in-image = 1. By conditioning on click = 1 and randomly sampling observations, observations are three times more likely to have pool-in-image = 1 and balcony-in-image = 1, as compared to either pool-in-image = 1 or balcony-in-image = 1. Hence, these co-occurring features are detected as interactions in the gRITs with a higher probability, assuming that they are stable.

gRITs parameters. In the above example, we assumed some values for the parameters but Basu recommends tuning these parameters using simulations with an assumed data generating process (DGP) that includes pre-specified interactions, and quantifying the gRITs' accuracy in detecting interactions through three metrics (i) interaction-AUC, (ii) recovery rate, (iii) false positive weight, where a true positive is defined as interactions of arbitrary-order between features in the DGP. For example, Basu et al. (2018) obtains gRIT parameter values of M = 500, D = 5, nchild = 2, B = 30. A difficulty in applying this tuning approach is that scholars apply machine learning to better understand the DGP and thus may not know which DGP to assume in simulations. We will return to this difficulty subsequently in our simulation studies.

2.2 H-statistic: Logic and implementation

The H-statistic was derived by Friedman and Popescu (2008) and assumes features do not correlate. We first explicate the H-statistic's underlying logic before elaborating on its empirical application. For brevity, we denote the partial dependence (PD) function by F and the H-statistic by H subsequently in our explanation. The PD function, F, calculates the average predicted target value as if all observations have the same value of a focal feature, for each value of the focal feature (Choudhury et al. 2021). For instance, suppose three listings have floor areas of 400, 800 and 1,200 square feet respectively, the PD for 400 square feet assumes that the two listings with a floor area of 800 and 1,200 square feet now have a floor area of 400 square feet, before averaging the predicted target value (i.e., click probability) across all three listings. Users who wish to reduce the computational time can calculate F using a certain number of equal-spaced values (also known as grid points) instead of all observed values of the focal feature. For instance, if x is a continuous feature ranging from [0,100], 10 grid points gives us 10 values (i.e., 10, 20, 30, .. , 100) which is used by F to calculate the averaged predicted target at these 10 values. Note that F does not require actual observations at these 10 values because it assumes that observations have these values when calculating PD.

Logic of H-statistic. To understand the H-statistic, we draw on the hypothetical Airbnb dataset used to explain gRITs earlier, and assume that the number of bedrooms interacts with floor area to predict the average clicks on a listing. For instance, suppose listing 3 has five bedrooms and a floor area of 1,200 square feet with an average predicted clicks of 0.5 from our tuned RF. To calculate listing 3's H-statistic, F first calculates the average predicted clicks for five bedrooms, and the average predicted clicks for 1,200 square feet which we assume to be 0.25 and 0.15 respectively. Since .25 and .15 sums up to 0.4, the joint combination of number of bedrooms and floor area, yield a higher average predicted click probability than the sum of average predicted clicks for each feature only. Hence, these features interact. Formally, the partial dependence of a pair of features for the i-th observation (i.e., F($x_{i,1}$, $x_{i,2}$)) is the sum of each feature's partial dependence (i.e., F($x_{i,1}$) + F($x_{i,2}$)) when these features do not interact (Molnar 2021). The difference between F($x_{i,1}$, $x_{i,2}$) and F($x_{i,1}$) + F($x_{i,2}$) is therefore the PD due to the interaction between x1 and x2, for observation i. Using an RF, we can obtain F($x_{i,1}$, $x_{i,2}$), F($x_{i,1}$), F($x_{i,2}$) and define the H-statistic as

$$\mathrm{H}(x_{i,1}, x_{i,2}) = \sum_{i=1}^{n} \frac{[\mathrm{F}(x_{i,1}, x_{i,2}) - \mathrm{F}(x_{i,1}) - \mathrm{F}(x_{i,2})]^2}{[\mathrm{F}(x_{i,1}, x_{i,2})]^2} \tag{1}$$

where i indexes the observation number with n total observations in our dataset, and the numerator is interpreted as the share of variance in PD accounted for by the interaction between x1 and x2 (Friedman and Popescu 2008, Molnar 2021). While we define the H-statistic using pairs of features, the statistic can also be derived for higher-order combinations of features (Molnar 2021).

Empirical implementation. To obtain the H-statistic, we calculate F for each observation. With n observations in the data for each feature, a pairwise interaction will require n times n calculations. Hence, calculating the H-statistic is computationally expensive. To reduce the computational time and resources, one could take enough small random samples (Molnar 2021) or reduce the number of grid points used (Greenwell 2017) when calculating PD in the RF.

2.3 Two contingencies for interaction detection by HIMPS, gRITs and H-statistic

A priori, there are two contingency scenarios which explain possible differences in how effectively gRITs and H-statistic detect interactions, relative to a baseline scenario where all features are equally important, and interactions are differentially important. In this baseline scenario, HIMPS is uninformative as it uses top-scoring features to construct interaction candidates as features are equally important. Literature suggests that gRITS detect true interactions effectively (Basu et al. 2018) and the H-statistic is well defined as the PD of differentially important interactions and the PD of each feature can be calculated, which suggests that gRITs and H-statistic might be better at detecting true interactions.

The first contingency scenario arises with the existence of a relatively more important feature that does not interact with other features. Since HIMPS relies on the top scoring features to construct interactions, it is likely to detect false positive interactions involving this feature. gRITs are less likely to detect false positives as it is a ML method that is specifically derived to detect interactions (Basu et al. 2018). The H-statistic is also less likely to detect false positives as the difference in PD between non-interacting features is likely to be small since the interaction between these features do not predict the target.

The second contingency scenario occurs when interacting features which predict the target are correlated with each other. Generally, it is hard for ML methods to recover interactions when interacting features are highly correlated (Basu et al. 2018) and raises the possibility that the three ML methods (HIMPS, gRITS and H-statistic) are less likely to detect interactions. Moreover, the H-statistic calculates each feature's PD by assuming that the feature is uncorrelated with other features. If the feature is instead correlated, the H-statistic is incorrectly calculated (Molnar 2019) which lowers its effectiveness in detecting true interactions with correlated interacting features.

2.4 Visualization using partial dependence plots (PDPs)

After detecting promising top-ranked interactions from gRITs and/or H-statistics, we visualize the relationships between these interactions and the target. Visualization lets us understand how our average predicted target varies with these interactions as the RITs and H-statistics quantifies these interactions' stability and strength but not the directionality of the target's variation with these interactions. Generally, management and marketing use PDPs to visualize interactions (e.g., Choudhury et al. 2021, Lemmens and Croux 2005). PDPs show the average predicted target value for each combination of feature values, conditional on other features (Hastie et al. 2009, Choudhury et al. 2021), where the same PD is used to calculate the H-statistic. Interpreting PDPs is relatively easy since they give the total effect of an interaction (Choudhury et al. 2021), i.e., the sum of the main and interaction effects.

Settings for PDPs. The key parameter in PDPs is the number of grid points which is inherited from the PD function and a higher number of grid points increases the amount of computational time needed to generate PDPs. For instance, a dataset with N observations, K grid points, and J features requires K^J x N calculations for PD plots (Apley and Zhu 2020). To reduce computational time, researchers can lower the number of grid points (i.e., K) which lowers the number of feature values used in calculating PD. However, researchers should bear in mind that fewer grid points may obscure the finer details of the relationship between the feature(s) and target.

2.5 Software implementation: Tuning RFs, gRITs, H-statistic, and PDPs

Except for gRITs and the H-statistic, stand-alone Python packages are available for tuning RFs and PDPs. gRITs are currently only available as part of the iRF package in Python (Yu Group 2020) while the H-statistic is only available as an R package (Molnar and Schratz 2020). For this study, we coded up a tuned RF, gRITs and the H-statistic into one complete Python workflow which is available upon request. We discuss the commonly used Python packages for tuning RFs, and PDPs in Appendix 6 Table A6 along with the required commands and their key parameters for the benefit of users who wish to use these methods.

3. Simulation study and design

In the following simulation studies, we compare the ability of HIMPS, gRITs, and the H-statistic to (i) detect all the true interactions, (ii) rank all the true interactions correctly, and (iii) infer the directionality of the true interactions correctly given a pre-specified data generating process (DGP), based on the baseline scenario and two contingency scenarios in section 2.3. Thereafter, we choose the best performing ML method and apply it to our empirical setting to detect, rank, and infer the directionality of interactions.

3.1 Baseline scenario. Equally important and uncorrelated interacting features

Our simulated dataset has one binary target (y) and seven uncorrelated features (x1 to x7) where the features follow a standard normal distribution N(0,1). We purposely generated x1 to x4 to be equally important as HIMPS will be uninformative and potentially recover more false positives since it uses top-scoring features to construct pairwise interactions. When applying HIMPS, literature does not advise on the number of features to be used for constructing two-way interactions. Hence, we use the top four features in terms of their feature importance scores to obtain six two-way interactions. For the H-statistic, we limit our attention to the top five interactions with the highest values of the statistic. Finally, we only focus on interactions with stability scores >= .5 from the gRITs, following Basu et al. (2018).

Simulation details. First, we define an intermediate variable z as

$$\mathrm{z} = -.1 + \mathrm{x1} + \mathrm{x2} + \mathrm{x3} - \mathrm{x4} + \mathrm{x5}_{\mathrm{positive}} + \mathrm{x1x2} \ - 2\mathrm{x3x4} \tag{2}$$

where $\mathrm{x5}_{\mathrm{positive}}$ is an indicator variable taking the value of 1 if x5 > 0 and 0 otherwise. As evident from z, we have four linear main effects (x1 to x4), two interaction effects (x1x2 and x3x4), and one nonlinear main effect ($\mathrm{x5}_{\mathrm{positive}}$). We deliberately omitted x6 and x7 from equation 1 so these variables have zero importance, to validate the RF's ability for correctly identifying features that are uninformative for predicting the target. Also, the coefficient is 1 for x1 to x4, which reflects the equal importance of these features as the relative coefficient of one feature to another indicates this feature's relative importance on the predicted probability of the target taking the value of one. Second, we define the probability of a successful draw pr using z

$$\mathrm{pr} \ = \ \exp(\mathrm{z})/(1 + \exp(\mathrm{z})) \tag{3}$$

Third, we generate y as a Bernoulli random variable taking the value of 1 with probability pr; we chose y as a binary target to mimic our empirical setting which observes if each consumer clicks on a given listing. We set β_0 at the value of -.1 to ensure that pr has a mean of approximately .50. To facilitate ease-of-reading, we summarize the distributional assumptions, parameter values and the correlation structure of the features in Table 2.

Table 2. Summary of Parameter Settings for Baseline Scenario

Distributional Assumptions, Parameter Values and Correlation Structure
1. x1 to x7 are distributed as standard normal variables
2. $\mathrm{x5}_{\mathrm{positive}} = 1$ if $\mathrm{x5} > 0$; 0 otherwise
3. $\mathrm{z} = -.1 + \mathrm{x1} + \mathrm{x2} + \mathrm{x3} - \mathrm{x4} + \mathrm{x5}_{\mathrm{positive}} + \mathrm{x1x2} \ - 2\mathrm{x3x4}$
4. x1 to x7 are uncorrelated with each other

We use the settings just introduced and simulate datasets using five different seeds (i.e., five datasets in all) to ensure that our results are robust to the chosen seed. For each dataset, we tune a RF following Choudhury et al. (2021) before applying and comparing HIMPS, gRITs and the H-statistic on the three metrics introduced earlier (i) correctly detecting interactions, (ii) correctly ranking detected interactions, and (iii) correctly inferring the directionality of the detected interactions. To implement gRITs, we adopt the parameter settings in Basu et al. (2018). Alternatively, one could tune gRITs by specifying a DGP and testing for its accuracy in detecting interactions (Basu et al. 2018). Since researchers would not know the true DPG, we adopt the baseline gRIT settings from Basu et al. (2018) as a neutral reference point. We compare the performance of each method in correctly (i) detecting, (ii) ranking and (iii) inferring the directionality of interactions, using the proportion of datasets in which (i) to (iii) was met.

Validation check. We validate our datasets by ranking each feature using its importance scores and verified that this ranking reflects the correct order of these features based on the DGP, where the scores are displayed in Figures 3a to 3e. From these figures, we see that in all but one dataset, x6 and x7 have the lowest scores and x4 has the highest scores. Also, x3 has the second highest score and highest score in two and one datasets respectively.

Figure 3a. Feature Importance for Dataset 1

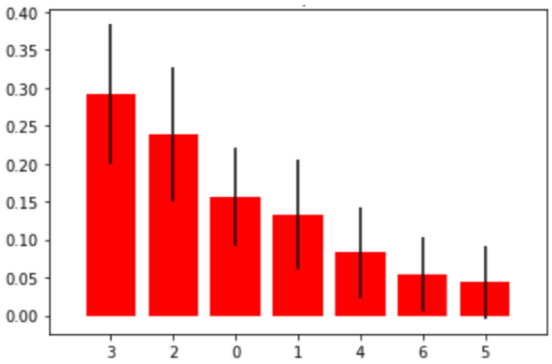


Figure 3b. Feature Importance for Dataset 2

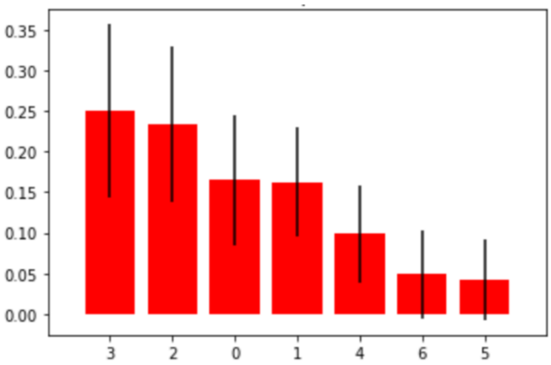


Figure 3c. Feature Importance for Dataset 3

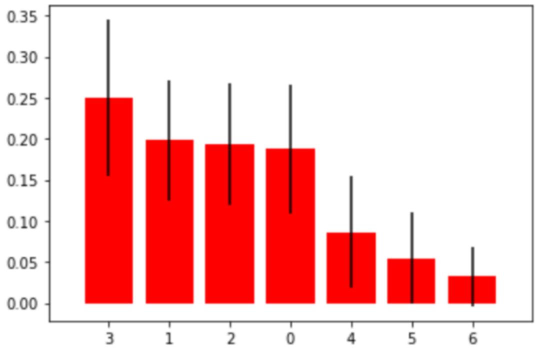


Figure 3d. Feature Importance for Dataset 4

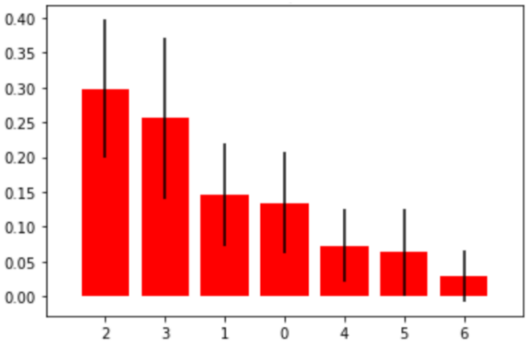
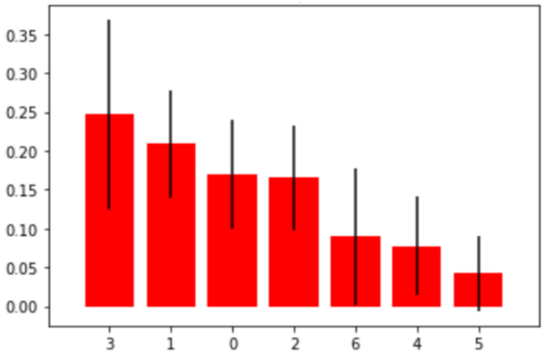


Figure 3e. Feature Importance for Dataset 5



Next, Figures 4a to 4e illustrate the PDPs for the informative features (x1 to x5) which are consistent with how we specified x1 to x5 in the DGP. For instance, the PDP for x5 is a step function and the PDPs for x1 to x4 are a linear function.

Figure 4a. PDPs for Informative Features, Dataset 1

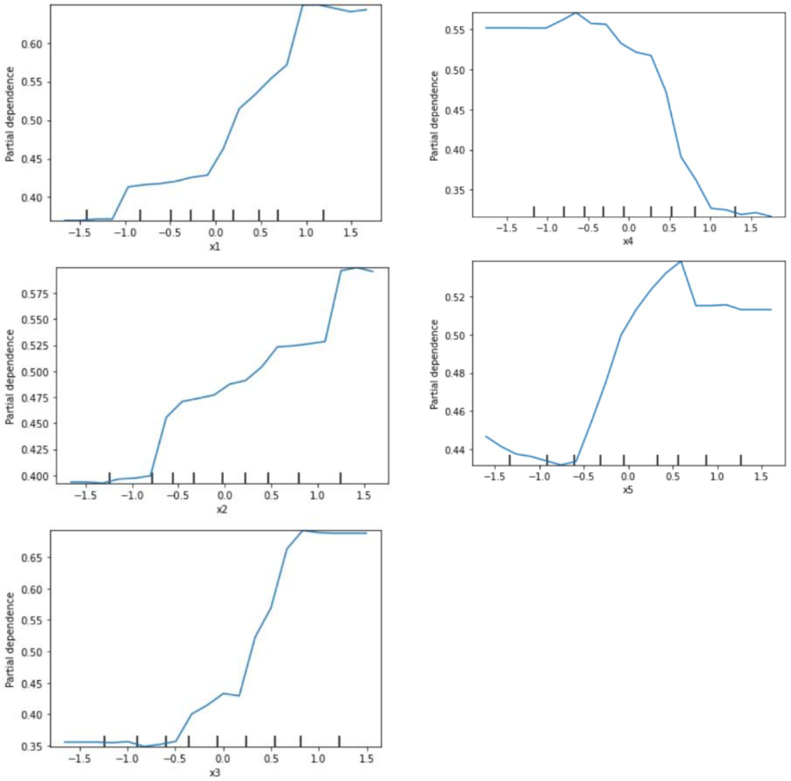


Figure 4b. PDPs for Informative Features, Dataset 2

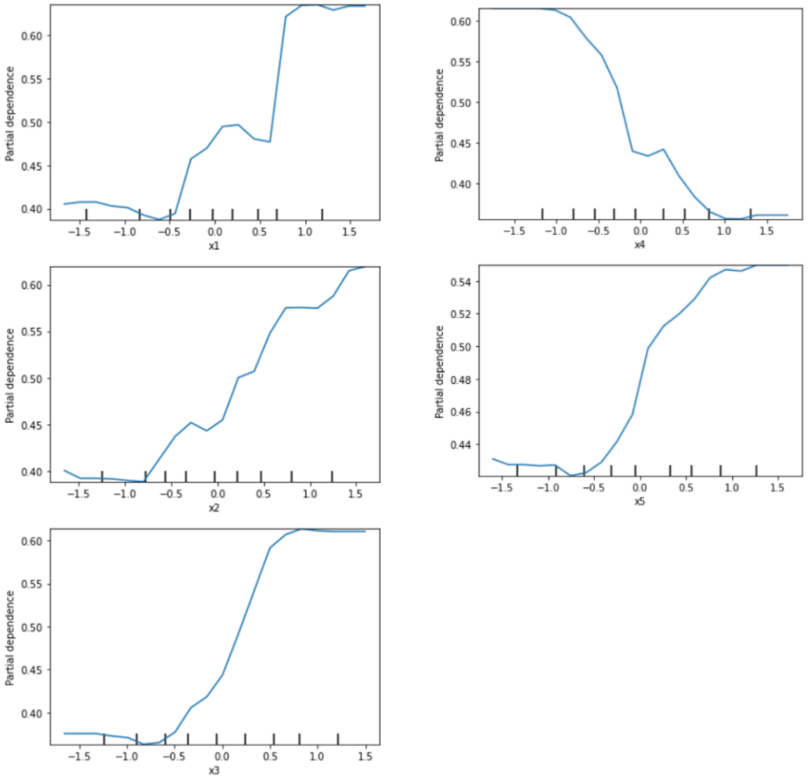


Figure 4c. PDPs for Informative Features, Dataset 3

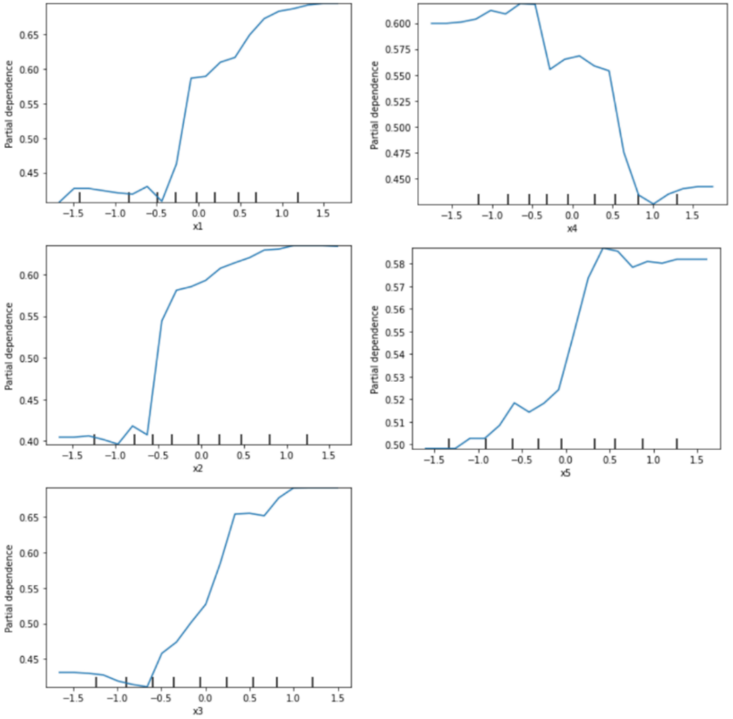


Figure 4d. PDPs for Informative Features, Dataset 4

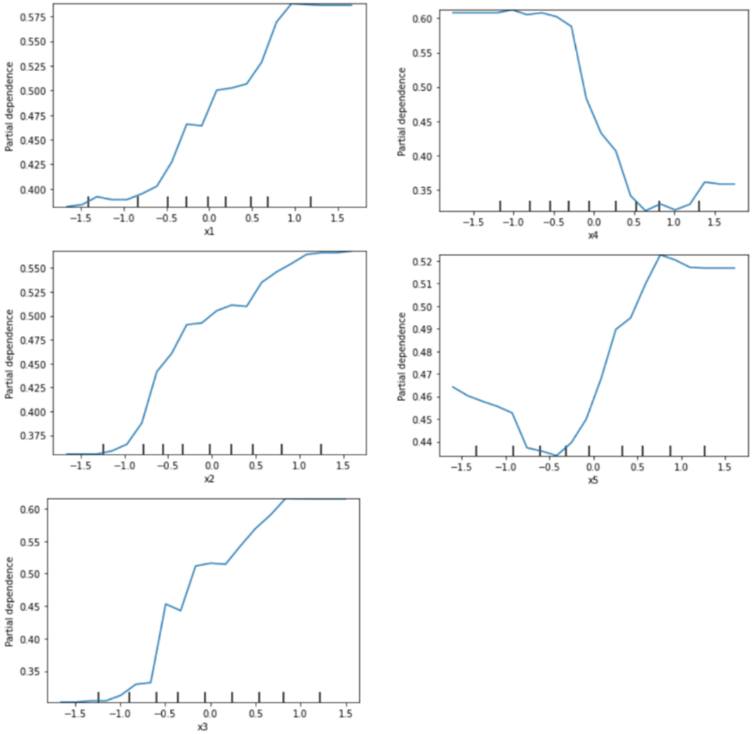
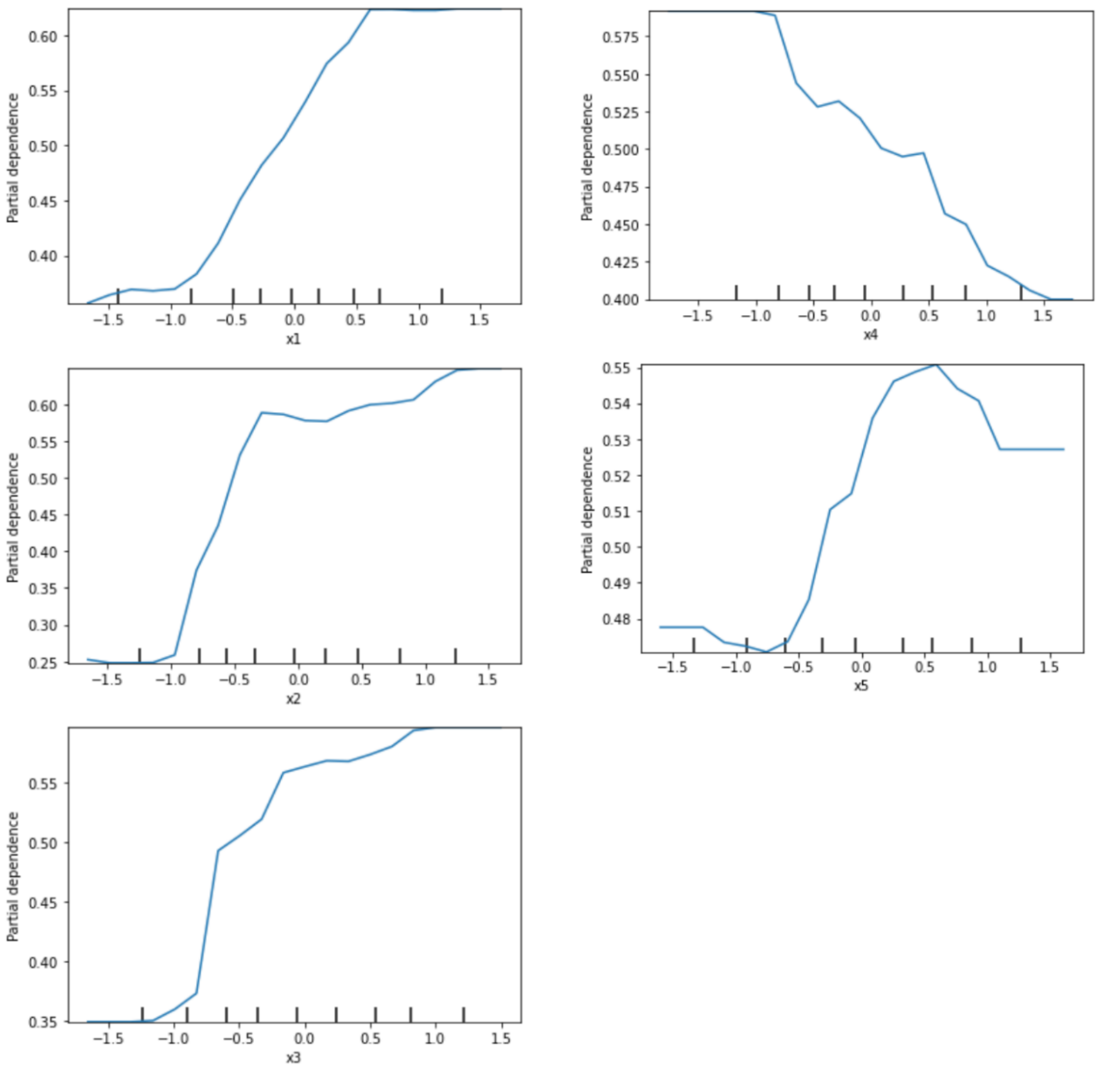


Figure 4e. PDPs for Informative Features, Dataset 5



Results from HIMPS, gRITs, and the H-Statistic. The performance of HIMPS, gRITs and the H-statistic on our three metrics (i.e., correctly detecting, ranking, and inferring the directionality) is given in Table 3 below, and we see that the H-statistic is the clear winner.

Table 3. Proportion of Datasets Satisfying Each Metric in Baseline Scenario

Metric	HIMPS	gRITs	H-statistic
Detection	1.0	.8	1.0
Ranking	N.A.	0	1.0
Directionality	1.0	.8	1.0

HIMPS detected both true interactions in all datasets whereas gRITs only recovered the true interactions for four out of five simulated datasets. Also, gRITs always ranked the true interactions incorrectly with x1x2 ranked higher than x3x4. Turning our attention to the H-statistic, it detected the true interactions (x3x4 and x1x2) and ranked these two interactions correctly across all five datasets. Hence, HIMPS and H-statistic outperforms gRITs in detecting both interactions, while it was a draw between HIMPS and the H-statistic. In terms of correctly ranking the true interactions, H-statistic outperform gRITs and HIMPS, as gRITS consistently ranked the true interactions incorrectly and HIMPS cannot rank interactions. For our third metric, which is correctly inferring the directionality of the interactions, HIMPS ties with H-statistic tie as correctly inferred the directionality of the true interactions between x3 and x4 (Figures 5a to 5e) and between x1 and x2 (Figures 6a to 6e). Readers who are interested in the details of the interactions detected and their rankings for each of HIMPS, gRITs, and the H-statistic can refer to Appendix 7, Tables A7 to A9 respectively.

Figure 5a. PDPs for x3 and x4, Dataset 1

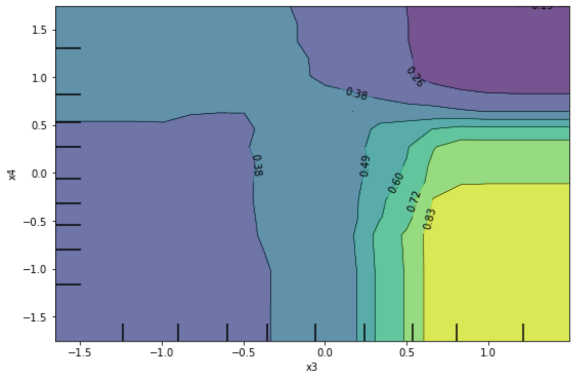


Figure 5b. PDPs for x3 and x4, Dataset 2

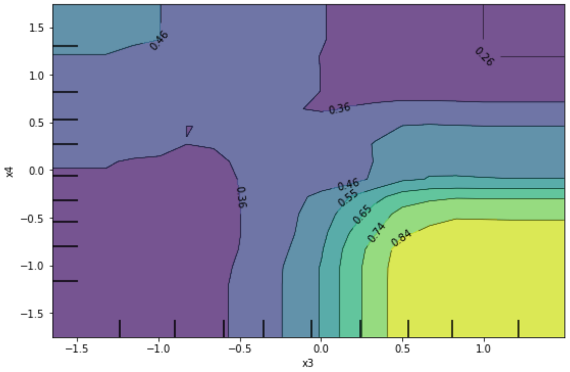


Figure 5c. PDPs for x3 and x4, Dataset 3

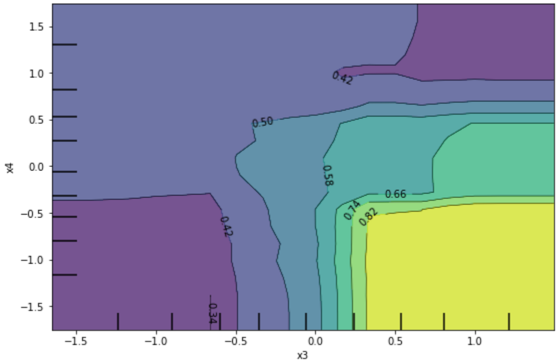


Figure 5d. PDPs for x3 and x4, Dataset 4

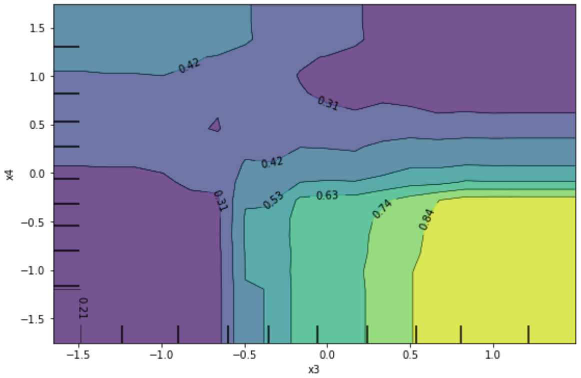


Figure 5e. PDPs for x3 and x4, Dataset 5

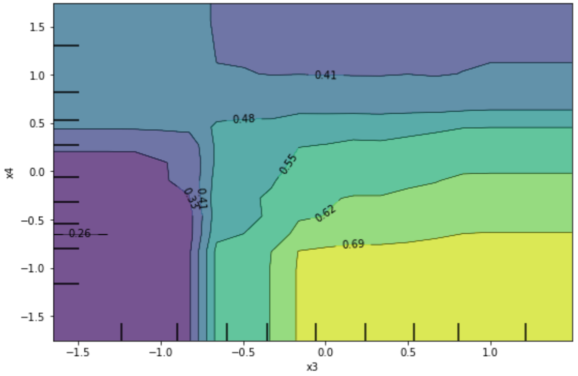


Figure 6a. PDP for x1 and x2, Dataset 1

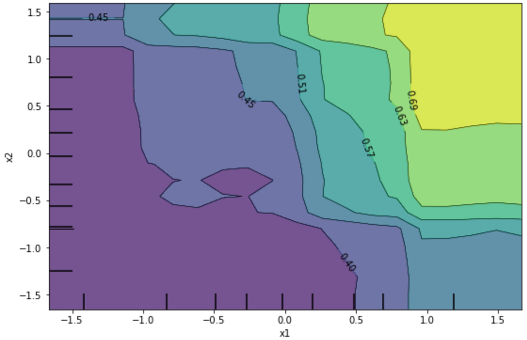


Figure 6b. PDP for x1 and x2, Dataset 2

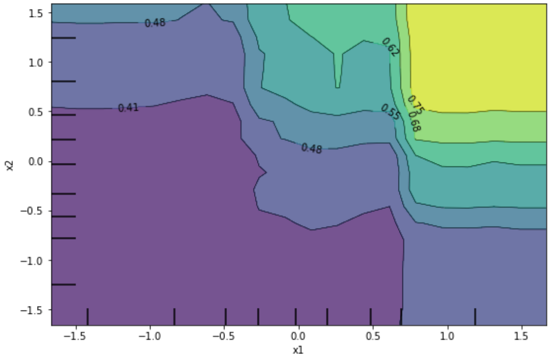


Figure 6c. PDP for x1 and x2, Dataset 3

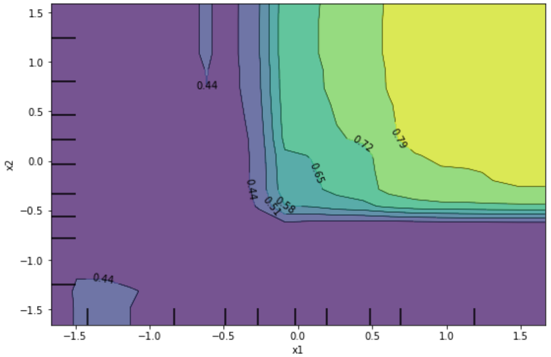


Figure 6d. PDP for x1 and x2, Dataset 4

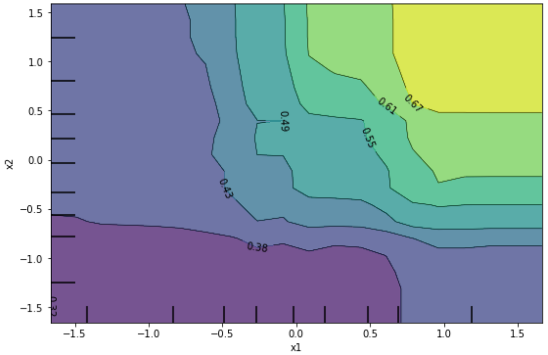
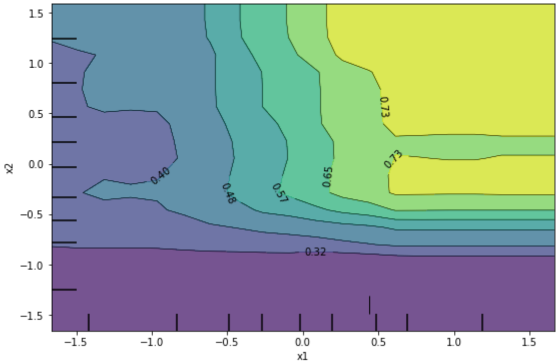


Figure 6e. PDP for x1 and x2, Dataset 5



Conclusions. With equally important features and uncorrelated interacting features, both HIMPs and H-statistic are better at detecting, ranking and correctly inferring the directionality of the true interactions than gRITs. However, the H-statistic triumphs over HIMPS as HIMPS is unable to rank interactions whereas the H-statistic ranks interactions.

3.2 Contingency scenario 1. A relatively more important non-interacting feature

In this study, we zoom in on the contingency that a relatively more important feature need not interact with other features by creating said feature in our simulated datasets. Clearly, HIMPS would fail in this setting because by construction this feature ranks the highest in terms of feature importance scores and has the highest chance of being falsely detected as an interaction. But how would gRITs and the H-statistic perform in this setting?

Simulation details. We retain the same distributional assumptions and correlation structure as per Study 1, and only change the coefficient on $\mathrm{x5}_{\mathrm{positive}}$ to 3 while retaining the other coefficients. Thus, $\mathrm{x5}_{\mathrm{positive}}$ is the relatively more important feature as the coefficient on $\mathrm{x5}_{\mathrm{positive}}$ (i.e., 3) is larger than the other features (i.e., 1) and our features have the same scale as they are drawn from an identical and independent N(0,1) distribution. Thereafter, we simulate five datasets and tune a RF on each dataset before applying HIMPS, gRITs and H-statistics to detect interactions. The details of the parameters used are summarized in Table 4.

Table 4. Summary of Parameter Settings for Contingency Scenario 1

Distributional assumptions, parameter values and correlation structure
1. x1 to x7 are distributed as standard normal variables
2. $\mathrm{x5}_{\mathrm{positive}} = 1$ if $\mathrm{x5} > 0$; 0 otherwise
3. $\mathrm{z} = .1 + \mathrm{x1} + \mathrm{x2} + \mathrm{x3} - \mathrm{x4} + 3\mathrm{x5}_{\mathrm{positive}} + \mathrm{x1x2} - 2\mathrm{x3x4}$
4. x1 to x7 are uncorrelated with each other

To apply HIMPS, we use the top four features in terms of their feature importance scores to obtain six two-way interactions. When applying the H-statistic, we limit our attention to the top five interactions with the highest values of the statistic. Finally, we only focus on interactions with stability scores >= .5 from the gRITs, following Basu et al. (2018).

Validation check. Analogous to study 1, we verify that the rankings of the feature importance scores and the PDPs of the features are consistent with our DGP. Specifically, we find that features x6 and x7 consistently have the lowest importance scores and x5 has the highest importance score across all simulated datasets. We also observe that the PDPs for each feature is consistent with the DGP, e.g., the PDP for x5 is a step function. We omit these the figures of these PDPs for brevity but they are available upon request.

Results from HIMPS, gRITs, and the H-statistic. HIMPS never detected both true interactions as it only detected x3x4 in three of the five datasets, and only detected x1x2 in the other two datasets. Most of the false positive interactions detected in four out of the five datasets include the relatively more important feature x5. gRITs consistently recover false interactions with the relatively more important feature x5 in all datasets. Only in one dataset is the true interaction x3x4 recovered, and it is ranked below four other false interactions. In contrast, the H-statistic recovers both true interactions in three datasets, and only one true interaction in the other two datasets. The ranking of the true interactions is also correct as x3x4 is ranked higher than x1x2. To sum up our findings, gRITs and HIMPS fared poorly in terms of correctly detecting both interactions as they failed to do so in all five datasets while the H-statistic detected true interactions in three datasets. Also, since gRITs and HIMPS never recovered both true interactions, it cannot rank and infer the directionality of the true interactions since ranking and directionality requires the interactions to be detected. Thus, the H-statistic is the clear winner. We summarize the results of each method's performance on the three key metrics in Table 5, which shows that the H-statistic is the clear winner. Detailed output for the interactions detected and their rankings is available for each method in Appendix 8, Tables A10 to A12. Figures for the PDPs of the true interactions (x3x4 and x1x2) are also available upon request.

Table 5. Proportion of Datasets Satisfying Each Metric in Contingency Scenario 1

Method	Detection	Ranking	Directionality
HIMPS	0	N.A.	N.A.
gRITs	0	0	N.A.
H-statistic	.6	.6	.6

Conclusions. The H-statistic is still effective in detecting, ranking and inferring the directionality of interactions even with a relatively more important feature that does not interact with other features, in three out of five datasets. gRITs and HIMPS performed abysmally as they detected false positive interactions involving the relatively more important feature. In empirical settings where a relatively more important feature is likely to exist, researchers should apply the H-statistic to lower the risk of detecting false-positive interactions involving said feature.

3.3 Contingency scenario 2. Correlated interacting features

This contingency scenario examines how correlated features which interact with each other affect the performance of HIMPS, gRITs and the H-statistic in detecting interactions as recovering interactions is difficult for any ML method when the interacting features are correlated with each other (Basu et al. 2018). Also, the H-statistic assumes independent features when calculating each feature's PD (Molnar 2019) which might affect its ability to detect interactions if the interacting features are correlated.

Simulation details. We retain the same distributional assumptions and coefficients on the main effects and interaction terms as per Study 1, but correlate x3 and x4 with a correlation of .50. To keep the mean of pr at around .50, we change the constant term to 1.

Thereafter, we generate five simulated datasets and tune our RF on these datasets before applying RITs and H-statistics to detect interactions. The details of the parameters used are summarized in Table 6. To apply HIMPS, we use the top four features in terms of their feature importance scores to obtain six two-way interactions. When applying the H-statistic, we limit our attention to the top five interactions with the highest values of the statistic. Finally, we only focus on interactions with stability scores >= .5 from the gRITs, following Basu et al. (2018).

Table 6. Summary of Parameter Settings for Contingency Scenario 2

Distributional assumptions, parameter values and correlation structure
1. x1 to x7 are distributed as standard normal variables
2. $\mathrm{x5}_{\text{positive}} = 1$ if $\mathrm{x5} > 0$; 0 otherwise
3. $\mathrm{z} = 1 + \mathrm{x1} + \mathrm{x2} + \mathrm{x3} - \mathrm{x4} + \mathrm{x5}_{\text{positive}} + \mathrm{x1x2} - 2\mathrm{x3x4}$
4. correlation between x3 and x4 is .50, all other features uncorrelated

Validation check. Similar to the previous simulation studies, we find that the ranking of the feature importances reflected the features which were included in z as the most important, followed by the features which were excluded from z as the least important. Finally, the shapes of the PDPs for x1 to x5 are consistent with their specification in z., we verify that the rankings of the feature importance scores and the PDPs of the features are consistent with our DGP. Specifically, we find that x4 has the highest importance score, and features x6 and x7 have the lowest importance scores. Additionally, the PDPs for each feature is consistent with the DGP and we observe that the PDP for x5 is a step function. Output for the validation check is available upon request.

Results from HIMPS, gRITs, and the H-statistic. HIMPS detects both true interactions in four out of five datasets but it also detects more false positives detected across all datasets than the gRITs and H-statistic. gRITs detected both true interactions in two datasets but only ranked the true interactions correctly in one dataset. The H-statistic detects both true interactions in four datasets and correctly ranks interactions in these four datasets. In terms of correctly inferring the directionality of the interactions, HIMPS ties with the H-statistic. The performance of each method on the three metrics is summarized in Table 7, from which we see that HIMPS and H-statistic outperforms gRITs for detecting the true interactions and correctly inferring the directionality of the true interactions, and the H-statistic dominates HIMPS as HIMPS is unable to rank the detected interactions. Hence, the H-statistic is the best performing method based on all three metrics. We provide the identities of the detected interactions and their rankings in Appendix 9, Tables A13 to A15 for HIMPS, gRITs and the H-statistic respectively.

Table 7. Proportion of Datasets Satisfying Each Metric in Contingency Scenario 2

Method	Detection	Ranking	Directionality
HIMPS	.8	N.A.	.8
gRITs	0	0	N.A.
H-statistic	.8	.8	.8

Conclusions. When there is a high correlation between the interacting features, both HIMPS and the H-statistic outperforms gRIT as both the H-statistic and HIMPS correctly detects and infers the directionality of both true interactions in more datasets than gRITs. H-statistic also dominates HIMPS as the H-statistic ranks the true interactions while HIMPS is unable to do so. Despite our initial concerns about the H-statistic's ability to detect both true interactions when interacting features are correlated, we find that the H-statistic is robust to this contingency as it still correctly detects, ranks, and infers the directionality in four out of five datasets.

3.4 Key findings from baseline scenario, and contingency scenarios 1 and 2

Overall, HIMPS and H-statistic outperforms gRITs for correctly detecting and inferring the directionality of both true interactions in the baseline scenario, and contingency scenarios 1 and 2. When there is a relatively more important feature (i.e., contingency scenario 1), the H-statistic outperforms HIMPS and gRITs as HIMPS and gRITs never detected both true interactions whereas the H-statistic statistic detected both true interactions in all but one dataset. Finally, the H-statistic trumps both HIMPS and gRITs in terms of correctly ranking the pre-specified interactions across all scenarios.

Ranking the detected interactions is arguably important as it allows scholars to focus on the more important interactions for robust theory building. Hence we recommend using the H-statistic over HIMPS due to availability of rankings in the H-statistic. In our subsequent empirical application, we adopt the H-statistic as there is a relatively more important feature in our empirical setting (i.e., contingency scenario 1 applies), and we cannot rule out a priori that interacting features are correlated with each other (i.e., contingency scenario 2 applies).

4. Empirical setting, variable operationalization, descriptives, and results

In this section, we first elaborate on the goal of our empirical study and the empirical setting before describing our variable operationalization. Thereafter, we provide descriptive statistics before applying our chosen approach (H-statistic) on the tuned RF to obtain substantive insights on the top five interactions between features.

4.1. Goal of study and empirical setting

The goal of our empirical study is to examine the interactions between features which predict a listing's clicks when consumers search for listings on property listing platforms. We posit that interactions are important because "marketers can better understand how to manipulate or arrange environments so that desired outcomes can be realized" (MacInnis 2001) and delimits the boundary conditions (Busse et al. 2015) where the relationship between a feature and a target applies. For instance, suppose that a listing's image distortion and image saturation interacts. We can determine the levels of distortion and saturation that yield the highest predicted click probability to guide owners and agents who are renting out their property. We focus on property listing platforms as they are heavily used by consumers who wish to buy, sell and rent their properties. For instance, industry reports indicate that the top three platforms in the six largest Southeast Asian economies collectively have 36 million site visits monthly for at least 2 million properties listed for sale (Umbelina 2019). In the US, the top three online property listing platforms averaged 77 million site visits monthly (Statista 2020). Thus, online search platforms are an important channel for property rentals and sales.

In this study, we obtained a unique online consumer browsing dataset from a large online property listing search platform in Southeast Asia, which recorded all actions taken by consumers during their browsing, and the image features and non-image features of all listings displayed. This dataset consists of a random sample of 6,568 consumers who began their search in the first week of August 2018, have at least two search sessions, and viewed at least two sets of impressions. We followed this sample until 31st December 2018 or until the date of their last search session, whichever date was earlier. The platform which gave us the dataset observed that nearly all consumers had a maximum search duration of 3 months, hence, the maximum amount of time with which we followed consumers was sufficient to capture their search activity. Consistent with this observation, the urban economics literature documented a median search duration of 12 weeks for housing (Genesove and Han 2012). Each visitor is assigned a unique consumer ID by Google analytics and each of the search sessions was assigned a session ID. Within each session, we observe the set of impressions shown to the consumer (i.e., impressions set). We also observed whether a listing that was exposed was clicked by the consumer, and the timestamps of clicks. In our dataset, we observe a total of 2.3 million listing exposures for private housing (i.e., condominiums) and public housing. In our empirical application, we focus on the condominium rental market as it displays more listings to consumers than public housing and property sales. Within the condominium rental market, we zoom in on the top 10 condominiums where we observe all listings belonging to each condominium. As the layout is constant for each condominium-bedroom-bathroom unit, we can control for the unit's objective quality by specifying condo-bedroom-bathroom dummies. These dummies which we term as condo-unit, control for all listing characteristics that do not vary with time, and all listings in each condo-unit thus have identical quality. Because non-image quality is held constant, we can study the effect of image features on clicks. In total, the 6,568 consumers engaged in 14,892 search sessions and were exposed to a total of 175,145 listings, where the number of unique listings equals 2,515. As the proportion of observations with clicks = 1 is very low (i.e., .02), the RF's ability to find informative features which predict observations with clicks = 1 will be adversely affected (Hasanin and Khoshgoftaar 2011, Japkowicz and Stephen 2002). Hence, we resampled this dataset using 1 to 1 resampling (King and Zeng 2001) which takes all observations with clicks = 1 and randomly samples an equal number of observations with clicks = 0. After resampling, our final dataset consists of 7,668 observations where our unit of observation is the consumer-listing level.

4.2. Variable operationalization

In this section, we outline how we use the data to operationalize our target, clicks, and features which are categorized into the listing's image features, listing's non-image features and consumer's search features.

Target. Our target is clicks (Click_{ijt}). This is a binary target with the value of 1 if consumer i clicked on listing j which was shown at time t; 0 otherwise. To avoid double counting, we only count clicks on unique listings for each set of impressions results displayed to the consumer. For instance, if a consumer clicked on the same listing twice, only the first click is counted. We operationalize our target using clicks as clicks reveal the attention allocated by the consumer (Matějka and McKay 2015) to the listing.

Features: listing's image features. We describe our operationalization of the seven image quality features used in our study. In choosing these features, we draw on extant literature in marketing and computer science. The first feature is distortion of listing j ($\mathrm{Distortion_j}$), which measures an image's loss of naturalness due to disturbances such as blur, watermarks etc., without requiring any reference image (Mittal et al. 2012). This feature calculates a measure of image quality using the "scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image." (Mittal et al. 2012). For each image, this evaluator calculates the empirical distribution of the mean subtracted contrast normalized ('MSCN') coefficient, and the empirical distributions of pairwise products of neighboring MSCN ('MSCN-pairwise') coefficients along four orientations of the image. Thereafter, the empirical distribution of MSCN coefficient and MSCN-pairwise coefficients are fitted to the generalized Gaussian distribution, and asymmetric generalized Gaussian distribution respectively to extract image attributes. These attributes are then mapped to a quality scores using support vector regression to calculate distortion. The score is bounded between [0,100] where a lower score indicates that the image is less distorted.

The second feature is the asymmetry of listing j ($\mathrm{Asymmetry_j}$) which is the extent by which an image lacks symmetry. Symmetry is the 'degree of visual balance' (Zhang et al. 2017) when an object is split into half along its y-axis. We operationalize asymmetry by transposing the image along its y-axis and calculating the information loss between the transposed and original images. If the image is symmetric, information loss is minimal after transposition. To calculate information loss, we represent each image as a sequence of 64-bit integers and calculate the number of disagreeing bits between these two sequences; this distance metric is known as the Hamming distance. This feature is bounded in the interval [0, 64] where 0 indicates no asymmetry and 64 indicates high asymmetry.

Our third, fourth, and fifth image features are derived from an image's hue, saturation, and value ('HSV'). We operationalized these features by calculating the mean value of each feature across pixels for each image where a pixel is an image's smallest unit area. We now explain how we define these three features. Hue (Hue_j) is the average color of an image across all its pixels for listing j and ranges from a value of 0 to 255, where for example, 0 to 40 represents the red color. Generally, consumers prefer warmer colors as these colors feel more welcoming (Zhang et al. 2017). Saturation (Sat_j) is the average intensity of an image's color (e.g., dull pink versus hot pink) across all its pixels, and is measured on a range of [0,255] where 0 indicates an extremely dull shade of the color and 255 indicates a fully intense shade of the color. Images with a higher saturation are more vivid and look more pleasing to the viewer (Zhang et al. 2017), which increases clicks on the listing. An alternative strand of research suggests that oversaturated images look artificial (Xu et al. 2015), which is less attractive and may lead to lower clicks on the listing. Value indicates the color's brightness where pure black has a value of 0 and pure white has a value of 255. Literature suggests that property listings with a higher level of brightness look bigger and roomier (Zhang et al. 2017, Li et al. 2019).

Features: listing's non-image features. Our first set of features comprises of three features of which the first two are listing j's rank (Rank_{ijt}) and price (Price_j) as consumers are more likely to click on listings with smaller ranks (Ghose et al. 2012, Ursu 2018) and on lower-priced listings (Ursu 2018). As consumers prefer a larger floor area (Sirmans et al. 2005), we included the listing's gross floor area (Area_j) in square feet as our third feature.

Second, we consider a vector of dummy variables indicating the unique condo-unit combination ($\mathrm{Condo\_unit}_{\mathrm{j}}$) for each listing j, where we specify the dummies using a condo-bedroom-bathroom triplet. As an example, suppose a condominium named ‘Oasis’ has 3 different unit types which are (i) 1 bedroom and 1 bathroom, (ii) 2 bedrooms and 1 bathroom, and (iii) 3 bedrooms and 2 bathrooms and a second condominium ‘Allure’ has 2 different unit types which are (i) 1 bedroom and 1 bathroom, (ii) 3 bedrooms and 3 bathrooms. This gives us 5 unique condo-unit dummies. In our data, each condominium has on average 7 unique unit types and we identify a total of 73 condo-units, and operationalize as 73 dummy features.

Our final set of two non-image features relate to the impressions page which displays said listing. The first feature is the impressions page number of the displayed listing ($\mathrm{Page}_{\mathrm{ijt}}$) as consumers incur search costs when scrolling to subsequent pages (Ghose et al. 2012, Ursu 2018). Our second feature is the number of identical condo-units shown on an impressions page to each consumer i for each listing j at time t ($\mathrm{Condo\_count}_{\mathrm{ijt}}$) as a greater number of identical condo-units increases the odds of a listing being clicked due to the exposure effect.

Features: consumer’s search features. The first feature we consider is consumer i’s cumulative search time ($\mathrm{Cum\_search\_time}_{\mathrm{it}}$) which is the total amount of time spent searching on the platform up to time t, in hours. With more time spent on search, consumers gain more information about a listing which increases their probability of clicks on the listing. Besides measuring consumer i’s search time across sessions, we also measure the amount of time spent searching (in hours) in a session and denote this variable as session time ($\mathrm{Session\_time}_{\mathrm{it}}$).

Our third feature is session count ($\mathrm{Session\_count}_{\mathrm{it}}$) which measures the cumulative session count of consumer i at time t. We include this variable as the cumulative session count proxies for the discrete information stock acquired by consumers during their search.

Turning our attention to the fourth feature, this feature is time of the day ($\mathrm{Time}_{\mathrm{it}}$) when a search was carried out by consumer i for the listings displayed at time t, coded using 24 hourly dummies. We include this feature as consumers focus on their mobile devices as a coping mechanism to deal with their loss of privacy in crowded places (Andrews et al. 2018), which increases the probability of clicks on listings displayed during peak hours, such as the daily commutes to and from workplaces.

Next, our fifth feature is the vector of dummies for the specific day of the week ($\mathrm{Day}_{\mathrm{it}}$) when consumer i carried out a search on the platform at time t. Consumers have lower search costs on weekends (Warner and Barsky 1995), and hence, clickthrough rates are higher for online advertisements on weekends than on weekdays (Narayanan and Kalyanam 2015).

Our penultimate feature is the vector of dummies indicating the device used by the consumer when visiting the online search platform (Device). There are three elements in this vector as consumers can use mobile phones, tablets, and desktops to access the platform. Since consumers face higher search costs on smaller devices (Ghose et al. 2012), their probability of clicks on a listing is lower for mobile devices such as mobile phones and tablets than desktops.

Finally, we have the vector of dummies indicating the channel by which the consumer arrives at the property listing platform's website ($\mathrm{Channel}_{\mathrm{it}}$) with 11 entries capturing all channels. These 11 channels can be categorized into six broad groups, i.e., arrived directly on search platform, paid display advertising, referrals, search engine, social media, and electronic customer relationship management ('e-CRM'). We include channel in our study as consumers self-select into the channel (e.g., display advertising, organic search, referrals) based on the perceived benefits versus costs of using said channel (Li and Kannan 2014), and each consumer’s propensity to click on a listing might depend on the channel from which they come from.

We provide a summary table of all variables used in this study, and their operationalization in Table A16 of Appendix 10 for ease of reference.

4.3. Descriptive statistics and correlation matrix

We present the descriptive statistics and correlation matrix of our features in Tables 8 and 9 respectively. As we have 131 features in our dataset, we only include the non-dummy features to avoid clutter.

Descriptive statistics. The average number of clicks is .50 as our resampled data has an equal number of observations with click = 1 and with click = 0. The distributions of distortion and asymmetry are relatively symmetric as indicated by the similar mean and median values for each feature. The average rank of a listing that is exposed to consumers is 21.28 and the average page exposed to consumers is the 2nd page of the impressions results. Because the median rank of a listing is 15 and the median page is 1, these two statistics taken together implies that 50% of listing exposures were on the 1st impressions page, or in the top-20 ranks. This is unsurprising as the platform displays twenty listings on each impressions page as a default setting. With regards to the distribution for price and floor area, we find the distributions are slightly right skewed as the mean exceeds the median for these features. Finally, the average and median cumulative search time is 31.67 hours and 13.96 hours respectively. This trend is consistent with the urban economics and real estate literature which documents a longer average search duration than the median search duration (Zumpano et al. 2003).

Table 8. Descriptive Statistics for all Features and Target

	Descriptive Statistics				
Variables	Mean	Min	Median	Max	Standard Deviation
1. Click	.50	.00	.50	1.00	.50
2. Distortion	46.83	.00	45.91	100.00	11.41
3. Asymmetry	24.62	.00	24.00	64.00	11.18
4. Hue	50.96	.00	45.49	128.25	21.90
5. Saturation	57.75	.00	51.23	191.18	30.40
6. Value	147.90	27.10	148.55	247.14	29.50
7. Rank	21.28	1.00	15.00	100.00	21.11
8. Price	3561.95	700.000	3200.00	17000.00	1583.09
9. Page	1.55	1.00	1.00	5.00	1.00
10. Condo count	4.83	1.00	3.00	27.00	4.87
11. Floor area	814.65	100.00	730.00	4252.00	380.93
12. Cum search time (hours)	31.67	.00	13.96	212.72	41.59
13. Session search time (hours)	.45	.00	.35	3.67	.39
14. Session count	7.74	1.00	3.00	155.00	13.86

Note: We suppress channel dummies, condo unit dummies, time of the day dummies and day dummies for brevity as there are 11 channels, 73 condo unit dummies, 24 time dummies, and 7 day dummies

Correlation table. From the correlation matrix, it is immediately apparent how weak each of the pairwise correlations are between the target (i.e., clicks) and the features. For instance, the strongest correlation (in absolute terms) is only .05 between clicks and condo-count. The weak correlation suggests that a linear relationship between clicks and features may not be an accurate portrayal of their true underlying relationships. The correlations between features are generally weak, except for the correlation between (i) price and floor area (.89) and between (ii) page and rank (-.95). (i) is intuitive as listings with a larger floor are more costly, and (ii) is consistent with the default platform setting of 20 listings displayed on each impressions page, and hence ranks with a larger number appear on the later pages.

Table 9. Correlation Matrix for Target and Features.

Target and features	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Click	1													
2. Distortion	-.03*	1												
3. Asymmetry	.02	.03*	1											
4. Hue	-.01	.12**	-.11**	1										
5. Saturation	-.05**	.05**	-.02	.19**	1									
6. Value	-.04**	.11**	-.05**	-.07**	-.12**	1								
7. Rank	-.02**	-.01	.00	-.02	-.02	-.04**	1							
8. Price	-.06**	-.07**	-.13**	-.03*	.02	.01	-.05**	1						
9. Page	.04**	-.01	.00	-.02	-.02	-.04**	.95**	-.05**	1					
10. Condo unit count	.15**	-.03**	-.04**	-.02*	-.03**	-.02	-.12**	.01	-.13**	1				
11. Floor area	-.05**	-.06**	-.12**	-.00	.02	-.02	-.02*	.89**	-.03**	-.00	1			
12. Cum search time	.08**	-.00	-.00	-.01	-.02*	-.03**	.11*	-.00	.11**	.22**	.02	1		
13. Session duration	.07**	.01	.01	.01	-.02	-.00	.04*	-.06**	.04**	-.01	-.05**	.56**	1	
14. Session count	.00	-.02	-.03**	.01	-.02*	-.02	.02*	.04**	.03*	.05**	.07**	.11**	.15**	1

Note: * and ** denotes significance based on *p* < .05 and *p* < .01 respectively

4.4. RF tuning, HIMPS and H-statistic results

To detect interactions between features, we first need to tune a RF. Hence, we first present our results for the tuned RF's parameters where we tune the RF's key tuning parameters following Choudhury et al. (2021). Conducting a sanity check on our results, we verify that the tuned RF recovers relationships between feature and targets, which are documented in extant literature. Thereafter, we illustrate the stark contrast between the interactions detected by HIMPS and H-statistic where none of the interactions detected in H-statistic were picked up in HIMPS. Finally, we visualize the effects of these two-way interactions on predicted clicks using PDPs.

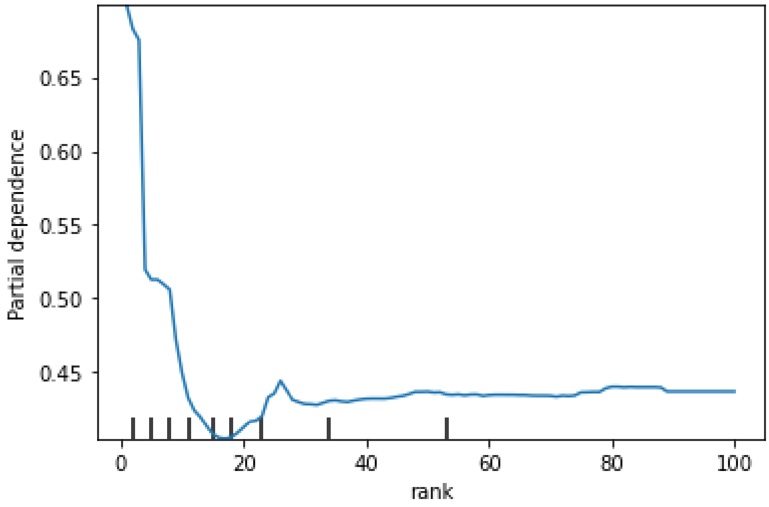
Results of our tuned RF. The finalized results from our RF tuning parameters are (i) minimum samples in leaf = 138, (ii) minimum decrease in impurity before a split occurs = 1.23x10^{-4}, (iii) max number of features = None, (iv) maximum depth = 10, (v) criterion for splitting at each node = 'entropy'. Amongst these values, we note that the large and small values for minimum samples and maximum depth respectively ensure that the RF is not overfit to idiosyncratic noise in the training dataset (Hastie et al. 2009). Fitting the tuned RF to the test dataset, we obtain the top ten features based on their importance scores which is displayed in Table 10. From this table, we see two features (rank and price) with well-documented main effects on click probability in the literature.

Table 10. Top Ten Features and Importance Scores from the Tuned RF

Rank	Feature	Feature importance scores
1	Rank	.13
2	Cumulative search time	.12
3	Session time	.10
4	Price	.08
5	Condo unit count	.08
6	Saturation	.06
7	Value	.06
8	Floor area	.06
9	Distortion	.06
10	Hue	.06

We conduct a sanity check on our fitted RF by visualizing the PDPs for rank and price. Figure 7 displays the PDP for rank which indicates a negative association between rank and average predicted click probability for ranks up to 20, a positive association from ranks 20 to 30, before resuming the negative association thereafter.

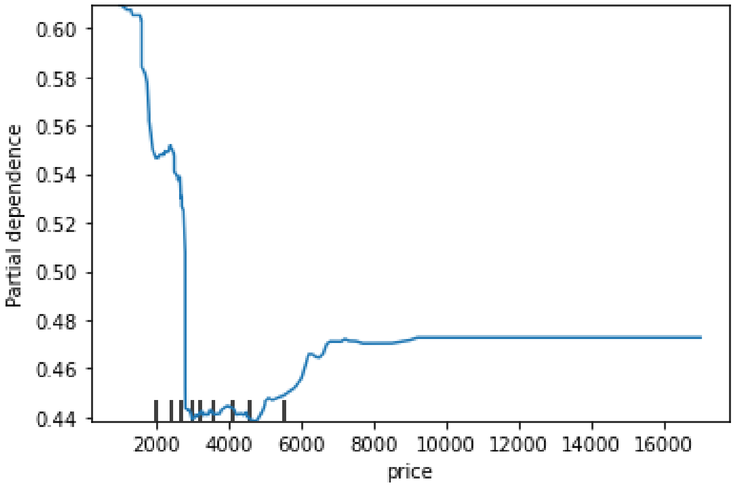
Figure 7. PDP for Rank



The negative association in Figure 7 is consistent with extant literature's results that rank is inversely related to click probability (e.g., Ghose et al. 2012, Ursu 2018). We reconcile the positive association by noting that listings are displayed in a set of twenty listings per page as a default option and hence, listings from 20 to 30 are the top few listings on the second page which attracts more clicks.

Viewing the PDP for price in Figure 8, we see a negative trend between price and average predicted click probability up to 3,000 dollars, and a positive trend between 4,500 dollars to 7,000 dollars before staying relative constant thereafter. The negative trend is consistent with the notion that listings with higher prices are associated with lower click probabilities (e.g., Ursu 2018) while the positive trend is rationalized if we consider that higher prices signal a higher quality (e.g., Zhao 2020) of the listing. Overall, we conclude that the PDPs for rank and price are consistent with the relationships in extant literature.

Figure 8. PDP for Price



Interactions from the H-statistic. Table 11 documents the top five interactions detected by the H-statistic of which four interactions included image features. If instead, we had used HIMPS and constructed pairwise interactions from the top four features, we would have missed out on interactions involving image features since none of the top four features by importance scores is an image feature.

Table 11. Top Five Interactions (scores) from H-statistic

Rank	H-statistic
1	Price * cumulative search time (.36)
2	Distortion * hue (.35)
3	Distortion * value (.29)
4	Distortion * saturation (.25)
5	Hue * saturation (.22)

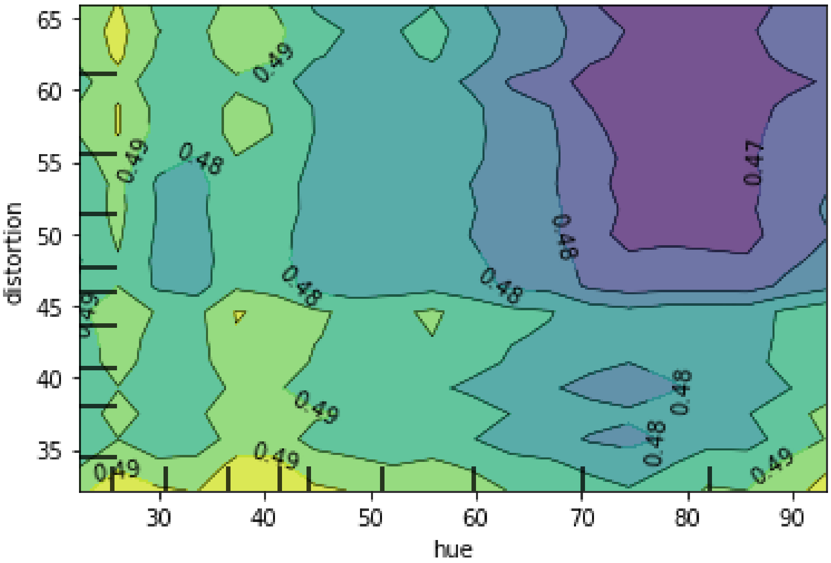
Visualizing top five interactions with image features from the H-statistic. We visualize the PDP for each interaction with an image feature in descending order of their H-statistic, and provide image examples to illustrate these interactions. To aid understanding, we first state the implications of these interactions on predicted click probabilities in Table 12, which shows the combinations of feature values yielding a higher average predicted click probability than other possible combinations of feature values.

Table 12. Managerial implications from the interactions with an image feature in H-statistic

Feature	Feature	Implication
Distortion	Hue	Set images to (i) low distortion and low hue or, (ii) high distortion and low hue
Distortion	Saturation	Set images to (i) low distortion and low saturation or, (ii) high distortion and low saturation
Distortion	Value	Set images to (i) low distortion and low value or, (ii) high distortion and low value
Hue	Saturation	Set images to (i) low hue and low saturation or, (ii) high hue and low saturation

Figure 9 shows the first PDP for the interaction between distortion and hue. In this figure, we find a lower average predicted click probability for images with (i) higher values of distortion and hue as compared to images with (ii) lower values of distortion and hue, and images with (iiii) lower values of hue but higher values of distortion.

Figure 9. PDP for Distortion and Hue



We illustrate (i) to (iii) using the requisite images in Figures 10, 11 and 12 respectively. Why does (i) occur? One potential explanation is that a distorted image is already less attractive to consumers and attracts less attention. With high hues (cold colors), the image’s ability to attract attention is further dampened. Hence, average predicted click probability is the lowest for images with high distortion and high hue. To explain (ii), first note when hue is small (i.e., warm colors), images with low distortion which are more attractive tend to grab more attention due to the warm colors. With regards to (iii), a combination of small hue and large distortion possibly allows the attention-grabbing effects of hue to overcome the unattractiveness of distortion, which leads to a higher average predicted click probability than a combination of large hue and large distortion.

Figure 10. Image with High Distortion (i.e., 69.15) and High Hue (i.e., 75.80)

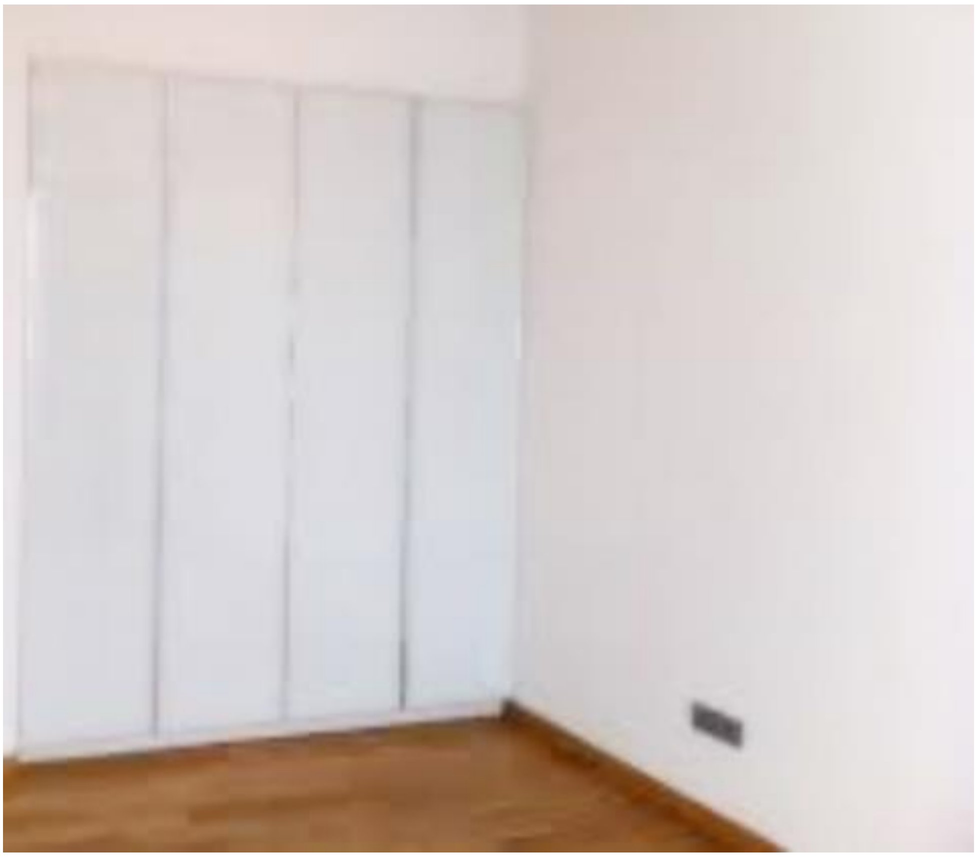


Figure 11. Image with High Distortion (i.e., 66.78) and Low Hue (i.e., 24.30)

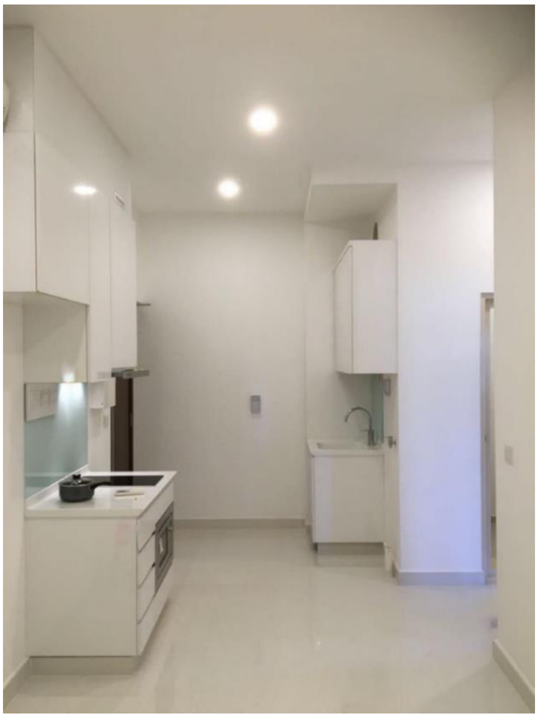
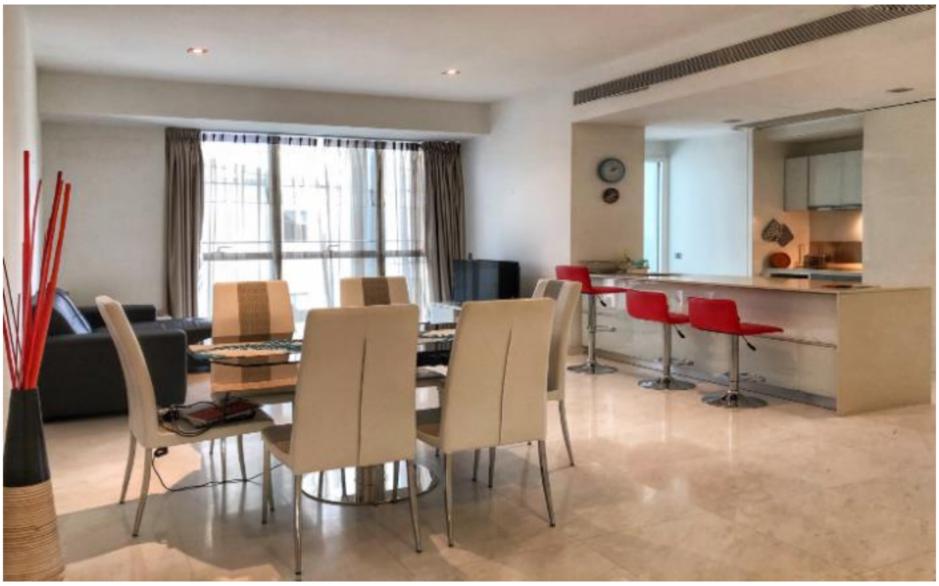
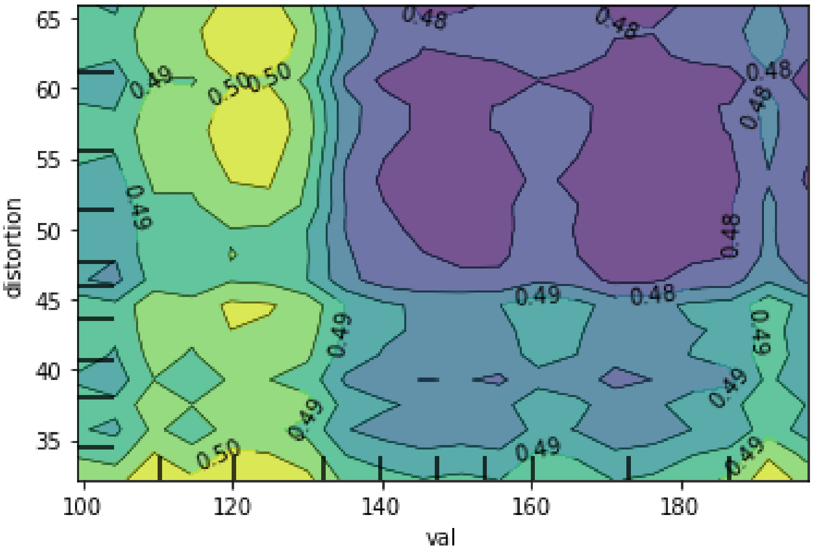


Figure 12. Image with Low Distortion (i.e., 27.72) and Low Hue (i.e., 26.61)



Our second PDP in Figure 13 visualizes the interaction between distortion and value. From this figure, we see that (i) higher distortion and value are associated with a lower average predicted click probability whereas both (ii) higher distortion and lower value and (ii) lower value and lower distortion is associated with a higher average predicted click probability.

Figure 13. PDP for Distortion and Value



Figures 14 to 16 illustrate a typical image with the appropriate feature values for distortion and saturation, corresponding to situations (i) to (iii) respectively. From Figure 14, we see that high distortion and high value makes the image even more unattractive as a bright image (i.e., high value) makes the distortion even more salient and obvious to the viewer. In comparison, an image with either (i) high distortion and low value (Figure 15) or, (ii) low distortion and low value (Figure 16) is less attention-grabbing as low value indicates that the image is less bright, which makes the distortion less salient to the viewer.

Figure 14. Image with High Value (i.e., 184.19) and High Distortion (i.e., 55.37)

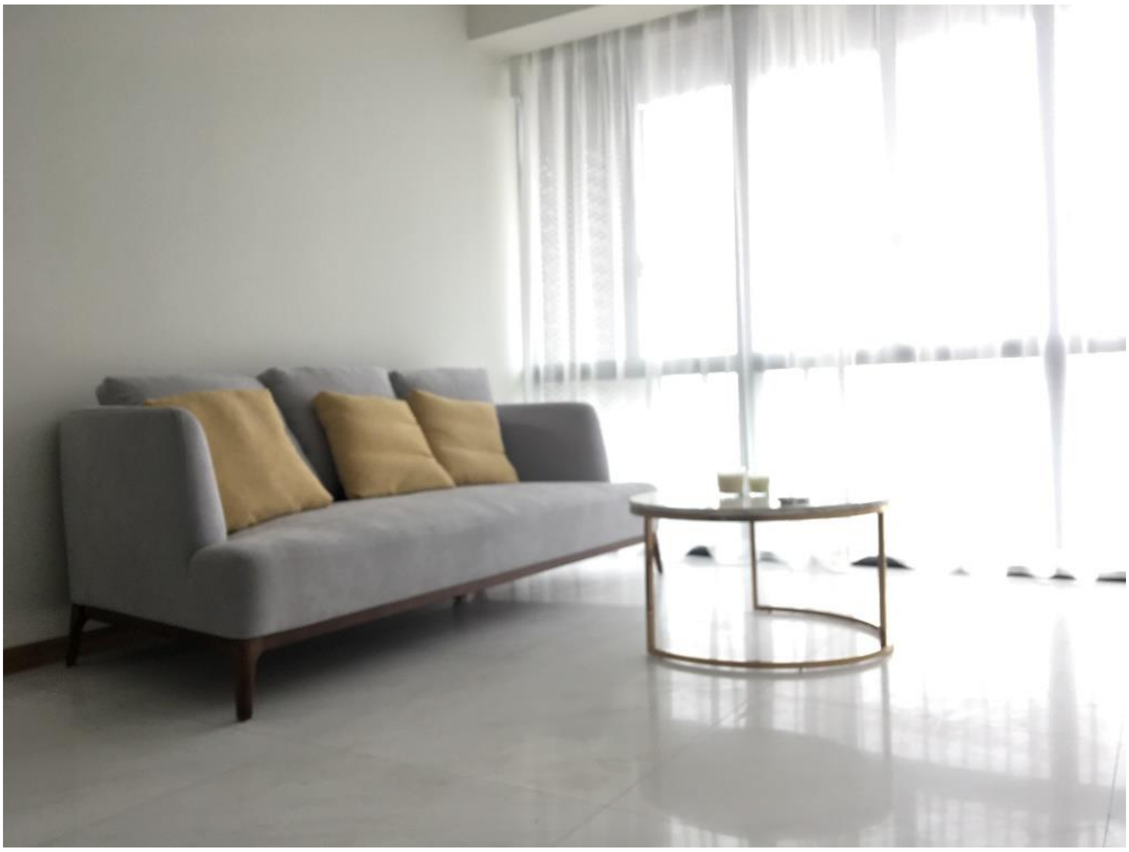


Figure 15. Image with Low Value (i.e., 127.60) and High Distortion (i.e., 53.63)

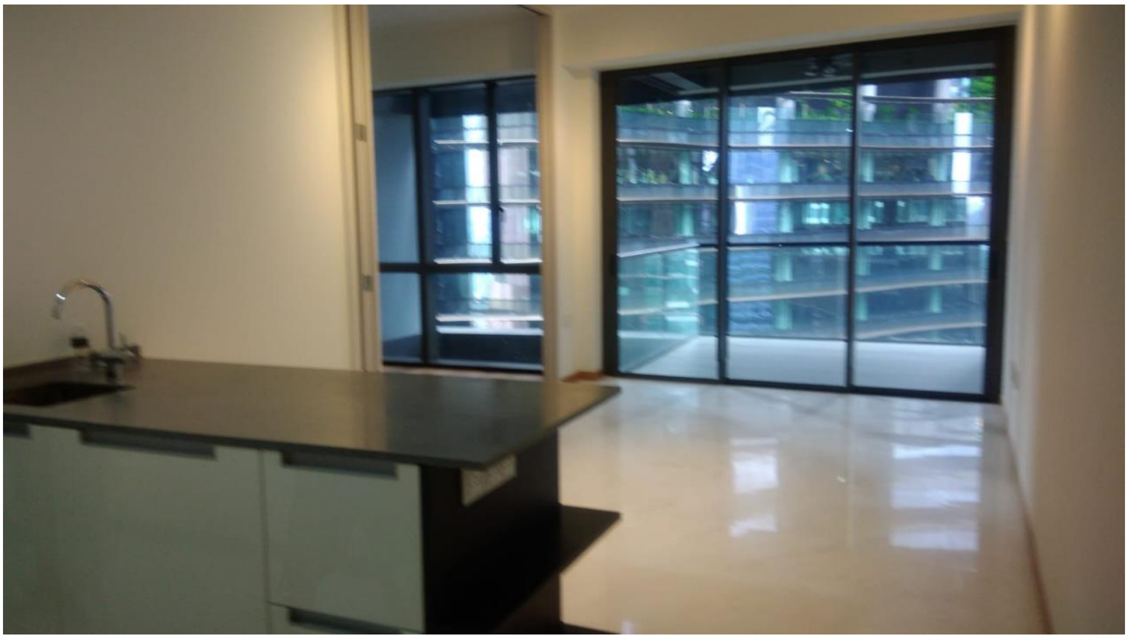
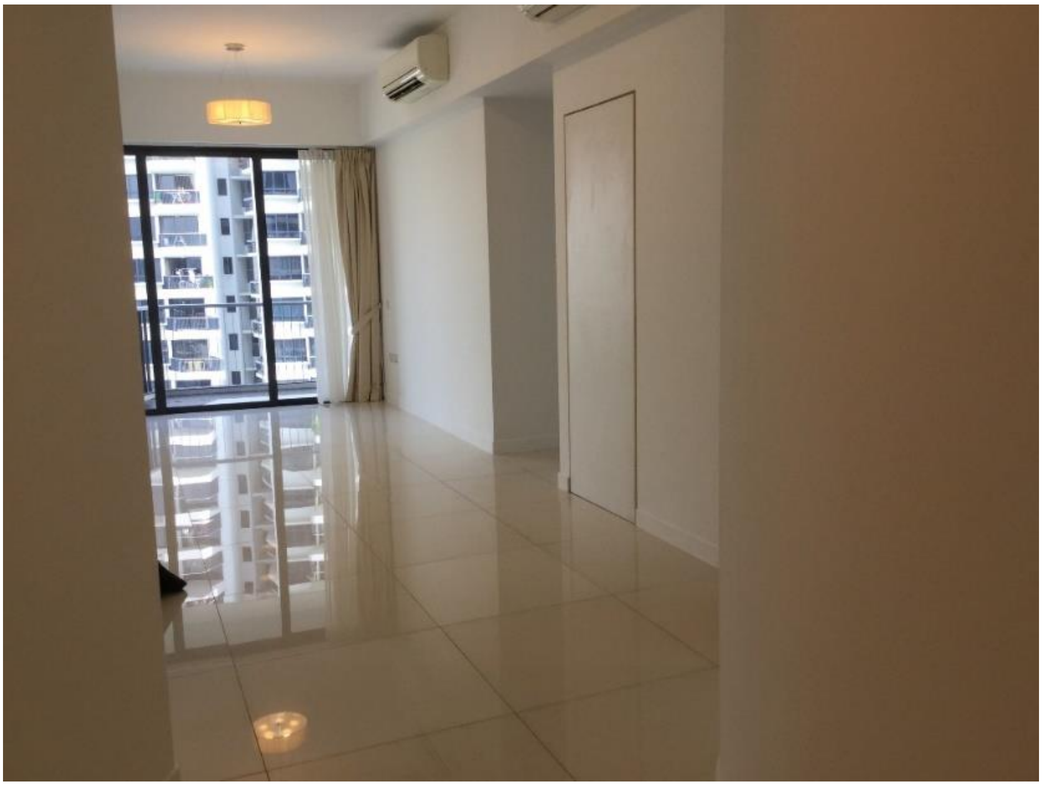
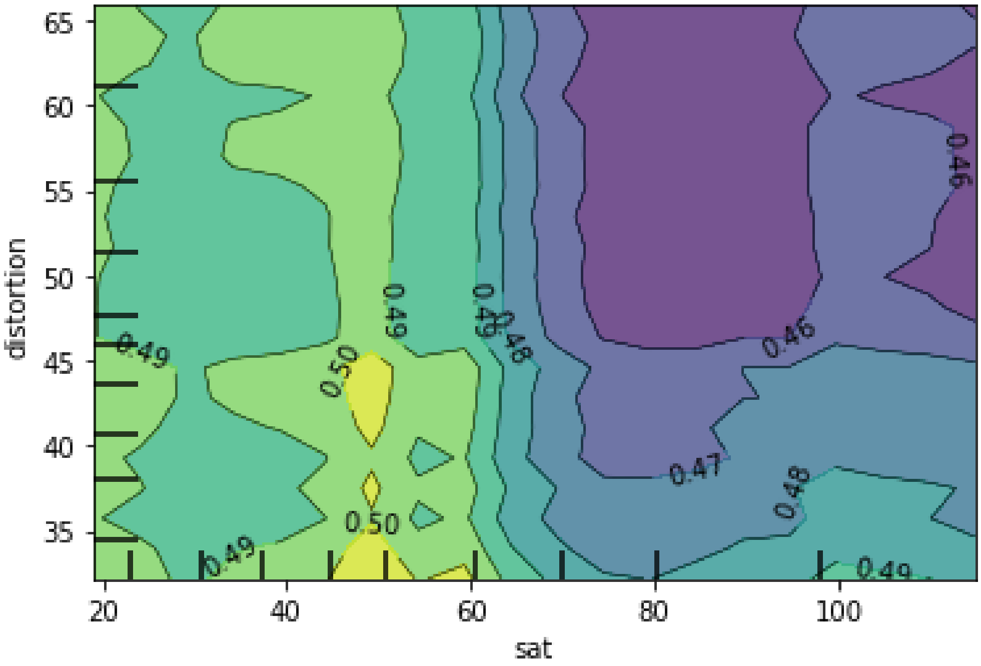


Figure 16. Image with Low Value (i.e., 129.68) and Low Distortion (i.e., 33.97)



Our penultimate PDP visualizes the relationship between distortion and saturation on average predicted click probability in Figure 17. Examining the PDP, we see that (i) higher values of distortion and saturation are associated with a lower average predicted click probability whereas both (ii) higher distortion and lower saturation and (ii) lower distortion and lower saturation are associated with a higher average predicted click probability.

Figure 17. PDP for Distortion and Saturation



Figures 18 to 20 illustrate a typical image with the requisite feature values for distortion and saturation, corresponding to situations (i) to (iii) respectively. Since saturation indicates the fullness of the colors, a simultaneously saturated and distorted image further highlights its distortion and reduces its attractiveness to consumers by an even greater extent.

Figure 18. Image with High Saturation (i.e., 79.06) and High Distortion (i.e., 53.17)

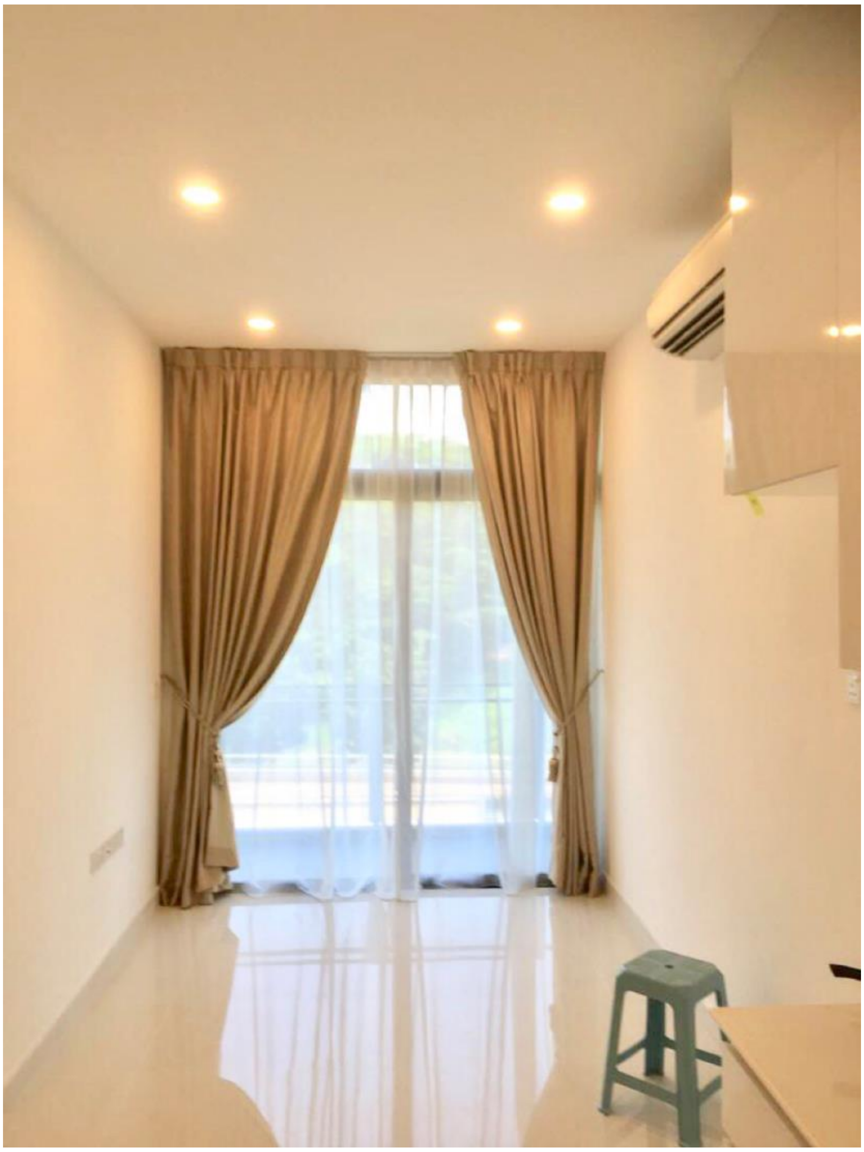


Figure 19. Image with Low Saturation (i.e., 5.33) and High Distortion (i.e., 61.83)

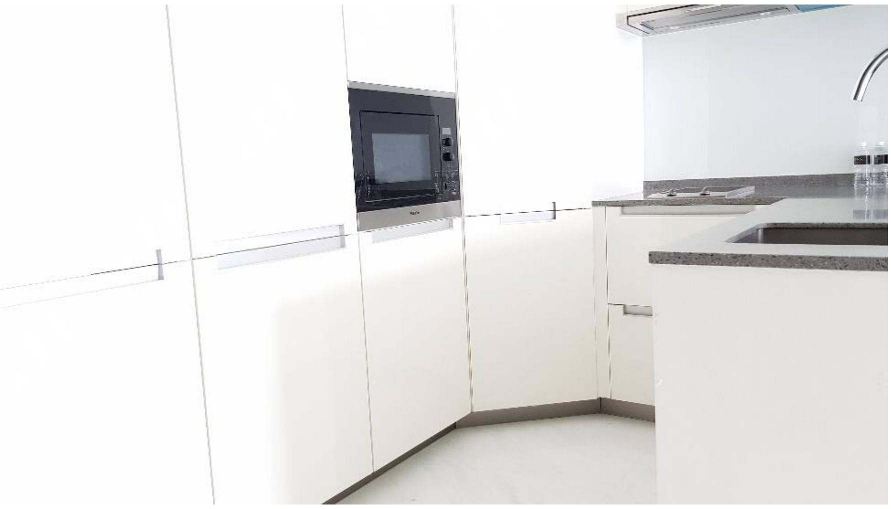
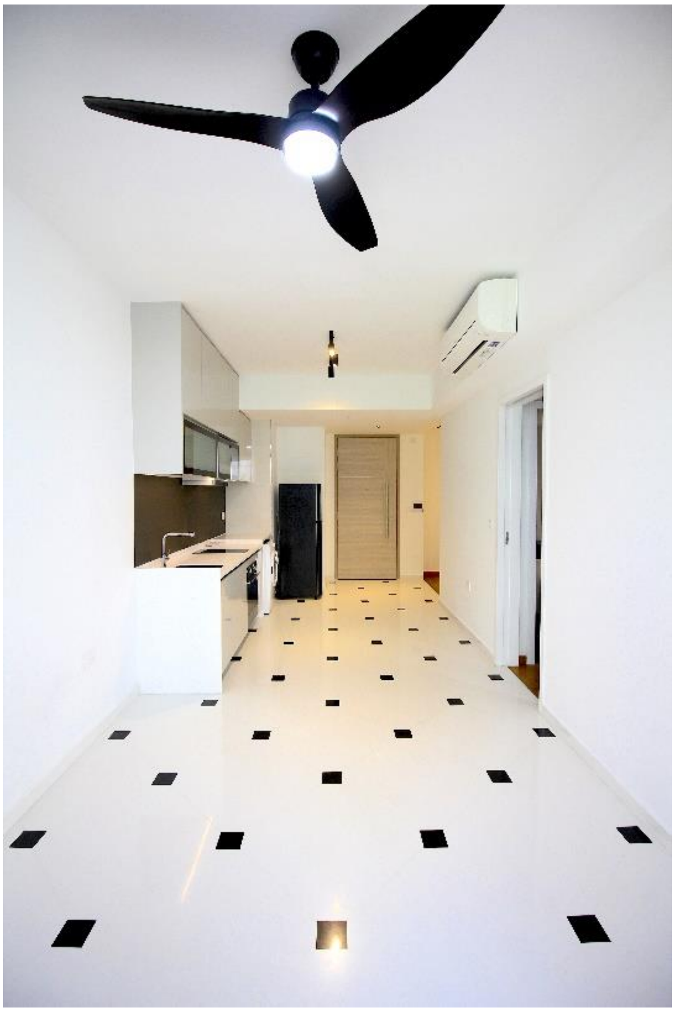
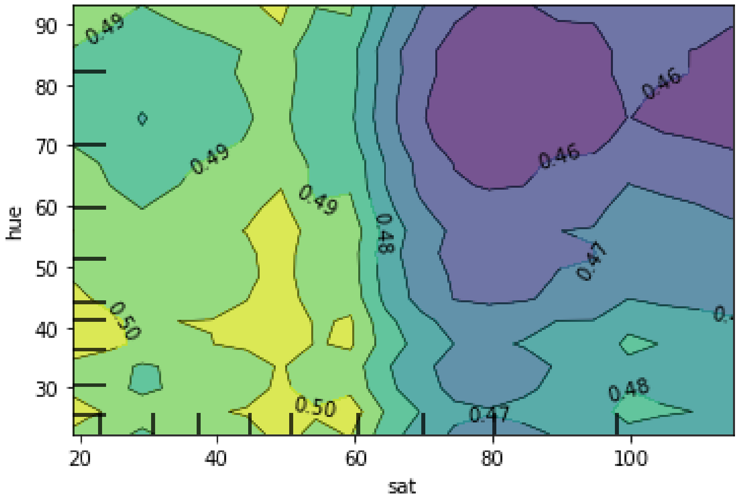


Figure 20. Image with Low Saturation (i.e., 20.99) and Low Distortion (i.e., 27.49)



Our final PDP examines the interaction between hue and saturation on average predicted click probability which is depicted in Figure 21. From Figure 21, we see that (i) higher values of hue and saturation are associated with a lower average predicted click probability whereas (ii) lower values of hue and saturation are associated with a higher average predicted click probability.

Figure 21. PDP for Hue and Saturation



We provide an image with feature values satisfying (i) and (ii) in Figures 22 and 23 respectively. As an image with cold colors is less attractive than an image with warm colors (Zhang et al. 2017), this image is rendered even less attractive when it is saturated as it makes the cold colors even more vivid. In comparison, when an image has a low hue (warm colors), the low saturation is less overpowering for the viewer as it is less vivid and less intense

Figure 22. Image with High Hue (76.52) and High Saturation (75.24)

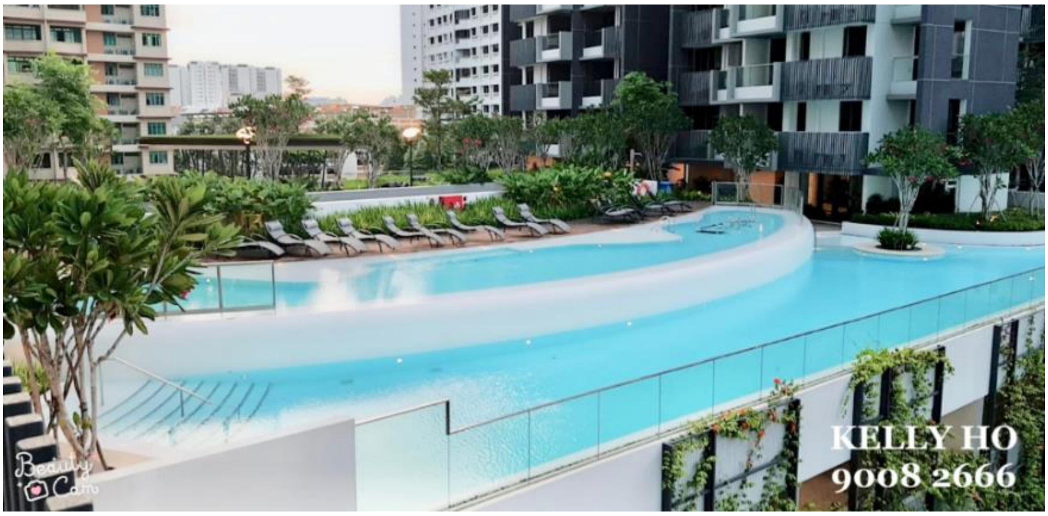
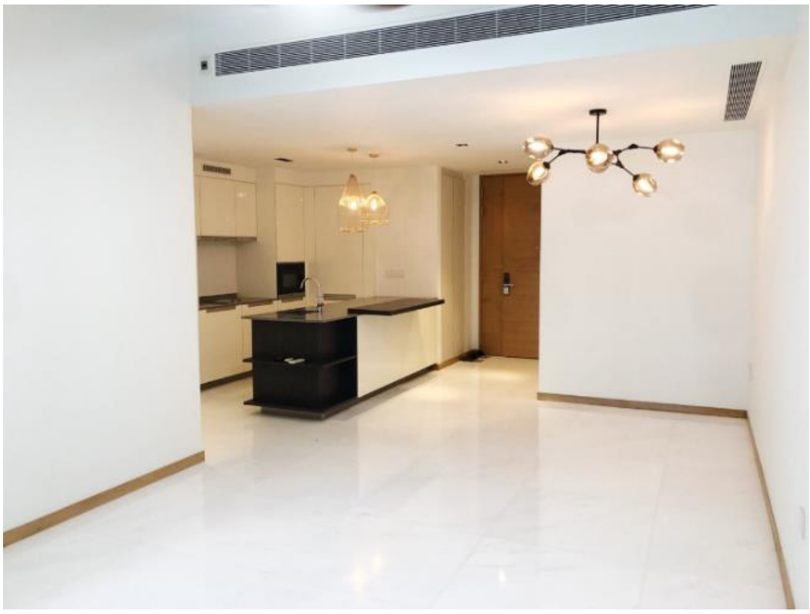


Figure 23. Image with Low Hue (76.52) and Low Saturation (75.24)



5. Discussion

In this section, we first briefly summarize our key findings. Next, we highlight the implications of our results for scholars and practitioners. We conclude with limitations and future research directions.

5.1. Summary of key findings

We compared the performance of HIMPS, gRITS and H-statistic using a series of simulation studies to ascertain their effectiveness in correctly detecting, ranking and inferring the directionality of the interactions. Overall, we find that the H-statistic is the most effective method across three scenarios. The H-statistic’s dominance over both HIMPS and gRITS when there is a relatively more important feature in the data that does not interact with other features, as both HIMPS and gRITS detects false interactions involving this feature. With correlated interacting features, the H-statistic ties with HIMPS for correctly detecting and inferring directionality of interactions although the H-statistic outperforms HIMPS in correctly ranking interactions as HIMPS cannot rank interactions.

Applying the H-statistic to consumer search data obtained from a large online property listing platform, we find that four of the top five interactions detected by the H-statistic involves image features and, none of the interactions detected by HIMPS involved image features. All of the top five interactions detected by the H-statistic exclude the relatively important feature, rank, which is detected in the top three interactions in HIMPS. Consistently, we find interaction effects between distortion and each of hue, value, and saturation, and between hue and saturation. For instance, an image with high distortion and high saturation has a lower average predicted click probability than an image with either (i) low distortion and low saturation or (ii) high distortion and low saturation.

5.2. Implications

Our study suggests three important implications for scholars and two implications for agents and managers. First, in contrast to HIMPS which detects but cannot rank interaction candidates, we show that the H-statistic is a superior method which not only detects but also ranks these interaction candidates across three scenarios, which are (i) equally important and uncorrelated interacting features, (ii) a relatively more important and non-interacting feature, (iii) correlated interacting features. As interactions facilitate theory building (MacInnis 2001), a desirable ML method should rank the interactions and allow users to focus on the most important interactions for theory building. As a robustness check, we also redid the simulations using an ad-hoc measure to rank interactions in HIMPS by summing the interacting features' importance scores. Even after including this ad-hoc measure, we find the H-statistic correctly ranks interactions at least as well as HIMPS, if not better (refer to Appendix 11 Tables A17 to A19). Hence, the H-statistic is a truly robust ML method for detecting, ranking and inferring the directionality of interactions and we recommend that scholars employ the H-statistic.

Second, the presence of interacting features allows us to explore important boundary conditions of the relationship between a listing's image features and the listing's average predicted click probability, thus refining the relationship between a listing's image features and its consumer outcomes. For instance, listings with highly saturated images are more vivid and capture the consumer's attention which consequently leads to higher demand (Zhang et al .2017). However, an important boundary condition is the image's distortion. With high distortion, the higher saturation makes the distortion even more vivid and further lowers the attractiveness of said image, thus lowering a listing's predicted click probability even more as we see from the PDP in Figure 17.

Third, when applying a ML method the user must recognize how the method defines interactions to ensure that the detected interaction is consistent with what the user is looking for. For instance, interactions in RITs are subsets "of all predictor variables that occur more often for observations in a class of interest than for other observations." (Shah and Meinshausen 2014). However, this definition is arguably necessary but not sufficient as predictor variables which occur more often can be main effects only but not interactions. For instance, suppose that our hypothetical Airbnb listings dataset only has two binary features, chair-in-image and face-in-image which do not interact in predicting the listing's click probability. Listings with either (i) chair-in-image = 1 or (ii) face-in-image = 1 have a click probability .2, while listings with (iii) chair-in-image = 1 and face-in-image = 1 have a click probability of .4. In this example, gRITs will sample an observation from (iii) twice as likely as an observation from either (i) or (ii) and detect these features as interactions; yet, these features do not interact. This rationale potentially explains why gRITs detects more false positive interactions when there is a relatively more important feature that does not interact (contingency scenario 1). The relatively more important feature has a higher chance of being indexed as an active feature in the gRITs for a randomly sampled observation, for a given class (i.e., click = 1), and lead to it being detected as a false interaction with other features.

For agents posting images, our findings suggest that they should adjust interacting image features jointly. For instance, we show that images with a (i) high distortion (e.g., >= 70) and high saturation (e.g., >= 45) have a lower average predicted click probability than an image with a (ii) high distortion (e.g., >= 50) but low saturation (e.g., <= 50), or a (iii) low distortion (e.g., < 50) and low saturation (e.g., <= 50). Hence, agents should edit both saturation and distortion jointly as opposed to editing either saturation or distortion individually. To facilitate this editing, platform managers can implement our machine learning model as a program on the platform. Based on each image feature's value chosen by the agent, this program suggests values for the other image features that yield the highest average predicted click probability. Higher predicted click probability benefits both agents and managers as a click on the listing resolves the consumer's uncertainty about the listing (e.g, Chen and Yao 2017) and raises the listing's likelihood of being transacted.

5.3. Limitations and future research

We highlight three limitations of our study which at the same time suggest avenues for further development. First, we did not translate our ML method into a parametric model which would allow for significance testing. To do so, we need to calculate the H-statistic under the null hypothesis of no interactions between each pair of features, i.e., it must be possible to set the interaction to zero and calculate the resulting PD in the ML model (Molnar 2021). This is an open question in the ML literature (Molnar 2021) which future research could address.

Second, we should apply our chosen ML method (H-statistic) in different empirical settings to ensure that the method generalizes for detecting interactions in different settings. As an example, consumer search for apparels might differ materially from consumer search for property listings, where images potentially play a more important role since a good looking piece of apparel makes the wearer happier.

Third, we apply ML methods on observational data which is correlational instead of causal. Future research should consider obtaining data using field experiments where a listing’s image features are manipulated and randomly exposed to consumers to ascertain the causal impact of image features can then be ascertained.

References

Andersson U, Cuervo-Cazurra A, Nielsen B (2014) From the editors: Explaining interaction effects within and across levels of analysis. *J. Int. Bus. Stud.* 45:1063-1071.

Andrews M, Luo X, Fang Z, Ghose A (2016) Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Sci.* 35(2):218-233.

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. *J. Royal Stat. Society: Series B (Statistical Methodology)* 82(4):1059-1086.

Basu S, Kumbier K, Brown JB, Yu B (2018) Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.* 115(8): 1943–1948.

Boulesteix AL, Janitza S, Hapfelmeier A, Van Steen K, Strobl C (2015) Letter to the Editor: On the term ‘interaction’ and related phrases in the literature on Random Forests. *Briefings Bioinformatics* 16(2):338-345.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and regression trees*. (Taylor and Francis, New York)

Breiman L (2001). Random forests. *Machine Learning* 45(1):5-32.

Chen X, Wang M, Zhang H (2011) The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery* 1(1): 55-63.

Chen Y, Yao S (2017) Sequential search with refinement: Model and application with click-stream data. *Management Sci.* 63(12):4345-4365.

Choudhury P, Allen RT, Endres MG (2021) Machine learning for pattern discovery in management research. *Strat. Management J.* 42(1): 30-57.

Cohen J, Cohen P, West SG, Aiken LS (2003) *Applied Multiple Regression/Correlation Analysis For The Behavioral Sciences*, 3rd ed. (Taylor and Francis, New York).

Cui D, Curry D (2005) Prediction in marketing using the support vector machine. *Marketing Sci.* 24(4):595-615.

Darst BF, Malecki KC, Engelman CD (2018) Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics* 19(1):1-6.

Dzyabura D, Yoganarasimhan H (2018) Machine learning and marketing. Mizik N, Hanssens DM, eds. *Handbook of Marketing Analytics: Methods and Applications in Marketing Management, Public Policy, and Litigation Support* (Edward Elgar, Massachusetts), 255–279.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Annals Stat.* 29(5):1189-1232.

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *Annals Applied Stat.* 2(3):916-954.

Genesove D, Han L (2012) Search and matching in the housing market. *J. Urban Econom.* 72(1):31-45.

Ghose A, Goldfarb A, Han SP (2012) How is the mobile Internet different? Search costs and local activities. *Inform. Systems Res.* 24(3):613–631.

Greenwell BM, Boehmke BC, McCarthy AJ (2018) A simple and effective model-based variable importance measure. Working paper, arXiv preprint arXiv:1805.04755.

Hagen L, Uetake K, Yang N, Bollinger B, Chaney A, Dzyabura D, Etkin J, et al. (2020) How can machine learning aid behavioral marketing research? *Marketing Lett.* 31(4):361-370.

Hasanin T, Khoshgoftaar T (2018) The effects of random undersampling with simulated class imbalance for big data. *Proc.* 2018 International Conference on Information Reuse and Integration (IRI) IEEE (Institute of Electrical and Electronics Engineers, Utah, Salt Lake City), 70-79.

Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: Data mining, inference, and prediction*. (Springer-Verlag, New York).

King G, Zeng L (2001) Logistic regression in rare events data. *Political Analysis* 9(2):137-163.

Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. *J. Marketing Res.* 43(2):276-286.

Li H, D Simchi-Levi, Wu MX, Zhu W (2019) Estimating and exploiting the impact of photo layout in sharing economy. Working paper, SSRN.

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: Exploiting interactions using random forests. *BMC genetics* 5(1):1-13.

Ma L, Sun B (2020) Machine learning and AI in marketing - connecting computing power to human insights. *Int. J. Res. Marketing* 37(3):481-504.

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *Amer. Econom. Rev.* 105(1): 272-98.

MacInnis DJ (2011) A framework for conceptual contributions in marketing. *J. Marketing* 75(4):136-154.

Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. *IEEE Transactions Image Processing*. 21(12): 4695-4708.

Molnar C (2021) Interpretable machine learning: A guide for making black box models explainable. Retrieved March 31, 2021 https://christophm.github.io/interpretable-ml-book/

Molnar C, Schratz P (2020) Package ‘iml’. Retrieved March 29, 2021 https://cran.r-project.org/web/packages/iml/iml.pdf

Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* 116 (44):22071-22080.

Narayanan S, Kalyanam K (2015) Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Sci.* 34(3):388-407.

Oh S (2019) Feature interaction in terms of prediction performance. *Applied Sci.* 9(23):5191-5203.

Shah RD, Meinshausen N (2014) Random intersection trees. *J. Machine Learn. Res.* 15:629-654.

Sirmans SG, Macpherson DA, Zietz EN (2005). The composition of hedonic pricing models. *J. Real Estate Lit.* 13(1):1–44.

Statista (2020). Most popular real estate websites in the United States as of January 2020 based on unique sites visits. Retrieved May 21, https://www.statista.com/statistics/381468/most-popular-real-estate-websites-by-monthly-visits-usa/

Thanei GA, Meinshausen N, Shah R.D (2018) The xyz algorithm for fast interaction search in high-dimensional data. *J. Machine Learning Res.* 19(37):1-42.

Tsang M, Cheng D, Liu H, Feng X, Zhou E, Liu Y (2020) Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection. *International Conference on Learning Representations 2020.*

Ursu R (2018) The power of rankings: quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Sci*. 37(4):530-552.

University of Michigan Library Research Guides (2020). All about images. Retrieved May 20, https://guides.lib.umich.edu/c.php?g=282942&p=1885350

Umbelina J (2019) Top property portals in Southeast Asia in August 2019. Retrieved May 21, 2020, https://www.josebaumbelina.com/real-estate-markets/top-property-portals-in-southeast-asia-in-august-2019/

Venkatraman N (1989) The concept of fit in strategy research: Toward verbal and statistical correspondence. *Acad. Management Rev.* 14(3):423-444.

Warner EJ, Barsky RB (1995) The timing and magnitude of retail store markdowns: Evidence from weekends and holidays. *Quart. J. Econom.* 110(2):321–352.

Wei Z, Li H (2007). Nonparametric pathway-based regression models for analysis of genomic data. *Biostat.* 8(2):265-284.

Wright M, Ziegler A, König IR (2016) Do little interactions get lost in dark random forests? *BMC Bioinformatics.* 17(1):1-10.

Whetten DA (1989) What constitutes a theoretical contribution? *Acad. Management Rev.* 14(4):490-495.

Xu Y, Liu D, Quan Y, Le Callet P (2015) Fractal analysis for reduced reference image quality assessment. *IEEE Transactions Image Processing*. 24(7):2098-2109.

Li H, D Simchi-Levi, Wu MX, Zhu W (2019) Estimating and exploiting the impact of photo layout in sharing economy. Working paper, SSRN.

Zhang S, Lee DK, Singh, Param V, Srinivasan K (2017) How much is an image worth? Airbnb property demand estimation leveraging large scale image analytics. Working paper, SSRN

Zhao H (2020) Raising awareness and signaling quality to uninformed consumers: A price-advertising model. *Marketing Sci.* 19(4):390-396.

Zhao Q, Hastie T (2021) Causal interpretations of black-box models. *J. Bus. Econom. Stat.* 39(1):1-10.

Zumpano LV, Johnson KH, Anderson RI (2003) Internet use and real estate brokerage market intermediation. *J. Housing Econom.* 12(2):134-150

Appendices

Appendix 6. Software Implementation for Tuning RF, PDPs

We recommend commonly used Python packages for tuning the random forest (RF), visualizing partial dependence (PD) plots and accumulated local effects (ALE) plots, their requisite commands and the key parameters for these commands in Table W1. We also provide guidance on using the requisite commands in these packages.

Table A6. Commonly Used Packages, Commands and Key Parameters

Goal	Package	Requisite commands	Key parameters	Website
Tune RF	1) Scikit-Learn (2021)) Choudhury t al. (2021)	random_search	(i) model, (ii) tuning parameters, (iii) tuning values, (iv) number of folds, (v) training dataset	1) https://scikit-learn.org/stable/ 2) https://tinyurl.com/3wfk8zb5
PD plots	Scikit-Learn (2021)	plot_partial_dependence	(i) model, (ii) features plotted, (iii) test dataset	https://scikit-learn.org/stable/

Before implementing the machine learning methods, users must specify the test and training dataset in the data pre-processing stage. We use the GroupShuffleSplit command in Scikit-Learn package which randomly generates train and test indices for each observation by sampling at the observation level. As our dataset contains multiple listings showed to the same consumer, this command samples the data by restricting the same consumer to show up in either train or test datasets, but not both. Doing so prevents leakage of information from the test dataset into the train dataset and avoids artificially inflating the performance of the machine learning model. There are three required parameters (i) proportion of dataset assigned to the testing dataset (‘test_size’), (ii) seed used to generate random train and test indices (‘random_state’), (iii) features, target, and group variable. Following Choudhury et al. (2021), we set the test_size to .3

We tune a RF by using the random_search command in the Scikit-Learn package. This command requires us to specify the (i) tuning parameters, (ii) values of parameters to be tuned, and (iii) number of folds (if using K-fold cross validation). Choudhury et al. (2021 supplementary section 1.1) provides a convenient Python code workflow which uses the random_search command and provides recommended parameters and values for (i) to (iii).

To visualize PD plots, we use the plot_partial_dependence command in the Scikit-Learn package. The key parameters for this command are the (i) model (e.g., a tuned RF), (ii) features to be plotted (e.g., single feature or combinations of features), (iii) test dataset. Besides the key parameters, we highlight three other optional parameters which the user might find useful to tweak, that are kept at default settings. The first parameter is the number of computer cores used for plotting (n_jobs) with a default of '1'. To use all cores, set this parameter to '-1' instead. The second parameter is the number of equally spaced points on the axes of the plots, for each target feature (grid_resolution) with a default value of '100'. To reduce computation time, users can lower the value for this parameter. The final parameter is the lower and upper percentile used to create the extreme values for the PD plot's axes ('percentiles') with a default value of (.05, .95). This parameter can set to (0,1) if users wish the axes to include the data's full range.

Appendix 7. Detected Interactions and Rankings in Baseline Scenario

This web appendix displays the detailed output for the baseline scenario (features are equally important and the interacting features are uncorrelated) with regards to the detection and ranking of interactions for HIMPS, gRITs and the H-statistic.

Table A7. Six Interactions from HIMPS

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
	x3x4	x3x4	x2x4	x3x4	x2x4
	x1x4	x1x4	x3x4	x2x3	x3x4
	x2x4	x1x2	x1x4	x1x3	x1x4
	x1x3	x2x4	x2x3	x2x4	x1x2
	x2x3	x2x3	x1x3	x1x4	x2x3
	x1x2	x1x3	x1x2	x1x2	x1x3

Note: HIMPS does not rank interactions

Table A8. Top Five Interactions from gRITs

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	x1x4 (.73)	x1x2 (.70)	x1x2 (.93)	x1x2 (.80)	x1x2 (.60)
2	x1x3 (.67)	x2x3 (.67)	x3x4 (.67)	x3x4 (.73)	x3x4 (.60)
3	x2x4 (.50)	x2x4 (.60)	x1x3 (.63)	x1x4 (.57)	
4	x2x3 (.50)	x3x4 (.60)	x2x3 (.60)	x1x3 (.53)	
5		x2x5 (.60)		x2x4 (.53)	

Note: only interactions with scores >= .50 are listed

Table A9. Top Five Interactions (Scores) from H-statistic

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	x3x4 (.82)	x3x4 (.77)	x3x4 (.75)	x3x4 (.84)	x3x4 (.83)
2	x1x3 (.44)	x1x2 (.34)	x1x2 (.63)	x1x2 (.36)	x1x2 (.47)
3	x1x2 (.33)	x1x4 (.32)	x1x3 (.34)	x1x4 (.31)	x1x4 (.27)
4	x2x3 (.23)	x2x3 (.30)	x1x4 (.25)	x5x6 (.17)	x1x5 (.20)
5	x6x7 (.22)	x2x5 (.30)	x2x4 (.18)	x1x3 (.15)	x5x6 (.16)

Appendix 8. Detected Interactions and Rankings in Contingency Scenario 1

This web appendix displays the detailed output for contingency scenario 1 with regards to the detection and ranking of interactions for HIMPS, gRITs and the H-statistic.

Table A10. Six Interactions from HIMPS

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
	x3x5 (.40)	x4x5 (.51)	x1x5 (.45)	x3x5 (.43)	x1x5 (.51)
	x4x5 (.38)	x3x5 (.45)	x2x5 (.44)	x3x4 (.41)	x2x5 (.49)
	x3x4 (.38)	x3x4 (.38)	x3x5 (.43)	x4x5 (.41)	x3x5 (.43)
	x1x3 (.30)	x1x4 (.35)	x1x2 (.34)	x1x5 (.36)	x1x2 (.37)
	x1x4 (.29)	x1x5 (.42)	x1x3 (.33)	x1x6 (.36)	x1x3 (.31)
	x1x5 (.32)	x1x3 (.29)	x2x3 (.33)	x1x4 (.34)	x2x3 (.30)

Note: HIMPS does not rank interactions

Table A11. Top Five Interactions from gRITs

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	x4x5 (.97)	x2x5 (.73)	x2x5 (.87)	x1x5 (.67)	x3x5 (.83)
2	x3x5 (.97)	x1x5 (.63)	x3x5 (.53)	x4x5 (.63)	x2x5 (.73)
3	x2x5 (.93)	x4x5 (.63)	x4x5 (.50)	x2x5 (.60)	x1x5 (.57)
4	x1x5 (.93)		x2x4 (.50)		x4x5 (.50)
5	x3x4 (.80)		x1x5 (.50)		

Note: only interactions with scores >= .50 are listed

Table A12. Top Five Interactions (Scores) from H-statistic

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	x3x4 (.88)	x3x4 (.48)	x3x4 (.56)	x3x4 (.61)	x1x2 (.50)
2	x3x5 (.45)	x4x5 (.38)	x1x2 (.36)	x2x5 (.44)	x3x4 (.37)
3	x6x7 (.43)	x1x5 (.27)	x4x5 (.25)	x1x2 (.30)	x4x5 (.23)
4	x4x5 (.43)	x2x5 (.24)	x1x5 (.22)	x6x7 (.26)	x1x3 (.17)
5	x1x2 (.39)	x1x3 (.17)	x3x5 (.19)	x4x5 (.25)	x1x5 (.16)

Appendix 9. Detected Interactions and Rankings in Contingency Scenario 2

This web appendix displays the detailed output for contingency scenario 2 (correlated features which interact with each other) with regards to the detection and ranking of interactions for HIMPS, gRITs, and the H-statistic.

Table A13. Six Interactions from HIMPS

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
	x3x4	x3x4	x3x4	x3x4	x2x4
	x2x4	x2x4	x2x4	x2x4	x3x4
	x2x3	x2x3	x1x4	x1x4	x1x4
	x4x5	x1x4	x2x3	x2x3	x2x3
	x3x5	x1x3	x2x3	x1x3	x1x2
	x2x5	x1x2	x1x2	x1x2	x1x3

Note: HIMPS does not rank interactions

Table A14. Top Five Interactions (Stability Scores) from gRITs

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	x3x4 (.83)	x3x4 (.70)	x1x2 (.73)	x1x2 (.63)	x2x4 (.60)
2		x2x4 (.60)	x2x4 (.57)		x3x4 (.60)
3			x3x4 (.53)		x1x2 (.57)
4			x2x3 (.53)		x2x3 (.53)
5					x2x5 (.50)

Note: only interactions with scores >= .50 are listed

Table A15. Top 5 Interactions (Scores) from H-statistic

Rank	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
1	x3x4 (.90)	x3x4 (.75)	x3x4 (.78)	x3x4 (.70)	x3x4 (.68)
2	x2x4 (.36)	x6x7 (.41)	x1x2 (.46)	x6x7 (.33)	x1x2 (.49)
3	x4x5 (.35)	x2x3 (.33)	x2x4 (.35)	x2x3 (.25)	x2x3 (.43)
4	x1x2 (.32)	x2x4 (.26)	x1x4 (.33)	x1x4 (.22)	x2x4 (.32)
5	x1x4 (.23)	x5x7 (.25)	x2x3 (.31)	x1x2 (.20)	x3x6 (.27)

Appendix 10. Summary Table

Table A16. Summary Table for Target and All Features

Target and Features	Description	Operationalization	Scale
$Clicks_{ijt}$	Did consumer click on the listing	If a listing was clicked	Binary; 1 for yes and 0 for no
Image Features			
$Distortion_j$	Distortion of listing; higher distortion is less desirable	BRISQUE (Mittal et al. 2012)	0 to 100; 0 is least distorted
$Asymmetry_j$	Asymmetry (lack of symmetry) of listing; higher asymmetry less desirable	Hamming distance between transposed and original image	0 to 64; 0 is least asymmetric
Hue_j	Average hue (i.e., color) of listing; warm hues more desirable	Mean of hue over all pixels	0 to 255; 0 is red hue
Sat_j	Average saturation (i.e., vividness) of listing; higher saturation more desirable	Mean of saturation over all pixels	0 to 255, 0 is unsaturated and 255 is fully saturated
$Value_j$	Average value (i.e., brightness of color) of listing; higher value more desirable	Mean of value over all pixels	0 to 255, 0 is pure white and 255 is pure black
Non-Image Features			
$Rank_{ijt}$	Rank of listing displayed to consumer	Observed rank	Continuous from 1 to 100
$Price_j$	Rental price of listing j	Observed monthly rental	Continuous from 700 to 17000
$Floor\ area_j$	Floor area of listing j	Observed area in squared meters	Continuous from 100 to 4252
$Condo\_unit_j$	Condo-bedroom-bathroom identifier for listing	1 dummy for each identifier	73 dummies
$Condo\_count_{ijt}$	Total number of condo units shown on each page	Count condo units for each listing on each page	Continuous. 1 to 30
$Page_{ijt}$	Page number of listing shown to consumer	Page number of listing	Continuous. 1 to 5
Consumer's Features			
$Cum\_search\_time_{it}$	Cumulative search time by consumer	Sum of total search time in hours	Continuous from 1 to 765,798
$Session\_time_{it}$	Search time for each session by consumer	Time spent in seconds for session i	Continuous from 1 to 15,243
$Session\_count_{it}$	Cumulative session count of the consumer	Count number of unique session ids for consumer	Continuous from 1 to 202
$Time_{it}$	Time of day when consumer searched on platform	Based on date-timestamp for each consumer, 1 dummy for each hour	24 dummies in all
Day_{it}	Day of week when consumer searched on platform	1 dummy for each day of week	7 dummies in all
$Device_i$	Device used by consumer when searching platform	1 dummy for mobile, 1 dummy for tablet, 1 dummy for desktop	3 dummies in all
$Channel_{it}$	Channel from which consumer arrives on platform	1 dummy for each channel	11 dummies in all

Note: index *i* refers to consumer *i*, index *j* denotes listing *j*, index *t* refers to time at which listing *j* was exposed to consumer *i*

Appendix 11. Detecting, Ranking, and Directionality of True Interactions with Ad-Hoc Measure in HIMPS

This appendix displays the performance of each method on the three metrics, when we specify an ad-hoc measure of ranking interactions detected from the gRITs using the sum of their feature importance scores, for the baseline scenario and contingency scenarios 1 and 2

Table A17. Proportion of Datasets Satisfying Each of the Three Metrics in Baseline Scenario

Metric	HIMPS	gRITs	H-statistic
Detection	1.0	0	1.0
Ranking	1.0	0	1.0
Directionality	1.0	N.A.	1.0

Note: Ad-hoc measure in HIMPS uses sum of importance scores

Table A18. Proportion of Datasets Satisfying Each of the Three Metrics in Contingency Scenario 1

Metric	HIMPS	gRITs	H-statistic
Detection	0	0	.6
Ranking	0	0	.6
Directionality	N.A.	N.A.	.6

Note: Ad-hoc measure in HIMPS uses sum of importance scores

Table A19. Proportion of Datasets Satisfying Each of the Three Metrics in Contingency Scenario 2

Metric	HIMPS	gRITs	H-statistic
Detection	.8	0	.8
Ranking	.8	0	.8
Directionality	.8	N.A.	.8

Note: Ad-hoc measure in HIMPS uses sum of importance scores

Discussion

I discuss three reflections arising from both of my essays where I (i) find a dual role of comparison site visits in the consumer's search in essay one and (ii) show the importance of image feature interactions for predicting clicks on two-sided search platforms in essay two.

My first reflection from my dissertation is the importance of using new forms of data to refine and deepen theories of consumer search. For instance, in my first essay I used data on the consumer's initial consideration set and expectations of obtaining a better deal where both constructs are measured prior to search, and the consumer's subsequent search behavior to understand when and how comparison sites are used in search. Similarly, my second essay utilizes new data which is the image features of the listings displayed to consumers and their observed search behavior to determine the boundary conditions of the relationship between image features and the consumer's clicks on a listing. Scholars who work on search in the future should continue obtaining new forms of data in order to extract new insights from consumer search behavior. While I used new data on previously unobserved consumer pre-search characteristics and image features in my first and second essays respectively, scholars should note that other instances of new data also exist in consumer search when studying this phenomenon. For instance, videos are commonly used in the Facebook livestream where sellers introduce their products, and viewers type in their comments to the seller's product introduction. Similarly, two-sided search platforms such as Taobao now allow videos to be uploaded in addition to a still image, which potentially allows the researcher to study how video features affects the consumer's search, using video features as a new data source.

My second reflection is on the impact of new methods available on the possible movement away from theory testing and towards theory development. In my first essay, I developed a priori hypotheses based on the extant search literature before testing my hypotheses on new data using parametric methods. In my second essay, I did not develop such a priori hypotheses. Instead, I utilized machine learning methods to learn from the data and discover interactions between image features that predict clicks on a listing. Interactions are important for theory building as they are used to infer the boundary conditions of a theory (Whetten 1989). With the increasing development and refinement of machine learning methods, scholars should consider using these methods to build theory, and uncover new and robust relationships between constructs. Arguably, such theory development is important to the field's long-run vitality as it allows us to generate home-grown theories by learning from the data.

My third reflection involves the implicit functional form assumptions when scholars utilize existing theories to derive a priori hypotheses. Specifically, scholars tend to posit either a linear or a curvilinear main effect between constructs (Haans et al. 2016), and the moderating effect of a construct as a product term with the focal construct (Kenny 2018). However, there is no reason to believe that the true main effect between constructs is necessarily linear or curvilinear and that the moderation effect is linear. For instance, in my second essay I utilize machine learning methods to infer the true functional form of the features (i.e., a listing's image and non-image features, and the consumer's features) and the interactions between features on a listing's predicted click probability. By doing so, I free myself of the implicit functional form assumptions used when developing a priori hypotheses. After all, as pointed out by Rumsfeld (2002) there are also "unknown unknowns – the ones we don't know we don't know." Utilizing machine learning methods thus allow us to bypass the limits of our current knowledge and let us derive a more accurate and realistic picture of the true relationship between the constructs that are studied.