Extended Conversion: Capturing Successful Interactions in Voice Shopping

Elad Haramaty
Amazon Research
Haifa, Israel

Zohar Karnin
Amazon Research
Haifa, Israel

Arnon Lazerson
Amazon Research
Haifa, Israel

Liane Lewin-Eytan
Amazon Research
Haifa, Israel

Yoelle Maarek
Amazon Research
Haifa, Israel

ABSTRACT

Being able to measure the success of online shopping interactions is crucial in order to evaluate and optimize the performance of e-commerce systems. It is especially challenging in the domain of voice shopping, typically supported by voice-based AI assistants. Unlike Web shopping, which offers a rich amount of behavioral signals such as clicks, in voice shopping a non-negligible amount of shopping interactions frequently ends without any immediate explicit or implicit user behavioral signal. Moreover, users may start their journey using a voice-enabled device, but complete it elsewhere, for example on their smartphone mobile app or a Web browser. We explore the challenge of measuring successful interactions in voice product search based on users' behavior, and propose a medium-term reward metric named *Extended ConVersion* (ECVR). ECVR extends the notion of conversion beyond the usual purchase action, which serves as an undisputed measure of success in e-commerce. More specifically, it also captures purchase actions that occur at a later stage during a same shopping journey, and possibly on different channel than the one on which the interaction started. In this paper, we formally define the ECVR metric, describe multiple ways of evaluating the quality of a metric, and use these to explore different parameters for ECVR. After selecting the most appropriate parameters, we show that a ranking system optimized for ECVR, set up with these parameters, leads to improvements in long-term engagement and revenue, without compromising immediate conversion gains.

CCS CONCEPTS

• **Information systems** → *Relevance assessment*; **Online shopping**.

KEYWORDS

Extended Conversion, Voice Shopping

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09…$15.00
https://doi.org/10.1145/3604915.3608836

ACM Reference Format:
Elad Haramaty, Zohar Karnin, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. 2023. Extended Conversion: Capturing Successful Interactions in Voice Shopping. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3604915.3608836

1 INTRODUCTION

E-commerce online systems have traditionally invested a great deal of efforts in automatically assessing user satisfaction from behavioral signals (clicks, dwell time, etc) so as to optimize their effectiveness [2, 9, 24, 29]. Shopping experiences over voice-based AI assistants however pose a challenge as such signals are not readily available, and a non-negligible amount of shopping interactions frequently end without any immediate explicit or implicit user behavioral signal. While manual annotations on a sample of the traffic have been traditionally used to assess satisfaction, they are costly and may even be inaccurate when annotators miss the subjective or implicit needs of users [10]. For example, [5] shows that users often purchase products that were annotated as irrelevant. In this work, we explore the challenge of measuring successful interactions in voice product search, based on users' behavior. We argue here that users' behavioral signals should not be limited to users' immediate actions but should include their entire shopping journey.

In e-commerce, a shopping journey typically refers to the path customers take to purchase a product. It includes several main stages: problem recognition, information search, evaluation of alternatives, purchase, and post-purchase [14, 20, 22]. While models such as the *decision analysis models* e.g., [13] and *hierarchy of effects models* e.g., [26] use different stages, they all recognize that the purchase action is not an independent event, but part of a decision making process. Thus, an interaction occurring at an early stage of a shopping journey will probably not be directly followed by a shopping action, even if customers received the appropriate information for that stage. Consequently, it is critical to be able to understand whether an action performed at a later stage of a shopping journey should be attributed to an appropriate behavior of the system at earlier stages. This is even more critical in voice-based shopping, where users typically initiate their journey on a voice-enabled device expressing informational or exploratory needs (e.g., 'find video games'), rather than transactional needs (e.g., 'buy iphone 14') [5]. As they advance in their journey and progress towards a purchase decision, they might need more detailed information that cannot be properly presented on a voice-only medium or small form-factor device. Similarly, for more expensive or complex products or with products that require visualization, users will often favor a larger form-factor and will continue their journey on their personal computer [27]. We argue here that voice shopping should not be considered in isolation: users may start their journey on a voice-based device but complete it elsewhere, raising the need for a success metric that takes the entire ecommerce system that supports the user's shopping journey into account.

The most common behavioral signal used for evaluation and optimization of online shopping systems is *ConVersion Rate* (CVR) e.g., [16, 17, 23], which is defined as the number of shopping requests that lead to conversion divided by the number of all shopping requests. Conversion captures the subjective notion of user satisfaction, and is a clear indication of success in an e-commerce system. It is scalable, as it can easily be collected from historical logs with no need of human annotations. However, as discussed above, it is sparse and fails to capture purchase actions that are part of a same shopping journey but conducted at a later stage and/or on a different interaction channel.

We introduce here the notion of *Extended Conversion* (ECVR), which considers an interaction successful if it leads to a successful completion of the shopping journey by a purchase action. ECVR attributes the purchase of a product to some earlier shopping interaction related to a similar product. This is achieved by broadening the criteria for product relevance and the time period within which the purchase is made. In the rest of this paper, we conduct data analysis to support our intuition that a large portion of voice shopping actions relate to interactions performed at early stages of the shopping journey. We provide a formal definition of ECVR based on two parameters for product similarity and time period. We then describe our experiments to select these appropriate parameters. Finally, we compare a ranking model in a real voice shopping assistant perform differently when optimized for ECVR vs CVR.

2 RELATED WORK

Voice Product search. Research on voice-based search [11] showed that voice based search queries are longer and closer to natural speech than textual queries, indicating that voice search differs from text search not only in available signals but in the nature of usage data. Another key difference is user's behavior, as shown in [5], which demonstrates that customers purchase or engage with objectively irrelevant search results in voice product search. A clear constraint of the voice channel is that it limits the amount of information that can be communicated to the user. This led to new approaches for conveying useful information about search results. Thus [4] focuses on improving conversational product search by effectively incorporating feedback on aspect-value pairs with the ranking model, while [18, 19] show that adding explainablity to voice product search results helps the decision making process of the customers.

In our paper, we focus on measuring the quality of a voice product search system. Unlike the latter, we do not focus on the presentation of results but rather on the ranking system returning a specific item for a request.

Optimizing for long term rewards. While [1] provide both theoretical and empirical evidence to the fact that in order to optimize for a long term reward optimizing for a short term surrogate can prove to be more effective (as less noisy), in our context of shopping and web engagement, we found that there is no consensus around how short-term signals can be used to optimize for long-term reward. The long-term reward varies based on business needs, and the nature of the short-term signals and the way they are used depends on the use case. In [15], the authors aim to maximize the number of conversions per month. They propose a reinforcement learning (RL) technique to do so, which extends episodic RL in order to take into account the effect of one session on another. In [25], the authors want to optimize long term engagement, and specifically focus on a way to convert infrequent users to frequent users. They list a handful of surrogate features based on short-term usage that are both less noisy, dense, and available within a short amount of time. They analyze the causal relationship between these surrogates and the true objective in order to choose a two surrogate feature (time to revisit and the diversity of selected items), and show empirical evidence that optimizing for either surrogate indeed improves the long-term reward. The authors of [30] optimize for another objective, namely long term revenue. They do so by identifying short term surrogates such as immediate revenue and user actions. They learn a model mapping these to long term revenue based on a small randomized experiment (alleviating the bias introduced in a live system), and optimize for the aggregation of the surrogates rather than for the long term signal directly. Another approach is proposed by [28], which jointly optimizes for click rate and reducing the time it would take a user to return to the website. They motivate the 'time to return' objective by showing it is not 100% correlated with click rate and propose a standard multi-objective framework. Finally, [12] takes a long-term view by identifying items with the potential of being trending/popular when they are new as well as learn how to best present them to users in order to make sure they indeed become popular.

The papers mentioned above thus validate a short term surrogate only via its connection to the long term reward. In our work, in addition to testing this direct connection, we propose additional ways to validate a short-term surrogate.

3 USERS' PURCHASE BEHAVIOR OVER THEIR SHOPPING JOURNEY

In order to verify that a large amount of voice shopping interactions belong to early stages of the shopping journey and thus are not considered as successful in CVR, we conduct here a quantitative analysis on the shopping traffic of a major voice-based digital assistant.

We built a dataset of user interactions with a leading commercial voice-based AI assistant. During these interactions, users searched for a product, and the AI assistant answered with a specific product offer. The data is represented as a list feature of vectors encoding information about the user, historical user actions, the user's request and its context, and the product returned by the assistant. The dataset[1] consists of a sample of more than a million interactions from a time period of 6 weeks.

[1] In order to protect user's privacy, all data was automatically processed by machines and no additional details are being shared here.

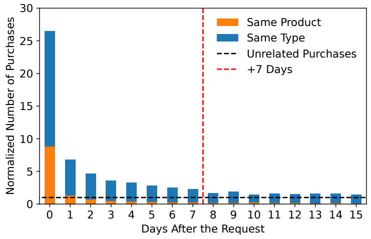

Figure 1: Normalized purchases of the same type, as function of the days from the request, in early shopping stages

We use a heuristic-based approach, leveraging utterance patterns to differentiate between early and later stages in the shopping journey. For example, utterances following the 'show me X' or 'what is the price of X' patterns are considered as early stages, while those following the 'add X to cart', or 'buy X' patterns are considered as late stages. We verified that a large portion of the traffic lies in the early stages of the shopping journey. For every request, we list the exact product offered by the system, and track whether the user purchased either this exact product, or a product of the same type[2], within the following 14 days.

Figure 1 presents statistics for early shopping stages. A 'same-product purchase' is a purchase of the exact product offered by the voice shopping assistant, a 'same-type purchase' corresponds to the purchase of a product of the same type as the offered product. We see that most purchases are not of the offered products and in particular do not occur within the same voice product search interaction. These purchases can be done either via voice, the mobile app, or a Web browser. In addition, a non-negligible portion of the purchases occur a day or even a week after the original interaction, showing the nature of the traffic, where the purchase decision-making process takes time. It could be argued that purchases of products of the same type, that are performed within a few days, may be unrelated to the original interaction. To answer that, we observe the dashed black line in the figure, representing the 'prior', which is defined as the probability that a user purchases an item of the same type without the condition of having queried it in the past. To obtain the amount of unrelated purchases, we calculated the amount of same-type purchases that occurred on an average day between 1 week to one month *before* the request. Intuitively, such purchases cannot be attributed to a request that is issued more than a week later. The y-axis in Figure 1 is normalized by the number of unrelated purchases.

We see that within the first 7 days, the same-type purchase fraction is much larger (over twice) than the prior, validating the strong connection these purchases have to the original interaction. Moreover, most of the purchases are not of the same offered product, but rather of a similar product. Thus, we see that the straightforward approach of measuring only direct conversions is missing most of the interactions that ended with a successful transaction.

Finally, in Figure 2 we consider late shopping stages and compare our findings to this case. We see that for this traffic, almost all purchases are of the offered product within the same day.

[2] Two products are of the same 'Type' if they belong to the same product-type, e.g. "bread" or "television". We provide more details in Section 4

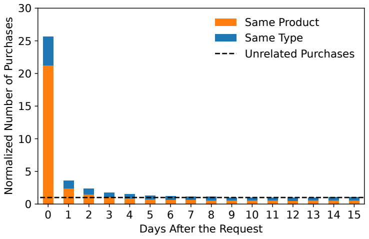

Figure 2: Normalized purchases of the same type, as function of the days from the request, in late shopping stages

4 DEFINING EXTENDED CONVERSION AS A PARAMETERIZED METRIC

To better capture successful interactions, we propose a parameterized metric called *Extented Conversion* (ECVR) that reflects the medium-term behavior of users in their shopping journey, as opposed to the immediate term as in CVR. ECVR extends CVR along two axes, time and product similarity. The time axis is extended by considering not only an immediate purchase but a purchase occurring within a consecutive time period (not necessarily using voice). Product similarity is extended by considering the purchase of similar products to the one offered by the voice AI assistant, for example if the item offered is a pack of batteries of one brand and the user eventually purchased batteries of a different brand, or a different size of pack. We discuss later in this paper how we propose to select the appropriate points on these axes as parameters of the ECVR definition.

We consider a hierarchy of five natural product similarity levels differing in their specificity, from the most specific similarity level, where a product is only similar to itself, to the most general similarity level, which includes all products:

- **Product**: a trivial similarity level in which a product is only similar to itself.
- **Substitutions**: products are considered similar if they can replace one another (e.g., in case one of them is out of stock). In other words, the customers are mostly indifferent between substituting products, e.g., different brands of paper towels with similar size, quality and price.
- **Type**: products are considered similar if they belong to the same product-type, e.g., "bread" or "television". In other words, if a customer just purchased a product of a specific 'Type', they are not expected to be in need of another product from the same 'Type'.
- **Department**: products are considered similar if they belong to the same department, which is a natural way to partition the universe of products, just like aisle descriptions in a physical store, e.g., electronics, food, apparel, etc.
- **All**: a trivial similarity level, in which a product is similar to all other products.

In Table 1, we report the number of clusters per similarity level in our sample of the traffic, where a cluster contains all the products that are similar according to the level. In addition, we report the average number of products per cluster.

Accurately attributing a purchase to a request is a challenging task. We have no way of knowing which specific purchase was triggered by a given request. Since we cannot accurately attribute a

Sim. Level	#Clusters	Avg Size
Product	1,475K	1
Substitutions	385K	3
Type	11K	128
Department	50	29K
All	1	1,475K

Table 1: Number of clusters and average cluster size (in terms of different products) for the different similarity levels.

purchase to a request (except from the case of immediate purchases), we use an approximation - attributing a purchase to a request if the purchased product is similar to the product originally offered for the request. Our method may be noisy and attribute an unrelated purchase to a request. To mitigate this issue, we attribute a purchase to a request by using this combination of similarity and reasonable time window, (a predefined time period), very much like users' sessions have been determined in Web search [8]. Formally, we define extended conversion as follows, with Δt and S being parameters of the metric:

Definition 4.1. Let $\mathcal{S}$ be similarity level, i.e., a partition of the product space into similarity sets. Let Δt be a time period, and $r = (u_r, t_r, p_r)$ be a request issued by user u_r at time t_r for which offer p_r was returned. The extended conversion of request r, with respect to t and $\mathcal{S}$, is a binary label which is positive iff u_r purchases a product p at time t such that

(1) $t_u \leq t \leq t_u + \Delta t$.
(2) p_u and p are similar according to $\mathcal{S}$. Namely, there exists a similarity set $S \in \mathcal{S}$ such that $p_u, p \in S$.

For example, consider a request "best LED television", for which a customer was offered an LG television first, and then purchased a Samsung neo-QLED television three days later. According to the definition this request resulted in extended conversion with respect to the type similarity level and time period of one week, since (1) the purchase was done after the request and less than a week later, and (2) there exists a product type "television" representing a similarity set in $\mathcal{S}$, that includes both the offered product and the purchased product.

4.1 Selecting the Time Period parameter

In Figure 1, we showed the normalized conversion statistics over time for the *Type* similarity level. Recall that the dashed black horizontal line represents the prior, with purchases below the line correspond to purchases unrelated to the request, while purchases above the line correspond to related purchases. As can be seen, after more than seven days from the request (marked by a red dashed line), most purchases are unrelated to the original request (below the black line), while in the first seven days from the request, most purchases are related to the request (above the black line).

The top left quadrant represents the number of true positives (TP), i.e., the number of purchases that are attributed to the request and are indeed related. Similarly, the bottom left corresponds to false positives (FP), the top right to false negatives (FN), and bottom right to true negatives (TN). Every choice of a time period represents a different partition (different red dotted line) to TP/FP/TN/FN quantities. Given these, we compute the F1-score[3] for each time period. We computed the F1 score for all five product-similarity levels and all time periods starting from one day (same day purchases) to 30 days, presented in Figure 3. The dashed black line represents a time period of seven days after the request. For Type, Department and All, the best F1-score is achieved after seven days. For Product and Substitutions, the best scores are on days 13 and 12, but they are similar to the seventh day F1-score, with an absolute difference of 1% and 0.3%, respectively.

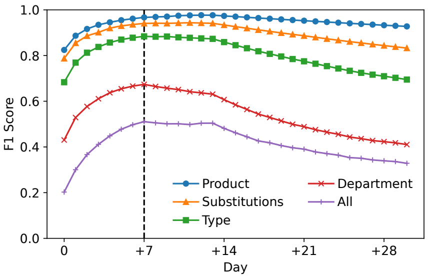

Figure 3: F1 score for different similarity levels as function of the number of days after the request.

Other than quantitative results, we take two more factors into account when deciding on the time period. First, multiples of seven days have an advantage as they reduce noise related to different behavioral patterns during different days of the week. Second, we give a slight preference to shorter time periods as longer delays create metrics that are more challenging to monitor and optimize for. Given the near optimality of the seven days time period across all similarity levels, we recommend selecting it for ECVR in our settings and experiments in the remaining of this paper. We note that if it was not the case, one would need to choose a different time period for each similarity level.

Next, having fixed the time period, we explore different product similarity levels showing that the 'type' level provides the best results overall in terms of validity and sensitivity.

4.2 Selecting the Similarity Level parameter

We consider the two following notions in order to identify the most appropriate similarity level. *Objective relevance* commonly used in the evaluation of ranking systems, is an aggregation of annotation-based labels indicating whether a product is relevant to a query. The *Long term effect* (LTE) can be represented either by user's engagement or revenue following an interaction (excluding the interaction itself) over a long period of time. We chose time periods of both 28 days and 90 days. For engagement, we measured the number of active days, meaning days during which the user interacted with the voice-based AI assistant with a product search request. For revenue, we measured the overall monetary amount spent on products purchased, for both voice and non-voice experiences.

During our experiments, we used a setup allowing the evaluation of product ranking within the product search experience. Specifically, we used an online serving system that is randomized so

[3] F1-score balances precision and recall by taking their harmonic mean, it is given by $\frac{TP}{TP+0.5(FP+FN)}$.

that the same request will not systematically get the same offered product as a result. This property allows for an offline unbiased estimation of different rankers, [21]. We use this system to compare different rankers measured by various parameter settings for ECVR.

Sensitivity. First, we analyze how different similarity levels affect the sensitivity of ECVR, that is, how well ECVR captures changes in the quality of user experience as a function of the similarity level. To this end, we used the setup described in Section 3 to simulate two ranking models for the product search task. We chose one ranker to be clearly superior to the other, so as to verify whether the expected performance gap would be captured by each parameter setting of ECVR. The two rankers considered are (1) uniform random – the ranking procedure selects one item from the set of retrieved products, uniformly at random, and is expected to provide a poor user experience; and (2) relevance-based – the retrieved products are ranked according to their relevance score. This relevance score is computed based on an ML model trained to predict objective relevance.

We measured the performance of both rankers in terms of ECVR based on a time period of 7 days and each of the similarity levels. In Table 2, we report for each similarity level, the corresponding ECVR score difference between the relevance and uniform random rankers, as well as a 95% confidence interval. It can clearly be seen that the ECVR instances parameterized by Department and All levels are not sensitive. Their lower confidence bound is below zero, meaning they might attribute a negative impact to the positive change (moving from random ranking to relevance based ranking). The more specific similarity levels are sensitive in that their confidence interval is entirely positive. Their lower absolute difference comes from the sparsity of purchases in early stages traffic. From a sensitivity viewpoint, we thus see that the best balance is achieved by the Type level, obtaining the largest absolute difference.

Sim. Level	Diff
Product	0.35 ± 0.07
Substitutions	0.37 ± 0.09
Type	0.51 ± 0.15
Department	0.27 ± 0.34
All	0.38 ± 0.46

Table 2: Sensitivity Experiment - Diff in ECVR between relevance and random ranking with 95% confidence interval.

Sim. Level	Norm. Precision
Product	87.9
Substitutions	85.4
Type	80.9
Department	43.4
All	7.7

Table 3: Normalized objective relevance precision score for different variants of ECVR.

Objective Relevance. Next, we consider the Objective Relevance metric mentioned earlier to test whether ECVR indeed captures a positive experience for each product similarity level. The main observation motivating our experiment is that objective relevance might not guarantee a positive experience but it does capture negative experiences, meaning that if the offered product is irrelevant to the query, user experience will be poor. With this in mind, we compute the *objective relevance precision* score for the different parameter settings of ECVR. The objective relevance precision score is defined as the fraction of interactions labelled positive by ECVR that are also positive according to objective relevance. Table 3 presents the *normalized objective relevance precision* of the various parameter settings of ECVR corresponding to the different product-similarity levels. Denoting the overall rate of objective relevance (ratio of interactions with a positive objective relevance score) as η, given an objective relevance precision of p we define the normalized counterpart as $100 \cdot \frac{p-\eta}{1-\eta}$. A trivial metric always providing a positive score will get a normalized precision score of 0 and a perfect metric (according to this test) will get a score of 100. Unsurprisingly, the precision correlates with the specificity of the similarity levels, where the most specific level (Product) gets the highest score, and the most general level (All) gets the lowest score. From an Objective Relevance perspective, we thus see a small deterioration from Product to Type (88% to 81%) and a steep drop when moving to Department and All.

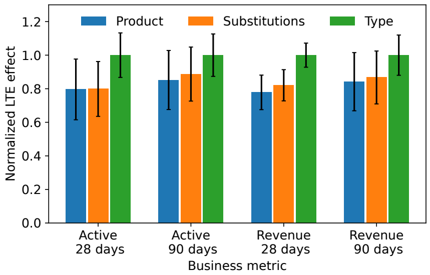

Figure 4: LTE using different similarity levels, 95% confidence intervals are marked in black .

Long Term Effect. Finally, we want to assess the long-term effect (with regrades to engagement and revenue) of different parameter settings of ECVR, using LTE metric defined earlier. To this effect, we apply a known technique called Double/Debiased ML [7] from the field of causal inference. This approach allows to remove the effect of confounding factors and identify causal, rather than spurious, relationships. Casting our problem to the terminology of causal inference, the treatment refers to whether the interaction resulted in a positive or negative ECVR, and the effect refers to LTE. The confounding factors are features related to the users history (e.g., past spending habits and engagement), and the nature of the query issued in the interaction. We use the LinearDML model of the econml [3] python library, with XGBoost [6] regressor and classifier to predict the LTE and the treatment probability. The process allows to estimate by how much a given increase in ECVR leads to an increase in LTE, along with a confidence interval[4]. In Figure 4, we show the effect of the different similarity levels on the long term business metrics. Although the confidence intervals do not always

[4] This estimate, in the field of causal inference is called the average treatment effect.

separate the different similarity levels, we do observe a clear trend across all four cases of {28/90 days, activity/revenue}, indicating that as the similarity level becomes less specific, the LTE becomes larger.

We conclude that using the Type level similarity in ECVR achieves the best overall results. It is the most sensitive metric, and displays the best long term effect. While it is not the best in terms of objective relevance precision, it provides good performance in that the results are comparable to the 'Product' and 'Substitution' levels. In the rest of this paper, we will consider for ECVR parameter settings, the purchase of a product of the same *Type* (similarity level) as the top offered product, within a *week* (time period).

5 ECVR VS CVR

After setting the parameters of the ECVR metric, we provide a deeper comparison between ECVR and the standard immediate conversion metric (CVR). We demonstrate ECVR superiority in terms of sensitivity and long term effect. In addition, we show that a ranker optimized for ECVR outperforms a ranker optimized for CVR.

Sensitivity. We repeat the experiment pertaining to sensitivity that was described in Section 4.2, this time measuring ECVR and CVR induced by applying random and relevance based ranking to voice shopping traffic corresponding to early phases of the shopping journey. We expect a sensitive metric to give a higher score to the relevance based ranker as it provides better user experience. Indeed, as both metrics are sensitive, they give a higher score to the relevance based ranker. The score difference between the relevance ranker and the random ranker is 0.51 ± 0.15% for ECVR and 0.09 ± 0.01% for CVR with a CI of 95%. As evident by the large and statistically significant difference, the ECVR metric is more sensitive to changes in user experience.

Business Metric	ECVR	CVR
Active 28 days	1 ± 0.13	0.44 ± 0.18
Active 90 days	1 ± 0.13	0.40 ± 0.17
Revenue 28 days	1 ± 0.07	0.49 ± 0.11
Revenue 90 days	1 ± 0.12	0.25 ± 0.18

Table 4: Normalized long term effect of CVR and ECVR.

Long Term Effect. As in Section 4.2, we estimate the causal effect of ECVR and CVR on long term business metrics, namely Active days and Revenue. For each metric we consider horizons of 28 days and 90 days. In Table 4 for each business metric and horizon we present the normalized LTE for both CVR and ECVR with confidence intervals of 95%. While both have a positive influence on the LTE, ECVR is superior across all tested long term metrics.

Ranker Performance. Our dataset (Section 3) contains 201 features describing the query, offered product, user and the relations between them, and is labeled with CVR and ECVR. We split the data uniformly 50:50 into train and test datasets. The train dataset is used to train two ranking models, one is optimized for conversion and one for extended conversion. In both cases we used Autogluon's Tabular Predictor[5]. As a baseline we consider an additional model optimized for Relevance, which is a natural baseline for exploratory traffic where conversions are relatively rare. In Table 5 we present the relative improvement of CVR and ECVR optimized models, with respect to a baseline model. Both models achieve better conversion than the baseline model but there is no statistical significant change between the CVR and ECVR optimized models . However, we see that the ECVR optimized model improves the extended conversion significantly with a difference of 21.1% relative to the CVR optimized model (with a confidence of 95%). Note, that by definition conversion only accounts for direct purchases while extended conversion accounts for both direct and indirect purchases. By improving extended conversion, we increase the total amount of purchases. It is interesting to see that when moving from conversion to extended conversion we increase the number of indirect purchases without harming the amount of direct purchases.

Ranker	CVR gain	ECVR gain
Relevance optimized	0%	0%
CVR optimized	23.9%	6.3%
ECVR optimized	**24.1%**	**27.4%***

* p value < 5%

Table 5: Relative improvement of ranker performance with respect to the relevance optimized ranker baseline.

6 NEXT STEPS AND CONCLUSIONS

In this paper, we introduced a new extended conversion metric, ECVR, aimed to better capture successful interactions in voice shopping and relying on two key parameters: time period and product similarity level. We conducted a series of experiments to both tune these parameters, and to compare ECVR to the standard CVR conversion metric. In addition to showcasing the usefulness of the newly proposed metric, our techniques provide (1) new insights into the domain of voice shopping and some of its challenges (2) a thought process that could potentially be applied in other instances of recommender systems, for which it is difficult to capture successful interactions, e.g., better attributing purchases to online advertisements. We plan to continue our work by moving towards an online system optimizing for ECVR. We will explore different methods for optimizing this system including ways to use actions reflecting customers' interests that are not available at inference time, such as adding an item to one's cart or continued conversation with a related query.

[5] We used the following global parameters: num_bag_folds=3, num_bag_sets=1, num_stack_levels=1. For GBM and CAT we restricted number of iterations/boost rounds to 10000, and for RF and XT we restricted number of estimators to 300.

REFERENCES

[1] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2019. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.* Technical Report. National Bureau of Economic Research.

[2] Wentian Bao, Hong Wen, Sha Li, Xiao-Yang Liu, Quan Lin, and Keping Yang. 2020. Gmcm: Graph-based micro-behavior conversion model for post-click conversion rate estimation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2201–2210.

[3] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. EconML: a Python package for ML-based heterogeneous treatment effects estimation. *GitHub* (2019).

[4] Keping Bi, Qingyao Ai, Yongfeng Zhang, and W Bruce Croft. 2019. Conversational product search based on negative feedback. In *Proceedings of the 28th acm international conference on information and knowledge management*. 359–368.

[5] David Carmel, Elad Haramaty, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. 2020. Why do people buy seemingly irrelevant items in voice product search? On the relation between product relevance and customer satisfaction in ecommerce. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 79–87.

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[7] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

[8] Deborah Donato, Francesco Bonchi, Tom Chi, and Yoelle Maarek. 2010. Do you want to take notes?: identifying research missions in Yahoo! Search Pad. In *Proceedings of WWW'2010* (Raleigh, NC, USA).

[9] Yali Du, Chang Xu, and Dacheng Tao. 2017. Privileged matrix factorization for collaborative filtering. In *IJCAI International Joint Conference on Artificial Intelligence*.

[10] Yaron Fairstein, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2022. External evaluation of ranking models under extreme position-bias. In *WSDM 2022*. https://www.amazon.science/publications/external-evaluation-of-ranking-models-under-extreme-position-bias

[11] Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 35–44.

[12] Luo Ji, Qi Qin, Bingqing Han, and Hongxia Yang. 2021. Reinforcement Learning to Optimize Lifetime Value in Cold-Start Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 782–791.

[13] Sahar Karimi, K Nadia Papamichail, and Christopher P Holland. 2015. The effect of prior knowledge and decision-making style on the online purchase decision-making process: A typology of consumer shopping behaviour. *Decision Support Systems* 77 (2015), 137–147.

[14] Katherine N Lemon and Peter C Verhoef. 2016. Understanding customer experience throughout the customer journey. *Journal of marketing* 80, 6 (2016), 69–96.

[15] Bogdan Mazoure, Paul Mineiro, Pavithra Srinath, Reza Sharifi Sedeh, Doina Precup, and Adith Swaminathan. 2021. Improving Long-Term Metrics in Recommendation Systems using Short-Horizon Reinforcement Learning. *arXiv preprint arXiv:2106.00589* (2021).

[16] Colin McFarland. 2012. *Experiment!: Website conversion rate optimization with A/B and multivariate testing*. New Riders.

[17] Risto Miikkulainen, Neil Iscoe, Aaron Shagrin, Ron Cordell, Sam Nazari, Cory Schoolland, Myles Brundage, Jonathan Epstein, Randy Dean, and Gurmeet Lamba. 2017. Conversion rate optimization through evolutionary computation. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 1193–1199.

[18] Gustavo Penha, Eyal Krikon, and Vanessa Murdock. 2022. Pairwise review-based explanations for voice product search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. 300–304.

[19] Gustavo Penha, Eyal Krikon, Vanessa Murdock, and Sandeep Avula. 2022. Helping Voice Shoppers Make Purchase Decisions. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[20] Nancy M Puccinelli, Ronald C Goodstein, Dhruv Grewal, Robert Price, Priya Raghubir, and David Stewart. 2009. Customer experience management in retailing: understanding the buying process. *Journal of retailing* 85, 1 (2009), 15–30.

[21] Yuta Saito and Thorsten Joachims. 2021. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 828–830.

[22] Susana Santos and Helena Martins Gonçalves. 2021. The consumer decision journey: A literature review of the foundational models and theories and a future perspective. *Technological Forecasting and Social Change* 173 (2021), 121117.

[23] Pim Soonsawad. 2013. Developing a new model for conversion rate optimization: A case study. *International Journal of Business and Management* 8, 10 (2013), 41.

[24] Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 17–22.

[25] Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H Chi, and Minmin Chen. 2022. Surrogate for Long-Term User Experience in Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4100–4109.

[26] Bambang Sukma Wijaya. 2015. The development of hierarchy of effects model in advertising. *International Research Journal of Business Studies* 5, 1 (2015).

[27] Karien Oude Wolbers and Nadine Walter. 2021. Silence Is Silver, but Speech Is Golden: Intelligent Voice Assistants (IVAs) and Their Impact on a Brand's Customer Decision Journey with a Special Focus on Trust and Convenience-A Qualitative Consumer Analysis in the Netherlands. *IUP Journal of Brand Management* 18, 1 (2021).

[28] Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. 2017. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1927–1936.

[29] Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. 2020. Privileged features distillation at Taobao recommendations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2590–2598.

[30] Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. 2020. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835* (2020).