Extended Conversion: Capturing Successful Interactions in Voice Shopping

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. As a baseline we consider an additional model optimized for Relevance, which is a natural baseline for exploratory traffic where conversions are relatively rare.

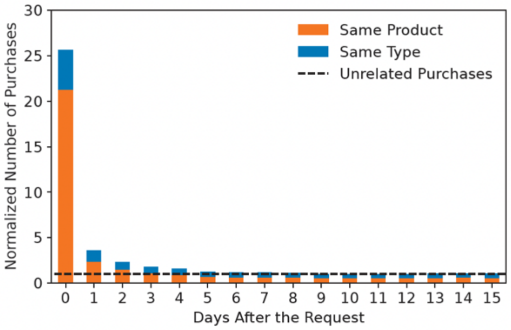

Figure 2: Normalized purchases of the same type, as function of the days from the request, in late shopping stages.

6 DEFINING EXTENDED CONVERSION AS A PARAMETERIZED METRIC

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. In other words, the customers are mostly indifferent between substituting products.
- **Type**: products are considered similar if they belong to
- the same product-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.
- **All**: a trivial similarity level, in which a product is similar to all other products.

REFERENCES

[1] 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] 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] EconML: a Python package for ML-based heterogeneous treatment effects estimation. GitHub (2019).

[4] Conversational product search based on negative feedback.