That only <30% of shoppers visit more than 2 times a year. Because of these constraints, the profit increase due to personalization strategies [10] has been often disappointing. In a context where user intentions are constantly changing, customers are mostly non-recurring and bounce rates are high [36], the competitive pressure towards real-time personalization requires machine learning models to work as early as possible in the customer journey and with as little data as possible. In this perspective, sessions emerge as the natural partition for building models with strong business impact [12, 23, 27, 39, 48] and session-based inference a crucial topic for retailers of all sizes [3, 41].

For practitioners reading about newer models, it is often hard to understand how those advances translate into their production context (if at all [26]), as researchers either use old (e.g. MovieLens [11] in [22, 27, 35]), industry-specific datasets (e.g. playlists [6]), or private datasets that may not be representative of common constraints [42]. For these reasons, we designed the Coveo Data Challenge with two principles in mind:

- advancing the field with the release of a new, fine-grained, session-based dataset. Our dataset (Section 3) presents product-level data from a mid-sized e-commerce, coupled with search-based interactions and content-based information; as such, it is a representative example of thousands of shops in the middle of the industry long tail, with realistic traffic, conversion rate, abandonment ratio;
- proposing inference tasks that are interesting for retailers of all sizes, but where models are evaluated on a mid-size shop. Given the increasing commoditization of tools and libraries, we expect more shops to develop ML capabilities in the coming years: as such, we believe in the value of research with tangible return on investment even outside of retail giants.

The rest of this document describes the data release, the chosen tasks and the evaluation procedure. Moreover, a brief survey of existing work is provided in Section 5.

1.1 About the organizers

The Coveo Data Challenge is organized by a collaboration between academic researchers and Coveo, a multi-tenant provider of
A.I. services, with a network of more than 1000 deployments for customer service, e-commerce and enterprise search use cases.

2 Task Description

For this data challenge, we release a new session-level dataset, containing 10M product interactions over 57K different products (Section 3). Starting from this dataset, we accept solutions addressing two different tasks:

1. A session-based recommendation task, where a model is asked to predict the next interactions between shoppers and products, based on the previous product interactions and search queries within the session;

2. A cart-abandonment task, where, given a session containing an add-to-cart event for a product X, a model is asked to predict whether the shopper will buy X or not in that very session. In particular, we are interested in modelling how abandonment prediction changes 1, 2, ..., n events after the original event, with the idea that the biggest business value comes from early prediction.

3 Data Description

The Coveo Data Challenge features the release of a dataset comprising three main files for training:

- **browsing interactions**: a text file containing fine-grained shopping tracking for more than 4M sessions. Each row is a separate interaction (page view or product interaction); events with the same session_id_hash are from the same session – please note that a product page may generate both a detail and a pageview event, and that the order of the events in the text file is not strictly chronological (refer to the session and timestamp information to reconstruct the actual chain of events for a given session). The fields are similar to the dataset in [33]: session_id_hash, event_type, product_action, server_timestamp_epoch_ms, hashed_url. Descriptive statistics for this file can be found in Table 1;

- **search interactions**: a text file containing more than 800k search-based interactions. Each row is a search query event issued by a shopper, which includes an array of (hashed) results returned to the client. We also provide which result(s) have been clicked from the result set, if any. By also reporting products seen but not clicked, we hope to inspire clever ways to use negative feedback. The fields are: session_id_hash, server_timestamp_epoch_ms, query_vector, product_sku_hash, clicked_sku_hash. The field query_vector is a vectorized representation of the original search query, obtained through standard pre-trained modeling and dimensionality reduction techniques;

- **product meta-data**: a text file containing a mapping between the SKUs (hashed) present in the entire dataset (training+testing) and their respective meta-data (when available), including a hashed representation of product categories, pricing information and vectorized representations of textual and image meta-data. The fields are: product_sku_hash, category_hash, price_bucket, description_vector, image_vector. The categories are hashed representations of a category tree, where each level of hierarchy is separated with a / . The pricing information is provided as a 10-quantile integer. The vectorized fields are a dense representation of the original catalog, obtained through standard pre-trained modeling and dimensionality reduction techniques.

All files are available for download, and additional important information and sample scripts are filed under the project repository. All training data comes from sampling in an unspecified ratio, several months of interactions in the life of Shop Z, a mid-size shop with Alexa Ranking between 25k and 200k; test data is sampled from a disjoint and adjacent time period. Data is fully anonymized and hashed, and collected in compliance with existing legislation: download and usage of the data implies acceptance of the terms and conditions (Section 6).

3.1 Comparison with existing datasets

Based on recent recommendation surveys and frameworks [24, 50] we can classify benchmarks in different categories including ratings (e.g. MovieLens [11], Anime [18], Netflix [29], Book-Crossing [51]), reviews (e.g. Epinions [49], Yelp [45], Steam [21], Amazon [28]), advertising (Criteo [17], Avazu [16]) and e-commerce. E-commerce datasets differ in terms of type of available data. Some only provide transactions (Ta-Feng, Belgium Retail Market, Foodarkt [31]) or product data for product classification [19, 20, 30], whereas others provide also interactions and product metadata (YOOCHOOSE (RCS15) [32], Retailrocket (RETAILR) [34], DIGINETICA [5]). Table 2 compares these three datasets with the data distributed for this challenge.

---

Table 1: Descriptive statistics for the training dataset containing browsing interactions.

| Property               | Value     |
|------------------------|-----------|
| # of SKUs              | 57,483    |
| # of sessions          | 4,934,699 |
| # of all events        | 36,079,307|
| # of all product actions| 10,431,611|
| # of add-to-cart       | 329,557   |
| # of purchases         | 77,848    |
| 25/50/75 pct. of session length | 2/3/8 |

---

5 The definition of "session" is the industry standard from https://support.google.com/analytics/answer/2731565?hl=en
6 The dataset is available at: https://github.com/coveooss/shopper-intent-prediction-nature-2021
7 Please note that removing a product from the cart might lead to multiple remove events, hence pre-processing might be needed with regard to this type of interactions.
8 As a further anonymization technique, the timestamp has been shifted by an unspecified amount of weeks, keeping intact the intra-week patterns.
9 The product categorization is given by the shop – in this case, product experts defined a custom hierarchy to classify products for navigation and marketing purposes.
10 The data challenge is hosted at https://github.com/coveooss/SIGIR-eCom-data-challenge
11 Exact traffic information is thus removed as a further anonymization strategy.
Table 2: Comparison of e-commerce datasets from the literature with the Coveo dataset.

|                        | Coveo | DIGINETICA | RETAILR | RCS15 |
|------------------------|-------|------------|---------|-------|
| Searches               | ✓     | ✓          |         |       |
| Impressions            | ✓     | ✓          |         |       |
| Clicks/Views           | ✓     | ✓          | ✓       | ✓     |
| Carts                  | ✓     | ✓          |         |       |
| Transactions           | ✓     | ✓          | ✓       | ✓     |
| Category               | tree  | single     | tree    | —     |
| Price                  | ✓     | ✓          | —       | ✓     |
| Description            | embedded | hashed   | name   | properties | — |
| Image                  | embedded | ✓        | —       | —     |
| Users                  | —     | 88K        | 1.4M    | —     |
| Items                  | 66K   | 184K       | 235K    | 52K   |
| Events                 | 37M   | 3.3M       | 2.7M    | 42M   |

Considering the variety of the available features and the fine-grained details included in the Coveo Data Challenge, we believe it to be a significant addition to the set of resources in the space. In particular, by releasing neural-friendly embeddings for text, images and search queries, we hope to strike a virtuous trade-off between anonymity and information content.

4 EVALUATION

In this Section, we describe the evaluation procedure and the key metrics that will be used to evaluate submissions in our two tasks\(^8\). While the final ranking will solely reflect quantitative metrics, we encourage papers to include a section on qualitative evaluation, following the suggestions below.

4.1 Quantitative evaluation

(1) Recommendation. We follow the evaluation protocol described in [24]: we reveal the start of a session to the model, and we ask it to predict future interactions. In particular, given a sequence of \(n\) events as input, we use two evaluation schemes:

- the model is evaluated at predicting the immediate next item, using the Mean Reciprocal Rank (MRR) as our metric;
- the model is evaluated at predicting all subsequent items of the session, up to a maximum of 20 after the current event. We take the F1 score from precision@20 and recall@20 as the final metric.

Moreover, we also measure additional quality factors defined in [14, 24]: coverage@20 and popularity bias@20. Finally, note that when a test session includes a search query, this event will be available to the model as an input (it can and should be taken into consideration by the model), but it will be ignored when computing the metrics: in other words, models are evaluated only on their ability to predict future interactions with products.

(2) Cart-abandonment. Models will be evaluated by their micro F1 score in predicting shopping cart abandonment at the first add-to-cart event (AC), and then 2, 4, 6, 8 and 10 events after it.\(^9\) Each model will be evaluated by performing a weighted combination of micro F1 scores at different clicks, with larger weights assigned to earlier predictions according to the following schema:

- micro F1 at AC * 1
- micro F1 2 clicks after AC * 0.9
- micro F1 4 clicks after AC * 0.8
- micro F1 6 clicks after AC * 0.7
- micro F1 8 clicks after AC * 0.6
- micro F1 10 clicks after AC * 0.5

The test set will not undergo any resampling to even the class distribution. Test sessions including purchases will be trimmed before the first purchase event to avoid trivializing the inference.

4.2 Qualitative evaluation

While we do recognize the importance of a standardized quantitative evaluation to have some measure of progress, we also agree with recent criticism of “leaderboard chasing”\(^9\), as there are many practical constraints not captured by metrics which are crucial for the real-world success of a model.

To encourage a deeper understanding of the underlying business problems and foster a healthy competition among models with broad applicability, we solicit submissions that provide new insights by mixing quantitative results with industry-relevant qualitative discussion. As an inspiration, we provide below some examples of interesting topics, divided by task: as a general recommendation, reflections on training cost, inference time and sample efficiency are greatly appreciated, together with connections to related issues in IR for products [2].

(1) Recommendation. We encourage papers to find solutions creatively using search and content data, together with browsing behavior. In particular, we encourage error analysis and comments on model performance for slices of the data that reflects practically important scenarios – new items, rare items, after-query prediction, interplay between search behavior and browsing patterns.

(2) Cart-abandonment. We encourage submissions to include a careful error analysis which considers the relation between classification accuracy at different target clicks and different session characteristics, such as the number of interactions prior to the cart event and the total length of the session [33]. For example, suppose there are 15 actions after the AC in a purchase session: a model which converges on consistently predicting a purchase from action 11 (not captured in our evaluation) would anticipate the right outcome more than a model which is undecided until the second-to-last action but still correctly predicts a purchase at the end of the session. A

\(^8\)The evaluation script can be found in the public repository. Please refer to the challenge website for latest news about important dates, rules and submission format.

\(^9\)By assessing performance at AC, moreover, we want to stress the importance of predicting intent from as little data as relevant.
measure of anticipation would be very informative although it will depend on some free parameters (how strong should the support for a given prediction be, for how many consecutive actions should it remain stable, etc.): are there important insights of this sort that we can uncover leveraging this new dataset?

5 RELATED WORK

Session-based models. Session-based inference is a well-studied topic in e-commerce, as it powers recommendation [4, 7, 15, 44, 47] and NLP/IR use cases [40, 46]: while neural models are now dominating the research agenda [27, 39], there is also some skepticism about the widespread applicability of these techniques and the progress made by the field [25, 26]; in particular, a significant number of recent models still goes through sub-optimal benchmarks [22, 27, 35], which may not be representative of the type of constraints faced by many practitioners [13].

Intent prediction. The ability to predict user intent (add to cart, purchase, abandonment, etc.) from click-stream data is a long-standing challenge for e-commerce, with older solutions modelling user browsing in a Markovian fashion [1] and recent attempts leveraging deep learning architectures [5]. As for in-session recommendations, the broad impact of recent findings is unclear: for example, the architecture proposed in [42] was tested with an unrealistic conversion rate (10x a “standard” shop). More recently, [33] simulated accuracy in deployment scenarios by artificially sub-sampling raw data, showing that models outperform a simple baseline only up to a class imbalance of 80% to 20%. More research into models that scale to extreme class imbalance is thus a priority, and with our release we hope to nudge the community to devote more resources to this problem.

6 LEGAL NOTICE

The Coveo Data Challenge comes with specific Terms and Conditions, as found in the public repository and bundled with the dataset archive: by downloading and accessing the data, you agree to comply with the terms — in addition, by submitting to the challenge, you agree to release your code under an open source license. In particular, the dataset is freely provided to the research community with the purpose of advancing the field: all data was collected and processed in an anonymized fashion, in compliance with existing legislation, and released with additional hashing of meta-data: participants agree to not use the dataset for any other purpose than what is stated in the Terms and Conditions, nor attempt to reverse engineer or de-anonymise the dataset by explicitly or implicitly linking the data to any person, brand or legal entity.

ACKNOWLEDGMENTS

The organizers wish to thank Luca Bigon for his outstanding support in data collection, Surya Kallumadi, Massimo Quadrana, Dietmar Jannach and Ajinkya Kale for the precious feedback on a previous version of this paper. Finally, special thanks to Coveo’s legal team for believing in this data sharing initiative.

REFERENCES

[1] Dimitris J. Bertsimas, Adam J. Marserea, and Nitin R. Patel. [n.d]. Dynamic Classification of Online Customers. https://doi.org/10.1371/journal.pone.0197733.10 arXiv:https://epubs.iamt.org/index.php/PositionPapers/article/view/44
[2] Federico Bianchi, Jacopo Tagliabue, and Bingqing Yu. 2021. Query2Prod2Vec: Grounded Word Embeddings for eCommerce. In NAACL-HLT. Association for Computational Linguistics.
[3] Federico Bianchi, Jacopo Tagliabue, Bingqing Yu, Luca Bigon, and Ciro Greco. 2020. Fantastic Embeddings and How to Align Them: Zero-Shot Inference in a Multi-Shop Scenario. In Proceedings of the SIGIR 2020 eCom workshop. July 2020. https://arxiv.org/abs/2007.14906
[4] Federico Bianchi, Bingqing Yu, and Jacopo Tagliabue. 2020. BERT Goes Shopping: Comparing Distributional Models for Product Representations. arXiv preprint arXiv:2012.09807 (2020).
[5] Luca Bigon, Giovanni Cassani, Ciro Greco, L. Lacasa, Mattia Pavoni, Andrea Polonioli, and Jacopo Tagliabue. 2019. Prediction is very hard, especially about conversion. Predicting user purchases from clickstream data in fashion e-commerce. ArXiv abs/1907.00400 (2019).
[6] Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. 2018. Word2vec Applied to Recommendation: Hyperparameters Matter. In Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18). Association for Computing Machinery, New York, NY, USA, 352–356. https://doi.org/10.1145/3240323.3240377
[7] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2019. A Dynamic Co-Attention Network for Session-Based Recommendation. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (Beijing, China) CIKM ’19. Association for Computing Machinery, New York, NY, USA, 1461–1470. https://doi.org/10.1145/3357384.3357964
[8] CIKM-Cup. 2016. Cikm cup 2016 track 2: Personalized e-commerce search challenge. https://competitions.codalab.org/competitions/11161
[9] Kasim Ethabrajah and Dan Jurafsky. 2020. Utility Is in the Eye of the User: A Critique of NLP Leaderboard Design. ArXiv abs/2009.13888 (2020).
[10] Penny Gillespie, Jason Daigler, Mike Lowndes, Kristina Klick, Yanna Dharmasthira, and Sandy Shen. 2018. Magic Quadrant for Digital Commerce. Technical Report. Garten.
[11] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Trans. Interact. Intell. Syst. 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872
[12] Balázs Hidasi, Alexandros Karatzoglou, L Baltrunas, and D Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. CoRR abs/1511.06939 (2016).
[13] Dietmar Jannach, Gabriel de Souza P. Moreira, and Even Oldridge. 2020. Why Are Deep Learning Models Not Consistently Winning Recommender Systems Competitions Yet? A Position Paper. In Proceedings of the Recommender Systems Challenge 2020 (RecSysChallenge’20). Association for Computing Machinery, New York, NY, USA, 44–49. https://doi.org/10.1145/3355864.3357336
[14] Dietmar Jannach and Malte Ludewig. 2017. When recurrent neural networks meet the neighborhood for session-based recommendation. In Proceedings of the Eleventh ACM Conference on Recommender Systems. 306–310.
[15] Yu jia Zheng, Siyi Liu, Zekun Li, and S. Wu. 2020. DGTN: Dual-channel Graph Transition Network for Session-based Recommendation. ArXiv abs/2009.16002 (2020).
[16] Kaggle. 2014a. Arzava click-through rate prediction. https://www.kaggle.com/c/avazu-ctr-prediction/data
[17] Kaggle. 2014b. Display advertising challenge. https://www.kaggle.com/c/criteo-display-ad-challenge/data
[18] Kaggle. 2016a. Anime Recommendations Database. https://www.kaggle.com/CooperUnion/anime-recommendations-database
[19] Kaggle. 2016b. e Commerce Item Data. https://www.kaggle.com/cclark/product-item-data/home
[20] Kaggle. 2017. Innerwear Data from Victoria’s Secret and Others. https://www.kaggle.com/promptcloud/innerwear-data-from-victorias-secret-and-others
[21] Wang-Cheng Kang and Julian McAusley. 2018. Self-attentive sequential recommendation. In 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 197–206.
[22] Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 1748–1757.
[23] Qiao Liu, Yifu Zeng, Refuse Mokhosi, and Haibin Zhang. 2018. STAMP: Short-Term Attention/Memory Priority Model for Session-based Recommendation. 1831–1839. https://doi.org/10.1145/3219819.3219950
[24] Malte Ludewig and Dietmar Jannach. 2018. Evaluation of session-based recommendation algorithms. User Modeling and User-Adapted Interaction 28, 4-5 (2018), 331–390.
[25] Malte Ludewig, Noemi Mauro, Sara Latifi, and D. Jannach. 2019. Empirical Analysis of Session-Based Recommendation Algorithms. ArXiv abs/1910.12781

Tagliabue et al.
