Delivery Hero Recommendation Dataset: A Novel Dataset for Benchmarking Recommendation Algorithms

YERNAT ASSYLBEKOV*, RAGHAV BALI*, LUKE BOVARD*, CHRISTIAN KLAUE*, Delivery Hero, Germany

In this paper we propose Delivery Hero Recommendation Dataset (DHRD), a novel real-world dataset for researchers. DHRD comprises over a million food delivery orders from three distinct cities, encompassing thousands of vendors and an extensive range of dishes, serving a combined customer base of over a million individuals. We discuss the challenges associated with such real-world datasets. By releasing DHRD, researchers are empowered with a valuable resource for building and evaluating recommender systems, paving the way for advancements in this domain.

ACM Reference Format:
Yernat Assylbekov*, Raghav Bali*, Luke Bovard*, Christian Klaue*. 2023. Delivery Hero Recommendation Dataset: A Novel Dataset for Benchmarking Recommendation Algorithms. 1, 1 (July 2023), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the importance of machine learning in industrial applications, coupled with the correlated increase of research publications and the success of deep learning techniques in fields like NLP and computer vision, one would expect that revolutionary progress resulting from these publications would be reflected in the field of Recommendation Systems (RS). While latest state-of-the-art approaches claim to outperform traditional methods, further investigations reveal that this is not the case, and fine-tuned simple ideas still produce comparable results [9].

The reasons for these findings point to a number of issues including poor reproducibility, low sample size, specific preprocessing techniques which do not generalise. Differences in training, validation and test split as well as removal of samples can be observed in papers using the popular MovieLens dataset (e.g., [14] vs [7]). Enabling researchers and practitioners to easily compare and reproduce results across works by providing them with good quality evaluation datasets is one step in the correct direction to enhance research.

A dataset which is used to evaluate different approaches should have the following attributes:

(1) **Accessibility**: The dataset should be open-sourced, easily accessible and well documented.

*These authors contributed equally to this work

Author's address: Yernat Assylbekov*, Raghav Bali*, Luke Bovard*, Christian Klaue*, Delivery Hero, Berlin, Germany, yernat.assylbekov@deliveryhero.com,raghav.bali@deliveryhero.com,luke.bovard@deliveryhero.com,christian.klaue@deliveryhero.com.

Manuscript submitted to ACM

(2) **Relevance**: The dataset should contain relevant features that are important for the analysis and modelling task at hand.
(3) **Quality**: The data should be of high quality, with minimal missing values, errors, or inconsistencies.
(4) **Reproducibility**: The training and test data should be clearly defined to ensure comparisons.
(5) **Representativeness**: While maintaining relevance and quality, the data should also reflect challenges faced in real-world settings (eg: inconsistent labels, changing identifiers, size/quantity information mixed with product names, etc.)
(6) **Multimodality**: The dataset should consist of different data types like text or images to maximise the horizon of potential applicable approaches.
(7) **Volume**: The dataset should be large enough to provide a representative sample of the population or phenomena being studied.

In order to enable researchers to correctly evaluate and compare approaches, it is fundamental to provide them with real world data samples. Given those samples and resulting research success, the outcome can then be used in the industry resulting in a fruitful cooperation. Therefore, it is the responsibility of companies to produce this real-world data and make it accessible to researchers. For a large proportion of the industries there exist datasets which are widely used and considered representative. A few examples of such areas include music (e.g., Spotfiy [13], lastfm [2]), movies (e.g., MovieLens [3], IMDb[5], Netflix [10]) and retail (e.g., Amazon[11], Tmall [15]).

In contrast, there are few datasets that relate to food recommendation and this domain offers many interesting challenges that have been under-represented in literature and, to the best of our knowledge, none of the leading food delivery companies across the world have officially released datasets for research purposes. There are datasets available that have a subset of food related data, but they either are a very small percentage of the whole dataset [6] or represent only a specific geographical area like US [16] or have been scraped [1, 8] with unspecified usage and license agreements.

Food recommendation has a variety of topics that make it a fruitful field of research and investigation, including, but not limited to delivery time, quality and fee, availability, nutritional breakdown, cuisine type, group orders, dietary considerations, as well as cultural, social and environmental factors. Recommendations within these parameters are also complex since dishes can be very different depending on the restaurants or a customer undergoes a dietary change and low-quality recommendations have long-term consequences. Additionally, food items are often time dependent, sometimes even ordered together and therefore influence each other.

2 DATASET DESCRIPTION

To address the concerns raised in the previous section, we have constructed a dataset from Delivery Hero's largest brands such as foodpanda and foodora. Delivery Hero brands process more than 5 million orders daily with operations in more than 400 cities [4]. This dataset, which encompasses millions of real world orders, will help researchers better evaluate recommendation algorithms in the food recommendation sector.

The data we are releasing here[1] covers a time period of three months with data consisting of 4.5 million orders processed on Delivery Hero platforms for three cities: Singapore, Taipei, and Stockholm. The data

[1] https://github.com/deliveryhero/dh-reco-dataset

consists of three parts for each location: orders, products, and vendor information. Orders data contains information about orders being placed, product data contains data about the products offered on a vendor's menu, and vendor data contains metadata about the vendor. Table 1 contains an overview of the columns in the dataset and table 2 presents summary statistics of the dataset. To ensure privacy and GDPR compliance, we have removed all personally identifiable information and the data has been anonymized through hashing with salt. Additionally, this data may not reflect the present availability of products and vendors in those cities. The normalised prices listed are based on those that the user would see and contain no proprietary discount, incentives, or commissions related information as those are beyond the scope of this data release. Due to privacy and business restrictions, we are not able to provide user demographic or vendor popularity information in this data release.

We have used a geohash [12] to approximate user/vendor locations. Using vendors that are within the same geohash / neighbouring geohashes is an effective way to determine all available vendors for a given user without real-time data. Internally, recommendation models use similar approximations and we have found it to give good results.

In contrast to the majority of open source datasets that are pre-processed, clean, and carefully curated to eliminate real-world challenges, this dataset emulates real-world settings. Our dataset comprises missing values, inconsistent IDs, and other typical issues encountered in real-world systems. The main motivation behind keeping the dataset as raw as possible were to enable the community to work towards novel pre-processing techniques, experience real-world issues, develop novel ranking and recommendation methods and also ensure that we keep any known biases while releasing such a dataset to the minimum. We list below a few potential challenges that we encourage researchers to explore.

(1) **Non-Trivial Cuisines**: Although e.g., Singapore offers an extensive variety of 78 cuisines, a detailed examination of these cuisines indicates a very interesting scenario. For instance, rice, noodles, tea, and sandwiches are some of the top primary cuisines listed in the dataset. Even though these are marked as primary cuisines, intuitively these do not correspond to typical understanding of cuisines. Another typical case is similar cuisines which cannot be directly treated as one. For example, bubble tea, coffee, and beverages could be grouped together as a single primary cuisines (say beverages), whereas identifying a common category for sandwiches and burgers could prove challenging.

(2) **Product Names**: Product/Dish/Beverage names play an important role in understanding user preferences and tastes. Vendors also benefit from understanding what products are in demand and which ones aren't to manage inventories, incentives and more. Vendors also use creative names, marketing labels, latest trending terms, and add quantity/size qualifiers to their offerings to make them more attractive to customers. This creates unique challenges to utilise this information for understanding ordering patterns, preferences and presents an unique opportunity for developing pre-processing techniques using NLP and domain understanding.

(3) **Unordered Products**: Another unique challenge is the difference between the available list of products in the system versus products being ordered. For instance, in Taipei there are details for 810K unique products out of which only 327K have associated orders (a mere 40% of the total). This situation frequently occurs due to a number of factors involving changes to product names (spelling corrections, marketing labels, etc.), portion/quantity changes, price modification, seasonal

Column	Description	Dataset
customer_id	uniquely hased customer_id	Orders
order_time	time order was placed	Orders
day_of_week	day of the week the order was placed, with 0 = Sunday,	Orders
order_day	day from the start of the dataset when the order was placed	Orders
geohash	hashed geographic location of where the order was placed given to 5 digits of precision	Orders, Vendors
order_id	unique order_id	Orders
vendor_id	unique hashed vendor_id of where the order was placed	Orders, Vendors, Product
chain_id	unique hashed identifier of the chain that a vendor belongs to. Not all vendors belong to chain and if null the vendor is not part of a chain	Vendor
geohash	hashed geographic location of where the order was placed given to 5 digits of precision	Orders, Vendor
primary_cuisine	vendor specified cuisine	Vendor
product_id	unique hashed product_id of the product ordered	Orders, Product
name	the product name as displayed on the menu	Product
unit_price	the normalised unit price of the item.	Product

Table 1. Schema of the three different dataset tables.

City	# of Orders	# of Customers	# of Vendors	# of Products	# of Products Ordered
Singapore	1.99M	512K	7411	1.06M	256K
Stockholm	400K	122K	1148	111K	41K
Taipei	2.0M	741K	9506	810K	327K

Table 2. Summary statistics of the dataset for each of the three cities

menu changes and more. This requires meticulous exploration of the dataset to come up with novel preprocessing techniques for different downstream tasks (not limited to recommendations alone). Multiple languages add an additional layer of complexity to such methods.

3 CONCLUSION

This paper presents the DHRD, a novel online food delivery dataset, which serves to expand and enhance the existing collection of public datasets for recommendation tasks across diverse domains. The dataset offers authentic information with minimal preprocessing to preserve real-world characteristics and inspire innovative approaches for preprocessing and recommendation techniques. It is conveniently available in a user-friendly format[2], accompanied by comprehensive details and descriptions outlined in this paper. While we anticipate that this dataset will introduce unique challenges to the research community, our objective is to continually enhance and enrich it through subsequent versions by incorporating additional features, modalities, and volume.

[2]https://github.com/deliveryhero/dh-reco-dataset

4 SPEAKER BIO AND ACKNOWLEDGEMENTS

The authors of this paper are a team of experienced Data Scientists who have been actively engaged in data science and related activities at Delivery Hero. Their primary focus has been on enhancing customer recommendations across diverse markets. Yernat and Luke, possessing doctorate degrees in Mathematics and AstroPhysics respectively, successfully completed their doctoral studies at the University of Washington and Goethe University Frankfurt. Raghav Bali, holding a Masters from IIIT-Bangalore, brings to the team over a decade of valuable experience gained from working with major technology companies. Christian Klaue, equipped with a Masters from the University of St. Thomas, boasts a versatile professional background spanning various domains. The authors extend their heartfelt gratitude to Ankur Kaul, Koray Kayir, and the entire squad for their unwavering support and invaluable inputs throughout this work. Their support and guidance have been instrumental in shaping the outcomes of this study.

REFERENCES

[1] Grubhub restaurant data. https://www.kaggle.com/datasets/polartech/grubhub-restaurant-data, 2022.
[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
[3] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.
[4] Delivery Hero. Q3 2021 trading update. https://web.archive.org/web/20211111082326/https://ir.deliveryhero.com/download/companies/delivery/Presentations/20211111_Delivery_Hero_SE_Trading_update_3Q21.pdf, 2021.
[5] IMDb. Imdb non-commercial datasets. https://developer.imdb.com/non-commercial-datasets/. Accessed: 2023-05-31.
[6] Michael Kechinov. ecommerce behavior data from multi category store. https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store, 2020.
[7] Malte Ludewig and Dietmar Jannach. Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, 28(4-5):331–390, oct 2018.
[8] Shruti Mehta. Zomato restaurants data. https://www.kaggle.com/datasets/shrutimehta/zomato-restaurants-data, 2018.
[9] Lorenz Muller, Julien Martel, and Giacomo Indiveri. Kernelized synaptic weight matrices. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3654–3663. PMLR, 10–15 Jul 2018.
[10] Netflix. Netflix prize data set. 2009.
[11] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics.
[12] G. Niemeyer. Geohash. https://web.archive.org/web/20180309054335/https://forums.geocaching.com/GC/index.php?%2Ftopic%2F186412-geohashorg%2F, 2008.
[13] Spotify R&D. Datasets. https://research.atspotify.com/datasets/. Accessed: 2023-05-31.
[14] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, 2019.
[15] Tmall.com. Repeat buyers prediction competition. https://ijcai-15.org/repeat-buyers-prediction-competition/, 2015. Accessed: 2023-05-31.
[16] Yelp. Yelp open dataset. https://www.yelp.com/dataset, 2015.