A Dataset for Learning Graph Representations to Predict Customer Returns in Fashion Retail

Jamie McGowan
j.mcgowan.18@ucl.ac.uk
University College London
London, UK

Elizabeth Guest
elizabeth.guest.21@ucl.ac.uk
University College London
London, UK

Ziyang Yan
ziyang.yan.17@ucl.ac.uk
University College London
London, UK

Zheng Cong
zheng.cong.20@ucl.ac.uk
University College London
London, UK

Neha Patel
neha.patel@asos.com
ASOS AI
London, UK

Mason Cusack
mason.cusack@asos.com
ASOS AI
London, UK

Charlie Donaldson
charlie.donaldson@asos.com
ASOS AI
London, UK

Sofie de Cnudde
sofiede.cnudde@asos.com
ASOS AI
London, UK

Gabriel Facini
g.facini@ucl.ac.uk
University College London
London, UK

Fabon Dzogang
fabon.dzogang@asos.com
ASOS AI
London, UK

ABSTRACT

We present a novel dataset collected by ASOS (a major online fashion retailer) to address the challenge of predicting customer returns in a fashion retail ecosystem. With the release of this substantial dataset we hope to motivate further collaboration between research communities and the fashion industry. We first explore the structure of this dataset with a focus on the application of Graph Representation Learning in order to exploit the natural data structure and provide statistical insights into particular features within the data. In addition to this, we show examples of a return prediction classification task with a selection of baseline models (i.e. with no intermediate representation learning step) and a graph representation based model. We show that in a downstream return prediction classification task, an F1-score of 0.792 can be found using a Graph Neural Network (GNN), improving upon other models discussed in this work. Alongside this increased F1-score, we also present a lower cross-entropy loss by recasting the data into a graph structure, indicating more robust predictions from a GNN based solution. These results provide evidence that GNNs could provide more impactful and usable classifications than other baseline models on the presented dataset and with this motivation, we hope to encourage further research into graph-based approaches using the ASOS GraphReturns dataset.

CCS CONCEPTS

• **Fashion Retail Dataset → Classification**; Customer Return Prediction; • **Graph Representation Learning → Neural message passing**; Edge Classification.

KEYWORDS

Recommendation Systems, Fashion Industry, e-commerce

ACM Reference Format:
Jamie McGowan, Elizabeth Guest, Ziyang Yan, Zheng Cong, Neha Patel, Mason Cusack, Charlie Donaldson, Sofie de Cnudde, Gabriel Facini, and Fabon Dzogang. 2022. A Dataset for Learning Graph Representations to Predict Customer Returns in Fashion Retail. In *Sixteenth ACM Conference on Recommender Systems (FashionXRecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Part of the unique digital experience that many fashion retailers deliver is the option to return products at a small or no cost to the customer. However, unnecessary shipping of products back and forth incurs a financial and environmental cost. With many fashion retailers having a commitment to minimizing the impact of the fashion industry on the planet, providing a service which can forecast returns and advise a customer of this at purchase time is in line with these goals.

With the continual development of e-commerce platforms, it is important that systems are able to model the user's preferences within the platform's ecosystem by using the available data to guide users and shape the modern customer experience. One approach to this challenge, which has sparked huge interest in the field of recommendation systems [15], are representation learning based

FashionXRecSys '22, September 18–23, 2022, Seattle, WA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

methods. Representation learning provides a framework for learning and encoding complex patterns present in data, which more naive machine learning (ML) approaches are unable to capture as easily. However at present, the available data that is able to facilitate such research avenues is scarce. Further to this, the number of available datasets which include anonymised customer and product information (and their interactions) is even less available.

E-commerce platforms in the fashion industry are in a unique position to contribute to this research by making data publicly available for use by the machine learning community. Of particular interest to ASOS is the application of machine learning to predicting customer returns at purchase time, due to this, we present the ASOS GraphReturns dataset in this article. The labelled purchase (return or not returned) connections between customers and products in this dataset naturally lends itself to a graph structure which has motivated our interest in encouraging the exploration of graph representation learning based solutions, which we provide an example of in Sect. 4. Graph Neural Networks (GNNs) have been the subject of immense success in recent years [3, 4, 10, 13, 14] and provide an intuitive way to exploit structured data. Another benefit of using GNNs is that they are able to make predictions for new instances not seen before. This is a particular attractive feature for industry environments where new products and customers are continually added.

In this work, we first present the ASOS GraphReturns dataset[1] and discuss some of the properties and features of this data. Using this data we then provide some examples demonstrating the use of GNNs with this data based on the downstream task of predicting customer returns. This information may then be used to inform customers based on their choice and make a personalised recommendation (i.e. a different size, style, colour etc.) at purchase time that has a lower probability of being returned.

The structure of the document is as follows: Sect. 2 describes the novel fashion retail dataset, Sect. 3 overviews the methodology and some example benchmark results are discussed in Sect. 4. Finally in Sect. 5 we summarise this contribution and provide some insights into potential further studies which could benefit from this dataset.

2 DATA DESCRIPTION

The train (test) data contains purchases and returns recorded by ASOS between Sept-Oct 2021 (Oct-Nov 2021), including the corresponding anonymous customer and product variant[2] specific information. The data is organised into customers (with hashed customer ID's to preserve anonymity), product variants and events (i.e. a purchase or return of a product by a customer). The training (testing) dataset includes $\sim 770,000$ ($\sim 820,000$) unique customers and $\sim 410,000$ ($\sim 410,000$) product variants, where every customer has at least one return and each product variant has been purchased at least once. To connect customers and products the data contains a total of 1.4M (1.5M) purchase events each labeled as a return (1) or no return (0) in both the training and testing datasets. The problem of predicting customer returns is then presented as an edge classification task as depicted in Fig. 1. This structure is similar to that of the Amazon reviews data [9] which also includes labeled links between customers and products.

Within each customer/product variant node, we also include specific node features, such as the average return rate, the ratios of different reasons for returns, and historical information relating to the number of purchases/returns made. Fig. 1 displays an exhaustive list of all the features included in this dataset. Fig. 2 (left) displays a subset of correlations between customer (top) and product (bottom) features. Within these correlations, one can observe strong associations such as male customers being less likely to make a return or a more expensive product in general having a higher return rate. Fig. 2 (right) summarises a selection of statistics related to the distribution of return labels across countries and brands included within the data. It can be seen that the data shows a larger proportion of returns across specific individual markets which could prove useful in ML based classification tasks[3].

Of particular interest to neural message passing techniques is the inherent graph structure that this dataset holds. In order to apply graph neural networks to data, one must first arrange the data into nodes that contain specific features and edges that link these node instances. This extra potential structure that can be constructed from the ASOS GraphReturns dataset further enhances the modality of the data from the raw structure and node features/correlations discussed above. In Fig. 3, we show the data in an undirected heterogeneous graph structure with 5 different edge types linking customers to their shipping countries and product variants to each other and their corresponding brands, product types and top return reasons by defining intermediate virtual nodes in all cases. These virtual nodes can be constructed in multiple ways, however in this paper the virtual nodes contain an averaged set of features for each instance i.e. a product type node will contain the average set of feature values for all products linked to this node.

3 METHODOLOGY

In this section, we present the methodology for a number of example baseline methods applied to the task of predicting customer returns in Sect. 4. The methods considered here aim to provide an early benchmark for future studies involving this dataset. For the graph representation learning based approach, the data is arranged into a highly connected structure with virtual nodes for: customer shipping countries, products, product types, product brands and top return reasons for product variants as described in Fig. 3.

We investigate the use of a Logistic Regression, a 2-layer MLP, a Random Forest [1], and an XGBoost [2] classifier trained directly on the raw data (i.e. not arranged into a graph) described in Sect. 2. For these models, the customer and product specific features are joined by each labelled purchase link in the data. Further to this, we also investigate a benchmark for a GNN based model trained in conjunction with the same baseline 2-layer MLP as a classifier head. In this case the output of the GNN is the learnt embeddings and the MLP provides a final classification layer for the downstream tasks.

To construct an embedding for an edge $\mathbf{e}_{ab}$ between two nodes a and b, in general one can perform an operation involving both

[1] The dataset can be found at https://osf.io/c793h/.

[2] Note that product variants include variations in size and colour and therefore a product may contain multiple variants.

[3] Due to the manner in which this dataset is constructed (i.e. only including customers who have at least one return), these statistics do not reflect the true ASOS purchase/return statistics.

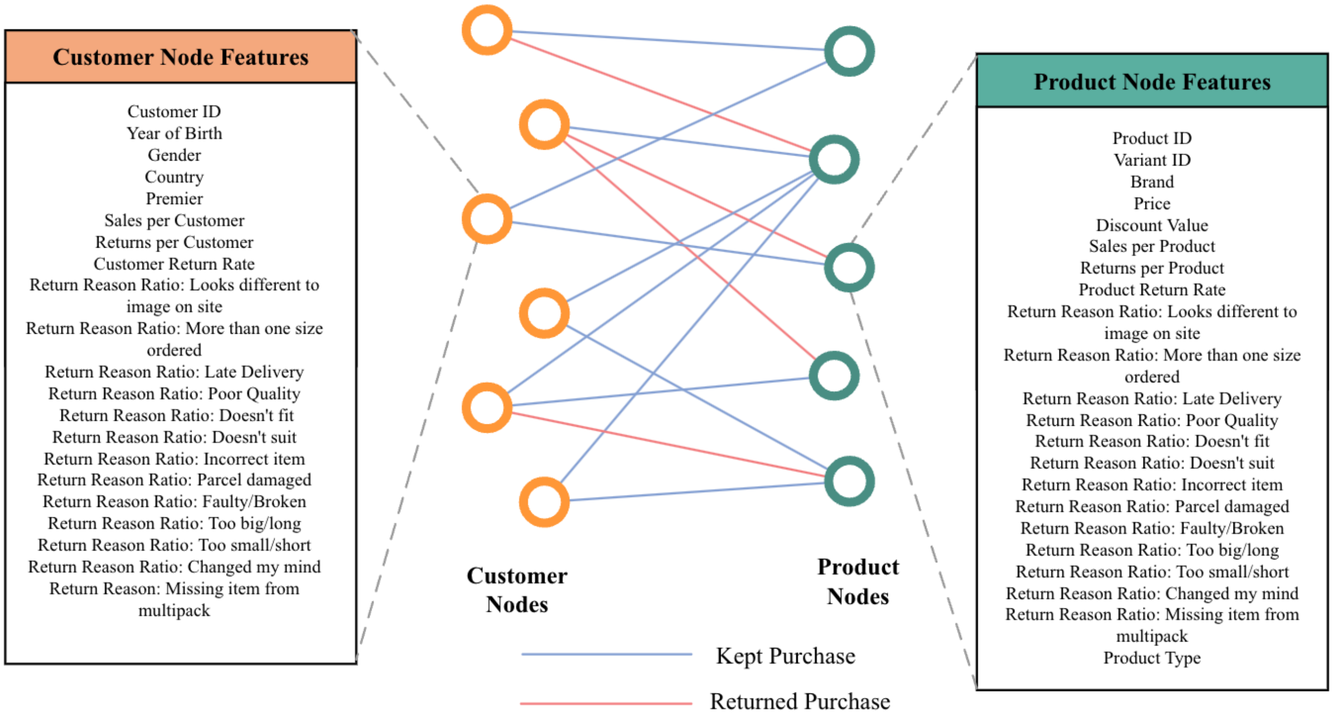
Customer Node Features
Customer ID
Year of Birth
Gender
Country
Premier
Sales per Customer
Returns per Customer
Customer Return Rate
Return Reason Ratio: Looks different to image on site
Return Reason Ratio: More than one size ordered
Return Reason Ratio: Late Delivery
Return Reason Ratio: Poor Quality
Return Reason Ratio: Doesn't fit
Return Reason Ratio: Doesn't suit
Return Reason Ratio: Incorrect item
Return Reason Ratio: Parcel damaged
Return Reason Ratio: Faulty/Broken
Return Reason Ratio: Too big/long
Return Reason Ratio: Too small/short
Return Reason Ratio: Changed my mind
Return Reason: Missing item from multipack
Customer Nodes
Product Nodes
Kept Purchase
Returned Purchase
Product Node Features
Product ID
Variant ID
Brand
Price
Discount Value
Sales per Product
Returns per Product
Product Return Rate
Return Reason Ratio: Looks different to image on site
Return Reason Ratio: More than one size ordered
Return Reason Ratio: Late Delivery
Return Reason Ratio: Poor Quality
Return Reason Ratio: Doesn't fit
Return Reason Ratio: Doesn't suit
Return Reason Ratio: Incorrect item
Return Reason Ratio: Parcel damaged
Return Reason Ratio: Faulty/Broken
Return Reason Ratio: Too big/long
Return Reason Ratio: Too small/short
Return Reason Ratio: Changed my mind
Return Reason Ratio: Missing item from multipack
Product Type

Figure 1: The raw data structure includes customer and product specific information linked by purchases. These purchase links are labeled with a no return (blue) or return (red) label. The entire list of node features for customers and products is also provided here.

representations for each node,

$$\mathbf{e}_{ab} = \mathcal{O}\left(\mathbf{h}_a^{(K)}, \mathbf{h}_b^{(K)}\right). \tag{1}$$

where in the case described above, $\mathcal{O}$ is described as a 2-layer MLP classifier which performs the final classification from the output of the GNN.

The output of the MLP classifier head is then the predicted probability for the two class labels (return or no return) which are fed into the cross entropy (CE) loss [6]:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i) \tag{2}$$

where N is the total number of predictions, y_i is the true class label (i.e. 0 or 1 for binary classification) of instance i and p_i is the predicted probability for the observation of instance i. Here we note that the CE loss takes into account the probability of each classification, whereas the F1-score only considers the final classification label. Therefore it is an important metric to consider when one is interested in robust predictions, as is needed for an effective fashion industry solution for reducing the number of returns.

In order to train the GNN discussed in the following section, an extra step is included into this methodology whereby the purchase events are only trained on if the product variant involved has an average return rate of higher than 80% or lower than 20%, in order to provide more robust positive and negative examples of return instances to the GNN. The reason for this is to investigate and avoid issues involving oversmoothing in the representations learnt by the GNN, however all results are quoted on the entire test set with no filtering. The result of this is a dataset with 200,000 purchase events and an average vertex degree for the real nodes of 5 for product variant nodes and 2 for customer nodes.

4 EXPERIMENT RESULTS

Model	Test Scores			
	Precision	Recall	F1-score	CE Loss $\mathcal{L}_{\text{CE}}$
Logistic Regression	0.723	0.726	0.725	0.602
Random Forest	0.788	0.712	0.748	0.630
MLP	0.870	0.656	0.748	0.582
XGBoost	0.805	0.745	0.774	0.561
GNN	**0.816**	**0.758**	**0.792**	**0.489**

Table 1: Results for models considered in this section evaluated on the full test data.

Table 1 displays the precision, recall and F1-scores each model evaluated on the full test dataset (1.5M purchase events). The final hyperparameter values are chosen based on a validation set, randomly and uniformly constructed from 10% of the training data and are listed as: Logistic Regression (C = 5.0, tol. = 10^{-4}), MLP (# of layers = 2, hidden dim. = 128), Random Forest (# of estimators = 100,

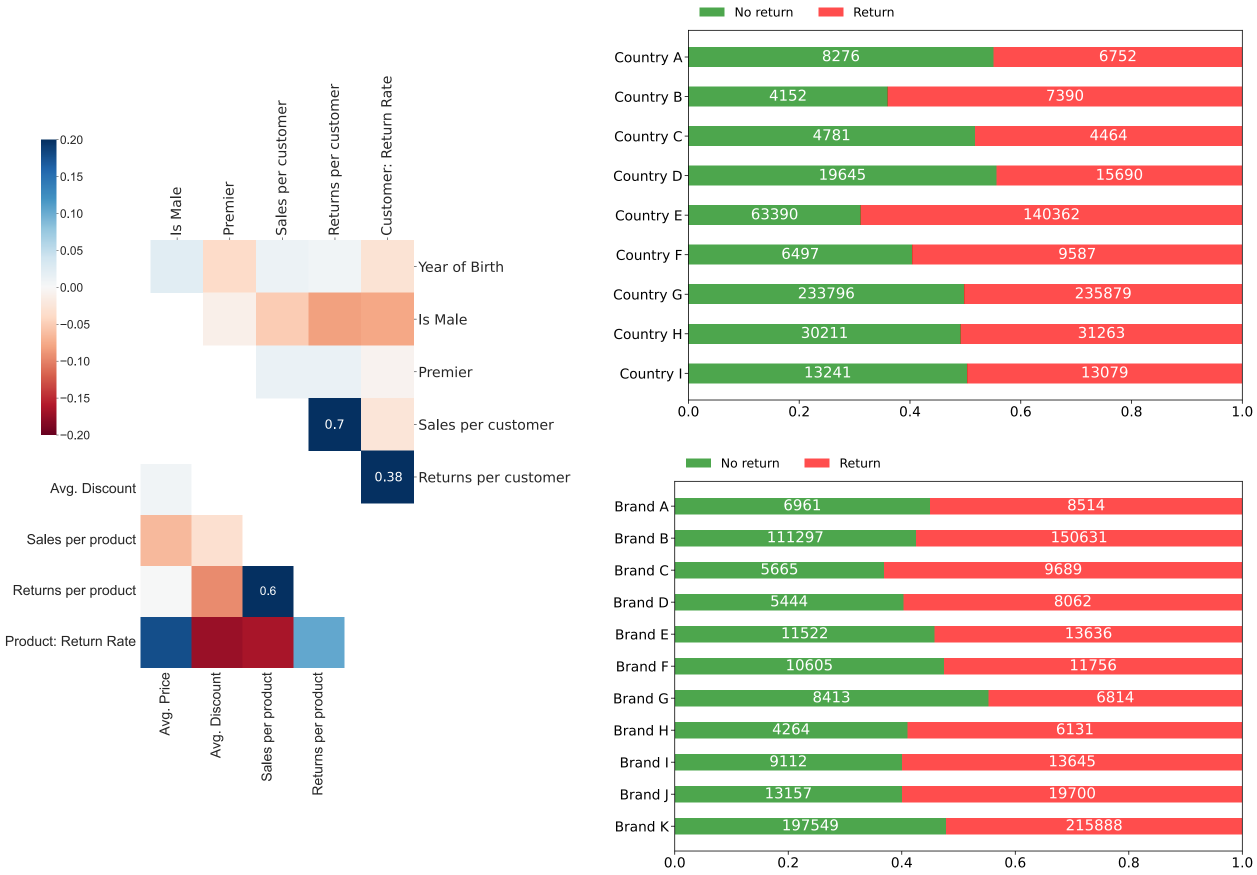

Figure 2: General summary of data statistics including correlations between customer and product specific features (left) and distributions of return labels (right) within each country (top) and brand (bottom).

max. depth = 6, min. samples split = 2, min. samples leaf = 1, max. leaf nodes = 10), XGBoost [2] (booster = gbtree, max. depth = 4, η = 0.1, γ = 1, min. child weight = 1, λ = 2, objective = Binary Logistic, early stopping rounds = 5), GNN (1 GraphSAGE [8] layer with dim. = 16, all aggregations = max. pool, dropout = 0.2, normalise = True)[4]. For the MLP (16,641 trainable parameters) and GNN (49,665 trainable parameters) models, an Adam optimizer is used with a learning rate of 0.01.

The results in Table 1 show a superior performance for a GNN based approach trained on high and low returning examples (described in Section 3) across all metrics considered, indicating that a graph-based approach yields a better performing and more robust classification model. For reference, when comparing the same GNN to one trained on all available data, an F1-score of 0.783 was found, suggesting the GNN's performance may suffer from oversmoothing when being trained on less discrete positive and negative examples. Furthermore, as mentioned in Sect. 3, the classifier head attached to the GNN is the same MLP model also present in Table 1, therefore supporting the expectation that the graph embeddings from the GNN are able to encode useful information from the data. Table 1 also suggests that the GNN's predictions are more robust, based on a lower final CE loss (Equation (2)) combined with a higher F1-score evaluated on the test set.

Table 2 displays the F1-scores evaluated on the test set for individual country markets. In all country instances, the GNN based approach obtains a superior F1-score to all other models considered. When comparing the results in these tables with the correlations discussed in Fig. 2 one can observe that those countries with higher correlations to a particular return label (1 or 0) are among the top performing F1-scores in Table 2.

Single market results are of particular interest to the wider e-commerce fashion industry in order to understand how to deliver the best service to customers and products across different individual markets. The ability to obtain results such as these are an important and unique feature in the novel ASOS GraphReturns dataset as it facilitates a level of understanding into how an ML model is performing across different areas and identify it's weaknesses. Note that a similar analysis can be done for different brands or product types.

[4] Any parameters not listed here are left at their default values provided by the packages `sklearn` [12] (Logistic Regression & Random Forest), `xgboost` [2] (XGBoost), `PyTorch` [11] (MLP). and `PyG` [5] (GNN).

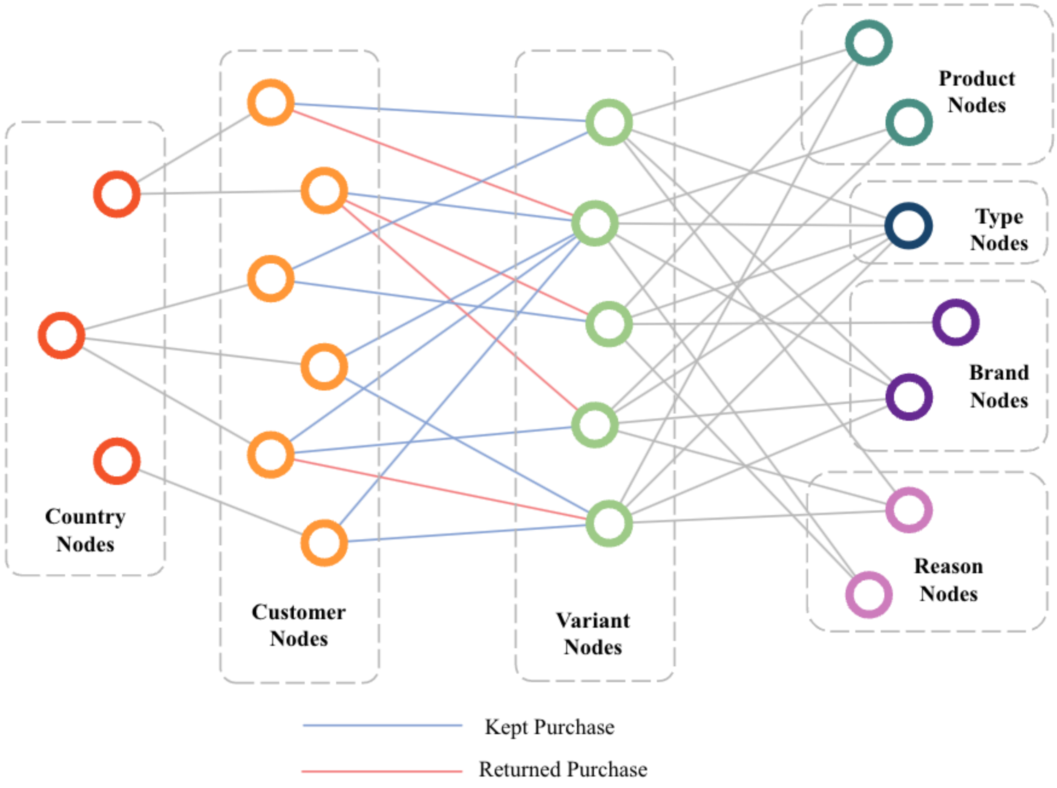

Figure 3: Representation of the richer graph structure contained within the ASOS returns data and how it can be recast into a form better suited to graph representation learning. Virtual nodes are shown for countries, products, product types, brands and return reasons with extra connections added to each customer and product variant node.

Model	Country A		Country B		Country C		Country D	
	Trained on all markets							
	F1-score	$\mathcal{L}_{\text{CE}}$	F1-score	$\mathcal{L}_{\text{CE}}$	F1-score	$\mathcal{L}_{\text{CE}}$	F1-score	$\mathcal{L}_{\text{CE}}$
Logistic Regression	0.635	0.611	0.776	0.606	0.658	0.611	0.593	0.608
Random Forest	0.655	0.633	0.785	0.633	0.672	0.635	0.606	0.633
MLP	0.680	0.527	0.792	0.527	0.691	0.528	0.626	0.518
XGBoost	0.731	0.556	0.806	0.567	0.717	0.567	0.664	0.561
GNN	**0.757**	**0.436**	**0.821**	**0.487**	**0.744**	**0.485**	**0.732**	**0.494**

Model	Country E		Country F		Country G		Country H	
	Trained on all markets							
	F1-score	$\mathcal{L}_{\text{CE}}$	F1-score	$\mathcal{L}_{\text{CE}}$	F1-score	$\mathcal{L}_{\text{CE}}$	F1-score	$\mathcal{L}_{\text{CE}}$
Logistic Regression	0.812	0.591	0.729	0.618	0.673	0.605	0.671	0.610
Random Forest	0.817	0.624	0.745	0.638	0.717	0.630	0.683	0.636
MLP	0.819	0.514	0.754	0.542	0.727	0.520	0.696	0.528
XGBoost	0.827	0.561	0.772	0.573	0.751	0.561	0.728	0.563
GNN	**0.842**	**0.487**	**0.801**	**0.500**	**0.774**	**0.489**	**0.744**	**0.505**

Table 2: Summary of F1-scores and CE losses ($\mathcal{L}_{\text{CE}}$) evaluated on the test data for each individual country market. In these results we use a GNN model with 1 SAGEGraph layer (dim. = 16) trained with all extra nodes considered from Sect. 3.

5 CONCLUSION

In this work we have presented a novel dataset to inspire new directions in fashion retail research. This dataset is particularly suited to graph representation learning techniques and exhibits a naturally rich geometrical structure.

The baseline models which have been presented here to provide an early benchmark trained on the presented data support the claim that a GNN based approach achieves a higher yield over the metrics considered. The best performing model is a GNN model described in Sect. 3 and 4 which obtained a final F1-score of 0.792 and a test CE loss score of 0.489 when evaluated on the test set. These results are an improvement from the next best performing model (2% higher F1-score and 6% lower CE loss) indicating the potential for graph based methods on this naturally graph structured data. Of particular interest for e-commerce companies is the level of confidence when making a prediction which will affect the likelihood of a customer being notified by the prediction. Therefore the final test CE loss value for the GNN being lower than other models implies that overall the GNN is likely more confident about its classifications than the other non-graph based approaches. In order to reinforce this point, a future analysis of these predictions could include the investigation of calibrated probabilities as in [7].

As discussed, our primary goal is to provide a novel dataset to facilitate future research studies in fashion retail. This data is presented with labeled purchase links between customers and product variants which can be used in a supervised learning setting (as in Sect. 4). However due to the graph structure of this data, it is possible to also use this data in the unsupervised setting with a wider range of transformer based models. Finally we wish to highlight the potential application of this dataset to advancements in recommendation systems. With the definite labels provided in this dataset which label a return, a future research direction would be investigating the universality of the GNN embeddings and how these translate into new recommendation systems for sustainable fashion.

REFERENCES

[1] L Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *CoRR* abs/1603.02754 (2016), 785 – 794. arXiv:1603.02754 http://arxiv.org/abs/1603.02754

[3] Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, Peter W. Battaglia, Vishal Gupta, Ang Li, Zhongwen Xu, Alvaro Sanchez-Gonzalez, Yujia Li, and Petar Velickovic. 2021. ETA Prediction with Graph Neural Networks in Google Maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM ’21)*. Association for Computing Machinery, New York, NY, USA, 3767–3776. https://doi.org/10.1145/3459637.3481916

[4] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. 2018. Pixie: A System for Recommending 3+ Billion Items to 200+ Million Users in Real-Time. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW ’18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1775–1784. https://doi.org/10.1145/3178876.3186183

[5] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

[6] I. J. Good. 1952. Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)* 14, 1 (1952), 107–114. http://www.jstor.org/stable/2984087

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML’17)*. JMLR.org, 1321–1330.

[8] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

[9] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883037

[10] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[13] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. 2020. Learning to Simulate Complex Physics with Graph Networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML’20)*. JMLR.org, ICML, Article 784, 10 pages.

[14] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. 2020. A deep learning approach to antibiotic discovery. *Cell* 180, 4 (2020), 688–702.

[15] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2022. Graph Neural Networks in Recommender Systems: A Survey. *ACM Comput. Surv.* (mar 2022). https://doi.org/10.1145/3535101 Just Accepted.