Koc. J. Sci. Eng., 5(ICOLES2021): (2022) 32-37 https://doi.org/10.34088/kojose.1019277

Kocaeli University

Kocaeli Journal of Science and Engineering

http://dergipark.org.tr/kojose

Brand Propensity Prediction with Click-Through Rate as a Target

Alptekin UZEL [1], Kaan PEKEL [2], Fatih ABUT [3,*], M. Fatih AKAY [4]

[1] *Trendyol, Istanbul, Turkey,* ***ORCID:*** *0000-0002-1563-743X*
[2] *Trendyol, Istanbul, Turkey,* ***ORCID:*** *0000-0001-5482-2999*
[3] *Department of Computer Engineering, Çukurova University, Adana, Turkey,* ***ORCID:*** *0000-0001-5876-4116*
[4] *Department of Computer Engineering, Çukurova University, Adana, Turkey,* ***ORCID:*** *0000-0003-0780-0679*

Article Info

Research paper

Received : November 04, 2021
Accepted : May 19, 2022

Keywords

Logistic Regression
Ensemble Model
Brand Propensity
Prediction

Abstract

Personalizing the e-commerce experience is vital since there are enormous amounts of products to offer customers. Each day new products are introduced into the ecosystem, and customer purchase behavior is dynamic as well. This mapping between products and customers needs to be optimized. E-commerce platforms try to funnel those products by a variety of methods like user clustering and product propensity analysis. The brand propensity metric is one of those key features for personalizing products offered to the customer. Once the brand propensity is calculated, it can be used to cluster customers or list products within the same brand. Since customers periodically interact with different products, these interactions (e.g., product visit, favorite, basket, search, and order) are aggregated to predict the next actions of the corresponding customer. Typically, the next action might be an order action or click. In this study, we develop Logistic Regression (LR) models to investigate the effect of the target variable on calculating brand propensity. For comparison purposes, models based on Decision Tree (DT), Random Forest (RF), and XGBoost (XGB) have also been developed. The target variable to be evaluated for the brand propensity model has been set to both order probability and click probability. The "Top N accuracy" metric has been used to evaluate the performance of the models. As the study's outcome, click as a target variable has been revealed to be more beneficial since it also shows that customers are more likely to explore what is inside that brand. In addition, the LR-based propensity models exhibit the best average performance for both Top 3 and Top 5 accuracies among the machine learning methods.

1. Introduction

One of the key problems in e-commerce is the mapping of products and customers: which product subset should be associated with which user cluster? In the personalization aspect, these clusters are the customers themselves. So, for each person, personalized product recommendations can be created. Then comes the next problem: the dimension of those product recommendations. The most common dimensions are brand, category, and price. All those factors can define user behavior such that specific sets of products can be recommended [1–3].

Services of e-commerce platforms reach the customer via applications and web pages. So, product recommendations have user experience (UX) components. In applications, those components are called widgets. Each widget can cover different dimensions of the product recommendation. Some can be associated with categories, whereas others with brands or prices.

This study proposes Logistic Regression (LR) models to investigate the effect of changing the target variable on brand propensity prediction. For comparison purposes, models based on Decision Tree (DT), Random Forest (RF), and XGBoost (XGB) have also been developed. The target variable to be evaluated for the brand propensity model has been set to both order and click probabilities. Personalized brands are sorted and recommended to the customers in a

* Corresponding Author: fabut@cu.edu.tr

brand slider widget. In this context, it is a ranking and sorting problem. Product and customer data of the e-commerce site Trendyol is used in this problem. Each customer interacts with different products through the application. These interactions (e.g., product visit, favorite, basket, search, and order) constitute the signals/features for the machine learning model to interpret. They are aggregated by brand and used to predict customers' actions on that brand. The problem also has a time domain, so it is also related to forecasting. The action features are aggregated as time-lagged features (i.e., one day, seven days, two weeks, etc.) to tune the effects of those signals in the expected actions.

The rest of the paper is organized as follows. First, the related works are summarized. Then, the details of the developed models and the methodology are introduced to predict the brand propensity for each customer. Next, the results are presented. Finally, the paper is concluded along with possible future works.

2. Related Works

The next purchase prediction of the customers in an e-commerce platform based on the customer and product interaction data has been investigated for a relatively long time. In recent years, deep learning models and ensemble methods are also adapted to the problem. The following studies helped to shape the idea presented in this paper.

Zhang [4] compared several LR and RF metrics for predicting customer propensity in an e-commerce platform. Valecha et al. [5] discussed consumer behavior by applying a predictive model to a dataset in Kaggle. Szabó and Geng [6] used a novel approach to create a sequence of numbers to represent customer behavior and then applied deep learning methods to use this as a feature to predict purchases. Liu and Li [7] used similar data to predict purchase behavior and applied the Support Vector Machine (SVM). SVM has high accuracies with high-dimensional data. Still, linear separability is always an issue for that family of algorithms, and new dimensions need to be introduced to solve the problem. Hu 38 and Shi [8] created a time-series sequence of customer behaviors, fed this into an LSTM model, and then used the outcome as a new feature to be fed into a Random Forest model. Zhai et al. [9] used an ensemble model combining XGB and LightGBM algorithms to predict customer purchases on extensive e-commerce customer interaction data. Policarpo et al. [10] provided a comprehensive and up-to-date survey of machine learning techniques used in e-commerce platforms. Stubseid and Arandjelovic [11] represented the difference between the Naive Bayes approach and the RF approach using real-world data, which consists of a user to product relation. Finally, Sasi et al. [12] applied RFM and Recurrent Neural Network using customers' previous purchases to predict the next purchase by including the time factor.

It is clear from the literature that prediction of the next purchase and propensity of customers are becoming more critical in e-commerce to optimize customer-product relations and provide better options to customers. Differently from the rest of the studies in the literature, we investigate the effect of changing the target variable from order probability to click probability in brand propensity prediction. We find that click as a target variable is more beneficial in brand propensity prediction.

3. Proposed Brand Propensity Models and Methodology

In most of the propensity models, the main features are the aggregated actions of the users. In e-commerce environments, these actions are related to product: visit, favorite, basket, search, and order are the main interactions. The first four interactions constitute the signals of the customer for buying a product. They all show the customer's interest in a brand, category, and specifically in that product. Those signals which translate into an interest might turn into an order. Finally, the order itself is a strong signal as well. After a completed order based on the product category, the customer might continue ordering the same product or from the same brand/category.

An essential problem in brand propensity is related to the replenishment of the products. Every product has a different purchase frequency. So, in this model, a brand-based repurchase ratio is calculated for each product. And then, these constants are multiplied by the main features to scale the features by interaction ratios. The input and the target variables of the baseline model are shown in Table 1.

The idea is to sort those brands for each customer so that if they order a product the next day, that brand is in the top 5 brands listed for that customer by the model. The model is responsible for gathering those signals, giving different weights to those signals, and then calculating the order probability from that brand in the next 24 hours as the output. More specifically, the created dataset includes 2.153.360 rows and 40 columns. This data has been gathered from the users who visited the application. There are around 2 million users. Numerous attributes have been collected for each user transaction, such as the number of basket actions, likes, and orders a user made in the last one week, last one day, and last two weeks.

These features were built into an LR model [13] for predicting whether a customer has ordered based on his/her signals. So, an LR model is fit into the data using the Sklearn package in Python with the parameters listed in Table 2. LR has been preferred due to its performance on brand

Table 1. The input and output variables used by the baseline model

Type	Variable	Description	Time period
Input variable	Basket count (user-based)	Number of times a brand is added to basket	1d, 1w, 2w, 1m
	Favorite count (user-based)	Number of times a brand is favorited	1d, 1w, 2w, 1m
	Search count (user-based)	Number of times a brand is searched	1d, 1w, 2w, 1m
	Visit count (user-based)	Number of times a brand is visited	1d, 1w, 2w, 1m
	Order count (user-based)	Number of times a brand is ordered	1d, 1w, 2w, 1m
	Basket interaction with brand repurchase ratio	Basket count * brand repurchase ratio	1d, 1w, 2w, 1m
	Favorite interaction with brand repurchase ratio	Basket count * brand repurchase ratio	1d, 1w, 2w, 1m
	Search interaction with brand repurchase ratio	Basket count * brand repurchase ratio	1d, 1w, 2w, 1m
	Visit interaction with brand repurchase ratio	Basket count * brand repurchase ratio	1d, 1w, 2w, 1m
	Order interaction with brand repurchase ratio	Basket count * brand repurchase ratio	1d, 1w, 2w, 1m
Target variable	Brand ordered	Brand ordered the next one day.	1d

Table 2. Logistics Regression parameters for the sklearn function

Parameter	Value
Penalty	l2
Dual	False
Tol	0.0001
C	2.0
Fit intercept	True
Intercept scaling	1
Class weight	“balanced”
Random state	None
Max iteration	100
Multi class	‘ovr’
Verbose	1
Warm start	False

Table 3. Input and target variables of the CTR propensity model

Type	Variable	Description	Time period
Input variable	Click propensity score	Propensity probability from the 1st propensity model	1d
	CTR rates for each brand	CTR ratios for each brand based on widget metrics	1d
Target variable	Brand clicked	Brand clicked the next day	1d

propensity prediction and quick training times. For comparison purposes, models based on DT, RF, and XGB have also been developed. Next, we attempted to improve the performance of the baseline model by changing the target from order to click. The same features have been used to predict the click on the brand. Additionally, click-through rate (CTR) has been added to the model to increase the conversion rates further. And then, another model used the

outcome of the first baseline model and CTR to predict click again. So, this turned into an ensemble model called the CTR propensity model, as illustrated in Figure 1.

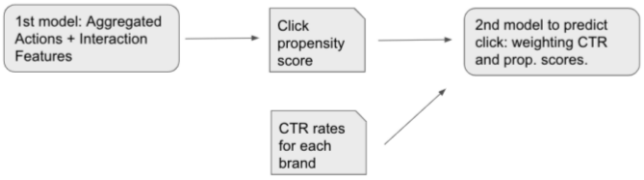

Figure 1. Improved ensemble CTR propensity model

In conclusion, the base model creates propensity scores for each user for the corresponding brands. But the ensemble model tunes the propensities using CTR data from the widget. This way, popular brand features are gathered and combined into the model. If some brands are more popular than the previous day, this overall aggregated feature is also integrated into the model. This feature is then used in the second model and thus tunes the propensity scores of the users.

CTR rates are calculated daily for the widget of the corresponding brand. Based on the signals obtained from the customers, raw propensity scores are weighted and rescored to produce the final output. Table 3 shows the input and target variables of the CTR propensity model.

The "Top N accuracy" metric [14] has been used to evaluate the performance of the models. The Trendyol application consists of different widgets. They correspond to a particular place in the application and have different functionalities. Figure 2 shows the brand slider widget in the Trendyol application. This widget is a slider in the application. It consists of brand logos that take the user to those brands if clicked. Brands in this slider are calculated and ordered by this algorithm in the paper. For each user, brand propensities are calculated and used to provide a personalized experience to the users. So, the brands and their order inside the widget are calculated by the corresponding user's interaction with the application. These interactions are used as features to predict the brand propensity of the user.

The "Top N" brands calculated are fed into this widget. Since the brand slider widget has limited slots for the brands, in both models, N was set to 5.

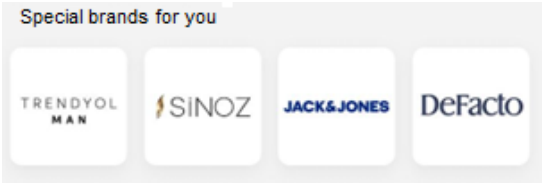

Figure 2. Brand slider widget in the Trendyol application

4. Results and Discussion

Table 4 shows the results achieved by evaluating the baseline and CTR propensity models. Since the order data is sparse, only customers with orders have been selected to assess the results of the models.

Table 4. Results achieved by evaluating the models using LR, DT, RF, and XGB (Top 3 and Top 5 product purchases - at least one transaction)

Model	Simulation Metric	Probability
LR-based Baseline model	Top 3 product purchase	57.36%
	Top 5 product purchase	65.05%
LR-based CTR propensity model	Top 3 product purchase	64.07%
	Top 5 product purchase	80.10%
DT-based CTR propensity model	Top 3 product purchase	61.09%
	Top 5 product purchase	69.33%
RF-based CTR propensity model	Top 3 product purchase	64.19%
	Top 5 product purchase	76.63%
XGB-based CTR propensity model	Top 3 product purchase	64.01%
	Top 5 product purchase	75.43%

According to the results obtained, when the "Top 3" product purchase metric is evaluated, the LR-based baseline and CTR propensity models yield an order probability of 57.36% and click probability of 64.07% for the next day, respectively. Similarly, when the "Top 5" product purchase metric is evaluated, the baseline and CTR propensity models produce an order probability of 65.05% and a click probability of 80.00% for the next one day, respectively.

In both "Top 3" and "Top 5" product purchase evaluations, it is observed that the CTR propensity model, where click is predicted as the target variable, clearly outperforms the baseline model predicting the order probability. The gain in probability obtained using the CTR propensity model instead of the baseline model is 11.69% and 22.98% for "Top 3" and "Top 5" product purchase metrics, respectively.

To compare and validate the accuracy of the LR-based CTR propensity prediction, models based on DT [15], RF [16], and XGB [17] have also been developed. When the "Top 3" product purchase metric is evaluated, the probabilities of DT-based, RF-based, and XGB-based models range from 61.09% to 64.19%. Similarly, when the "Top 5" product purchase metric is evaluated, the probabilities of DT-based, RF-based, and XGB-based models vary between 64.19% and 75.53%. Although LR shows comparable performances to other alternative methods in terms of "Top 3" product purchase metric, it clearly outperforms other methods in terms of the "Top 5" product purchase metric.

We can conclude that click as a target variable, an indirect way to increase the order rates, came out as more beneficial. The results have shown that the CTR propensity model can sort brands in such a way that the calculated "Top 5" brands for users with at least one order have an acceptable high probability (i.e., 80.00%) of being that ordered brand.

5. Conclusion and Future Work

This study investigated the effect of the target variable on brand propensity prediction using LR. For comparison purposes, models based on DT, RF, and XGB have also been developed. The target variable to be evaluated has been set to both order and click probabilities. The users' interactions (i.e., product visit, favorite, basket, search, and order) constitute the signals/features for the models to interpret. They are aggregated by brand and used to predict the brand's order and click probabilities. The results show that click as a target variable, an indirect way to increase the order rates, has been revealed to be more beneficial in predicting brand propensity. In addition, the LR-based CTR propensity model exhibits the best average performance among the machine learning methods for both Top 3 and Top 5 product purchases.

In the future, we plan to evaluate other promising methods, such as general regression neural networks and multilayer perceptron, which can be leveraged to improve the prediction accuracy of brand propensity. Also, other candidate potential features, such as customers' past purchase history and location, can be integrated into our prediction models to investigate the correlation of these variables with brand propensity.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Hussien F. T. A., Rahma A. M. S., Abdulwahab H. B., 2021. An e-commerce recommendation system based on dynamic analysis of customer behavior, Sustainability, **13**(19), 10786.

[2] Abdul Hussien F. T., Rahma A. M. S., Abdul Wahab H. B., 2021. Recommendation Systems for E-commerce Systems An Overview. Journal of Physics: Conference Series, **1897**(1), 012024.

[3] Daoud M., Naqvi S. K., Ahmad A., 2014. Opinion Observer: Recommendation System on E-Commerce Website. International Journal of Computer Applications, **105**(14), pp. 975–8887.

[4] Zhang Y., 2021. Prediction of Customer Propensity Based on Machine Learning. In: Proceedings of Asia-Pacific Conference on Communications Technology and Computer Science, pp. 5–9.

[5] Valecha H., Varma A., Khare I., Sachdeva A., Goyal M., 2018. Prediction of Consumer Behaviour using Random Forest Algorithm. In: Proceedings of 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering, pp. 1-6.

[6] Szabó P., Genge B., 2020. Efficient Conversion Prediction in E-Commerce Applications with Unsupervised Learning. In: Proceedings of 28th International Conference on Software, Tele-communications and Computer Networks.

[7] Liu X., Li J., 2016. Using support vector machine for online purchase predication. In: Proceedings of International Conference on Logistics, Informatics and Service Sciences.

[8] Hu W., Shi Y., 2020. Prediction of online consumers' buying behavior based on LSTM-RF model. In: Proceedings of 5th International Conference on Communication, Image and Signal Processing, pp. 224–228.

[9] Zhai X., Shi P., Xu L., Wang Y., Chen X., 2020. Prediction Model of User Purchase Behavior Based on Machine Learning. In: Proceedings of IEEE International Conference on Mechatronics and Automation, pp. 1483–1487.

[10] Micol P. L. et al., 2021. Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review, Computer Science Review, **41**, 100414.

[11] Stubseid S., Arandjelovic O., 2018. Machine Learning Based Prediction of Consumer Purchasing Decisions: The Evidence and its Significance. In: Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence, pp. 100–106.

[12] Sasi R. K., John H., Jerard B., Sudheer S., A. Shaju, 2020. Customer Behaviour Prediction using Propensity Model. IPEM Journal of Computer Application & Research, **5**, pp. 38–43.

[13] Peng C. Y. J., Lee K. L., Ingersoll G. M., 2010. An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, **96**(1), pp. 3–14.

[14] Cremonesi P., Koren Y., Turrin R., 2010. Performance of recommender algorithms on top-N recommendation tasks. In: Proceedings of the 4th ACM Conference on Recommender Systems, pp. 39–46.

[15] Lakshminarayanan B., Decision Trees and Forests: A Probabilistic Perspective, Ph.D. Thesis, 2016.

[16] Ali J., Khan R., Ahmad N., Maqsood I., 2012. Random Forests and Decision Trees. International Journal of Computer Science Issues, **9**(5), pp. 272–278.

[17] Chen T., Guestrin C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.