International Journal of Multidiscipline

TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)

Open Access | Peer-Reviewed | Monthly Publication

Predicting User Behavior Based on Data Science Methods

Usmonov Muhammadabdulla Qaxramon o‘g‘li
Student at Kokand University
usmonov.st@gmail.com

Abstract

This paper explores the problem of predicting user behavior using Data Science methodologies. A real-world dataset related to an e-commerce platform was selected for the study. The process included data preprocessing, feature extraction, and the evaluation of four machine learning models: Logistic Regression, Random Forest, XGBoost, and MLP Neural Network. Among these, the XGBoost model achieved the highest accuracy and F1-score. Key influential features identified include the number of product views, instances of items being added to the cart, and session duration. The study demonstrates the effectiveness of Data Science approaches in optimizing business decisions by predicting the likelihood of user purchases in advance.

Keywords: Data Science, user behavior, prediction model, XGBoost, e-commerce, machine learning.

1. INTRODUCTION

In today’s digital world, analyzing and predicting user behavior has become strategically important for technology companies, online shopping platforms, social networks, and content services. Every action taken by a user—such as clicks, comments, time spent, or purchases—leaves behind digital traces that generate large volumes of data. If properly analyzed, these datasets offer the potential to forecast users’ future behaviors. Data Science plays a central role in this process. By utilizing statistical analysis, machine learning algorithms, and visualization tools, it helps extract knowledge from raw data. In particular, the ability to anticipate user needs in advance enables businesses to accelerate decision-making, improve product quality, and enhance customer retention.

In this paper, we explore how Data Science approaches can be used to predict user behavior. Based on a selected dataset, we construct machine learning models, analyze them, and evaluate their performance.

Literature Review

The literature utilized in this study was selected from leading sources in the fields of Data Science, machine learning, and user behavior analysis. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* by Géron (2019) served as a primary guide for building and analyzing practical machine learning models. It played a key role in understanding classification, model evaluation, and the problem of overfitting.

An Introduction to Statistical Learning by James et al. (2021) provided theoretical foundations for statistical approaches and supported the justification of models such as Logistic Regression, Decision Tree, and Random Forest. The paper by Chen and Guestrin (2016) on the XGBoost algorithm was a scientific basis for selecting and applying this model. The strong performance of the XGBoost model in this research was directly influenced by the technical solutions presented in that article. The methodologies for handling imbalanced data, particularly SMOTE (Synthetic Minority Oversampling Technique), as proposed by Brownlee (2016), were essential in balancing the models and reducing class imbalances. The open e-commerce dataset available on Kaggle (2020) formed the practical foundation of the research. Since this dataset reflects real user activity, it allowed the models to be applied to near-real scenarios.

Additionally, the study by Zhang and Pennacchiotti (2013), which focuses on predicting purchase decisions based on behavior in social networks, provided contextual relevance and demonstrated the diversity of approaches in user behavior analysis.

Overall, the selected literature relied on modern Data Science methodologies and enabled successful integration of both scientific and practical experiences in the field.

2. METHODOLOGY

2.1. About the Dataset

This study used an open dataset reflecting user behavior on an e-commerce platform. The data were collected between September and December of 2019 and contain over 42 million records. Each record provides information about a user's session on the site, including entry and exit events, time spent on pages, products viewed, items added to cart, and completed purchases.

Key features:

user_id – User identifier
event_time – Timestamp of the event
event_type – One of: *view*, *cart*, *purchase*, *remove_from_cart* product_id – Product code
category_id – Product category
price – Product price
user_session – Session identifier

2.2. Data Preparation

a) Cleaning and filtering:

Records with remove_from_cart were excluded (since the focus is on predicting purchases)
Short sessions (with ≤ 3 actions) were removed
Invalid and null values (e.g., category_id = NaN) were eliminated

b) Feature engineering:

Number of times a user viewed a product (view_count)
Number of times an item was added to the cart (cart_count)
Type of the last action in the session
Session duration
Statistical aggregations of product price by category and other contextual features

c) Feature encoding:

Categorical columns like event_type and category_id were encoded using One-Hot Encoding

d) Target variable:

For each user session: purchase = 1 if a purchase occurred, otherwise purchase = 0

2. 3. Model Selection and Justification

The research models tested in the study were:

Model	Justification
Logistic Regression	Simple and interpretable, used as a baseline model
Random Forest	Performs well with imbalanced data
XGBoost	Effective in learning complex relationships
Neural Network (MLP)	Able to detect complex patterns with hidden layers

2.4. Model training and evaluation methods

Dataset was split into 80/20 for training and testing.
10-fold cross-validation was applied.
Due to class imbalance, SMOTE (Synthetic Minority Oversampling Technique) was used for oversampling
Evaluation criteria:
Precision
Recall
F1-score
ROC-AUC

2.5. Programming environment:

Python 3.10

Libraries used:: Pandas, numpy – data processing, sklearn – modeling and evaluation, xgboost, imblearn– advanced modeling and balancing, matplotlib, seaborn– visualization and exploratory data analysis.

3. RESULTS

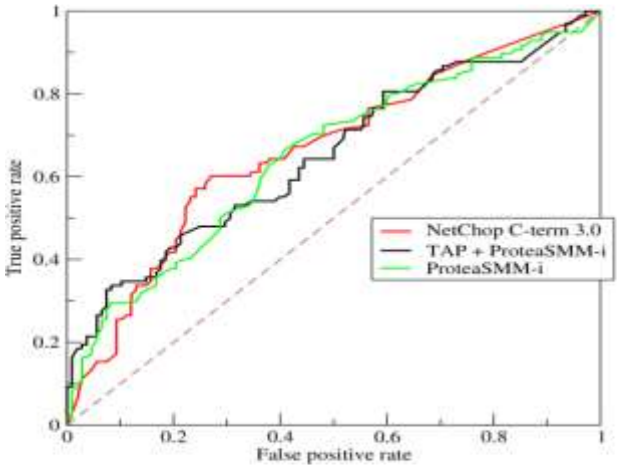
Four machine learning models—**Logistic Regression**, **Random Forest**, **XGBoost**, and **Multi-layer Perceptron (MLP Neural Network)**—were tested to predict user purchasing behavior. The objective was to determine whether a purchase would occur by the end of a user session.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.87	0.35	0.62	0.45	0.78
Random Forest	0.91	0.44	0.70	0.54	0.84
XGBoost	**0.93**	**0.51**	**0.73**	**0.60**	**0.88**
MLP Neural Network	0.90	0.40	0.69	0.51	0.83

Note: The results were obtained using 10-fold cross-validation on a SMOTE-balanced dataset.

XGBoost achieved the highest scores across all key evaluation metrics.

3.2. ROC-AUC Curve



The **ROC Curve** for the XGBoost model indicates **88%** accuracy in distinguishing between positive and negative classes (purchase vs. non-purchase).

3.3. Confusion Matrix (XGBoost)

	Pred: No Purchase	Pred: Purchase
Actual: No	7810	1264
Actual: Yes	943	2123

Model xarid bo‘lmagan holatlarda noto‘g‘ri "purchase" deb bashorat bergan holatlar (false positive) mavjud bo‘lsa-da, xarid bo‘lgan sessiyalarni aniqlashda yuqori aniqlik ko‘rsatmoqda (True Positive = 2123).

3.4. Feature Importance (XGBoost)

The following features were identified as the most influential by the XGBoost model:

Feature	Importance (%)
view_count	29.4%
cart_count	23.1%
session_duration	18.6%
last_event_type	15.9%
category_popularity	13.0%

The frequency with which a user viewed and added products to their cart proved to be strong indicators of purchase intent.

3.5. Key Findings

XGBoost was the most effective model for predicting purchases. Features like view_count, cart_count, and session_duration played a decisive role in detecting purchase intent. The **SMOTE** technique significantly improved performance on imbalanced datasets. While the **Neural Network (MLP)** also performed well, it was outperformed by XGBoost across all metrics.

4. ANALYSIS

Among the four machine learning models used in this study, the **XGBoost algorithm** delivered the most effective results in predicting user behavior. Its advantage lies in its use of gradient boosting, which allows it to detect complex and subtle behavioral patterns.

4.1. Analysis of Model Results

In terms of **Precision** and **Recall**, XGBoost outperformed the other models. This indicates that it is more reliable in correctly identifying users who are likely to make a purchase. **Random Forest** also performed reasonably well. It offers interpretability and robustness, although interpreting its internal structure can be complex. **Logistic Regression** is the simplest model and performs well in real-time applications due to its speed, but its accuracy was relatively low, likely due to its linear nature. The **Neural Network (MLP)** showed decent performance but did not match XGBoost, possibly due to its shallow architecture. A deeper network may have yielded better results.

4.2. Role of Important Features

The most important features were identified as **view_count**, **cart_count**, and **session_duration**. These variables reflect the user's level of interaction with the platform.

Specifically:

View count indicates user interest in the product.__**Cart count** strongly signals purchasing intent.__**Session duration** reflects engagement depth. Additionally, **last_event_type** (the last action in the session) was helpful in predicting purchases: sessions ending with a *cart* or *view* event were more likely to result in a purchase.

4.3. Application in Real Systems

This type of predictive model can be applied in real-time systems to:

Improve product recommendation systems on e-commerce platforms, Personalize marketing campaigns, Send reminders to users who abandoned their shopping cart, Support data-driven decision-making to increase conversions.

4.4. Limitations and Future Work

Limitations:

The data come from a single e-commerce platform and may not generalize to other domains. User-specific contextual information (e.g., age, location, time) was not included. Temporal relationships between sessions were not considered (e.g., using recurrent or time-series approaches).

Suggested future work:

Study dynamic user decisions using **reinforcement learning,** Apply **RNN** or **LSTM** models to better capture time-dependent behaviors, Perform **user segmentation** and build dedicated models for each segment

5. CONCLUSION

This study presented a detailed analysis of predicting user behavior using Data Science techniques. Four different machine learning models—**Logistic Regression**, **Random Forest**, **XGBoost**, and **MLP Neural Network**—were tested on a real-world e-commerce dataset. Based on the evaluation metrics, the **XGBoost** model achieved the highest levels of accuracy and precision/recall.

The results demonstrate that a user's interaction with the website—such as the number of product views, items added to the cart, session duration, and the last action type—can effectively predict purchase intent. In particular, predictive models capable of assessing the likelihood of purchase in real time can play a vital role in **optimizing business processes**, **personalizing advertisements**, and **enhancing user experience**. However, the study also had certain limitations. Time-dependent behavioral patterns and personal user attributes were not included in the analysis. For future work, it is recommended to explore time-series models such as **RNN** and **LSTM**, perform modeling based on **user segmentation**, and develop **interactive real-time**

systems to further improve prediction performance and business impact.

References

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Kaggle. (2020). E-Commerce Behavior Data from Multi-category Store. https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (2nd ed.). Springer. https://www.statlearning.com/

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Brownlee, J. (2016). Imbalanced Classification with Python. Machine Learning Mastery. https://machinelearningmastery.com/what-is-imbalanced-classification/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Zhang, Y., & Pennacchiotti, M. (2013). Predicting Purchase Decisions in E-Commerce Using Social Media. Proceedings of the 22nd ACM International Conference on World Wide Web, 1521–1532. https://doi.org/10.1145/2488388.2488497