UDC 004.8:658.89
DOI: https://doi.org/ 10.17721/3041-2323.2024.388-398

Alona BOHOLIEPOVA, Student
e-mail: abogolepova@gmail.com
GoIT, Redmond, USA

Tetyana FILIMONOVA, PhD (Phys . & Math.), Assoc. Prof.
ORCID ID: 0000-0001-9467-0141
e-mail: tatyana0377@gmail.com
State University of Trade and Economics, Kyiv, Ukraine

PREDICTING CONSUMER PURCHASING BEHAVIOR USING MACHINE LEARNING METHODS

This article explores the application of machine learning methods for predicting consumer purchasing behavior based on data analysis. Using a dataset from Kaggle, data was analyzed and prepared, including removal of duplicates, feature scaling, and correlation analysis. Various machine learning models, such as Random Forest and Gradient Boosting, were tested using different techniques, including hyperparameter optimization and class balancing. The results showed that incorporating feature correlation and hyperparameter optimization significantly improves the accuracy of the models, making them effective tools for predicting consumer behavior in online sales.

Keywords: ***machine learning, consumer behavior prediction, data analysis, Random Forest, Gradient Boosting, hyperparameter optimization, correlation analysis, online sales.***

Background

In the modern world, the demand for selling and buying goods online is growing every year. Therefore, it is crucial for e-commerce websites to predict whether a user will make a purchase. Fortunately, machine learning effectively handles this task.

Machine learning (ML) allows for the study of raw data to quickly solve complex business tasks. It helps automate routine tasks, process large volumes of data, improve customer service quality, and gain a competitive advantage in the market. ML enables businesses to make more informed decisions, predict future demand, and optimize business processes (Intelliarts, 2024; PixelPlex, 2023).

This paper examines the prediction of consumer purchasing behavior using machine learning methods. The "Predict Customer Purchase Behavior Dataset" (Kaggle, n. d.)

Machine learning methods are increasingly being applied to solve various economic problems, particularly in the financial sector. A significant amount of research is focused on using recurrent neural networks, such as LSTM (Long Short-Term Memory), for stock price prediction. For example, the study (Simplilearn, 2024 b) examines the application of LSTM networks for predicting Google stock prices, which illustrates the effectiveness of deep learning in financial analysis. Another study (Simplilearn, 2024 a) analyzes the process of developing machine learning models for stock price prediction based on historical data, emphasizing the importance of using machine learning (ML) methods in forecasting financial indicators.

Further studies demonstrate the application of the LSTM algorithm along with technical indicators for predicting price trends in the Vietnamese stock market, indicating the potential for using ML in emerging markets (Phuoc et al., 2024). In a systematic review (Sonkavde et al., 2023) of various machine learning and deep learning methods for stock market price prediction, a comparative analysis of the effectiveness of different approaches is presented. This study provides valuable recommendations for choosing the optimal method depending on specific conditions and data.

Furthermore, the study (ProjectPro, 2024) focuses on predicting stock closing prices using machine learning methods, confirming the potential of ML for analyzing specific financial indicators. The combination of these studies highlights the growing role of machine learning in financial analysis and opens up prospects for further improvement of existing models and approaches.

After downloading and reading the dataset, it is necessary to conduct data analysis. In this dataset, there are no missing values, and all data is presented in numerical format. However, 112 duplicates were detected. Considering the negative impact of duplicates on the machine learning process, a decision was made to remove them. The main reasons for this are:

• Duplicates can lead to model overfitting, as the model "learns" from the same data repeatedly, reducing its ability to generalize.

• The presence of duplicates can distort the data distribution, leading to model bias and decreased accuracy on new data.

• Duplicates increase the volume of data, which can result in longer model training times without actually improving its performance.

• Duplicates can affect quality metrics such as accuracy and F1-score, creating a false impression of high model performance (Zhao et al., 2021).

The next stage of the research is the construction and selection of the best model for predicting consumer purchasing behavior. This stage consists of three key parts. First, the best base model is selected, focusing on its ability to accurately predict consumer behavior. Second, the correlation between features is taken into account to improve the quality of the model. Third, various techniques are applied to enhance the chosen model, improving its performance and adaptability to changing market conditions. This approach will not only determine the optimal model but also ensure its high efficiency in real-world conditions.

The dataset includes the following features:

• Age: customer'sage;

• Gender: customer's gender (0: male, 1: female);

• Annual Income: customer's annual income in dollars;

• Number of Purchases: total number of purchases made by the customer;

• Product Category: category of the purchased product (0: electronics, 1: clothing, 2: home goods, 3: beauty, 4: sports);

• Time Spent on Website: time spent by the customer on the website, in minutes;

• Loyalty Program: customer's participation in the loyalty program (0: no, 1: yes);

• Discounts Availed: number of discounts used by the customer (range: 0-5);

• PurchaseStatus (target variable): likelihood of the customer making a purchase (0: no, 1: yes) (Kaggle, n. d.).

In classical machine learning, there are two main types of tasks: classification and regression. Classification is used to predict categorical labels (for example, determining whether an email is spam), while regression is used to predict numerical values (for example, predicting real estate prices). Given the objective of the task (PurchaseStatus), classification methods were chosen (Murphy, 2012).

Initially, we focus on selecting the best base model among several popular machine learning algorithms, specifically RandomForest-Classifier, GradientBoostingClassifier, SVC (Support Vector Classifier), and LogisticRegression. Each of these models has its unique characteristics and areas of application, making them suitable for different types of tasks. The research was aimed at identifying the model that provides optimal performance and accuracy for this dataset (Scikit-learn, n. d. b).

Data scaling is important in machine learning as it ensures the same scale for all features, allowing models to learn faster and improve accuracy. Using StandardScaler() from the sklearn library allows scaling of data, reducing the impact of different feature scales and increasing the efficiency of algorithms (Scikit-learn, n. d. d). For this purpose, StandardScaler was specifically chosen because it standardizes the data to zero mean and unit standard deviation (Pedregosa et al., 2011).

All features were selected for scaling, except for PurchaseStatus. This is because PurchaseStatus already has binary values (1 or 0), making it suitable for direct use in classification tasks. The main focus was on scaling other features to ensure their compliance with machine learning algorithm requirements and to improve the overall accuracy of the model.

To split the data into training and test sets, we used the train_test_split function (Scikit-learn, n. d. c). In this process, all features were split into training and test sets, and the target variable PurchaseStatus was split into corresponding labels (y_train_b and y_test_b). The test set comprised 30% of the total data volume, and the stratify parameter ensured the preservation of class proportions of the target variable in the samples.

After training the base models, we conducted additional training taking into account the correlation between features. Correlation is a statistical measure that reflects the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where a value close to 1 indicates a strong direct relationship, and a value close to -1 indicates a strong inverse relationship. Understanding correlation is important for making informed decisions in data analysis and model building. This will allow for a better understanding of the relationships

in the data and improve model accuracy by eliminating redundant information and focusing on the most significant features (Correlation analysis of data: What it is and how to apply it, n. d.). Fig. 1 presents the study of feature correlation with the target variable.

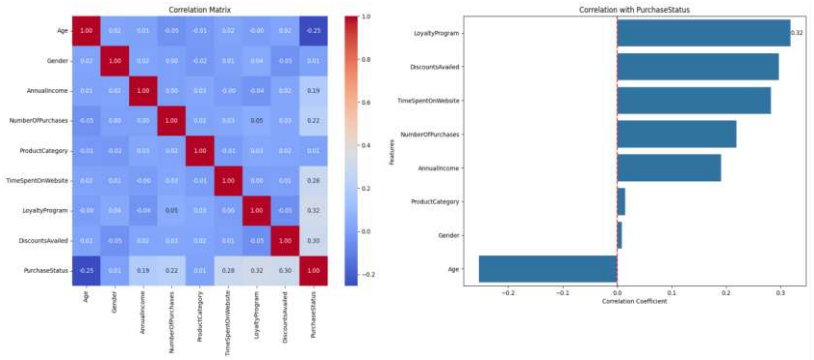

Fig. 1. *Correlation of features with the target variable PurchaseStatus*

The analysis of feature correlation with the target variable PurchaseStatus showed that the features LoyaltyProgram, Discounts-Availed, and TimeSpentOnWebsite have the strongest positive influence. At the same time, the features ProductCategory and Gender were found to be almost uncorrelated with the target variable, so it was decided not to include them in further model training.

Based on the research results presented in Tables 1 and 2, the following conclusions can be drawn regarding the selection of models for further work.

Among all the models considered, Random Forest with Correlation and Gradient Boosting with Correlation showed the highest accuracy, precision, recall, and F1-score, reaching 92.33%. This indicates that considering feature correlation significantly improves the performance of these models. At the same time, SVM and Logistic Regression demonstrated less impressive results, with accuracies of 86.09% and 82.73% respectively. Even accounting for correlation could not substantially improve their performance, indicating their lower effectiveness in this context. Given these results, two models were

selected for further work: Random Forest with Correlation (Table 3) and Gradient Boosting with Correlation (Table 4). These models demonstrate a high capacity for data generalization and are the most promising for solving tasks that require high accuracy and reliability.

Table 1

Performance results of base models

№	Model	Accuracy	Precision	Recall	F1-Score
1	Random Forest Base Model	0.908873	0.912551	0.905994	0.907894
2	Gradient Boosting Base Model	0.908873	0.911062	0.906618	0.908069
3	SVM Base Model	0.860911	0.861859	0.858766	0.859808
4	Logistic Regression Base Model	0.827338	0.830673	0.823493	0.825116

Table 2

Performance results of models considering correlation

№	Model	Accuracy	Precision	Recall	F1-Score
1	Random Forest with Correlation	0.923261	0.925669	0.921067	0.922584
2	Gradient Boosting with Correlation	0.923261	0.926409	0.920755	0.922512
3	SVM with Correlation	0.856115	0.855674	0.855198	0.855416
4	Logistic Regression with Correlation	0.810552	0.809724	0.80991	0.809812

After selecting the models, we focused on applying various techniques to improve them. Initially, hyperparameter optimization methods such as Grid Search and Random Search were used. These methods allow finding optimal values for model parameters, which

increases its accuracy and performance. Grid Search conducts an exhaustive search of all possible parameter combinations, while Random Search selectively tests random combinations, which can be more efficient in large parameter spaces (Scikit-learn, n. d. a).

Next, ensemble methods were used, particularly bagging (Bootstrap Aggregating). Ensemble methods combine predictions from multiple models to improve overall accuracy and stability of results. Bagging involves creating multiple subsets of data through random sampling with replacement and training individual models on these subsets. The results of these models are then averaged (or combined in another way) to obtain the final prediction. This approach reduces variability and increases the model's resistance to overfitting (Breiman, 1996).

Table 3

Results for Random Forest Models

№	Model	Accuracy	Precision	Recall	F1-Score
1	Random Forest Base Model	0.908873	0.912551	0.905994	0.907894
2	Random Forest using Bagging	0.920863	0.922287	0.919127	0.920269
3	Balanced Random Forest	0.920863	0.922898	0.918815	0.920201
4	Random Forest with Correlation	0.923261	0.925669	0.921067	0.922584
5	Random Forest using Random Search	0.930456	0.932625	0.928448	0.929873
6	Random Forest using Grid Search	0.932854	0.934132	0.931324	0.932377

The problem of class imbalance, which often arises in machine learning tasks when one category of the target variable significantly outweighs others, was also considered. Class imbalance can lead to the model predicting the dominant class well, but performing poorly on less represented classes. To address this issue, various approaches were applied, such as data resampling (increasing the number of examples in the less represented class or decreasing the number of

examples in the dominant class) and the use of specialized algorithms that account for class imbalance (He, & Garcia, 2009).

Table 4

Results for Gradient Boosting Models

№	Model	Accuracy	Precision	Recall	F1-Score
1	Gradient Boosting Base Model	0.908873	0.911062	0.906618	0.908069
2	Gradient Boosting using Grid Search	0.918465	0.919588	0.916875	0.917887
3	Gradient Boosting using Random Search	0.918465	0.920154	0.916563	0.917818
4	Gradient Boosting with Correlation	0.923261	0.926409	0.920755	0.922512
5	Balanced Gradient Boosting	0.925659	0.927762	0.923631	0.925037
6	Gradient Boosting using Bagging	0.930456	0.933341	0.928136	0.92981

Results

Random Forest models:

• The base model showed the lowest accuracy (90.89%).

• The best results were demonstrated by the model using Grid Search (93.29% accuracy).

• All improved versions of Random Forest outperformed the base model.

• Gradient Boosting models:

• The base model had the lowest accuracy (90.89%).

• The best results were shown by the model using Bagging (93.05% accuracy).

• All modifications of Gradient Boosting outperformed the base version.

The main objective of the study was to predict consumer purchasing behavior using machine learning methods. The research analyzed the performance of Random Forest and Gradient Boosting models using various improvement techniques, including feature correlation consideration, Grid Search, Random Search, Bagging, and class balancing. The following results were obtained:

Random Forest:

• The base Random Forest model showed an initial accuracy of 90.89%, serving as a baseline for further improvements.

• Adding feature correlation to Random Forest increased accuracy to 92.33%, highlighting the importance of considering feature correlations.

• Using Bagging and Balanced Random Forest methods led to the same accuracy of 92.09%, demonstrating the effectiveness of these approaches in reducing variance and increasing model stability.

• Random Search and Grid Search methods provided the highest results, with accuracies of 93.05% and 93.29% respectively, significantly improving model performance through hyperparameter optimization.

• Gradient Boosting:

• The base Gradient Boosting model also showed an initial accuracy of 90.89%.

• Adding feature correlation to Gradient Boosting improved performance to 92.33%, confirming the importance of considering feature correlations.

• Using Grid Search and Random Search led to an accuracy of 91.85%, indicating the effectiveness of these methods for fine-tuning the model.

• Balanced Gradient Boosting and Gradient Boosting with Bagging showed the highest accuracy among all Gradient Boosting models, reaching 92.57% and 93.05% respectively, emphasizing their potential in improving model accuracy and robustness.

Discussion and conclusions

Based on the conducted research, it can be concluded that for tasks of predicting consumer purchasing behavior, adding feature correlation and using hyperparameter optimization methods such as Grid Search and Random Search significantly improve the performance of Random Forest and Gradient Boosting models. Class balancing and Bagging also demonstrated their effectiveness, especially in conditions of data imbalance. For tasks requiring high accuracy and reliability, it is recommended to use models that take into account correlation and have optimized hyperparameters.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.

Correlation analysis of data: What it is and how to apply it. (n. d.). Retrieved from https://ua5.org

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* 21(9), 1263–1284.

Intelliarts. (2024). *Machine learning implementation in business [10 use cases].* Retrieved from https://intelliarts.com/blog/machine-learning-business-applications/

Kaggle. (n. d.). *Dataset for predicting consumer purchasing behavior.* Retrieved from https://www.kaggle.com/datasets/rabieelkharoua/predict-customer-purchase-behavior-dataset

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* MIT Press.

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.

Phuoc, T., Anh, P. T. K., Tam, P. H. et al. (2024). Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam. *Humanities and Social Sciences Communications*, 11, 393.

PixelPlex. (2023). *Top 10 machine learning applications in business [2023 list].* Retrieved from https://pixelplex.io/blog/machine-learning-applications-in-business/

ProjectPro. (2024). *Stock closing price prediction using machine learning techniques.*

Scikit-learn. (n. d. a). *Comparing randomized search and grid search for hyperparameter estimation.* Retrieved from https://scikit-learn.org/stable/auto_examples/-model_selection/plot_randomized_search.html

Scikit-learn. (n. d. b). *Comparing random forests and histogram gradient boosting models.* Retrieved from https://scikit-learn.org/stable/auto_examples/ensemble/-plot_forest_hist_grad_boosting_comparison.html

Scikit-learn. (n. d. c). *Data preparation.* Retrieved from https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Scikit-learn. (n. d. d). *StandardScaler.* Retrieved from https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Simplilearn. (2024a). *Master guide on machine learning for stock prediction.*

Simplilearn. (2024b). *Stock market prediction using machine learning in 2024.*

Sonkavde, G., Dharrao, D. S., Bongale, A. M. et al. (2023). Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies*, 11(3), 94.

Zhao, Y., Li, L., Wang, H. et al. (2021). On the impact of sample duplication in machine-learning-based Android malware detection. *ACM Transactions on Software Engineering and Methodology*, 30(3), Art. 40. Retrieved from https://chapering.github.io/pubs/tosem21impact.pdf

Отримано редакцією журналу / Received: 12.09.24
Прорецензовано / Revised: 23.09.24
Схвалено до друку / Accepted: 01.10.24

Олена БОГОЛЄПОВА, студ.
e-mail: abogolepova@gmail.com
GoIT, Редмонд, США

Тетяна ФІЛІМОНОВА, канд. фіз.-мат. наук, доц.
ORCID ID: 0000-0001-9467-0141
e-mail: tatyana0377@gmail.com
Державний торговельно-економічний університет, Київ, Україна

ПРОГНОЗУВАННЯ КУПІВЕЛЬНОЇ ПОВЕДІНКИ СПОЖИВАЧІВ ЗА ДОПОМОГОЮ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Досліджено застосування методів машинного навчання для прогнозування поведінки споживачів на основі аналізу даних. З використанням набору даних із Kaggle, проведено аналіз і підготовку даних, включаючи видалення дублікатів, масштабування ознак і кореляційний аналіз. Протестовано різні моделі машинного навчання, такі як Random Forest і Gradient Boosting, із застосуванням різних технік, зокрема й оптимізації гіперпараметрів і балансування класів. Результати показали, що використання кореляції ознак та оптимізації гіперпараметрів значно покращує точність моделей, що робить їх ефективними інструментами для прогнозування поведінки споживачів у сфері онлайн-продажів.

Ключові слова: ***машинне навчання, прогнозування поведінки споживачів, аналіз даних, Random Forest, Gradient Boosting, оптимізація гіперпараметрів, кореляційний аналіз, онлайн-продажі.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.