Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28896

Machine Learning-Driven Analysis of Factors Influencing Consumer Purchase Decision

Changlin Chen

School of Social Science, University of Manchester, Manchester, United Kingdom
Andrewchen0923@outlook.com

Abstract. Nowadays, as artificial intelligence (AI) develops, an increasing number of fields are starting to adopt it to support their work. However, using AI and models that analyze vast amounts of data, which humans cannot handle, seems to be the best way to utilize the technology. Nowadays, influencing factors ranging from the personal level to the broader environmental context are taken into account when concluding. However, the combined impact of these factors jointly exert a combined impact on such conclusions has not yet been thoroughly and extensively researched. The paper aims to use a machine learning model to uncover the conditions that have an effects on the purchase decisions of people from five cities in America. The research finds that consumers loyalty to specific stores or markets is the most important factor influencing purchase behavior, as predicted by logistic regression and random forest models. The research has further shown that the cities in which consumers reside do affect their purchase behavior.

Keywords: Machine Learning, Random Forest, Logistic Regression, Purchase Behavior

1. Introduction

In recent years, the widespread use of artificial intelligence (AI) and machine learning model are increasing, and predicion of results across various fields have become essential and crucial. Using the machine learning model as a method of improving customer likelihood of making a purchase is highly important, owing to the models' high accuracy and current market demands. Recent studies have shown the feasibility of using machine learning models for in analyzing and predicting purchase behaviors. Study predicting the purchase behavior online has used models such as decision trees, neural network and ensemble methods which reaches high accuracy [1]. Another study, which uses using logistic regression to predict repeated purchase behavior, also achieves high accuracy, further demonstrating the importance and necessity of machine learning models in marketing [2]. Most of the studies ignores conditions including the cities of customers, local item price comparing with local average salary or other personal status.

To analyze the influence of customer personal atrributes on their purchase behavior, the paper uses a machine learning model based on algorithms such as random forest and logistic regression. Through the paper, conditions influencing customer purchase behavior can be identified to provide offline shops with a competitive edge and enhance their commercial capabilities in the highly developed online market today.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28896

2. Literature review

Purchase behavior problem is a simple binary classification, as the result of the problem is whether it is purchased or not. Analyzing and predicting binary classification is one of the primary uses of machine learning models in marketing nowadays, and two kinds of models seem to be the best fit for the task among all the others. Logistic regression and random forest are two foundational algorithms in the machine learning model. Both models are widely used in the same type of tasks across many fields.

2.1. Logistic regression

The core principle of the logistic regression is to learn the linear relationship between input variables and the binary outcome. Through transformation via the logistic function, it outputs a possibility from 0 to 1 representing the likelihood of a purchase. Many studies addressing binary classification tasks opt to use the logistic regression model, as it delivers high accuracy with datasets from various fields and is easy to implement, requiring only input features and variables to generate predictive probabilities. For example, one study employs machine learning models (including support vector machine, random forests, and logistic regression) to identify potential customers by analyzing their purchase behavior [3]. Additionally, scholars have used logistic regression, random forests, and other models to predict consumer furniture purchase behavior, achieving high accuracy and identifying key influencing factors such as furniture type and purchase timing [4]. Compared to other models, logistic regression offers advantages, including its high accuracy, simplicity, and strong suitability for binary classification tasks.

2.2. Random forest

Different from logistic regression, random forest does not rely on the linear relationship between input variables. While random forest needs longer training time, it excels at both predicting purchase behavior and quantifying the contributing factors. The accuracy of it is higher than that of logistic regression as well. Advancements in machine learning models have led to a growing number of studies using random forests for predictive tasks. A study uses the model to predict cloud service purchase behaviors based on factors such as whether users clickes on advertisement, prior cloud service purchase and other three more factors, which is very close to the digital marketing environment nowadays [5]. Studies also apply to purchase behavior prediction in specific sectors, such as the cosmetics industry [6]. Others are interested in the model itself, for example, improving its predictive efficiency [7].

3. Methodology

3.1. Theoretical framework

The two models used in the experiment are basic and easy to understand compared to other models. Since the research aims to find the conditions that influence consumers purchase behavior, logistic regression and the random forest algorithm are more suitable than other algorithms, as the purchase or not type of binary classification problem. These two models are also widely used in solving these kinds of questions.

DOI: 10.54254/2755-2721/2025.LD28896

3.1.1. Logistic regression

Logistic regression is used to discover a linear relationship between variables. The theory involves importing a dataset with numerous columns of variables, and one column is the dependent variable, where the other columns are the independent variables. By testing the importance of each independent variable, the model can give out the possibility between 0 and 1 by transforming the linear combination using the sigmoid function. In the experiment of uncovering purchase behavior, logistic regression is a good choice since it can explain the linear relationship between customer attributes and purchase conditions.

3.1.2. Random forest

Random forest is the other model used in the experiment, and it is more generally used for predicting various outcomes. The core idea of the model is creating an ensemble of decision trees. By using randomness and voting or averaging in the algorithm, it can make high-variance weak trees together and become a strong predictive model. The abilities, such as modeling nonlinearity and interactions, of decision trees are incorporated into the random forest model. However, it is different from a single decision in the way of reducing overfitting of single trees in the combination.

3.2. Data process

3.2.1. Dataset

The dataset used to train the models contains six columns of variables. There are 500 rows of variables in the dataset, and they are divided into age, income, city, marital status, score, and purchased. The age of the customers ranges from 18 to 64, with the majority falling between 41 and 50. The minimum and maximum income in the dataset is 20.1 thousand to 120 thousand, and the samples are fairly divided into every section. City column records where the samples are from, including New York, Phoenix, Chicago, Los Angeles, and Houston. The marital status of the samples is divided into single, married, and divorced, with each category consisting of the same percentage of the samples. Score refers to the credit score of the samples, which is a numerical value ranging from 0 to 1. Purchased is the purchase behavior of the samples. Purchase behavior is the dependent variable in the experiment, and the goal is to find the most important variable that can influence the behavior. The behavior is recorded as 0 and 1 where 0 is not purchased and 1 is purchased in the dataset.

3.2.2. Data preprocessing

Before training the model, it is necessary to find and remove the missing values in the dataset. Since the number of columns is small, using one-hot encoding to divide them into more variables is considered and used in the model training. The dataset is split into a 75% training set and a 25% testing set, using the same random split during training.

3.3. Model design

One-hot encoding strategy spreads the city column in the dataset into five different factors, while another kind of strategy regards the city as a single factor. To reduce the affecting factors, hyperparameter tuning is used with the fundamental value in the experiment. However, the accuracy

DOI: 10.54254/2755-2721/2025.LD28896

and other evaluation metrics can be low since an increasing number of these is not the most important target. In the logistic regression model training, the features are standardized, and L2 regularization is used with regularization strength C at the default value 1.0.

The random forest model uses the same encoding method as the logistic regression does in order to help compare the two models under the same conditions. The number of decision trees in the training is 100 with fixed random seeds. Since the dataset is not large, 100 decision trees are considered sufficient for the experiment. The hyperparameter tuning strategy for the random forest model focuses on controlling model complexity, diversity, and model size. ROC-AUC is more important in determining model is a better fit for the experiment than other evaluation values, so it is the tuning target of the strategy.

4. Experiment

4.1. Experiment design

Since the final goal of the experiment is finding the most important factor that affects on the purchase behavior, ROC-AUC is considered the primary metric. After the model outputs the results, quantifying the uncertainty is needed for testing metrics, and using bootstrap resampling with a 95% confidence interval and DeLong's test to find the differences in AUC between models. To verify the results, robustness testing is necessary. The test includes repeated random splits, sensitivity analysis of smooth parameters, and comparison of imbalance-handling strategies.

4.2. Evaluation metrics

The metrics of the logistic regression and random forest models using two different encoding strategies, viewing the 5 cities as individual variables or as one single variable, are compared due to the metrics, as shown in Table 1. The accuracy of the models, however, is lower than expected, and it is clear that it is hard to use the dataset to predict purchase behavior, as there is too little data in the dataset. Furthermore, comparing the metrics of one-hot encoding and basic encoding method, no identical improvement can be found from the data in the logistic regression model, which means different cities can be the factor with the most effect on the purchase behavior.

Table 1. Evaluation metric

Encoding Strategy	Model used	Accuracy	Precision	Recall	F1 Score	AUC
One-hot	Logistic Regression	0.4960	0.5227	0.6866	0.5935	0.4581
One-hot	Random Forest	0.5600	0.5750	0.6866	0.6259	0.5643
Normal	Logistic Regression	0.4960	0.5222	0.7015	0.5987	0.4583
Normal	Random Forest	0.5360	0.5600	0.6269	0.5915	0.5376

According to Table 1, the one-hot encoding strategy still yields higher accuracy in the random forest algorithm, but has almost no effect on the logistic regression model cases. To improve the accuracy, precision, and other metrics, the result is based on the one-hot encoding strategy. The comparison between the two models is also based on the outcome of one-hot encoding.

DOI: 10.54254/2755-2721/2025.LD28896

4.3. Result analysis and comparison

From Table 1, it is evident that the random forest model better fits the experiment than the logistic regression. Comparing metrics in the table, almost all of the metrics for the random forest are higher than those of the other models. For example, the ROC-AUC of the random forest is 0.56, while that of logistic regression is 0.45. The accuracy of the two models is 0.56 and 0.49 respectively. In this way, random forest can do more work in finding conditions that influence purchase behavior. The feature importance of logistic regression in rank is shown in Table 2, and that of random forest is shown in Table 3. The tables are made up of the top 5 features that the model is important to the purchase behavior.

Table 2. Feature importance of logistic regression

Feature	Odds Ratio	Rank
City_Phoenix	1.012	1
Score	0.999	2
City_Houston	0.996	3
Age	0.973	4
City_Los Angles	0.936	5

Table 3. Feature importance of Random forest

Feature	Importance	Rank
Income	0.296	1
Score	0.292	2
Age	0.253	3
Marital_Status_Single	0.035	4
Marital_Status_Married	0.029	5

Loyalty score ranks the second in both models, which means it is one of the most important conditions influencing purchase behavior. Odds ratio of loyalty score is 0.999 ,and importance is 0.292. No city features rank top 5 in random forest's feature importance, showing the city features only have a linear relationship with purchase behavior, however, cannot be the most important factor since none of them are among the top 5 factors in random forest.

5. Discussion

5.1. Summary

Overall, the research can helps to identify the key conditions of the behavior, with loyalty score being the most important factor in influencing purchase behavior. Loyalty score ranks high in both models, although not the top among other metrics, illustrating that the score plays an important role in customers' behavior. With a higher loyalty score, the consumer is more likely to make a purchasing in the store. Other features do not show such high importance in both models, while the city ranks high in logistic regression but low in the random forest model. The importance of the city is only shown in the logistic regression model, so that can influence the result. However, further experiments are considered to be taken to test the factor's importance.

DOI: 10.54254/2755-2721/2025.LD28896

5.2. Limitations

The model's accuracy is lower than the normal rate and exhibits underwhelming performance across other metrics (e.g., precision, ROC-AUC). These limitations primarily manifest in two aspects: data quality and scale, as well as model methodology.

First, the number of samples suffers from insufficient scale and incomplete features. A larger training set can lead to higher accuracy since models can better learn and predict the purchase behavior than when using a dataset with less data. Second, the dataset may have an imbalance in the target variable, a common issue in binary classification that can degrade logistic regression performance [8]. The model methodology lacks comprehensiveness and refinement, limiting its predictive capacity [9]. Furthermore, the study ignores the adoption of active learning algorithms, which shows that active learning enhances classification accuracy by prioritizing high-information samples for annotation [10].

6. Conclusion

The study focuses on identifying the factors influencing customer purchase behavior using logistic regression and random forest models to identify key influencing factors amid the boom of online retail. It processes a dataset of 500 customer samples, including age, income, loyalty score, and purchase behavior, among others. Experimental results indicate that the loyalty score emerges as the most important factor affecting purchase decisions, based on the outcomes of the machine learning models. Additionally, the random forest model outperforms logistic regression in predictive performance, with higher accuracy and ROC-AUC, reflecting its advantage in capturing non-linear relationships between factors. These findings can help stores develop more effective sales strategies and attract more customers.

However, this study has several limitations. The relatively low accuracy and precision of the models are caused by the dataset with a limited number of features, imperfect model selection, and inadequate data preprocessing. Further studies can bring more accurate and varied results than the present ones by expanding the dataset through the collection of both online and offline customer data. Additionally, exploring and integrating other models will be an essential direction for developing higher-accuracy prediction systems.

References

[1] Wei, Z., Chen, L., & Xu, L. (2024). Prediction Model for E-Commerce User Purchase Behavior Based on Data Mining. In 2024 IEEE International Conference on Electronic Information Engineering and Computer Science (ICEIB).

[2] Li, S. (2024). Machine Learning-based Prediction Mechanism of Repeated Purchase Behavior of E-commerce Customers. In 2024 IEEE International Conference on Electronic Information Engineering and Computer Science (ICEIB).

[3] Sharma, S. and Kumar Soni, H. (2020). Discernment of Potential Buyers Based on Purchasing Behaviour Via Machine Learning Techniques. In 2020 International Conference on Advances in Computing and Data Sciences (ICACDS).

[4] Tian, E. (2025). Prediction of Furniture Consumer Behavior Based on Machine Learning and Comparative Analysis of Models. In 2025 IEEE Conference on Computer Information and Big Data Applications (CITBDA)

[5] Ghosh, S. and Banerjee, C. (2020). A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment. In 2020 IEEE 4th Conference on Information & Communication Technology (CICT).

[6] Nair, S. S., Raj, S., Kumar, A., & Patel, R. (2025) Segmenting and Predicting Consumer behavior in cosmetic industry: A Machine learning approach using Latent class and Random Forest techniques. In 2025 IEEE 3rd

DOI: 10.54254/2755-2721/2025.LD28896

International Conference on Robotics, Automation, Electrical and Electronics Engineering (RAEE-UCCI).

[7] Dhali, S., Ahmed, T., & Rahman, M. (2020). An efficient predictive analysis model of customer purchase behavior using random forest and XGBoost algorithm. In 2020 IEEE 4th Conference on Information & Communication Technology (CICT).

[8] Zhang, H., Li, Y., & Wang, J. (2019). Improving prediction accuracy for logistic regression on imbalanced datasets. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC).

[9] Bujang, S. D. A., Selamat, A., Krejcar, O., & Fujita, H. (2021). Multiclass prediction model for student grade prediction using machine learning. IEEE Access, 9, 95608–95621.

[10] Sakthi, K., Lekha, M. and Manesha, P.A. (2023) Exploring Active Machine Learning Techniques to Boost Classification Accuracy in Image and Text Models. In 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSS).