

Predicting Consumer Preferences for Online Versus Offline Shopping Channels Based on Machine Learning Models

Yanru Chen

*Department of Marketing, Southwestern University of Finance and Economics, Chengdu, China
42351032@smail.swufe.edu.cn*

Abstract. With the advent of the omnichannel retail era, identifying consumer shopping channel preferences has become critical for retailers to achieve precision marketing and enhance market competitiveness. Based on a consumer dataset from the Kaggle platform, which contains 11,420 valid samples and 18 predictive features, this study employed five machine learning models, namely Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and XGBoost, to predict consumers' preferences for online versus offline shopping channels. The predictive performance of each model was systematically evaluated using metrics including recall, F1-score and AUC, and the key factors influencing consumers' channel choices were identified. Experimental results showed that Logistic Regression achieved the highest AUC (0.7147) and F1-score (27.34%), demonstrating the advantage of linear models in overall discriminative ability. XGBoost, with an F1-score (27.03%) close to that of Logistic Regression and a significantly higher recall (72.73%), exhibited more practically valuable comprehensive performance. In terms of feature importance, need for touch, tech-savviness, online payment trust and daily internet hours were identified as the four most critical features for identifying consumer channel preferences. This study reveals the core driving role of psychological and digital habits factors in channel selection, and provides decision support for omnichannel retailers to identify potential online consumers and formulate differentiated marketing strategies.

Keywords: Omnichannel Retail, Consumer Channel Preference, Machine Learning, XGBoost

1. Introduction

With the development of digital technologies and the rise of e-commerce, consumers are no longer confined to a single shopping channel but switch and choose between online and offline channels [1]. To meet consumer demands and enhance market competitiveness, retailers have begun to widely adopt omnichannel retail strategies. Today, omnichannel retail has become a mainstream trend in the retail industry [2].

Omnichannel retail is a strategy in which retailers integrate online and offline stores to improve customer purchase experiences [3]. Burdin proposed a four-stage progression for retail channel development, namely single channel, multi channel, cross channel, and omnichannel [4]. The first three stages are business-oriented retail strategy phases [5], whereas omnichannel retail is a

consumer-experience-oriented new retail strategy stage [6]. Therefore, under the omnichannel retail model, identifying consumers' shopping channel preferences and improving their shopping experiences have become key to enhancing retailers' competitiveness.

Currently, machine learning models have been applied to consumer behaviour prediction, mainly covering purchase behaviour prediction, customer churn prediction, precision marketing and related areas [7]. For example, in non-contractual purchase behaviour research, Martínez et al. constructed a dynamic data-driven prediction framework that used customers' historical purchase timing, transaction amounts and other features, together with gradient boosting trees and other machine learning algorithms, to dynamically predict future purchase behaviour [8]. In customer churn research, Höppner et al. proposed a profit-driven decision tree method for churn prediction, which significantly improved profit performance on a telecommunications dataset compared with traditional accuracy-based tree methods [9]. In precision marketing, Nguyen et al. developed a recommendation system integrating collaborative filtering, popularity, and Bayesian personalised ranking, and evaluated the results using LightGBM and deep neural networks. The results showed that LightGBM significantly outperformed deep neural networks in terms of MAP@K and MAR@K metrics [10]. However, existing research still has limitations. First, relatively little attention has been paid to consumers' preference choices between online and offline channels. Second, psychological factors are underutilised as features in consumer behaviour prediction models. Therefore, this study aims to fill these gaps and provide a basis for retailers' channel strategies.

This study comprehensively employed five machine learning models, including Logistic Regression, Decision Tree, SVM, Random Forest and XGBoost, to predict consumers' online versus offline shopping channel preferences. The predictive performance of each model was comprehensively evaluated using accuracy, recall, F1-score and AUC. In addition, the relative importance of four categories of features, namely demographics, digital habits, psychological and attitudinal factors and shopping behaviour, in predicting channel preferences was assessed. The results are expected to help retailers identify potential online or offline consumers and provide decision support for their resource allocation.

2. Methodology

2.1. Data description

The dataset used in this study originated from the Kaggle platform. The original dataset contained 11,789 records and 26 features, covering demographics, digital habits, shopping patterns, logistics preferences and psychological factors. To avoid data leakage and reverse causality, feature filtering was performed, resulting in 11,420 valid records with 18 predictor variables and 1 target variable. The target variable is binary, with values of Online or Store. Table 1 presents the specific names and dimensional groupings of the predictor variables.

Table 1. Feature dimensions and descriptions

Dimension	Features
Demographics (4)	age, monthly_income, gender, city_tier
Digital Habits (4)	smartphone_usage_years, social_media_hours, tech_savvy_score (self-assessed tech proficiency), daily_internet_hours

Table 1. (continued)

Psychological & Attitudinal (7)	online_payment_trust_score (trust in online payment security), discount_sensitivity, impulse_buying_score (tendency for unplanned purchases), brand_loyalty_score, need_touch_feel_score (preference for tactile product evaluation), time_pressure_level (perceived time constraint in shopping), environmental_awareness
Shopping Behavior (3)	delivery_fee_sensitivity, product_availability_online (perceived variety of products online), free_return_importance

The distribution of the target variable is shown in Figure 1: offline consumers account for 89.7%, while online consumers account for 10.3%. This imbalance suggests that physical retail may still dominate, and also indicates that class weight adjustment strategies should be adopted in subsequent modelling.

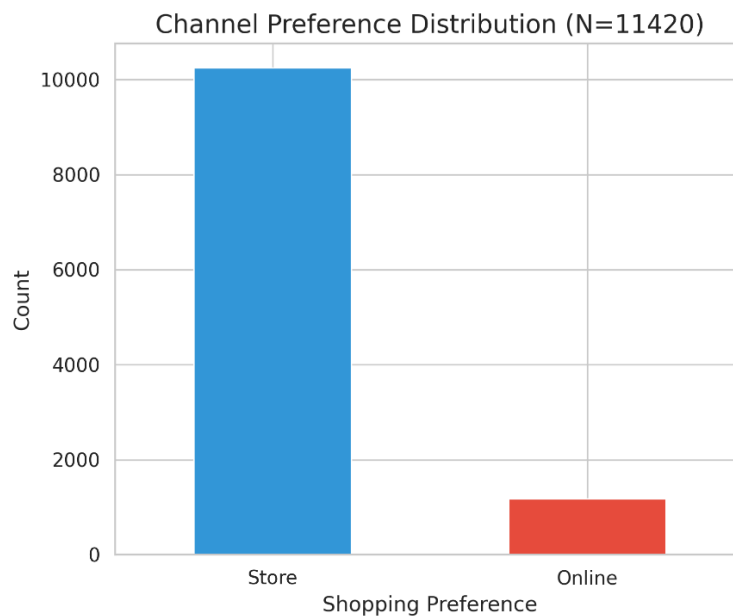


Figure 1. Distribution of shopping channel preference (picture credit: original)

2.2. Data preprocessing

2.2.1. Categorical variable encoding and feature standardisation

After inspection, all variables in the dataset had no missing values, so no imputation or sample deletion was required. Because *gender* and *city_tier* are categorical variables of nominal type, One-Hot Encoding was used. To eliminate the influence of different scales on model training, especially for Logistic Regression and SVM, all numerical features were standardised using Z-score normalisation to have a mean of 0 and standard deviation of 1. Although tree-based models are not affected by feature scaling, standardisation was applied uniformly to all model inputs for experimental consistency.

2.2.2. Multicollinearity check

During preprocessing, multicollinearity among numerical features was examined using the Pearson correlation coefficient matrix. The absolute correlation coefficients between any pair of features did not exceed the critical threshold of 0.8, indicating no severe multicollinearity (Figure 2). This result confirms that the 18 core features retained are not highly redundant, and no feature merging or further removal is needed. In subsequent machine learning modelling, these features will not interfere with model stability due to multicollinearity.

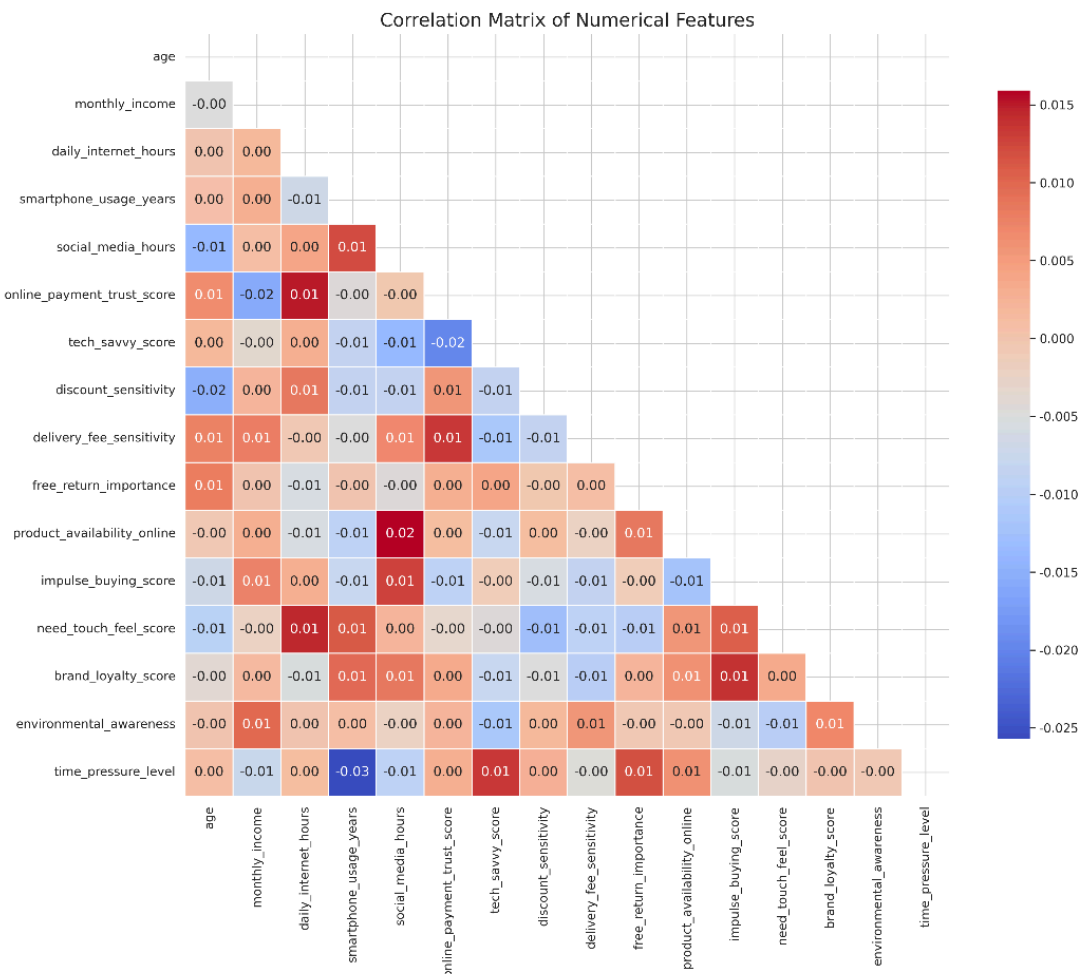


Figure 2. Pearson correlation matrix of numerical features (picture credit: original)

2.2.3. Data splitting

The data were divided via stratified sampling, with 70% allocated to the training set, 15% to the validation set, and the remaining 15% to the test set. The class proportions were kept consistent across the three subsets through stratification. The resulting split sizes were: training set 9,136 records, validation set 1,142 records and test set 1,142 records. The training set served for parameter learning, the validation set served for hyperparameter tuning and early stopping decisions, while the test set served for final performance evaluation. The test set was used only once after all model training and tuning were completed, ensuring objectivity of the evaluation results.

2.3. Model selection

Five representative machine learning models were selected, covering linear models, non-linear single models and ensemble models, for systematic comparison. For Logistic Regression, SVM and Random Forest, the `class_weight='balanced'` parameter was used to handle class imbalance. For XGBoost, the `scale_pos_weight` parameter was used, with hyperparameter optimisation via grid search and early stopping.

2.3.1. Logistic Regression

Logistic Regression is a generalised linear model that maps the linear combination through the sigmoid function to the $[0,1]$ interval, outputting the probability that a sample belongs to the positive class:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}} \quad (1)$$

where $\mathbf{x}$ is the feature vector, $\mathbf{w}$ is the weight vector and b is the bias term. When $P \geq 0.5$, the sample is predicted as an online consumer, whereas, it is predicted as an offline consumer.

2.3.2. Decision Tree

Decision trees build a tree-structured decision process via recursive partitioning of the feature space, in which the optimal feature and cutoff value are selected at every node. The splitting criterion uses the Gini Index:

$$Gini(D) = 1 - \sum_{k=1}^{|y|} p_k^2 \quad (2)$$

where p_k is the proportion of samples of class k in node D , and $|y|$ is the total number of classes. A smaller $Gini(D)$ indicates higher node purity. To prevent overfitting, the maximum tree depth was limited to 5.

2.3.3. Support Vector Machine

SVM seeks a hyperplane that maximizes the class separation and then employs it as the decision boundary for classification. In non-linearly separable cases, the data are mapped into a high-dimensional space via kernel functions, after which the optimal hyperplane is found by SVM in that transformed space. This study uses the Radial Basis Function (RBF) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \quad (3)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are feature vectors of two samples, $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ is the squared Euclidean distance between them, and γ is the bandwidth parameter of the kernel.

2.3.4. Random Forest

Random Forest employs bagging as its core ensemble learning strategy. It generates multiple training subsets via bootstrap sampling, trains multiple decision trees in parallel, and determines

final predictions by majority voting. Ensemble averaging reduces the variance of individual decision trees, which in turn enhances generalisation performance.

2.3.5. XGBoost

XGBoost is an ensemble learning algorithm based on gradient boosting, employing a sequential optimisation strategy to progressively reduce prediction error. Its objective function includes both loss and regularisation terms:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

where $\sum_{i=1}^n l(y_i, \hat{y}_i)$ measures the model's fit on the training set, and $\sum_{k=1}^K \Omega(f_k)$ is the sum of complexity penalties for all trees, used to control overfitting. The complexity of the k -th tree is defined as: $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|\mathbf{w}\|^2$. Here, T is the number of leaf nodes, γ controls the conservativeness of leaf splitting, $\|\mathbf{w}\|^2$ is the squared sum of leaf weights (L2 norm), and λ is the L2 regularisation coefficient that penalises large leaf weights.

2.4. Evaluation metrics

Since online consumers account for only 10.3% of the dataset, Accuracy may be misleading and is therefore not used as the primary evaluation metric. This study adopts recall, F1-score, and AUC as the core evaluation metrics. Recall measures the model's ability to identify the minority class (online consumers), reflecting its practical value in real scenarios. The F1-score, as the harmonic mean of precision and recall, comprehensively evaluates the model's identification quality for the minority class. AUC evaluates the model's discriminative ranking capability between positive and negative instances, without depending on classification threshold, thus rendering it highly applicable to the imbalanced dataset used in this work. In addition, precision, together with the mean accuracy and standard deviation from five-fold cross-validation, was taken into account for a thorough evaluation of the models' overall effectiveness and stability.

3. Results

3.1. Model performance comparison

A systematic evaluation of the five models was conducted on the test set. Table 2 summarises the performance of the five models.

In terms of overall discriminative ability, Logistic Regression achieved the highest AUC (0.7147), followed by SVM (0.7002). The superior AUC performance of Logistic Regression suggests that the core predictive features in this dataset have a relatively reasonable linear relationship with consumer channel preferences, so that linear models can more effectively distinguish between online and offline consumers.

Recall directly measures the model's ability to identify online consumers and is therefore one of the most practically meaningful metrics in this business scenario. The Decision Tree achieved the highest recall (85.96%), indicating that it can identify the vast majority of potential online consumers with almost no omission of users inclined toward online shopping. However, its precision was only 14.53% and its AUC was 0.6803, the lowest among all models, implying that this model

sacrifices specificity for sensitivity at the cost of a large number of false positives, which may lead to significant resource waste in actual marketing campaigns.

As the harmonic mean of precision and recall, the F1-score reflects a model's ability to balance false positives and false negatives. The experimental results showed that Logistic Regression had the highest F1-score (27.34%), followed extremely closely by XGBoost (27.03%), with the two being almost identical. Although Logistic Regression held a slight numerical advantage in F1, XGBoost outperformed Logistic Regression by nearly 7 percentage points in recall (72.73% vs. 65.96%) without a substantial decline in precision (16.60% vs. 17.24%). This indicates that XGBoost achieves a better balance than Logistic Regression between maximising the identification of online consumers and avoiding excessive disturbance to offline consumers, which aligns better with the comprehensive performance sought in this study.

Table 2. Performance comparison of five machine learning models

Model	Precision (Online)	Recall (Online)	F1-score	AUC	Accuracy
Logistic Regression	17.24%	65.96%	27.34%	0.7147	63.92%
Decision Tree	14.53%	85.96%	24.86%	0.6803	46.54%
SVM	16.02%	67.61%	25.96%	0.7002	60.25%
Random Forest	17.16%	55.68%	26.24%	0.6846	67.83%
XGBoost	16.60%	72.73%	27.03%	0.6901	59.66%

To further verify generalisation performance and stability, 5-fold cross-validation was performed for all models. Results are shown in Table 3. For all models, the CV Accuracy was very close to the test set Accuracy, with differences all below 3.5 percentage points, indicating reliable evaluation results and no severe overfitting. Among them, Random Forest performed best in cross-validation, with both high accuracy and high stability.

Table 3. Five-fold cross-validation results

Model	CV Accuracy	CV Std
Logistic Regression	63.20%	1.19%
Decision Tree	48.81%	7.11%
SVM	60.21%	1.44%
Random Forest	69.63%	1.10%
XGBoost	59.89%	1.39%

3.2. Feature importance analysis

To reveal the key factors driving consumers' shopping channel choices, feature importance was analysed for each model (except SVM), as shown in Table 4.

Need for touch ranked first across all four models, being the strongest predictor distinguishing online from offline consumers. Its coefficient in Logistic Regression was -0.520, indicating that consumers with higher need for touch are more inclined to shop offline. This is highly consistent with the haptic information processing theory of Peck & Childers, which states that reliance on

physical product experience (touching, trying on) is a core psychological driver of offline consumption.

Tech-savviness ranked second in all models, with a Logistic Regression coefficient of +0.435, suggesting that consumers with higher digital literacy are more inclined to shop online. This result accords with the Technology Acceptance Model, which suggests that a consumer's command of digital skills directly shapes their adoption behavior regarding online services.

Table 4. Feature importance rankings across models

	Logistic Regression	Decision Tree	Random Forest	XGBoost
1	need_touch_feel (-0.520)	need_touch_feel (0.407)	need_touch_feel (0.259)	need_touch_feel (0.150)
2	tech_savvy (+0.435)	tech_savvy (0.349)	tech_savvy (0.182)	tech_savvy (0.106)
3	online_payment_trust (+0.259)	online_payment_trust (0.122)	online_payment_trust (0.092)	online_payment_trust (0.066)
4	daily_internet_hour (+0.221)	daily_internet_hour (0.122)	daily_internet_hour (0.088)	daily_internet_hour (0.053)
5	product_availability_online (+0.121)	—	age (0.046)	brand_loyalty (0.048)

4. Discussion

4.1. Limitations

This study has several limitations. First, the dataset is sourced from the Kaggle platform as third-party public data, which may suffer from sample selection bias. Moreover, the collection year and geographic information are unspecified, limiting the timeliness and regional applicability of the conclusions. On the one hand, consumer behaviour has distinct temporal characteristics. If the dataset was collected around 2020, it may not reflect post-pandemic consumer channel preferences. On the other hand, e-commerce penetration rates vary greatly across countries and regions. The lack of regional labels makes it impossible to directly apply the findings to a specific market. Second, online consumers account for only 10.3% of the dataset, resulting in class imbalance. Despite the use of class weight adjustments, the model remains constrained in its capacity to thoroughly learn the minority class patterns. Third, the feature set is limited to static variables such as digital habits and psychological factors, lacking dynamic behavioural data that could capture the temporal evolution of consumer preferences.

4.2. Future research directions

Given the aforementioned shortcomings, several directions are proposed for future research enhancement. At the data level, sequential data such as consumers' historical purchase sequences and page-view trajectories can be integrated, and deep learning architectures such as LSTM or Transformer can be employed to model the dynamic evolution of consumption behaviour, thereby enhancing predictive accuracy and the ability to capture changing trends. In parallel, to validate the generalisability of the models, retail datasets from different countries or regions can be adopted for comparative studies. At the model level, explainable AI techniques such as SHAP and LIME can be

introduced to reveal, from different dimensions, the impact of each feature on individual predictions, thus offering more transparent decision support for the formulation of precision marketing strategies.

5. Conclusion

This study addresses the problem of predicting consumers' online versus offline shopping channel preferences in the omnichannel retail context. The predictive performance of five algorithms, namely Logistic Regression, Decision Tree, SVM, Random Forest and XGBoost, was systematically evaluated, and the key features driving channel choice were identified. In terms of model performance, Logistic Regression achieved the highest AUC and F1-score, while XGBoost, with a close F1-score and higher recall, achieved a better balance between recall and precision, demonstrating robust comprehensive performance. Regarding key features, need for touch and tech-savviness were the two most predictive factors, ranking top two across all models. Online payment trust and daily internet hours also showed stable predictive power. This study provides empirical evidence and decision references for omnichannel retailers to identify consumer channel preferences and optimise resource allocation, and also offers a methodological reference for subsequent research on consumer behaviour prediction. Future research could incorporate dynamic behavioural data and cross-regional samples to further enhance the applicability of the models.

References

- [1] Wolf, L., & Steul-Fischer, M. (2023). Factors of customers' channel choice in an omnichannel environment: A systematic literature review. *Management Review Quarterly*, 73(4), 1579–1630.
- [2] Sharma, N., & Dutta, N. (2023). Omnichannel retailing: Exploring future research avenues in retail marketing and distribution management. *International Journal of Retail & Distribution Management*, 51(7), 894–919.
- [3] Rigby, D. (2011). The future of shopping. *Harvard Business Review*, 89(12), 64–75.
- [4] Burdin, T. (2017). Omni-channel retailing: The brick, click and mobile revolution. *IDC Retail Insights Report*, 53(1), 12–15.
- [5] Berman, B., & Thelen, S. (2004). A guide to developing and managing a well-integrated multi-channel retail strategy. *International Journal of Retail & Distribution Management*, 32(3), 147–156.
- [6] Xia, Y., & Zhang, P. (2010). The impact of the online channel on retailers' performances: An empirical evaluation. *Decision Sciences*, 41(3), 517–546.
- [7] Kostopoulos, G., Stefani, A., Vasiliadis, V., & Kotsiantis, S. (2026). Deep learning for e-commerce: Recent developments in prediction, personalization and decision intelligence. *Applied Sciences*, 16(5), 2263.
- [8] Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3), 588–596.
- [9] Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., & Verdonck, T. (2020). Profit driven decision trees for churn prediction. *European Journal of Operational Research*, 284(3), 920–933.
- [10] Nguyen, D. N., Nguyen, V. H., Trinh, T., Ho, T., & Le, H. S. (2024). A personalized product recommendation model in e-commerce based on retrieval strategy. *Journal of Open Innovation: Technology, Market, and Complexity*, 10(2), 100303.