Applied Mathematics and Computational Intelligence
Volume 15, No. 2, 2026 [59-74]

Consumer Subscription Behavior Prediction using Machine Learning

Guang Wei Zheng[1], Wei Chien Ng[2*], Yu Qing Soong[3]

[1,3]School of Management, 11800, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[2,3]Department of Accountancy and Business, Tunku Abdul Rahman University of Management and Technology, Penang Branch, 11200 Tanjung Bungah, Pulau Pinang, Malaysia.

[*] Corresponding author : ngweichien@usm.my

Received: 10 March 2025
Revised: 12 July 2025
Accepted: 5 August 2025

ABSTRACT

Nowadays the e-commerce market is becoming increasingly competitive in the era of industry 4.0, which highlights the importance of customer retention strategies, especially through consumer subscription rates. In order to survive in the competitive environment, companies should understand the purchasing behavior of users, especially in the subscription model business as it is becoming one of the most important elements of revenue generation in business. This study utilized six supervised machine learning algorithms namely K-Nearest Neighbor, Support Vector Machine, Logistic Regression, Decision Tree, Random Forest and Extreme Gradient Boosting in predicting whether consumers will choose a subscription service through 3705 sample size datasets. Among the six supervised machine learning, the champion model found which shows that the Random Forest model has the highest accuracy with 77.1% and ROC of 0.843. It is also found that promo codes used, discount applied, and previous purchases are the most influential features for customers to choose to subscribe. This will enable subscription service providers to revise their subscription plan in order to attract more customers.

Keywords: Customer Subscription, Feature Importance Analysis, Machine Learning, Predictive Analytics.

1 INTRODUCTION

With the advancement of technology and the popularity of personalized services, the subscription model has gradually become an emerging business model to attract consumers, especially in the field of e-commerce. The e-commerce field has undergone rapid changes due to the introduction of new technologies and changes in consumer preferences [1]. The diversification of consumer behavior and personalized needs have become the core issues of concern to the industry [2]. The subscription economy has completely changed the way companies interact with customers, marking a major shift towards customer-centric marketing [3]. Consumer behavior research has been a critical issue in the field of marketing and retail, the core of which is to understand the psychology and behavior patterns of consumers in the process of purchasing decision-making.

Machine learning is a powerful tool for analyzing large datasets, particularly in identifying patterns and generating predictions based on consumer behavior [4]. Despite its widespread use across industries, limited research exists on applying machine learning to model consumer decisions related

to subscription services in e-commerce. The e-commerce industry faces intense competition from both established brands and new entrants. This often leads to aggressive pricing strategies that erode profit margins and make customer retention a pressing concern. Consumers, empowered by choices and highly sensitive to price, frequently shift between platforms, pushing companies to invest heavily in marketing and personalization to maintain loyalty [5].

Subscription-based e-commerce has emerged as a promising model to enhance customer retention and improve profitability. Retaining existing customers is often more cost-effective than acquiring new ones, and can significantly increase long-term revenue [6]. However, predicting whether a customer will opt into a subscription service remains challenging due to the complexity of consumer behavior and the diverse factors influencing purchasing decisions. To address this gap, the present study employs a comparative machine learning approach to predict consumer subscription decisions. Specifically, models such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) are evaluated. The best-performing model is further analyzed to identify key features that drive subscription behavior.

2 LITERATURE REVIEW

2.1 Supervised Machine Learning

This paper uses machine learning algorithms to develop predictive models for consumer subscription services. Based on the target variables, which is a binary classification, six supervised learning algorithms namely KNN, SVM, LR, DT, RF and XGboost will be adopted in the analysis.

The K-Nearest Neighbor (KNN) algorithm is a simple, yet powerful supervised learning technique widely used for classification and regression tasks. It operates on the principle that data points with similar characteristics tend to exist in proximity to the feature space. In a classification context, KNN classifies a data point by identifying its k nearest neighbors and assigning it the most common class among them, using distance metrics like Euclidean or Manhattan distance to determine proximity. The algorithm is non-parametric, meaning it makes no prior assumptions about the underlying data distribution, making it flexible for various applications including recommendation systems, pattern recognition, and medical diagnostics [7].

SVM is a supervised machine learning algorithm for classification problems and regression problems. The main goal of the SVM algorithm is to use a surface to separate multiple classes in the training data by optimizing the margins between them [8]. The SVM algorithm was chosen for this study because it has always been considered to have high accuracy. LR goes a step further than “normal” linear regression. It is useful when the dependent variable Y is categorical [9]. This study focuses on whether consumers subscribe, which is a binary classification problem. Therefore, this study will adopt binary LR because the predicted outcome is binary which is yes or no. In addition, LR does not require the optimization of any hyperparameters. Finally, LR is effective even when the sample size is small, there are few events, and there are few predictors.

For each sample, RF creates a tree and predicts the results from it. The higher number of trees gives the greater prediction accuracy. In addition, RF can also be used to select features and perform feature importance interpretation [10, 11]. They work well even when there are many features and

few observations. Therefore, this is why this study chooses to adopt the RF algorithm. Decision Tree (DT) is a widely used supervised machine learning model that applies a tree-like structure to perform both classification and regression tasks [12]. It represents the process of combining the tree growth model and its promising results, including occurrence significance, cost elements and performance characteristics. It is just a flowchart-like structure whose network represents the test of the function. In particular, the learning time is shorter compared to other algorithms. Therefore, this study selects the DT algorithm as one of the models.

XGBoost which is known as Extreme Gradient Boosting is an advanced ensemble learning algorithm based on the gradient boosting framework. XGBoost has gained significant popularity in both academia and industry for its high performance, scalability, and flexibility in handling structured or tabular data. It constructs additive decision trees sequentially, with each new tree aiming to correct the errors made by the preceding ones. It introduces regularization terms in the objective function to control model complexity, thereby reducing overfitting and improving generalization [13, 14].

2.2 Related Work on Machine Learning Prediction

The following sections present past research that used machine learning algorithms for predictive analytics. Chang et al. [15] focused on the problem of customer churn prediction in the telecommunications industry. It uses a variety of machine learning models to conduct research and finds that the RF model performed best, with the largest number of correct predictions in the confusion matrix and a prediction accuracy of 86.94%. The study conducted by [16] explored the use machine learning algorithms to predict customer churn, and on-time delivery of products in the e-commerce industry. An empirical analysis using real e-commerce data was conducted to predict customer churn and on-time product delivery. The DT algorithm achieved an accuracy of 83.38% on the test dataset for customer churn prediction, with a precision of 0.80, recall of 0.83, and F1-score of 0.77. These results demonstrate that machine learning is highly effective in forecasting customer churn and delivery performance in the e-commerce sector.

Pamina et al. [17] tried to improve the accuracy of customer churn prediction using KNN, RF, and XGBoost. The results show that XGBoost is the champion model with an accuracy of 79.80% and an F-score of 58.2% respectively. The research provides a model explanation for XGBoost, and feature importance analysis shows that customers using fiber options and paying higher monthly fees have a greater impact on churn. Another study showed that using RF and LR algorithms to predict heart disease, the results showed that the RF algorithm was more accurate than LR, reaching 90.16%.

Naveen et al. [18] focuses on early detection of cardiovascular disease (CVD) using machine learning methods—LR and RF—based on patients' health parameters. The models achieved precision rates of 85.25% and 90.16% respectively, enabling accurate identification of high-risk individuals and aiding in timely intervention. Another study developed a churn prediction model using LR and Fisher discriminant analysis based on data from three major Chinese telecom companies. The results indicate that LR exhibits superior performance, achieving an accuracy rate of 93.94% in the churn prediction model designed for telecom companies [19].

In the predictive analysis of heart disease based on human physical indicators, the accuracy of the SVM algorithm was 83.25%, the DT algorithm was 83.89%, and the KNN algorithm was 86.45%. The RF algorithm was 88.35%, and the LR algorithm was 84.22%. The results show that the RF algorithm is higher prediction accuracy in binary classification problems than other algorithms [20].

Kooptiwoot et al. [21] investigates the key factors influencing students' preferences for online learning during the COVID-19 pandemic, using data exploration and DT-based machine learning algorithms. Results identified seven major prerequisites, with "internet difficulties" being the most significant barrier to online learning. The study concludes that offering basic computer and internet courses can enhance students' readiness and engagement in online education.

Many past studies have used machine learning to predict customer churn, product delivery, or health risks. However, few have focused on customer subscription behavior in e-commerce. Most studies looked at different industries or outcomes. This study is different because it focuses on predicting subscription choices using six machine learning models, mainly KNN, SVM, LR, DT, RF, and XGBoost. It not only compares the performance of these models but also identifies the most important factors that influence customers to subscribe. These make the study unique and add value to the current research in this area.

3 METHODOLOGY

3.1 Research design

Figure 1 shows the research framework of this paper. The process is mainly completed using Python programming language.

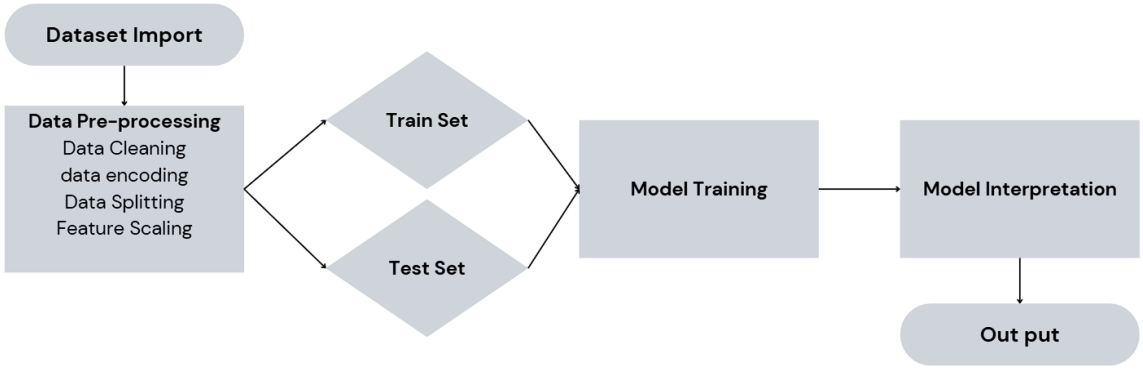

Figure 1: Research framework.

3.2 Data Collection

The online shopping customer purchase behavior dataset is a public dataset obtained from Kaggle. The dataset extracts a portion of data from each state in the United States, with a total of 18 columns and 3705 rows, including a target column for whether the customer has chosen to subscribe to the service. Table 1 shows the feature description for all the variables.

Table 1: Features Description

Variables	Description	Value Type
Customer ID	The unique ID of the consumer	Numeric
Age	Age of consumers	Numeric
Gender	Gender of consumers	Categorical
Item Purchased	Products purchased by consumers	Categorical
Category	Categories of products purchased by consumers	Categorical
Purchase Amount (USD)	Consumer spending amount	Numeric
Location	State of consumer	Categorical
Size	Size purchased by the consumer	Categorical
Color	Color purchased by the consumer	Categorical
Season	Consumer purchasing season	Categorical
Review Rating	Consumer Ratings	Numeric
Subscription Status	Whether the consumer chooses to subscribe	Categorical (YES, NO)
Shipping Type	Customer pickup method	Categorical
Discount Applied	Whether the consumer uses discount	Categorical (YES, NO)
Promo Code Used	Whether the consumer uses Promo Code	Categorical (YES, NO)
Previous Purchases	Number of previous purchases by the consumer	Numeric
Payment Method	Consumer payment methods	Categorical
Frequency of Purchases	Consumer purchase frequency	Categorical

3.3 Data Preprocessing

3.3.1 Data Cleaning

Data cleaning plays a crucial role in machine learning, as it enhances the quality of the dataset and significantly improves the accuracy and reliability of the model. The Customer ID column is usually a unique identifier used to label each row of data, which is usually meaningless in model training. The drop function was used to delete this column, and the dataset remains 17 columns. Date cleaning can improve the accuracy of the model, which is important for machine learning. The df.isnull().sum() function was used to detect whether there are missing values in the data set. The results are shown in Figure 2.

```
missing_values:
                          missing_values
Age                                    0
Gender                                 0
Item Purchased                         0
Category                               0
Purchase Amount (USD)                  0
Location                               0
Size                                   0
Color                                  0
Season                                 0
Review Rating                          0
Subscription Status                    0
Shipping Type                          0
Discount Applied                       0
Promo Code Used                        0
Previous Purchases                     0
Payment Method                         0
Frequency of Purchases                 0
```

Figure 2: Missing Value Detection

3.3.2 Data Encoding

Machine learning models cannot work directly with non-numeric data, such as strings. The dataset contains non-numeric data, which needs to be encoded before building a machine learning model. Based on Figure 3, binary encodes the Subscription Status, Discount Applied, and Promo Code Used columns (Yes = 1, No = 0), where Subscription Status is the target variable. One-hot encoding for categorical variables was used and encodes categorical variables by creating dummy variables.

```
binary_columns = ['Subscription Status', 'Discount Applied','Promo Code Used']
for column in binary_columns:
    df[column] = df[column].apply(lambda x: 1 if x == 'Yes' else 0)
categorical_columns = ['Gender','Category', 'Size', 'Season',
                       'Payment Method', 'Frequency of Purchases',
                       'Shipping Type','Color','Location','Item Purchased']

df = pd.get_dummies(df, columns=categorical_columns)
```

Figure 3: Data Coding

3.3.3 Data Splitting

Machine learning requires determining feature variables and target variables. First, use df_balanced. drop function to remove the Subscription Status column from the feature column and then set it as the target variable (y). Second, use the train_test_split function to generate a test set and a training set, with the ratio of the two sets to 70:30 and the random_state set to 42 as shown in Figure 4.

```
X = df.drop('Subscription Status', axis=1)
y = df['Subscription Status']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Figure 4: Data Splitting

3.3.4 Feature Scaling

This study uses SVM, KNN and LR algorithms. Differences in feature scales can cause computational bias in the algorithms. Therefore, the StandardScaler function is used for feature scaling as shown in Figure 5.

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Figure 5: Data Standardization

3.4 Model Training

First, build the KNN model. After importing the KNN classifier from the library, set the k value to cycle between 1-26 to find the best value as shown in Figure 6. The best performance is achieved when the k value is 18 as shown in Figure 7.

```
k_range = range(1,26
                )
scores = {}
scores_list = []
for k in k_range:
    classifier = KNeighborsClassifier(n_neighbors = k)
    classifier.fit(X_train, y_train)
    y_pred = classifier.predict(X_test)
    scores[k] = metrics.accuracy_score(y_test, y_pred)
    scores_list.append(metrics.accuracy_score(y_test, y_pred))
```

Figure 6: KNN classifier

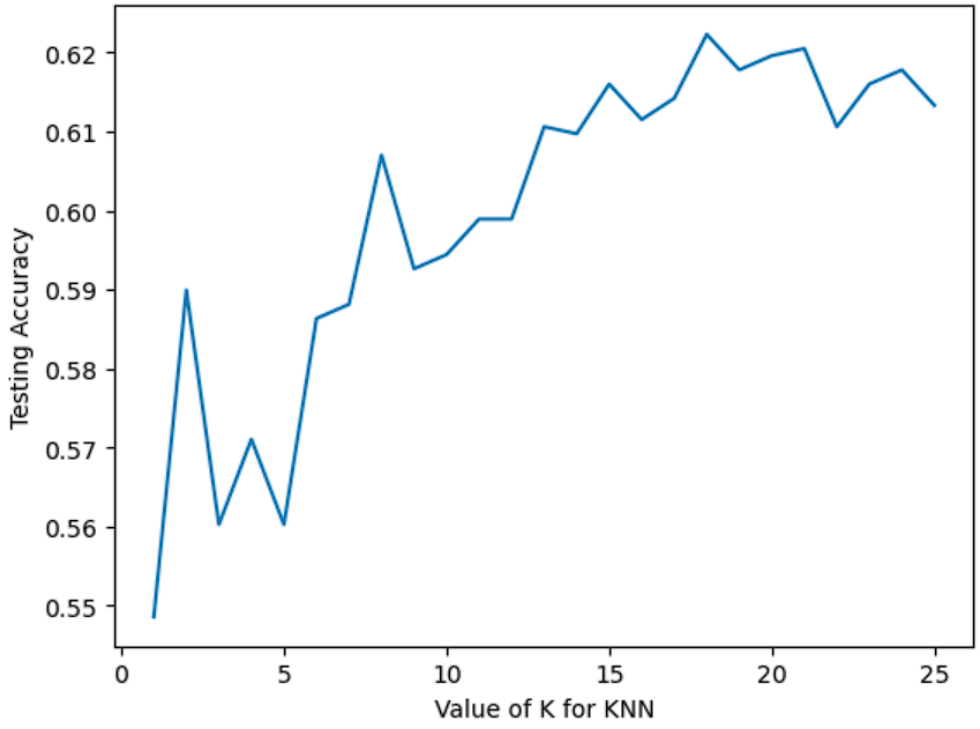

Figure 7: K value selection

By creating a dictionary containing 5 algorithms to store different machine learning models, the loop function is used to train the model using X_train and then use the trained model to predict the X_test, as shown in Figure 8.

```
models = {
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier(),
    'Logistic Regression': LogisticRegression(max_iter=1000),
     "SVM": SVC(),
    'XGBoost': XGBClassifier(use_label_encoder=False, eval_metric='logloss')
}
results = {}
for model_name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
```

Figure 8: Modes Building

3.5 Evaluation

There are many ways to measure the performance of machine learning. Confusion matrix and AUC-ROC are used in this paper.

3.5.1 Confusion matrix

In order to find the champion model, this study compares the performance of different algorithms by generating a confusion matrix. The prediction results of machine learning models can be calculated based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as shown in Table 2 [22].

Table 2: Confusion matrix

		Predication	
		No subscribe (0)	**Subscribe (1)**
Actual	**No subscribe (0)**	TN	FP
	Subscribe (1)	FN	TP

The following are descriptions of the four performance indicators:

i. Accuracy means the proportion of correct predictions to the total number of prediction results.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

ii. Precision is the proportion of all samples predicted to be in the positive class that are truly in the positive class.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

iii. Recall is concerned that the model correctly predicts the proportion of positive class. in all samples that are actually positive class.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

iv. F1-Score is the reconciled average of precision and recall, which is a combination of precision and recall.

$$F - measure = \frac{2*p*r}{p+r} \quad (4)$$

3.6 ROC curve

The ROC curve is a widely used tool for evaluating the performance of machine learning models, particularly in classification tasks. It measures the model's ability to distinguish between positive

and negative classes, with values ranging from 0 to 1. The higher ROC scores indicate better classification performance. An ROC score of 0 suggests that the model is completely inaccurate in distinguishing [23, 24].

4 RESULTS AND DISCUSSION

4.1 Comparison of Confusion Matrix

Figure 9 shows the confusion matrix generated for each algorithm using the confusion_matrix function. The KNN model performs well in identifying unsubscribed customers but performs the worst in identifying the number of TP (85). This shows that the KNN algorithm is likely to predict

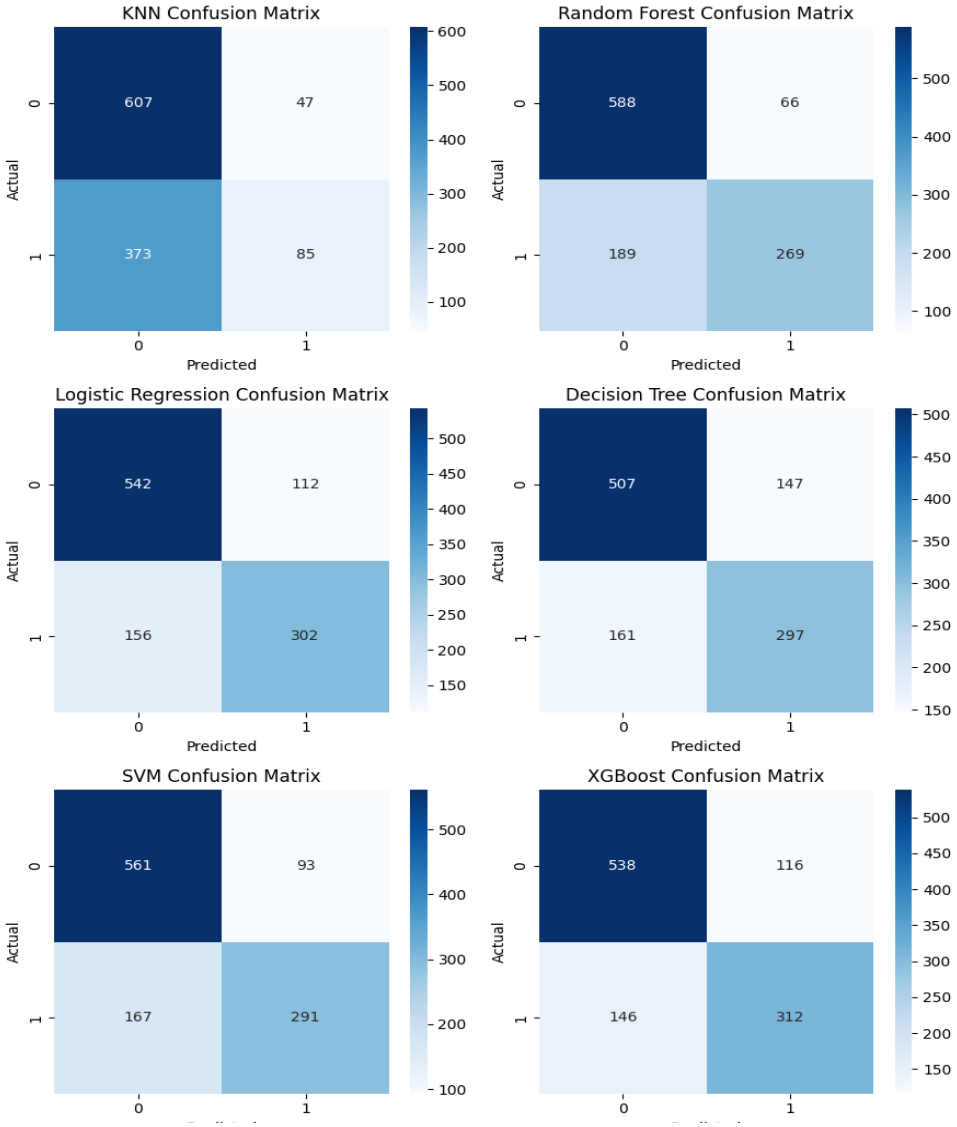

Figure 9: Confusion matrix of six models

subscribed customers as unsubscribed customers. For RF model test data set, 269 consumers were correctly classified as subscribers, 588 consumers were correctly classified as non-subscribers, 66 subscribers were misclassified as subscribers, and 189 consumers were mistakenly considered non-subscribers. The RF model has the least the number of FP (66) and the most tp plus tn. This indicates that the accuracy of this model may be relatively high.

In order to better compare the performance of each model, the accuracy, precision, recall, and F1-score calculation formulas were used, and the results are shown in Table 3. The RF model has the highest accuracy of 77.1%. LR, SVM and XGBoost also performed well, with 75.9%, 76.6% and 76.4% respectively. In terms of precision, the RF model is also much higher than the other models. For the KNN model, its accuracy is the lowest among the six models, and its prediction accuracy is obviously biased towards non-subscribing consumers. This is very unfavorable for consumer behavior analysis.

Table 3: Results of models

Model	Accuracy	Precision	Recall	F1-score
KNN	62.2%	64.4%	18.5%	28.8%
RF	77.1%	79.3%	59.6%	68.1%
DT	71.8%	65.6%	65.9%	65.8%
LR	75.9%	72.9%	65.9%	69.2%
SVM	76.6%	75.7%	63.5%	69.1%
XGBoost	76.4%	72.9%	68.1%	70.4%

4.2 Comparison of ROC score

Figure 10 shows the ROC curve comparison of six machine learning algorithms, in which XGBoost scores the highest, 0.844; The second was RF with 0.843; LR also performed well, with 0.832; KNN and DT had lower scores of 0.636 and 0.7, respectively, indicating that they were poor at distinguishing between subscribed and non-subscribed customers.

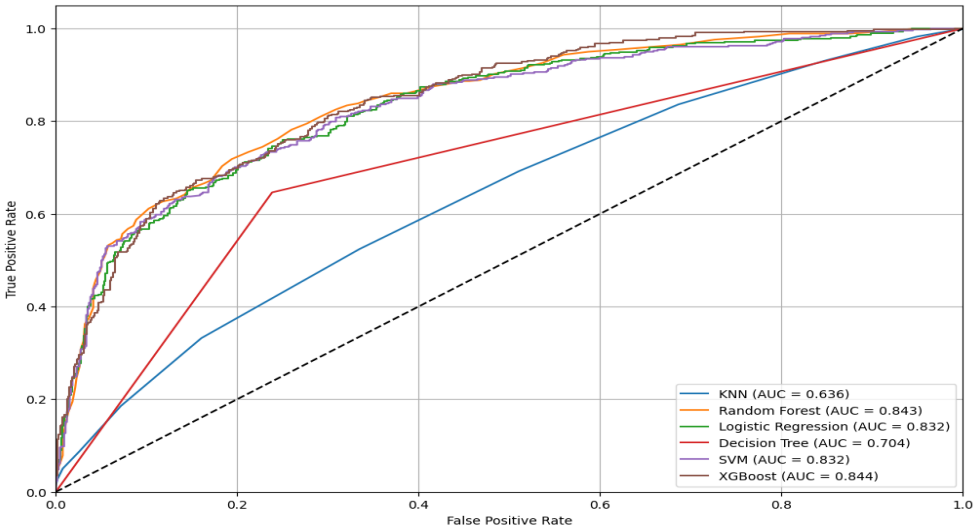

Figure 10: ROC curve comparison of six models

4.3 Champion model

Table 4 shows the comparison of the accuracy and ROC scores of the six models. The RF model has the highest accuracy and a good ROC score, so it is the champion model among the six machine learning algorithm models.

Table 4: Accuracy and ROC scores of six machine learning models

Models	Accuracy	ROC score
KNN	62.2%	0.636
RF	77.1%	0.843
DT	71.8%	0.832
LR	75.9%	0.704
SVM	76.6%	0.832
XGBoost	76.4%	0.844

4.4 Model interpretation

The prediction of machine learning models will produce black box problems [25]. The black box problem can be understood as the decision-making process inside the model is difficult to understand and explain. This study selects the RF model as the champion model and the Gini importance and Rapidminer tool are used to perform feature importance analysis on RF model. Gini Importance can provide a ranking of the relative importance of each feature in the model. Through this ranking, users can know which features have a greater impact on the prediction results of the model. Rapidminer can maximize the prediction target through automatic modeling technology to view the most important influencing factors. Figure 11 shows the Gini importance analysis. Promo code used and

discount applied are the two most important features, with important scores of 0.1525 and 0.1169, respectively. Next are the purchase amount, number of purchases, age, and consumer ratings, which contribute much more to consumers choosing to subscribe than other features.

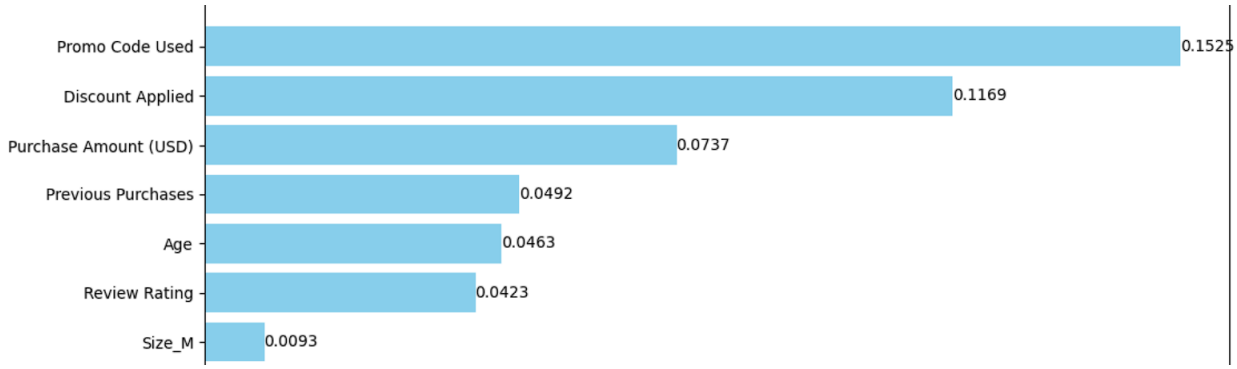

Figure 11: Gini importance analysis

As shown in Figure 12, the use of promo codes and discount applied and previous purchases are the most influential features for customers to choose to subscribe. This is consistent with the Gini Importance analysis. Next is the previous purchased feature.

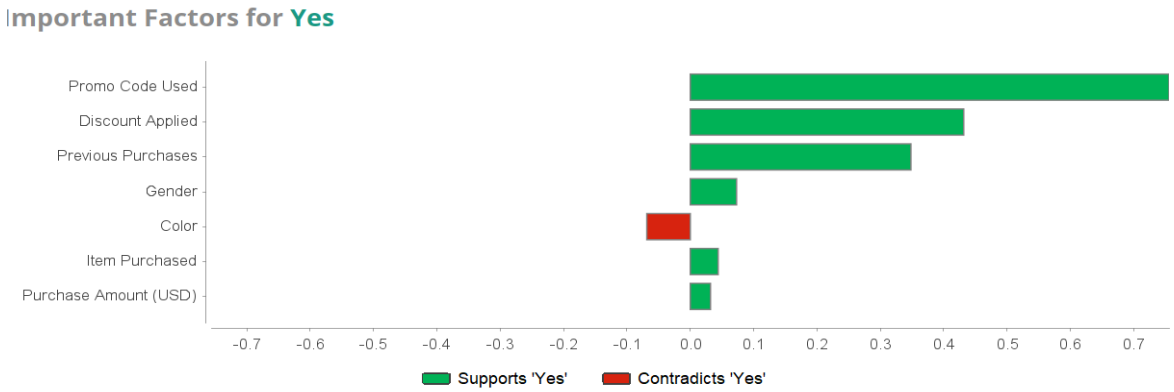

Figure 12: RapidMiner factors analysis

4.5 Discussion

This paper ensured the feasibility of creating machine learning models by preprocessing the data such as data cleaning, data encoding, data splitting, and data standardization. Six machine learning models were built. The data was divided into training and test set at a ratio of 70:30, and the number of test sets is 1112. The RF model correctly predicted 857 samples out of 1112 samples, with a prediction accuracy of 77.1%. Among the six models, the worst performance is the KNN model. Although it identified the largest number of TN (607) in the confusion matrix, indicating that it has a good performance in identifying unsubscribed customers, it identified the least number of TP (85), and the accuracy was only 62.2% after calculation. The LR model correctly identified 844 out of 1112

samples with an accuracy of 75.9%. The DT model correctly identified 804 subscribed and unsubscribed consumers with an accuracy of 71.8%. The SVM model and XGBoost model correctly predicted 852 and 850, with an accuracy of 76.6% and 76.4% respectively. In the ROC curve analysis, the RF model and the XGBoost model showed good performance, with ROC scores of 0.843 and 0.844 respectively. The KNN model performed the worst, with a score of only 0.636. The RF model performed well and was the champion model among the six models.

In the model interpretation, Gini importance and Rapidminer tools were used to analyze the feature importance of the RF model. The results show that promo code used, and discount applied have the greatest contribution to influencing consumers to choose to subscribe. Secondly, the number of previous purchases also has a certain influence on consumers. Several suggestions can be made for online clothing operations. Discount codes and discounts are important for attracting consumers to subscribe. Taking advantage of consumers' sensitivity to discounts, operations should carefully plan a variety of promotional activities and regularly launch attractive discount codes. In addition, for consumers who are good at using discount codes, a member-exclusive promotional information push should be established to predict upcoming promotional activities and corresponding discount codes in advance and cultivate consumers' continued attention to and use habits of subscription services. For people with a high repurchase rate of fast-moving consumer goods, they pay more attention to product updates, which also shows that companies should continue to innovate products to meet the needs of this group of people; for people with a low repurchase rate, the operation department can cooperate with preferential measures to attract repeat buyers to buy and increase the subscription rate.

5 CONCLUSION

Competition in the online shopping market is becoming increasingly intense, with growing rivalry among e-commerce platforms. As a result, increasing brand subscription rates has become a key strategy for enhancing customer loyalty and engagement. This study employs Python to build six machine learning models based on consumer behavior datasets to predict whether consumers will choose to subscribe. The dataset underwent pre-processing steps such as data cleaning, encoding, train-test splitting, and standardization—crucial processes for improving model accuracy and preparing for model construction. The six machine learning models developed were KNN, SVM, LR, DT, RF, and XGBoost.

Model performance was evaluated using the confusion matrix and ROC curve, with metrics such as accuracy, precision, recall, and F1-score calculated from TP, TN, FP, and FN. Experimental results showed that the RF model outperformed the others in terms of accuracy and ROC score, earning the title of champion model. Following this, Gini importance and RapidMiner were used to analyze feature importance. The top contributing factors to subscription behavior were the use of promotional codes, applied discounts, and previous purchase history. Based on these findings, the operations team can refine marketing strategies to boost customer subscriptions.

Although this study focuses on the e-commerce sector, the findings and methodology may be applicable to other industries that operate on subscription models, such as telecommunications, media streaming, and mobile applications. The predictive features identified such as response to discounts, previous purchase behavior, and demographic variables which are also relevant in these

sectors. However, generalizability would require retraining the models using domain-specific datasets to account for contextual differences in customer behavior, service types, and pricing structures. Future research may explore cross-sector comparisons to validate the robustness of these machine learning models in broader subscription contexts.

However, this study has certain limitations. The dataset lacked some relevant features, and the highest model accuracy achieved was only 77.1%. Additionally, there is a scarcity of related research on subscription behavior, limiting the broader understanding of consumer decision-making. Future research should incorporate richer datasets with more feature variables to enhance prediction accuracy and provide deeper insights into consumer behavior.

REFERENCES

[1] S. Panasenko, M. Seifullaeva, I. Ramazanov, E. Mayorova, A. Nikishin, and A. Vovk, "Impact of the pandemic on the development and regulation of electronic commerce in Russia," *International Journal of Advanced Computer Science and Applications,* vol. 13, 2022.

[2] P. Gazzola, E. Pavione, R. Pezzetti, and D. Grechi, "Trends in the fashion industry. The perception of sustainability and circular economy: A gender/generation quantitative approach," *Sustainability,* vol. 12, p. 2809, 2020.

[3] F. A. Al-Zahrani, "Subscription-based data-sharing model using blockchain and data as a service," *Ieee Access,* vol. 8, pp. 115966-115981, 2020.

[4] H.-S. Le, T.-V. H. Do, M. H. Nguyen, H.-A. Tran, T.-T. T. Pham, N. T. Nguyen *et al.*, "Predictive model for customer satisfaction analytics in e-commerce sector using machine learning and deep learning," *International Journal of Information Management Data Insights,* vol. 4, p. 100295, 2024.

[5] S. Chandra, S. Verma, W. M. Lim, S. Kumar, and N. Donthu, "Personalization in personalized marketing: Trends and ways forward," *Psychology & Marketing,* vol. 39, pp. 1529-1562, 2022.

[6] R. Sudharsan and E. Ganesh, "A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy," *Connection Science,* vol. 34, pp. 1855-1876, 2022.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory,* vol. 13, pp. 21-27, 1967.

[8] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing,* vol. 408, pp. 189-215, 2020.

[9] M. Fritz and P. D. Berger, "Will anybody buy? Logistic regression," *Improving the user experience through practical data analytics,* pp. 271-304, 2015.

[10] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[11] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine,* vol. 47, pp. 31-39, 2017.

[12] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of applied science and technology trends,* vol. 2, pp. 20-28, 2021.

[13] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2,* vol. 1, pp. 1-4, 2015.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.

[15] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of customer churn behavior in the telecommunication industry using machine learning models," *Algorithms,* vol. 17, p. 231, 2024.

[16] M. A. Al Rahib, N. Saha, R. Mia, and A. Sattar, "Customer data prediction and analysis in e-commerce using machine learning," *Bulletin of Electrical Engineering and Informatics,* vol. 13, pp. 2624-2633, 2024.

[17] J. Pamina, B. Raja, S. SathyaBama, M. Sruthi, and A. VJ, "An effective classifier for predicting churn in telecommunication," *Journal of Advanced Research in Dynamical & Control Systems,* vol. 11, 2019.

[18] S. Naveen, S. K. Ravindran, G. Shreya, and S. N. Ameen, "Effective Heart disease prediction framework using Random Forest and Logistic regression," in *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 2023, pp. 1-6.

[19] T. Zhang, S. Moro, and R. F. Ramos, "A data-driven approach to improve customer churn prediction based on telecom customer segmentation," *Future Internet,* vol. 14, p. 94, 2022.

[20] T. Ramesh, U. K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," *Malaysian Journal of Computer Science,* pp. 132-148, 2022.

[21] S. Kooptiwoot, S. Kooptiwoot, and B. Javadi, "Application of regression decision tree and machine learning algorithms to examine students' online learning preferences during COVID-19 pandemic," *International Journal of Education and Practice,* vol. 12, pp. 82-94, 2024.

[22] A. Y. W. Chong, K. W. Khaw, W. C. Yeong, and W. X. Chuah, "Customer churn prediction of telecom company using machine learning algorithms," *Journal of Soft Computing and Data Mining,* vol. 4, pp. 1-22, 2023.

[23] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?," vol. 34, ed: BMJ Publishing Group Ltd and the British Association for Accident …, 2017, pp. 357-359.

[24] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters,* vol. 27, pp. 861-874, 2006.

[25] A. M. Musolf, E. R. Holzinger, J. D. Malley, and J. E. Bailey-Wilson, "What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics," *Human Genetics,* vol. 141, pp. 1515-1528, 2022.