Product potential user prediction based on machine learning

Jingyi Yu *

Institute of Education, University College London, London, United Kingdom

* Corresponding Author Email: stnzyux@ucl.ac.uk

Abstract. In today's ever-changing business environment, it is crucial for businesses to market their products with the needs of users in mind to increase profitability. With consumers increasingly receiving services through digital platforms, analyzing potential users of products through big data has attracted extensive attention from researchers. The obvious differences in characteristics between users have greatly increased the difficulty of predicting potential users of a product. Fortunately, machine learning-based data analysis methods offer solutions to this complex task. However, there are obvious differences in usage scenarios and performance between different algorithms, which brings inconvenience to practical applications. This study delves into machine learning-based methods for product prospecting, including k -nearest neighbours (KNN), support vector machine (SVM), decision tree, XGBoost and random forest. Meanwhile, the performance of these algorithms is compared with data to analyse their advantages and disadvantages. Finally, the full paper is summarized and future research directions are envisaged.

Keywords: Machine Learning, Product Potential User Prediction, SVM, KNN, Decision Tree, XGBoost, Random Forest.

1. Introduction

Suitable marketing methods are the key to business competition. Recommender systems have become indispensable in commercial marketing with the rise of e-commerce. At the same time, facing the increasing variety of products, users often find it challenging to find suitable products in a short time, which increases the difficulty of shopping. Based on the above problems, product potential user prediction has gradually become a popular research topic. However, there are huge differences in characteristics between users and products, significantly increasing the difficulty of product prediction.

In recent years, machine learning, as an emerging technology, has gained excellent performance on tasks such as image, text, and speech. Some researchers have referenced machine learning methods to produce potential user predictions and achieved good results. Cheng-Ju Liu [1] proposed an online shopping behaviour analysis and prediction system based on machine learning methods, which predicts the probability of users buying again by learning their transaction data. Arif Furqon Nugraha Adz Zikri [2] investigated Indonesian shopping website's user behaviour to reduce. Pay Later defaults and reduces losses. Chen Qiaoshan [3] constructed a personalised marketing strategy model based on the data of Chinese consumers' buying habits and the characteristics of luxury goods. T K. Das [4] built a predictive model to evaluate customers' responses to a company's products based on past buying trends to increase product sales. However, these machine learning algorithms have significant differences in different scenarios, increasing the difficulty of application.

This paper summarises commonly used machine learning methods in the field of product prospecting. These include Decision Trees, k-nearest neighbours (KNN), Random Forests, XGBoost and Support Vector Machines (SVM). The datasets commonly used in the field of product potential users are first introduced. Then, the principles of these machine learning algorithms are presented, and finally, the data is analysed to compare the advantages and disadvantages of these algorithms.

2. Datasets

The study employed the XYZ PT e-commerce transaction data as the dataset for testing the proposed framework model. Ten characteristics total are included in the dataset (Table 1.), including categorical and numerical features, with 11,289 sessions belonging to different users.

Table 1. Dataset Feature

Features name	Description
Verification Account	Check the account using the "Already" or "Not yet" Boolean data type caption.
Gander	Gender Description
Age	Age Description
Education	Description of Last Education
E-commerce Transaction	Description of transactions in the last 6 months and 1 year
Salary	Description of Salary
Delivery Location	Delivery location consistency of Boolean data type
Length of use of E-Commerce platforms	Time spent using the E-Commerce platform
BI Check	Boolean data type: User's credit score, either good or negative

Regarding the evaluation approach, the study employed AUC and ROC. The ROC curve displays the false positive rate (FPR) and true positive rate (TPR), with the area under the curve (AUC) representing the space beneath it. FPR and TPR values are obtained from a confusion matrix that categorizes predictions into true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). The AUC is computed based on the following method:

$$AUC = \int_0^1 TPRd_{FPR} = \frac{1}{(TP+FN)(TN+FP)} \int_0^1 TPd_{FP} \quad (1)$$

The calculation of Accuracy is below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

3. Method

3.1. Decision Tree

A decision tree is an algorithm involving regression and classification. In order to draw conclusions or predictions, it recursively divides the dataset into subsets based on the values of the input features [5]. The tree structure is specifically composed of three nodes: the leaf node, the branch node, and the root node. Furthermore, it is a decision node that typically represents a particular sample characteristic that the dataset needs to be classes for. A leaf node offers a potential categorization outcome, whereas a branch indicates an alternate value for the root node. The algorithm creates the decision tree recursively after splitting the training set into subsets of comparatively pure features. In Cheng-Ju Liu's [1] study, the calculation equation using the decision tree algorithm is shown below:

$$Entropy(S) = -\sum_{i=1}^{m} p_i log_2 p_i \quad (3)$$

The information gain (S, A) represents the division of the information gain of attribute A within dataset S. This is expressed as the difference between the entropy of S and the entropy of the subset resulting from the split based on attribute A:

$$Gain(S, A) = Entropy(S) - Entropy_A(S) \quad (4)$$

Both Arif Furqon Nugraha Adz Zikri's [2] and Chen Qiaoshan's [3] studies state that the benefit of decision trees is that they can simplify complex decision-making processes. Enables decisions to explain solutions to problems better. Decision trees are also used to exploit data to discover many hidden relationships between expected input variables and target variables. The primary drawback is that when the data have unequal sample sizes for every decision tree category, the information gain conclusions are biassed towards values with greater values. Furthermore, a decision tree needs help to handle missing data and can occasionally cause overfitting issues [3].

3.2. XGBoost

XGBoost is an effective machine-learning technique that may be applied to tasks involving both regression and classification. As a member of the gradient-boosting algorithm family, it builds predictive solid models by combining weak learners (typically decision trees) iteratively. Gradient enhancement is constructed by the use of the boosting approach, which minimises mistakes or residuals in the model by starting with a new model and adding it until no errors remain [2]. Shopping cart decision trees make up the underlying layer of the XGBoost algorithm, according to Cheng-Ju Liu [1]. The overfitting issue of individual decision trees is resolved by treating these decision trees as fundamental operational "units" and combining them to produce collective decisions. The calculation formula of the merchandise purchase behaviour prediction model based on the XGBoost method is as follows.

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; \alpha_m) \quad (5)$$

In this formula, $F_m(x)$ denotes the prediction made by the mth decision tree in the XGBoost ensemble for a given input x. The prediction from the previous $m-1$th model is denoted by $F_{m-1}(x)$ The term ρ_m represents the learning rate associated with the m-th decision tree, and $h(x; \alpha_m)$ is the prediction made by the m-th decision tree with its associated parameters α_m.

XGBoost is known for its high performance, efficient handling of large datasets and speed optimization. Regularization techniques like L1 and L2 help prevent overfitting, and tree pruning controls the depth of the tree [6]. Feature importance analysis helps in understanding the contribution of variables. XGBoost supports parallel and distributed computation making it suitable for large-scale tasks and can handle missing values. However, its complexity and potential as a black-box model may present interpretation challenges for users unfamiliar with its complexity.

3.3. Random Forest

When building many decision trees during training, Random forests—an integrated learning technique in machine learning—produce a pattern of classes for classification missions or an average prediction aiming for regression. The "random" aspect of Random Forests comes from the randomness introduced in selecting data samples and the features considered in constructing individual decision trees [7]. Specifically, during training, a random portion of the training data is used to build each tree, and at each node, a random subset of attributes is taken into consideration for splitting. This unpredictability improves generalization and lessens overfitting. All of the different decision trees' predictions are combined to create the final forecast, which results in a reliable and accurate model that is less vulnerable to noise and volatility than individual decision trees.

3.4. Support Vector Machine

SVM is suitable for problems with a clear separation between different classes. In a support vector machine, the algorithm aims to find a hyperplane that best classifies the data. The hyperplane is selected to optimise the margin—the distance, also known as the support vector, between the hyperplane and the closest data point in each class.

For a linear Support vector machine: w is the weight vector, x^erepresents the input features, b is the bias term, and $\| w \|^2$ is as small as possible.

$$\begin{cases} wx^e + b \geq 1, \ positive\ cases \\ wx^e + b \leq -1, \ negative\ cases \end{cases} \tag{6}$$

The SVM classification algorithm has several benefits, such as the efficient resolution of high-dimensional and machine learning problems, nonlinear difficulties in situations with small sample sizes, and the avoidance of local minima and neural network structure selection issues [3]. In contrast to its counterparts, SVM has the largest error rate due to a high number of false positive and false negative cases [4]. SVM only takes into account the error distance from their support vectors, which is a point near the input data's decision limit. The support vector machine selects the type with the maximum error tolerance by adaptively modifying the constraints to satisfy the requirements of each category and excluding excessively large outliers [8]. As a result, the SVM automatically ignores any outliers present in the training data, which has an impact on the data.

3.5. K-Nearest Neighbors

KNN algorithm is an algorithm based on similarity decision-making. In classification and regression tasks, KNN aims to predict the outcome of new data points by using the majority class or mean of nearest neighbours in the feature space. The method first calculates the distance, usually using a distance metric like the Euclidean distance, between each data point in the training set and the input data points. Subsequently, the k data points with the shortest distance are identified and established as nearest neighbours based on feature similarity.

In order to assign the most popular class labels to the input data points for classification, KNN employs a majority vote method among these k neighbours. The regression job's method predicts the target value of the input data points by averaging (or weighting) the target values of the k nearest neighbours [9].

The prediction process is an integral part of the KNN algorithm, and the final output is either a predicted class label for classification or a predicted target value for regression. It is important to point out that the algorithm's performance is greatly impacted by the choice of parameter k, which strikes a compromise between noise reduction and model sensitivity. In addition, the choice of an appropriate distance metric is crucial to address feature scale variations and optimise the performance of the algorithm on different datasets. The formula for calculating the distance (d) is shown below:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \tag{7}$$

The KNN algorithm has obvious advantages and limitations in application. On the positive side, KNN is a simple and intuitive algorithm with minimal assumptions about data distribution. It effectively handles classification and regression tasks, making it versatile [10]. In addition, its lazy learning approach allows it to adapt to changes in the data set without retraining the model [2]. However, KNN also has its disadvantages. This algorithm can be computationally expensive, especially for large data sets, as it requires calculating distances to all data points. The choice of parameter k, which represents the number of neighbours, is critical and needs careful consideration, which will affect the sensitivity and performance of the algorithm.

3.6. Comparison

In the study conducted, five machine learning algorithms including Decision Tree, XGBoost, Random Forest, SVM and KNN - were compared based on their performance in handling tasks. The accuracy level of each algorithm was evaluated, and Random Forest was the best-performing algorithm with 93.4% accuracy. Decision Trees had the lowest accuracy at 82.8%, while XGBoost, SVM and KNN had accuracies of 89.12%, 87% and 90%, respectively [2].

Table 2. Accuracy Level of Algorithms

Algorithm	Accuracy rate
Decision Tree	82.8%
Random Forest	93.4%
XGBoost	89.12%
KNN	90%
SVM	87%

The superior accuracy of Random Forest can be attributed to its integrated learning approach, which combines multiple decision trees and introduces randomness in sample and feature selection to mitigate overfitting. While intuitive, decision trees face challenges such as biased information acquisition and difficulty dealing with missing data. XGBoost demonstrates high performance, utilising gradient enhancement and regularization techniques, but its complexity can present interpretation challenges. SVM excel in scenarios with precise classification but have high error rates, particularly in the case of false positives and false negatives. KNN, while intuitive and adaptable, faces computational challenges, especially with large datasets.

4. Conclusion

This paper focuses on machine learning methods for predicting potential users of products. It mainly includes algorithms such as KNN, SVM, Decision Tree, XGBoost, and Random Forest. A comparative analysis of five machine learning algorithms revealed their performance in handling classification tasks, with Random Forest being the best-performing algorithm. However, each algorithm has its own unique advantages and limitations, highlighting the complexity of algorithm selection in real-world applications.

The poor interpretability of some machine learning algorithms, such as Random Forest and XGBoost, increases the risk in practical applications. Future research can start from the interpretability of the model to improve the stability of the model. In addition, with the popularity of edge devices, the demand for models with low power consumption gradually increases. Therefore, the lightweight training and deployment of models is a worthy research direction.

References

[1] Cheng-Ju Liu, Tien-Shou Huang, Ping-Tsan Ho, Jui-Chan Huang, Ching-Tang Hsieh. Machine learning-based e-commerce platform repurchase customer prediction model. PLoS ONE. 2020, 15 (12): 1 – 15.

[2] Arif Furqon Nugraha Adz Zikri1, Wiwin Suwarningsih. Pay Later Risk Management: A Review of FMECA and Potential Customer Prediction Frameworks Through the Application of Machine Learning. International Journal of Advances in Data and Information Systems. 2023, 4 (2):167 - 180.

[3] Qiaoshan Chen, Shousong Cai, Xiaomin Gu. Construction of the Luxury Marketing Model Based on Machine Learning Classification Algorithm, Scientific Programming. 2021, (2021): 11.

[4] T.K. Das. A Customer Classification Prediction Model Based on Machine Learning Techniques. International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).2015, 321 - 326.

[5] Kyoungok Kim. A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. Pattern Recognition. 2016, 60: 157 – 163.

[6] Pesantez-Narvaez J, Guillen M, Alcañiz M. Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. Risks. 2019, 7 (2): 70.

[7] Louppe, Gilles. Understanding random forests: From theory to practice. arXiv preprint arXiv. 2014: 1407. 7502.

[8] Chun-fu Lin, Sheng-de Wang. Training algorithms for fuzzy support vector machines with noisy data. Pattern Recognition Letters. 2004, 25: 1647 – 1656.

[9] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. Learning k for kNN Classification. ACM Trans. 2017, 8: 1 – 19.

[10] Gavin Hackeling. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd. 2017, 33.