ISSN 2959-6157

Applications of Machine Learning Algorithms in Predicting User's Purchasing Behavior

Ranzhi Sun[1,*]

[1]High School Affiliated to Nanjing Normal University, Nanjing, China
*Corresponding author: 002142@nuist.edu.cn

Abstract:

With the rapid development of big data in the Internet era, accurately identifying consumers' purchase intention and predicting their future purchase behavior among the massive user behaviors are crucial for business decisions. The purpose of this paper is to analyze the advantages and disadvantages of multiple supervised learning algorithms and integrated learning algorithms, as well as their applications and performances in predicting users' purchasing behaviors. The paper concludes that some traditional algorithms have been consistently used due to their simplicity and interpretability, while the more cutting-edge algorithms have a greater advantage in characterizing specific aspects around an innovative core idea. Different algorithms have their own highlights and limitations in prediction, and researchers can choose them according to the dataset and prediction needs. At the same time, this paper emphasizes that combining models that complement each other's strengths will maximize efficiency and accuracy when using fusion methods. This paper compiles and compares practical machine learning algorithms today, and analyzes the future direction of predictive modeling and areas worthy of further exploration, such as language and image processing, which can provide a reference for enterprises with the need of user behavior prediction in the development of marketing plans.

Keywords: Machine learning, User behavior, Purchasing prediction, Model integration

1. Introduction

In recent years, e-commerce platforms have developed in a diversified direction with the updated iteration of the Internet. This has attracted more and more customers to learn about products and make purchases on the Internet. As with the offline model, customers rely on their own needs, preferences, and other factors when making decisions, which allows each customer to develop a set of unique behavioral patterns. Big data makes these personalized behaviors easier to record, integrate and analyze. On the basis of such a huge amount of data, it is inevitable to consider predicting the next movements of customers, such as the collation of products preferred for browsing, the range of pricing accepted, or whether there will be repurchase behavior. Such predictions are personalized recommendations to customers, which can be described as „tailored", and enhance customer experience, maintain customer loyalty, tap potential customers and increase merchant revenue.

Machine learning, as an intelligent means, can be used to discover patterns between data by learning from historical data to achieve automated predictions. At the same time, no matter whether the data is large or small, there are corresponding models to improve the accuracy of the prediction, through which more accurate recommendations can be realized. Recommendation systems were first proposed by Resnick in 1997[1]. Starting from the historical behavior data of customers, these systems analyze to determine the types of products the user is interested in and then recommend other related products that the user might be interested in to them. Traditional recommendation system algorithms mainly include five kinds of content-based recommendation, association rule-based recommendation, knowledge-based recommendation, collaborative filtering recommendation and hybrid recommendation [1]. Based on the recommendation system, the research focusing on "user purchase behavior prediction" using machine learning knowledge is developing rapidly. Lv Zhipeng et al. deeply mined decision tree model found that the algorithm is easy to understand and practical and effective, in the help of decision-making has significant application value [2].

Meanwhile, many researchers have improved and integrated the traditional algorithms to improve the accuracy and interpretability of predictions further. With the popularity of deep learning algorithms, deep forest was proposed by Professor Zhou Zhihua in 2017, and Fu Hongyu

et al. then introduced integration algorithms such as XGboost with the concept of time sliding window to enhance the variability of input features and the model's ability to process the features and established a user purchase behavior prediction model based on deep forest [1]. In terms of feature engineering, Zhou Chengji proposed a new feature algorithm SSP based on Bagging strategy, which improves the prediction accuracy while reducing the complexity [3]. In terms of neural network, Hu Xiaoli et al. focus on the defect that the assumption of independence of users' purchasing behavior sometimes does not match with the actual purchasing situation and propose a CNN-LSTM neural network combination model to simplify the process of feature selection [4]. In terms of integration learning, Chen Long fused three models, XGB, LGB, and CatBoost, to obtain a combined model for predicting the repurchase problem and explored how to maximize the enhancement through the fusion method [5].

The purpose of this paper is to explore the application of machine learning algorithms in predicting user buying behavior. First, the concepts and advantages and disadvantages of various machine learning algorithms are introduced. Second, based on the factors influencing customer behavior, such as purchase history and product information, the application and performance of different machine learning algorithms in predicting user purchasing behavior are analyzed. Finally, possible improvements are suggested for the algorithms themselves and their applications, and the future directions in which this research area can be studied in depth are looked forward to.

2. Overview of Machine Learning Algorithms

Machine learning is when a computer learns from data and improves. Machine learning algorithms include supervised learning algorithms, unsupervised learning algorithms, deep learning algorithms, and integrated learning algorithms. In essence, these algorithms look for patterns in given data, integrate it, and then use the identified patterns to make predictions about unknown data.

2.1 Supervised Learning Algorithms

The basic idea of supervised learning is to analyze the relationship between input features and corresponding output markers from labeled training data, while detecting the degree of error and adjusting its own parameters to form a set of models. Then, based on this model, new unlabeled data will be classified and predicted. Supervised learning algorithms include linear regression, logistic regression, decision trees, support vector machines, neural network algorithms, etc. This paper focuses on decision tree and neural network algorithms.

Although logistic regression has "regression" in its name, it is a nonlinear classification model and is often used to solve binary classification problems. Logistic regression is based on the Sigmoid function, which maps a combination of input features to a probability value between 0 and 1, and then forms a continuous probability distribution. The model estimates the parameters by great likelihood, and after training is completed, new samples are classified and mapped to between 0 and 1 for binary prediction. Logistic regression is popular with many machine learning researchers due to its simplicity and interpretability.

The decision tree algorithm is a supervised learning algorithm for solving classification problems in a tree-like structure with root nodes, inner nodes (secondary root nodes), leaf nodes and branches. Where each root node represents a feature attribute and each leaf node represents a category or label. The principle of the decision tree lies in the gradual division of the input features, which are traversed from top to bottom until a specific stopping condition according to the decision rule. As shown in Fig. 1, the tree structure of the decision tree model is intuitive and easy to understand and does not require forced standardization or normalization of data. However, when the depth of the tree is too large, it is prone to overfitting.

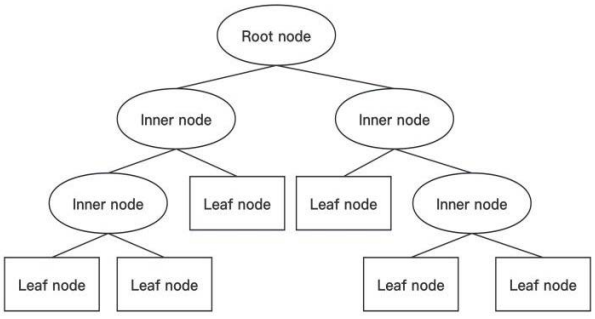

Fig.1 Decision tree

Support vector machines are supervised learning algorithms and are particularly effective in classification tasks. The core principle of a support vector machine is to find a decision boundary, i.e., a hyperplane, such that the distance from the closest sample points to the decision boundary is maximized. These sample points are the support vectors. At the same time, support vector machines are good at dealing with nonlinear problems, and their kernel trick is able to cope with more complex data by mapping a low-dimensional nonlinearly differentiable problem to a high-dimensional linearly differentiable problem. However, the model is highly dependent on appropriate regularization parameter settings otherwise the performance is easily affected.

Neural network algorithms also belong to the category of supervised learning, through which classified labels can

be used to adjust neural network parameters to improve prediction accuracy. The neural network model consists of a lattice of neurons in multiple layers (input, hidden, and output), and the principle is that the neurons in the next layer receive the inputs from the previous layer, are weighted and, summed and processed with an activation function, and then passed to the next layer, and so on (Fig. 2).

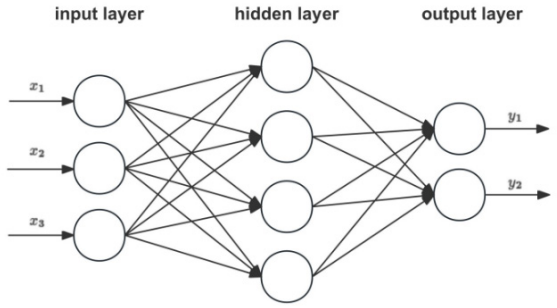

Fig.2 Neural network structure

The core of the neural network model lies in the back-propagation algorithm, a method that allows the prediction results to approximate the true label constantly. Compared with traditional machine learning algorithms, neural network algorithms benefit from their strong fitting ability and perform better in speech recognition, image recognition, recommendation systems and other fields. Common neural network structures include convolutional neural networks, recurrent neural networks, feedforward neural networks, and so on.

2.2 Integrated Learning Algorithms

The core idea of integrated learning algorithms is to improve the prediction performance by combining multiple underlying prediction models. So, it becomes one of the challenges of integrated learning algorithms to effectively integrate the prediction results of multiple weak learners to get a better-integrated model. Usually, integrated learning models have good robustness and generalization ability, and are widely used in multi-class machine learning projects. The integrated learning algorithms generally include automatic aggregation, gradient boosting, and stacked generalization. In this paper, we introduce the corresponding algorithms for each of the above three methods.

Random forest constructs multiple decision trees by random sampling and random feature selection to accomplish regression or classification tasks, with no connection between different decision trees. Subsequently, the random forest will count the results of all decision trees and use voting to determine the final category to get the prediction result. Usually, Random Forest has a high accuracy rate, and the combination of multiple decision trees prevents the risk of overfitting, but this also increases the model complexity and training time.

Gradient boosting decision tree (GBDT) is a popular integrated learning algorithm which corrects the residuals by adding decision tree models in an orderly manner. It improves the performance compared with the traditional base model. The mainstream models of GBDT mentioned in this paper include XGBoost, LightGBM and CatBoost.

The core principle of XGBoost lies in serially training multiple decision tree models and weighting and summing the results. At the same time, L1 regularization (Lasso) and L2 regularization (Ridge) are introduced into the objective function to prevent overfitting. XGBoost combines the greedy algorithm to select the optimal split point of each node and can process the node splits of each tree in parallel, which greatly improves the training efficiency.

LightGBM was open-sourced and proposed by Microsoft in 2017[6], which also belongs to the GBDT algorithm. The core of LightGBM's high efficiency and accuracy lies in using a histogram-based decision tree algorithm, which constructs a histogram by binning continuous features to improve the data robustness and stability, and to reduce the memory occupation. For node splitting, LightGBM selects the leaf node splitting with the largest absolute value through the Leaf-wise growth strategy to minimize the depth of the tree. Therefore, compared with traditional GBDT, LightGBM is faster in training and superior in dealing with large-scale data and high-dimensional features.

CatBoost belongs to the mainstream GBDT models, XGBoost, and LightGBM[6]. The core principle of CatBoost is a symmetric tree structure, where each node remains symmetric during splitting, which reduces computational complexity and prevents overfitting. In addition, the advantage of CatBoost is that there is no need to tune the model or perform preprocessing such as solo thermal encoding, which improves the efficiency of model debugging.

3. User Purchasing Behavior Prediction and Application

3.1 Influencing Factors

Large and reliable data is indispensable in model training. Therefore, in order to make predictions about the future behavior of users more accurately, researchers are targeted to find datasets that meet their needs. In other words, the variables in these data are influences that have a strong correlation with user behavior and its effects. Since the prediction focuses on the relationship between customers and goods, the influencing factors can be discussed based

on both users and goods.

The factors that influence users can be divided into two main categories. One category is user information (or user profile), including user id, age, gender, city, device model, and so on. These objective basic information can help researchers determine user preferences more intuitively and conduct preliminary analysis of user behavior[7]. The other category is user behavior, including user id, login status, clicking behavior, add purchase or order behavior, historical purchase information, and whether it is a repeat buyer.

For commodities, the influencing factors that can be used for analysis include product id, category, brand, price, number of reviews, and rate of bad reviews. For example, the category tag can be linked to the user's historical purchase preference, the price can be compared to the user's purchasing power, and the bad review rate can reflect the popularity of the item. In short, all these factors can be used to refine the product characteristics, thus refining the recommendation algorithm and improving the model relevance.

3.2 Feature Engineering

In general, the original dataset consisting of the influencing factors mentioned above does not serve as a training sample for the model, and further extraction of effective features is required to specify a mineable data sample, i.e., feature engineering[8]. Feature engineering involves selecting and transforming variables in the raw data to construct potential problems. Better feature clusters can shorten parameter search time and reduce model complexity while better describing the data to improve model performance[9].

Feature construction is one of the core aspects of feature engineering and is usually performed early in the project to provide more effective features for subsequent feature selection. Feature construction aims to obtain more information from the original data. Currently, the more rudimentary methods include grouping construction, i.e., grouping numerical features using category features and forming new features in the process; and combination construction, which combines two or more features through operations such as addition, subtraction, multiplication, and division. In addition, as an enhancement, many researchers have introduced the time sliding window as a method of feature construction, which is used to count the cumulative features in different time ranges[10], reflecting the temporal relationship between user data, such as the average browsing time interval and the average purchasing time interval[1].

Feature selection is another core aspect, aiming to filter out a subset of eligible features for training, while removing negatively correlated features to simplify the model and improve accuracy[3]. The more popular feature selection methods include recursive feature elimination (RFE), whose core idea lies in repeating the iteration and selecting the most qualified features, with the advantage of automatic identification, which can effectively reduce the number of irrelevant features. However, RFE is mostly used in regression models and SVMs, and its default sequential ordering makes it difficult to compare the score gap between different features and unsuitable for integrated algorithms such as GBDT[3]. The other is the stable selection method, which combines a regularization method with multiple runs on different subsets to count how often each feature is selected, with scores ranging from useless to useful, from 0 to 1[11-12]. The method is more widely applicable. Based on XGBoost, Zhou Chengji combined the stable selection method and Pearson correlation coefficient in order to get a determinant, constructed a new feature algorithm SSP, the basic idea lies in summing the Pearson correlation coefficients of a certain feature with other features, and then dividing it by the stability selection influence factor of the feature, which solves the problem of covariance and reduces the difficulty of the final feature selection[3].

3.3 Comparison of the Application and Performance of Different Machine Learning Algorithms in Prediction

3.3.1 Decision tree

As a traditional machine learning algorithm, a decision tree analyzes the correlation between users' purchasing decisions and related features through classification. Although the accuracy is not high compared to cutting-edge algorithms, the model is simple and easy to interpret and the results are more intuitive.

The model can be based on Python language. Firstly, the training set and test set are divided, the model is built through the training set and the performance is evaluated through the test set. In this regard, Lv Zhipeng et al. explored the user's purchase decision on customized home furnishings, using the brand as the root node, and the important internal nodes, including age, gender, selection, occupation, price, etc., and went through six layers of judgments to arrive at a prediction of the purchase intention. As for the results, the accuracy of both the training and test sets is above 85%, indicating that the decision tree model can better predict the user's purchase behavior[2].

3.3.2 Support vector machine

Support vector machine is also a more maturely developed classification model, which is relatively simple, but

its hyperplane idea can also cope with complex high-dimensional data. For the prediction of user repurchase behavior, Zhang Zongyao, in order to study and correctly predict the repurchase behavior of customers, based on the support vector machine, divides the linearly divisible data in the training set as much as possible and constructs the model under the spark computation engine by using the local vector LabeledPoint of the scala language, and iterates 5000 times for learning. The final score cannot be observed as a distinctive feature, so it needs to be normalized to analyze the purchase intention with a threshold of 65 points[13], which is very intuitive and easy to compare.

3.3.3 Combined random forest and logistic regression models

Random forest is an integrated algorithm consisting of multiple decision trees, and Logistic regression specializes in nonlinear classification. In order to accurately identify users, predict future demand, and formulate business strategies, Zhang Feng et al. combined the two and added the prediction results of Random Forest as a new variable to the Logistic model while retaining the original input variables of Logistic regression for the dataset. Finally, through the ROC curve, the combination model was analyzed to reach an accuracy of 0.9485, higher than the other models[14].

3.3.4 Neural network algorithms

Neural network algorithms are more accurate but also more complex than the above-mentioned algorithms. Taking CNN as an example, its convolutional layer can effectively capture data and extract high-impact features to better cope with large-scale data, especially in today's era of data explosion. Based on CNN, Hu Xiaoli et al. applied a combined CNN-LSTM framework to predict the purchase behavior of users on e-commerce platforms for the purpose of improving user experience and marketing effectiveness, with the original features including the number of views, number of purchases, and browsing-buying conversion rate, etc. The LSTM predicts important features extracted with the CNN, and then converts them into feature vectors through the high-dimensional outputs of the LSTM at the full join layer, which expresses the results of whether users purchase results. The accuracy of this composite framework is higher than 0.85 for both the training and test sets, which is an improvement compared to the benchmark model, and at the same time outperforms supervised or integrated learning models such as Random Forests, Support Vector Machines, XGBoost, etc [4]. This segmented sampling method effectively solves the problem of imbalance between purchased and unpurchased samples.

3.3.5 Gradient boosting decision tree

Gradient boosting decision tree (GBDT) contains three mainstream integrated algorithms: XGBoost, LightGBM, and CatBoost.XGBoost belongs to one kind of gradient boosting tree, which has the ability of multi-thread processing. It first adds new trees through multiple rounds of iterations to predict the residuals between the true value and the current prediction and then uses a greedy algorithm to obtain the optimal split point, but such a large number of traversals also take up more memory. Because XGBoost relies on parameter tuning, the researchers need to determine the number of classifiers and the early stopping condition, followed by sequentially tuning the Booster class parameter, the gamma parameter, and the subsample and colsample_bytree parameters, and finally tuning the learning rate to near 0.01 to obtain the best model performance[15]. LightGBM uses the histogram algorithm instead of the sorting algorithm, greatly improving the training speed. In addition, the model's one-sided sampling (GOSS) allows selective sampling based on the gradient, retaining samples with large prediction errors. Its mutually exclusive feature bundle (EFB) can also effectively reduce the feature dimensions, and CatBoost can automatically process the category features, eliminating the preprocessing time, and using the structure of a symmetric tree to maximize the speed of model construction.
Chen Long explored this issue in depth and modeled it using XGBoost, LightGBM and CatBoost respectively, and found that all three GBDT models outperformed a single logistic regression model. The average AUCs of the three models on the five-fold hierarchical training and test sets are 0.6767, 0.6766, and 0.6789, respectively, which is not a big difference[5].

3.3.6 Stacking integration

Stacking integration is training multiple models and combining the results with certain strategies. The model generally consists of two layers of learning up, the base learner and the secondary learner, and the principle is to use the output of the base learner as the input of the secondary learner, so as to correct the prediction bias of different models and improve the accuracy[16]. In terms of application, in order to study how much the combined model improves the machine learning ability, Li Linyan used four models, namely Support Vector Machines, Random Forest, XGBoost, and LightGBM, for the base learner of Stacking fusion, which yielded an AUC value of 0.7804 for the Stacking model, which is superior to the independent AUC values of the remaining four baseline models, indicating that the Stacking integration has better results

in user repeat purchase behavior prediction[14]. Also using this fusion method, Chen Long substitutes the three primary classifiers of GBDT and obtains an AUC value of 0.6766 for the test set, which is almost no improvement compared to a single model[5]. Thus, models with similar functionality are sometimes not suitable for integration, and often benchmark models with some differences in principles can be enhanced in Stacking integration.

4. Suggestions

From the previous discussion, it can be observed that in most cases, improved models perform better in predicting user buying behavior than a single baseline model. These methods include using different models in segments or fusing multiple models to complement each other's strengths. Therefore, after verifying the feasibility, researchers can try to innovatively fuse individual models to take advantage of the mature theoretical foundation of traditional models while utilizing the superior learning capabilities of frontier models.

The data used in many of the references in this paper come from real data from major e-commerce platforms. In the era of big data, these online shopping platforms are closely related to people's lives, and the prediction methods obtained through the study of machine learning models can be used as a solid foundation, and after configuring the appropriate extensions, they can be applied to the back-end of each merchant to provide accurate and efficient predictions in terms of sales forecasts, pricing strategies, customer segmentation, personalized recommendations, etc.

With the development of machine learning technology, the prediction of user buying behavior will become more and more accurate. From the customer's perspective, this provides behavioral patterns. Further, it helps merchants optimize their product lines, maintain customer satisfaction and loyalty, grow their business and enhance market competitiveness. Meanwhile, timeliness will also become a worthy direction for future in-depth research, with improved real-time data processing capabilities representing dynamic recommended content, facilitating the capture of more available marketing opportunities.

In addition, neural networks and deep learning-related technologies have already demonstrated strong computing power in the field of image recognition. For example, Convolutional Neural Networks (CNNs) can accurately recognize and analyze images uploaded by users on social media in terms of objects, emotions, etc., to speculate on products that may interest them. On the other hand, Natural Language Processing (NLP) can analyze massive user reviews to determine the hotness of the product, user satisfaction, and purchase intention. The convergence of these technologies is expected to facilitate multimodal data analysis and make prediction results more accurate and comprehensive. As a result, future predictions can more frequently incorporate image recognition to analyze user preferences or language processing technology to dissect user feedback, which can enhance user experience while supporting enterprises in formulating strategies.

5. Conclusion

With the theme of machine learning algorithms, this paper parses several popular supervised learning algorithms and integrated learning algorithms by laying out the influencing factors and feature engineering, aiming at comparing the application and performance of different machine learning algorithms in predicting users' purchasing behaviors. This paper finds that even with the rapid change of times, some traditional machine learning algorithms, such as logistic regression and decision trees, can still serve as benchmark models and play an irreplaceable role in prediction due to their intuitive nature and strong interpretability. They combine the advantages of the base algorithms for numerous integrated learning algorithms. Then each is based on a specific core idea, such as XGBoost using the greedy algorithm to find the optimal splitting point. Lightgbm introduces the histogram algorithm, which effectively improves the model prediction efficiency and accuracy, and has a strong generalization ability. In addition, innovative fusion methods such as Stacking integration can drastically improve the machine learning capability. However, this also requires a suitable combination of methods, the principle of similar algorithms sometimes can not be fused to produce a more powerful model, only to take into account a number of different categories of algorithms, in the case of complementary advantages, in order to play the strengths of fusion algorithms truly. For the research of machine learning in prediction, there are more deep learning algorithms for researchers to discuss. Using neural network algorithms as an example, image and language recognition are expected to be further promoted in future models to analyze and predict consumer purchasing behavior from multiple angles.

References

[1] Fu H Y, He H. The application of deep forest in the prediction of user's purchase behavior[J]. Computer Applications and Software,2023,40(01):298-305.

[2] Lv Z P, Zheng D D, Guo Q et al. Research on the prediction of the whole customized home purchase disicion by decision tree algorithm[J]. China Forest Products Industry,2023,60(05):88-92.

[3] Zhou C J. Design of predictive model of commodity

purchase behavior based on machine learning[D]. Guangzhou University,2019.
[4] Hu X L. Prediction model of user buying behavior based on CNN-LSTM[J]. Computer Applications and Software,2020,37(06):59-64.
[5] Chen L. Prediciton of user repurchase behavior based on machine learning methods[D]. Nankai University,2022.
[6] Shao Y L. Research on prediction of E-commerce user buying behavior based on multi-model fusion[D]. Dongbei University of Finance and Economics,2023.
[7] Qi W Y. Prediction of online course purchasing behavior based on integrated learning[D]. Huazhong Agricultural University,2023.
[8] Yang L. Research on the application of machine learning in E-commerce user's purchase behavior prediction[D]. Tianjing University of Commerce,2022.
[9] Bahnsen A C, Aouada D, Stojanoviic A, et al. Feature engineering strategies for credit card fraud detection[J]. Expert Systems with Applications, 2016, 51: 134-142.
[10] Wang Y H. Research on user purchase behavior prediction based on LightGBM[D]. Lanzhou University,2022.
[11] Kalousis A, Prados J, Hilario M. Stability of Feature Selection Algorithms[C]. IEEE International Conference on Data Mining. IEEE Computer Society, 2005: 218-225.
[12] Hofner B, Boccuto L, Göker M. Controlling false discoveries in high-dimensional situations: boosting with stability selection[J]. Bmc Bioinformatics, 2015, 16(1): 144.
[13] Zhang Z Y. Prediction of user repurchase behavior based on support vector machines[J]. Computer Programming Skills & Mantenance,2020(06):3-6+21.
[14] Zhang F, Zhang L N, Li J J. Prediction of user consumption behavior based on machine learning combinantion model[J]. Technology of IoT & AI,2022,5(02):19-27.
[15] Jing X L, Shi M X. Repurchase prediction of E-commerce user based on XGBoost[J]. Journal of Liaoning University(Natural Sciences Edtion,2023,50(02):134-145.
[16] Li L Y. Research on user repeat purchase behavior prediction based on stacking fusion model[D]. Zhongnan University of Economics and Law,2023.