Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/47/20241338

Analysis of buying intension online based on decision tree and support vector machine

Linjing Zhang

School of Medical information engineering, Jining Medical University, Qingdao, 266000, China

zhanglinjing@stu.mail.jnmc.edu.cn

Abstract. With the development of the Internet, online shopping has gradually emerged and changed people's living habits. On this basis, the purchase intention of users has become a hot topic. In this study, the data set on Kaggle is used to train and analyse the model with decision tree as well as support vector machine (SVM). To be specific, 30% of the data is used for testing. According to the feature correlation analysis, one can see that some features, such as ExitRates and BounceRates, are highly correlated, and PageValue has a great impact on income. In addition, precision, recall, f1-score and other data were processed to evaluate the model, and its limitations and future development were explained. It is of great significance to forecast the income of online merchants and how merchants purchase goods and reduce inventory, and to understand and predict the intention of users to purchase online from the income of merchants.

Keywords: decision tree, support vector machine, buying intension, data analysis.

1. Introduction

The rapid development of science and technology has changed people's lives. People have electronic products and cannot live without the Internet. Among them, website has changed the living habits, people can browse the website and buy products at home. Before the 1980s, most consumers learned about product information through TV advertisements, and the market was relatively small and not much concerned by people. In 1978, Amazon, the first B2C outlet in the United States, was established and began to operate online business in Seattle in 1979, and the first distribution centre was established in California in 1981. In 2001, Germany's largest e-commerce district was listed on the London Stock Exchange, and in 2003, Taobao was officially launched in China, and international brands entered the Chinese market, which made the development of online shopping reach its peak.

More and more people accept online shopping because it is more comfortable than traditional shopping, which is often considered anxious, crowded and so on [1]. Under the background of the steady increase of online sales, there is a huge untapped market potential for online shopping [2]. One of the challenges facing e-commerce is to predict the purchasing power of customers so that businesses more reasonable purchases [3]. Researchers use machine learning techniques to study such problems. The performance of the model can be improved by first identifying the problem and collecting data, then preprocessing the data, and selecting and building a set of features that are meaningful to the problem. Models are then selected, trained, evaluated, and tuned. Finally, the model is applied to the reality to predict the demand of consumers, so that merchants can purchase and sell reasonably, reduce inventory

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/47/20241338

maintenance costs, and increase online shopping revenue. Suthep Nimsai and other researchers used decision tree and other four methods to find the factors influencing the purchase intention of premium beef consumers, and finally came up with price and product age [4]. Bikram Prasad and his team considered attributes of online buying behavior such as transaction security, innovative product availability, and product quality, using ANN to gain insight into market architecture and customer choice [5]. Šebalj, Dario, Franjković and Jelena Hodak, Kristinato created a model. They find that the monetary value and the most important variable in the classification of shopping intentions is the price level within the store [6].

Therefore, it is necessary to obtain certain data sets first, prepare to use decision tree and support vector machine (SVM) to predict, and find important influencing factors. For example, how to compare the same type of products, or how to choose different stores under the same product and so on. The remainder of this paper is organized as follows. The second section is data and methods. The third section analyses the results and discusses, including feature engineering, model performance and comparison. The fourth section gives the limitation of the method and the prospect of the future. This paper gives a summary of the conclusion.

2. Data and method

The data set comes from Kaggle, and it is the purchase intention data set of online shoppers. The data set consists of 14 numerical attributes and 4 categorical attributes, but this article uses numerical attributes only, and the last 30% is used to test the model. The revenue attribute can be used for class labels. First, code to ensure the integrity of the data, and then think about what ICONS to use to describe the data set. Then understand the data values such as: administrative, information, product-related and their time all describe the different pages visited by the user and the total time spent browsing the different pages. Through the study of merchant revenue, it is necessary to analyse which characteristics have an important impact on users' online purchase intention.

Decision tree is a machine learning method that can continuously segment data according to a certain parameter. A tree is made up of nodes and leaves. The decision node is the location of the split data [7]. The essence of this method is to use Gini index. Suppose the dataset D has K classes, and the probability that the sample points belong to class k isp_k, then the Gini value of the probability distribution is defined as:

$$Gini(D) = \sum_{K=1}^{K} p_k(1 - p_k) = 1 - \sum_{K=1}^{K} p_k^2. \quad (1)$$

The abbreviation of support vector machine is SVM, and it can be used kernel tricks to transform the data for classification or regression problems [8]. The goal of it is to determine an optimal separation hyperplane to maximize the margin between different classes of training data. SVM is useful for text and hypertext classification, face detection, image classification, recognition of handwritten characters, cancer diagnosis and prognosis. SVM is essentially a classification method, with $w^T + b$ to define the classification function, so find $w$and$b$ in order to find the maximum interval, derive $||w||^2/2$, and then introduce Lagrange factor, become the solution of Lagrange multiplier a. One chooses a_j^* and the component vector satisfies$0 < a_j^* < K$ ($K$is the kernel function) computes

$$b^* = y_j - \sum_{i=1}^{N} a_i^* y_i K(x_i, x_j) \quad (2)$$

Finally, classification is performed by classification decision function:

$$f(x) = sign\left(\sum_{i=1}^{N} a_i^* y_i K(x, x_i) + b^*\right) \quad (3)$$

3. Results and discussion

Due to the large number of features, the decision tree cannot be used for non-data classification, so the following 14 features are selected for training the model and data analysis. Std is the standard deviation, and the remaining four rows represent quantiles of different degrees. By observing the maximum, minimum and quartile points of the data in Table 1 and Table 2, one can know: the maximum value of

DOI: 10.54254/2755-2721/47/20241338

Administrative_Duration, Informational_Duration, ProductRelated_Duration greatly affects the corresponding average, making the average far greater than the median. The maximum value of PageValues, ProductRelated and BounceRates affects the average value of ProductRelated to some extent. The maximum values of other features have little effect on the mean value. In order to understand the correlation between the features, the correlation coefficients of the features were calculated with python code. The correlation coefficients with stronger correlations are rounded to keep the higher ones, and the correlation coefficients with less correlations are kept the lower ones, as shown in Figure 1.

Table 1. Description of features Part1.

	Administrative	Administrative_Duration	Informational	Informational_Duration	Product Related	ProductRelated_Duration	Bounce Rates
count	12330	12330	12330	12330	12330	12330	12330
mean	2.3152	80.8186	0.5036	34.4724	31.7315	1194.7462	0.0222
std	3.3218	176.7791	1.2702	140.74934	44.4755	1913.6693	0.0485
min	0	0	0	0	0	0	0
25%	0	0	0	0	7	184.1375	0
50%	1	7.5	0	0	18	598.9369	0.0031
75%	4	93.2563	0	0	38	1464.1572	0.0168
max	27	3398.75	24	2549.375	705	63973.522	0.2

Table 2. Description of features Part2.

	ExitRates	PageValues	SpecialDay	Operating Systems	Browser	Region	TrafficType
count	12330	12330	12330	12330	12330	12330	12330
mean	0.0431	5.8894	0.0614	2.1240	2.3571	3.1473	4.0696
std	0.0486	18.5684	0.1989	0.9113	1.7173	2.4016	4.0252
min	0	0	0	1	1	1	1
25%	0.0143	0	0	2	2	1	2
50%	0.0252	0	0	2	2	3	2
75%	0.05	0	0	3	2	4	4
max	0.2	361.7637	1	8	13	9	20

DOI: 10.54254/2755-2721/47/20241338

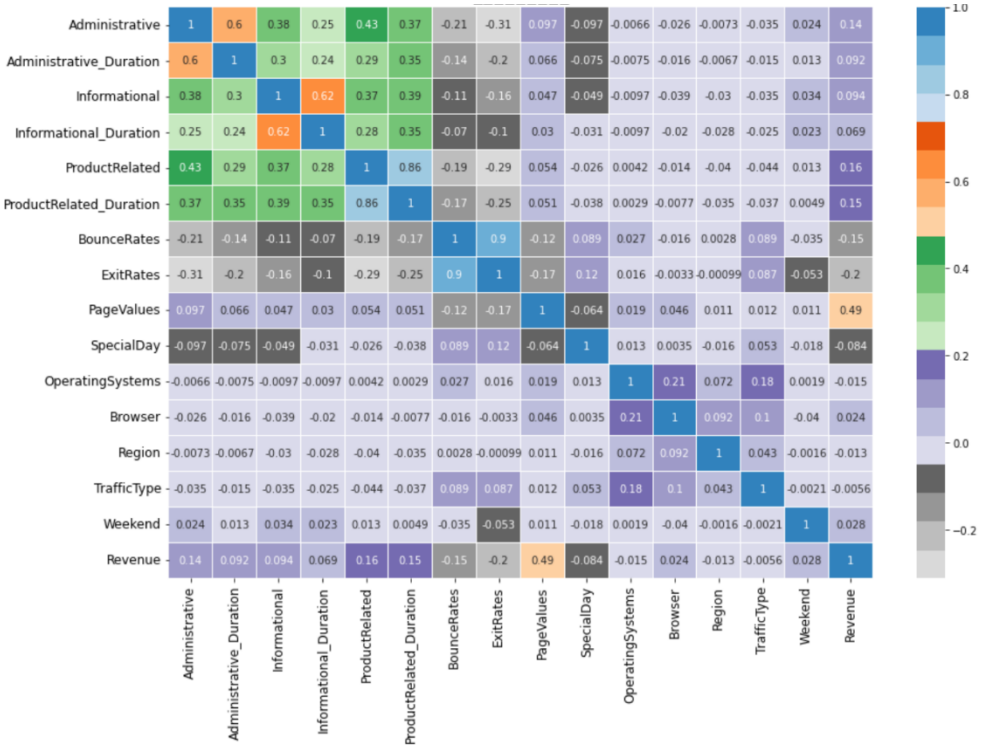


Figure 1. Correlation coefficient heat map (Photo/Picture credit: Original).

As shown in the Fig. 3, the correlation coefficient matrix thermal map can be intuitively seen. The strongest correlation between ExitRates and BounceRates is about 0.9. The correlation coefficient between ProductRelated_Duration and ProductRelated is 0.86. The Administrative_Duration has a strong correlation with Administrative, and the data is 0.62. The correlation coefficient between Administrative_Duration and Administrative is 0.6, ranking fourth. The above characteristics have strong correlation and great influence on each other. ExitRates have a weak negative correlation with Administrative and ProductRelated data of -0.31 and -0.29, respectively. The strongest correlation between Revenue and PageValues was 0.49. Therefore, the value of the page plays a decisive role in the user's online shopping. Therefore, merchants can modify the page to increase the value of the page, reduce exit rates and bounce rates, and make users stay on the page for a longer time, which helps to enhance users' purchase intention and increase merchants' income. Promotional activities are also held on special festivals to improve the negative correlation between special festivals and income.

After training the model with 70% of the data, use the remaining 30% as the test set. Decision tree accuracy table is shown in Table 3. The relevant definitions and formulas of precision and recall are:

- precision = TP/(TP+FP) recall = TP/(TP+FN)
- TP(True Positive): predict the right answer.
- FP(False Positive): incorrectly predicts another class as this class.
- FN(False Negative): this class tag predicts other class tags.

F1-score is a statistical index used to measure the accuracy of a binary classification model. The closer the f value is to 1, the better the model [9]. The first column is false, true, accurary, macro avg and weighted avg. Accurary is 0.86, indicating that the Revenue accuracy measured by the classifier test set is 86%. macro avg and weighted avg are both metrics used to measure classifier performance. macro avg averages each category without considering the number of samples, but ignores the differences in the number of samples. In this case, weighted avg provides the weighted average of each category, and considers the percentage of samples to avoid the impact of sample imbalance on the evaluation result. The table shows macro avg as 0.75 and weighted avg as 0.86. SVM accuracy table is shown in Table 4.

DOI: 10.54254/2755-2721/47/20241338

Table 3. Training the accuracy of the set in the decision tree.

	precision	recall	f1-score	support
false	0.92	0.92	0.92	3071
ture	0.58	0.57	0.57	591
accuracy			0.86	3662
macro avg	0.75	0.75	0.75	3662
weighted avg	0.86	0.86	0.86	3662

Table 4. Training the accuracy of the set in SVM.

	precision	recall	f1-score	support
false	0.84	1.00	0.91	3071
ture	1.00	0.02	0.04	591
accuracy			0.84	3662
macro avg	0.92	0.51	0.48	3662
weighted avg	0.87	0.84	0.77	3662

The first row and first column of the table are the same as in Table 2. The accuracy of the test set was 84%, and the macro avg and weighted avg data from f1-score were 0.48 and 0.77, respectively. Moreover, in the Decision tree model, macro avg-precision equals macro avg-recall, and weighted avg-precision equals weighted-recall. In the Decision tree model, macro avg-precision is different from macro avg-recall, and weighted avg-precision is also inconsistent with weighted avg-recall. Compared with the two models, the accuracy of decision tree is 0.86, which is 0.02 higher than that of SVM. There is little difference between the precision of Decision tree and the true or False of recall, while there is a big difference between SVM and the maximum of False- precision is 1.00 and False-recall is 0.02. As a result, the corresponding f1-score is only 0.04. This is because the decision tree is generated according to the rules of a hierarchical structure, and it can be made more complex by increasing the size of the tree, simulating annular boundaries with more and more partitions, and considering the interactions between variables. SVM makes the classes linearly separable by mapping the feature space to the kernel space and can handle the interaction between nonlinear features. The decision boundary mapped back to the original feature space is a nonlinear decision boundary.

4. Limitations and prospects

To sum up, only two models are used to analyse this problem, each has advantages and disadvantages, and both have certain limitations for predicting user purchase intention. Since the data comes from Kaggle and there are 12,330 pieces of data, the time of data collection has a certain range, and the results and predictions obtained may be different from the current situation, which may lead to the inaccurate forecast data, with the highest being only 86%. In addition, the determination and collection of the feature points have a certain limited range, which will lead to some features not strong correlation, or the features are not typical. Decision trees are continuous data that need to be discretized before they can be used. The data continuity of this dataset is strong, so the accuracy of testing the model is not high. In addition, decision trees have limited application scope and cannot be applied to some decisions that are not represented by numbers, such as weekend data and special day in the data set. Using this method to determine the probability of occurrence of the parties is sometimes subjective, which can lead to decision-making errors. SVM decision boundaries are complex and do not perform well for scaling data. When the amount of data used by the SVM model is greater than 100,000, there can be challenges in terms of runtime and memory usage. There is room to improve the accuracy of the test set, improve the code, and use other methods to predict this problem. The highest correlation of feature points is about 0.6, and the lowest correlation is -0.31. Therefore, in order to predict the income of merchants, that is, the influencing factors of consumers' online shopping intentions, researches can use other data sets to

DOI: 10.54254/2755-2721/47/20241338

find more intuitive and direct feature points. Decision tree software has always been one of the heroic issues of decision tree, and more and more people participate in this method, such as is used to generate and evaluate software resource analysis [9]. The decision tree method can study simpler and more effective simplification methods, such as DI pruning method [10]. Some people have also proposed new methods to improve accuracy, but will greatly reduce its efficiency, that is, the time complexity is greatly increased. SVM can improve its algorithm and combine SVM model with practical problems, so it has great development potential.

5. Conclusion

In this paper, decision tree and SVM are used to analyse the intention of online shopping. The data set last 30% used to a test set. Through the two models, decision tree method is more reliable in predicting online shopping intention, and the accuracy of the data is 86%, which is slightly higher than that of 84% obtained by SVM. Therefore, the benefits of decision tree can be summarized as follows. It is easy to understand and it performs classification without much computation. It clearly shows which areas are most important for forecasting or classification. The downside of the decision-making method is that the decision tree is not suitable for estimating tasks designed to predict continuous properties. They tend to be wrong when it comes to sorting. There's a high probability of reconfiguration in the decision tree. SVM benefits: effective when larger than the sample size. It works well and has a good memory. Works in large sizes. The downside of SVM is that when big data are available, it takes longer to learn. The SVM cannot provide a direct probability calculation. It cannot be performed with data set noise. The decision tree method can be improved in overfitting and pruning. For example, scikit-learn only realizes pre-pruning but not post-pruning. The SVM method can be improved in many ways. Future research can focus on optimizing regularization parameter C and selecting appropriate kernel functions and functions to enhance and perfect the classification algorithm. However, the data set of this study is small and the correlation gap of feature points is large, which may lead to inaccurate prediction results. In general, merchants can consider these two aspects to improve, to attract old users to buy, new users to browse the page, so as to achieve the purpose of increasing revenue and reducing inventory. Through this model, users' online purchase intention and future development trend can be predicted. Merchants can reasonably modify the page introduction and page functions to purchase goods reasonably.

References

[1] Islam M A and Daud K A K 2011 Factors that influence customers' buying intention on shopping online International Journal of marketing studies vol 3(1) p 128.

[2] M S Islam 2023 Prediction of Buying Intention: Factors Affecting Online Shopping International Conference on Next-Generation Computing IoT and Machine Learning (NCIM) Gazipur Bangladesh pp 1-6.

[3] Yap C T and Khor K C 2022 Utilising Sampling Methods to Improve the Prediction on Customers' Buying Intention International Conference on Decision Aid Sciences and Applications (DASA) Chiangrai Thailand pp 352-356.

[4] Rangsaritvorakarn N, Nimsai S, Fakkhong K and Jongsureyapart C 2022 An Experimental Comparison of Classification Algorithms for Premium Beef Customer Buying Intention International Conference on Decision Aid Sciences and Applications (DASA) Chiangrai Thailand pp 1078-1082.

[5] Prasad B and Ghosal I 2022 Forecasting buying intention through artificial neural network: an algorithmic solution on direct-to-consumer brands Business Review vol 11(4) pp 405–421.

[6] Šebalj D, Franjković J and Hodak K 2017 Shopping intention prediction using decision trees Millenium vol 4 pp 13-22.

[7] Quinlan J R 1986 Induction of decision trees. Machine learning Machine learning vol 1 pp 81-106.

DOI: 10.54254/2755-2721/47/20241338

[8] Somvanshi M, Chavan P, Tambade S and Shinde S V 2016 A review of machine learning techniques using decision tree and support vector machine International conference on computing communication control and automation (ICCUBEA) pp 1-7.
[9] Selby R W and Porter A A 1988 Learning from examples: generation and evaluation of decision trees for software resource analysis IEEE Transactions on Software Engineering vol 14(12) pp 1743-1757.
[10] Fournier D and Crémilleux B 2002 A quality index for decision tree pruning Systems vol 15(1-2) pp 37-43.