Prediction of Customer Purchase Satisfaction and Influencing Factors Investigation Based on Machine Learning

Tianyi Ouyang[a]
Department of Math, The Ohio State University, Columbus, U.S.A.

Keywords: Shopping Satisfaction, Machine Learning, Feature Selection, Predictive Model.

Abstract: In the field of e-commerce, shopping satisfaction is a key indicator for measuring consumers' overall perception of their shopping experience and is crucial for merchants to attract and retain customers. This study utilized the Amazon consumer behavior data from Kaggle and applied machine learning techniques to construct an efficient predictive model for accurately forecasting shopping satisfaction. The core of the research method involved the application of three classic machine learning algorithms: K-Nearest Neighbors (KNN), Decision Trees, and Random Forests. The specific research steps included data preprocessing, feature selection, dataset partitioning, and model prediction. In the data preprocessing phase, missing values in the dataset were removed. The feature selection phase employed the ExtraTreesClassifier algorithm for importance analysis, thereby determining the relative importance of each feature for model prediction. After feature selection, the study chose the most important features for the model and used cross-validation to evaluate the performance of the algorithms. Finally, after the model construction, this paper conducted hyperparameter tuning to optimize the model, resulting in the best predictive model, with the Decision Tree and Random Forest models showing excellent performance due to their high accuracy in classification tasks. The research results indicated that rating accuracy and personalized recommendation frequency are the two most important factors affecting shopping satisfaction. These findings provide guidance for online platforms to improve services and recommendation systems, which can help increase customer satisfaction and sales.

1 INTRODUCTION

With the increasingly fierce competition among e-commerce platforms, improving shopping satisfaction has become a key strategy for merchants to attract and retain customers and enhance brand loyalty. Shopping satisfaction, as an important indicator to measure the overall perception of the shopping process and results, is not only related to consumers' immediate purchase decisions, but also has a profound impact on their future purchase behavior and word-of-mouth communication. Therefore, exploring the influencing factors of shopping satisfaction and accurately predicting shopping satisfaction is of immeasurable value for merchants to optimize service processes, precision marketing, and improve customer experience.

Over the years, extensive research has been carried out to explore the influencing factors for Shopping Satisfaction. These studies encompassed several characteristics, including age, personal purchase characteristics, search methods etc. These studies revealed that there is a relationship between these factors and Shopping Satisfaction (Rajesh, 2018; Katta, 2016; Jahwari, 2018). In recent years, the rapid development of machine learning technology has provided a powerful tool for predicting shopping satisfaction. Machine learning models can automatically learn from massive amounts of data and extract complex patterns and patterns to make accurate predictions about unknown data. In the field of e-commerce, machine learning models have been widely used in product recommendation, price prediction, fraud detection and other aspects, and have achieved remarkable results. Using machine learning models, it is possible to build efficient and accurate prediction models by in-depth analysis of multi-dimensional data such as age, consumer behavior, and purchased items, so as to predict Shopping Satisfaction. Previous studies have

[a] https://orcid.org/0009-0003-6923-7663

Ouyang, T.
Prediction of Customer Purchase Satisfaction and Influencing Factors Investigation Based on Machine Learning.
DOI: 10.5220/0013207100004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 92-96
ISBN: 978-989-758-726-9

extensively explored the influencing factors and predictions of Shopping Satisfaction. For instance, a study conducted by Ludin collected data on the factors influencing customer satisfaction among consumers in the Klang Valley and analysed the data through descriptive analysis and regression analysis (Ludin et al., 2014). Another related work by Lin used Structural Equation Modeling (SEM) as the main analysis tool to analyze the shopping experience of users of major shopping websites in Taiwan, and the results showed that website service quality can directly have a positive impact on customer satisfaction (Lin et al., 2009). Another related work of Gim used methods such as association analysis and multiple regression analysis to analyze the influencing factors of shopping satisfaction of Viet Nam consumers (Gim, 2014).

However, a few studies have been conducted on Shopping Satisfaction in the previous literature. In this case, the purpose of this paper is to further build an accurate model through a classification algorithm to make a more accurate prediction of Shopping Satisfaction. This research will contribute to existing research by using EDA and machine learning algorithms to predict customer satisfaction with shopping. By analyzing a comprehensive data set, this study will identify the factors that will have the greatest impact on purchase satisfaction and build a predictive model to predict their purchase satisfaction. The conclusions of the study are very meaningful for online platforms, according to which they can improve the mechanism and recommendation methods of their platforms, and adopt different strategies to increase customer purchase satisfaction, thereby increasing sales and increasing revenue.

2 DATASET

This section will include data preprocessing, feature engineering, feature selection, and data partitioning. These steps will better assist this study in exploring the influencing factors of shopping satisfaction.

2.1 Dataset Preparation

The dataset used in this article is the Amazon consumer behaviour dataset, which comes from the Kaggle website and is sourced from (Kaggle, 2023). This is a publicly available dataset created by Swathi Menon that contains 22 characteristics of amazon consumer behavior. This paper analyzes 22 features in the dataset to explore the influencing factors of shopping satisfaction and establishes a model to predict shopping satisfaction.

Analysis of this dataset can help companies identify the most influential factors of purchase satisfaction and explore the interrelationships between these factors to form the causes. These findings can provide valuable insights for companies to help them develop effective sales strategies to increase consumer satisfaction with their purchases, thereby increasing sales.

After carrying out the analysis for missing values count, it can be found there are fewer missing values in the dataset, with only two missing values in Product_Search_Method. So, this study just drops out the missing values.

2.2 Feature Engineering

Feature engineering involves enhancing the performance of predictive models by transforming the feature space of a dataset (Nargesian, et al., 2017). In order to make effective use of variables in the analysis, this article performs label encoding using a library in Python using “sklearn.preprocessing”. Label encoding is a critical pre-processing step for categorical data, as it allows to represent categories numerically, which is necessary for many machine learning algorithms.

Label encoding requires a total of two steps. First, this paper imported the required libraries. After that, eighteen categorical variables such as "age", "gender", etc were chose for analysis. Eventually, all the categorical variables were expressed in numerical form in this study.

2.3 Importance and Feature Selection

In order to examine the variables that are related to purchase satisfaction, a correlation analysis was performed in this paper.

Importance analysis can explore the relationships between variables. Search accuracy is highly importance to purchase satisfaction, which means that if a customer searches for exactly what he wants to buy, the probability of his satisfaction will be very high. The importance of shopping type, customer evaluation, and personalized recommendation and purchase satisfaction is relatively high. However, there are some variables with very low importance with purchase satisfaction, suggesting that there is no significant correlation between them.

Therefore, in this case, the feature selection was carried out. Feature selection is used to clean up the noisy, redundant and irrelevant data (Venkatesh et al., 2019). This study picked the top 10 most important

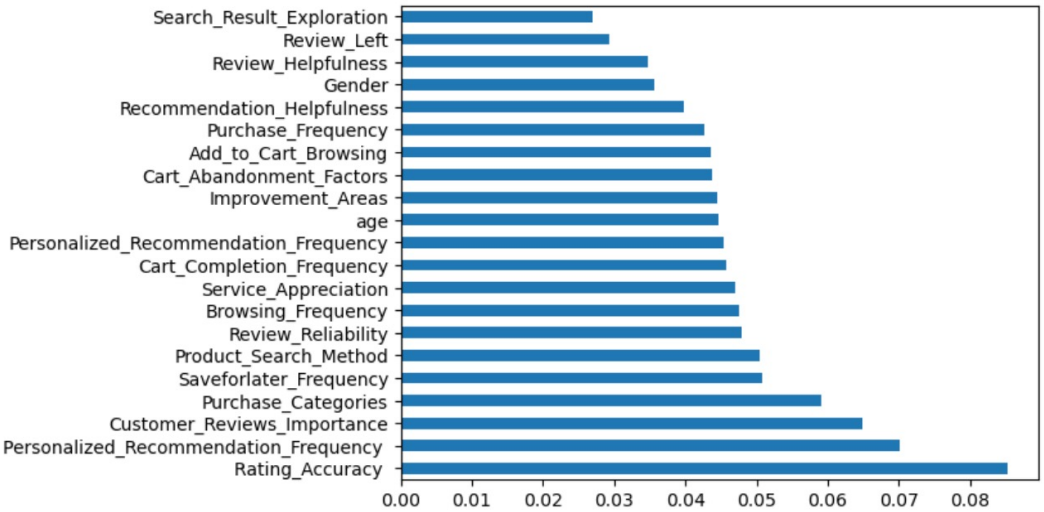

Figure 1: Importance with Shopping satisfaction (ExtraTreesclassifier) (Photo/Picture credit: Original).

variables and prepared them for the modeling analysis.

2.4 Splitting Data

The dataset X and purchase satisfaction y are divided into a training set and a test set according to the ratio of 9:1, in which 10% of the data is used as a test and the remaining 90% is used as a training. By setting random_state parameters, it can be ensured that the result of the split is the same every time the code is executed.

3 MACHINE LEARNING MODELS

To predict purchase satisfaction, three classical algorithms were chosen: K-nearest neighbors, random forests, and decision trees. In this study, these three algorithms will be trained, hyper tuning, and ultimately their effectiveness will be judged by evaluating their accuracy in classification tasks.

The k-Nearest-Neighbours (KNN) is a simple but effective method for classification (Guo et al., 2003). KNN is a classification algorithm that looks for the K training samples closest to the test samples and predicts the class of the test samples based on those samples. KNN's simple implementation and no assumptions about the distribution of data make it excellent when dealing with nonlinear problems.

The decision tree methodology is a widely utilized approach in data mining for creating classification systems or prediction algorithms based on multiple covariates (Song, 2015). This method constructs a tree-like structure where each internal node denotes a feature test, each branch signifies the outcome of the test, and each leaf node delivers the final prediction.

Random forest, an ensemble learning technique, enhances the decision tree approach by constructing multiple decision trees and merging their outputs to boost model accuracy and stability (Biau, 2016). This method, which involves generating several randomized decision trees and averaging their predictions, has demonstrated superior performance in scenarios where the number of variables significantly exceeds the number of observations. Additionally, random forests can evaluate feature significance, aiding in feature selection.

This study first performed data preprocessing, including data observation and cleaning. After that, this paper did feature selection and data standardization, and split the data into a training set and a test set. Subsequently, this paper used cross-validation to evaluate the performance of each algorithm and adjusted their parameters to obtain optimal results. By comparing the accuracy, recall, F1 score, and other relevant metrics of the three algorithms, this study can determine which algorithm is best suited to solve the research question.

This paper trained the three models in three steps. First, this paper created a classifier, all of which are default. After that, this study trained three classifiers with data from the training set using the FIT method. In the training phase, the KNN algorithm does not actually learn the model parameters, but simply stores the training data, while the other two algorithms learn the model parameters. In the third step, this paper predicted the X_test and got the prediction result y_pred.

After that, hyper tuning was carried out. This paper inputs the value range of various parameters and use GridSearchCV in sklearn to obtain the optimal parameters. Then this paper re-inputs the data of the training set into the model, adjusts the parameters to the optimal parameters, and trains the classifier again by the FIT method. After that, this study used the newly adjusted model to predict the y_pred again. Accuracy is a measure of the performance of a classification model, especially in binary or multiclassification problems. It represents the number of samples correctly predicted by the model as a proportion of the total number of samples. The closer the Accuracy is to 1, the better the model.

4 RESULTS AND DISCUSSION

The research first explores the shopping sanctification's influential factors. By exploring the importance of 21charqcters to shopping satisfaction , the relationship between shopping satisfaction and each factor were clearly presented according to the importance of each feature, the features have been listed from high to low. The following Figure 1 represents the importance of each variable to shopping satisfaction.

From the Figure 1, it is evident that the importance between rating accuracy and shopping satisfaction is the greatest, reaching 0.08, far exceeding other features. This may be because when the rating accuracy is high, it implies that the platform can provide more accurate ratings. By providing accurate ratings for products, customers can more precisely match their needs to purchase the required products. After purchasing suitable products, customers' shopping satisfaction will naturally increase. The second most important factor is personalized recommendation frequency. Its importance to shopping satisfaction reaches 0.07. When the frequency of personalized recommendations increases, it means that customers will receive more frequent personalized recommendations tailored to them. Through these personalized recommendations, customers can purchase products suitable for themselves to meet their needs. Once the customers' needs are met, their shopping satisfaction will naturally rise.

At the bottom of the list is search result exploration. This means that looking through more purchase pages does not effectively enhance shopping satisfaction. This may be because the products on the first page are always the most relevant to the customer's search terms, so continuing to flip through subsequent pages is unlikely to find more suitable products. Therefore, flipping through more pages does not directly affect whether customers can find suitable products, so it is unrelated to shopping satisfaction.

Subsequently, this study conducted modeling analysis, hoping to predict shopping satisfaction through modeling. This paper involves a total of three models, namely KNN, decision tree, and random forest. This paper adjusted the hyperparameters to find the optimal parameters for the three models and obtained the results based on the optimal parameter models.

Table 1: The performance of different models.

model	accuracy
KNN	0.88
decision tree	1
random forest	1

The accuracy of the KNN model is 0.88 shown in Table 1, which means that the KNN model can predict shopping satisfaction with a relatively high degree of precision. The accuracy of both the decision tree model and the random forest model is 1, indicating that these two models can predict shopping satisfaction with 100% accuracy in the test set.

5 CONCLUSIONS

Through the research, this paper concludes which features are most important for shopping satisfaction. Among them, rating accuracy is the most important factor; the higher the rating accuracy, the higher the shopping satisfaction of the customers. Personalized recommendation frequency follows closely behind. Under different standards of importance, different features will be selected. This study provides meaningful insights into shopping satisfaction and offers directions for improving customer satisfaction on the platform. In the model prediction part, the decision tree and random forest models are clearly superior to the KNN model. The random forest model is a model that combines multiple decision tree models, and a single decision tree model can accurately predict shopping satisfaction, which implies that the decision tree model is the best predictive model. In the future, more challenging scenarios will be investigated based on more advanced machine learning models.

REFERENCES

Al-Jahwari, N. S., Khan, F. R., Al Kalbani, G. K., & Al Khansouri, S. 2018. Factors influencing customer

satisfaction of online shopping in Oman: Youth perspective. Humanities & Social Science Reviews, eISSN, 2395-7654.

Biau, G., & Scornet, E. 2016. A random forest guided tou. Test, 25, 197-227.

Gim, G. 2014. Evaluating factors influencing consumer satisfaction towards online shopping in Viet Nam. Journal of Emerging Trends in Computing and Information Sciences, 5(1), 67-71.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. 2003. KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg

Kaggle, 2023, Amazon Consumer Behaviour Dataset, https://www.kaggle.com/datasets/swathiunnikrishnan/amazon-consumer-behaviour-dataset

Katta, R. M. R., & Patro, C. S. 2016. Online Shopping Behavior: A Study of Factors Influencing Consumer Satisfaction on Online viz-a-viz Conventional Store Shopping. International Journal of Sociotechnology and Knowledge Development (IJSKD), 8(4), 21-36.

Lin, G. T., & Sun, C. C. 2009. Factors influencing satisfaction and loyalty in online shopping: an integrated model. Online information review, 33(3), 458-475.

Ludin, I. H. B. H., & Cheng, B. L. 2014. Factors influencing customer satisfaction and e-loyalty: Online shopping environment among the young adults. Management Dynamics in the Knowledge Economy, 2(3), 462-462.

Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. S. 2017. Learning Feature Engineering for Classification. In Ijcai (Vol. 17, pp. 2529-2535).

Rajesh, R. 2018. Evaluating the factors influencing online shopping and its consumer satisfaction in Pune Area. PEOPLE: International Journal of Social Sciences, 4(1), 54-76.

Song, Y. Y., & Ying, L. U. 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

Venkatesh, B., & Anuradha, J. 2019. A review of feature selection and its methods. Cybernetics and information technologies, 19(1), 3-26.