Using the XGBoost Model Experiment to Predict the Car Brand Purchase Desire of the Crowd

Yuhan Wang

China Automotive Technology and Research Center Co., Ltd., China

wangyuhan@catarc.ac.cn

Abstract

In recent years, with the rapid economic development, the rapid improvement of material living standards, and the continuous popularization and application of big data technology. Big data technology is applied to online e-commerce and physical stores, allowing store owners to understand who potential customers are, what customers need, and what products customers need to advertise to potential customers accurately. The XGBoost model can well assist us in predicting the user's purchase desire. The idea of the XGBoost algorithm is to add trees continuously. Each split will add a new tree. Each new tree uses a new function to fit the residual of the previous prediction data. Until the entire algorithm finishes training, it will get many K's trees. The method of predicting the score of a sample is as follows: according to the characteristics of the model, each tree will correspond to a leaf node, each leaf node will get a score, and finally, adding each corresponding score, we can calculate the predicted value of the sample.

Keywords

Desire to Purchase; Prediction; XGBoost Model.

1. Research Background

In recent years, with the rapid economic development, the rapid improvement of material living standards, and the continuous popularization and application of big data technology. Big data technology is applied to online e-commerce and physical stores, allowing store owners to understand who potential customers are, what customers need, and what products customers need to advertise to potential customers accurately. Therefore, it has become an important research topic in industry and academia to deeply mine and analyze users' purchase behavior through technical means to provide reliable technical support for user choice and business operation. The mutual win is of great significance. In this context, targeted technologies such as collaborative filtering and content recommendation emerge as the times require. The principle of collaborative filtering is to recommend items that may be of interest to the user based on the rating data of the nearest neighbors with similar ratings. However, there are still some scenarios where the user rating data cannot be obtained, so the recommendation cannot be completed. The general idea of content-based recommendation is to mine the user's characteristics to match the product's features and recommend the item with the highest similarity between the two to the user. These recommendation technologies mainly complete the recommendation based on the actual purchase behavior results of the user while ignoring the operation behavior generated by the user during the purchase process. Therefore, the targeted recommendation technology can determine which type of product the user is likely to buy but cannot accurately predict when the user will finally buy the product. In recent years, many user operation behavior data have been generated in-service platforms of all walks of life. By digging deeply into the user operation behavior data, users' shopping habits and preferences can be found. The operation behavior generated during the user's purchase process will generate relevant data. Using this data for a recommendation has become a feasible method, which many scholars have adopted and studied.

This paper proposes a user purchase behavior prediction model based on deep forest. Section 2 introduces the principles and algorithms of traditional recommendation algorithms; Section 3 introduces problem scenarios, data, and results; Section 4 summarizes the full text.

2. Introduction to Related Models

2.1. Multiple Linear Regression Model

Multiple linear regression models are often used with two or more influencing factors as independent variables to explain changes independent variables. When there is a linear relationship between multiple independent and dependent variables, the regression analysis performed is multiple linear regression. The formula for calculating the multiple linear regression model is as follows:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \xi$$

Among them $\hat{y}$ is the prediction result, β_0 is the constant term, and x_i β_i (i=1, 2, 3, ..., n) represent the eigenvalues and regression coefficients, respectively, and ξ represent the error term.

2.2. Random Forest Model

The random forest model is an important learning method established based on the classic bagging algorithm, which is used to solve problems such as classification and regression. The random forest model has many advantages, such as extremely high accuracy, not being easy to overfit, can handle high-dimensional data, and quickly realizing program parallelization. Figure 1 describes the whole process of data calculation by the random forest model: the first step is to randomly perform sampling with replacement in the original data set to form n different sample data sets; the second step, for each sample data set, construct a decision tree and build n other decision tree models; finally, obtain the final result according to the average value of the decision tree models.

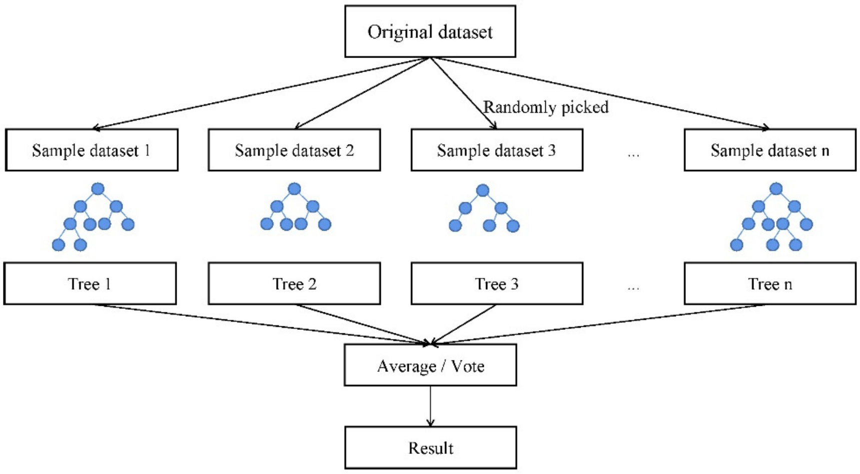

Fig 1. Random forest prediction process

2.3. XGBoost Model

XGBoost is a Boosting algorithm. As shown in Figure 2, the idea of Boosting algorithm is to integrate many weak classifiers to form a robust classifier. Because XGBoost is a boosting tree model, it is a strong classifier developed by integrating many tree models, and the tree model used is the CART regression tree model.

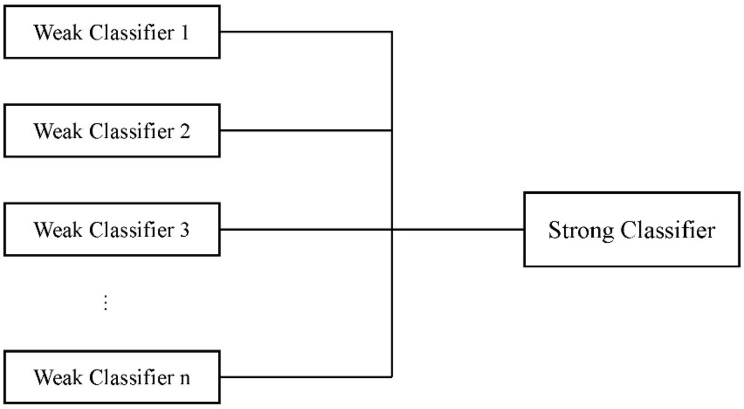

Fig 2. Multiple weak classifiers form a robust classifier

The idea of the XGBoost algorithm is to add trees continuously. Each split will add a new tree, and each new tree uses a new function to fit the residuals of the last prediction. After the algorithm is trained, several K trees will be obtained. Predicting the score of a sample is based on the characteristics of the model. Each tree corresponds to a leaf node, and each node corresponds to a score. Finally, the sum of each corresponding score is the predicted value of the sample.

The objective function of XGBoost consists of two parts: the training loss function and the regularization term. The objective function is defined as follows:

$$Obj=\sum_{i=1}^{N} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k), \tag{1}$$

Among them $\hat{y}_i$ is the predicted value of the ith sample, y_i is the ith real value, and $l(y_i,\hat{y}_i)$ is the training loss function. There are two commonly used loss functions:

(1) Square loss function: $l(y_i,\hat{y}_i)=(y_i-\hat{y}_i)^2$

(2) Logistic regression loss function: $l(y_i,\hat{y}_i)=y_i\ln(1+e^{-\hat{y}_i})+(1-y_i)\ln(1+e^{\hat{y}_i})$

Because XGBoost is an additive model, the prediction score is the sum of the prediction result data of each tree, and the value is $\hat{y}_i=\sum_{k=1}^{K} f_k(x_i)$, where $\sum_{k=1}^{K} f_k(x_i)$ is the tree of the i-th sample and f_k is the function of the k-th tree. $\Omega(f_k)$ This formula expresses the complexity of the tree. Summing all complexity data from 1 to numbered K-tree and then adding it to the objective function as a regularization term can be used to prevent the model from overfitting.

2.3.1. XGboost Classification Algorithm

XGBoost is a gradient boosting decision tree that combines multiple weak classifiers into a strong classifier to minimize the target loss function. The objective function O of XGboost comprises a loss function and a regularization term. Its definition expression is as follows:

$$\text{O} = \text{L} + \Omega\ (\text{ft}) \tag{2}$$

L in this formula refers to the loss function. The loss or error between the predicted value and the actual value in the model is judged by the weight of L. Ω(ft) as the regularization term; ft is each classification tree; t is the number of trees. It is Used to control the complexity of the model and avoid overfitting.

2.3.2. Loss Function

The XGBoost classification algorithm is equivalent to an additive model consisting of K trees. Based on keeping the original model unchanged, the error generated by the previous prediction is used as a reference to build the next tree, which is the difference between the predicted value and the actual value. The residuals are used as input to the next tree. The addition process is as follows:

Initialize $\hat{y}_i^{(0)} = 0$

Add the first tree to the model $\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$

Add a second tree to the model $\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$

...

The above formula $\hat{y}_i^{(t)}$ represents the predicted value of the i-th sample at the t-the time. It retains the model prediction result of t - 1times and then adds a new function $f_n(x_i)$. The unique part added in each round can minimize the loss function to the greatest extent. At this time, the loss function is:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)) \tag{3}$$

Expand the loss function by the second-order Taylor formula, $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}), h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ omitting the constant term $l(y_i, \hat{y}_i^{(t-1)})$, then the loss function becomes:

$$L \approx \sum_{i=1}^{n} [g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)] \tag{4}$$

2.3.3. Regularization Term

The regularization term can effectively control the overfitting of the model, which is related to the number of leaf nodes T and the weight of leaf nodes ω, which are defined as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \tag{5}$$

In summary, the objective function formula is as follows:

$$O \approx \sum_{i=1}^{n} [g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \lambda T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) \omega_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) \omega_j^2] + \gamma T \tag{6}$$

In the formula, γ λ the leaf node coefficient and regular coefficient I_j are the samples on the jth leaf node; ω_j they are the weight of the jth leaf node. Let $G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$, and take the partial derivative ω_j to get the optimal weight:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{7}$$

The optimal objective function is as follows:

$$O = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j}{H_j + \lambda} + \gamma T \tag{8}$$

3. Results and Discussion

3.1. Prediction Process

The software used in this experiment is python3.8, and the libraries and packages used are pandas, NumPy, XGBoost, and Sklearn. The specific process is as follows: Input the sample data set in python, preprocess the data in the sample, Preprocessing includes missing value handling and conversion of feature parameter data types, generate a 5:1 training set, and test from the preprocessed data set. Set the commonly used training parameters, such as Max_Depth, etc.; use grid search and cross-validation for tuning. The grid search algorithm is an exhaustive search algorithm for the specified parameters, arranging and combining each parameter's possible choices and values, listing all possible decisions and combination results, applying the combination of each item to For XGBoost training cross-evaluate its table to obtain the optimal parameters.

3.2. Experimental Results

3.2.1. Data Sources

The data used in this article is the satisfaction survey data of a car brand in 2021. The survey aims to understand the brand's brand funnel health. The survey objects are buyers of different brands. The information of 3030 consumers was collected as a data sample. Before modeling the dataset, we preprocessed and cleaned the dataset because there will always be some missing values and outliers in the original dataset and even some features that we do not need. We preprocess and clean the dataset to ensure its accuracy of the data. Table 1 shows the characteristic parameters extracted from this dataset. The data in this table are all typical personal data of the user. The user's decision to purchase this brand depends on individual factors, such as living area, family monthly income, gender, age, etc.

Table 1. Characteristic parameters

Title	Options
Q1. What is your city level?	1. First-tier cities 2. New first-tier cities 3. Second-tier cities 4. Third-tier cities 5. fourth-tier cities 6. Fifth-tier cities

Q2. where do you live?	1.Beijing 2.aiyuan 3.Jilin 4.Baotou 5. Shanghai 6.Suzhou 7. Ningbo 8.Linyi 9. Guangzhou 10.Fuzhou 11.Haikou 12.Yangjiang 13. Wuhan 14.Zhengzhou 15.Changsha 16.Ganzhou 17. Chengdu 18.Xi'an 19.Kunming 20.Urumqi
Q3. What's your birth year?	1.post-80s 2.post-85s 3.post-90s 4.post-95s 5.post-00s
Q4. Your marital status?	1. Single 2. Married without children 3. Married with children 4. Refused to answer
Q5. What is your highest education?	1. Uneducated 2. Primary school 3. Junior high school 4. High school 5. Technical secondary school/technical school/vocational school 6. College 7. Undergraduate 8. Graduate or above 9. Others
Q6. What's your working period?	1. Less than one year 2. 1-2 years 3. 2-5 years 4. 5 years 5. Never worked
Q7. What is your job?	1. Agriculture, forestry, animal husbandry, fishery, and industry 2. Energy extraction and supply 3. Industrial manufacturing 4. Construction industry 5. Transportation, warehousing, postal service 6. Information transmission, computer service, the software industry 7. The wholesale and retail industry 8. Accommodation and catering industry 9. Financial industry 10. Real estate development, property 11. Leasing and intermediary industry 12. Consulting, law, auditing, etc. 13. Scientific research and exploration 14. Public administration and community administration 15. Education, health, social security, and social welfare industry 16. Sports and cultural entertainment industry 17. International affairs

	18. Freelancers 19. Pharmaceutical industry 20. Others
Q8. Which of the following is your job title?	20. Government agencies and institutions 21. Foreign enterprises - business executives 22. Foreign enterprises - middle and senior managers 23. Foreign enterprises - general staff 24. State-owned enterprises - senior management personnel 25. State-owned enterprises - middle-level personnel 26. State-owned enterprises - general employees 27. Private - business owners 28. Personal, private - middle, and senior managers 29. Personal, remote - available employees 30. Individual bosses 12. Freelancers
Q9. What is your combined monthly household income (before taxes)?	1. Below RMB 5,000 2. RMB 5,000-8,000 2. RMB 8,001-12,000 4. RMB 12,001-15,000 5.RMB 15,001-20,000 6. RMB 20,001-25,000 7.RMB 25,001-30,000 8. RMB 30,001-40,000 RMB 9.RMB 35,001-40,000 RMB 10. RMB 40,001-45,000 RMB 11. RMB 45,001-50,000 RMB 12. RMB 50,000 and above
Q10.Please enter your gender	1. Male 2. Female

3.2.2. Model Prediction Effect

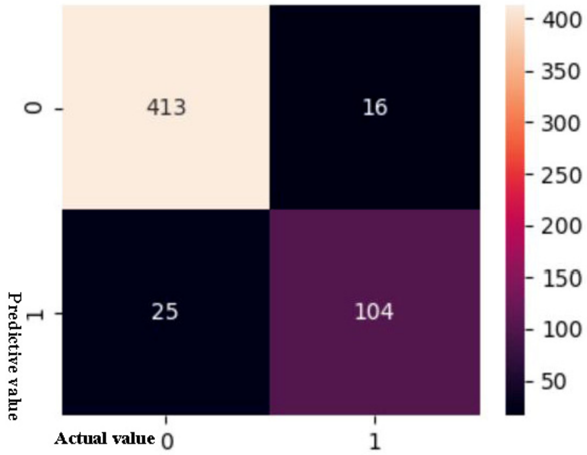

Fig 3. Model prediction results (in the form of a heatmap)

Figure 3 shows the statistics of the predicted values of the XGBoost model. As can be seen from Figure 3, there are 517 samples with correct predictions, in which the expected value and the actual value are both 1, the accuracy rate of predicting the user to purchase the brand is 86.7%, and the expected value and the absolute value are both 0. The user is expected to buy the brand. The accuracy rate of other commodities is 94.3%, indicating that the model has higher prediction accuracy and more reliable prediction results.

Two representative machine learning algorithms, logistic regression and random forest were selected as the comparison model to evaluate further the accuracy of the prediction model of user purchase desire. The accuracy rate is an evaluation indicator of the performance of the model. Table 2 summarizes the comparative data of the three models. Table 2 shows that the accuracy rate of non-purchase prediction and purchase prediction accuracy of the XGBoost model is 94.3% and 86.7%, respectively. Both sets of data of the XGBoost model are better than other algorithms, indicating that the XGBoost model has the best effect.

Table 2. Prediction results of each algorithm model

Model	Accuracy of non-purchase intention predictions	Accuracy of Purchase Intent Prediction
XGBoost	94.3%	86.7%
Logistic regression	77.2%	63.4%
Random forest	88.7%	79.5%

4. Conclusion

This article establishes a purchase prediction model based on the XGBoost classification algorithm, data preprocessing, feature selection, and other works to improve the model calculation's prediction rate. The prediction results show that the accuracy rates are 0.943 and 0.867, respectively. The data results are better than the logistic regression and random forest model algorithms, showing high prediction accuracy. In the follow-up work, we will continue to add some subjective questions, such as satisfaction with the appearance, interior, comfort, and other indicators of a particular car, to make the prediction results more realistic.

References

[1] Fotheringham A S, Crespo R, Yao J. Exploring modeling and predicting spatiotemporal variations in house prices. Anr-nails of Regional Science,vol.02(2015),No.54,p.417-436.

[2] Dansko D, Hoffman M M. Classification and interaction in random forests.Proceedings of the National Academy of Sciences of the United States of America, vol.08(2018), No.115,p.1690-1692.

[3] PRAKS P, KOPUSTINSKAS V, MASERATI M. Probabilistic modeling of security of supply in gas networks and evaluation of new infrastructure. Reliability Engineering and System Safety, vol.07 (2015), No.144,p.254-264.

[4] LOBANOVA G, Fath BD, ROVENSKAYA E. Exploring simple structural configurations for optimal network mutualism.Communications in Nonlinear Science and Numerical Simulation, vol.04(2009), No.14, p.1461-1485.

[5] CUMMING GS, BODIN O, ERNSTONH, et al. Network analysis in conservation biogeography: challenges and opportunities. Diversity and Distributions,vol.03(2010),No.16,p.414-425.

[6] GRAYK R, ALJABAR P, HECKEMANN R A, et al. Randomforest-based similarity measures for multi-modal classification of Alzheimer's disease.NeuroImage,vol.09(2013),No.65,p.167-175.

[7] STRUBF, MARY J. Collaborative filtering with stacked denoising-autoencoders and sparse inputs. NIPS workshop on machine learning for eCommerce.vol.08(2015)No.22,p.77-78.

[8] Chen Shipeng, Jin Shengping. Prediction of house price based on random forest model. Technological innovation and application, vol.04 (2016)No.33,p.52-53.

[9] Zhang Jingmiao. A comparative study on the spatial differentiation of urban housing prices and its influencing factors based on the GWR model--Taking Kunming and Chengdu as examples. Kunming: Kunming University of Science and Technology, vol.07(2017)No.16,p.89-90.

[10] Xu Gang, An Qiqi, Yang Jie, et al. PMV-PPD human thermal comfort assessment model considering individual differences and its application.Journal of Xi'an University of Science and Technology, vol. 01 (2021)No.41,p.55-61.

[11] Gong Hongliang. An Empirical Study on the Price Prediction Model of Second-hand Housing in Wuhan Based on XGBoost Algorithm.Wuhan: Central China Normal University, vol. 05 (2018) No. 66, p.96-99.