Prediction of Consumer Behavior Based on Machine Learning Algorithm

Yan Jiang*[1]

*Corresponding author: 2966348454@qq.com

[1] School of Business Administration, Zhejiang Gongshang University, Hangzhou, China

Abstract: Under the background of digital economy, the Internet industry pays more and more attention to the importance of user consumption behavior, but how to distinguish high-quality user channels and optimize marketing costs has always been the pain point of companies. In this paper, we propose an analysis and prediction method of consumer behavior based on machine learning algorithm. Therefore, based on the statistical analysis of platform user data indicators, this paper establishes an improved random forest model based on unbalanced samples. This paper proposes a prediction method of consumer behavior based on machine learning algorithm, through the value analysis of user consumption behavior, hoping to provide reasonable suggestions for enterprises in precision marketing based on user consumption behavior analysis in the context of the era of big data. First of all, this paper preprocesses the data, matches the data fields, and integrates the missing values and outliers of the data. Secondly, this paper codes the city fields according to the frequency to reduce the problems caused by parallelism and multicollinearity. Finally, machine learning modeling is carried out to obtain the optimal model, with an accuracy rate of 99% and an F1 score of 0.92, and the prediction results are output. The results show that the prediction results are basically consistent with the actual data, the model has high reliability, and the value of consumer behavior can be analyzed based on machine learning method.

Keywords: consumer behavior; precision marketing; random forest; Machine learning; unbalanced sample; principal component analysis

1. INTRODUCTION

Today, with the rapid development of the Internet, how to obtain high-quality data information from massive data is particularly important. Companies in various fields expand online channels to introduce fresh and active users to their products. Traditional marketing strategies have been eclipsed, and precision marketing based on big data has emerged as the times require[2].

At present, experts and scholars at home and abroad have carried out relevant research and analysis in data mining and user consumption behavior research, which is an important research hotspot in the field of AI. Data can be divided into structured data and unstructured data. Data mining refers to finding potential valuable information from a large amount of data and providing decision support for problem solving. This research method has been widely used in all walks of life, such as prediction and evaluation, financial engineering, risk management and so on. Data mining algorithms can be generally divided into two categories: supervised learning and unsupervised learning. [1][3]

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334304

The generation of consumer behavior generally needs five stages: demand motivation, information acquisition, evaluation, decision-making and feedback. If enterprises want to respond to the market quickly in the competition, they must grasp the behavioral characteristics of consumers. Behind the consumption behavior, it is affected by various factors. Consumers make different purchasing decisions because of their age, gender, occupation, income, geographical location and so on. It is very important to be familiar with the characteristics of consumers' purchase, which is helpful for enterprises to formulate marketing strategies according to the digital characteristics of users' consumption behavior in product promotion, dynamically adjust to achieve greater market share, establish a good brand image, and create greater commercial value.[1][3]

This paper proposes a method of consumer behavior trends based on machine learning, which can accurately locate users and form a stable prediction process in a complex prediction environment. The main content of this paper is: in the complex prediction environment, based on the machine learning algorithm to predict the user's consumption behavior, predict whether the target user will have a purchase behavior, and then in the context of big data, precise marketing, precise push of goods for users, which can accurately locate the users who are easy to lose, and form a stable prediction process.

2. PREPARATION OF THE MODEL

2.1 Meanings of important terms and indicators

(1) Basic information of the user. The user's main information includes age, city, equipment model, etc.; the course selection has different pertinence for different age groups, and the audience groups are inconsistent. Geographical differences are also a factor leading to the selection and purchase of courses.

(2) Number of days logged in, etc. After the user logs in the APP, the number of days and time interval of login largely reflect the user's activity; whether the user pays attention to the official account and adds sales friends, and the number of lessons learned is an important indicator of the subsequent purchase of the course.

(3) Access information. Users may browse a lot of web courses, and the time spent on each course and the number of clicks are very important to the analysis of user consumption behavior.

(4) User potential indicators. Users stay in the course for a long time, often click on advertisements and visit relevant courses, which shows that the user has potential value and is an important customer of precision marketing.

3. MODEL ESTABLISHMENT AND SOLUTION

3.1 Initial data information preprocessing

According to the given four tables of user information table (user _ info. CSV), user login situation table (login _ day. CSV), user access statistics table (visit _ info. CSV) and user order table (result. CSV), in order to complete the follow-up task and analyze the consumer behavior,

The four tables are explored and analyzed, and it is found that the dimensions of the four tables are inconsistent, and there are a large number of missing values and some errors, so the missing values are filled and deleted, and the outliers are deleted. Finally, we perform dimensional unification.

(1) Field cleaning. By analyzing the four tables, we will judge the relationship between the user _ ID field and each table. Through analysis and comparison, we find that there is only a user's personal information table in the user _ ID (2000001566047613) of the reference sample, but not in the user's login table and the user's access statistics table. So we can think that this user will not place an order; moreover, there are several user _ID (2000001563151338, 2000001563163750, 20000015632661) in the given user order table19, 2000001566046975, 2000001566153564, 2000001568947583, 2000001569852746) They are stored in the table that needs to be forecasted, so we can assume that they must be ordered, which is the implicit certainty of the problem.

After the above analysis, we removed the above user _ ID, linked the table, and deleted and filled in some missing values, time stamp processing and abnormal value processing. When dealing with the type of city, we first considered the unique coding to deal with the dummy variable, but because the number of cities is up to 361, if it is treated as a dummy variable, it will cause a dimension explosion, so we chose to replace the type of each city with frequency, and finally assigned the null city to the city with more users (Chongqing).There are a lot of null values, and we can see from the title that this part indicates that the user has not placed an order. Examples of missing values are shown in Table 1.

Table 1. Missing values

Field	Frequency
First _ order_ time	458
First _ order_ price	458
Age _ month	458
City _ num	28435
Platform _ num	458
Model _ num	458
result	130785

From the above table, we can see that most of them are missing 458, which is caused by the extra user information, so here we directly choose to delete. The rest is filled and processed.

(2) Data standardization. In the multi-index evaluation system, each evaluation index usually has different dimensions and orders of magnitude due to its different nature. When the level of each index is very different, if the original index value is directly used for analysis, the role of the index with higher value in the comprehensive analysis will be highlighted, and the role of the index with lower value will be relatively weakened. Therefore, in order to ensure the reliability of the results, it is necessary to standardize the original index data. We adopt the Z-Score normalization here.

3.2 Visual analysis

This step is to analyze the distribution and login of users in each city, and visualize the results. Therefore, we only need to use the city field and the login information related field in the user information table (user _ info. CSV) and the user login information table (login _ day. CSV).

(1) Combine the information in the user information table (user _ info. CSV) and the user login information table (login _ day. CSV). First, let's draw a map of the distribution of user cities. In the process of drawing the map, the following missing value and outlier errors occur. In order to ensure the authenticity of the data, we delete them. Examples of missing values and outliers in the city field are shown in Table 2 below.

Table 2. Missing and outlier statistics for the city field

City	Frequency
error	412
Pu'er	28
Kizilsu Kirgiz Autonomous Prefecture	23
Beitun	11
Null value	28209

(2) According to the word frequency statistics of the city field, the user city distribution map is drawn through pyecharts. as shown in fig. 1.It can be seen from Figure 1 that the cities with less than 50 users are evenly distributed, mainly in minority areas and areas with less developed communication foundation; the distribution of cities with 51 to 100 users begins to shift to the central and eastern regions; the distribution of cities with 101 to 300 users begins to be dense in the central and eastern regions; The cities with more than 301 users are more concentrated in the central and eastern regions, and the distribution is more dispersed.

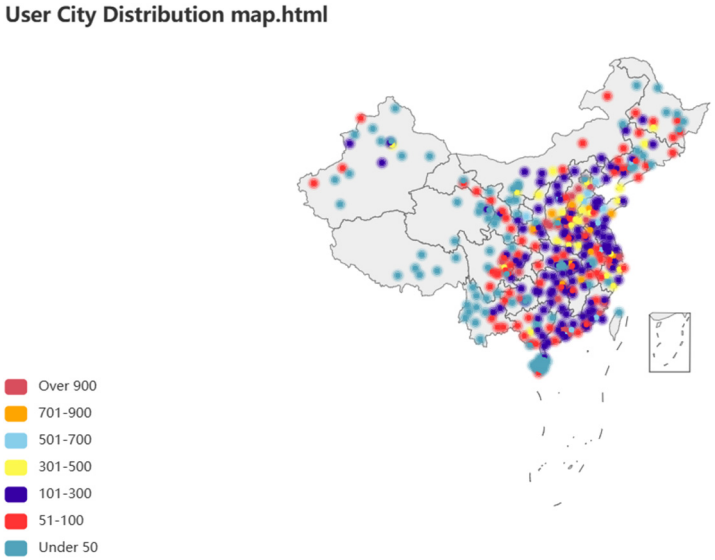

Figure 1. User city distribution map

(3) In order to see the city distribution of the number of users more specifically, we selected the top ten cities based on the word frequency statistics of the city field, and drew the distribution map of the top ten cities with the number of users. There are 1662 users in Shanghai, 1887 users in Shenzhen, 1920 users in Quanzhou, 2160 users in Baoding, 2444 users in Luoyang, 2589 users in Beijing, 3184 users in Guangzhou and 3540 users in Yuncheng. There are 3604 users in Chengdu and 12411 users in Chongqing, with the largest number of users in Chongqing.

(4) Based on the above basis, the age distribution of users in the top ten cities is further analyzed to see if there is any difference in the age distribution. From Figure 2, we can see that the age distribution of users in the top ten cities is mainly young users and middle-aged users, and there is an excessive phenomenon.

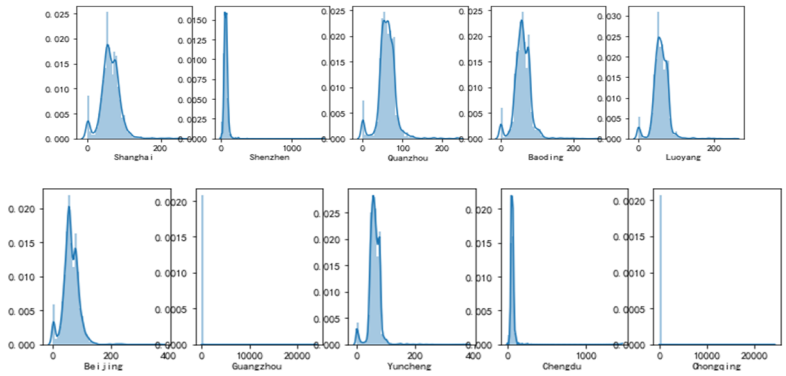

Figure 2. Number of Users TOP10 Urban Age Distribution

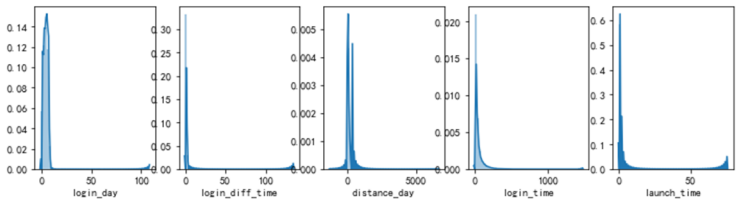

Figure 3. Distribution map of user login

(5) Finally, we analyze the user login situation. A distribution chart is drawn for the number of user login days, the user login interval, the number of days from the last user login to the end of the term, the user login duration and the number of landing pages visited by the user again, as shown in Figure 3. Basically, there is a peak phenomenon in these charts, and there is a slight upward trend in the number of days from the last login to the end of the period and the number of landing pages visited by users again.

3.3 Prediction of user purchase behavior

After completing the processing of the original data set, this step is to determine whether the user will eventually place an order by building a model. Here, we combine the current popular models (such as XGBoost, CatBoost and Random Forest) and improve and evaluate the model, using precision, recall, The accuracy and F1 scores were compared with a variety of indicators.

3.3.1 Principal component analysis

(1) The basic idea of principal component analysis. Principal component analysis (PCA) refers to the method of using the idea of dimensionality reduction to convert multiple characteristic indicators into a few indicators to represent the overall data. Mapping N-dimensional features to K-dimensional features is an unsupervised learning algorithm commonly used in machine learning. Its advantages are that it can make data easier to visualize, reduce the probability of overfitting in some cases, and speed up model training. Var is the variance, n is the sample, xi is the specific value, and xˆis the mean. As shown in (1)[4].

$$Var = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{x})^2 \quad (1)$$

(2) Establishment and Solution of Principal Component Analysis Model.

The theoretical development of matrix decomposition is unique in the industry, Minka, T. P. While doing research at the MIT Media Lab, I found a way to let PCA use maximum likelihood estimation to select hyperparameters. This method can be called by inputting "MLE" as the parameter input of n _ components. Based on this method, p-dimensional random vectors are collected from the standardization of the training set index data. After dimension reduction, the cumulative interpretable variance contribution rate is called, which is 1. The feature after dimension reduction can fully express the original feature. The PCA model based on the training set also reduces the dimensionality of the test set.

3.3.2 Oversampling algorithm

(1) The basic idea of the algorithm. The SMOTE algorithm was proposed by Chawla for unbalanced classification data, and its basic principle is detailed in Figure 4. In unbalanced classification data, the majority unit class is called the positive class, and the minority unit class is called the negative class. Due to the large difference in the number of positive and negative units, the classification accuracy of the traditional classification model often declines, especially for negative units, the classification model can not fully fit its inherent law through the training set data, resulting in relatively low classification accuracy. SMOTE algorithm is a classical over-sampling method to deal with imbalanced classification data, which is different from the random over-sampling method that simply copies the negative class units, but generates new synthetic units by linear interpolation between the negative class units which are close to each other, so as to balance the classification data set and improve the accuracy of the classification model.[5]

(2) Basic assumptions of the algorithm. The units between the closer negative class units are still negative classes, and the balance degree of the data set is improved through the synthetic units of the negative classes.

(3) Algorithm flow. The cell synthesis rate r is determined. Assuming that the number of positive class units is n+ and the number of negative class units is n-, in order to balance the categories of the data set, it is necessary to generate a number of negative class synthesis units ns = n+ - n-, then the unit synthesis rate is formula (2).

$$r = \frac{n_s}{n_-} = \frac{n_+ - n_-}{n_-} \quad (2)$$

Calculate the distance between the negative class units and select the neighboring units. Without loss of generality, let dij denote the Euclidean distance between negative class units xi and xj. For each negative class unit xi (i=1,2,..,n-), the distance vector from other negative class units is denoted by Di = (di1 ,⋯, dij ,⋯,di(n- -1)), from which the smallest b units in dij are selected as the neighboring units.

Resulting in a negative class of synthetic units. Among the b neighboring units selected by the negative class unit xi, r units are randomly selected and denoted as xl (l=1,2,…,r), using xi and xl to generate a new synthesis unit pil according to formula (3), where rand (0,1) represents a random number between 0 and 1. Finally, the r synthetic units of each negative class unit xi are merged into the original data set to form a new data set.

$$p_{il} = x_i + \text{rand}\,(0,1) \times (x_l - x_i) \tag{3}$$

(4) Establishment and Solution of SMOTE Algorithm. In the original data set, the proportion of users placing orders is 3%, and the proportion of users not placing orders is 97%, which is an unbalanced sample, and SMOTE algorithm is used.

SMOTE was performed on the original training set and the overall data set, respectively. The sampling strategy is 0.2, too much sampling will produce redundant data, and too little sampling has little effect. The sample size changes before and after SMOTE are shown in Table 3.

Table 3. Sample size changes before and after SMOTE

	1	0
Before using SMOTE(train)	3208	91263
After using SMOTE(train)	18252	91263
Before using SMOTE(all)	4632	130327
After using SMOTE(all)	26065	130327

3.3.3 Random forest algorithm

(1) The basic idea of the algorithm. Random forest is a kind of combined classifier, which combines a certain number of CART decision trees (base classifiers) through Bagging method. Different decision trees are independent and identically distributed. Finally, the classification results of each decision tree are collected, and the optimal classification results are determined by voting selection mechanism. The essence of Bagging algorithm is to use Bootstrap method to sample the sample set with replacement, and continuously extract samples until multiple groups of different training samples are constructed. Each set of samples is assigned to a weak classifier for training. All weak classifiers summarize the classification results and report them to the strong classifier in the form of votes, and the strong classifier selects the category with the most votes as the final classification result, which is the idea of random forest Bagging algorithm[10][11].

(2) Algorithm flow

The random forest model is based on a tree-type classifier. A large number of trees generate a random forest from a subset of data sets, and the regression model is based on stability and classification accuracy to reduce variance and avoid over-fitting synchronization. The final

decision is to select the winner class by totaling the votes of the component predictors of each class, and then based on the number of votes. Random forest constructs multiple binary decision trees with Bootstrap samples from the learning sample, and randomly selects a set of explanatory variables x at each node. A random forest is a tree-structured classifier {h(x,θk),k=1,2,…}, where x is the input vector and {θk} are independent and identically distributed random vectors depending on the training set, each tree votes for the input vector to decide its class. In the classification problem, the combined model of random forest is shown in (4) [6][7][8].

$$H_R(x) = \arg \max_j \sum_{m=1}^{M} \{I[h(x;\theta_m) = j]\} \quad (4)$$

Generalization error. Given classifiers {hk(x),k=1,2,…}, for a randomly sampled training sample input vector (X, Y), define the marginal function as shown in formula (5).

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \quad (5)$$

The former term is the average number of votes to judge X correctly, and the latter term is the average number of votes of the most misjudged category. The marginal function measures the lowest positive error deviation of the random forest on the input X. The larger the marginal value is, the greater the classification credibility is, thus defining the generalization error of random forest, as shown in formula (6).

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \quad (6)$$

In the formula (6), the subscripts X and Y indicate that the probability value of the error is obtained in the X and Y space. It can be proved by the law of large numbers that when the number of trees in a random forest increases, for all random sequences θ1, θ2, … , θm, PE* converges almost everywhere to formula (7), where θ is a random vector identically distributed with θm and h(X)is a random variable identically distributed with hk(X). Therefore, it can be proved that the generalization error of the random forest algorithm can converge to a limit with the increase of the number of trees, which shows that the random forest algorithm will not appear overfitting phenomenon with the increase of the number of classification trees in the forest[6][7][8].

$$P_{X,Y}\left(P_\theta(h(X) = Y) - \max_{j \neq Y} P_\theta(h(X) = j) < 0\right) \quad (7)$$

3.3.4 Data training and result evaluation

(1) First of all, XGBoost model, CatBoost model and random forest model are used to fit the oversampled training set data, and then the respective models are used to predict the test set. The model evaluation results are shown in Table 4. It can be seen that the CatBoost model performs well in the original model. Continue to evaluate.

Table 4. Evaluation of the original model

		precision	recall	f1-score	support
XGBoost	0	0.99	0.99	0.99	39064
	1	0.78	0.76	0.77	1424
	accuracy			0.98	40488
CatBoost	0	0.99	0.99	0.99	39064
	1	0.79	0.78	0.78	1424

	accuracy			0.98	40488
Random forest	0	0.99	0.99	0.99	39064
	1	0.78	0.71	0.74	1424
	accuracy			0.98	40488

(2) Secondly, considering that there is still room for improvement in the F1 score of Class 1, the parameters of the model are adjusted. The CatBoost model parameters are set to loss _ function = "Logloss", eval _ metric = "AUC", task _ type = "GPU", learning _ rate = 0.01, iterations = 2020, Random _ seed = 2020, od _ type = "Iter", depth = 12 and early _ stopping _ rounds = 500; The XGBoost model parameters are set to n _ estimators = 300, learning _ rate = 0.1, colsample _ bytree = 0.8 and n _ jobs = -1, booster = 'dart'; The random forest model parameters are set to n _ estimators = 300 and criterion = 'entropy'. Through simple parameter adjustment, we find that the F1 score of Class 1 in the evaluation of XGBoost model and random forest model has a significant increase, and the results are shown in Table 5:

Table 5. Model evaluation after parameter adjustment

		precision	recall	f1-score	support
XGBoost	0	0.99	0.99	0.99	39064
	1	0.78	0.77	0.78	1424
	accuracy			0.98	40488
CatBoost	0	0.99	0.99	0.99	39064
	1	0.79	0.77	0.78	142
	accuracy			0.98	40488
Random forest	0	0.99	0.99	0.99	39064
	1	0.78	0.71	0.75	1424
	accuracy			0.98	40488

However, we have been fitting the training set and testing the test set. Finally, we directly use the model after parameter adjustment to fit the whole data set and test the whole data set. The results are shown in Table 6:

Table 6. Holistic Data Model Assessment

		precision	recall	f1-score	support
XGBoost	0	1	1	1	130327
	1	0.92	0.92	0.92	4632
	accuracy			0.99	134959
CatBoost	0	1	1	1	130327
	1	0.92	0.91	0.91	4632
	accuracy			0.99	134959
Random forest	0	1	1	1	130327
	1	0.93	0.92	0.92	4632
	accuracy			0.99	134959

Because the overall data also has the phenomenon of sample imbalance, we use the XGBoost model and the random forest model to fit the oversampled overall data, and the roc _ auc area of the random forest model also reaches 1. The results are shown in Table 7 and Figure 4:

Table 7. Oversampling whole data model evaluation

		precision	recall	f1-score	support
XGBoost	0	1	1	1	130327
	1	0.92	0.92	0.92	4632
	accuracy			0.99	134959
Random forest	0	1	1	1	130327
	1	0.93	0.92	0.92	4632
	accuracy			0.99	134959

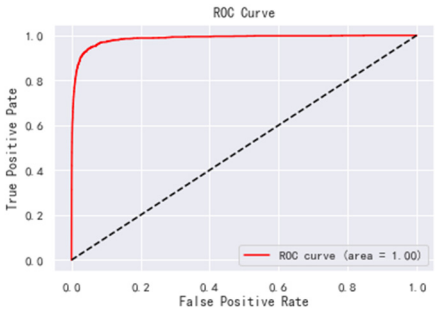

Figure 4. ROC Curve

To sum up, we choose the random forest model with good indicators, the accuracy rate is 99%, the precision of class 1 is 0.93, and the F1 score of class 1 is 0.92. Then, the input will refer to the information in the sample table, and the output forecast results are shown in Table 8. Most users will not place an order.

Table 8. Model prediction results

obs	user_id	result
1	2000001555945280	0
2	2000001556645220	0
3	2000001558047800	0
4	2000001558146460	0
5	2000001558146870	0
…	...	...
197	2000001569847870	0
198	2000001569852740	1
199	2000001569853220	0
200	2000001569853290	0
201	2000001569855100	0

3.4 Suggestions after Value Analysis of Consumer Behavior

Based on the above design, we have a deeper understanding of the overall data, and predict whether the user will place an order to buy. However, in the actual marketing, whether

consumers place orders or not is affected by various external factors, so we give some suggestions on the precise marketing of enterprises.

In the past, marketing was a static way to judge the consumer's consumption ability and consumer behavior orientation through the characteristics of age, gender, occupation and so on, so as to formulate corresponding strategies to promote products. In the digital age, consumers are generating data all the time, behind the data is the mining of potential consumer demand, to achieve precise marketing based on consumers, which is obviously a dynamic way.

(1) Customer segmentation and market positioning

In the marketing promotion of products, enterprises need to subdivide the target user groups, so as to adopt differentiated marketing methods, and then achieve the maximization of product sales profits. In the past, enterprises segmented the market through geographical, demographic and other factors, but now they need to use big data technology to mine useful value information from structured or unstructured data, establish a complete user portrait, and improve the decision-making effect. Therefore, it achieves a greater output with less capital and precise input.

(2) Establishing an Effective Evaluation System of Internet Marketing Effect

The evaluation and feedback of marketing effect is very important after enterprises invest in marketing, and the effective control of this link is conducive to adjusting the marketing strategy of enterprises.

(3) Carry out precision marketing and marketing strategy design

The era of big data provides enterprises with favorable conditions for precision marketing. The channels for obtaining data are diversified. Enterprises can find potential business opportunities in different customer groups and develop personalized marketing services by mining the links between data. Of course, enterprises can also carry out precise advertising through the current popular APP, combined with their own situation. In the marketing process, enterprises can promote through the current hot short video, which helps consumers to understand their own products, thereby improving the conversion rate of users. Of course, good after-sales service is the premise for consumers to consume many times, so behind the success of precision marketing, enterprises can not do without the support of after-sales team[9].

4. EVALUATION AND EXTENSION OF THE MODEL

4.1 Advantages of the model

In this paper, the selection of indicators is more comprehensive. Through variance filtering and principal component analysis, 44 features are finally selected to enlarge 46 features, and the cumulative contribution rate of explained variance is 1.

The assumptions of the model are reasonable, so the model is accurate, can better reflect the actual situation, has a good practical application ability, and the prediction results should be in line with expectations.

The establishment of the model, from the pretreatment of the data, and then according to the establishment and analysis of indicators, can better analyze the consumer behavior layer by layer.

In this paper, we use SMOTE algorithm to solve the imbalance problem of label 0 samples. Through the analysis and judgment of several commonly used algorithm models, we use the improved random forest sample model based on imbalanced samples.

For the improved random forest, the processed data set is trained, and the final accuracy rate is 99%, and the F1 score also reaches 0.92, so it has a good prediction effect on the data.

4.2 Disadvantages of the model

The data of the model is structured, but there may be unstructured data in practice, so it is very important to realize the fusion analysis of multi-modal data and improve the accuracy of marketing.

In reality, data is dynamically generated, so when the user's consumption behavior data is constantly changing, how to update and load is a problem that needs to be considered.

4.3 Generalization of the model

In this paper, an improved random forest model based on unbalanced sample is constructed to analyze the statistical data of user consumption behavior, and variance filtering and principal component analysis are used to screen the indicators, and the impact of user information and behavior characteristics on whether users buy courses is established. In the fourth task, we give some suggestions, which have certain reference significance for the precise marketing of the company's products, enhancing the willingness of users to buy products, improving the conversion rate of users, and enhancing the brand influence of the company.

REFERENCES

[1] Zhu Xuehong. Research on Consumer Behavior Intention of Mobile Internet Users [D]. Nanjing University of Posts and Telecommunications, 2011.

[2] Zhai Jinzhi. Analysis of Internet Users' Consumption Behavior Based on Big Data [J]. Journal of Commercial Economics,2020(24):46-49.

[3] J. Han and M. Kamber et al., Data Mining: Concepts and Techniques [M]. Beijing China Machine Press, 2012.

[4] Li Yanshuang, Zeng Zhenxiang. Application of Principal Component Analysis in Multi-index Comprehensive Evaluation Method [J]. Journal of Hebei University of Technology,1999(1):94-97.

[5] Yang Guijun, Du Fei, Sun Lingli. Propensity Score Matching lmputation Based on Synthetic Minority Over-sampling Technique [J]. Journal of Statistics and Information,2021,36(01):3-12.

[6] Liaw A, Wiener M . Classification and Regression by random Forest[J].R News,2002,23(23).

[7] Ma Xiaojun, Dong Biying, Wang Changxin. Design of a Credit Rating Model of Quoted Companies Based on the PSO Optimized Weighted Random Forest Algorithm [J].Journal of Quantitative & Technological,2019,36(12):165-182.

[8] Leo Breiman. Random Forests.[J]. Machine Learning, 2001, 45(1) : 5-32.

[9] Wang Bo, Wu Ziyu. Research on Precision Marketing Model in the Era of Big Data [J].China Economist, 2013(05):16-18.

[10] Yang Guijun, Sun Lingli, Li Lu. Response Propensity Score Matching lmputation [J]. Journal of Statistics and Information,2018,33(08):3-11.

[11] Wei Chen et al. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility[J].Catena,2017,151：147-160.