International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-13

https://doi.org/10.70917/ijcisim-2026-0401
Article

A Study of Digital Marketing in the Housing Market and the Prediction of Consumer Purchasing Behavior

Lanlan Zhou [1,*]

[1] Sichuan University of Science & Engineering, Yibin, Sichuan, 644002, China
* Correspondence author: zhoulan13785@163.com

Abstract: Driven by the wave of digital economy, the real estate industry is facing the transformation of marketing methods. Traditional marketing means have been difficult to meet the precise and diversified consumer demand, and digital marketing means have gradually become an important strategy in the housing market due to its advantages of high efficiency and wide coverage, which has pushed the research on consumer behavior prediction to become an emerging hot spot. This study focuses on the application of digital marketing in the housing market and explores how to predict consumer purchase behavior through multi-model fusion. First, the behavioral data of 13 million users of the real estate e-commerce platform are preprocessed, and the SMOTE and Borderline-SMOTE methods are used to achieve data balance. In terms of modeling, three single models, logistic regression, random forest and XGBoost, are selected for prediction, and the prediction performance is enhanced by Voting fusion algorithm. The empirical analysis results show that the F1 value of Random Forest among the single models is the highest up to 73.68%, while the weighted voting fusion model performs the best in terms of comprehensive performance, with the F1 value elevated to 81.51%, and the AUC value up to 0.6723. The results verify the effectiveness of the fusion model in predicting the consumer purchasing behaviors, and provide data support and technical basis for the implementation of precision marketing in the housing market. The conclusion states that digital marketing tools combined with multi-model fusion prediction technology can significantly improve the conversion rate of home purchase and user identification.

Keywords: digital marketing, housing market, consumer behavior prediction, random forest, XGBoost, model fusion

1. Introduction

At present, the wave of digitalization is sweeping the world, practicing digital transformation and upgrading and improving the operation level and competitiveness of real estate development enterprises has become an important wrestling field for real estate development enterprises [1-2]. With the rapid innovation and application of digital technologies such as mobile Internet, big data, artificial intelligence and other digital technologies, the pace of digital transformation of real estate development enterprises has accelerated significantly, and digital marketing has become a key link in the digital transformation of real estate development enterprises [3-6]. How to carry out systematic digital marketing management based on the user's personalized needs has become an urgent problem for real estate development enterprises.

The continuous development of the housing market has made essential changes in the relationship between supply and demand, and users have put forward higher requirements for housing quality and living environment [7]. At the same time, the changes in the housing market have also changed the user

consumption behavior, the user both consumption behavior and decision-making path has changed dramatically [8]. User-centered personalized needs continue to emerge, and the relationship between users and real estate development enterprises has also shifted from a single offline contact point to a dual online and offline contact point [9-11]. Therefore, in-depth excavation, accurate understanding of the user's personalized needs, timely tracking and response to changes in user demand has become the center of gravity of real estate development enterprises [12-13]. If the enterprise wants to more real and close to understand the consumer's willingness to buy and demand, and thus enhance their core competitiveness, in the fierce and cruel real estate market to win [14-15]. Only through the analysis and prediction of consumer behavior and psychological needs, so as to carry out a series of housing product development, positioning and sales and other follow-up work for the needs of potential customers [16-18].

Under the background of accelerated development of social informatization and digital economy, traditional industries have explored the path of digital transformation, real estate as a high-value low-frequency typical industry, the mastery of the user's decision-making behavior has a very high demand. Consumers' decision-making cycle is long and has many influencing factors, so how to achieve personalized recommendation and accurate reach has become an important issue in digital marketing. Especially in the era of rapid development of big data and artificial intelligence, the large amount of behavioral data left by users on e-commerce platforms provides a solid foundation for the prediction of housing consumption behavior. By analyzing users' clicking, consulting and viewing behaviors, it is possible to build more accurate user profiles and recommendation models, thus improving service experience and transaction efficiency. In the past, many studies have focused on the impact of traditional economic variables on the real estate market, while the systematic mining and application of user micro-behavioral data has not been sufficient. Therefore, introducing machine learning methods into the modeling of consumer home-buying behavior and constructing a multi-model fusion mechanism can help break through the limitations of traditional marketing prediction and empower the digital transformation of real estate enterprises.

In this paper, firstly, based on the large-scale behavioral data of the real estate e-commerce platform, we complete the cleaning, label setting and over-sampling processing of unbalanced data. Subsequently, three single prediction models, logistic regression, random forest and XGBoost, are constructed respectively, and the Voting fusion algorithm is applied to synthesize the advantages of multiple models in order to enhance the prediction effect. Finally, the optimal model is selected through the comparison of indicators and empirical verification, and the prediction results are used to provide technical support for the precise marketing strategy of the housing market. The whole research path emphasizes the integrated logic of “data-model-application”, and is committed to building an effective bridge between theory and practice.

2. Precision marketing strategies for the housing market in the context of digital marketing

In the era of digital economy, the real estate industry must follow the pace of digital marketing in time and implement precision marketing. The article focuses on exploring the application of digital marketing in the housing market and studying the countermeasures of the real estate industry to implement precision marketing in the era of digital marketing.

2.1. Digital marketing in the housing market

In the digital era, digital marketing has become an important strategy for the real estate industry to promote its products and services. Specifically, with the help of communication technology and computer network technology, digital marketing digitizes the marketing process to help the real estate industry anticipate changes in the housing market and guide the development and implementation of marketing strategies, which covers a wide range of factors including network marketing technology, Internet communication channels and traditional media channels, such as television, SMS, and so on.

Among them, network marketing technology to electronic advertising, search engine optimization, content marketing and other means, so that the real estate industry in the Internet to establish a brand image, enhance product exposure, to attract the attention of the target audience, network marketing technology also uses big data analysis and accurate positioning technology, to achieve in-depth mastery of consumer behavior, and thus improve the marketing effect.

The rise of social media has made the Internet communication channel an indispensable part of people's daily life, the real estate industry can establish a social platform account, direct interaction with consumers, disseminate brand concepts, improve consumer awareness and loyalty, while the Internet communication channel allows the real estate industry to quickly respond to consumer demand

and feedback, so that the two sides to establish a closer relationship, to achieve the emotional connection between the brand and consumers.

2.2. Real Estate Precision Marketing Countermeasures under the View of Digital Marketing

2.2.1. Accurately analyze customers' home-buying needs

In today's rapidly changing housing market, the real estate industry needs to accurately analyze the needs of customers to buy a home, this paper here will be divided into three categories of demand for home ownership: self-occupancy, preservation of value, speculation.

(1) Owner-occupied homebuyers

Owner-occupied home buyers is one of the most robust groups in the home-buying market, more important to the comfort and convenience of the living environment, the primary goal of buying a home is to meet the needs of their own and their families to live, focusing on the quality of housing construction, renovation standards, and the community's supporting facilities and perfect sex. Providing high-quality residential products and thoughtful after-sales service will win the trust and loyalty of homebuyers.

(2) Value-preservation type home buyers

Value-preserving home buyers mainly buy real estate as an asset to hold for a long time, to realize asset preservation and appreciation, pay more attention to the return on property investment and the potential for future appreciation. For this type of buyers, the real estate industry should focus on the value of property investment and potential value-added space, in the project promotion and sales process, the need to fully demonstrate the advantages and value to the preservation-type buyers.

(3) Speculative home buyers

Speculative homebuyers usually pursue rapid speculation and high returns, focusing on independent property rights and market liquidity, hoping to realize rapid growth in property value in a short period of time. For this type of buyers, the real estate industry should choose projects with obvious appreciation potential and provide flexible purchase policies and efficient transaction processes to attract speculative buyers.

2.2.2. Precise customer orientation

(1) Market segmentation

Market segmentation refers to the division of the market into different consumer groups based on factors such as consumer purchasing behavior and differences in demand. The housing market is mainly subdivided according to different criteria, such as grade, product category and household size, and developers need to have a deep understanding of the characteristics of each subdivided market in order to satisfy customers' needs so as to enhance customer satisfaction and sales conversion rate.

(2) Target Market Selection

On the basis of project planning and market characteristics, the real estate industry needs to focus on selecting the target market, digging deep into its potential and opportunities, and formulating corresponding marketing strategies for each market segment, i.e. pricing strategies, promotional activities, channel selection, etc., to ensure acceptance and recognition by target customers.

(3) Market Positioning

When determining the market positioning of the project, it is necessary to clarify the core competitive advantages of the product. In the promotion process, according to the characteristics of different customer groups, we should accurately select the appropriate marketing channels and communication methods to ensure the efficiency and accuracy of the information conveyed. Feedback data and market response should be followed up in a timely manner to optimize the strategy and program to achieve excellence and continuous innovation.

2.2.3. Precise formulation of product strategy

(1) Project Development Strategy

Under the digital marketing perspective, the real estate industry should not only focus on the quality and innovative development of the product itself, but also implement precise marketing strategies in planning and design, property management and services. Developers should be aware of market trends and adjust planning and design programs in a timely manner, and utilize data analysis and market research to grasp the preferences and needs of target customer groups. In terms of property management and services, developers should ensure efficient and high-quality operations to improve

community quality and living experience. On the basis of refined management and services, developers should also pay attention to details to create a perfect living space.

(2) Landscape Design

In the digital marketing perspective, the success of real estate precision marketing cannot be separated from the importance of landscape design and integration.

First, precision marketing needs to deeply grasp the needs and preferences of the target customer groups, and highlight the characteristics and advantages of landscape design in the development of marketing strategies to attract customer attention.

Secondly, the real estate industry should integrate the concept of people-oriented into landscape design and precision marketing, and customize landscape design solutions to enhance user experience and satisfaction for the needs of different groups.

Third, in the era of digital marketing, the real estate industry should make use of social media, mobile applications and other channels to show potential customers the project's landscape design concepts and features, and use VR technology, so that customers can feel the charm of the project's landscape design without actually looking at the house, and improve the willingness to buy a house.

(3) Household design

Scientific and reasonable house type design can promote the sales performance of the housing market. In order to ensure that the house type design meets the needs of customers and contributes to the purchase behavior, it is necessary to integrate the concept of precision marketing into real estate marketing, and tailor the house type design to meet the needs of customers, so as to enhance the willingness of home buyers to purchase.

(4) Realize one-to-one communication

In the context of digital marketing, how to implement precision marketing strategy has become a key point in the real estate industry to compete for market share. The core of precision marketing is one-to-one communication, sales staff and customers face-to-face exchanges, a comprehensive understanding of customer needs and real ideas, and the establishment of a long-term stable relationship to enhance customer goodwill, leading to sustained cooperation. In this process, the real estate industry should establish a perfect tracking reception system to ensure that every customer interaction is properly handled.

(5) Accurate marketing promotion

In the era of digital marketing, the real estate industry must combine traditional offline marketing means and online networks to meet the diversified and personalized needs of consumers and enhance competitiveness with precise marketing promotion. Online services such as VR viewing and providing project information allow customers to intuitively grasp real estate projects at home, saving time and energy, while offline precision marketing activities are organized to strengthen interaction with customers and improve service quality.

3. Voting fusion-based model for predicting consumer housing purchase behavior

In this chapter, based on oversampling consumer behavioral data, Voting method is used to fuse multiple single prediction models in order to predict consumer housing purchase behavior and realize digital precision marketing.

3.1. Oversampling treatment methods

3.1.1. SMOTE oversampling

SMOTE oversampling achieves the effect of balancing the data distribution by linear interpolation in the original data, where the interpolation space is located in the original data space. The basic idea of the method is to analyze the minority class samples and artificially synthesize new samples based on the minority class samples to achieve the balancing effect. The basic steps of the algorithm are as follows:

Step 1: Take each individual X_i of the smaller class as the root sample for synthesizing a new sample.

Step 2: Select a sample X_i among the fewer class samples, and search for its nearest K samples in the root sample.

Step 3: Assuming a sampling multiplicity of n on the dataset, randomly select a sample from the K nearest-neighbor samples as an auxiliary sample for synthesizing a new sample, and repeat n times ($n < K$), denoted as $y_1, y_2 ... y_n$. The sampling multiplicity n usually depends on the degree of

imbalance between the majority and minority classes of the original data.

Step 4: Construct n synthetic samples by a random linear interpolation operation of X_i with y_j via equation (1):

$$Z = X_i + (y_j - X_i) \times \gamma, i = 1,2,...,N, j = 1,2,...,n \tag{1}$$

where X_i is the sample data in the minority class, y_j is the j th auxiliary sample of the sample X_i, γ is the random number between $[0,1]$, and Z is the synthesized new sample.

3.1.2. Borderline-SMOTE oversampling

The Borederline-SMOTE oversampling algorithm is an improvement on the SMOTE oversampling algorithm, which performs on-the-fly linear interpolation of only a few classes of samples distributed near the classification boundaries, and increases the number of few classes of samples included in the boundaries used to determine the classification, making the newly synthesized samples more reasonable. The basic steps of the algorithm are as follows:

Step 1: Take each individual X_i of the smaller class as the root sample for synthesizing a new sample.

Step 2: Select a sample X_i in the minority class sample and search for its nearest K samples in the root sample.

Step 3: Classify the minority class samples into three categories: safe samples, invalid samples and boundary samples. If more than half of the K nearest neighbor samples are minority samples, they are safe samples. If more than half of the K near-neighbor samples are majority class samples, they are boundary samples. If K near-neighbor samples are all majority class samples, the samples are all noise and are classified as invalid samples.

Step 4: Apply the same method as the SMOTE algorithm, only for the boundary samples for linear interpolation operation, after the synthesis of the minority class sample distribution is relatively more reasonable.

3.2. Single prediction model

3.2.1. Logistic regression

Logistic regression is a classical machine learning algorithm, a predictive model for classification, often used for binary prediction [19]. Logistic regression is essentially: if a certain data obeys this distribution, use the great likelihood estimation to estimate the parameters of this distribution. The distribution of logistic regression is a continuous probability distribution, the distribution function and probability density function of the regression are shown below:

$$F(X) = P(X \le x) = \frac{1}{1 + e^{-(x-\mu)/\gamma}} \tag{2}$$

$$f(x) = F'(X \le x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma(1 + e^{-(x-\mu)/\gamma})^2} \tag{3}$$

where μ denotes the location parameter and $\gamma > 0$ is the shape parameter. The logistic regression distribution is defined as a continuous function by its positional and shape parameters. The shape of the logistic regression distribution is similar to that of the familiar normal distribution, but the right-hand side of the logistic regression distribution is narrower, so for data with longer tails and peaks, we can use the logistic regression distribution to model the data compared to the normal distribution. The Sigmoid function, which is common in machine learning, is a special form of the distribution function of logistic regression in $\mu = 0, \gamma = 1$.The formula for the Sigmoid function is shown below:

$$\sigma(t) = \frac{1}{1 + e^{-z}} \tag{4}$$

Logistic regression function also has strong robustness, and the input range of the function $(\infty, -\infty)$, the final output range $(0,1)$, so that the between and has a probabilistic meaning. Input a sample into the logistic regression function, the output result is 0.7, meaning that this sample has a 70%

probability of being a positive case, when 1-70% is a 30% probability of being a negative case.

3.2.2. Random Forests

Random Forest (RF) [20] has a promising application as a relatively new and highly flexible machine learning algorithm model. Random forest is a further extension of decision making, so decision trees are introduced first. Decision tree is to use the tree structure to carry out the construction of a classification model, each of its nodes represents a feature, and then based on this feature is divided, the children of this node is split into leaf nodes, each leaf node represents a feature category, and ultimately can be classified. Commonly used decision trees are ID4, C4.5, CART and so on. And in the process of tree generation, you need to choose which feature to analyze first. In general, in principle, the separation needs to improve the correlation accuracy as much as possible. This step can be measured by evaluation metrics such as information gain, gain rate and Gini coefficient. However, if a tree is generated, a pruning operation is also performed to further avoid the overfitting problem. Decision tree is shown in Figure 1, decision tree prediction, at the internal node of the tree with a certain attribute value for judgment, and finally according to the results of the tree model judgment, to decide which sub-node to enter, until it reaches a leaf node, to get the final classification results.

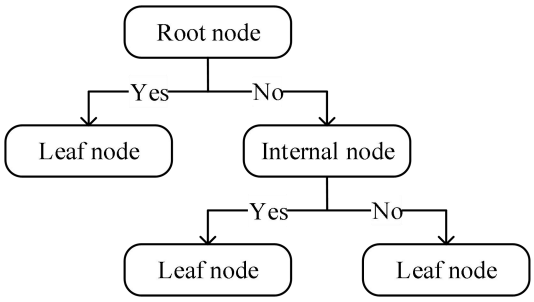


Figure 1. Decision Tree

Random forest is built from multiple decision tree models, and there is no connection between different decision trees. When the classification task is performed, each decision tree in the random forest judges and classifies the input sample data, and then each decision tree will get a final result of classification, and finally count the classification results of all the decision trees, which one of them has the largest number of classifications, and then the random forest will take the result of this category as the final result. Overall, the random forest model is still a new approach to bagging, which essentially adds a decision tree algorithm to the bagging model. The model first uses the bootstrap algorithm to generate N training datasets and then constructs a decision tree for each training dataset. When a node finds a feature to be segmented, it randomly takes a portion of the feature data and finds the optimal solution among the taken features and applies it to each node to be segmented. The random forest method has the bagging process, which is equivalent to the process of sampling each row and each column with the training data as a matrix, which can avoid the overfitting phenomenon that may occur during the model training process to a certain extent.

3.2.3. XGBoost

XGBoost [21] is a type of GBDT, which is an iterative decision tree algorithm whose basic idea is to reduce the residuals from the model training process for classification and regression. The algorithm combines decision tree with integration thinking. XGBoost is a more efficient implementation based on GBDT, which achieves a great improvement in the speed and efficiency of training.XGBoost is an optimized distributed gradient enhancement model, which aims to achieve efficiency, flexibility and portability. The biggest difference between XGBoost and GBDT lies in the definition of the objective function. The objective function for model training consists of two parts, one is the training error and the other is the regularization as shown below:

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) \tag{5}$$

$L(\Theta)$ is used to represent the training error function, $\Omega(\Theta)$ represents the training regularization function, when the model starts training, the prediction is good or bad in time by the training error function, at the same time, in order to reduce the overfitting, the regularization function is introduced. The XGBoost model, after the forward division addition and the Taylor expansion, its objective function is is shown below:

$$Obj^{(t)} \simeq \sum_{i=l}^{n}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{6}$$

The model first needs to go through a training process to find the first and second order derivatives of the loss function at each step, and then further come to regularization to optimize our objective function so that $f(x)$ can be found at each time, and finally use the additive model to find the overall model. With the objective function, which represents the score of a specified tree structure in the model, the objective function is utilized to find the first-order derivatives g_i and second-order derivatives h_i of each sample at each node, and then the corresponding samples at each node are summed up to G_j and H_j, and finally the final value of the objective function is obtained by iterating through all the nodes can get the final value of the objective function.

The XGBoost algorithm will also utilize the rate of greedy algorithm and approximation algorithm to determine the optimal way to slice the leaf nodes during the actual training process. The actual project appears in the case of sparse data, such as default values and unique hot coding, XGBoost only considered no default value data modeling, but at the same time for each node to increase the default direction, when there is a corresponding missing, you can put this data in the default direction, and the optimal default direction can ultimately be learned from the data model. This treatment makes the model faster to model on sparse data.

3.3. Voting-based multi-model fusion approach

Voting fusion method [22] is a fusion method based on the single model prediction results and adopts the majority voting method to select the final prediction results. The basic idea of this method is that the single model prediction results are summarized and counted, and the classification result with the highest percentage is taken as the final prediction result of the fusion model. The voting method usually contains three voting methods: absolute majority voting method, relative majority voting method and weighted voting method. Since the absolute majority voting method will not output the final prediction results, the relative majority voting method and the weighted majority voting method are mainly adopted for the research of this paper.

3.3.1. Relative majority voting

The core idea of the Relative Majority Voting (RMV) method is that the minority follows the majority. In a single model prediction result, the prediction category with the highest percentage is used as the final category of the sample to be tested. If more than one category has the highest percentage, one of the categories with the highest percentage is randomly selected as the final categorization category of the sample to be tested. In the application of the binary classification problem when the number of selected single models is odd, the phenomenon of multiple multinomials can be effectively avoided. Its specific prediction expression is as follows:

$$result = c_{\arg\max\left(\sum_{i=1}^{m} h_i^j(x)\right)} \tag{7}$$

where $\{c_1, c_1, \ldots, c_N\}$ is all possible prediction categories of the sample, and $\left\{h_i^1(x), h_i^2(x), \ldots, h_i^N(x)\right\}$ is the N-dimensional vector predicted by the single model h_i on the sample, $h_i^j(x)$ is the category labeling c_j on the category labeling c_j.

3.3.2. Weighted voting method

The weighted voting method is a voting method that adds model weight information on the basis of the relative majority voting method and still obeys the minority to the majority. Its idea is similar to the weighted average method, according to the prediction effect of a single model to determine the weight of each model w_i, the better the model prediction effect the greater the corresponding weight. Multiply the prediction result of each model by the corresponding weight w_i, and sum it up. The category corresponding to the maximum value after weighted summation is taken as the final prediction effect of the sample to be tested. Its specific prediction expression is as follows:

$$result = c_{\arg\max\left(\sum_{i=1}^{m} w_i h_i^j(x)\right)} \tag{8}$$

where $w_i > 0$, $\sum_{i=1}^{m} w_i = 1$.

4. Empirical analysis of forecasts of consumer purchasing behavior in the housing market

This chapter provides a practical application of the proposed multi-model fusion approach for predicting consumer housing purchase behavior to validate the validity of the model.

4.1. Data sources and data pre-processing

4.1.1. Data sources

In order to better empirically exercise the established model, this paper uses the official dataset derived from a real estate e-commerce platform, the dataset contains the platform from October 24, 2022 to November 5, 2024, the original dataset contains the number of consumers 134,268,872, the number of commodity houses 5,053,246, and the number of all behaviors is more than 1 billion, of which the purchase behavior accounts for only than about 0.1%. It involves about 13 million consumers who have behaviors, and the behaviors include clicking, consulting, viewing, and purchasing. Other data fields include Consumer ID, Property ID, Property Category ID, Behavior Type, and the timestamp of when the behavior occurred.

4.1.2. Forecasting objectives

The main problem to be solved in this paper is how to predict the future purchase behavior of consumers through their behavioral data generated on real estate e-commerce platforms. How to analyze the huge amount of implicit consumer feedback data, predict consumer behavioral goals, better recommend properties of interest to consumers, and improve the service level to facilitate transactions is the key to improving the service level of real estate e-commerce platforms.

In this paper, we hope to establish a machine learning-based prediction model of consumer housing purchase behavior by using the historical data of 13 million consumers' behaviors in different categories of commodity houses and at different moments provided by the real estate e-commerce platform. Taking the commodity house as the smallest unit, by studying the data of the behavioral sequences generated by consumers under a certain category of commodity house, and combining the other characteristic indicators of consumers and commodity houses, we predict whether consumers will generate purchasing behaviors for that commodity house after generating that kind of behavioral sequences.

Judging whether a consumer will buy a commodity house through the behavioral data of the consumer can be transformed into binary classification in machine learning. The classification objectives are divided into two types, which are labeled using 0 and 1 in the specific data: purchase, will not purchase. If the consumer produces a purchase behavior under this type of behavioral sequence, the record is marked as 1.

4.1.3. Data analysis

In order to better model the selected data and process the original data, this paper first conducts statistics on the original data, and describes and analyzes the statistical results, and adopts appropriate pre-processing methods to filter and visualize the statistics according to the characteristics of the data and the statistical situation.

First of all, the click behavior occurs at the point of time to carry out statistics. Consumer clicking behavior statistics are shown in Figure 2, where the number of clicks is the daily average of the selected time period. From the statistical results, it can be seen that consumers are highly active between 19:00 and 23:00, which means that consumers are most active in logging in and browsing the website after work, and reach the highest at night before going to bed, which is in line with people's daily work and rest patterns. Secondly, the daytime peak is from 11:00 to 15:00, consumers are more active near the time of lunch break, and the clicking behavior is smoother from 14:00 to 17:00, which shows that the time point of consumers' online browsing is also related to the behaviors generated by consumers.

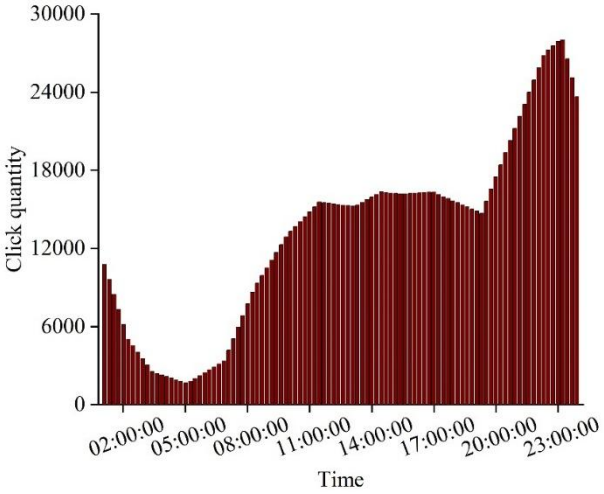
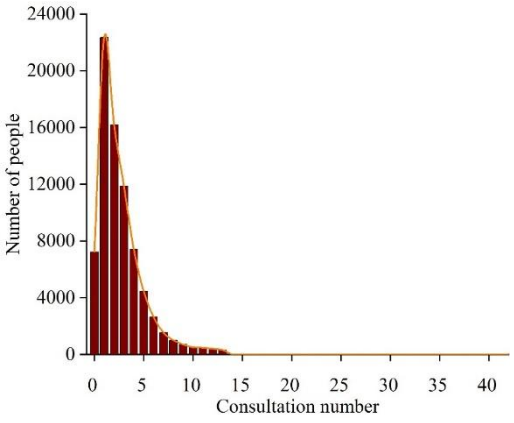
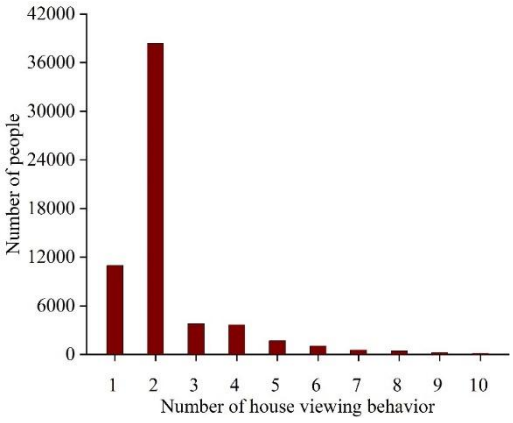


Figure 2. Statistics of consumers' click behaviors

Secondly, from a qualitative perspective, consumers will generate certain clicking, consulting or viewing behaviors before generating purchasing behaviors on real estate e-commerce platforms. The statistics of pre-purchase consulting and viewing behavior are shown in Figure 3, where (a) indicates consulting behavior and (b) indicates viewing behavior. An analysis of the behavioral data of consumers who generate purchase behavior from the dimension of commodity houses reveals that before generating purchase behavior, most consumers generate about 0-6 times of consultation behavior to learn about the details of commodity houses and other information, of which the vast majority generates 1 time of consultation behavior. After making inquiries, consumers will conduct 1-4 viewings, with the majority of them having 2 viewings. Basically, all consumers will conduct at least one viewing behavior, which is in line with the daily understanding of consumer home buying behavior. In summary, consumers tend to make purchases of commercial properties only after generating a longer sequence of operations when they are active on real estate e-commerce platforms.



(a) The consultation behavior before the purchase behavior



(b) The house viewing behavior before the purchase

Figure 3. Behaviors statistics before purchase

In the analysis of consumer behavior, the conversion rate of consumers is a part that cannot be ignored. An analysis of the behavior of the 13 million consumers in the original dataset and the number of housing categories in which the behavior was generated reveals that the vast majority of the behavior generated by consumers before making a purchase is clicking behavior, and the number of housing categories involved in the clicking behavior is less than one-fourth of the number of clicking behaviors generated by the clicking behavior. In the consultation and viewing behavior, the number of consumer behaviors is still greater than the number of housing categories generated by their behaviors. It can be seen that consumers tend to operate multiple times on the selected category when purchasing on the platform, and repeatedly click, consult and look at the same type of housing before purchasing, with the

clicking behavior accounting for the vast majority. Since the data applied in this paper is the behavioral data within two years, it indicates that consumers may make few purchases of the same kind of commodity houses in a longer period of time, reflecting the long-term nature of the consumer purchasing cycle in the housing market.

4.2. Analysis of model prediction effects

4.2.1. Comparison of the predictive effectiveness of single models

After tuning the parameters of a single model, we run the model to obtain the prediction effect of each model as shown in Table 1, where TP refers to the number that the model predicts as purchased and actually purchased, FP refers to the number that the model predicts as purchased and actually did not purchase, TN refers to the number that the model predicts as not purchased but actually purchased, and FN refers to the number that the model predicts as not going to purchase and actually did not purchase.

Since there is a negative correlation between precision and recall, we take the F1 value of the reconciled average of precision and recall as the evaluation criterion, thus we can see that the prediction effect of random forest is better, and the F1 value is 73.68%, which is greater than the other two models. The logistic regression has a higher AUC of 0.6325, indicating that it is more classifiable, but its ACU is not much different from that of the random forest, so the prediction effect of the random forest will be relatively better in these three single models, but the overall prediction effect is low.

Table 1. Comparison of the prediction effects of a single model

Indicator	Logistic regression	XGBoost	Random forest
TP	25	32	35
FP	10	16	17
TN	14	9	8
Precision	71.43%	66.67%	67.31%
Recall	64.10%	78.05%	81.39%
F1	67.57%	71.92%	73.68%
AUC	0.6325	0.5954	0.6231

4.2.2. Fusion-based model predictions

Due to the general prediction effect of a single model, in order to improve the accuracy of the prediction model in predicting user purchase behavior. In this paper, we will combine the advantages of XGBoost in overfitting processing and training speed, the robustness of logistic regression algorithm to small noise in data and the advantages of random forest in prediction accuracy, and combine the three models mentioned above through the method of Voting classifier integrated learning for prediction, so as to get better prediction effect.

In this paper, we will use the two methods of relative majority voting and weighted voting for model fusion respectively, and compare the prediction effect of these two fusion models, and choose the one with better prediction effect to compare with the single prediction model. In this case, in order to avoid the phenomenon of overfitting, in the choice of the weighting method regarding weighted voting, equal weights were chosen to fuse each single model.

The prediction effect of the fused prediction model by relative majority voting and weighted voting is shown in Table 2, from which it can be seen that there is not much difference in the prediction effect of the two fusion methods, but relatively speaking, the prediction of the weighted voting is a little bit better, with its F1 value and AUC value of 81.51% and 0.6723, respectively. Although relative majority voting and weighted voting have one advantage and one disadvantage in the comparison of the results of precision and recall, but since these two are constrained by each other. Then for the comparison of the prediction effect of relative majority voting and weighted voting, the reconciled mean F1 value and AUC value of the two will be used as the basis of comparison, so that it can be seen that the value of weighted voting is higher than the relative majority voting, which indicates that its prediction and classification is more effective, so the prediction effect of weighted voting will be used to compare with each single model.

Table 2. Comparison of the prediction effects between plurality voting and weighted voting

	Precision /%	Recall /%	F1 /%	AUC
Plurality voting	79.82	82.65	81.21	0.6485

Weighted voting	79.74	83.36	81.51	0.6723

4.2.3. Comparison of forecast results

The prediction effect of the fusion prediction model and each single prediction model is shown in Fig. 4, the fusion model has improved the prediction effect relative to a single prediction model, although the recall rate of the random forest is higher than that of the fusion model, but the comprehensive effect of prediction is still better than the fusion model, in terms of the F1 value of the fusion model is improved by 9.61% relative to the best prediction effect of the random forest prediction model.

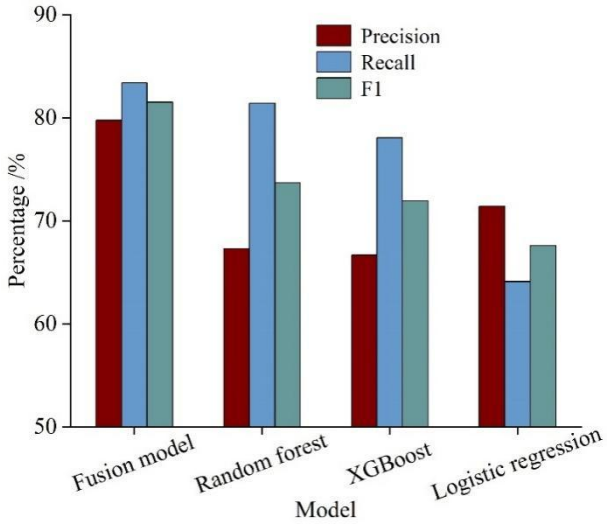


Figure 4. Comparison of the prediction effects between the fusion model and the single model

The above conclusion can also be corroborated by the ROC curve, and the comparison of the ROC curves of the fusion model and the single model is shown in Figure 5. It can be seen that the AUC value of the fusion model is higher than that of all other single models, indicating that the classifiability of the fusion model is better than that of the single model, which in turn proves that the prediction effect of the fusion model is improved compared with that of the single model.

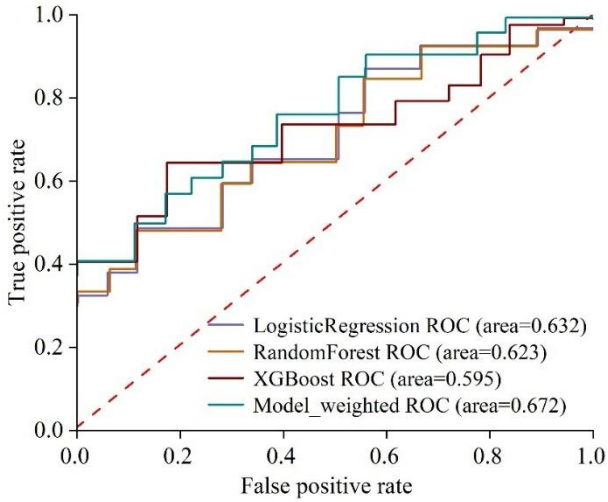


Figure 5. Comparison of the ROC curves of the fusion model and the single model

4.2.4. Predicted Results of Platform Consumer Home Buying Behavior

After completing the selection work of the prediction model, extracting the predicted user feature information, processing the sample data in accordance with the data cleaning process of the training model, and outputting the user selection label as well as the probability according to the model, it can be learned that more than 60% of the consumers in the selection of the purchase behavior of the

consumer will choose a new house for purchase. The prediction results of consumer purchasing behavior of some platforms are shown in Table 3. The label output rule is: when the probability of the model output is greater than or equal to 0.6 is predicted as the resident will choose a new home to purchase, otherwise it is predicted that the resident will choose a second-hand home to purchase.

Table 3. Some platform consumer buying behavior forecast results

User ID	Label	Probability
1	0	0.5943
2	0	0.5412
3	1	0.7958
4	1	0.8524
5	0	0.5762
6	0	0.5241
7	0	0.4976
…	…	…
274	1	0.8854
275	1	0.7957
276	0	0.3948
277	1	0.6593
278	1	0.8256
279	0	0.3145
280	1	0.8374

5. Conclusion

The fusion model in this study outperforms all single models in terms of predictive effectiveness and improves the accuracy of consumer home buying behavior identification. Specifically, the prediction model under the weighted voting method achieves 83.36% recall and 79.74% precision, which improves the performance of the F1 value by 19.26% and 11.69% compared to logistic regression and XGBoost, respectively. Through the ROC curve analysis, the AUC value of the fusion model is 0.6723, which is better than the average value of 0.6170 of the single model, indicating that its classification ability is stronger. In terms of user behavior prediction, after the model is applied to platform consumer data, it shows that more than 60% of users prefer to choose new houses rather than second-hand houses, with a probability threshold of 0.6. These results show that the fusion strategy based on Voting integrated learning not only significantly improves the model stability and generalization ability, but also provides the real estate industry with the theoretical support and decision-making tools for accurate marketing and resource allocation.

References

1. Matidza, I., Ping, T., & Nyasulu, C. (2020). Use of digital marketing in estate agency industry in Malawi. E-Learning and Digital Media, 17(3), 253-270.
2. Ullah, F., Sepasgozar, S. M., & Wang, C. (2018). A systematic review of smart real estate technology: Drivers of, and barriers to, the use of digital disruptive technologies and online platforms. Sustainability, 10(9), 3142.
3. Kaur, H. (2017). Digital marketing and its impulsiveness in real estate. International Journal of Management IT and Engineering, 7(12), 147-153.
4. Rabby, F. (2022). Digital transformation in real estate marketing: A review. Available at SSRN 5101903.
5. Bansude, S., & Hittalmani, V. (2021). Impact of digital marketing on real estate business. International journal of health sciences, 5(S1), 172-183.
6. Low, S., Ullah, F., Shirowzhan, S., Sepasgozar, S. M., & Lin Lee, C. (2020). Smart digital marketing capabilities for sustainable property development: A case of Malaysia. Sustainability, 12(13), 5402.
7. Stroebel, J., & Vavra, J. (2019). House prices, local demand, and retail prices. Journal of Political Economy, 127(3), 1391-1436.

8. Vuković, M. (2024). Generational differences in behavioral factors affecting real estate purchase intention. Property Management, 42(1), 86-104.
9. Wang, H. (2021, July). Research on Precision Marketing Strategies of Real Estate Companies Based on Big Data. In 2021 International Conference on Education, Information Management and Service Science (EIMSS) (pp. 145-148). IEEE.
10. Huang, R., & Mao, S. (2022). Research on precision marketing of real estate market based on data mining. Scientific Programming, 2022(1), 8198568.
11. Keleş, A. E., & Arıkan, Y. C. (2023). Analyzing Customers' Demands for Different Housing Features in Buildings Using a Data Mining Method. Buildings, 13(2), 555.
12. Landvoigt, T. (2017). Housing demand during the boom: The role of expectations and credit constraints. The Review of Financial Studies, 30(6), 1865-1902.
13. BuHamdan, S., Alwisy, A., & Bouferguene, A. (2021). Drivers of housing purchasing decisions: a data-driven analysis. International Journal of Housing Markets and Analysis, 14(1), 97-123.
14. Zhang, Y., Yuan, J., Li, L., & Cheng, H. (2019). Proposing a value field model for predicting homebuyers' purchasing behavior of green residential buildings: A case study in China. Sustainability, 11(23), 6877.
15. Kabir, S., Jamal, Z. B., & Kairy, B. P. (2024). How much to invest for house purchase? The consumer purchase intention perspective of real estate investment decision. International Journal of Housing Markets and Analysis, 17(4), 881-903.
16. Judge, M., Warren-Myers, G., & Paladino, A. (2019). Using the theory of planned behaviour to predict intentions to purchase sustainable housing. Journal of cleaner production, 215, 259-267.
17. Zhao, S., & Chen, L. (2021). Exploring residents' purchase intention of green housings in China: an extended perspective of perceived value. International Journal of Environmental Research and Public Health, 18(8), 4074.
18. Subagya, Y. H. (2021). The effect of price variables, location variables, and promotion variables on consumer decisions to purchase housing. International Journal of Seocology, 065-070.
19. Zhengya Guo. (2024). Accurate prediction of purchasing behaviour of cross border e-commerce consumers under social media marketing. International Journal of Web Based Communities,20(3-4),340-354.
20. Hamed GhorbanTanhaei, Payam Boozary,Sogand Sheykhan, Maryam Rabiee, Farzam Rahmani & Iman Hosseini. (2024). Predictive analytics in customer behavior: Anticipating trends and preferences. Results in Control and Optimization,17,100462-100462.
21. Song Peiyi & Liu Yutong. (2020). An XGBoost Algorithm for Predicting Purchasing Behaviour on E-Commerce Platforms. Tehnički vjesnik,27(5),1467-1471.
22. Zhihui Hu, Ailong Fan, Wengang Mao, Yaqing Shu, Yifu Wang, Minjie Xia... & Bin Li. (2025). Ship energy consumption prediction: Multi-model fusion methods and multi-dimensional performance evaluation. Ocean Engineering,322,120538-120538.