Saint Petersburg State University

Graduate School of Management

Master of Business Analytics and Big Data

**Consumer Repurchase Behavior Prediction Based on Different Fusion Models**

Master’s Thesis by the 2nd year students

Concentration — ВМ.5783.2019

Master in Business Analytics and Big Data

Du Shaohui

Research Advisor:

Dmitry Kudryavtsev

Associate Professor, Information Technologies in Management Department

Saint Petersburg 2021

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Я, Ду Шаохуэй, студент 2 курса магистратуры ВМ.5783.2019 «Бизнес-аналитика и большие данные (Master in Business Analytics and Big Data – MiBA)» , подтверждаем, что в моем магистерской диссертации на тему «Прогнозирование повторных продаж с помощью диффузных моделей», представленной в службу обеспечения программ магистратуры для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее курсовых и выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Нам известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

Ду Шаохуэй

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF THE MASTER THESIS

I, Du Shaohui, 2nd year master student of the ВМ.5783.2019 master program “Master in Business Analytics and Big Data - MiBA”,, state that my master thesis on the topic “Consumer Repurchase Behavior Prediction Based on Different Fusion Models” which is presented to the Master Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from master theses, PhD and doctorate theses which were defended earlier, have appropriate references.

I am aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St. Petersburg University «А master thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Professional Education Saint-Petersburg State University «a student can be expelled from St. Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».

Du Shaaohui

АННОТАЦИЯ

|  |  |
| --- | --- |
| Авторы | Ду Шаохуэй |
| Название магистерской диссертации | Прогнозирование повторных продаж с помощью диффузных моделей |
| Факультет | Высшая школа менеджмента |
| Направление подготовки | Бизнес-аналитика и большие данные |
| Год | 2021 |
| Научный руководитель | Кудрявцев Дмитрий Вячеславович |
| Описание цели, задач и основных результатов | Цель этого исследования - улучшить способность моделей машинного обучения предсказывать поведение потребителей при повторной покупке за счет слияния моделей.  Задачи исследования:  1. Обработка данных и разработка функций.  2. Моделирование и сравнительная оценка моделей.  3. Применение и улучшение модели слияния.  Основные результаты включают систематическое сравнение широко используемых моделей машинного обучения и методов объединения моделей в отрасли, а также усовершенствование метода объединения моделей с накоплением. |
| Ключевые слова | Прогнозирование выкупа, слияние моделей, классификация, классификационная модель машинного обучения. |

ABSTRACT

|  |  |
| --- | --- |
| Master Student Names | Du Shaohui |
| Master Thesis Title | Consumer Repurchase Behavior Prediction Based on Different Fusion Models |
| Title Faculty | Graduate School of Management |
| Main field of study | Business Analytics and Big Data |
| Year | 2021 |
| Academic Advisor’s Name | Dmitry Kudryavtsev |
| Description of the goals, tasks, and main results | The goal of this study is to improve the predictive ability of machine learning models on consumer repurchase behavior through model fusion..  The research tasks are:   1. Data processing and feature engineering. 2. Modeling and model comparison evaluation. 3. Model fusion application and improvement.   The main results incluse systematically compared the widely used machine learning models and model fusion methods in the industry, and improved the stacking model fusion method. |
| Keywords | Repurchase prediction, model fusion, classification, machine learning classification model. |

**CONTENTS**

[1.Introduction 7](#_Toc73501846)

[*1.1. Background* 7](#_Toc73501847)

[*1.2. Research motivation* 8](#_Toc73501848)

[2. Theoretical background 10](#_Toc73501849)

[*2.1. Research Status of Influencing Factors of User Repurchase Behavior* 10](#_Toc73501850)

[*2.2. Research Status of Predicting Repurchase Behavior* 14](#_Toc73501851)

[*2.3. Research Gap* 18](#_Toc73501852)

[3. Data preprocessing 18](#_Toc73501853)

[*3.1. Data Sources* 18](#_Toc73501854)

[*3.2. Data description* 19](#_Toc73501855)

[*3.3. Data Integration* 21](#_Toc73501856)

[*3.4. Missing value processing* 22](#_Toc73501857)

[4. Feature engineering 22](#_Toc73501858)

[*4.1. Extract statistical features* 22](#_Toc73501859)

[*4.2. Data vectorization* 24](#_Toc73501860)

[5.Common machine learning classification models 27](#_Toc73501861)

[*5.1 Logistic regression* 27](#_Toc73501862)

[*5.2. Bayesian Classification* 29](#_Toc73501863)

[*5.3. KNN* 32](#_Toc73501864)

[*5.4. SVM* 33](#_Toc73501865)

[*5.5. CART* 35](#_Toc73501866)

[5.6. *Random forest* 37](#_Toc73501867)

[*5.7. Extreme random tree* 38](#_Toc73501868)

[*5.8. GBDT* 39](#_Toc73501869)

[*5.9. AdaBoost* 40](#_Toc73501870)

[*5.10. XGBOOST* 42](#_Toc73501871)

[*5.11. LightGBM* 45](#_Toc73501872)

[6.Model fusion method 48](#_Toc73501873)

[*6.1. Soft-Voting* 48](#_Toc73501874)

[*6.2. Hard-Voting* 48](#_Toc73501875)

[*6.3. Stacking* 49](#_Toc73501876)

[*6.4. Improvements to the stacking model: soft-stacking model* 50](#_Toc73501877)

[7.Model comparison and evaluation 50](#_Toc73501878)

[*7.1. Sampling method* 50](#_Toc73501879)

[*7.2. Model evaluation method* 51](#_Toc73501880)

[*7.3. Feature optimization based on algorithm model* 52](#_Toc73501881)

[*7.4. Resualts* 54](#_Toc73501882)

[Conclusion 56](#_Toc73501883)

[*Theoretical contribution:* 56](#_Toc73501884)

[*Practical contribution:* 56](#_Toc73501885)

[Reference 58](#_Toc73501886)

[*Academic Literature in English* 58](#_Toc73501887)

[*Academic Literature in Other Languages* 60](#_Toc73501888)

# 1.Introduction

## 1.1. Background

With the development of e-commerce in China, major local e-commerce platforms such as Tmall, JD.com, and Vipshop have emerged one after another, and more and more people are participating in online shopping. Compared with traditional shopping models, online shopping has great advantages such as convenience, real-time, and no space restrictions. However, while online shopping brings convenience to people, it also causes problems such as information overload. On the one hand, users are struggling to find their favorite products in the face of a large amount of product information. On the other hand, businesses often get lost in user groups and cannot achieve accurate marketing information delivery.

In order to help users and merchants establish a wide range of contacts, and promote e-commerce platforms to reach more transactions, each e-commerce platform has adopted intelligent systems to analyze users' needs and preferences and recommend products for them. Consumers’ repurchase behavior prediction is to analyze the their past historical behavior data to analyze the rules of their purchase behavior, and then predict the their future repeat purchase behavior. This technology can be applied to the recommendation system of e-commerce platforms to help merchants identify users with repeated purchase intentions, so that merchants can take more targeted marketing measures, improve user experience, and enable users and merchants to establish a lasting and reliable Relationship, prompting users to buy more products.

## 1.2. Research motivation

Consumer repurchase behavior is an important measure of their loyalty, and it has always been a research hotspot in the field of marketing. Before the popularization of online shopping, limited by data, the research hotspots of repeat purchase behavior generally focused on its influencing factors, trying to estimate which type of user group would be most likely to produce repeat purchase behavior through the research of influencing factors.

With the development of big data technology and the continuous growth of e-commerce platforms, personal information such as consumer interests and hobbies, and daily shopping behavior information are accumulated in the databases of major e-commerce platforms, gradually forming massive amounts of data. It has been discovered that through the mining of online shopping behavior big data, users' repeated purchase behavior can be predicted in advance, and it can even be specifically predicted to which merchants each user has repeated purchase intentions. Therefore, the research on repurchase behavior has a new direction. Some studies have begun to pay attention to the technology related to the prediction of user repeat purchase behavior, and gradually apply relevant algorithms in the field of machine learning to the prediction problem of repurchase behavior.

The research of consumer repurchase prediction problem can be applied to the intelligent recommendation system of e-commerce platform to help merchants identify users with repeated purchase intentions. Obviously, a high-precision forecasting system can help more merchants and users establish consumer relationships, thereby bringing huge profits to the e-commerce platform. Therefore, the prediction accuracy and generalization performance of related models used to predict repeat purchase behavior are particularly important.

However, when the existing machine learning model is applied to repeat purchase behavior prediction, its prediction accuracy is low, and the prediction effect is only slightly better than random guessing. Such an effect is difficult to produce valuable effects when applied to a recommendation system. Through combing the literature on the influencing factors of repeat purchase behavior, it is found that consumers' purchase behavior is controlled by a series of complicated subjective factors, including perceived value, satisfaction, trust, etc. . These influencing factors make the repeat purchase behavior of users diversified and different. When these complex and diverse behavior patterns are hidden in the data, it will bring great difficulties to the fitting of the machine learning model. The difference in user behavior rules is an important reason that restricts the accuracy of prediction, and none of the existing machine learning models can solve this problem well.

1.3. Research goal

Improve the predictive ability of machine learning models on consumer repurchase behavior through model fusion.

1.4. Research questio***n***

1. How to mine user data characteristics in multiple dimensions.

2. How to evaluate the prefetching ability of the model fusion method.

3. How to improve the model fusion method.

# 2. Theoretical background

## **2.1. Research Status of Influencing Factors of User Repurchase Behavior**

Many theories of factors affecting user repurchase behavior are improved from research on user purchasing behavior, usually involving economics, psychology, management, behavior, and sociology. User repurchase behavior includes traditional offline repurchase behavior and current online repurchase behavior. The difference is that e-commerce users may change their repurchase intentions due to psychological or external factors that the merchant cannot detect.

(1) Perceived service quality theory: originated from the perceived risk theory in the study of factors affecting user purchase behavior. Perceived risk theory was proposed by Professor Bauer of Harvard University in 1960. This theory believes that when consumers perceive when making purchase decisions The risk of behavior, such as the price trend of commodities, the quality assurance of commodities, and other uncertain consequences that may bring losses. Afterwards, the theory of perceived risk is further subdivided into six risk theories: time, body, finance, function, society, and psychology[[1]](#footnote-1). Li Rixu (2017) found that perceived risk has a negative impact on consumers' purchasing intention and trust in online purchase of fresh agricultural products [[2]](#footnote-2). Consumers’ buying behavior occurs in a state of risk-free or low-risk under their perception, so the factors that affect consumers’ repeated buying behavior turn to the perception of service quality theory, such as the pre-sales and after-sales experience of the product, and the experience of using the product Factors in the motivation of repurchase behavior. Huang Qian (2011) found that the level of service quality will change the impression of products in consumers' minds, and the perceived service quality positively affects users' repurchase behavior in the study of repeated purchase behavior of economy hotel customers[[3]](#footnote-3). Xiong Xiaoyuan (2014) believes that different interactions between consumers and websites will affect perception variables and thus affect users' repurchase behavior. Consumer satisfaction is a direct variable that affects users' online repurchase behavior[[4]](#footnote-4).

(2) Perceived value theory: Perceived value theory has always been an important factor influencing consumers' purchase behavior, and it also affects consumers' repurchase behavior. Zhang Yanjing (2014) constructed a five-dimensional model of online festival promotion in a specific context based on quality value, price value, service value, emotional value and social value, and established a relationship model for user perceived value, user satisfaction, and user purchase behavior which were verified[[5]](#footnote-5) . Zhang Chuntao (2016) took the consumer group of college students as an example, divided the perceived value into 8 dimensions, and the research results showed that the information value, emotional value, etc. positively affect the user's purchasing behavior[[6]](#footnote-6). Zhao Dandan (2014) demonstrated through empirical research that the three consumer online interaction modes of B2C interaction, C2C interaction and content interaction have a positive impact on consumers' internal and external perceived value and willingness to repurchase[[7]](#footnote-7) .

En-Chi Chang, Ya-Fen Tseng (2013) divide consumer perceived value into utilitarian and hedonic values. The former focuses on logical reasoning, which is the overall assessment of consumers’ functional benefits and losses of purchasing behavior; the latter focuses on the emotional response of consumers ,which is the overall evaluation of consumers' online shopping experience or emotion[[8]](#footnote-8). At the same time, they also found that the image of online shopping malls influences consumers' repurchase willingness through perceived value, in which functional value has a greater impact than hedonic value[[9]](#footnote-9). TG Wang (2014) et al. used Yahoo Qimo’s user transaction behavior data as the research object, and proposed that the user’s approval value is a factor that positively affects their repeat purchase behavior, and the user’s risk judgment will have a negative impact on the repeat purchase intention. And they try to establish an MEC model by linking the value and behavior of users on the e-commerce platform. The MEC model theory believes that the user's purchasing behavior is driven by values, so the perceived value will eventually change the user's choice and judgment[[10]](#footnote-10).

(3) Perceived transaction cost theory: Perceived transaction cost theory is based on the theory of transaction cost economics. Transaction Cost Economics (TCE) theoretically explains why transaction entities tend to engage in specific forms of transactions. It is not other transaction subjects, and its main influencing factors are transaction frequency, asset specificity and uncertainty[[11]](#footnote-11).

Mukherjee A, Banerjee S, BandyopadhyayS (2012) constructed a simulation model to study the impact of social networks on consumers’ online shopping behavior. The purpose is to examine the impact of social networks on consumers’ personal perception of transaction costs, which determines consumers’ online The tendency of buying behavior[[12]](#footnote-12).Dragan Benazić (2015) conducted an online survey based on the framework of how consumers' perceived risks and costs establish trust relationships with websites, and found that perceived costs affect consumers' willingness to repurchase to a certain extent[[13]](#footnote-13).Myung Ja K, Choong-Ki L, Namho C and others (2014) found that products with transaction cost advantages, under the regulation of consumer emotional loyalty (EL), consumer satisfaction will also increase , Thereby affecting consumers' intention to repeat purchases[[14]](#footnote-14).Lei-Yu Wu et al. (2012) studied the impact of transaction costs on consumers’ repurchase intentions from the perspective of online shoppers. They conceptualized transaction costs into information search costs, moral hazard costs, and specific asset investment costs. The results show that consumers each transaction cost is positively correlated with its repurchase intention, among which the cost of information search has the most significant impact [[15]](#footnote-15).

(4) Consumer's subjective decision-making: Consumer's subjective decision-making means that consumers make behavioral decisions on the price, quantity, and method of repurchasing commodities based on subjective conditions such as life experience, preference needs, and psychological state.

The user's purchasing behavior decision generally goes through five stages: needs identification, information search, selection evaluation, purchase decision, and post-purchase evaluation[[16]](#footnote-16) . Hong Shuo (2018) constructed a research model with purchase decision as the dependent variable on the basis of technical model, and verified through empirical research that consumers' actual purchase decision will be affected by subjective purchase intention[[17]](#footnote-17) .

Luo Sheng (2018) constructed a theoretical model of consumer impulsive purchase of cosmetics based on the S-O-R viewpoint of environmental psychology. Consumers will significantly influence consumers' purchasing behavior decisions under marketing stimulus scenarios such as product promotion[[18]](#footnote-18) .

Xu Xuefeng (2012) constructed a structural equation model based on the summarized cognitive variables, attitude variables, and decision-making behavior variables. The results verified that the three variables have a significant impact on consumer purchasing decisions and merchant trust. Among them, consumer attitude factors Directly affect decision-making behavior variables[[19]](#footnote-19) .

Y Zhang and YQ Feng (2011) divide consumer demand identification into INR (Initial Demand Identification) affected by traditional shopping and FNR (Final Demand Identification) affected by online shopping. They believe that the offline experience and online shopping experience may affect consumers’ purchasing decisions[[20]](#footnote-20) .

## **2.2. Research Status of Predicting Repurchase Behavior**

Through a large amount of literature review, it is found that in the e-commerce environment, there are mainly user purchase behavior prediction, click behavior prediction, and user activity prediction, etc., and the user's repurchase behavior is essentially the user's purchase behavior. Traditional buying behavior prediction models mainly include SMC model, RFM model and BG/NBD model. Most researchers improve on the basis to achieve better prediction results. Among them, SMC and BG/NBD models both verify model assumptions to make the results obey a certain function distribution, such as Gamma distribution, exponential distribution and so on. The RFM model is extended from the marketing field to predict online purchase behavior, and there are relatively a few applied researches in the e-commerce field. At present, with the rise of big data and artificial intelligence, a few scholars have tried to use the combination of big data technology and machine learning algorithms, machine learning systems and neural networks to build models, aiming to improve the prediction accuracy and robustness of the model .

**2.2.1 Related research on SMC and BG/NBD model**

Zhang Chunlian (2006) analyzed the traditional customer buying behavior prediction model, improved the SMC model based on the BG/NBD model, and verified its effectiveness on a medical product purchase behavior data set[[21]](#footnote-21) .

Chen Jie (2011) and others believe that it is more difficult to predict online consumer behavior in online consumer behavior prediction by purchasing rate indicators. Therefore, they built an online consumer purchase rate prediction model based on the BG/NBD model. The user's consumption behavior data in the first 26 weeks of a shopping mall is used to predict the user's purchase rate in the next 27 weeks[[22]](#footnote-22).

Y Tian (2015) et al. built a prediction model that uses user transaction frequency and time based on the significant differences in C2C e-commerce online and offline purchases and sales, and verified that its accuracy is better than traditional transaction data in real transaction data. The Pareto/NBD model provides a simple and powerful tool for predicting repeat purchase behavior of users under C2C e-commerce[[23]](#footnote-23).

Shu Fang (2015) et al. proposed a HIPP model, and used genetic algorithms to combine the model with the traditional SMC model to optimize the weights. Through empirical analysis, she verified the effectiveness and superiority of this combined forecasting method[[24]](#footnote-24).

Li Meiqi and Qi Jiayin (2016) used the user data of Dianping.com as the research object, constructed a Pareto/NBD prediction model, and improved the prediction effect of the model on user purchase behavior by introducing covariates[[25]](#footnote-25).

**2.2.2 Related research on RFM model**

HJ Chang (2007) et al. proposed a predictive model for the purchase behavior of potential customers through cluster analysis and association rule analysis. Cluster analysis collects personal information data of loyal customers to locate potential customer attributes, and association rule analysis extracts loyal customer purchases. Behavioral characteristics are used to detect customer interest in hot-selling products[[26]](#footnote-26) .

YS Cho (2013) et al. proposed a new volume-weighted mining method based on the most recent consumption (Recency), consumption frequency (Frequency) and consumption amount (Monetary), that is, the RFM model, which is used for consumer purchase behavior. To predict, and to verify its effectiveness, collect the consumer behavior data of cosmetics users in online shopping malls to conduct experiments[[27]](#footnote-27).

**2.2.3 Related research on machine learning models**

In the field of e-commerce marketing, D Cui (2005) et al. conducted comparative experiments on the prediction accuracy of support vector machine (SVM) prediction models and logistic regression models, evaluated the advantages and disadvantages of their models, and studied the prediction methods and standards of SVM models. Differences in parametric model methods[[28]](#footnote-28).

R Gupta (2014) uses machine learning algorithms to enhance customers' correct price purchases on e-commerce platforms, and predicts consumers' purchase decisions based on adaptive or dynamic pricing of goods[[29]](#footnote-29).

Zuo Y (2014) and others used the SVM model to process the RFID (radio frequency identification) data generated by consumers in the supermarket. The study found that the SVM model significantly improved the prediction accuracy of consumer purchase behavior compared with linear regression and other predictive models[[30]](#footnote-30).

Du Gang, Huang Zhenyu (2015) based on the decision tree to improve the algorithm to build a new user purchase behavior prediction model, and use the Teradata platform to achieve the decision tree model construction in the big data environment, from the local optimal to the global optimal[[31]](#footnote-31).

Silahtaroglu G (2015) and others extract information from the collected consumer clickstream data, and use decision trees and multi-layer neural networks to build consumer buying behavior prediction models. This model is effective in predicting whether consumers will buy goods[[32]](#footnote-32).

Zhu Xin (2017) et al. took the prediction of customer buying behavior in the context of online shopping as the research object, and used big data analysis methods—machine learning algorithms to build a predictive model from real consumer online shopping behavior data. By trying a single algorithm and a fusion algorithm to build the model, it is proved that the prediction result of the fusion model is better than the single model under certain circumstances[[33]](#footnote-33).

Zhou Chengji (2018) takes the interaction behavior data of consumers and commodities in a certain period of time as the analysis object, proposes a data preprocessing method based on time series and a feature selection method based on SSP algorithm, and constructs Xgboost based on the Bagging model fusion strategy The mixed model improves the prediction accuracy of the model[[34]](#footnote-34).

## **2.3. Research Gap**

(1) The traditional SMC, BG/NBD and other single models are widely used as prediction models. This type of model is to verify the model assumptions to make the results obey a certain function distribution, such as Gamma distribution, exponential distribution and so on. When the amount of data is large and the data cannot meet the assumptions, the predictive ability is poor.

(2) Most of the research on machine learning algorithm models is limited to a certain algorithm or a certain model fusion method, and there is a lack of comparison between models.

(3) Most researches involving model fusion methods lack innovation in model fusion methods.

# 3. Data preprocessing

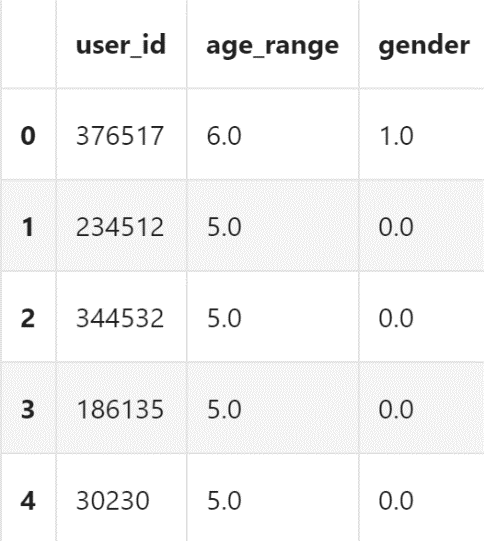
## **3.1. Data Sources**

The data in this thesis comes from the Tianchi big data competition held by Alibaba. This data set is provided by the Tmall platform and contains the shopping logs of anonymous users on the day of "Double Eleven" and the past 6 months before it. A total of 3986 merchants and their corresponding 260,000 new buyers (these new buyers have all bought the products of the corresponding merchant on the day of "Double Eleven"). The task of this article is to predict whether these new buyers will buy goods from the same merchant again in the next 6 months. Due to privacy issues, data is sampled in a partial sampling manner, so the statistical results of this data set will deviate from the actual situation of Tmall.com. But it will not affect the applicability of the solution.

## **3.2. Data description**

The data set contains three tables: user personal information table, user behavior log table and training set table.

1.The user personal information table contains the basic information of new ustomers of the business during the "Double 11" period. The information of some customers is incomplete, and the age range or gender is empty. The following table shows the first 5 rows of data in the user's personal information table.



The description of each column of the user personal information table is as follows:

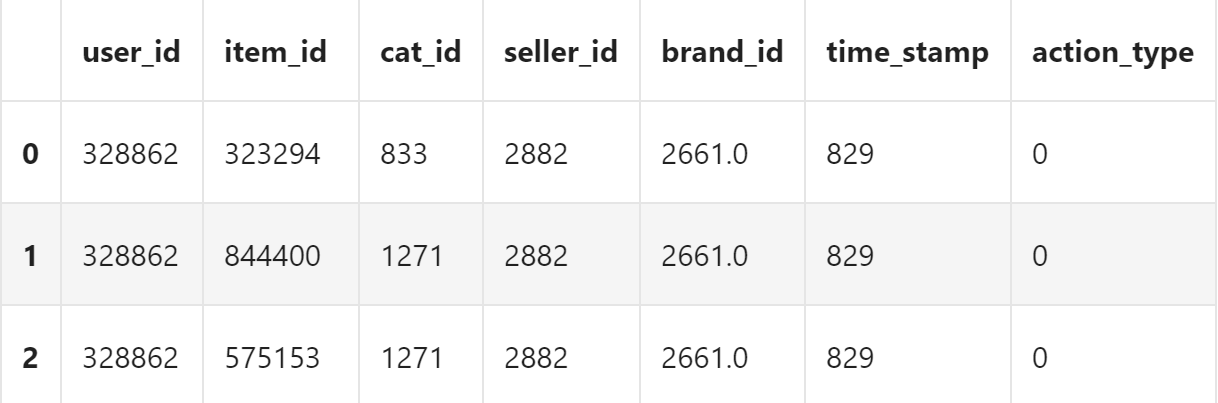
**user\_id:**  Unique user ID

**Age\_range:**  1:<18; 2: [18,24]; 3: [25,29];4: [30,34]; 5: [35,39];6: [40,49];

7and8:>=50; 0 and null: unknown age.

**Gender：** 0: female; 1: male

2. The user behavior log table contains the shopping logs of new customers during the first 6 months of the "Double 11" period, and the time span is from May 12 to November 11. The first 3 rows of data in the user behavior log table are as follows:



The description of each column of the user behavior log table is as follows:

**user\_id:**  unique user ID

**merchant\_id:** merchant unique identifier

**Item\_id:** Item ID.

**Cat\_id:** Commodity category ID.

**Seller\_id:** The ID of the online store where the behavior occurred.

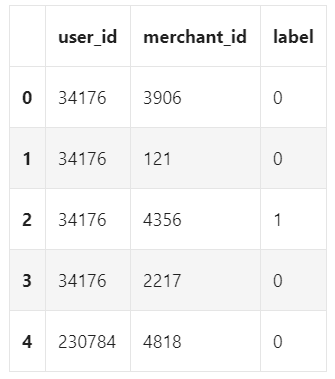
**Brand\_id:** product brand ID.

**Time\_stamp:** time when the behavior occurred.

**Action\_type:** user behavior: 0: Click. 1: Add to shopping cart. 2: Purchase.

3: Collection.

3. The training set table includes the identification of the user, the merchant, and whether the user purchases again at the merchant. The first five rows of data are as follows:



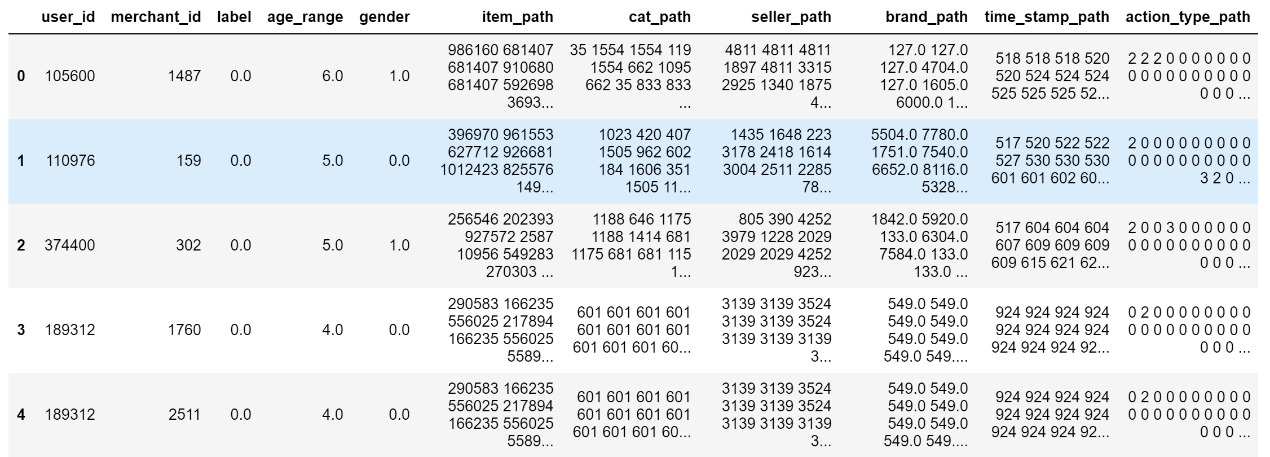
The description of each column of the training set table is as follows:

**user\_id:** unique user ID

**merchant\_id:** merchant unique identifier

**label：** 0: No repurchase; 1: Repurchase

## **3.3. Data Integration**

In the above three data tables, two fields, user ID and merchant ID, appear. In order to improve the accuracy and speed of the subsequent feature mining process, it is necessary to merge the user's personal information, behavior log information, and purchase behavior information in the training table. And integrate the behavior of the same user. The integrated table is called the user purchase behavior table. The first five rows of data are as follows:

The new features after integration are described as follows:

**Item\_path:** Collection of item IDs

**Cat\_path:** Collection of item categories IDs

**Seller\_path:** Collection of seller IDs

**Brand\_path:**  Collection of brand IDs

**Time\_stamp\_path:** Collection of time that the action occurred

**Action\_type\_path:** Collection of actions

## **3.4. Missing value processing**

Missing values will affect the integrity of the data. If there are too many missing values for an attribute, the attribute should be eliminated. If the missing value is relatively small, you can use an appropriate missing value processing method to complete the data.

Based on statistics of the data in the three data tables, it is found that there are missing values in the age range and gender fields of the user. However, the proportion of missing values is relatively small, the proportion of missing values in the age range is 0.56%, and the proportion of gender missing values is 0.43%. Therefore, the missing value filling method can be used to complete the data. Since the data of these two characteristics are non-continuous data, this paper uses the largest possible number to fill. The gender and age range with the largest number of users are counted, and the corresponding gender and age range values are filled into the missing value data.

# 4. Feature engineering

## **4.1. Extract statistical features**

**4.1.1 User-based feature**

The user's own demand preferences and the law of purchase behavior are the most important factors affecting repeat purchase behavior. Therefore, this article conducts a comprehensive analysis of the user's characteristics from the following different perspectives.

(1) The degree of diversity of products purchased by users

When users participate in online shopping, the types of behaviors they produce can be divided into clicking, buying, adding to favorites, and so on. Therefore, the four types of behaviors can be used to count how many products each user has paid attention to. Such feature are mainly used to measure the degree of diversity of users' purchase of goods. The higher the degree of diversity, the more the user likes to buy different products, so the probability of repeating purchases of the same merchant's products may be lower. In this paper, a total of 12 features are selected to measure the degree of diversity, as shown in the following:

|  |  |
| --- | --- |
| Feature Name | Feature description |
| f1 | How many different products have users clicked |
| f2 | The user has clicked on several categories of products |
| f3 | How many different businesses have the user clicked |
| f4 | Users have clicked on products of several brands |
| f5 | How many different products have users bought |
| f6 | The user has purchased several categories of goods |
| f7 | How many products from different businesses have been purchased by the user |
| f8 | The user has purchased several brands of goods |
| f9 | How many products have been added to favorites by users |
| f10 | How many categories of products the user has added to the collection |
| f11 | How many different businesses have been added to favorites by users |
| f12 | How many brands of products the user has added to the collection |

(2) User activity level of online shopping

By counting the number of occurrences of various behaviors of users, it is possible to discover the active degree of users participating in online shopping. Generally speaking, users who frequently participate in online shopping may have a higher probability of repeated purchases than users who do not frequently participate in online shopping. The features selected in this article to indicate the user's online shopping activity are as follows.

|  |  |
| --- | --- |
| Feature Name | Feature description |
| f13 | The total number of clicks by the user |
| f14 | The total number of purchases by the user |
| f15 | User's total number of favorites |
| f16 | The total number of user behaviors (including clicks, purchases, favorites, and adding shopping carts) |

**4.1.2 Features based on the relationship between users and businesses**

The relationship between the user and the corresponding business can be described by a series of features. This article selects the attention degree to reflect the degree of the relationship between the user and the corresponding business. The user's attention behavior to the corresponding business refers to any behavior that the user has made at the business, such as buying, clicking, or adding to favorites. The higher the degree of attention, the more attractive the merchandise of the merchant is to the corresponding users, and the greater the possibility that users will repeat purchases. The five characteristics that reflect the user's attention to the business are as follows.

|  |  |
| --- | --- |
| Feature Name | Feature description |
| f17 | The total number of clicks by the user |
| f18 | The total number of purchases by the user |
| f19 | User's total number of favorites |
| f20 | The total number of user behaviors (including clicks, purchases, favorites, and adding shopping carts) |
| f21 | How many different products did the user click on in the corresponding merchant’s store |

## **4.2. Data vectorization**

Vectorize path data with unequal dimensions (processing Item\_path, Cat\_path, Seller\_path, Brand\_path, Time\_stamp\_path).

The five features are all sets of multiple numbers. Different users have different numbers of elements in their collections. Therefore, if you want to apply them to the algorithm model, you must convert them into vectors with consistent dimensions.

This problem is essentially the same as generating word vectors in natural language processing. Therefore, both the TF-IDF algorithm and the CBOW model can be used to extract features.

**TF-IDF**

TF-IDF for a word in a document is calculated by multiplying two different metrics:

The **term frequency (TF)** of a word in a document. There are several ways of calculating this frequency, with the simplest being a raw count of instances a word appears in a document. Then, there are ways to adjust the frequency, by length of a document, or by the raw frequency of the most frequent word in a document.

The **inverse document frequency (IDF)** of the word across a set of documents. This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

So, if the word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1.

Multiplying these two numbers results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document.

To put it in more formal mathematical terms, the TF-IDF score for the word t in the document d from the document set D is calculated as follows:

Where:

**Continuous Bag of Words Model (CBOW) Model**

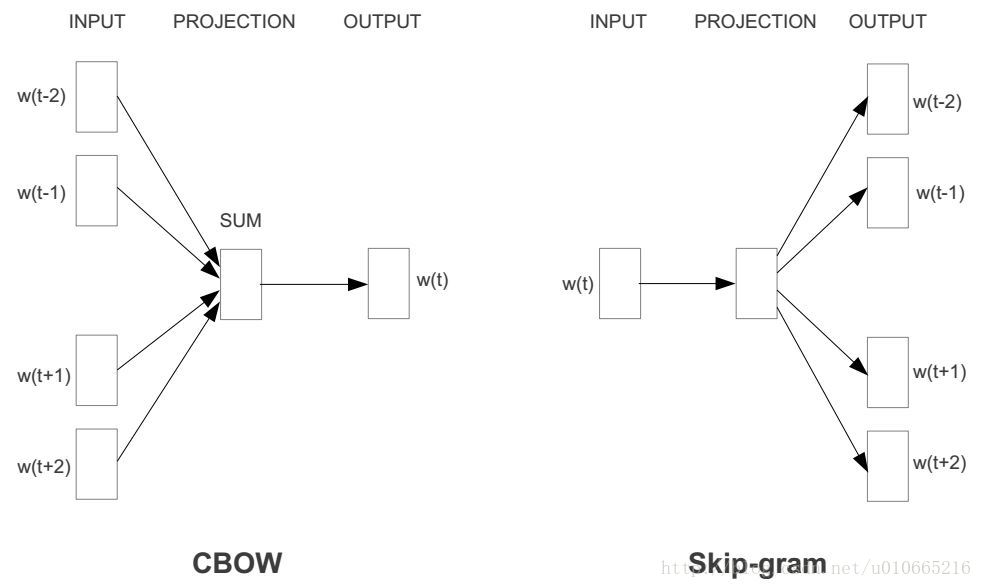
In the CBOW model, the distributed representations of context (or surrounding words) are combined to predict the word in the middle.

In the CBOW model, the distributed representations of context (or surrounding words) are combined to predict the word in the middle.

1) Input layer: one-hot codes of context words. Assume that the word vector space dim is V and the number of context words is C.

2) All one-hot codes are respectively multiplied by the shared input weight matrix W.

3) Define the loss function (usually the cross-entropy cost function), and use the gradient descent algorithm to update W and W'. After training, the vector obtained by multiplying each word in the input layer with the matrix W is the word vector.



TF-IDF is suitable for columns with a small number of different elements, so I use it to vectorize Seller\_path and Cat\_path.

The advantage of the CBOW Model is that we can customize the dimensions of the vector to suit columns with large different elements. Therefore, I vectorized Item\_path, Brand\_path, and Time\_stamp\_path.

# 5.Common machine learning classification models

## **5.1 Logistic regression**

Logistic regression assumes that the data obeys the Bernoulli distribution, and uses the method of maximizing the likelihood function and using gradient descent to solve the parameters to achieve the goal of dicing the data.

Logistic regression is a generalized linear regression model, which has many similarities with multiple linear regression analysis. Their model forms are basically the same. Regarding the characteristic data , the models all contain, where and are the parameters to be solved. The difference is that the dependent variable y is different. Multiple linear regression directly uses as the dependent variable y, that is, , while Logistic regression uses the function L to map Into a hidden state p, that is,, and then determine the value of the dependent variable y according to the size of p and 1-p. If L is a logistic function, it is logistic regression, and if L is a polynomial function, it is polynomial regression.

The binomial logistic regression model is a classification model, represented by the conditional probability distribution P(Y|X), in the form of a parameterized Logistic distribution. The value of the random variable X is a real number, and the value of the random variable Y is 1 or 0. The conditional probability distribution form of the binomial Logistic regression model is as follows:

Among them, is the input, Y∈{0.,1} is the output, and are parameters, w is the weight vector, b is called the bias, and is The inner product of w and x.

For a given input instance x, P(Y=1|x) andP (Y=0|x). Logistic regression compares the size of the two conditional probability values, and it classifies the instance x into the category with the larger probability value.

The logistic regression model uses the maximum likelihood method to estimate the model parameters. When the sample size is large enough, the estimation is more accurate, and hypothesis testing and interval estimation can be performed.

Assuming that there are m mutually independent observation events , if the probability of occurrence of the event is p, then for or , can be expressed as:

The principle of maximum likelihood estimation is to write the joint probability distribution of the parameter sample to be estimated, and then maximize the log-likelihood function to solve the corresponding parameter estimation value. For this reason, for , the corresponding , the logistic regression likelihood function is:

Take the logarithm and reduce it to a likelihood function:

In order to maximize the likelihood function, the partial derivative( the gradient) can be set to 0, namely:

The above formula is a system of equations composed of d+1 nonlinear equations about , which cannot be solved analytically. To solve this problem, the Newton-Raphson iterative method is commonly used. The iterative formula is:

In the formula, is the Hessian matrix, and the estimated value of the Logistic regression parameter is obtained by the right iteration of the above formula.

## **5.2. Bayesian Classification**

**5.2.1 Bayesian method related theories**

The Bayesian classification method is based on Bayes' theorem and adopts probabilistic reasoning methods. The principle of Bayesian classification is to calculate the posterior probability of a given sample in each category, and then determine the sample as the category corresponding to the largest posterior probability. In the process of calculating the posterior probability, it is necessary to know the prior probability of each category in the data set and the conditional probability of the attribute. The prior probability of a category can be known in advance through statistical means, and the conditional probability of an attribute can also be estimated through statistical methods or assumed distribution models.

**5.2.2 Basics of Probability Theory**

In the sample space S, let A and B be two random events. Under the condition that event B occurs, the probability of event A is called the conditional probability of event B in the given situation of A, also called posterior probability, Denoted as P(B∣A). Correspondingly, P(A) is called unconditional probability or prior probability. The conditional probability can be calculated by the following formula:

From the multiplication theorem of probability, we can get:

Suppose there are n events, then:

**5.2.3 Total probability formula and Bayes' theorem**

In the sample space S, A is a random event, is a division of S, and , then the full probability formula can be obtained:

Bayes' theorem can be derived from the definition of conditional probability and the formula of total probability:

**5.2.4 Independence of events:**

Suppose A and B are two random events. Generally, the probability of occurrence of A and B has an influence on each other, that is, P(B∣A)≠P(A). Only when this influence does not exist will there be P (B∣A)=P(B), if at this time:

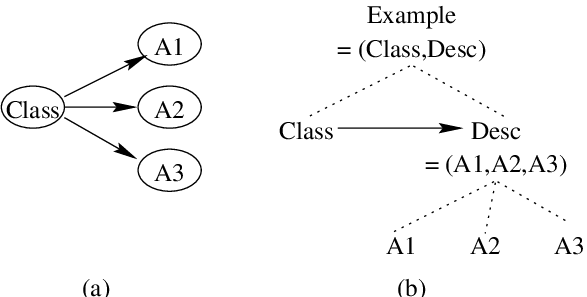
then A and B are independent events.

Similarly, for n events, if there are:

then are independent events.

**5.2.5 Naive Bayes Classifier**

Naive Bayes classifier is one of the most widely used models in Bayesian classifiers. The model description is shown in the figure below:



Suppose there is a variable set , where includes n conditional attributes, and includes m labels. The naive Bayes classification model assumes that all conditional attributes are used as leaf nodes of class variables, and a given sample to be classified assigned to class, if and only if:. According to Bayes' theorem, there are:

If the probability of a class in the data set is unknown in advance, it can be assumed that the probabilities of each class are equal. That is:

According to this, is maximized. Otherwise, maximize . Since P(X) is constant for all categories, there are:

The assumptions that the conditional attributes of the naive Bayes classification algorithm are mutually independent are as follows:

Where , is the number of instances of class in the training sample, and S is the total number of training samples. Then the formula expression of the naive Bayes model is:

Based on the statistical characteristics of the Tmall repurchase data set, it is assumed that they obey the Gaussian distribution, namely:

Where  is the Gaussian density function of the attribute , and and are the mean and standard deviation, respectively.

## **5.3. KNN**

Cover and Hart proposed the KNN (*k*-nearest neighbors) algorithm in 1968, which is a lazy learning method that judges the class of sample points according to the types of nearby points. The idea of this method is: Calculate the distance from a point of a known type in the feature space to a sample point, select the K smallest distance points, and count which type of points in the K points have the largest number, then the sample point is which category it belongs to. For binary data, we generally choose an odd number when choosing K, so that there will not be the same number of sample points in the two types, so as to avoid the defect of not being able to select the sample point category at a time[[35]](#footnote-35). The steps of the traditional KNN algorithm are as follows:

(1). Construct the training sample set and the test sample set, and calculate the distance between the samples in the test sample set and all the samples in the training sample set. Training sample set , test sample set . There are many ways to calculate the sample distance. Commonly used are Euclidean distance, Manhattan distance, angle cosine, Mahalanobis distance, etc. Here, the Euclidean distance is taken as an example, and the distance calculation formula is:

(2). Compare the distances from all training sample points to a certain test sample point, and select K training sample points with the smallest distance.

(3). Calculate the number of each category of the K sample points, find the category with the largest number, and classify the sample to be tested into this category.

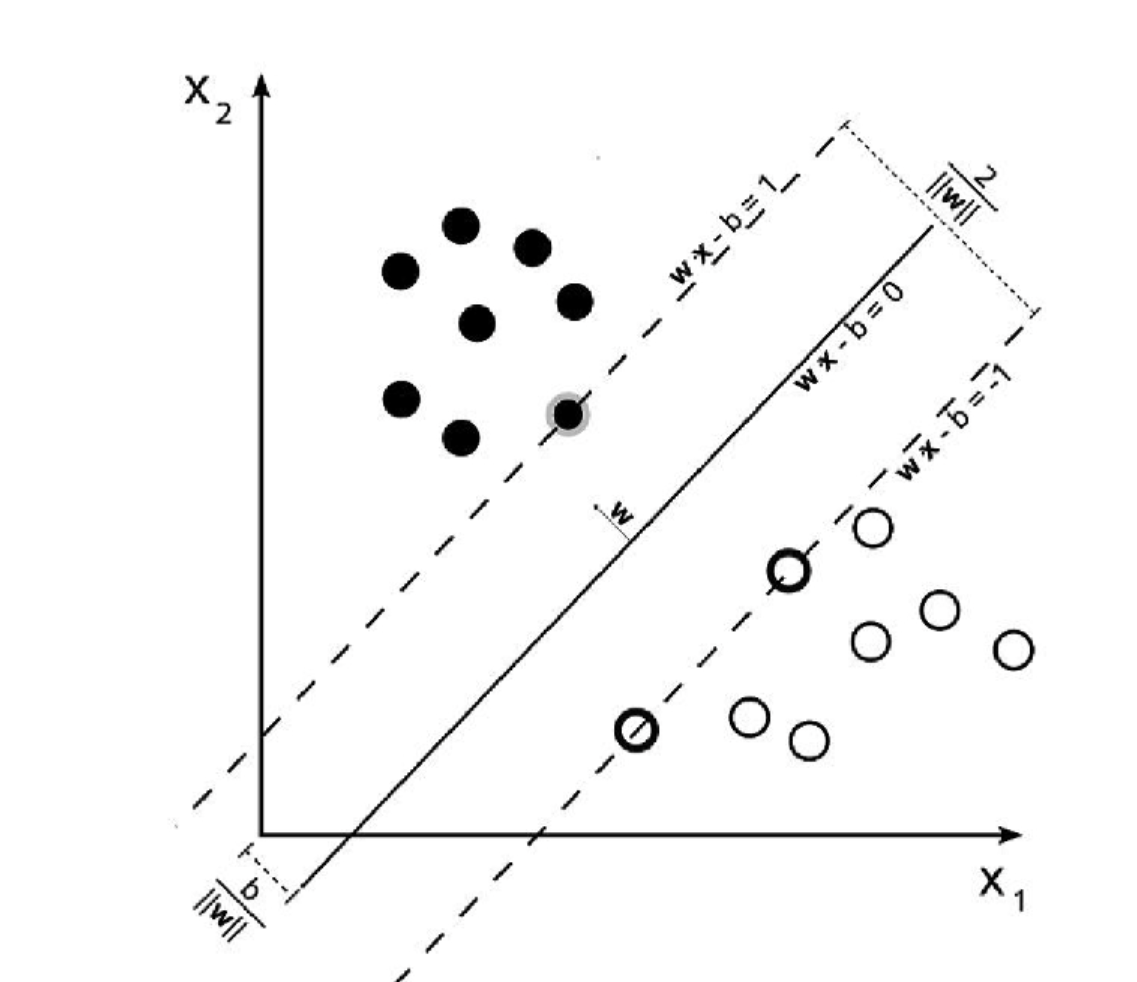
(4). Repeat this step until all test sample points have been calculated.

## **5.4. SVM**

Support Vector Machine (SVM) is a classic classification model in machine learning. The core idea is to find a hyperplane, segment the sample data based on the principle of maximizing the interval, and finally transform the objective function into a quadratic function. , The constraint condition is linear convex quadratic programming problem solving. The optimal classification plane should meet the following conditions:

Among them, represents the size of the classification interval of the classifier, and its size indicates the classification effect of this classification plane. The constraint condition s.t. indicates that the distance between the sample point and the classification plane is greater than or equal to 1.

As shown in the figure below, the black solid ball and the white hollow ball respectively represent the two types of samples that need to be classified. The sample point on the dotted line is the sample closest to the solid line. Based on the core idea of SVM, we need to determine the position of the solid line, that is, to ensure The interval between the sample and the optimal classification surface is maximized.



Converting the problem of solving the maximum value of d into the problem of solving the minimization of can obtain the equivalent formula:

It can be seen from the original constraints that the interval between all sample points in the SVM classification and the classification hyperplane is greater than or equal to 1, that is, hard interval classification. Sometimes, in order to avoid overfitting of the model or noise in the data, the SVM classifier is allowed to misclassify some sample points, that is, soft interval classification. We introduce a slack variable to change the original constraint condition into:

Due to the introduction of the slack variable, the interval between the sample point and the classification hyperplane may be less than 1, so as to achieve the effect of sample misclassification, so the original objective function becomes:

Among them, C>0 is the penalty factor in the objective function, which is the same as that in logistic regression, and its size indicates the degree of punishment of the classifier for incorrectly classified samples. At the same time, the dual problem of the original problem is obtained by introducing the Lagrangian function:

For the linearly inseparable low-dimensional feature data, by mapping it to the high-dimensional space, to achieve the purpose of linear separability, applying it to the SVM algorithm is to introduce the kernel function:

Therefore, the soft interval SVM classification only needs to change the kernel function to support the classification of nonlinear samples. Our commonly used SVM classifier is this kind of soft interval classification. This paper studies the Gaussian Kernel commonly used in nonlinear classification, also known as the Radial Basis Function (RBF). Its function form is:

## **5.5. CART**

CART (Classification And Regression Tree) is a decision tree model proposed by Breiman et al. in 1984. CART can be used for both classification and regression. It is a widely used non-parametric classification and regression method. It builds a binary decision tree recursively to generate a predictive model.

CART uses a binary tree to judge the eigenvalues of the sample. The left branch of the binary tree takes the value "Yes" and the right branch takes the value "No". By recursively judging each feature value, CART can divide the feature space into finite units and determine the conditional probability distribution of the output variable Y on these units.

Therefore, for any input variable X, CART can divide X into a certain type of feature space according to the constructed binary tree, and give the conditional probability distribution of output variable Y under the condition of input variable X.

The modeling process of CART mainly consists of the following two steps:

(1) Generation of decision tree: A decision tree is generated based on the training data set. During the training process, try not to limit the depth of the decision tree and the number of leaf nodes.

(2) Pruning of decision tree: Prune the generated decision tree with an independent verification data set, and use the decision tree with the smallest loss function as the optimal subtree.

The classification tree selects the optimal feature and the optimal segmentation point based on the Gini index. In the classification problem, assuming that the output value has k categories, the probability that the sample point x belongs to the k-th category is , then the Gini index of the probability distribution is defined as:

For a given sample set D, |D| is the total number of samples in the set D, and is the number of samples belonging to the k-th category, then the Gini index of the set D is:

If set is divided into two parts and based on whether feature A takes a certain possible value a, then the Gini index of set D under the condition of feature A is defined as:

The Gini index Gini(D) represents the uncertainty of the set D, and the Gini index Gini(D,A) represents the uncertainty of the set D after the division of A=a. The greater the value of the Gini index, the greater the uncertainty of the sample set. Therefore, the optimal feature and optimal segmentation point can be selected through the Gini index.

## **5.6.** *Random forest*

Random Forest is a very flexible and easy-to-use supervised learning algorithm. The random forest classifier is composed of a group of decision trees. Each tree is generated by independent sampling random vectors, and each tree votes to find the most popular category to classify the input.[[36]](#footnote-36)

It can be said that the random forest algorithm is an improved version of the Bagging algorithm. The idea of the random forest is essentially the Bagging idea. Bagging is the bagging method. It is the most famous representative of the parallel integrated learning method. The algorithm process is as follows:

1. Extract the training set from the original sample set for the first time. Each round uses the Bootstrapping method to extract n training samples from the original sample set (for each sample, some may be drawn multiple times, or it may not be drawn once). M training sets are obtained through m rounds (the training sets are independent of each other).
2. Perform model training on each training set, where m training sets will get m models (there is no specific classification algorithm or regression algorithm here, the type of model algorithm is determined according to the problem).
3. In the classification problem, the final classification result will be obtained by voting according to the prediction results of the m models; for the regression problem, the average value of the above m models will be calculated as the final result. The framework of the Bagging method is shown in the figure below:



There are two main improvements made by random forest on the basis of Bagging: first, the random forest algorithm uses the CART decision tree as the weak learner; second, on the basis of the use of the weak learner, the random forest also makes the establishment of the decision tree. It has been improved. Based on the construction of Bagging ensemble learning with decision tree as the base learner, it introduces the selection of random attributes in the learning process of decision tree.

For the ordinary decision tree model, we usually select an optimal feature from all n samples on the node as the left and right subtree division of the decision tree, while the random forest model randomly selects some sample features on the node, assuming some samples The number of features is . If , an optimal feature is selected among these randomly selected 1 sample features to divide the left and right subtrees of the decision tree, thereby further improving the generalization ability of the model.

If , then there is no difference between the CART decision tree of the random forest and the ordinary CART decision tree at this time. The smaller the , the stronger the generalization ability of the model, which means that the model is more robust, but the model fitting ability of the training set becomes worse. In actual projects, the most suitable value of is generally obtained through cross-validation tuning.

## **5.7. Extreme random tree**

Geurts et al. (2006) proposed the extreme random tree method. According to the classic top-down method, the extreme random tree constructed a series of "free-growing" regression tree sets. Similar to the random forest method, the extreme random tree method is also composed of multiple decision trees, but it is different from the random forest method in that the extreme random tree method obtains the bifurcation value completely randomly, so as to perform the bifurcation of the regression tree, which is different from the random forest method. The best bifurcation attribute of the forest is obtained in a random subset. In addition, each regression tree in the extreme random tree method uses all training samples[[37]](#footnote-37).

## **5.8. GBDT**

The GBDT (Gradient. Boosting. Decision. Tree) algorithm is based on the improvement of the Grading Boosting. Tree, which solves the optimization problem of the general loss function.The GBDT algorithm is an iterative decision tree algorithm, which is composed of multiple decision trees. The final prediction result is to accumulate the prediction results of all decision trees, so it can avoid overfitting the prediction results of a single decision tree. problem. The core idea of GBDT is to continuously reduce the loss function by fitting residuals. In the process of generating trees, each tree will learn the residuals of the prediction results of the previous tree and repeat this process until the residuals are small enough.

The GBDT algorithm mainly contains three concepts: Respective Tree, Gradient Boosting, and Shrinkage.

The idea of gradient boosting is to generate multiple base classifiers iteratively, and then add the prediction results of each base classifier. If the prediction results of the base classifiers added at a certain time are wrong, then its negative impact is easily affected. Corrections are made in the generation of subsequent base classifiers. Controlling the shrinkage speed is one of the classic methods to control the complexity of the model. The shrinkage speed is usually used in the ridge regression algorithm to reduce the influence of the potentially unstable regression coefficient by reducing the regression coefficient. In the gradient boosting algorithm, the impact of the base classifier added each time on the overall model can be reduced by reducing the shrinkage speed.

The training steps of GBDT algorithm are as follows:

1. The initial model is a constant value:
2. Calculate the negative gradient of i=l,. . . , n based on the loss function space:
3. Generate base classifier based on .
4. Through the negative gradient of the loss function, optimize the decision tree parameters :
5. Update the model:
6. Repeat steps (1) to (5) until the specified number of base classifiers are generated.

## **5.9. AdaBoost**

AdaBoost, short for "Adaptive Boosting" in English, was proposed by Yoav Freund and Robert Schapire in 1995. Its self-adaptation lies in the fact that the wrong samples of the previous basic classifier will be strengthened, and all the weighted samples will be used to train the next basic classifier again. At the same time, a new weak classifier is added in each round until it reaches a predetermined sufficiently small error rate or reaches the pre-specified maximum number of iterations[[38]](#footnote-38).

The AdaBoost algorithm trains the same basic classifier (weak classifier) for different training sets, and then combines these classifiers obtained on different training sets to form a stronger final classifier (strong classifier). The theory proves that as long as the classification ability of each weak classifier is better than random guessing, when its number tends to infinity, the error rate of the strong classifier will tend to zero.

The different training sets in the AdaBoost algorithm are achieved by adjusting the weight corresponding to each sample. At the beginning, the weight corresponding to each sample is the same, and a basic classifier h1(x) is trained under this sample distribution. For samples that are misclassified by h1(x), the weight of the corresponding sample is increased; for samples that are correctly classified, the weight is decreased. In this way, the misclassified samples can be highlighted and a new sample distribution can be obtained.

At the same time, h1(x) is given a weight according to the misclassification, which indicates the importance of the basic classifier. The less the misclassification, the greater the weight. Under the new sample distribution, the basic classifier is trained again to obtain the basic classifier h2(x) and its weight. By analogy, after T such cycles, T basic classifiers and T corresponding weights are obtained. Finally, the T basic classifiers are accumulated according to a certain weight, and the final desired strong classifier is obtained.

The specific description of the AdaBoost algorithm is as follows:

Step 1: Given a training sample , where .

Step 2: Initialize the weight coefficient of each sample .

Step 3: In each cycle t=1.... , T, do the following steps:

Step 3.1: Use the weak classifier to train the weighted training samples to obtain an appropriate member classifier .

Step 3.2: Calculate the weight training error of . If the description is correct, I=1; otherwise, I=0.

Step 3.3: If or , set and skip to step 4.

Step 3.4: Let the weight of the weak classifier .

Step 3.5: Update the weight coefficient, get:

is the normalization coefficient, which can make .

Step 4: Output the strong classifier:.

## **5.10. XGBOOST**

The full name of XGBoost is the limit gradient boosting algorithm (extreme Gradient Boosting), which is a large-scale algorithm proposed by Dr. Chen Tianqi. The XGBoost algorithm supports a variety of algorithms as the base classifier. When the tree algorithm is used as the base classifier, it has made some improvements on the basis of the GBDT algorithm[[39]](#footnote-39).

The XGBoost algorithm pre-prunes each tree. The traditional GBDT only uses the first-order derivative information of the loss function to optimize. The XGBoost algorithm performs a second-order Taylor expansion on the loss function. At the same time, it also uses the first and second-order derivative information. To use Newton's method to optimize the loss function.

The XGBoost algorithm trains the model by means of an additive model. Each time a new base model is added, it is trained on the basis of the previous base model. The generation process of the base classifier is as follows:

The model prediction value of the i-th sample in the t-th round retains the model prediction value in round, and a new function is added. Each iteration adds a new function to minimize the error of the objective function as much as possible. The following formula is the final prediction result of the model after generating k trees:

Among them: is the set space of the tree; represents the feature vector of the ith data point; corresponds to the structure of the kth independent tree and the relative status of the leaf weight.

The XGBoost algorithm introduces a regular term in the loss function, which improves the generalization ability of a single tree. The following formula is the calculation method of the loss function. The loss function contains the error evaluation of the prediction result and the regular term. The regular term is the sum of the complexity of the tree and is used to control the complexity of the model:

The expanded form of the regular term is as follows, the parameter T represents the number of leaf nodes, and the parameter w represents the predicted value of the leaf nodes. The parameter γ can limit the number of leaf nodes, and the parameter X can adjust the predicted value of the leaf node to prevent the prediction value from being too large. These two parameters can effectively prevent overfitting.

In the training process of the XGBoost algorithm, the loss function is minimized as much as possible. The loss function of the t-th round of training is as follows:

Definition represents the sample set of the leaves of the jth tree, the first derivative of the loss function: , the second derivative of the loss function:. The way of the t-order Taylor expansion of the objective function is:

Define , the objective function can be transformed into:

After obtaining the partial derivative of w, the weight vector can be obtained:

The form of the objective function after substituting the weight vector into the objective function is as follows:

The core of the decision tree algorithm is the growth of the tree. The most important thing in the tree building process of the XGBoost algorithm is to find the best split point. When the node is split, XGBoost uses the greedy method to calculate the gain before and after the split point is added to determine whether a node grows. For the tth step, calculate the gain of segmentation to determine whether to segment:

Similar to the ED3 algorithm and the CART algorithm, the gain is obtained by subtracting a certain value before the split from a certain value after the split. represents the score of the left leaf node after segmentation, is the score of the right leaf node after segmentation, is the score calculated without segmentation, and γ represents the complexity cost introduced by adding a leaf node. The optimal segmentation criterion of the decision tree is to enumerate all possible segmentation methods for each node, calculate the gain value, and then use the linear scan method to give the optimal segmentation point.

## **5.11. LightGBM**

LightGBM (Light Gradient Boosting Machine Method) is an improved algorithm of GBDT algorithm. Through the histogram optimization algorithm, the leaf generation strategy with depth limitation, the GoSS sampling method, independent feature bundling and the optimization of histogram difference acceleration, the GBDT-based LightGBM can have higher training efficiency, better accuracy and support the advantages of parallel learning[[40]](#footnote-40).

(1) Histogram optimization algorithm

The histogram optimization algorithm transforms continuous features into discrete features and constructs a histogram. When training the model, the histogram algorithm uses the discretized value as a new index to calculate cumulative statistics while traversing the entire training data, helping to finally find the best split point. Since the computer consumes a certain amount of memory space when processing data, discretizing the continuous feature vector not only reduces the use of storage space, but also reduces the number of calculations for information gain. Although the use of the histogram reduces the accuracy of the data, the rough split node will bring a certain regularization effect.

(2) Leaf generation strategy with depth limitation

Generally, decision tree generation is grown in layers, which ensures that the complexity is easy to control and thus alleviates the occurrence of over-fitting. However, this approach leads to invalid calculations. In the process of decision tree generation, LightGBM and GBDT repeatedly split the feature with the largest split gain. Although this can ensure that better accuracy can be obtained under the same number of splits, due to the increase of the number of splits, the over-fitting phenomenon is more likely to occur than the previous method, and it is necessary to limit the depth of the decision tree.

(3) GOSS sampling method

For a single data instance, the gradient value is very small, and the training error is correspondingly small. However, if these examples are ignored, the data structure and distribution will change, which will affect the accuracy of the model. The GOSS (Gradient-based One-Side Sampling) sampling method retains all large gradient data instances and performs random sampling for small gradient data instances. Then use the small gradient instance to multiply the appropriate real number to enlarge the sampling sample to ensure that the data distribution is not affected, and to increase the performance of LightGBM.

Suppose that during the decision tree generation process, the data set O is the training set of a fixed node of the decision tree. The variance gain at node d of split feature j is defined as follows:

Where , , , represents the negative gradient value of the loss function corresponding to the output of each training sample of the training set O. The decision tree algorithm selects the split point by maximizing . For the GOSS method, the first step is to sort in descending order according to the absolute value of the gradient, and then keep the top a% of the samples and construct the sample set A. For the remaining sample set , we randomly select the number of samples as ,which is the sample subset B. Finally, maximize on the sample set A∪B according to the above formula, and select the optimal split point .

In the above formula,,,,and .

(4) Independent feature binding

Since the feature is divided into a smaller number of mutually exclusive bindings, this is an NP-hard problem, that is, it is impossible to find an accurate solution in polynomial time. LightGBM uses an approximate solution, allowing a small number of sample points between features that are not mutually exclusive (for example, there are some corresponding sample points that are not non-zero at the same time), allowing a small number of conflicts to get smaller Feature binding number.

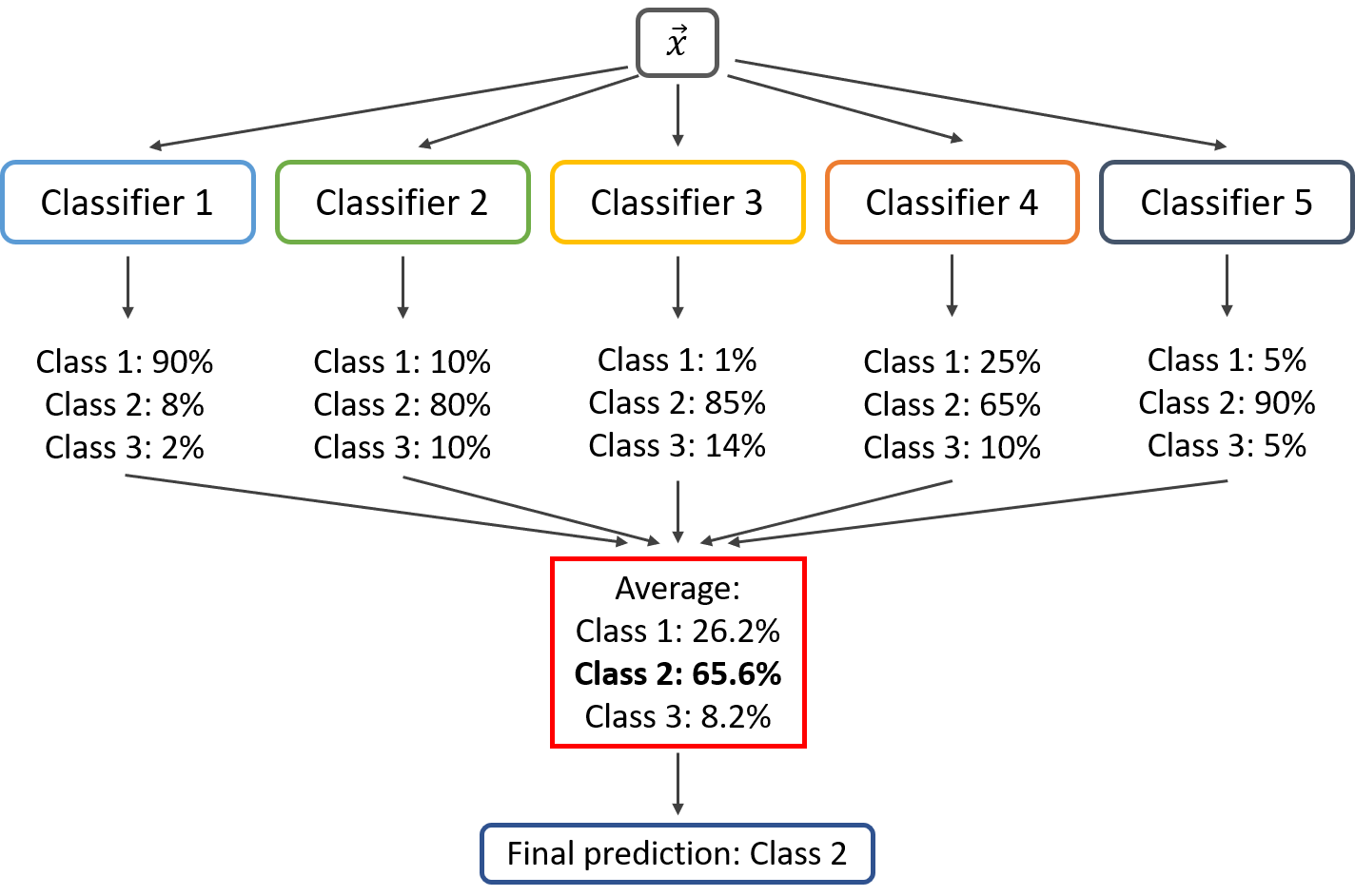
(5) Acceleration of histogram difference

The basic principle of LightGBM's difference acceleration when doing histograms is to use the difference between the histogram of the parent node and one of the child nodes to get the histogram of the other child node to improve the running speed.

# 6.Model fusion method

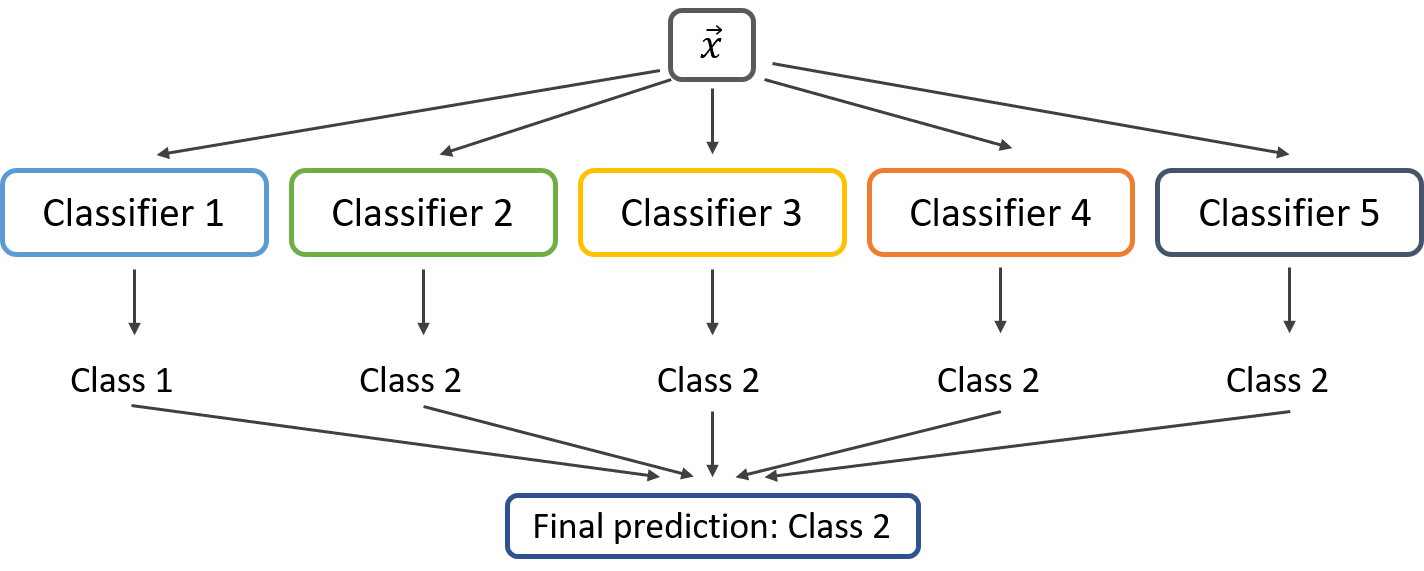
## **6.1. Soft-Voting**

In soft voting, every individual classifier(model) provides a probability value that a specific data point belongs to a particular target class. The predictions are weighted by the classifier's(model) importance and summed up. Then the target label with the greatest sum of weighted probabilities wins the vote.



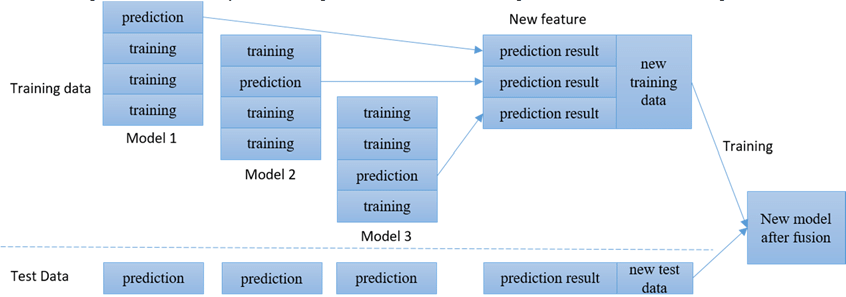
## **6.2. Hard-Voting**

In hard voting (also known as majority voting), every individual classifier(model) votes for a class, and the majority wins. In statistical terms, the predicted target label of the ensemble is the mode of the distribution of individually predicted labels.



## **6.3. Stacking**

The Stacking model fusion is generally composed of two-layer classifiers. The classifier in the base layer is called the base classifier and consists of multiple classifiers. The classifier in the meta-layer is called the meta classifier and is generally a classifier. The Stacking model fusion method is to use the output result of the base classifier as the input of the meta-classifier learning, so that the training process of the meta-classifier can fully learn the learning result of the base classifier, thereby correcting the prediction deviation of the base classifier and improving The generalization and prediction accuracy of the fusion model. The flow of Stacking model fusion is shown in the figure below:

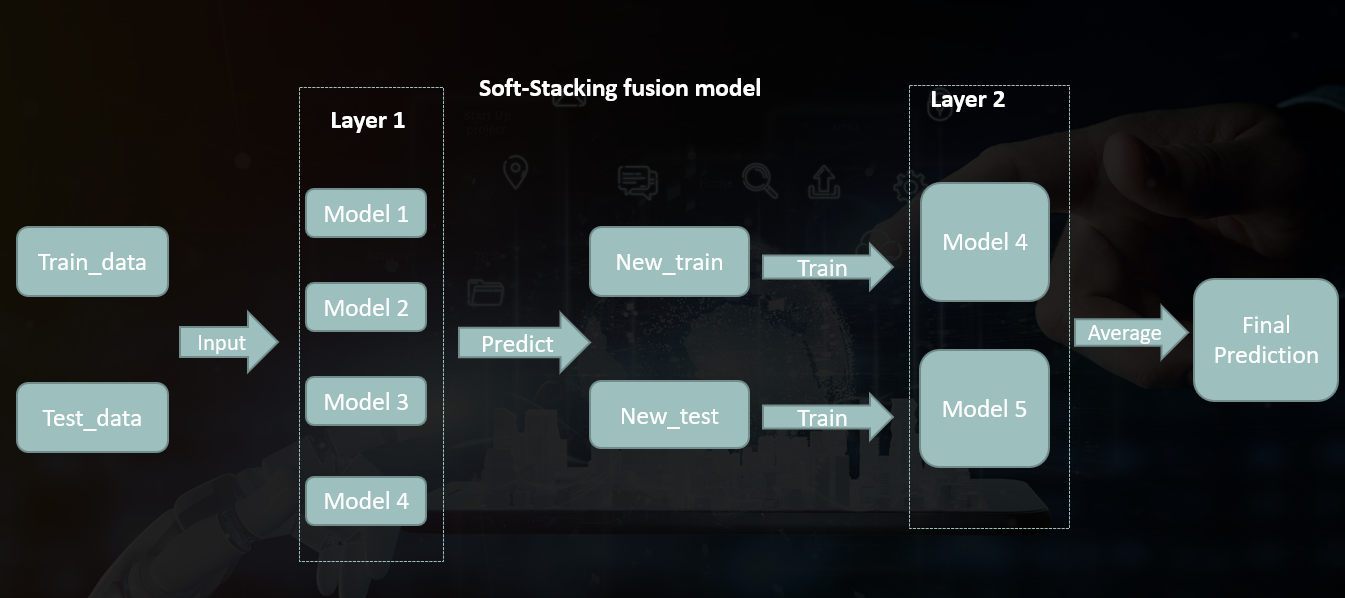


Train a number of different base-level models first, and then use the output of each base-level model trained in the previous step as input to train a sub-model. In order to improve the generalization ability of the model, the base classifier of the fusion model can choose a prediction algorithm with large differences.

## **6.4. Improvements to the stacking model: soft-stacking model**

In the stacking model fusion, the output result of the first layer classification model is used as the input of the second layer, so that the second layer classification model fully learns the result of the first layer classification, and the output result of the second layer is the final prediction. In the traditional stacking model, there is only one classifier in the second layer, and the ability to learn the output results of the first layer is limited.

My method is based on soft-voting, improving the traditional stacking model fusion method. That is, in the second layer, multiple classification models are added to learn the output results of the first layer, and the second layer's classification model adopts a soft-voting method to fuse. In other words, the output results of all models in the second layer are averaged as the final output. I name this fusion method soft-stacking . The flow of Soft-Stacking model fusion is shown in the figure below:



# 7.Model comparison and evaluation

## **7.1. Sampling method**

The total number of users in the data set is 260864, of which the number of users who have repeated purchases is 15,952, and the number of users who have not repeated purchases is 244912. Users with repeat purchase behavior accounted for about 6% of the total number of users. If users with repeated purchase behaviors are regarded as positive samples, and users who do not have repeated purchase behaviors are regarded as negative samples, then the positive and negative samples will be seriously imbalanced. The imbalance of sample categories will cause the machine learning model to be biased towards a larger number of categories, resulting in overfitting, and seriously affecting the prediction performance of machine learning.

According to the above requirements, before running the model, first divide the data set into a training set and a test set. among them

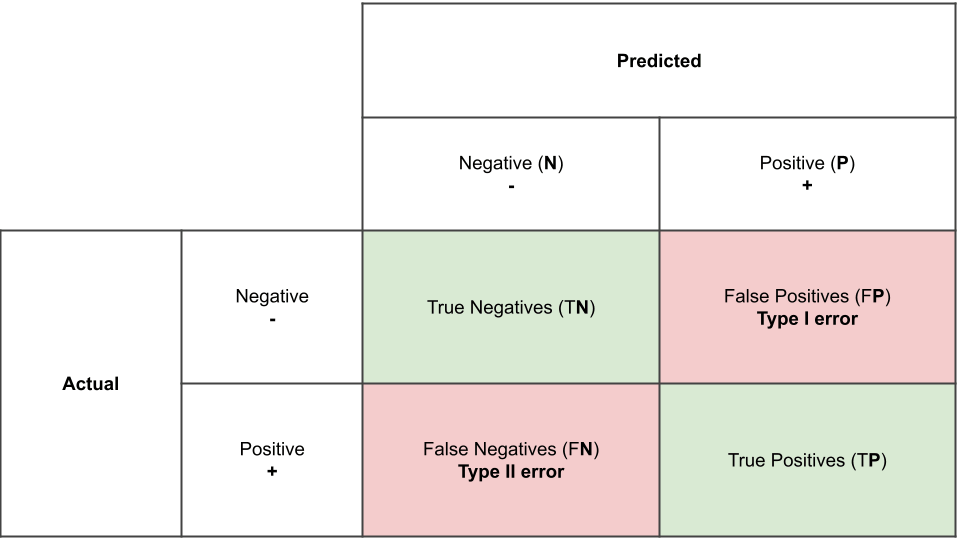
The test set is sampled without replacement. In each experiment, 1000 positive samples and 100 negative samples are randomly selected from the data set.

The sample is used for the evaluation of the model. The training set is composed of 10,000 positive samples randomly selected with replacement and 10,000 negative samples randomly selected without replacement from the rest data set.

## **7.2. Model evaluation method**

This article uses accuracy and AUC value as the model evaluation criteria. Accuracy and AUC value are the two most commonly used evaluation methods in classification tasks, which are suitable for both binary classification tasks and multi-classification tasks.

The direct meaning of the AUC value is the area under the ROC curve. ROC is called the receiver operating characteristic curve. In classification problems, the samples are sorted according to the prediction results of the classifier, and the samples are used as positive examples to predict in this order, and two important values are calculated each time, respectively Use them as the horizontal and vertical coordinates to plot, and get the "ROC" curve. The vertical axis of the ROC curve is "True Positive Rate" (TPR), and the horizontal axis is "False Positive Rate" (FPR). Confusion matrix based on the following table:



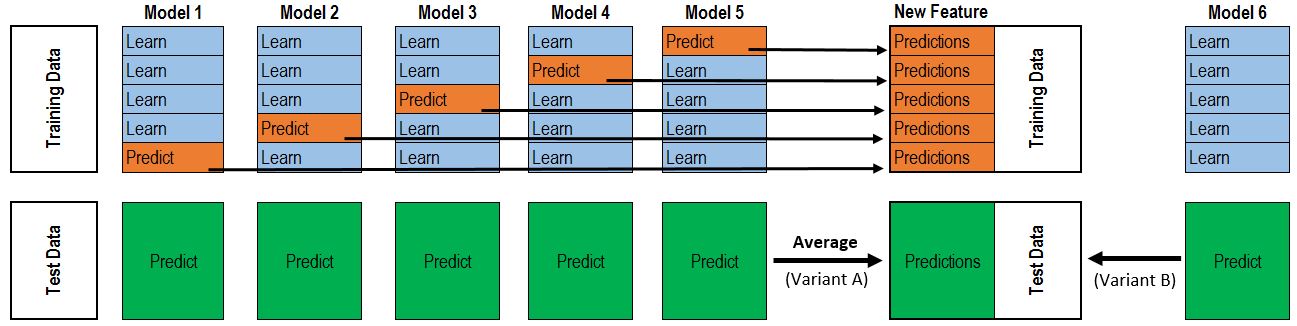
The calculation methods of TPR and FPR are as follows:

AUC can be obtained by summing the area of each part under the ROC curve. Assuming that the ROC curve is formed by sequentially connecting points with coordinates, the AUC value can be estimated for:

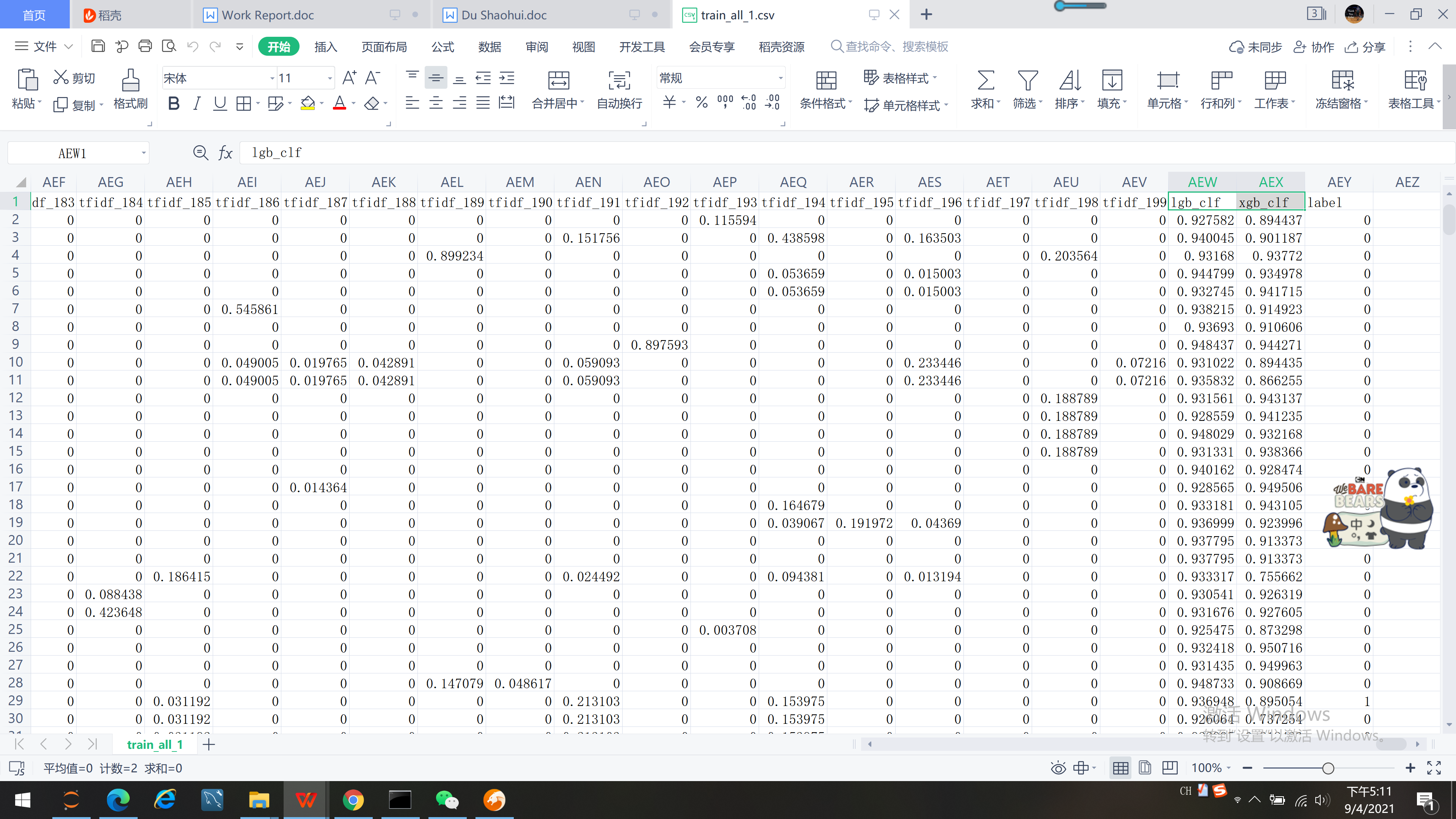
## **7.3. Feature optimization based on algorithm model**

**7.3.1 Feature generation based on stacking method**

Divide the training set into five folds, select one fold each time, and use the remaining four folds to train the model, and then predict the selected one fold data and the test set respectively. This is done 5 times to get the predicted value of 50% off and the predicted value of 5 test sets. Stack up the predicted values of the five-fold training set as the new feature of the training set; take the average of the predicted values of the five test sets as the feature of the predicted set. The specific process is as follows:



This paper uses lightgbm and xgboost models to construct two new features, lgb\_clf and xgboost\_clf, to improve the model’s predictive ability.



**7.3.2 Feature selection**

After the data feature extraction, vectorization, and new feature generation operations, the dimensionality of the data has reached 829. However, not every column contributes significantly to the final result. In order to save computational power and time cost, feature selection is often required for data sets with too high dimensions, that is, to filter out features that have a significant impact on the final result.

There are many ways to select features, such as deleting features with small variance, SelectKBest model, recursive elimination, and selection through classification models.

In actual implementation, I mainly select data through classification models. I used Logistic Regression (LR), Exta-Tree (ET) and LightGBM (LGB) models to filter features. The screening method is to retain only 200 features. Input all the data sets before and after feature selection into the decision tree classifier, and compare the accuracy of the two. The results are as follows:

|  |  |  |
| --- | --- | --- |
| Model | Before feature selection | After feature selection |
| LR | 0.85 | 0.82 |
| EM | 0.84 | 0.84 |
| LGB | 0.87 | 0.85 |

The results show that only the ET model has no effect on the output after feature selection. Therefore, the ET model is used to filter out the input 200-dimensional features.

## **7.4. Resualts**

**7.4.1 Comparison of common classification algorithm models**

Build logistic regression (LR), KNN, GaussianNB (GaussianNB), random forest (RF), Extra-Tree (ET), Adaboost, GBDT, XGboost, Lightgbm (LGB), SVM models respectively. Sampling 10 times, generate four sets of training set and test set, calculate the average and standard deviation of AUC of each model, and use it as an indicator for evaluating the model. The result is as follows

|  |  |  |
| --- | --- | --- |
| **Model** | **Avg(AUC)** | **Std(AUC)** |
| LR | 0.795 | 0.156 |
| KNN | 0.788 | 0.0976 |
| GaussianNB | 0.533 | 0.134 |
| RF | 0.801 | 0.172 |
| ET | 0.781 | 0.166 |
| Adaboost | 0.819 | 0.154 |
| GBDT | 0.813 | 0.154 |
| Xgboost | 0.826 | 0.132 |
| SVM | 0.811 | 0.122 |
| LGB | 0.826 | 0.131 |

**7.4.2 Comparison of fusion models**

Whether it is Voting or Stacking, the base model should choose a classifier (model) with greater difference. Adaboost, GBDT, XGboost, Lightgbm (LGB), are all based on Boosting algorithm. Among them, XGboost and LGB are tied for the highest accuracy, but LGB is faster in computing speed, so LGB is chosen as the base model.

Both RF and ET are tree models, but RF has higher AUC and faster calculations, so RF is chosen as the base model. GaussianNB has the lowest AUC, so it can not be used as bse model.

Therefore, the base models are: LR, KNN, RF, LGB, SVM.

Based on them, the soft-voting and hard-voting are constructed respectively, and the average and variance of AUC is as follows:

|  |  |  |
| --- | --- | --- |
| **Model** | **Avg(AUC)** | **Std(AUC)** |
| Soft-voting | 0.831 | 0.087 |
| Hard-voting | 0.827 | 0.062 |
| Stacking  (first-layer-models:LGB,RF,KNN,SVM;  second-layer-model:LR) | 0.835 | 0.074 |
| Soft-Stacking  (first-layer-models: LGB,RF,SVM;  second-layer-model:LR,KNN) | 0.837 | 0.063 |

It can be seen from the above results that the prediction ability and stability of the fusion model is better than that of a single classification model; stacking fusion is better than soft-voting and hard-voting; improved stacking (soft stacking) performs best.

# Conclusion

Based on user online behavior data, this paper studies and compares the application effects of industry-common classification algorithms and model fusion in user purchase behavior prediction, and improves the reliability of prediction by improving the fusion method. The specific work of this paper is summarized as follows:

(1) Based on the user characteristics and the relationship between the user and the merchant, this paper deeply digs into the effective features behind the original data, and applies the stacking method to generate new and effective features to lay a solid foundation for the classification model.

(2) This paper compares the prediction effects of 10 machine learning classification models commonly used in industries on user repurchase behavior, and applies different model fusion methods to improve the reliability of prediction.

(3) This paper combines stacking and soft-voting two model fusion methods to improve the predictive ability and predictive stability of traditional stacking fusion methods.

## **Theoretical contribution:**

1. This article systematically compares and demonstrates the predictive capabilities of common classification algorithm models.

2. This article scientifically shows the basic selection and construction of common fusion model algorithms.

3. Through the combination of soft-voting and stacking methods, this article breaks through the limitation of only one classifier in the second layer of the stacking hybrid algorithm, and improves the predictive ability of the stacking algorithm.

## **Practical contribution:**

Through the comparison and construction of fusion models, this paper successfully and accurately predicts the probability of users repurchasing after receiving advertisements. Combining the prediction results, this article gives an advertisement push strategy, that is, in the case of limited budget, preferentially push advertisements to users with higher repurchase probability. This effectively improves the efficiency and conversion rate of advertising push.

As the amount of data grows, more and more noise in the data and the differences in data patterns caused by the differences in user behavior patterns have caused existing machine learning methods to encounter difficult bottlenecks in solving the problem of repeat purchase behavior prediction. The emergence of new problems has caused great obstacles to the promotion and application of machine learning technology on e-commerce platforms. In this paper, through the mining of consumer data characteristics, the comparison of a variety of commonly used machine learning classification algorithms, and the application and improvement of model fusion methods, a fusion model with accurate and reliable prediction capabilities is constructed. The repurchase behavior prediction technology can be applied to the recommendation system of current e-commerce platforms. Due to the huge user base and transaction amount of the e-commerce platform, a slight improvement in the prediction effect will bring huge profits to the e-commerce platform.

# Reference

## **Academic Literature in English**

1.Bauer R A.(1960).Consumer behavior as risk taking.In R.S.Hancock(Ed.), Dynamic marketing for a changing world(389- 398).Chicago: American Marketing Association.

2.Overby J.W., Lee E.J.The effects of utilitarian and hedonic online shopping value onconsumer preference and intentions[J].Journal of Business Research,2006,59(10-11):1160-1166.

3.Chang E.C.,Tseng Y.F.Research note: E-store image, perceived value and perceived risk[J].Journal of Business Research,2013,66(7):864-870.

4.Wang T.G.,Fang Y.H.,Huang H.Y.Understanding customers'repeat purchase intentions in B2C e-commerce:the roles of utilitarian value,hedonic value and perceived risk[J].Information System Journal,2014,24(1):85-114.

5.Teo T.S.H., Wang P, Leong C.H.Understanding Online Shopping Behaviour Using a Transaction Cost Economics Approach[J].International Journal of Internet Marketing &Advertising,2004,1(1):62-84.

6.Mukherjee A,Banerjee S,Bandyopadhyay S.A.Simulation Model Using Transaction Cost Economics to Analyze the Impact of Social Media on Online Shopping[C].ICSI.Springer-Verlag,2012:43-53.

7.Benazić D，Tanković A.Č .Impact of Perceived Risk and Perceived Cost on Trust in the Online Shopping Websites and Customer Repurchase Intention[C].Cromar Congress:Marketing Theory and Practice-Building Bridges and Fostering Collaboration.2015.

8.Myung J.K., Choong K.L., Namho C,et al.Factors Affecting Online Tourism Group Buying and the Moderating Role of Loyalty.[J].Journal of Travel Research,2014,53(3):380-394.

9.Wu L.Y.,Chen K.Y.,Chen P.Y.,et al.Perceived value, transaction cost, and repurchase-intention in online shopping: A relational exchange perspective[J]. Journal of Business Research, 2014,67(1):2768-2776.

10.Yoon C,Gonzalez R,Bechara A,et al.Decision neuroscience and consumer decision making[J].Marketing Letters, 2012, 23(2):473-485.

11.Zhang Y,Feng Y.Q.Factors that Influence a Buyer's Decision Process of Shopping Online: The Effects of Tradition and Virtual Community[M]. 2011.

12.Tian Y,Ye Z,Yan Y,et al.A practical model to predict the repeat purchasing pattern of consumers in the C2C e-commerce[J]. Electronic Commerce Research, 2015, 15(4):571-583.

13.Chang H.J.,Hung L.P., Ho C.L.An anticipation model of potential customers’purchasing behavior based on clustering analysis and association rules analysis.[J]. Expert Systems with Applications, 2007, 32(3):753-764.

14.Cho Y.S.,Moon S.C., Oh I.B.,et al.Incremental Weighted Mining based on RFM Analysis for Recommending Prediction in u-Commerce[J]. International Journal of Smart Home, 2013,7(6):133-144.

15.Cui D,Curry D.Prediction in Marketing Using the Support Vector Machine[M]. INFORMS,2005.

16.Gupta R,Pathak C.A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing[J]. Procedia Computer Science, 2014, 36:599-605.

17.Zuo Y,Ali A.B.M.S.,Yada K.Consumeer Purchasing Behavior Extraction Using Statistical Learning Theory[J].Procedia Computer Science,2014,35:1464-1473.

18.Hart, P.E. (1968) The condensed nearest neighbor rule. IEEE Transactions on Information Theory, 14, 515-516.

19.D. Shaohui, G. Qiu, H. Mai and H. Yu, "Customer Transaction Fraud Detection Using Random Forest," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 144-147, doi: 10.1109/ICCECE51280.2021.9342259.

20.Geurts P，Ernst D， Wehenkel L.2006.Extremely randomized trees.Mach Learn，63（1）：3-42.

21.Freund, Yoav; Schapire, Robert E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. 1995.

22.Chen T，He T，Benesty M，et al．xgboost：Extreme Gradient Boosting [J].2017.

23.Brownlee, Jason (March 31, 2020). "Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost".

## **Academic Literature in Other Languages**

1.李日旭.网购生鲜农产品中感知风险对购买意愿的影响研究[D].哈尔滨工业大学,2017.

2.黄茜.经济型酒店顾客重复购买行为影响因素实证研究[D].湖北大学,2011.

3.熊晓元.基于互动和感知理论的网络重购行为研究[D].西南交通大学,2014.

4.张延静.网上节日促销中顾客感知价值、顾客满意与顾客网购行为关系研究[D].青岛理工大学,2014.

5.张 春 桃 . 顾 客 感 知 价 值 对 消 费 者 网 购 行 为 影 响 研 究 —— 以 大 学 生 群 体 为 例 [J].商,2016(01):281.

6.肇丹丹.基于 SEM 的消费者网购再惠顾意愿度量分析[J].统计与决策,2014(01):89-92.

7.洪硕.旅游移动电商环境下的游客购买决策影响因素研究[D].广西师范大学,2018.

8.罗胜.电子商务环境下化妆品消费者冲动性购买行为研究[D].华南理工大学,2018.

9.徐雪枫.网络消费者购物决策行为影响因素的实证研究[D].东北财经大学,2012.

10.张春莲.客户购买行为的 BG/NBD 预测模型及其应用研究[D].哈尔滨工业大学,2006.

11.陈洁,谢文昕,杨升荣.在线渠道消费者动态品牌选择购买率预测[J].工业工程与管理,2011,16(03):52-56.

12.舒方,马少辉.客户重复购买的组合预测方法[J].计算机与现代化,2015(05):67-70.

13.李美其,齐佳音.基于购买行为及评论行为的用户购买预测研究[J].北京邮电大学学报(社会科学版),2016,18(04):18-25.

14.杜刚,黄震宇.大数据环境下客户购买行为预测[J].管理现代化,2015,35(01):40-42.

15.刘柏鳞.基于客户网络购物行为分析及预测系统研究[D].北京工业大学,2016.

16.祝歆,刘潇蔓,陈树广,李静,张天宇.基于机器学习融合算法的网络购买行为预测研究[J].统计与信息论坛,2017,32(12):94-100.

17.周成骥.基于机器学习的商品购买行为预测模型设计[D].广州大学,2018.

1. Bauer R A.(1960).Consumer behavior as risk taking.In R.S.Hancock(Ed.), Dynamic marketing for a changing world(389- 398).Chicago: American Marketing Association. [↑](#footnote-ref-1)
2. 李日旭.网购生鲜农产品中感知风险对购买意愿的影响研究[D].哈尔滨工业大学,2017. [↑](#footnote-ref-2)
3. 黄茜.经济型酒店顾客重复购买行为影响因素实证研究[D].湖北大学,2011. [↑](#footnote-ref-3)
4. 熊晓元.基于互动和感知理论的网络重购行为研究[D].西南交通大学,2014. [↑](#footnote-ref-4)
5. 张延静.网上节日促销中顾客感知价值、顾客满意与顾客网购行为关系研究[D].青岛理工大学,2014. [↑](#footnote-ref-5)
6. 张 春 桃 . 顾 客 感 知 价 值 对 消 费 者 网 购 行 为 影 响 研 究 —— 以 大 学 生 群 体 为 例 [J].商,2016(01):281. [↑](#footnote-ref-6)
7. 肇丹丹.基于 SEM 的消费者网购再惠顾意愿度量分析[J].统计与决策,2014(01):89-92. [↑](#footnote-ref-7)
8. Overby J.W., Lee E.J.The effects of utilitarian and hedonic online shopping value onconsumer preference and intentions[J].Journal of Business Research,2006,59(10-11):1160-1166. [↑](#footnote-ref-8)
9. Chang E.C.,Tseng Y.F.Research note: E-store image, perceived value and perceived risk[J].Journal of Business Research,2013,66(7):864-870. [↑](#footnote-ref-9)
10. Wang T.G.,Fang Y.H.,Huang H.Y.Understanding customers'repeat purchase intentions in B2C e-commerce:the roles of utilitarian value,hedonic value and perceived risk[J].Information System

    Journal,2014,24(1):85-114. [↑](#footnote-ref-10)
11. Teo T.S.H., Wang P, Leong C.H.Understanding Online Shopping Behaviour Using a Transaction Cost Economics Approach[J].International Journal of Internet Marketing &Advertising,2004,1(1):62-84. [↑](#footnote-ref-11)
12. Mukherjee A,Banerjee S,Bandyopadhyay S.A.Simulation Model Using Transaction Cost Economics to Analyze the Impact of Social Media on Online Shopping[C].ICSI.Springer-Verlag,2012:43-53. [↑](#footnote-ref-12)
13. Benazić D，Tanković A.Č .Impact of Perceived Risk and Perceived Cost on Trust in the Online

    Shopping Websites and Customer Repurchase Intention[C].Cromar Congress:Marketing Theory and Practice-Building Bridges and Fostering Collaboration.2015. [↑](#footnote-ref-13)
14. Myung J.K., Choong K.L., Namho C,et al.Factors Affecting Online Tourism Group Buying and the Moderating Role of Loyalty.[J].Journal of Travel Research,2014,53(3):380-394. [↑](#footnote-ref-14)
15. Wu L.Y.,Chen K.Y.,Chen P.Y.,et al.Perceived value, transaction cost, and repurchase-intention in online shopping: A relational exchange perspective[J]. Journal of Business Research, 2014,67(1):2768-2776. [↑](#footnote-ref-15)
16. Yoon C,Gonzalez R,Bechara A,et al.Decision neuroscience and consumer decision making[J].Marketing Letters, 2012, 23(2):473-485. [↑](#footnote-ref-16)
17. 洪硕.旅游移动电商环境下的游客购买决策影响因素研究[D].广西师范大学,2018. [↑](#footnote-ref-17)
18. 罗胜.电子商务环境下化妆品消费者冲动性购买行为研究[D].华南理工大学,2018. [↑](#footnote-ref-18)
19. 徐雪枫.网络消费者购物决策行为影响因素的实证研究[D].东北财经大学,2012. [↑](#footnote-ref-19)
20. Zhang Y,Feng Y.Q.Factors that Influence a Buyer's Decision Process of Shopping Online: The Effects of Tradition and Virtual Community[M]. 2011. [↑](#footnote-ref-20)
21. 张春莲.客户购买行为的 BG/NBD 预测模型及其应用研究[D].哈尔滨工业大学,2006. [↑](#footnote-ref-21)
22. 陈洁,谢文昕,杨升荣.在线渠道消费者动态品牌选择购买率预测[J].工业工程与管理,2011,16(03):52-56. [↑](#footnote-ref-22)
23. Tian Y,Ye Z,Yan Y,et al.A practical model to predict the repeat purchasing pattern of consumers in the C2C e-commerce[J]. Electronic Commerce Research, 2015, 15(4):571-583. [↑](#footnote-ref-23)
24. 舒方,马少辉.客户重复购买的组合预测方法[J].计算机与现代化,2015(05):67-70. [↑](#footnote-ref-24)
25. 李美其,齐佳音.基于购买行为及评论行为的用户购买预测研究[J].北京邮电大学学报(社会科学版),2016,18(04):18-25. [↑](#footnote-ref-25)
26. Chang H.J.,Hung L.P., Ho C.L.An anticipation model of potential customers’purchasing behavior based on clustering analysis and association rules analysis.[J]. Expert Systems with Applications, 2007, 32(3):753-764. [↑](#footnote-ref-26)
27. Cho Y.S.,Moon S.C., Oh I.B.,et al.Incremental Weighted Mining based on RFM Analysis for Recommending Prediction in u-Commerce[J]. International Journal of Smart Home, 2013,7(6):133-144. [↑](#footnote-ref-27)
28. Cui D,Curry D.Prediction in Marketing Using the Support Vector Machine[M]. INFORMS,2005. [↑](#footnote-ref-28)
29. Gupta R,Pathak C.A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing[J]. Procedia Computer Science, 2014, 36:599-605. [↑](#footnote-ref-29)
30. Zuo Y,Ali A.B.M.S.,Yada K.Consumeer Purchasing Behavior Extraction Using Statistical Learning Theory[J].Procedia Computer Science,2014,35:1464-1473. [↑](#footnote-ref-30)
31. 杜刚,黄震宇.大数据环境下客户购买行为预测[J].管理现代化,2015,35(01):40-42. [↑](#footnote-ref-31)
32. 刘柏鳞.基于客户网络购物行为分析及预测系统研究[D].北京工业大学,2016. [↑](#footnote-ref-32)
33. 祝歆,刘潇蔓,陈树广,李静,张天宇.基于机器学习融合算法的网络购买行为预测研究[J].统计与信息论坛,2017,32(12):94-100. [↑](#footnote-ref-33)
34. 周成骥.基于机器学习的商品购买行为预测模型设计[D].广州大学,2018. [↑](#footnote-ref-34)
35. Hart, P.E. (1968) The condensed nearest neighbor rule. IEEE Transactions on Information Theory, 14, 515-516. [↑](#footnote-ref-35)
36. D. Shaohui, G. Qiu, H. Mai and H. Yu, "Customer Transaction Fraud Detection Using Random Forest," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 144-147, doi: 10.1109/ICCECE51280.2021.9342259. [↑](#footnote-ref-36)
37. Geurts P，Ernst D， Wehenkel L.2006.Extremely randomized trees.Mach Learn，63（1）：3-42 [↑](#footnote-ref-37)
38. Freund, Yoav; Schapire, Robert E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. 1995. [↑](#footnote-ref-38)
39. Chen T，He T，Benesty M，et al．xgboost：Extreme Gradient Boosting [J]．2017． [↑](#footnote-ref-39)
40. Brownlee, Jason (March 31, 2020). "Gradient Boosting with Scikit-Learn, XGBoost, LightGBM, and CatBoost" [↑](#footnote-ref-40)