

# Classification and Prediction of Bank Marketing Activity by Machine Learning

Sibo Zhao\*

School of Management, Ocean University of China, Qingdao, China

\*Corresponding author: zhaosibo@stu.ouc.edu.cn

**Abstract.** As the competition in the financial industry intensifies, the customer experience has become the key to whether the financial institutions can continue to develop. As the traditional banking services are difficult to meet the needs of customers, more and more people begin to choose to trade through self-service channels or online banking. With the popularity of mobile devices, people are increasingly dependent on mobile banking and mobile payment. Banks analyze the customer behavior data to realize the effective interaction between customers and banks, which has become a hot issue in the financial industry. From the perspective of machine learning, this paper will analyze the ability of a bank customer to buy products, mainly including customer purchasing power analysis, product classification, customer stratification and other aspects. On this basis, this paper will use the decision tree algorithm to segment the customers with different purchasing power, so as to achieve precision marketing.

**Keywords:** Machine learning; banking; customer purchasing power.

## 1. Introduction

One of the main businesses of banks, as the main representatives of the financial sector, is time deposits, and the goal of banks is precision marketing, which is to find customers with lifetime value, and then to maintain a long-term relationship between the customers and the bank, so that they choose the bank's financial products. And bank managers always hope that risk control analysts can predict very accurately whether customers will subscribe to the bank's products, what are the influencing factors affecting customers' subscription and how to improve them, so as to truly realize the precision marketing of bank financial products [1].

In recent years, scholars from all walks of life have different views on analyzing operational activities in the financial sector, and one representative approach is machine learning. The field of machine learning research has changed significantly, from the academic and theoretical research direction, the goal of machine learning has been changed to solve the complex problems in real life, so that our decision-making becomes easier, and machine learning-related technologies have been utilized successfully in a variety of domains, including computer vision, natural language processing, image recognition, and speech recognition. And many researchers and scholars in the field of machine learning have proposed a lot of methods for the interpretability of machine learning models, and published a lot of expository articles in domestic and international journals. Notably, in his new book *Interpretable Machine Learning*, statistician and data scientist Christoph Molnar provided a comprehensive introduction to machine learning interpretability and explained the need for machine learning interpretability, how to use models to make decisions, and how to account for black box models are discussed. In the book, the author introduces the following methods for interpreting machine learning models, including starting from a dataset, introduction of several models with strong interpretability, model-independent interpretation, example-based interpretation, and several other methods. In China, Wu Fei et al. discussed deep learning interpretability in the article "Interpretability of Deep Learning," arguing that deep learning interpretability research can begin with aspects of causality, inference, cognition, and the intelligent interactivity of the model to establish theoretical methods and models with strong interpretability [2]. Shou-Ling Ji et al. in their paper "A Review of Machine Learning Model Interpretability Methods, Applications, and Security Research" provide an overview and discussion of machine learning interpretability methods, analyze the security problems

of machine learning interpretability, and discuss the challenges of machine learning interpretability research as well as the feasibility of future research directions.

The main work of this paper is to propose an interpretable machine learning model using bank customer data as an example, with the aim of predicting whether a bank customer will subscribe to a time deposit or not, giving the bank staff an effective decision to realize accurate marketing and improve the bank's revenue. This paper focuses on three main aspects of interpretable analysis:

The first is the interpretable analysis before building the model. Before building the model, we have to understand the data very well and use the method of data analysis to mine the laws from the data. Then, using data visualization methods, accurately and fully comprehend the overall distribution of the data, which involves summarizing all of the features in the data and then mining those features that are more relevant to the characteristics to be predicted for feature extraction. Secondly, because the data used in this paper is characterized by imbalance, the processing method of unbalanced data is used to make the data present a balanced nature.

Then is the interpretability analysis in modeling, this paper uses two machine learning models, Logistic Regression and KNN, to predict the customer purchase behavior through bank marketing data, and compare the results of the two models.

(3) Finally, the interpretability approach used after modeling helps us to understand the predictive shape between the input features and the output features. Interpretability after modeling is associated with the model built and focuses on understanding how these features affect the predicted results. The main focus is to use Feature Importance through this tool to understand how these features act on the prediction and how much the model built affects and is sensitive to some of the important features etc.

In this paper, the data sources used are related to the marketing activities of the bank, and by comparing the built machine learning model with the single and combined models, it is proved that the constructed model can improve the accuracy of the classification prediction, and the proposed model also has the interpretability, which gives credible decision making to the bankers.

## **2. Machine Learning Theory and Methodology**

### **2.1. Machine Learning and Deep Learning**

Machine learning is one of the various ways for humans to realize "artificial intelligence". We should try to find a way to "predict" unknown data through empirical data. Traditional machine learning requires multi-domain knowledge, and the appropriate methods such as elephant decision tree, random forest, and Bayesian learning are selected according to the characteristics of the data. Deep learning is one of the many ways of machine learning, inspired by how the human brain works. This method constructs a black box in which units are arranged in layers, each given a weight to determine how much the signal input to this unit affects the result [3].

### **2.2. Neural Network-related Concepts**

Neural network is one of the methods to realize deep learning. It is a mathematical model that takes neurons as the basic unit and mimics the behavioral characteristics of animal neural network. Its key elements include learning rules and topology. Among them, BackPropagation (BP) algorithm is a common learning rule of neural network model, which has good persistence and timely prediction. The topology of the neural network model consists of the input layer, the hidden layer and output layer. Each node of the input layer corresponds to the predictor variables, the output layer node corresponds to the target variable, between the input layer and output layer is hidden layer, between each layer through activation function and node weights together, form a learning network topology, and then use this structure for repeated neural network learning, until reach the desired goal, the weight is the optimal result by BP algorithm [4].

In machine learning models, some models are difficult to explain, such as deep neural networks. Deep neural networks can fit highly complex data with a large number of parameters that are very

difficult to interpret, such as algorithms such as SVM, RNN, and LSTM. However, some algorithms can be explained relatively easily, such as logistic regression LR, ensemble learning model, K-nearest neighbor and other algorithms, among which ensemble learning model includes random forest RF, GBDT, Xgboost, LightGBM, GcForest and other algorithms. This paper uses logistic regression and KNN for prediction, and first introduces interpretable models: logistic regression and ensemble learning model [5].

### 3. Model Construction and Results

#### 3.1. Model Research

##### 3.1.1 Dataset

In this paper, stratified sampling method is adopted when data set is divided. This method can ensure that positive and negative samples are evenly divided into training set, verification set and test set, and will not lead to only positive or negative samples in a certain data set, so as to avoid the deviation in prediction of training set, verification set and test set due to unbalanced data distribution. It is helpful to improve the overall prediction accuracy of the model. The layered sampling method is adopted to evenly divide the positive and negative samples in the whole data set into the training set, verification set and test set, in which the training set accounts for 80%, the verification set accounts for 10% and the test set accounts for 10%. 80% of the training set is used to train the parameters of the model, 10% of the verification set is used to verify the prediction performance of the model, so that the parameters of the model are optimized, and the remaining 10% is used to test fitting, and finally the best classification prediction model is obtained.

##### 3.1.2 Five-fold Cross Validation

In this paper, K-fold cross-validation method is used to evaluate the predictive performance of ensemble learning model when it is trained. Cross-validation is a model validation method that evaluates the adaptability of machine learning algorithms to test sets, verifies whether the model is reliable, and prevents overfitting of the model [6].

- a) Divide the whole Data set into two parts, training set D and test set S;
- b) The training set D was divided into 5 parts, D1, D2, D3, D4 and D5;
- c) Use one of the 5 copies each time as the verification set, and the remaining 4 copies as the training set;
- d) After 5 times of cross-training, 5 different results were obtained;
- e) Finally, the average value of the 5 experimental results is taken.

##### 3.1.3 Model Evaluation Index

The sample data used in this experiment showed a serious imbalance, with a small amount of positive sample data accounting for 11.23%, a large amount of negative sample data accounting for 88.77%, and a ratio of positive and negative samples approaching 1:8. In Chapter 3 above, the unbalance of data was dealt with by the combination of undersampling and oversampling. In general classification problems, ACC (accuracy) is used as an evaluation index. However, ACC cannot accurately reflect the accuracy of the model in the classification of unbalanced data, because the imbalance of data will lead to the bias of the prediction to the side with large amount of data [7]. Although the accuracy of the output is high, the accuracy of the output is poor on a small number of samples. Therefore, this paper selects the method of confusion matrix on the evaluation index of the model to evaluate the prediction effect of the model. Confusion matrix can well reflect the accuracy of classification and avoid the above problems [8].

In this experiment, 1 represents subscription and is a positive sample; 0 indicates no subscription and is a negative sample.

- 1) If 1 is predicted to be 1, it is True Postive (TP);
- 2) If 1 is predicted to be 0, it is False Negative (FN);

- 3) If 0 is predicted to be 1, it is False Positive (FP);
- 4) If 0 is predicted to be 0, it is True Negative (TN);

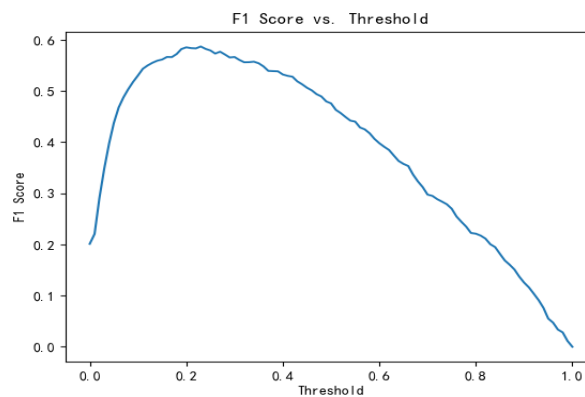
### 3.2. Logistic Regression

A generalized linear regression analysis model that is a part of supervised learning in machine learning, logistic regression is also known as logistic regression analysis. Although its computation and derivation procedures resemble those of regression, its primary application is the solution of binary classification problems (although it may also address problems involving many categories). One or more pieces of data (test set) are categorized for a given set or sets of data (training set) after the model has been trained using a given n sets of data (training set).

When building a Logistic Regression model, we train and test the model in three steps:

1. Isolate important variables by recursive feature elimination (RFE)
2. Run Logistic regression to understand the beta of important variables
3. Find the best parameters and thresholds

Then Calculate the model's performance at different thresholds to determine the appropriate threshold to use. The optimal threshold obtained during the validation process was 0.23 (See Fig. 1).

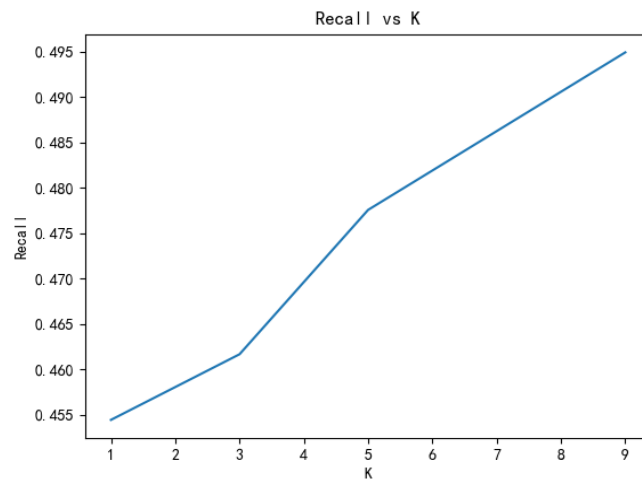


**Fig. 1** F1 Score of Logistic Regression

### 3.3. KNN

K-Nearest Neighbor algorithm implementation principle: in order to determine the category of the unknown sample, to all known categories of samples as a reference, calculate the distance between the unknown sample and all known samples, from which the closest to the unknown sample of the K known samples, according to the majority-voting, the unknown sample and the K closest samples belonging to the category accounted for a larger number of classified as a class.

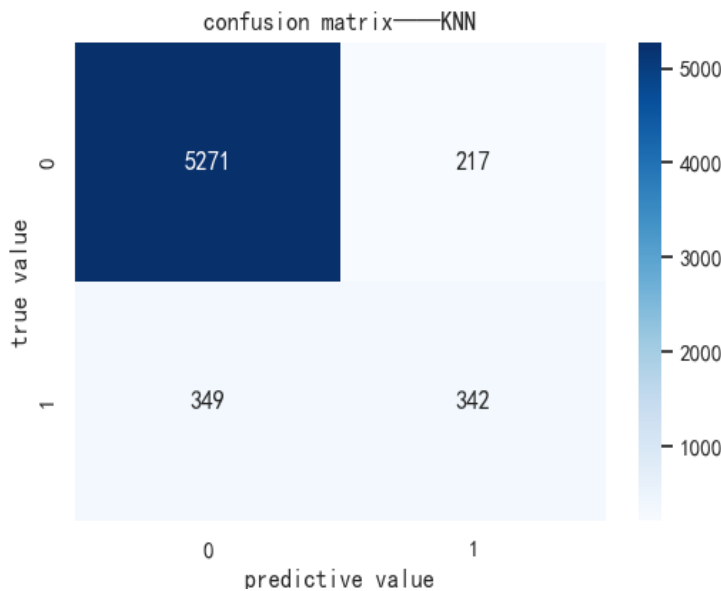
One key aspect of the KNN algorithm lies in determining the value of K. Since KNN does not support calculating Feature Importance. The final choice of K value is decided by calculating the F1 score corresponding to different K values. The optimal K value obtained during the validation process was 9 (See Fig. 2).



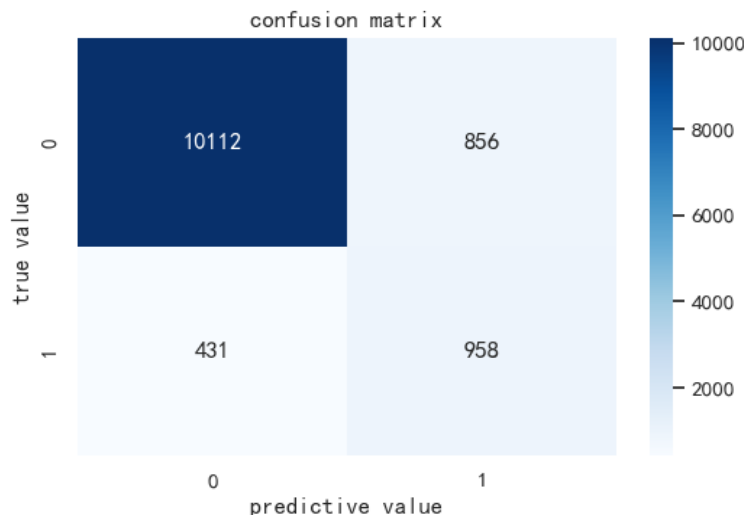
**Fig. 2** Recall Rate of KNN

**3.4. Results**

It can be seen from the confusion matrix that there are 417 customers in the test set who will subscribe for time deposits, and the Ensemble LSTM model is used for prediction. Among the 417 customers, 384 customers are successfully predicted, with the success rate of positive prediction reaching 92.09% and the success rate of negative prediction reaching 86.33%. In other words, 33 customers who agreed to subscribe to time deposits were wrongly predicted to disapprove; Among the customers who did not agree, 450 were wrongly predicted to agree to subscribe deposits, and the AUC of the whole model reached 95.08%. The effect of classification prediction was very good, and it played a positive role in the bank's marketing decisions (See Fig. 3, Fig. 4, and Fig. 5).

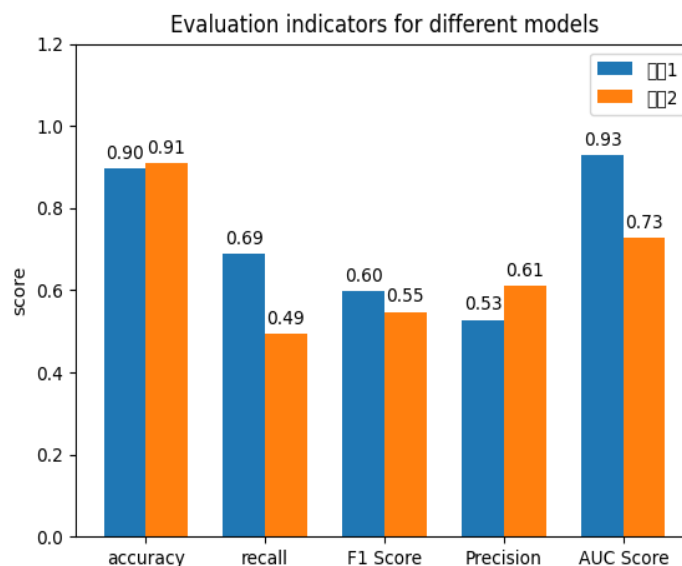


**Fig. 3** Confusion Matrix-KNN



**Fig. 4** Confusion Matrix-Logistic Regression

By comparing the results of these 2 models, we can draw the conclusion without question: Logistic regression has higher prediction accuracy than KNN.

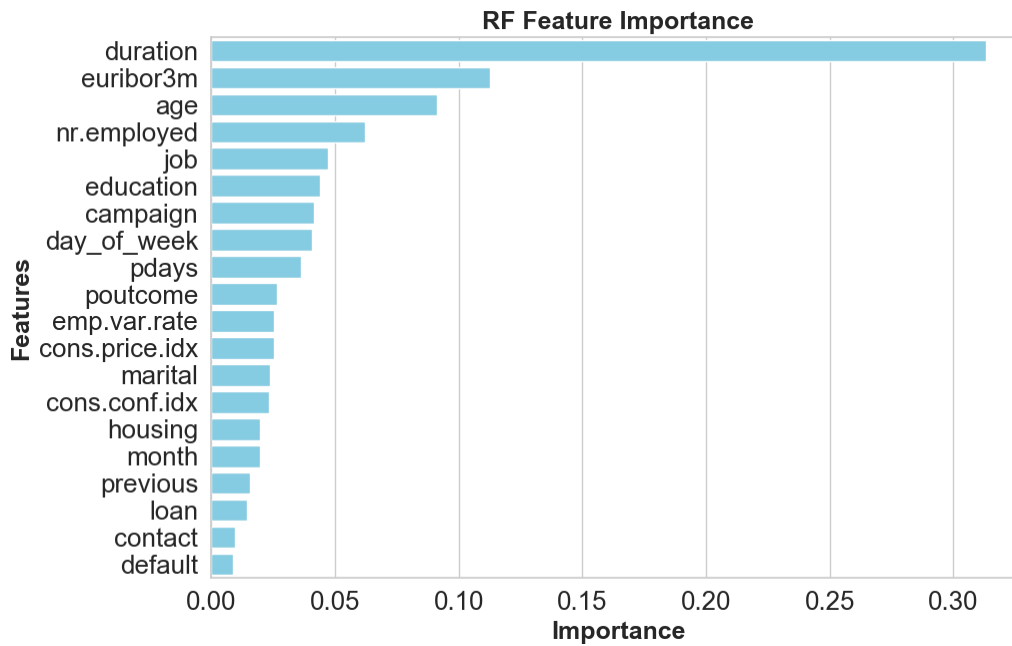


**Fig. 5** Evaluation Indicators for Different Models

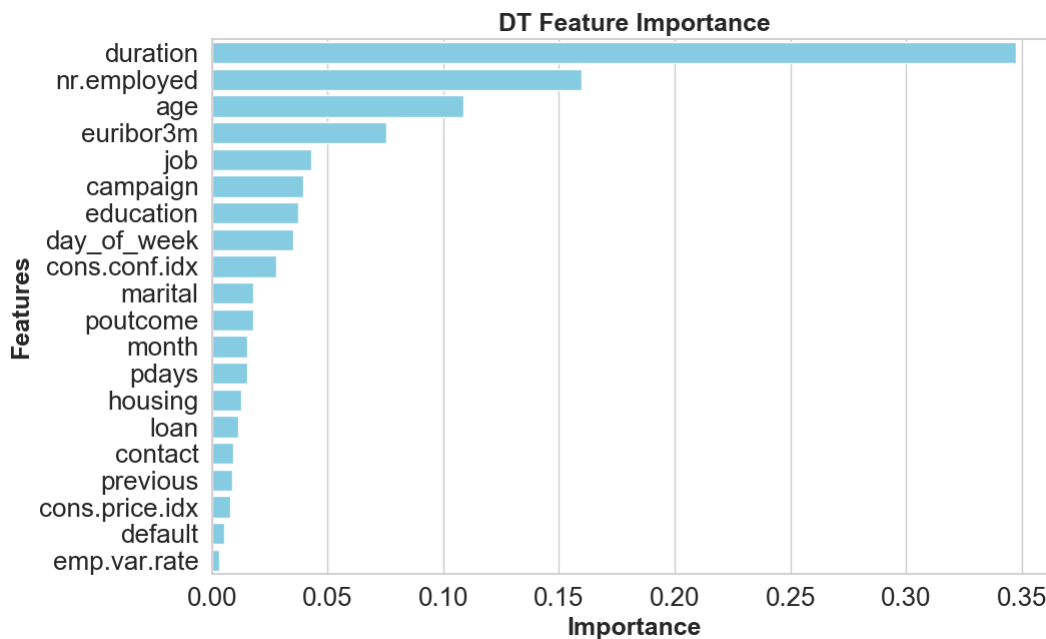
#### 4. Evaluation

Feature Importance refers to the influence value of the reaction features on the prediction of the model. It is a definite value and an interpretable method that is model independent.

Here, we use the PermutationImportance method in the Sklearn package to calculate the feature importance by randomizing all the features in the dataset. Then, we use Logistic Regression and KNN algorithm for training, and use PermutationImportance method to calculate the characteristic importance or contribution rate of these two algorithms to the target variable y respectively [9]. Then the feature importance weights are sorted respectively. If the feature weights of the two models are not different, it means that the feature shows a lower importance. If the feature weight difference between the two models is significant, it indicates that the feature is of high importance to the model (See Fig. 6 and Fig. 7).



**Fig. 6** RF Feature Importance



**Fig. 7** DT Feature Importance

Each feature has a weight value for  $y$ , so each feature has a certain degree of influence on whether customers subscribe for time deposits. Among them, the weight of the last contact duration of these two models is the largest, and the difference is significant, indicating that the last call duration has the greatest impact on whether customers subscribe for fixed deposits, and can be used as an important feature of  $y$ . There are also some features that have not obvious weight differences between the two models, such as the consumer confidence index (`cons.conf.idx`), which shows that the consumer confidence index shows little weight difference between different models, so the change of consumer confidence index has little impact on whether customers subscribe for deposits. Can be dropped as a non-essential feature [10].

Through the importance of features, we can easily see the weight value ranking of each feature, understand which features are more important, those are not important, can give a clear explanation of the weight value of each feature, arrange the weight value does not need to retrain the model, it will become simple and easy to understand.

However, the measure of feature importance is related to the error of the model, and different models will output different feature importance rankings, which will increase the randomness of the feature importance measure. In addition, when two features are associated, the feature importance ranking will not be reflected, on the contrary, it may be biased by untrue or abnormal data points, giving us wrong information. Therefore, it is not enough to discuss the impact of feature importance on prediction, and we need to explore further.

## 5. Conclusion

In recent years in the financial field, machine learning models have shown an increasingly large decision-making effect, even more than humans, due to the customer's data information is very large, it is necessary to use an interpretable machine learning model to process the data to make decisions and judgments, to combine big data and financial risk control models, and to really use the data and the model to solve the actual problems of finance and to give the data scientists a better decision-making feedback. In this paper, we take bank customer data as an example to predict whether a customer will subscribe to a time deposit or not and illustrate the interpretability of machine learning models at two different levels. The results highlight the usefulness of Logistic regression in analyzing the bank marketing activities.

## References

- [1] Industrial and Commercial Bank of China. Year-end bonus "spend more and more" guide. Modern Commercial Bank, 2022, (3): 70-72.
- [2] Jiang, X.W., & Yang, D.F. Purchase prediction study based on the Random Forest classification method. Jiangsu Communications, 2021, 37(06): 72-74.
- [3] Wang, S.B. Suggestions for improving the maintenance ability of private banking customers. Modern Finance, 2021, (5): 56.
- [4] Li, B. Analysis of customer loyalty and its influencing factors of modern commercial Banks. Volkswagen Technology, 2021, 23(3): 155-157.
- [5] Sun, Y. C. Research on customer value evaluation and customer classification of commercial banks. Commercial Economy, 2017, (12): 162-164.
- [6] Zhao, H. Application of customer information in customer value measurement of commercial banks — Based on the analysis of customer relationship life cycle. Modern Finance, 2017, (5): 19-21.
- [7] Nayak, B., Ojha, R.K., Subbarao, P.S., & Batth, V. Machine learning finance: Application of machine learning in collaborative filtering recommendation system for financial recommendations. Int J Recent Technol Eng, 2019, 8: 905-909.
- [8] Li Qing,Zhao Shuai,Zhao Shancheng & Wen Jinming. Logistic Regression Matching Pursuit algorithm for text classification. Knowledge-Based Systems, 2023.
- [9] Martín-Baos José Ángel,García-Ródenas Ricardo & Rodríguez-Benítez Luis. A Python package for performing penalized maximum likelihood estimation of conditional logit models using Kernel Logistic Regression. Transportation Research Procedia, 2021.
- [10] ZHANG, Xu, et al. Logistic regression with network structure. Statistica Sinica, 2020, 30(2): 673-693.