An Integrated Learning-Based Prediction Model for Purchasing Propensity of Jingdong Visitors

Yizheng Liu *, Hengkui Zhang, Hangjun Ren

School of Finance, University of International Business and Economics, Beijing, China, 100105

* Corresponding Author Email: lyz1230731@163.com

Abstract. The detection of purchasing behaviour of consumer customers belongs to the classical machine learning binary classification problem, which requires the use of a new dataset after downsampling to build a machine learning model and the selection of parameters for the visual display of the model.In this paper, we first performed data cleaning and pre-processing of the data, explored the complexity, non-linearity and other properties of the original data, performed logarithmic transformation and normalisation of the data pre-processing, and then selected the data features, established multiple machine learning models for modelling, adjusted the parameters using cross-validation and grid tuning, further optimised the models, calculated the evaluation metrics of the models, and compared the models, and finally used cross-validation and grid tuning to adjust parameters, further optimise the models, and calculate evaluation metrics of the models. The models were compared and finally the selected models were integrated using the Stacking method. Then the probability calibration was performed and the models were interpreted using SHAP values and PDP plots. This paper also adopts a rich visualisation approach for data and model presentation during data processing and modelling.

Keywords: Random Forest, GBDT, XGBoost, LightGBM, CatBoost.

1. Introduction

Jingdong is one of the largest comprehensive e-commerce platforms in China, offering a wide range of goods and services. On Jingdong, visitors can browse and purchase a wide range of products, including electronic devices, home furnishings, clothing, and food. Jingdong employs sophisticated algorithms to optimise user experience and sales effectiveness, one of which is the purchase model [1]. The steps to build the model are as follows: construct a new dataset using downsampling, further mine and analyse the new dataset, try multiple methods to adjust the model hyperparameters to find the optimal solution, and compare the performance of each model as well as draw feature importance maps [2], visualise the results, and calculate the probability of a purchase occurring for each sample in the test set [3]. We find that Random Forest, GBDT, XGBoost, LightGBM, and CatBoost, which have high performance on various performance metrics, are all integrated tree learners, which shows that integrated tree models are naturally suitable for this type of binary classification task. We use these five models as base learners and the multiple linear regression model as a meta-learner, and the Stacking method as a combining strategy to construct the integrated learning model [4-7]. The accuracy, Kappa coefficient, accuracy, completeness, F1 and AUC of the models on the test set are 0.99999, 0.99939, 0.99879, 1.0, 0.99939 and 0.99999, respectively, which are very close to the perfect classification model and can fully the classification model be very close to the perfect classification model and is able to find out the fraudulent samples completely [8]. This paper combines the idea of integrated learning by combining Random Forest, GBDT, LightGBM and CatBoost to form a learning model, validates the computational parameters of these models respectively, and then applies them to the customer consumption behaviour model in this paper, which provides theoretical support for the operation strategy of the large-scale online shopping platforms, and also provides reasonable decision-making suggestions for the consumer's consumption behaviour.

2. The fundamental of data analysis

The dataset used in this paper is the Jingdong Visitor Purchase Key Indicator dataset from the Kaggle website. The features contained in this dataset cover factors closely related to purchasing behaviour such as visitor type, web page value, month, etc. This paper uses python to empirically analyse this dataset, and successively explores the best algorithmic combinations of dataset sampling and classification model construction under the evaluation indexes of each classification model through data exploration, data preprocessing, feature engineering, model training, and model integration to more accurately predict the status of Jingdong visitors' purchasing behaviours. accurately predict the status of Jingdong visitors' purchasing behaviour. The dataset used in this modelling task comes from the telecom fraud data provided by an anonymous data collection agency and contains 14 features. Among them, three are continuous variables and the others are discrete variables after data processing. And "Whether or not to purchase is the target variable. The dataset contains 1,000,000 observations on the characteristics of e-commerce purchases of different users, and in the next section, we will explain the features of the data.

Although there are only 14 features present in the original data, there are a total of 1,000,000 observations, which makes it extremely difficult to analyse the data directly. This is because too many observations will contain more information, leading to a greater range of variation in each feature and reduced stability, which in turn reduces the instability and generalisation performance of the model. As shown in Table 1, the values of the first three continuous features fluctuate greatly, with the smallest less than a single digit and the largest exceeding five digits. The last five discrete features are all 0-1 variables, but there is a large imbalance problem. To deal with these problems, corresponding methods are needed.

Table. 1 Kolmogorov-Smirnov Test

diagnostic property	statistic	P-value
X_1	0.8532	<0.0001
X_2	0.5006	<0.0001
X_3	0.5217	<0.0001

The results of the Kolmogorov-Smirnov Test are shown in Table 1. X_1, X_2, X_3 represent different variable characteristics, respectively. Normal distribution is a highly symmetric ideal distribution with a series of excellent properties. The assumption of normality is present in most of the classical statistical models. Therefore, before statistical modelling, it is essential to test whether the data conforms to a normal distribution. In the case of large samples, we chose to use the Kolmogorov-Smirnov test. The results show that none of the three continuous variables satisfy the normal distribution and are far off. Therefore, statistical methods like Pearman's correlation coefficient, ANOVA and linear regression which require the assumption of normality cannot be used. In the previous subsection, it has been suggested that the raw data does not satisfy the normality assumption and Pearman correlation coefficient cannot be used. Therefore, we utilise the Spearman correlation coefficient as an indicator of the magnitude of correlation between features, which is calculated as shown in equation (1). In the process of observing 1 4 features, we found that the correlation between features is low. This may be due to the low dimensionality of the variables and the high independence of the data from the collection agency establishing the search. Too high correlation increases the variance and error of the model, which in turn affects the stability and robustness of the model. Obviously, the original dataset does not have such a problem, and there is no need to deal with the correlation of the data. At the same time, through the correlation, we can also see that the correlation between the values of the 15 independent variables and purchasing behaviour is poor, which is likely to indicate that the relationship between the independent variables and the dependent variable is a non-linear relationship that is difficult to show rather than a linear relationship that can be represented by a simple regression model. This reveals that we need to focus on machine learning and deep learning models when selecting models after constructing the feature variables [9].

Of the 1,000,000 observations in the dataset, only 87,403 purchases were recorded. The ratio of purchases to non-purchases is approximately 1:10.4, which suggests a more serious imbalance between the number of purchases and non-purchases in the dataset. Many models predict categories that are actually predictions of purchases. For example, the most classic Logistic regression model essentially predicts the probability that a sample is a positive example. If the probability is greater than 0.5, it is a positive example; if the probability is less than 0.5, it is considered a negative example. The same is true for multiclassification. In the case of a neural network used for 3-classification, for example, there are three values of the classification variable, 0, 1, and 2, and after one-hot coding, the probability of a sample belonging to one of the three categories is also predicted. When the data is unbalanced and the default threshold is not changed, the model will be overfit to the category with more data and will be more inclined to output such categories. Therefore, if the raw data are brought directly into the model, it may reduce the sensitivity of the model to a smaller number of samples, which in turn may seriously affect the classification performance of the model [10].

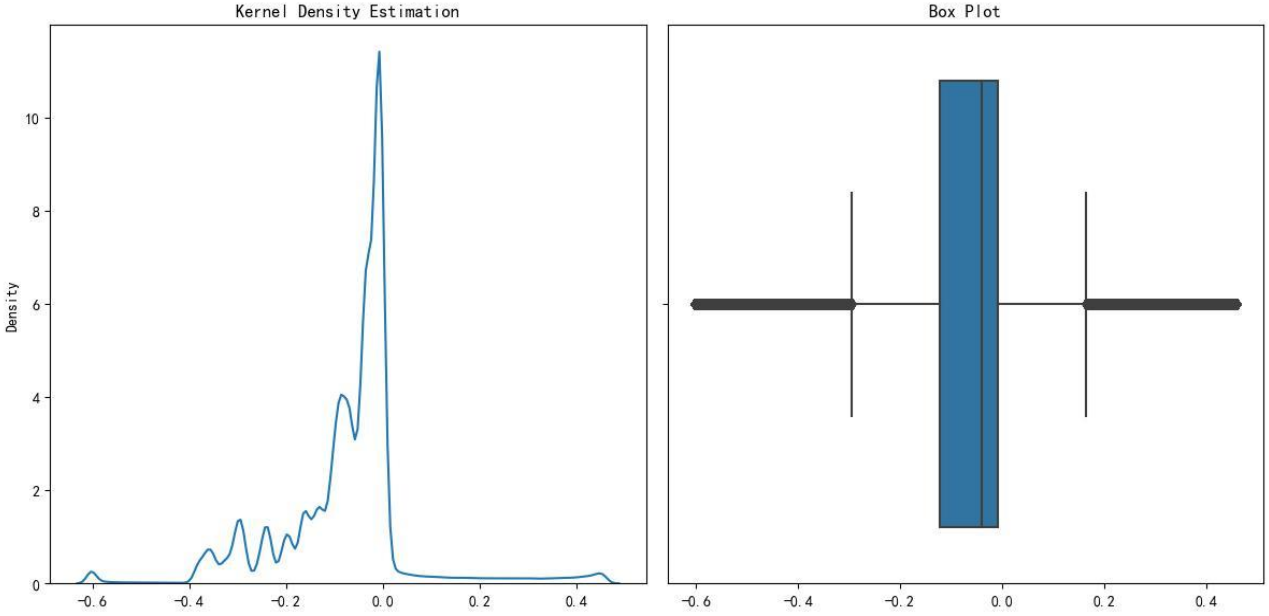

Figure. 1 Kernel density plots and box plots for variables before and after logarithmic transformation

Table. 2 Kolmogorov-Smirnov Test

diagnostic property	statistic	P-value
X_1	0.0006	0.7985
X_2	0.0006	0.9195
X_3	0.0005	0.9615

X_1, X_2, X_3 represent different variable characteristics, respectively. Anomalies in datasets generally occur in two ways. One, the anomaly is due to human loss, destruction or other human factors. The second is anomalies caused by excessive fluctuations in the actual values. The data in this paper comes from a data collection organisation, and under the premise of assuming that all the data are real and valid, it is unreasonable to delete the abnormal values directly, and it is easy to cause artificial loss and damage of information. Therefore, it is necessary to use reasonable methods to deal with abnormal values under the premise of retaining the data information is not lost. The kernel densities of the data samples studied in this paper are larger within the values of [0, 100], and most of the outliers are right-skewed severe points, which is also proved by the box plots. Doing local zoom-in and zoom-out on the three kernel density curves, it is found that the kernel density curves of the variables are very similar to the right-skewed normal distribution, which is further confirmed by the arrangement of the anomalies in the box plots. This suggests that the adjustments and corrections

made to the variables for the logarithmic treatment are likely to have some of the good characteristics of a normal distribution. The kernel density plots, and box plots are shown in Figure 1. In order to further verify the results of standardisation, Kolmogorov-Smirnov normality test was performed on the standardised variables, and the results are shown in Table 2. The original hypothesis of the Kolmogorov-Smirnov test is that "the data conforms to the normal distribution", and since the P-values are much larger than 0.05, it is not possible to reject the original hypothesis. 0.05, the original hypothesis cannot be rejected, and the variables are considered to be normally distributed.

It is important to note that in the actual modelling process, it is not possible to standardise all the data beforehand. This is because normalisation requires the mean and standard deviation of the data, which would result in information from the test set flowing into the model, which is not allowed.

3. Results

3.1. The establishment of data computation

In general, corrections to the sampling method can effectively address the problem of category imbalance. Class weighting: By assigning different weights to different classes, the model is made to focus more on a few classes of samples. This can be achieved by setting the algorithm parameters or using a loss function with class weights. Threshold shifting is an alternative to undersampling and oversampling when dealing with class imbalance. Assuming a dataset with 999 counterexamples and only 1 positive example, the model only needs to output all counterexamples to achieve 99.9% accuracy. This requires adjusting the model's probability thresholds for the categories. At the same time, different machine learning models are likely to learn different information from the data, so adjusting the threshold can also significantly improve classification performance.

Due to the non-linear nature of the data presented in this paper, classification of the data will likely require the use of machine learning and deep learning models. When developing such models on the training set, hyperparameters need to be given artificially. Models trained using different hyperparameters are likely to differ in their effectiveness on the validation set, and we definitely tend to choose the hyperparameters that perform better on the validation set. This process is also essentially a kind of learning, i.e., finding globally optimal or locally optimal parameter combinations in a wide parameter space. In general, a machine learning model will easily be overfit on the training set, and then quickly overfit on the validation set when artificially given hyperparameters that perform better on the validation set. This can lead to information about the validation set leaking to the model during the training of the model, and every time the hyperparameters are adjusted based on the model's performance on the validation set, this can lead to some information about the validation set leaking into the model's parameter tuning. Therefore, it is likely that the model's good performance on the validation set is simply due to human parameter tuning, but we are concerned with the model's performance on brand new data, not on the validation set. Therefore, we keep 840,000 pieces of data, a dataset that is itself extremely unbalanced and brand new to the model. Dividing the dataset into training, validation, and test sets provides a better measure of the model's generalisation performance.

We divide the dataset into two datasets containing 160,000 and 840,000 observations, respectively, and use the latter as the test set. For the first 160,000 data used for training and validation, we use the cross-validation method to evaluate the model performance.

The cross-validation method first divides the dataset into k mutually exclusive subsets of the same size, and then for each training session, k-1 subsets are selected as the training set, and the remaining subsets are used as the validation set. The final evaluation result is the average of the k test results. In order to get the evaluation results with better stability and fidelity, we take a value of 10 for k and use the 10-fold cross-validation method to evaluate the performance of the model.

Machine learning models and deep learning models often need to be artificially given hyperparameters in advance, and the choice of hyperparameters often has a large impact on the performance of the model. For example, neural networks generally need to be given hyperparameters such as the number of neurons in each layer and the learning rate, XGBoost needs to modify the L1

and L2 regularisation, and random forests need to be given the maximum depth of each tree and the number of trees. Generally speaking, the choice of hyperparameters is based on the experience of the researcher, but human beings inevitably make mistakes. Lattice search is the method born to deal with the choice of hyperparameters. Grid search is essentially an exhaustive search method. Among all the candidate parameter combinations, each parameter combination is brought into the model by loop traversal, and finally the optimal hyperparameters are obtained.

3.2. Analysis of experimental results

After finding the better parameters and training, the performance metrics of all models are summarised and the results are shown in Table 3. The Kappa coefficient is generally lower than the accuracy, while the other three performance metrics are similar. Moreover, the tree model is significantly better at classifying such problems, with a performance close to 1.0, which is very close to the perfect classification model.

Finally, the models were applied to the test set and ROC and P-R curves were plotted. Similar to the results on the validation set, the tree models Random Forest, GBDT, and XGBoost have clearly the best classification performance, followed by LightGBM and CatBoost.

Table. 3 Performance measures of different models under 10-fold cross-validation

model	accuracy	Kappa	F1	AUC
Logistic	0.8566	0.7645	0.8812	0.8775
KNN	0.9237	0.9726	0.9876	0.9877
Bayesian	0.8315	0.7145	0.8323	0.8565
SVC	0.9873	0.9435	0.9724	0.9828
decision tree	0.9740	0.9479	0.9739	0.9854
random forest	0.9999	0.9998	0.9999	0.9999
AdaBoost	0.9842	0.9683	0.9844	0.9845
GBDT	0.9999	0.9998	0.9999	0.9999
XGBoost	0.9999	0.9996	0.9997	0.9998
LightGBM	0.9998	0.9994	0.9999	0.9999
CatBoost	0.9998	0.9997	0.9998	0.9998
BP	0.9614	0.9247	0.9632	0.9723
TabNet	0.9624	0.9238	0.9631	0.9647
LCE	0.9999	0.9999	0.9998	0.9999

For classification models, the probability that a sample belongs to each category is generally calculated, and if the probability is greater than a certain threshold, it is considered to belong to the category corresponding to that probability. The general threshold is set to 0.5, and adjusting the threshold can lead to large changes in the model's predictions. In this regard, we draw the confusion matrix of all classification models and set the thresholds from 0.1 to 0.9 for 9 cases. Taking the XGBoost algorithm as an example, with the threshold value from small to large, the false positive cases predicted by the XGBoost type decreases, the false negative cases are always 0, and the true cases are always all correctly predicted.

Integration learning method is a method to build a strong learner by fusing several weak learners and taking the best of all. Combination strategy is a method to summarise and generalise the learning results of the base learner, in general, the combination strategies for regression and classification problems are averaging method and voting method. However, when there is more training data, there exists a more powerful method, i.e., the learning method. In this paper, we use the Stacking method to construct the integrated learning model. When using the Stacking method to construct an integrated model, it usually contains two layers, the first layer uses two or more models to predict the endpoints separately, called the base learner, and generally chooses non-linear models. The second layer has

only one model, which is responsible for fusing the predictions of the first layer models, and is called the meta-learner, and OLS is generally chosen as the model.

In McNemar's test, the performance of all 12 models differed significantly, so the five models with the best performance measures, Random Forest, GBDT, XGBoost, LigthGBM, and CatBoost, were chosen as the base learner, and multivariate linear regression was chosen as the meta-learner. After building the model, the prediction is re-run using the integrated model and the classification of the integrated learning is excellent and close to the perfect classification model.

4. Conclusions

In the previous section, a machine learning model was trained with the processed data to achieve near-perfect classification, and predictions were made on a test set using this model. However, this model is still a black-box model, which predicts that there is no precise knowledge of each feature of each sample, and similar patterns of how fraud occurs are not yet revealed. In this paper, we have developed a complete workflow with a high degree of freedom to support data scientists to fully observe and mine the data, arbitrarily try out various classical or state-of-the-art models and comprehensively evaluate their performance and decide on their own to integrate them in order to come up with a model that is best suited for the task. After refinement, this workflow can be used not only for e-commerce prediction, but also for risk identification of other financial behaviours, with probabilistic calibration. With probability calibration, it can even be used to perform expected risk measurement.

In addition, from the perspective of anti-telco fraud agencies, mobile communication companies and major banks, the model can be deployed in financial industry scenarios (online banking, ATM transactions) specifically for detecting credit card fraud, and in the future deployment, consideration can be given to simplifying the model to make the model's temporal and spatial complexity convenient for large-scale deployment. At the same time, if new data are constantly generated, they can be added to the model in reverse for further training, and the deployment can be considered to increase the interface for new data entry.

References

[1] Su Yuteng, Lv Shiyun, Xie Wenhan, Li Yuan, Ouyang Yixin, Xue Yongxi, Hu Meiling, Li Shuting, Zhou Hang, Liu Xiangtong. Analysis of risk factors for the development of type 2 diabetes mellitus based on LASSO regression and random forest algorithm [J]. Journal of Environmental Health,2023,13(07):485-495.

[2] KONG Yaqi, LIU Yu. Design and implementation of fitness counting system based on BlazePose and KNN [J]. Software Engineering,2023,26(07):58-62.

[3] JIA Ying,ZHAO Feng,LI Bo,GE Shiyu. Bayesian optimisation of XGBoost credit risk assessment model [J]. Computer Engineering and Applications:1-15.

[4] ZHOU Wentao, WEI Guangtao,WANG Zeli, ZHANG Xiaochen,REN Lizhi. A probabilistic short-term load forecasting method based on LightGBM at night economy user level [J]. Frontiers of Data and Computing Development,2023,5(03):160-168.

[5] Li Hongsheng,Li Kuniu,Wang Yang,Gao Fei,Zhang Yu,Xie Hongfu. Robust optimal scheduling model for grid-connected electric vehicle clusters based on SVC [J]. High Voltage Technology:1-9.

[6] Zheng Lijia,Song Bing. Pre-pruning and optimisation of decision tree classification algorithms [J]. Automation Instrumentation,2023,44(05):56-62.

[7] HOU Tianbao,WANG Aiyin. Personal credit assessment based on Stacking feature-enhanced multi-granularity cascade logistic [J]. Journal of Henan Normal University (Natural Science Edition), 2023,51(03):111-122.

[8] JIN Cui, LIU Yang,LI Qi,ZHAO Molin,MO Xianyao,WANG Ying. CatBoost-based arcing fault identification method for commonly used electrical loads [J]. Electrical Measurement and Instrumentation,2023,60(07):193-200.

[9] ZHAO Miao, WANG Xiaolei, ZHU Liwen, SHEN Jie, ZHANG Jian. Research on user repurchase behaviour prediction and marketing strategy scheme based on CatBoost-RBF fusion algorithm [J]. Inner Mongolia Science and Economy,2023, (07):74-77+90.

[10] Wang Shaoqing. Research on hotel pricing under online tourism business model [D]. Beijing university of industry and commerce, 2019.