

Subscribe to DeepL Pro to edit this document.  
Visit [www.DeepL.com/pro](https://www.deepl.com/pro?cta=edit-document) for more information.

**internship report**

June 6, 2024

**catalogs**

[I. Data preparation 3](#_Toc12300)

[II. Data processing 4](#_Toc29482)

[(i) Sample definitions 4](#_Toc30846)

[(ii) Data pre-processing 4](#_Toc29481)

[(iii) Feature engineering 4](#_Toc16164)

[(iv) Characterization screening 5](#_Toc21355)

[III. Model construction 6](#_Toc26974)

[(i) Modeling dataset slicing and dicing 6](#_Toc15582)

[(ii) Modeling training 6](#_Toc6468)

[IV. Analysis of model results 7](#_Toc31377)

[(i) Model prediction effects 7](#_Toc16791)

[(ii) Model interpretability 11](#_Toc109)

[(iii) Model stability 18](#_Toc18981)

[V. Model Strategy Analysis 18](#_Toc21636)

[VI. Appendices 20](#_Toc20550)

[(i) Evaluation indicators 20](#_Toc22546)

# I. Data preparation

This modeling makes a prediction of whether a municipal bond will default or not, which is a classification problem. Using the client's data related to the year, to predict whether the client year whether to default or not. The data source is based on wind data, extracting the national data from 2018 to 2021, and the extracted data range is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Training Set & Test Set** | **X** | 2018 | 2019 |
| **Y** | 2018 | 2019 |
| **out-of-time validation set** | **X** | 2020 |  |
| **Y** | 2020 |  |

The amount of data totaled 2793 entries, and a total of 417 features were extracted as shown in the table below. The total number of missing fields is 240, accounting for 8.59% of the total number of features, and the data quality is relatively good for subsequent modeling.

|  |  |  |
| --- | --- | --- |
| **Characteristic dimensions** | **give an example** | **data source table** |
| Regional economic finance | Last 3 years + Regional Economy\_Gross Domestic Product (GDP) ($) + Regional Finance\_Tax Revenue ($) + Regional Spread\_1-year AAA + Regional Loans\_Regional Bank Loan Balance ($ billion) | . /Town Investment Modeling Training Data/Regional Economic and Fiscal Data.xlsx sheet1, . /Town Investment Modeling Training Data/Regional List.xlsx, . /TPIT Model Training Data/Regional Changes Table.xlsx, . /TPIC Model Training Data/Regional Economic Data.xlsx, . /TPIC Model Training Data/Regional Fiscal Data.xlsx, . /TPIC Model Training Data/Regional Spreads.xlsx |
| Land area data | Last 3 years + Regional Economy\_Gross Domestic Product (GDP) ($) + Regional Finance\_General Public Budget Revenue ($) + Regional Loans\_Regional Bank Loan Balance ($bn) + Abortive Auction Ratio (Abortive Auction/(Sold + Abortive Auction)) | /Town Modeling Training Data/Land\_Data\_20220106.xlsx |
| Financial data | Last 3 years + Per Share Metrics\_EBITDA per share + Capital Structure\_Capital Fixation Ratio + Operational Capacity\_Accounts Payable Turnover Days + Solvency\_Equity Attributable to Shareholders of the Parent Company / Total Liabilities + Profitability\_Return on Personnel Input (ROP) | . All xlsx files in the folder / TPD Model Training Data / 20220211 Additional Financial Data |
| Financially derived data | Last 3 Years + Fiscal Derivation\_Long-Term Asset Suitability Ratio + Fiscal Derivation\_Year-over-Year Net Asset Growth Rate + Fiscal Derivation\_Net Income | Data corresponding to different years . /City Investment Modeling Training Data/Financial Report Data/{year} financial indicators of city investment organizations.csv |

# II. Data processing

## Sample Definitions

1. define

After communicating with the demand side, this project uses the For: whether the municipal bonds are in default

## Data preprocessing

An exploratory data analysis of the X features was conducted to guide the data processing methods. The analysis methodology consisted of descriptive statistical analysis (mean, maximum, standard deviation) and visual analysis to view the distribution of discrete and continuous features and to understand the missing, 0-value, and outlier rates for each feature.

Missing value filling in conjunction with business implications and data processing:

Fill the column Pre-Change Area Code with -9999999 for missing values and convert to int. 18, 19, and 20 Year Characteristics missing values are filled with -9999999 .

## feature engineering

Feature engineering consists of two main aspects, which are feature processing methods and feature derivation methods:

Feature processing methods:

*Growth of general budget revenues of* regional finances *=* Growth rate of general public budget revenues of *regional finances* (%) \* General public budget revenues of regional finances \_ ($)

*Share of general budget revenues of* regional finances *(%) = 100 \** General public budget revenues of *regional finances* ($) / Revenues of regional finances\_ ($)

Regional Fiscal *Fund Revenue Share (%) = 100 \* Regional Fiscal* Governmental Fund Revenue ($) / Regional Fiscal \_ Revenue ($)

Characteristic derivation method: the method of derivation cycle used and the aggregation method of characteristic derivation into multiple periods, the derivation cycle used is 3, 6, 12 months, and the derivation method samples the average value, summation, maximum value, growth value, and growth ratio.

## Feature Screening

The feature screening method is a method to find out and eliminate the features that are useless or even negative or unstable in the model, in order to improve the predictive effect of the model, improve the interpretability of the model and reduce the computational cost of the model. At the same time, the regression also needs to consider the multicollinearity between the features, so it is necessary to screen out the features with predictive ability and appropriate number of features before entering the model.

The total number of unscreened features is 417, and the total number of screened features suitable for modeling is 30, with the following screening logic:

1. Remove poor quality features: remove features that are not suitable for modeling, such as original date and customer number; remove features with a missing rate greater than 50% ;
2. Eliminate features with poor discriminatory power: Calculate the IV value of each feature and eliminate features with IV<0.1;
3. Excluding features with weak importance: the XGBOOST model was chosen to calculate the importance of each feature, and features with an importance ranking >60 were excluded;
4. Eliminate features with strong correlation: calculate the correlation coefficient between features and eliminate features with correlation coefficient greater than 0.95.

The final molded 30 feature names are shown below:

|  |  |
| --- | --- |
| **Incoming feature dimensions** | **Incoming feature name** |
| Regional economic finance | Regional Economy\_Gross Domestic Product (GDP) ($), Regional Loans\_Regional Bank Loan Balance ($ billion), Regional Finance\_General Budget Revenue Share (%), Regional Economy\_Gross Domestic Product Year-on-Year Growth (%), Regional Finance\_Fiscal Revenue ($) |
| Land area data | Transaction\_Average price of transacted land (yuan per square meter), Regional finance\_Government debt ratio (%), Regional finance\_Fiscal self-sufficiency rate (%), Regional finance\_Surviving municipal debt size/general budget revenue, Regional spread\_1-year unlimited rating (%), Regional finance\_Local government debt ratio (%), Regional finance\_General budget revenue growth, Regional finance\_Funds revenue share (%) |
| Financial data | Asset Size Data\_Total Assets, Solvency\_Money Funds/Short-Term Debt, Solvency\_Cash Ratio, Per Share Metrics\_Net Cash Flow Per Share, Solvency\_Tangible Net Worth Debt Ratio, Growth\_Compound Annual Growth Rate of Net Income, Growth\_Operating Profit (N-Year, Growth), Growth\_Equity Attributable to Shareholders of the Parent Company (Relative to Year-to-Start Growth Rate) |
| Financially derived data | Fiscal Derivation\_Net Assets (i.e. Owners' Equity), Fiscal Derivation\_Current Assets Fiscal Derivation\_Fixed Assets, Fiscal Derivation\_Net Income, Fiscal Derivation\_Other Payables/Current Liabilities, Fiscal Derivation\_Other Receivables/Total Assets, Fiscal Derivation\_Intangible Assets, Fiscal Derivation\_Inventory Turnover Days, Fiscal Derivation\_Other Payables/Current Assets |

# III. Model construction

## Modeling dataset slicing

In order to train the model scientifically and effectively, the dataset is sliced according to the year, where training set:test set=7:3, and the specific slicing results are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data use** | **Type of data set** | **X data time** | **Y data time** | **sample size** |
| model training | training set | 2018, 2019 | 2019 | 2424 |
| test set | 2018, 2019 | 2019 | 1039 |
| Model Effect Verification | Out-of-time sample validation set | 2020 | 2020 | 2792 |

## Modeling Training

The CatBoost model was chosen as the best modeling tool for this project due to its good training speed, generalization ability, and scientific treatment of category-based variables based on target variables. The prediction principle is based on GBDT, which minimizes the loss function by iteratively constructing a decision tree, and each moment a new tree is constructed to correct the prediction error of the previous tree.

The model parameters are categorized into hyper-parameters and internal parameters, where the hyper-parameters need to be set manually by using the grid search method and introducing Cross-Validation to reduce overfitting. Grid search means enumerating all possible combinations of hyperparameters and arriving at the best combination of parameters. The results are as follows:

|  |  |  |
| --- | --- | --- |
| **hyperparameterization** | **Parameter Meaning** | **optimal parameter** |
| iterations | Maximum number of iterations/tree/boosting rounds | 300 |
| depth | Maximum tree depth | 4 |
| l2\_leaf\_reg | L2 regularity coefficient | 3 |
| learning\_rate | learning rate | 0.01 |

The remaining internal parameters of the model are as follows:

|  |  |  |
| --- | --- | --- |
| **internal parameter** | **Parameter Meaning** | **default parameter** |
| random\_state | random seed | 0 |
| scale\_pos\_weight | quality-ratio | 2.618599791013584 |
| loss\_function | loss function | *CrossEntropy* |

# Model Analysis of results

The model results are analyzed in three dimensions, namely, prediction effect, interpretability and stability. The model prediction effect is composed of six-dimensional evaluation indexes, KS and ROC curves, and distribution of prediction results; the model interpretability is analyzed around the analysis of the incoming model, which is composed of feature contribution, multi-indicator analysis of feature contributions, analysis of the ordering of feature values, and univariate analysis; and the model stability is composed of PSI.

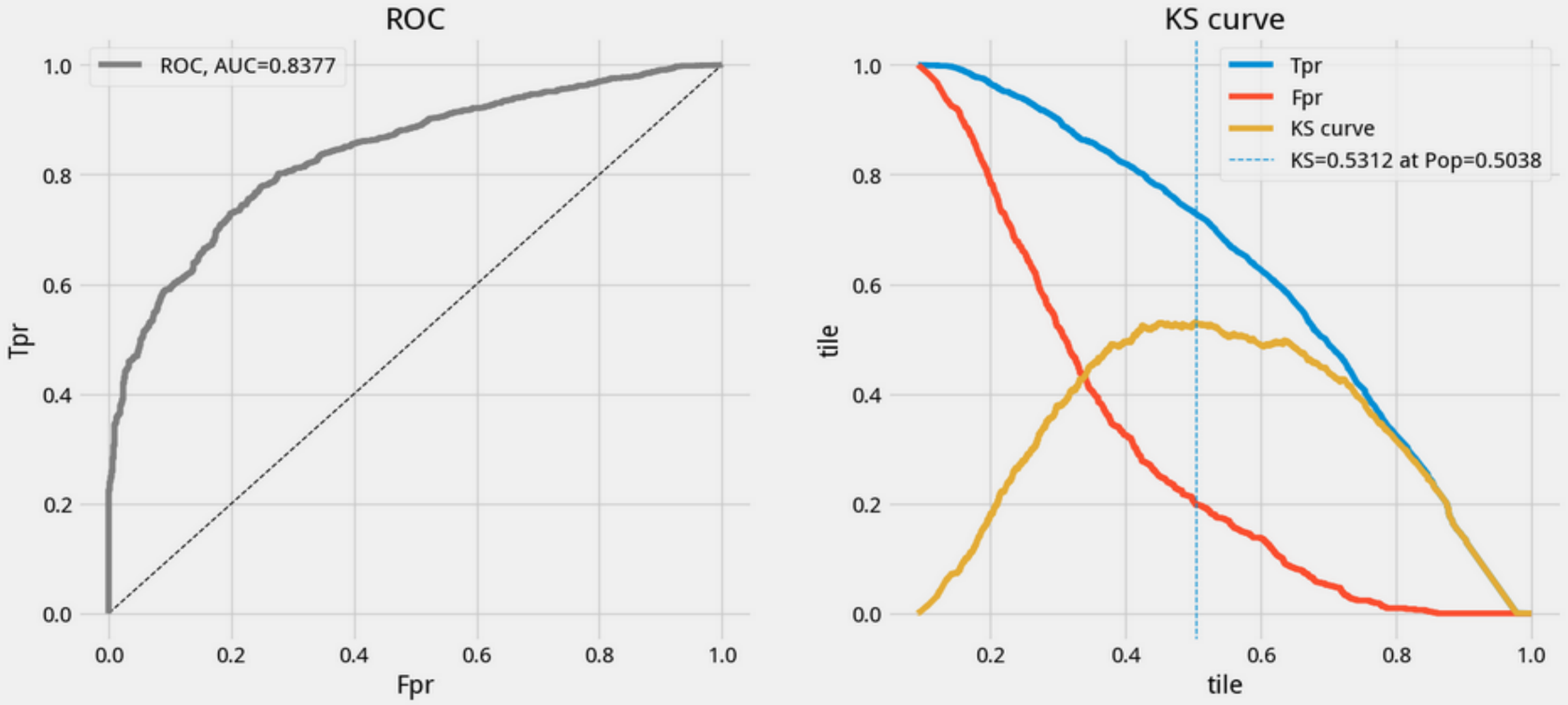
## Model Prediction Effect

1. Six-dimensional evaluation indicators

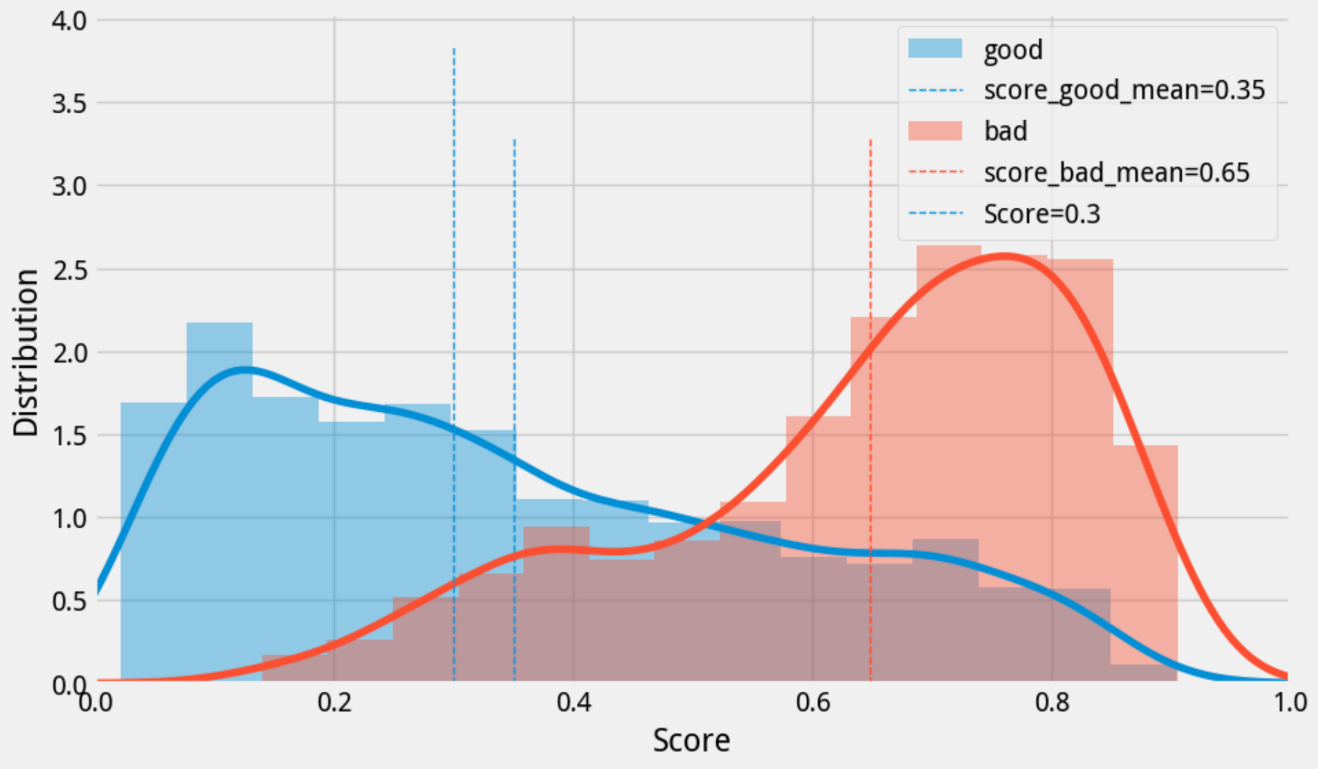
Using the above parameters for model prediction, the model prediction effect evaluation indexes obtained are listed in the following table, the model finally achieved a classification level of KS about 0.53 and AUC 0.84 on the test set, which is a good performance of the model; and the difference between the six-dimensional indexes of the model on the test set and the validation set is very small, and the model did not produce overfitting phenomenon.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Training** | **Validation** | **Testing** | **Validation - Training** | **Testing - Training** | **Testing - Validation** |
| **Accuracy** | 0.776815 | 0.769971 | 0.746777 | -0.006844 | -0.030039 | -0.023195 |
| **Precision** | 0.565615 | 0.558252 | 0.467652 | -0.007363 | -0.097963 | -0.0906 |
| **Recall** | 0.829851 | 0.801394 | 0.794349 | -0.028457 | -0.035502 | -0.007045 |
| **F1** | 0.672716 | 0.658083 | 0.588714 | -0.014633 | -0.084002 | -0.069369 |
| **AUC** | 0.881921 | 0.849454 | 0.837653 | -0.032467 | -0.044268 | -0.011801 |
| **KS** | 0.589898 | 0.571795 | 0.5312 | -0.018103 | -0.058698 | -0.040594 |

1. KS vs. ROC curve

Taking the October out-of-time sample validation set as an example, it can be seen from the figure that the ROC curve bulges towards the upper left corner and separates significantly from the diagonal dashed line. The yellow KS curve peaks at 0.5038 on the x-axis, where the difference between FPR and TPR reaches a maximum, and this maximum KS is 0.5312.

1. Distribution of forecast results
   1. Overall model score distribution

Taking the 2020 out-of-time sample validation set as an example, the probability distribution of predicting good and "bad" customers is shown in the following figure, which shows that the model is well differentiated.

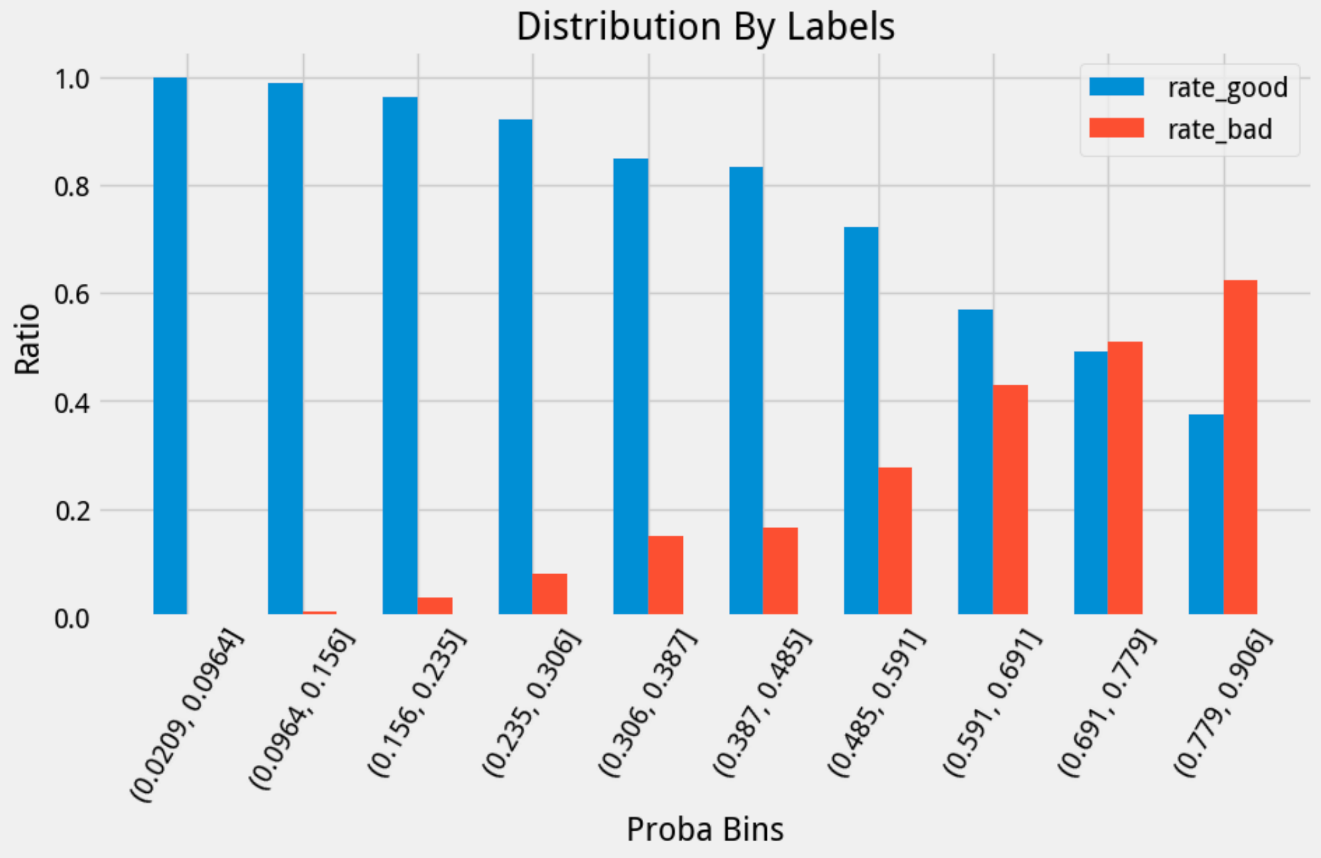
* 1. Interval model subdistribution

The perspective of the final predictive ability of the model is an extremely important metric in addition to the model classification ability KS metric, model sortability. This metric shows whether the model has business logic, i.e., the higher the model score the higher (or lower) the customer risk, the lower the model score the lower (or higher) the customer risk.

It is expected that the model sorting is the higher the model score, the better the customer. Take the 2020 out-of-time sample validation set as an example to analyze, the actual sortability is shown in the figure below, it can be seen that, as the probability of default becomes larger "bad" customers account for a larger proportion of good customers account for a smaller proportion of the final sortability of the model is in line with the expected design, and therefore it can be considered that there is no abnormality in the model in terms of the sortability.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **interval (math.)** | **interval (math.) bad** | **Total number of people between zones** | **Interval good percentage** | **Interval Bad Percentage** | **Cumulative total** | **Cumulative headcount as a percentage of** | **Cumulative good percentage** | **Cumulative Bad Percentage** | **Current interval bad percentage** |
| **0.95-1.0** | 0 | 0 | 0 | 0.00% | 0.00% | 2,792 | 100.00% | 100.00% | 100.00% | 22.82% |
| **0.9-0.95** | 1 | 5 | 6 | 16.67% | 83.33% | 2,792 | 100.00% | 100.00% | 100.00% | 22.82% |
| **0.85-0.9** | 11 | 46 | 57 | 19.30% | 80.70% | 2,786 | 99.79% | 99.95% | 99.22% | 22.68% |
| **0.8-0.85** | 62 | 84 | 146 | 42.47% | 57.53% | 2,729 | 97.74% | 99.44% | 91.99% | 21.47% |
| **0.75-0.8** | 60 | 82 | 142 | 42.25% | 57.75% | 2,583 | 92.51% | 96.57% | 78.81% | 19.43% |
| **0.7-0.75** | 80 | 87 | 167 | 47.90% | 52.10% | 2,441 | 87.43% | 93.78% | 65.93% | 17.21% |
| **0.65-0.7** | 90 | 74 | 164 | 54.88% | 45.12% | 2,274 | 81.45% | 90.07% | 52.28% | 14.64% |
| **0.6-0.65** | 84 | 52 | 136 | 61.76% | 38.24% | 2,110 | 75.57% | 85.89% | 40.66% | 12.27% |
| **0.55-0.6** | 89 | 47 | 136 | 65.44% | 34.56% | 1,974 | 70.70% | 82.00% | 32.50% | 10.49% |
| **0.5-0.55** | 99 | 29 | 128 | 77.34% | 22.66% | 1,838 | 65.83% | 77.87% | 25.12% | 8.71% |
| **0.45-0.5** | 121 | 23 | 144 | 84.03% | 15.97% | 1,710 | 61.25% | 73.27% | 20.57% | 7.66% |
| **0.4-0.45** | 109 | 21 | 130 | 83.85% | 16.15% | 1,566 | 56.09% | 67.66% | 16.95% | 6.90% |
| **0.35-0.4** | 127 | 35 | 162 | 78.40% | 21.60% | 1,436 | 51.43% | 62.60% | 13.66% | 6.06% |
| **0.3-0.35** | 164 | 19 | 183 | 89.62% | 10.38% | 1,274 | 45.63% | 56.71% | 8.16% | 4.08% |
| **0.25-0.3** | 172 | 18 | 190 | 90.53% | 9.47% | 1,091 | 39.08% | 49.10% | 5.18% | 3.02% |
| **0.2-0.25** | 184 | 9 | 193 | 95.34% | 4.66% | 901 | 32.27% | 41.11% | 2.35% | 1.66% |
| **0.15-0.2** | 171 | 4 | 175 | 97.71% | 2.29% | 708 | 25.36% | 32.58% | 0.94% | 0.85% |
| **0.1-0.15** | 235 | 2 | 237 | 99.16% | 0.84% | 533 | 19.09% | 24.64% | 0.31% | 0.38% |
| **0.05-0.1** | 199 | 0 | 199 | 100.00% | 0.00% | 296 | 10.60% | 13.74% | 0.00% | 0.00% |
| **0.0-0.05** | 97 | 0 | 97 | 100.00% | 0.00% | 97 | 3.47% | 4.50% | 0.00% | 0.00% |

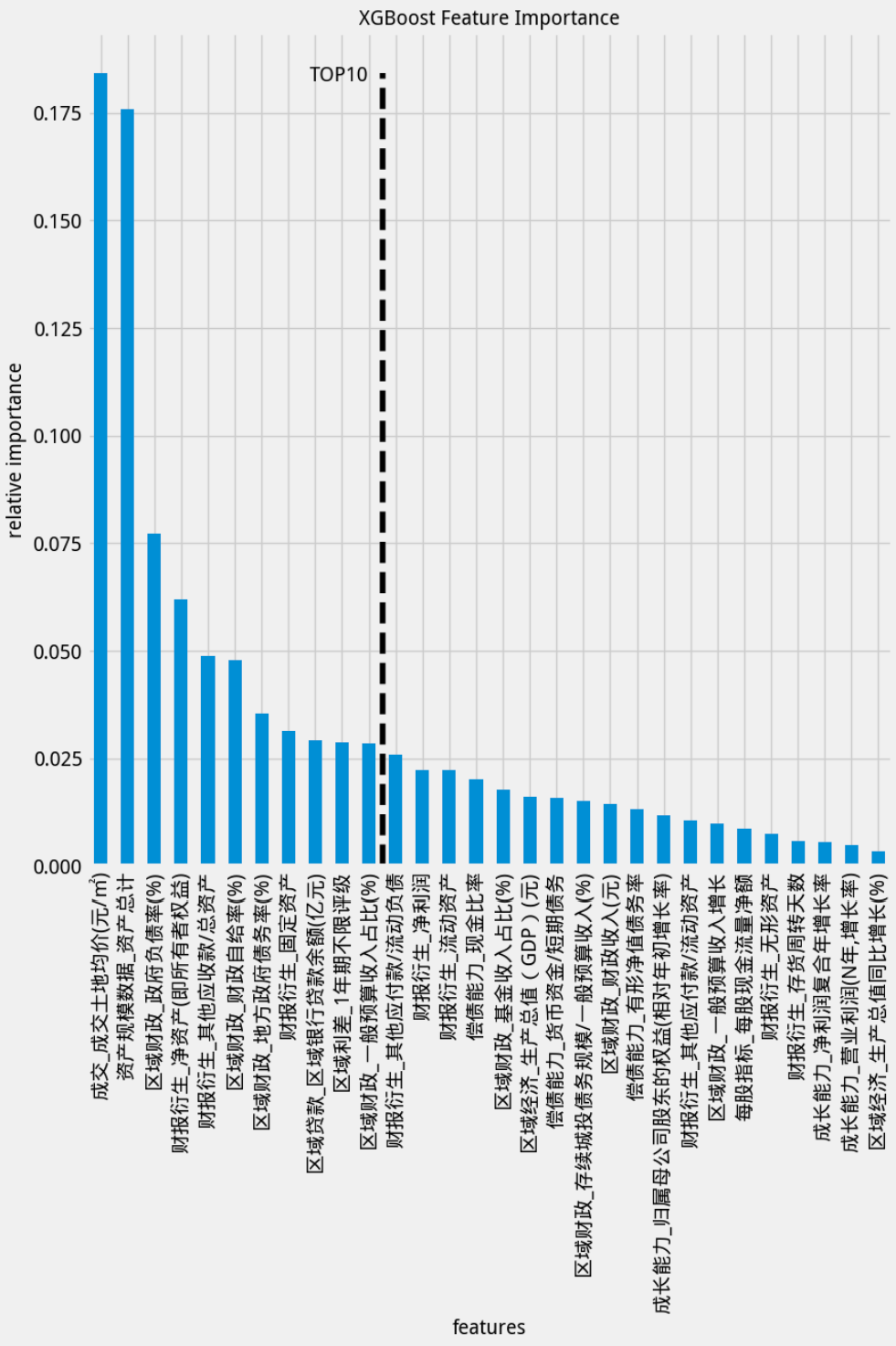
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **bins** | **interval (math.) bad** | **Total number of people between zones** | **Interval Bad Percentage** | **Cumulative total** | **Cumulative headcount as a percentage of** | **Recall** | **Precision** |
| **0** | (0.779,0.906] | 175 | 280 | 62.50% | 280 | 10.03% | 27.47% | 62.50% |
| **1** | (0.691,0.779] | 142 | 279 | 50.90% | 559 | 20.02% | 49.76% | 56.71% |
| **2** | (0.591,0.691] | 120 | 279 | 43.01% | 838 | 30.01% | 68.60% | 52.15% |
| **3** | (0.485,0.591] | 77 | 279 | 27.60% | 1117 | 40.01% | 80.69% | 46.02% |
| **4** | (0.387,0.485] | 46 | 279 | 16.49% | 1396 | 50.00% | 87.91% | 40.11% |
| **5** | (0.306,0.387] | 42 | 279 | 15.05% | 1675 | 59.99% | 94.51% | 35.94% |
| **6** | (0.235,0.306] | 22 | 279 | 7.89% | 1954 | 69.99% | 97.96% | 31.93% |
| **7** | (0.156,0.235] | 10 | 279 | 3.58% | 2233 | 79.98% | 99.53% | 28.39% |
| **8** | (0.0964,0.156] | 3 | 275 | 1.09% | 2508 | 89.83% | 100.00% | 25.40% |
| **9** | (0.0209,0.0964] | 0 | 284 | 0.00% | 2792 | 100.00% | 100.00% | 22.82% |

Equally-scored bin interval plots also illustrate the model's ability to differentiate, and ordering is strong:

## Model Explanatory

1. Characteristic contribution

The importance ranking of the final 30 features in the model is shown below. From the technical point of view, the importance ranking of features decreases gently, and there is no heavy dependence of the model on specific features or steep decrease in feature importance, which ensures the stability of the subsequent production effect. From the business perspective, the modeled variables are all consistent with business experience and can be explained. Therefore, it can be considered that there is no abnormality in the modeling features.



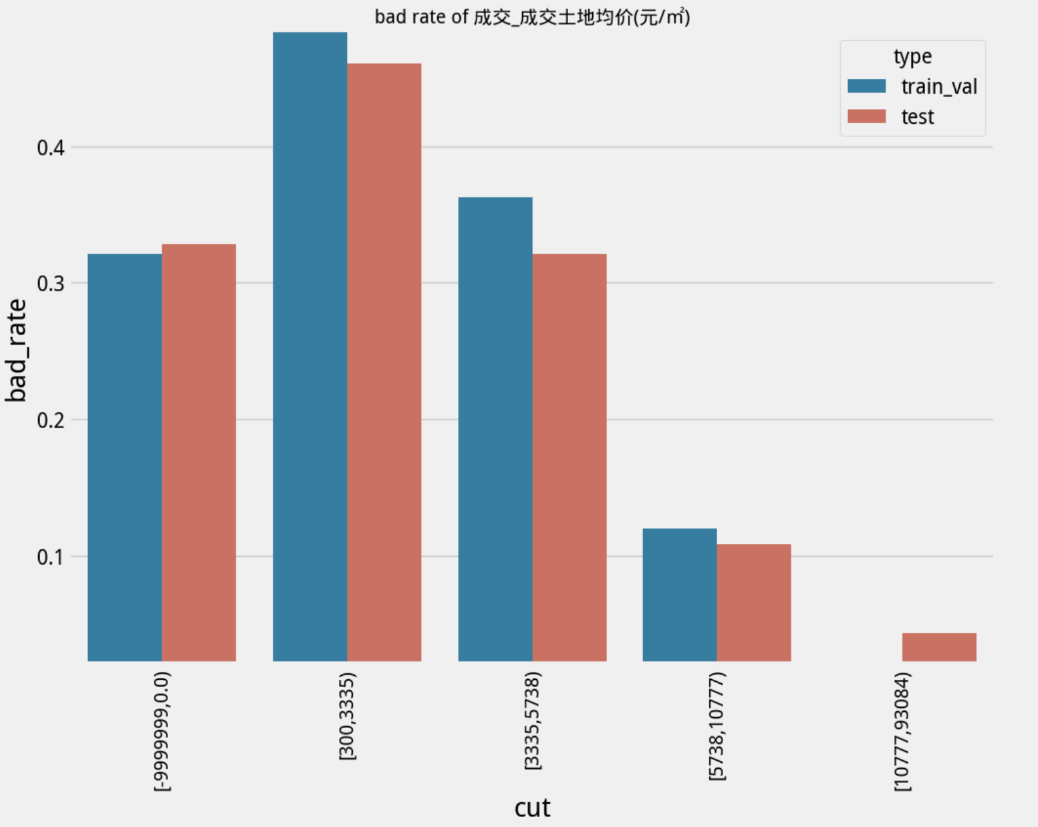
1. Multi-Indicator Analysis of Characteristic Contributions

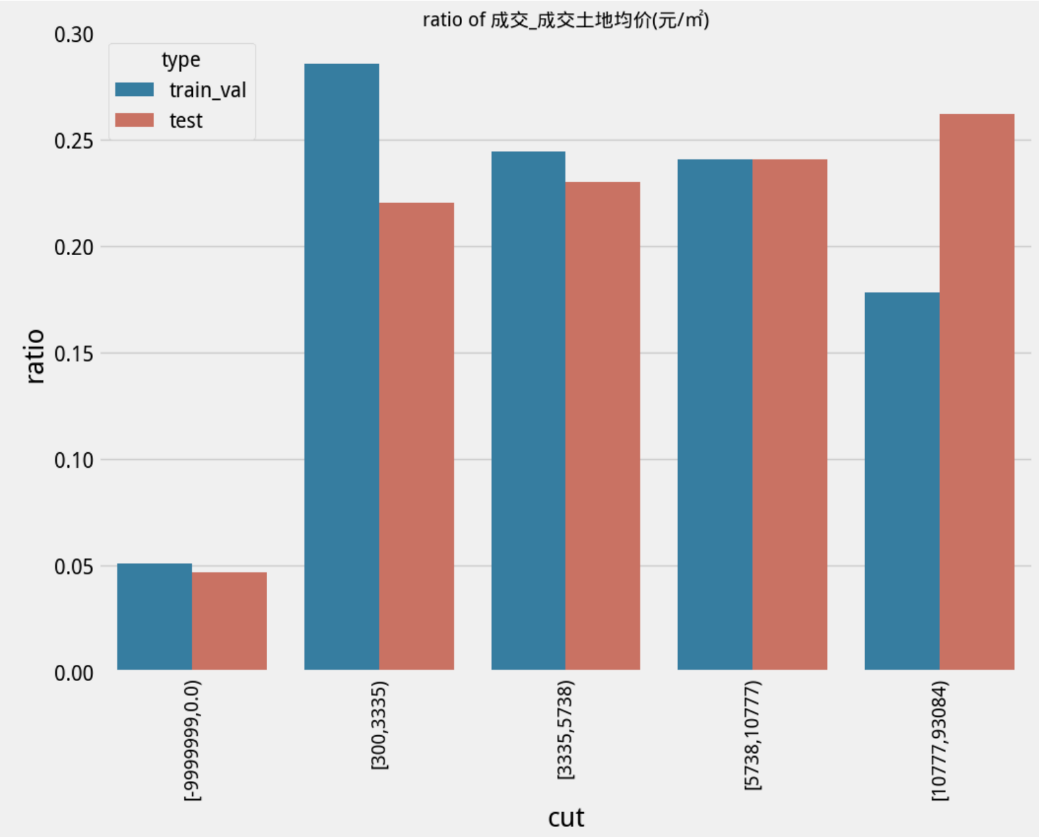
As can be seen from the MCA plot of the incoming features, AUC, IV, and Importance all show high levels.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **features** | **iv** | **importance** | **KS** | **AUC** |
| **0** | Sold\_Average price of land sold (yuan/m2) | 1.02 | 0.18 | 0.44 | 0.77 |
| **8** | Asset Size Data\_Total Assets | 0.44 | 0.18 | 0.24 | 0.64 |
| **10** | Regional fiscal\_government debt ratio (%) | 0.38 | 0.08 | 0.35 | 0.70 |
| **11** | Financial Reporting Derivation\_Net Assets (i.e. Owners' Equity) | 0.30 | 0.06 | 0.21 | 0.61 |
| **19** | Financial Report Derivation\_Other Receivables/Total Assets | 0.09 | 0.05 | 0.13 | 0.57 |
| **1** | Regional financial \_ financial self-sufficiency (%) | 0.94 | 0.05 | 0.40 | 0.76 |
| **12** | Regional Finance\_Local Government Debt Ratio (%) | 0.27 | 0.04 | 0.32 | 0.69 |
| **15** | Derivation of Financial Reports\_Fixed Assets | 0.23 | 0.03 | 0.17 | 0.59 |
| **2** | Regional Loans\_Regional Bank Loan Balances (Billions of dollars) | 0.66 | 0.03 | 0.38 | 0.72 |
| **5** | Regional Spreads\_1 Year Unlimited Rating | 0.59 | 0.03 | 0.23 | 0.52 |
| **7** | Regional Finance\_General Budget Revenue Share (%) | 0.48 | 0.03 | 0.33 | 0.71 |
| **13** | Derivation of financial statements\_Other payables/current liabilities | 0.23 | 0.03 | 0.19 | 0.61 |
| **16** | Derivation of Financial Report\_Net Income | 0.22 | 0.02 | 0.16 | 0.57 |
| **14** | Financial Report Derivation\_Current Assets | 0.23 | 0.02 | 0.17 | 0.59 |
| **18** | Solvency\_Cash Ratio | 0.11 | 0.02 | 0.20 | 0.62 |
| **9** | Regional fiscal\_fund revenue share (%) | 0.43 | 0.02 | 0.31 | 0.69 |
| **3** | Regional economic\_gross domestic product (GDP) ($) | 0.61 | 0.02 | 0.36 | 0.72 |
| **26** | Solvency\_Money funds/short-term debt | 0.06 | 0.02 | 0.15 | 0.57 |
| **28** | Regional finance\_size of surviving municipal debt/general budget revenue (%) | 0.02 | 0.02 | 0.09 | 0.50 |
| **4** | Regional Finance\_Revenue ($) | 0.59 | 0.01 | 0.35 | 0.71 |
| **17** | Solvency\_tangible net worth debt ratio | 0.15 | 0.01 | 0.14 | 0.59 |
| **22** | Growth capacity\_Equity attributable to shareholders of the parent company (growth rate relative to the beginning of the year) | 0.08 | 0.01 | 0.15 | 0.56 |
| **29** | Financial Reporting Derivation\_Other Payables/Current Assets | 0.02 | 0.01 | 0.09 | 0.55 |
| **6** | Regional finance\_General budget revenue growth | 0.50 | 0.01 | 0.30 | 0.65 |
| **25** | Per Share Indicators\_Net Cash Flow per Share | 0.07 | 0.01 | 0.12 | 0.55 |
| **21** | Derivation of Financial Reports\_Intangible Assets | 0.09 | 0.01 | 0.11 | 0.56 |
| **24** | Financial Report Derivation\_Inventory Turnover Days | 0.08 | 0.01 | 0.14 | 0.57 |
| **23** | Growth Capacity\_Compound Annual Growth Rate of Net Profit | 0.08 | 0.01 | 0.12 | 0.52 |
| **27** | Growth Capacity\_Operating Profit (Year N, Growth Rate) | 0.04 | 0.00 | 0.06 | 0.51 |
| **20** | Regional economy\_GDP growth (%) | 0.09 | 0.00 | 0.11 | 0.54 |

1. Univariate analysis of mode entry
2. Due to the large number of variables in the modeling, this section analyzes only the more important and clearly ranked features from the different feature dimensions. Observe the land area data, the financial data dimension, and the regional economy and finance dimension with the the trend of change. Observe whether the trends of the predicted and real values are consistent, and the impact of each feature on whether default is predicted.
3. "Land area data" dimension

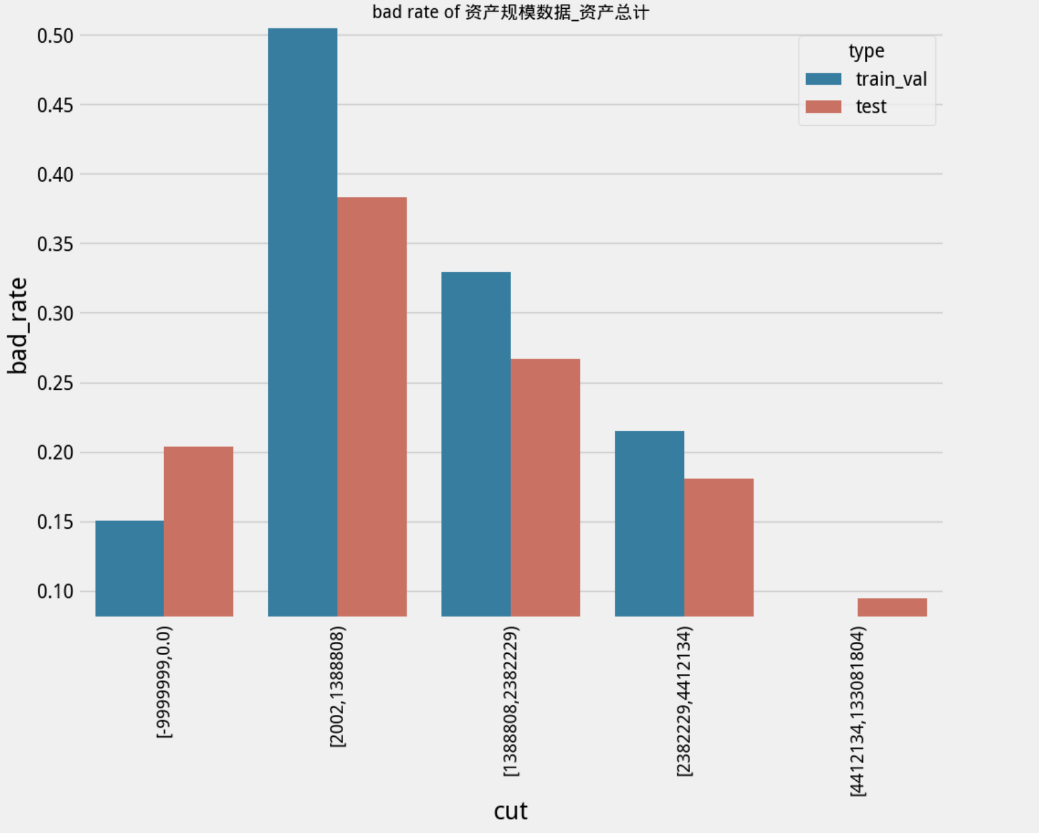
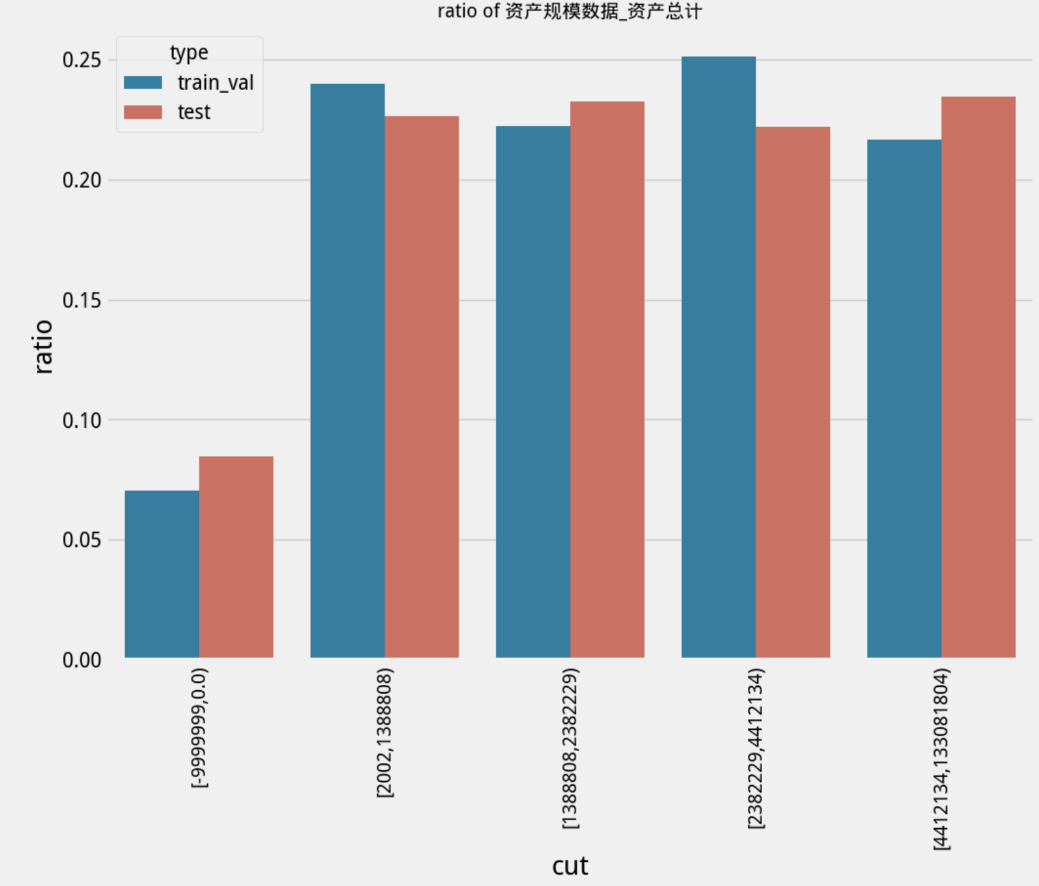
"Transaction\_Transaction\_Average\_Price\_of\_Land" feature shows that



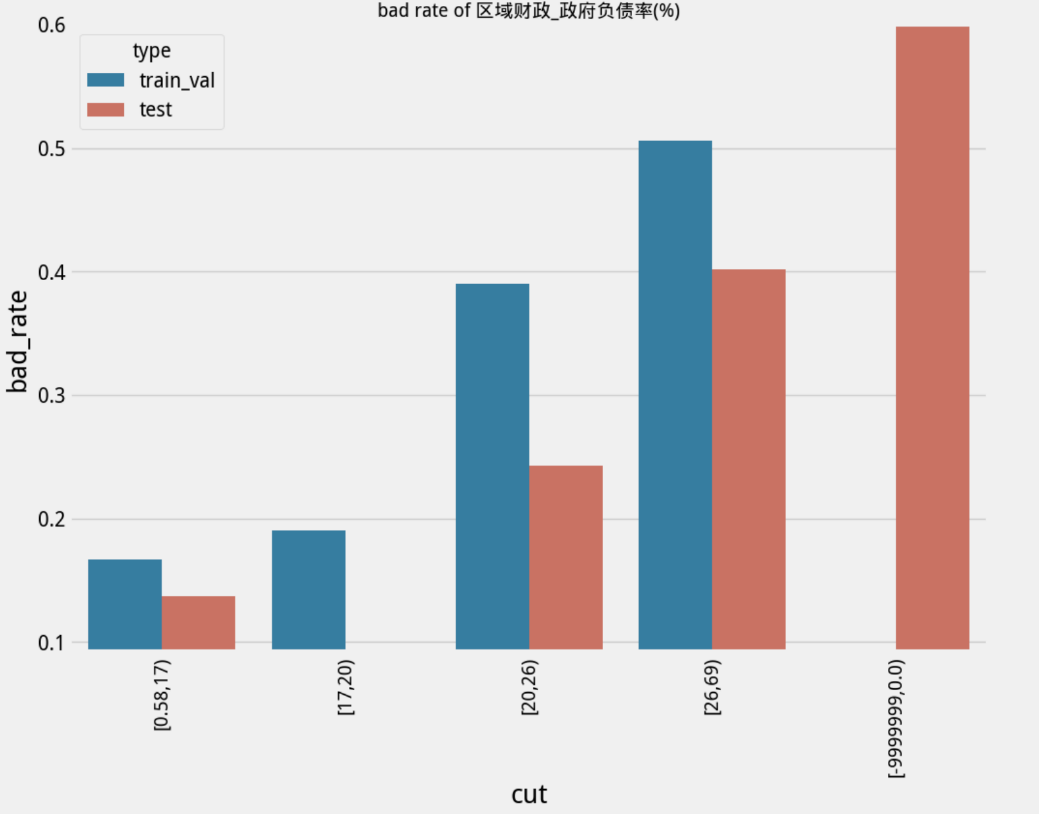
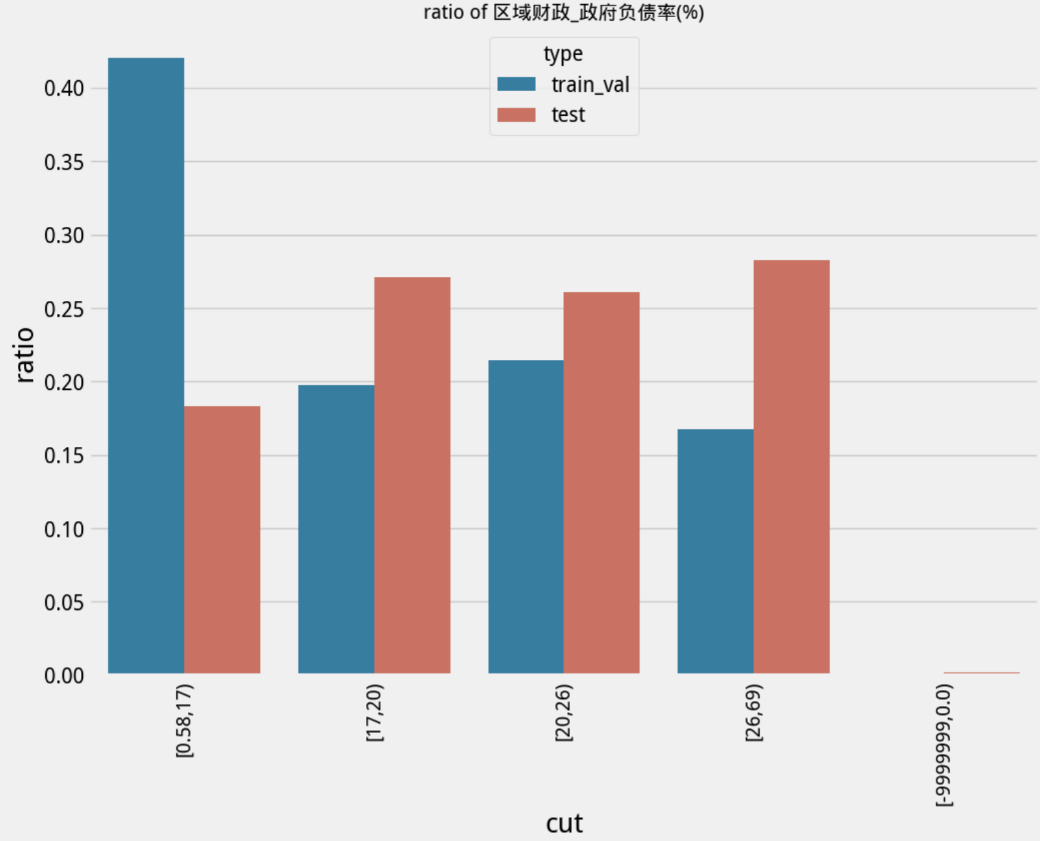


1. "Financial data" dimension

"Asset Size Data\_Asset Totals" feature presents



1. "Regional economic finance" dimension

"Regional fiscal\_government indebtedness" is characterized by

## model stability

PSI is a measure of the size of the difference between two distributions.

The PSI between the validation set and the test set of the model was calculated to be 0.025071, with insignificant changes in the model score distribution. Since the test set is a dataset outside the modeling sample time, the PSI between the validation set and the test set, in addition to the significance of the validation of the model score distribution, also represents that the model's prediction on the inter-temporal range also retains stability. Therefore, it can be considered that **the model output score distribution is stable and can be reused in the long term**.

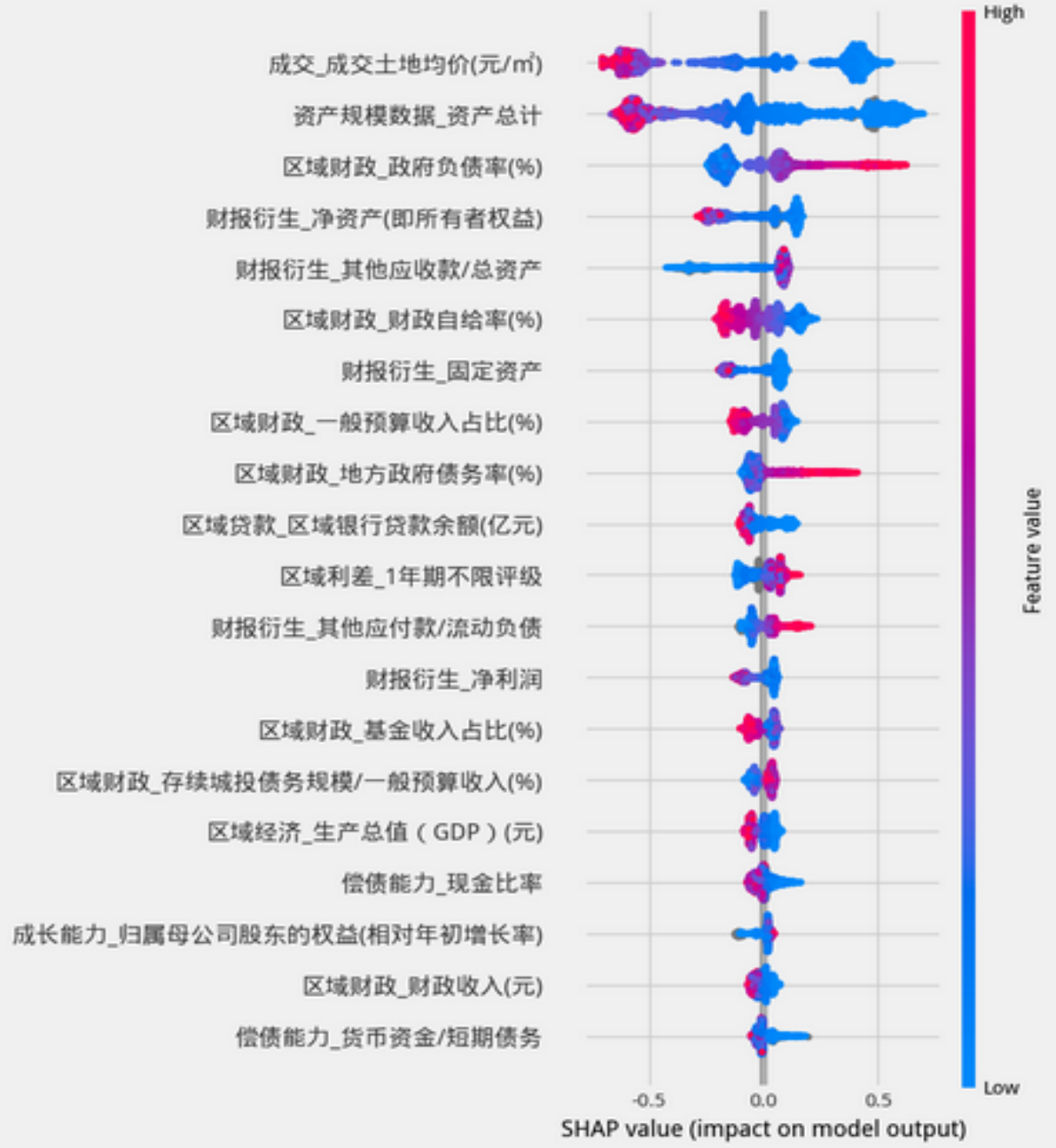
# Model Strategy Analysis

1. Segmentation strategy: Under the evaluation criterion of prioritizing the recall rate, the 5th interval is taken, i.e., the warning list considering the intervals of proba (0.306, 0.387] and above. This is because the number of bad numbers in the intervals after this interval is low, which has less impact on the overall recall rate; in addition, the size of the precision rate is more reasonable.
2. Early warning list for 2021.
3. List shap analysis

The horizontal axis is the SHAP value, which indicates the magnitude of the impact of each feature on the model's predicted results. The larger the value, the greater the influence of the feature on the prediction result.A positive SHAP value indicates that the feature has a positive influence on the prediction result; a negative SHAP value indicates that the feature has a negative influence on the prediction result.

The location of the red dots indicates the direction and magnitude of the effect on the model output when the eigenvalue is high. For example, if a red dot is in the positive direction of the SHAP value, this indicates that a high eigenvalue has a positive effect on the model output (i.e., increases the model prediction); if a red dot is in the negative direction of the SHAP value, this indicates that a high eigenvalue has a negative effect on the model output (i.e., decreases the model prediction).

From the results, the greatest influence on the model prediction is the "average price of land sold" and "total assets". Specifically, for "average price of land sold", the higher the average price, the lower the model prediction, that is to say, "average price of land sold" has a negative effect on the probability of default. In addition, for "total assets", its effect on the probability of default is similar to that of "average land price".

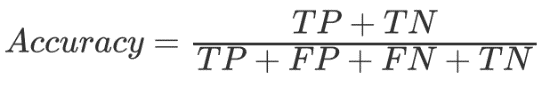


# VI. APPENDIX

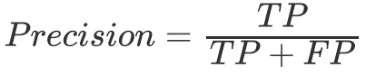
## Evaluation indicators

1. Confusion matrix: Confusion matrix (Confusion matrix) is composed of the following table.

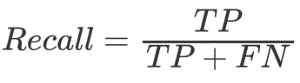


1. Accuracy. The percentage of classifications predicted by the model that are correctly categorized.

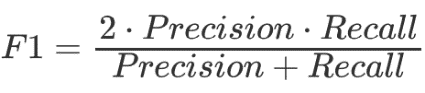
* Defined as:

1. Precision. The percentage of a classification determined by the model that actually belongs to that classification.

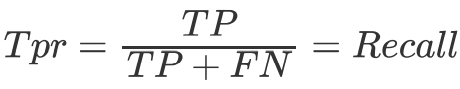
* Defined as:

1. Recall. The percentage of categories that truly belong to a category that are recognized as such by the model.

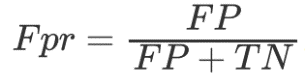
* Defined as:

1. F1: A reconciled average of accuracy and recall.

* Defined as:

1. Tpr: True positive rate.

* Defined as:

1. Fpr: False positive rate.

* Defined as:

1. ROC curve: receiver operating characteristic curve (ROC). In a certain classification judged by the model, it consists of Tpr as the axis and Fpr as axis.
2. AUC: Area under the curve of ROC. The probability that the model correctly determines a positive sample if one positive and one negative sample are randomly selected.The model works as well as random guessing when AUC = 0.5.
3. KS: Kolmogorov-Smirnov test (KS), which indicates the maximum difference between Tpr and Fpr, is used to test whether the two empirical distributions are different or not.KS is between 0.4 and 0.6 when the model's differentiation ability is good, and KS is greater than 0.75 when it suggests that the model is overfitting.
4. PSI: Model stability was tested using the Population Stability Index (PSI). This index measures the difference in the distribution of classification probability scores between the test sample and the model sample, and is the most common index for assessing model stability.

* Defined as:

