Smart Default Prediction and Evaluation System in Credit Loan

Instructors：Deng-Yang Huang、Chih-yung Tsai

Presenters： Guan-Ting Li, Zi-Hsuan Yang

Team Members：Yan-Xiang Huang, Shih-Guo Hung,
Guan-Ting Li, Zi-Yun Su,
Hui-Ying Tsou, Fan-Yu Wang,
Tian-Zhou Wang, Zi-Hsuan Yang
(list by the alphabetical order of family name)

Table of Contents

0 Prologue

Pain Points

1. For the borrowers identified in a non-traditional sense and capable of repayment, banks or loan distribution firms might overestimate their high default risk and refuse to provide credit loans.
2. How to effectively analyze the applications of past borrowers based on limited imbalanced data.
3. For customers who have never banked, it is difficult to estimate the default probability because of the lack of historical records.

Resolutions

1. AI is used to predict more precisely whether customers will default or not.
2. Banks or loan distribution firms can identify potential customers with income and loan repayment capacity among those who are traditionally considered unable to lend. They can promote proper credit plans to different types of customers.

Analysis Process

Data Examination → Data Cleaning → First Stage of Model Test → Feature Engineering → Second Stage of Model Test → Conclusion & Prospect

kaggle
Excel
jupyter
colab

python
NumPy
pandas
matplotlib
seaborn

scikit learn

python
NumPy
pandas
matplotlib
seaborn

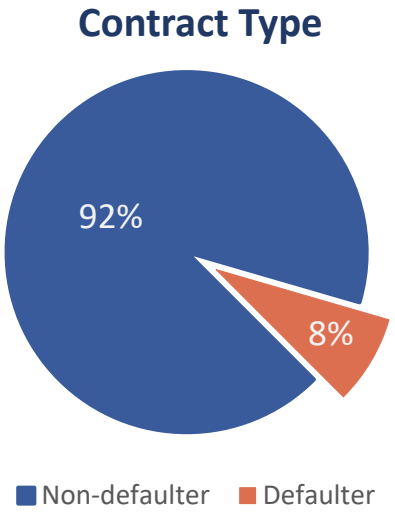
scikit learn

AMCHARTS

Flask
web development, one drop at a time

1 Introduction to Dataset

Data Examination



- This dataset includes a total of 300,000 pieces of data in 122 fields.
- One of the fields is about default or not.
- It's an extremely imbalanced dataset. The ratio of non-defaulters to defaulters is 92:8.

Data Examination

Personal Information

- SK_ID_CURR
- ORGANIZATION_TYPE
- OCCUPATION_TYPE
- CODE_GENDER
- DAYS_BIRTH
- DAYS_EMPLOYED
- NAME_EDUCATION_TYPE
- NAME_INCOME_TYPE
- AMT_INCOME_TOTAL
- NAME_FAMILY_STATUS
- CNT_CHILDREN
- CNT_FAM_MEMBERS
- NAME_TYPE_SUITE

Personal Asset

- FLAG_OWN_CAR
- OWN_CAR_AGE
- FLAG_OWN_REALTY

Loan State

- TARGET
- NAME_CONTRACT_TYPE
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- WEEKDAY_APPR_PROCESS_START
- HOUR_APPR_PROCESS_START

Residence Information

- NAME_HOUSING_TYPE
- REGION_POPULATION_RELATIVE
- REGION_RATING_CLIENT
- REGION_RATING_CLIENT_W_CITY
- INFORMATION ABOUT BUILDING

Application Documents

- REG_CITY_NOT_LIVE_CITY
- FLAG_DOCUMENT

Credit Information

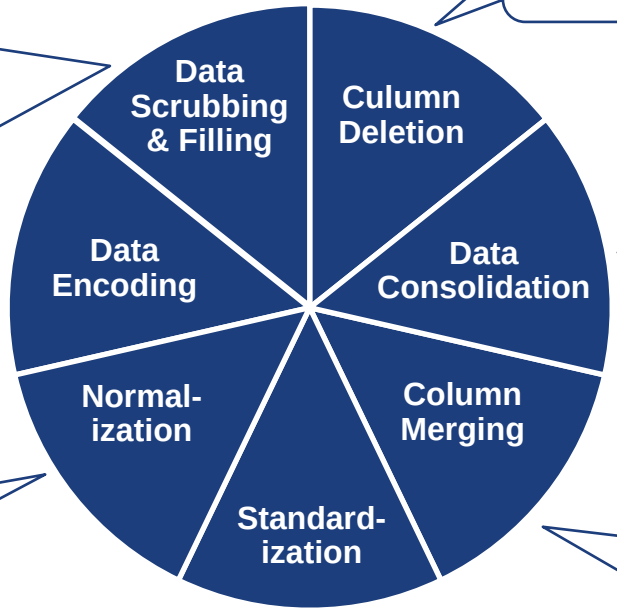
- OBS_30_CNT_SOCIAL_CIRCLE
- DEF_30_CNT_SOCIAL_CIRCLE
- OBS_60_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE
- AMT_REQ_CREDIT_BUREAU

Group fields into 6 categories.

2 Data Preprocessing

Data processing

- Fill null values with zero in the fields of inquiries to a credit bureau, car age, and building rating.
- Fill with the string "other" if the null value of the classification field exceeds 10% of the data.
- The null values in the remaining fields are filled with the median and mode respectively.



- Delete the fields if the correlation between X variables is greater than 0.8.

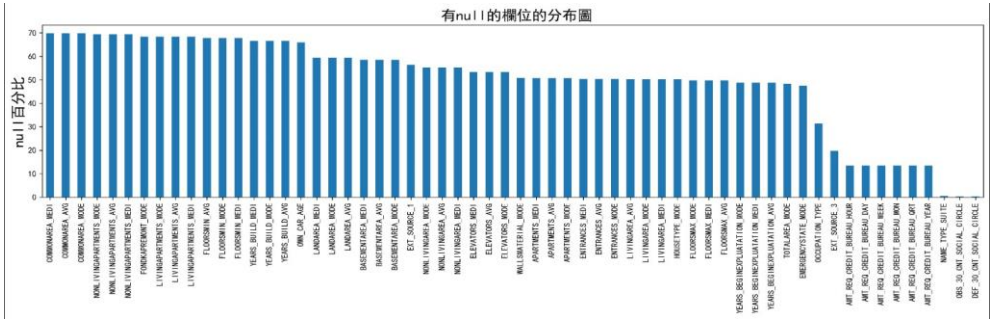
- Consolidate the application day for loans to workdays and weekends.
- Consolidate age to five-year groups.

- Consolidate building-related information to one rating field for the entire house.

- When the skewness exceeds 0.5, the field value is converted to logarithm, square root, or reciprocal.

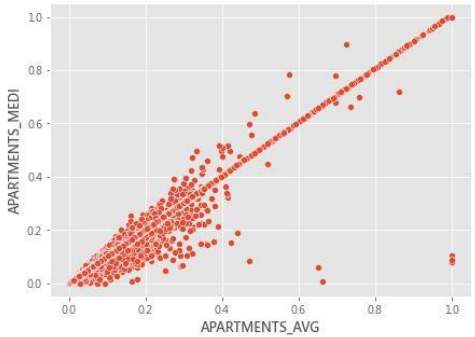
Data Filling

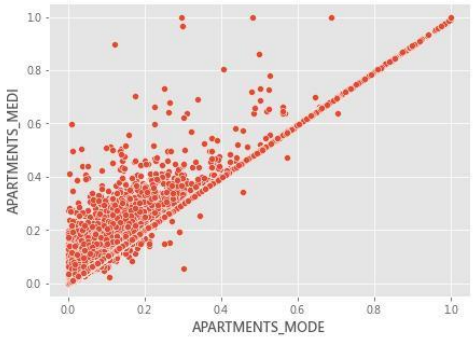
1. Fill null values with zero in the fields of inquiries to a credit bureau, car age, and building rating.
2. Fill with the string “other” if the null value of the classification field exceeds 10% of the data.
3. The null values in the remaining fields are filled with the median and mode respectively.

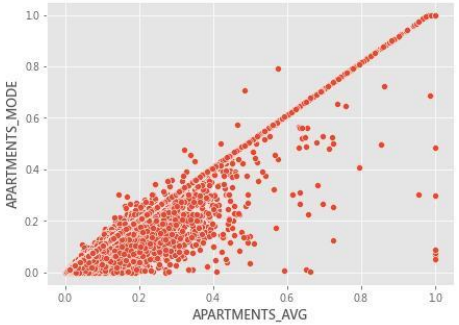


Column Cleaning

- Delete the fields if the correlation between X variables is greater than 0.8.
- When the scatter plot shows a straight-line trend from lower left to upper right, there is a positive correlation between the two variables.







3 Machine Learning

Oversampling before & after

Dataset：CreditV2-2

Before

Model	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
DNN	92%	91.9%	0	0
KNN	92.18%	91.46%	0.01	0.03
Random Forest	91.91%	91.96%	0	0
Bayesian Classifier	89.53%	89.63%	0.1	0.13
XGBoost	91.94%	91.99%	0.01	0.02

After

Model	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
DNN	68.65%	66.79%	0.62	0.24
KNN	95.97%	79.42%	0.22	0.15
☆Random Forest	66.21%	65.39%	0.68	0.24
Bayesian Classifier	61.44%	57.19%	0.67	0.20
XGBoost	65.49%	67.19%	0.63	0.24

Data Scrubbing-Cluster & Column Consolidation

Loan Information	Change of Personal Information	Income State	Building Rating	Numbers of Credit Inquiry
AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE	DAYS_REGISTRATION DAYS_ID_PUBLISH DAYS_LAST_PHONE_CHANGE	AMT_INCOME_TOTAL DAYS_BIRTH DAYS_EMPLOYED	Building-related rating	AMT_REQ_CREDIT_BUREAU by hour, day, week, month, season, year

Cluster before & after

資料集：CreditV2-2、CreditV3-3

Before

Model	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
DNN	68.65%	66.79%	0.62	0.24
Random Forest	66.21%	65.39%	0.68	0.24
Bayesian Classifier	61.44%	57.19%	0.67	0.20
XGBoost	65.49%	67.19%	0.63	0.24

After

Model	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
DNN	67.04%	76.31%	0.55	0.27
Random Forest	71.62%	69.92%	0.64	0.25
Bayesian Classifier	64.36%	61.92%	0.67	0.22
☆ XGBoost	69.19%	68.85%	0.67	0.26

Feature Engineering--Wrapper

After the cluster, there are thirty-five fields left. By using the wrapper method, 17 features are selected for model tests.
The test results in the XGBoost model are shown in the table below.

Dataset	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
creditV3-4(35)	69.19%	68.85%	0.67	0.26
creditV3-4(17)	**77.12%**	**73.10%**	**0.61**	**0.26**

Feature Engineering --SelectKBest

By using the SelectKBest method, fifteen, seven, and five features are selected respectively. The results are:

Dataset	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
creditV3-4(35)	69.19%	68.85%	0.67	0.26
creditV3-4(15)	**70.25%**	**68.75%**	**0.64**	**0.25**
creditV3-4(7)	57.44%	52.65%	0.6	0.17
creditV3-4(5)	56.29%	47.75%	0.65	0.17

Feature Engineering --Feature Importance

By using the feature importance method, five features are selected for testing. The table below shows the results of the XGBoost model.

Dataset	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
creditV3-4(35)	69.19%	68.85%	0.67	0.26
creditV3-4(5)	**73.31%**	**71.98%**	**0.65**	**0.27**

Similar accuracy to original data can be achieved with only a few fields.

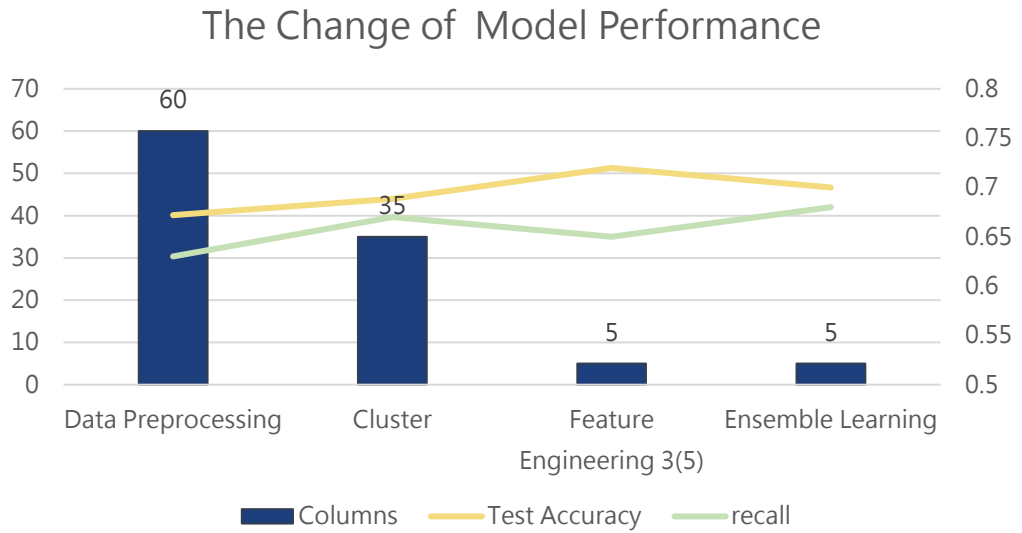
Ensemble Learning--Voting

To solve the problem caused by high bias or high variation in a single model, five selected features are used to integrate with multiple models in the soft voting method. Compared with other methods of feature engineering, the recall of test results rises.
The table below shows the test results in ensemble learning.

Dataset	Train Accuracy	Test Accuracy	recall(y=1)	F1 score(y=1)
creditV3-4(5)	73.31%	71.98%	0.65	0.27
creditV3-4 (Ensembling 5)	**68.00%**	**70.00%**	**0.68**	**0.26**

The Winner

Finally, XGBoost is selected in this project to predict the default rate of credit loans. The following figure shows the trend of the number of columns and model performance after cleaning and machine learning.



4 Conclusion

Conclusion in Management

Key decision factors:

Education Category, Credit Agency Score, Years Employed, and Score of Residence Region

Applications:

In the financial industry, the default system helps find reliable customers and sell more products.
It can also apply to "fraud detection", "defective rate detection", "preventive medicine".

Members & Tasks

Name	Expertise	Task
Shih-Guo Hung	System Analysis, Programming, Project Management, AI Applications, Data Science, Modeling, System Architecture	Modeling, Program Development, Task Dispatch, Technical Support
Guan-Ting Li	Python, MySQL, Data Analysis, Data Cleaning, Data Science, Algorithms	Data cleaning, Feature Engineering, Programming, Oral Presentation
Zi-Hsuan Yang	Data analysis, Statistical Applications, Python, MySQL, HTML	PPT Preparation, Model Testing, Website Front-end Design, Oral Presentation
Hui-Ying Tsou	Foreign Languages (English, French, German, Japanese), Event Management, Project Execution	PPT Preparation, Text Editing

Members & Tasks

Name	Expertise	Task
Yan-Xiang Huang	Python, MySQL, HTML, Data Analysis, Data Science, Information Management	Website Front-end Design, Model Testing
Fan-Yu Wang	Python, HTML, MySQL, Information Management	PPT Preparation, Model Testing, Website Front-end Design
Tian-Zhou Wang	Team Coordination, Project Management, Data Analysis, System Analysis, Algorithm Model, Data Science	Team Leader, Task Coordination, Model Testing
Zi-Yun Su	Finance, Statistical Science, Brand Marketing, Python, MySQL	PPT Preparation, Model Testing

5 Interactive Default System

Visual Present

Scan and Enter
Our Website