CZ4032: Data Analytics & Mining

Home Credit Default Risk Prediction

Group 6

Liu Zhemin
Alvin Lee Yong Teck
Tan Yong Kuan
Liu Jihao Bryan
Pang Qi Kai Jarald
Zhuo Yue Zheng

PROBLEM DESCRIPTION

PROBLEM DESCRIPTION

- Many people **struggle to get loans** due to insufficient credit histories
- Aim to make use of alternative data (e.g. transactional info) to **predict repayment ability**
- Ensure customer with capability to repay loans are **NOT** rejected

OBJECTIVE

Predict if each application is NOT capable of repaying a loan

Classify into 2 Classes:

Negative – 0: Able to repay a loan ✓

Positive – 1: Unable to repay a loan ✗

DATASET DESCRIPTION

DATASET DESCRIPTION

Total No. of Features: 121

Total No. of Observations: 304531

Accuracy Paradox

Imbalanced Class:

- More loans were **repaid** than loans that were **not repaid**

Dataset VISUALIZATION

Non trivial problem – overlapping classes

PCA

ICA

DATA PRE-PROCESSING

ANOMALY DETECTION ①

- **Maximum number of days employed is 1000 years (Anomaly)**

```
[13] app_train['DAYS_EMPLOYED'].describe()

count    307511.000000
mean      63815.045904
std      141275.766519
min      -17912.000000
25%       -2760.000000
50%       -1213.000000
75%        -289.000000
max      365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```

ANOMALY DETECTION ①

- **Observation:** All the anomalies have the same value and have lower rate of default

- **Solution:** Set the anomalies to median

IMPUTATION - MISSING VALUES ②

- **Observation:** Total of 67 columns with missing values

- **Solution:** Median imputing

Your selected dataframe has 122 columns.
There are 67 columns that have missing values.

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
AMT_REQ_CREDIT_BUREAU_HOUR	41519	13.5
AMT_REQ_CREDIT_BUREAU_YEAR	41519	13.5
NAME_TYPE_SUITE	1292	0.4
DEF_30_CNT_SOCIAL_CIRCLE	1021	0.3
OBS_60_CNT_SOCIAL_CIRCLE	1021	0.3
DEF_60_CNT_SOCIAL_CIRCLE	1021	0.3
OBS_30_CNT_SOCIAL_CIRCLE	1021	0.3
EXT_SOURCE_2	660	0.2
AMT_GOODS_PRICE	278	0.1
AMT_ANNUITY	12	0.0
CNT_FAM_MEMBERS	2	0.0
DAYS_LAST_PHONE_CHANGE	1	0.0

LABEL ENCODING & ONE HOT ENCODING ③

Label Encoding

- For columns with **<u>2 or fewer</u>** categories (binary values)
- Not done for columns >=3 categories because it will imply that there is order
- A total of 4 columns were labelled encoded

```
# Number of each type of column
app_train.dtypes.value_counts()

float64    65
int64      41
object     16
dtype: int64
```

One Hot Encoding

- For columns with **<u>more than 3 categories</u>**
- Except the 4 columns, the rest of the columns were **One Hot Encoded**

NAME_CONTRACT_TYPE	2
CODE_GENDER	3
FLAG_OWN_CAR	2
FLAG_OWN_REALTY	2
NAME_TYPE_SUITE	7
NAME_INCOME_TYPE	8
NAME_EDUCATION_TYPE	5
NAME_FAMILY_STATUS	6
NAME_HOUSING_TYPE	6
OCCUPATION_TYPE	18
WEEKDAY_APPR_PROCESS_START	7
ORGANIZATION_TYPE	58
FONDKAPREMONT_MODE	4
HOUSETYPE_MODE	3
WALLSMATERIAL_MODE	7
EMERGENCYSTATE_MODE	2

FEATURE SELECTION ④

```
Most Positive Correlations:
 NAME_EDUCATION_TYPE_Secondary / secondary special    0.049824
REG_CITY_NOT_WORK_CITY                                0.050994
DAYS_ID_PUBLISH                                       0.051457
CODE_GENDER_M                                         0.054713
DAYS_LAST_PHONE_CHANGE                                0.055218
NAME_INCOME_TYPE_Working                              0.057481
REGION_RATING_CLIENT                                  0.058899
REGION_RATING_CLIENT_W_CITY                           0.060893
DAYS_BIRTH                                            0.078239
TARGET                                                1.000000
Name: TARGET, dtype: float64

Most Negative Correlations:
 EXT_SOURCE_3                          -0.178919
EXT_SOURCE_2                          -0.160472
EXT_SOURCE_1                          -0.155317
NAME_EDUCATION_TYPE_Higher education  -0.056593
CODE_GENDER_F                         -0.054704
NAME_INCOME_TYPE_Pensioner            -0.046209
ORGANIZATION_TYPE_XNA                 -0.045987
DAYS_EMPLOYED                         -0.044932
FLOORSMAX_AVG                         -0.044003
FLOORSMAX_MEDI                        -0.043768
Name: TARGET, dtype: float64
```

FEATURE ENGINEERING ⑤

Polynomial Feature Engineering

- **Top 5 features with highest correlation selected**
- **Do polynomial feature engineering with degree-3**
  - If an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features are [1, a, b, a^2, ab, b^2].
  - Resultant **56** features
- **Goal:** To **increase correlation** to target

FEATURE ENGINEERING ⑤

Correlations after feature engineering

- **New features correlations to target increased**

```
Most Negative Correlations:
EXT_SOURCE_3 EXT_SOURCE_2                -0.193939
DAYS_BIRTH EXT_SOURCE_3 EXT_SOURCE_2     -0.189605
EXT_SOURCE_3^2 EXT_SOURCE_2              -0.176428
EXT_SOURCE_3 EXT_SOURCE_2^2              -0.172282
DAYS_BIRTH EXT_SOURCE_3                  -0.166625
DAYS_BIRTH EXT_SOURCE_2                  -0.164065
EXT_SOURCE_3                             -0.160295
DAYS_BIRTH EXT_SOURCE_3^2                -0.156867
EXT_SOURCE_2                             -0.155892
DAYS_BIRTH EXT_SOURCE_2^2                -0.150822
Name: TARGET, dtype: float64

Most Positive Correlations:
DAYS_BIRTH EXT_SOURCE_1                   0.104882
EXT_SOURCE_2^2 EXT_SOURCE_1               0.141777
EXT_SOURCE_3^2 EXT_SOURCE_1               0.149313
EXT_SOURCE_2 EXT_SOURCE_1                 0.150109
DAYS_BIRTH EXT_SOURCE_2 EXT_SOURCE_1      0.151816
DAYS_BIRTH EXT_SOURCE_3 EXT_SOURCE_1      0.155891
EXT_SOURCE_3 EXT_SOURCE_1                 0.156873
EXT_SOURCE_3 EXT_SOURCE_2 EXT_SOURCE_1    0.181283
TARGET                                    1.000000
1                                              NaN
Name: TARGET, dtype: float64
```

UPSAMPLING + SMOTE ⑥

Using SMOTE

- **Imbalanced dataset:** 10x number of negative class compared to positive
- **Problems:** Overall high accuracy but low recall for minority class (positive)
  - High prediction for person who are able to repay
  - Low prediction for person who are unable to repay
- **Synthetic Minority Oversampling Technique**
  - Uses KNN to generate new synthetic instances
  - Instead of cloning the same data
- **Upsampling resolves this issue by adding more minority class**
  - Increases sensitivity to the minority class

NORMALISATION ⑦

- Different features have **non-standardised** weightage
- Prevent dominance in dimensions, weighs all features **equally** in their representation
- Various **scaling methods** were used
- MinMax scaling works **BEST** due to difference in test and training dataset

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
0	135000.0	568800.0	20560.5
0	99000.0	222768.0	17370.0
0	202500.0	663264.0	69777.0
2	315000.0	1575000.0	49018.5
1	180000.0	625500.0	32067.0

ALGORITHMS USED

Decision Tree

Linear SVM

Naive Bayes

Logistic Regression

Random Forest

Adaboost

Neural Network (MLP)

XGBoost

BASELINE MODELS – PERFORMANCE COMPARISON

Models	Accuracy	F1-Score	ROC AUC
Logistic Regression	91.93	0	0.62
Decision Tree	85.33%	0.15	0.54
Linear SVM	91.76%	0.001	0.50
Naive Bayes	91.23%	0.01	0.60
Adaboost	91.91%	0.025	0.70
Random Forest	91.92%	0.002	0.69
Neural Network (MLP)	91.91%	0.001	0.52

REASON FOR POOR F1-SCORE

PROCESSED MODELS – PERFORMANCE COMPARISON

Models	Accuracy	F1-Score	ROC AUC
Logistic Regression	66.23% (↓25.7%)	0.24 (↑0.24)	0.72 (↑0.10)
Decision Tree	76.78% (↓8.55%)	0.18 (↑0.03)	0.56 (↑0.02)
Linear SVM	67.34% (↓24.42%)	0.247 (↑0.246)	0.67 (↑0.17)
Naive Bayes	49.77% (↓41.46%)	0.20 (↑0.19)	0.69 (↑0.09)
Adaboost	66.82% (↓25.09%)	0.243 (↑0.218)	0.72 (↑0.02)
Random Forest	81.82% (↓10.10%)	0.20 (↑0.198)	0.66 (↓0.03)
Neural Network (MLP)	65.25% (↓26.66%)	0.24 (↑0.239)	0.72 (↑0.20)
XGBoost	73.97%	0.25	0.72

BETTER F1-SCORE

KAGGLE LEADERBOARD PERFOMANCE

BEFORE

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
log_reg_baseline.csv	just now	0 seconds	0 seconds	0.62079

Complete

Jump to your position on the leaderboard

AFTER

Your most recent submission

+0.09955

Name	Submitted	Wait time	Execution time	Score
xg_boost.csv	just now	1 seconds	1 seconds	0.72034

Complete

Jump to your position on the leaderboard

CONCLUSION

- **Knowing the complexity of the dataset**

  - Missing values, multiple sources etc.

- **Feature Selection, Engineering & Normalization**

  - Curse of dimensionality

  - Dominance of dimensions

- **Imbalance Dataset**

  - Accuracy Paradox

  - Insensitivity to minority class, resulting in poor F1-Score

Questions & Answers

Group 6

Liu Zhemin
Alvin Lee Yong Teck
Tan Yong Kuan
Liu Jihao Bryan
Pang Qi Kai Jarald
Zhuo Yue Zheng