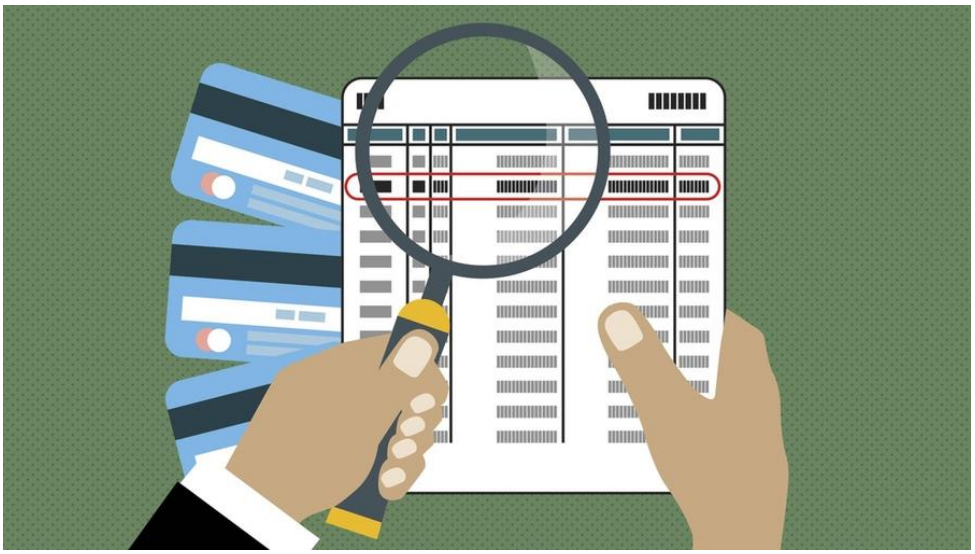


Default Credit Card Clients Prediction

CAPSTONE PROJECT REPORT

Dasun Kehelwala (DSA_0392)

Machine Learning Foundations Training
2022-11-20

Introduction (Problem Definition)

This project is focusing on predicting credit card clients who will default on their next month payment based on their demographic characteristics, past spending patterns and past repayment patterns. This is one of important business problems for banks which provide credit card facilities for Customers. This will be specifically useful to manage credit risks. This challenge is addressed as classification problem in this project. After deployment, customers should be able to access the service through API and also by sending batch input as csv file. Derived machine learning model predicts whether customer is going to default next month payment or not.

Dataset

The dataset used in this project was named as "Default of credit card clients Data Set". This dataset contains the default payment details in Taiwanese banking industry in year 2005. Dataset is available to download downloaded from UCI Machine learning repository through following link.

Link: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#

This dataset is Multivariate dataset with 24 attributes and 30,000 instances. All the attributes available in dataset is converted to Real Integer values. This dataset was donated to public access in 2016-01-26. Most notable observation about this dataset is its class imbalance. Table 1 Describes Attribute Details.

Table 1: Attribute Details

Attribute	Description
ID	Identifier for data entry
X1 (LIMIT_BAL)	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. →Numerical
X2 (SEX)	Gender (1 = male; 2 = female). →Categorical variable mapped to integers
X3 (EDUCATION)	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). →Categorical variable mapped to integers

X4 (MARRIAGE)	Marital status (1 = married; 2 = single; 3 = others). →Categorical variable mapped to integers
X5 (AGE)	Age (year) →Numerical
X6 - X11 (PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6)	History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows: X6 = the repayment status in September 2005; X7 = the repayment status in August, 2005; . . .; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -2: No consumption; -1 = pay duly; 0: The use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above. → Categorical variables mapped to integers, but have ordinal nature as per definition
X12-X17 (BILL_AMT1 to BILL_AMT6	Amount of bill statement (NT dollar). X12 = amount of bill statement in September 2005; X13 = amount of bill statement in August 2005; . . .; X17 = amount of bill statement in April 2005. →Numerical
X18-X23 (PAY_AMT1)	Amount of previous payment (NT dollar). X18 = amount paid in September 2005; X19 = amount paid in August 2005; . . .; X23 = amount paid in April, 2005. →Numerical
Y (default payment next month_	default payment (Yes = 1, No = 0) → class variable - Categorical variables mapped to integers

Methodology (Solution Approach, Tools used)

As per problem definition, ML model need to predict this solution is approached as binary classification problem. Solution approach is described below as step-by-step process

During Exploratory Data Analysis, author has noticed following observations

Positive class (Defaulted CX or value 1) represents only 22.1% of whole dataset. Figure 1 depicts the imbalanced Nature of Dataset.

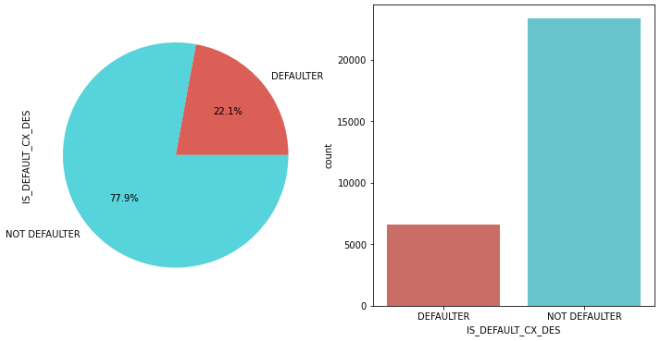


Figure 1 : Imbalanced Nature of Dataset

Results

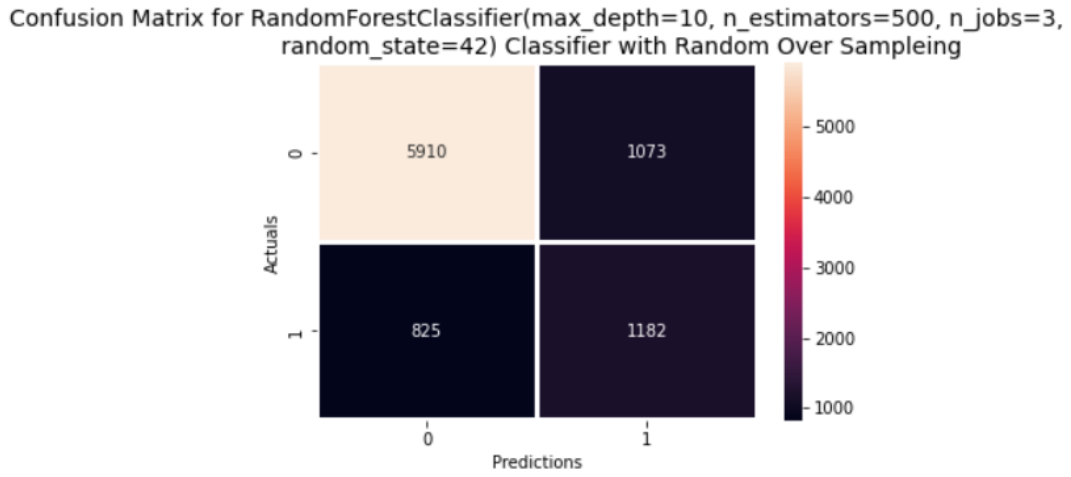
Initially three models were developed using Three classifiers: Logistic Regression, XGBoost and Decision Tree. For these three models, default hyperparameters were used and test set with all 23 features was used for training without using resampling techniques. Since all three F scores were not satisfactory, Models with same settings was trained again with resampled dataset using Random Oversampling. Table 2 shows the results obtained through this initial run. XGBoost and Decision Tree Classifiers with Random Oversampling gave better results.

Table 2: Results Through Initial Run

Model ID	Model	Resampeling method	Feature count	Accuracy score	Precision score	Recall score	F1 score	F1 score weighted	ROC AUC score
lgr_ns_01	LogisticRegression(random_state=42)	NO RESAMPELING	23	0.813126	0.714286	0.271550	0.393502	0.778805	0.727441
xgb_ns_01	XGBClassifier(random_state=42)	NO RESAMPELING	23	0.823248	0.697543	0.367713	0.481566	0.801508	0.789493
rf_ns_01	(DecisionTreeClassifier(max_depth=20, max_feat...	NO RESAMPELING	23	0.818020	0.672558	0.360239	0.469176	0.796197	0.782405
lgr_rovs_01	LogisticRegression(random_state=42)	Random Over Sampleing	23	0.720245	0.415333	0.620827	0.497703	0.737279	0.727212
xgb_rovs_01	XGBClassifier(random_state=42)	Random Over Sampleing	23	0.770523	0.489313	0.638764	0.554139	0.780456	0.786756
rf_rovs_01	(DecisionTreeClassifier(max_depth=20, max_feat...	Random Over Sampleing	23	0.807675	0.583534	0.483807	0.529011	0.800995	0.779527

Since XGBoost and Decision Tree Classifiers with Random Oversampling have given better results in initial run, These two classifiers were used for second level evaluation. At second level, performance was tested using different hyperparameter settings.

Model ID	Model	Resampeling method	Feature count	Accuracy score	Precision score	Recall score	F1 score	F1 score weighted	ROC AUC score
lgr_ns_01	LogisticRegression(random_state=42)	NO RESAMPELING	23	0.813126	0.714286	0.271550	0.393502	0.778805	0.727441
xgb_ns_01	XGBClassifier(random_state=42)	NO RESAMPELING	23	0.823248	0.697543	0.367713	0.481566	0.801508	0.789493
rf_ns_01	(DecisionTreeClassifier(max_depth=20, max_feat...	NO RESAMPELING	23	0.818020	0.672558	0.360239	0.469176	0.796197	0.782405
lgr_rovs_01	LogisticRegression(random_state=42)	Random Over Sampleing	23	0.720245	0.415333	0.620827	0.497703	0.737279	0.727212
xgb_rovs_01	XGBClassifier(random_state=42)	Random Over Sampleing	23	0.770523	0.489313	0.638764	0.554139	0.780456	0.786756
rf_rovs_01	(DecisionTreeClassifier(max_depth=20, max_feat...	Random Over Sampleing	23	0.807675	0.583534	0.483807	0.529011	0.800995	0.779527




```
Model ID : rf_rovs_04
Model : RandomForestClassifier(max_depth=10, n_estimators=500, n_jobs=3,
                       random_state=42)
Resampeling method : Random Over Sampleing
Feature count  : 23
Accuracy score : 0.7888765294771969
Precision score : 0.5241685144124169
Recall score : 0.5889387144992526
F1 score : 0.5546691694040357
F1 score weighted : 0.793110632502284
ROC AUC score : 0.7851551504433039
```

Conclusion

aaaa

Discussion

References

- ✓ https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#
- ✓ Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.
- ✓ https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset
- ✓