HCI Credit Scoring

Timothy E. Ledwinardi

Background

Customer defaults often result from **inaccurate predictions** due to **excessive data**. This is a **common challenge for credit institutions**. **Credit scoring models** emerge as a **solution** to mitigate credit risks and **enhance revenue** efficiency through **strategic approaches**

Problem Statement

How to **decrease the default rate of 10 %** on next period?

Objectives

- Creating model to predict loan status
- Determining strategies based on the predictions of the model

DATASET DESCRIPTION

A dataset is provided by **Home Credit Indonesia**, consisting of **several tables**. The primary datasets to be utilized are "**application_train**" with **300 thousand** rows and "**application_test**" containing **48 thousand** rows, both containing **numerous features**.

For more detailed information, please refer **here**

Workflow

Data Preprocessing

- **Joining Dataset**
  - Merging dataset according to table schema
- **Cleaning Data**
  - Treating missing value
  - Treating unknown value
  - Dropping irrelevant value
- **Feature Engineering**
  - Create new columns
  - Encoding categorical column

Workflow

Data Preprocessing

- **Feature Selection**
  - Using Correlation and dropping strong-correlated value
  - Using KBest and Feature Importance using tree-based model
- **Modeling**
  - TrainTest Split
  - Evaluating model based on classification report and AUC Score
  - Parameter setting
- **Fitting Dataset**
  Using the model to predict test dataset and define default chance

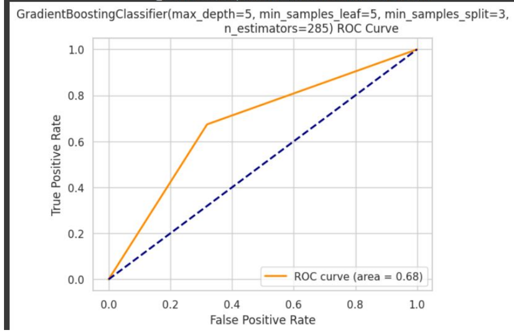
Machine Learning Evaluation

- Due to **highly imbalanced data** we evaluated the model using **recall score** and **ROC score**
- Due to the **large datasets** and **limited time**, we should **explore further** to try **best feature** to **get best score**
- The best model to be selected are **GradientBoostingClassifier** with average **recall and ROC score of 69 %** after **Tuning the Parameter** using **GridSearchCV** and **5-fold Cross Validation**

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.68	0.80	84841
1	0.16	0.67	0.25	7413
accuracy			0.68	92254
macro avg	0.56	0.68	0.52	92254
weighted avg	0.90	0.68	0.75	92254

```
GradientBoostingClassifier(max_depth=5, min_samples_leaf=5, min_samples_split=3,
                           n_estimators=285) accuracy is 0.680230667504932
GradientBoostingClassifier(max_depth=5, min_samples_leaf=5, min_samples_split=3,
                           n_estimators=285) ROC AUC score is 0.6773652535232612
```



Insight Found

- After fitting to the test dataset, **most of the customer (test)** are predicted to be **defaulted (56% test, increase of 270% in data train)**. Most **defaulted customers** exhibit **low external scores**, indicative of **poor financial habits**.
  - **External scores serve as a reference**, establishing an **internal scoring** is **essential for prevention of future defaults**
  - It is **essential to clearly define maternity leaves** as either **temporary** or **permanent**, as this **distinction significantly** on **default risk**

- **Unmarried customer** are also **prone** to be **default** looking deeper, probably **married customers having a higher total income**, combining the **financial resources of both partners**
  - Customers with **higher education levels** are **less likely to default** due to their typically **enhanced employment**
  - Developing programs that **support education financing** likely **leads** to **customers with improved financial habits**, resulting in a **lower risk of default**

Thanks!

Links & References:

1. [Repository]
2. [Reference Code]
3. HCI: Final Task Tips & Trick
4. ChatGPT

Any Comments or Feedback are highly appreciated

timothyevanno@gmail.com