Rakamin Academy X id/x partners

BY: CYNTIA ANGELICA

ID/X X RAKAMIN ACADEMY DATA SCIENCE PROJECT

DATASET OVERVIEW

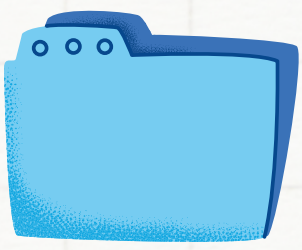
The dataset is a loan application data comprised of 466285 rows and 74 columns. The target variable is denoted by the column 'loan_status'. This target consists of several subtargets, and we will split it into good_loan (0) and bad_loan (1). Our goal is to build a model that has a high recall (1) score to minimize financial risk of the business.

FEATURE ENGINEERING

The previously cleaned data has to be engineered so that it can be fitted into the machine learning models. After feature engineering the data, we get a data consisted of 89 engineered columns.

MODEL TRAINING

In this project we try 4 models: Logistic Regression, Random Forest, AdaBoost, and GradientBoost. We fit and evaluate the result using testing data and search for the model with the highest recall score, minimizing number of the false negative.



EDA

After cleaning the data, Exploratory Data Analysis is done to gain insight of the dataset, such as:

- People that borrow the loan in order to support their small business, to move to a new house, and to get married are less likely pay back the loan
- Homeless people and those who rent a place are less likely to repay.

TREATING IMBALANCED DATA

After splitting the data into training and testing, we fit the training data and transform it with RandomOversampler(), since the data is imbalanced (target variable 1 is much less than target variable 0).

RESULTS

The classification reports concluded that logistic regression has the highest recall, with 0.78 and an AUC score of 0.848. Hence, to satisfy our goal, the model we pick is **Logistic Regression**.

https://www.linkedin.com/in/cyntia-angelica-6b5439217

https://github.com/Cyntiahp31