Project Proposal

Yi Wang

Banner ID: B01629467

Brown ID: 140304323

Because of the advance of financial crediting system, customers are able to make an expense with their creditability and pay off in the future. Yet, it is challenging yet vital for banks to determine whether their customers would have the ability to pay off their credits to minimize the risk of capital loss. Therefore, one of the ways of doing so is to track historical transactions and payments to target potential customers who will not meet payoff requirements. These customers will be classified and will have a greater change of fedault in the future.

In this project, the dataset was obtained from banks in Taiwan and was meant to target the case of customers default payments. This is a binary classification problem and the target valuable is whether acustomer will default or not based on different features and payment histories. This project is interesting and meaningful because it will help banks to target potential default payments and make a better decision with the customers.

GitHub Page: https://github.com/yahowang/Data1030Project

Source: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Number of Instances: 30000

Number of Features: 24

Original Features in the dataset:

X_1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X_2: Sex (1 = male; 2 = female).

X_3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X_4: Marriage (1 = married; 2 = single; 3 = others).

X_5: Age (year).

X_6 - X_{11}: History of past payment (with 10 categories for each feature).

- The measurement scale for the repayment status is:
  - -1 = pay duly
  - 1 = payment delay for one month
  - 2 = payment delay for two months
  - 3 = payment delay for three months
  - 4 = payment delay for four months
  - 5 = payment delay for five months
  - 6 = payment delay for six months
  - 7 = payment delay for seven months
  - 8 = payment delay for eight months
  - 9 = payment delay for nine months and above
- X_6 = the repayment status in September, 2005
- X_7 = the repayment status in August, 2005
- X_8 = the repayment status in July, 2005
- X_9 = the repayment status in June, 2005
- X_{10} = the repayment status in May, 2005
- X_{11} = the repayment status in April, 2005

X_{12}-X_{17}: Amount of bill statement (NT dollar)

- X_{12} = amount of bill statement in September, 2005
- X_{13} = amount of bill statement in August, 2005
- X_{14} = amount of bill statement in July, 2005
- X_{15} = amount of bill statement in June, 2005
- X_{16} = amount of bill statement in May, 2005
- X_{17} = amount of bill statement in April, 2005

X_{18}-X_{23}: Amount of previous payment (NT dollar).

- X_{18} = amount paid in September, 2005
- X_{19} = amount paid in August, 2005
- X_{20} = amount paid in July, 2005
- X_{21} = amount paid in June, 2005
- X_{22} = amount paid in May, 2005
- X_{23} = amount paid in April, 2005.

X_{24}: Whether the customer will default in the next session (1 = Yes; 0 = No)

Public Paper:

1) 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.

- This paper applies six different approaches to the dataset and concludes that artificial neural network is the only one that canaccurately estimate the real probability of default.

2) Islam, Sheikh Rabiul, William Eberle, and Sheikh Khaled Ghafoor. "Credit default mining using combined machine learning and heuristic approach." arXiv preprint arXiv:1807.01176 (2018).

- The paper presents and validates a heuristic approach to mine potential default accounts in advance where a risk probability is precomputed from all previous data and the risk probability for recent transactions are computed as soon they happen.

Data Preprocessing

1) For X_1(Limit of Balance) and X_5(Age), I used MinMaxScaler because they are believed to fall within a certain interval.
2) For X_2(Sex), X_4(Marriage), X_6(Payment in September), X_7(Payment in August), X_8(Payment in July), X_9(Payment in June), X_{10}(Payment in May), X_{11}(Payment in April), I used OneHot Encoder because they do not have ordinality but just unordered categories.
3) For X_{24}(Response Variable), I changed the name of the feature to Y.
4) In the original dataset, the first column contains IDs and it is dropped.
4) As the result, there are 86 features (columns) in the preprocessed data in total.

```
[ ]
```