Credit Card Clients

Domain

This dataset describes and default payments of credit card clients in Taiwan from April to September 2005. It was sourced from the Department of Information Management, Chung Hua University, Taiwan and the Department of Civil Engineering, Tamkang University, Taiwan.

Past usage of the dataset include various proposals suggesting a classification problem such as the following:

https://www.linkedin.com/pulse/default-payment-prediction-system-duy-hoang-ly

Problem Statement

The dataset compares the predictive accuracy of probability of default based on 23 feature describing demographics such as gender, marital status and educational background as well as financial features such as payment lateness for the past 6 months, payment amount and line of credit. For any give set of features we will use a classification model to predict if a credit card client will default or not.

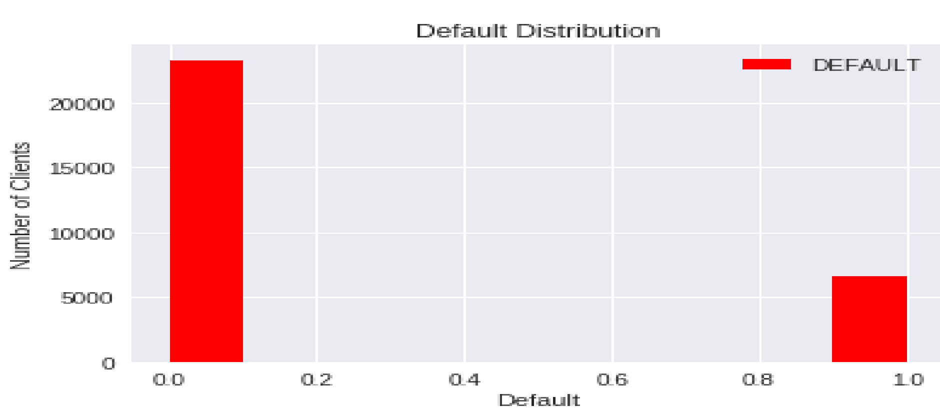
Dataset and Inputs

The data set contains 30k observations and 23 explanatory variables (56 categorical, 23 numeric) that are involved in assessing the category for default as 0 for not defaulting and 1 for defaulting.

The data set uses up 5.28 MB in memory.

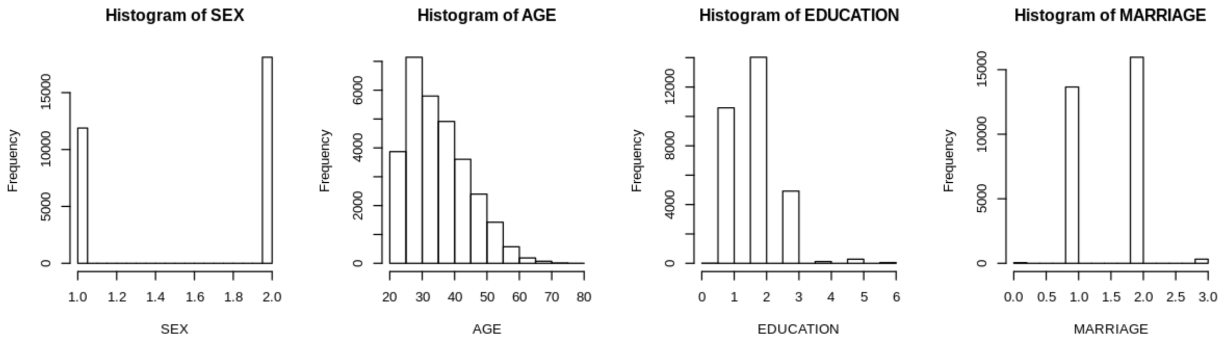
Evaluation Metrics

The dataset is not evenly distributed as most credit card client do not default. We can use the success of the model measuring against the F1 score and determine if our model predicts the correct default class more reliably than the F1 score.



Data Exploration

As seen above the distribution of defaulting is about 80% to 20%. In the below plot we can see the demographic distribution of the dataset:



Gender (1 = male; 2 = female)
Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
Marital status (1 = married; 2 = single; 3 = others)

The dataset has a few more female clients than male with the majority of clients between the age of 25-40 and an almost evenly split between married and single.

Here is the summary of all features:

```
       ID          LIMIT_BAL            SEX          EDUCATION
 Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000
 1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000
 Median :15000   Median : 140000   Median :2.000   Median :2.000
 Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853
 3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000
    MARRIAGE          AGE             PAY_0             PAY_2
 Min.   :0.000   Min.   :21.00   Min.   :-2.0000   Min.   :-2.0000
 1st Qu.:1.000   1st Qu.:28.00   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median :2.000   Median :34.00   Median : 0.0000   Median : 0.0000
 Mean   :1.552   Mean   :35.49   Mean   :-0.0167   Mean   :-0.1338
 3rd Qu.:2.000   3rd Qu.:41.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000
 Max.   :3.000   Max.   :79.00   Max.   : 8.0000   Max.   : 8.0000
     PAY_3             PAY_4             PAY_5             PAY_6
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
 Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000
   BILL_AMT1         BILL_AMT2        BILL_AMT3         BILL_AMT4
 Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000
 1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327
 Median :  22382   Median : 21200   Median :  20088   Median :  19052
 Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263
 3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506
 Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586
   BILL_AMT5        BILL_AMT6          PAY_AMT1         PAY_AMT2
 Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :      0
 1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:    833
 Median : 18104   Median :  17071   Median :  2100   Median :   2009
 Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :   5921
 3rd Qu.: 50190   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:   5000
 Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259
    PAY_AMT3         PAY_AMT4         PAY_AMT5           PAY_AMT6
 Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0.0
 1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8
 Median :  1800   Median :  1500   Median :  1500.0   Median :  1500.0
 Mean   :  5226   Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5
 3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0
 Max.   :896040   Max.   :621000   Max.   :426529.0   Max.   :528666.0
    DEFAULT
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.2212
 3rd Qu.:0.0000
 Max.   :1.0000
```

We have to turn the various numerical features representing categories such as gender, education, marital status, age and history of past payment into factors and later one need to be one hot encoded. Also, a few values in the demographic features are zero, which is not defined in the legend. We can remove these outliers as they only make up 68 instances out of the 30k we have available.

Exploratory Visualization

When looking at the various demographic feature we can see that the distribution of defaulting to not defaulting is fairly similar to the dataset as a whole (about 80/20)

```
   DEFAULT
SEX         0         1
  1 0.7583277 0.2416723
  2 0.7922372 0.2077628
         DEFAULT
EDUCATION          0          1
        0 1.00000000 0.00000000
        1 0.80765234 0.19234766
        2 0.76265146 0.23734854
        3 0.74842384 0.25157616
        4 0.94308943 0.05691057
        5 0.93571429 0.06428571
        6 0.84313725 0.15686275
        DEFAULT
MARRIAGE          0          1
       0 0.90740741 0.09259259
       1 0.76528296 0.23471704
       2 0.79071661 0.20928339
       3 0.73993808 0.26006192
```

This is suggesting that payment history, payment delay and pf given credit are better predictors for defaulting. This can be confirmed by below correlations to the target:

LIMIT_BAL	-0.153519876
SEX	-0.039960578
EDUCATION	0.028006078
MARRIAGE	-0.024339216
AGE	0.013889834
PAY_0	0.324793728
PAY_2	0.263551202
PAY_3	0.235252514
PAY_4	0.216613637
PAY_5	0.204148914
PAY_6	0.186866362
BILL_AMT1	-0.019644197
BILL_AMT2	-0.014193218
BILL_AMT3	-0.014075518
BILL_AMT4	-0.010156496
BILL_AMT5	-0.006760464
BILL_AMT6	-0.005372315
PAY_AMT1	-0.072929488
PAY_AMT2	-0.058578707
PAY_AMT3	-0.056250351
PAY_AMT4	-0.056827401
PAY_AMT5	-0.055123516
PAY_AMT6	-0.053183340

Solution Statement

A solution to this problem will be a classification model such as a logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, one-vs-rest.

We are going to fit the dataset on a decision tree which is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. It is one way to display an algorithm that only contains conditional control statements. With that we are identifying the F1 score as a measure precision and recall and then using hyper parameter tuning, boosting and/or a different model to improve the score.

Algorithms and Techniques

The decision tree is fit with the standard algorithms. The quality of a split is measured by the Gini impurity with the best splitter and no maximal depth. The minimum number of samples required to split 2 and the minimum number of samples required to be at a leaf node is 1. The samples have equal weight and we do not limit the amount of features when looking at the best split. Since the decision tree requires minimal data transformation and feature engineering it is a good model to start with.

Benchmark Model

When fitting a decision tree with above described default algorithm we are getting a 0.99 score for the train set which is to be expected as we didn't specify maximum depth and decision trees can easily over fit on a data set. We are also getting a score of 0.72 on the test set and an F1 score of 0.40 for the test set's actual target values and the predicted values by the model. This has room for improvement and next we would apply a boosting method such as gradient boosting, try to set a maximum depth for the decision tree model fitting and identify other hyper parameters that can improve the model.

Project Design

In order to identify the best model for this project we should fit on various models such as logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, etc instead of juts fitting the decision tree. Once we have scores for the various models we can select the most promising to further improve upon with hyper parameter tuning and/or boosting.