CIE 417 Project – Task 1

Names

Ahmed Mahmoud – 201800683

Ahmed Elghamry – 201801254

Problem definition

Nowadays, credit cards have become a crucial part of our lives due to the easy usage and pay-back options offered by credit card companies. Also, discounts and offers on credit cards make them alluring to people. Although it may seem that using credit cards would ease human lives, it could be a debt trap if not used wisely. Hence, we chose this dataset to monitor the behavior of adults using credit cards and how they pay their debts and bills.

Dataset description

This dataset carries information about the default payments, credit data, history of payment, bill statements, and demographic factors of credit card clients in Taiwan from April to September 2005. The data set contains 25 features:

ID: ID of each client

LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit

SEX: Gender (1=male, 2=female)

EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

MARRIAGE: Marital status (1=married, 2=single, 3=others)

AGE: Age in years

PAY_0: Repayment status in September, 2005 (-1=pay duly, 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)

PAY_2: Repayment status in August, 2005 (scale same as above)

PAY_3: Repayment status in July, 2005 (scale same as above)

PAY_4: Repayment status in June, 2005 (scale same as above)

PAY_5: Repayment status in May, 2005 (scale same as above)

PAY_6: Repayment status in April, 2005 (scale same as above)

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)

BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)

BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)

BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)

BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)

PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)

PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)

PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)

PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

default.payment.next.month: Default payment (1=yes, 0=no)

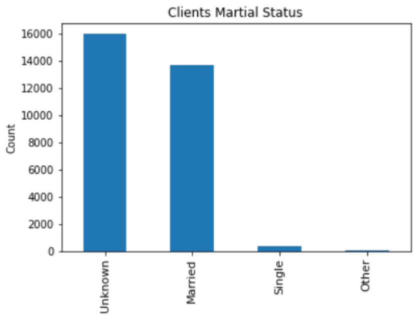
The features used in this dataset seem to be sufficient because all the information needed to predict whether a client is going to pay the default payment or not is available.

Approach and Methodology

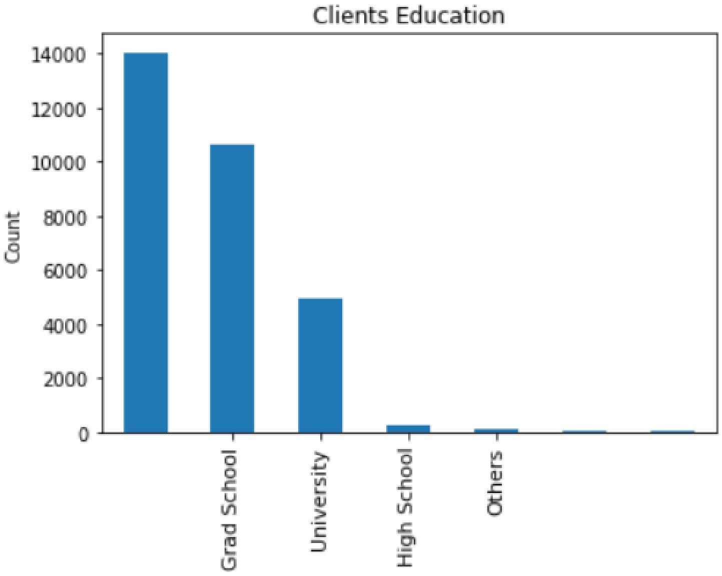
1. Data Preprocessing

The dataset included information about the clients (Age, Education, and Marital Status) and banking information regarding the payment and billing amounts in the previous six months. After visualizing this data, we have found some features that needs processing and cleaning to be used to fit our model like,

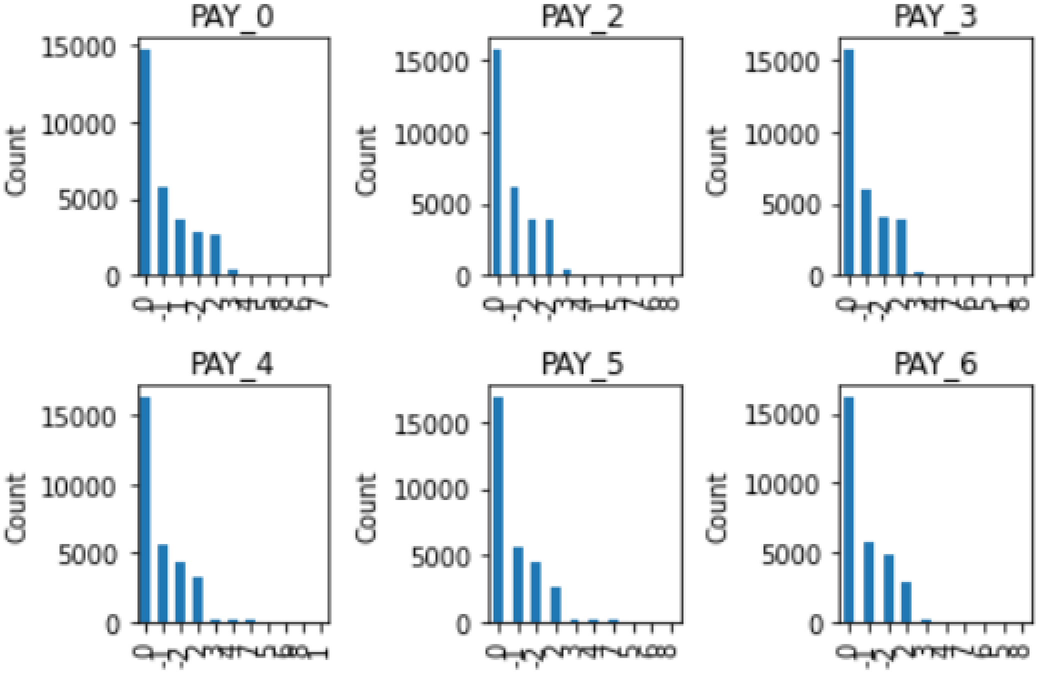
- **<u>MARRIAGE</u>**



- **EDUCATION**



- **PAY_0 and PAY_2 to PAY_6**



The Marital Status has unknown values which have value (0). Education also has unknown values of (0,5,6). Hence, data is cleaned by defining 0 as unknown in marital status, letting (0,5,6) be represented as 0 in Education. Moreover, we changed PAY_0 to PAY_1 to be consistent with PAY_AMT and BIL_AMT.

Moreover, after research, we have found that having 0 in PAY_1 to PAY_6 means the customer paid enough for their account to remain in good standing, but did revolve a balance as he/she did not pay the whole amount. And, having -2 means the credit card is inactive. This information is not detailed by the dataset owner, however, it exists in the dataset itself.

2. Model Selection

The target variable is a categorical binary variable to be classified. Hence, we chose to select models that work best on binary classification: Logistic Regression, KNN Classification, Boosting, and Bagging through Linear SVC and Decision Tree.

3. Model Development and Evaluation

To work on the best model, we first discarded all samples having unknown values, then, we dropped columns which are not correlated with the target variable, leaving only PAY_1 to PAY_6. Moreover, we oversampled the dataset to overcome the unbalanced target variable.

After testing all selected models, Decision Tree was found to be the best performing model to fit in our data with or without oversampling.

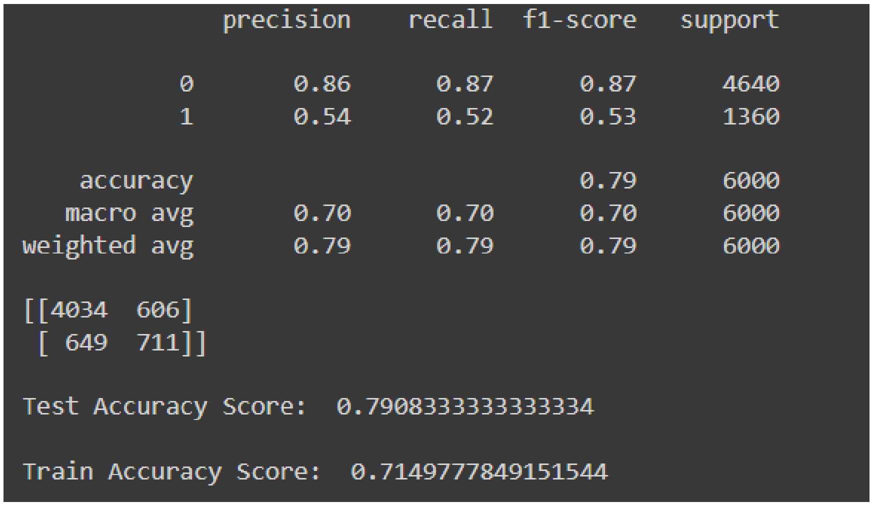
Implementation

Through this dataset we have used various libraries such as Pandas, Numpy, Seaborn and matplotlib.pyplot. In order to predict the target variable we used five models such as KNN, Logistic Regression, Boosting, Linear SVC, and Decision Tree. In the Decision Tree model, we used grid search to tune our model.

Discussion

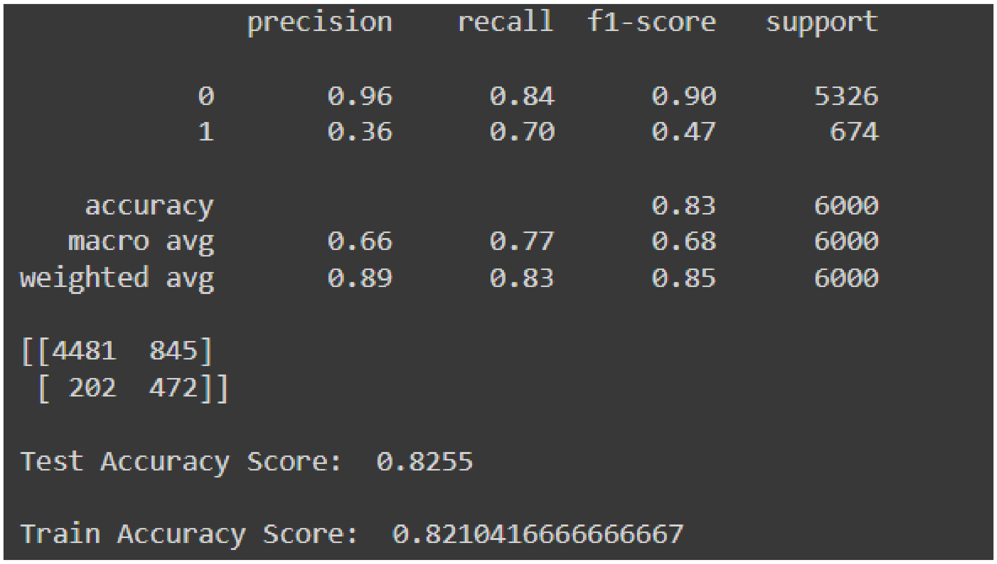
We tested the selected models with and without oversampling and found that the F1 score and the precision of having a target variable of 1 (the lower count) increased and the Recall decreased compared to those without oversampling. On the other hand, we found that the opposite happened to having a target variable of 0 (the higher count). Moreover, the accuracy of predicting the test data is higher than without oversampling is higher than that with oversampling.

After choosing the best features that is correlated with the target variable and trying two other models such KNN and Logistic Regression we found that Decision tree model after tuning out perform KNN and logistic regression as the accuracies were 82.55% (Decision Tree), 80.9% (KNN), 81.05% (Logistic Regression). Here are the scores of the Decision Tree with oversampling:



```
              precision    recall  f1-score   support

           0       0.86      0.87      0.87      4640
           1       0.54      0.52      0.53      1360

    accuracy                           0.79      6000
   macro avg       0.70      0.70      0.70      6000
weighted avg       0.79      0.79      0.79      6000

[[4034  606]
 [ 649  711]]

Test Accuracy Score:  0.7908333333333334

Train Accuracy Score:  0.7149777849151544
```

Here are the scores without oversampling:



```
              precision    recall  f1-score   support

           0       0.96      0.84      0.90      5326
           1       0.36      0.70      0.47       674

    accuracy                           0.83      6000
   macro avg       0.66      0.77      0.68      6000
weighted avg       0.89      0.83      0.85      6000

[[4481  845]
 [ 202  472]]

Test Accuracy Score:  0.8255

Train Accuracy Score:  0.8210416666666667
```

Compared to other models that use the same dataset, we successfully reached higher scores in accuracy, precision, recall, and F1 score. Most of the models have reached almost 80% accuracy. Not all models look for other scores, but most of them did not reach higher scores than 60%

Conclusion

According to our model testing and evaluation, Decision Tree is the best model to fit our dataset to predict whether a client will defaultly pay or not. A high accuracy and recall are obtained without oversampling while a high precision and F1 score are obtained while oversampling. For the sake of improvements one should tune the hyperparameters and replace or eliminate missing data,and finally try to manipulate the data in order to remove any obstacle that would decrease the accuracy of your model.