# UCLA Extension Data Science Certificate

## Final Exam

### Project: Default of credit card clients:
## Proposal

In this given classification problem I'm planning to use different modeling techniques and after that using pipeline to boost the models. For this I will first look at the data and see if needs any cleaning. I will try to normalize the data using mean, standard deviation. Then using correlation I will find the most correlated features. After these exploratory analyses I will identify the most important features and divide the data to test and training sets.

## Domain Background

The Credit Card Default dataset contains 30,000 records of customer information for a Taiwanese bank. The customer data included in the dataset are education, gender, age, marriage status, and the payment actions for a six-month period.

Any publications based on this dataset should acknowledge the following: Lichman, M. (2013). UCI Machine Learning Repository http://archive.ics.uci.edu/ml . Irvine, CA: University of California, School of Information and Computer Science.

The original dataset can be found here at the UCI Machine Learning Repository.

#### Problem Statement

One the best application of data science is to predict something before it happens and prevent it from happening. In this dataset we have some data to predict what are some risk factores resulting in an individual failing to make a payment on their credit card. We look at some information such as age, marital status, gender , etc. , to make a prediction and reduce the risk factors for the bank. Here I'll use the decision tree modeling to make the prediction.

### Metrics

Since this problem is a classification problem, F1 Score and Accuracy will be the metrics that I use for this dataset

#### Data Exploration

| Name | Explantion |
|---|---|
| limit_bal | Amount of the given credit (NT dollar): it includes both the individual consumer credit |

and his/her family (supplementary) credit.

sex Gender

(1 = male; 2 = female)

education Education

(1 = graduate school; 2 = university; 3 = high school; 4 = others)

marriage Marital status

(1 = married; 2 = single; 3 = others)

age Age (years)

pay_1 - pay_6 History of past payment. Past monthly payment records

From April to September, 2005 as follows:

pay_1 = the repayment status in September, 2005

pay_2 = the repayment status in August, 2005

...

pay_6 = the repayment status in April, 2005

The measurement scale for the repayment status is:

-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

bill_amt1-bill_amt5 Amount of bill statement (NT dollar).

bill_amt1 = amount of bill statement in September, 2005

bill_amt2 = amount of bill statement in August, 2005

...

bill_amt6= amount of bill statement in April, 2005

pay_amt1-pay_amt6 Amount of previous payment (NT dollar)

pay_amt1 = amount paid in September, 2005

pay_amt2 = amount paid in August, 2005

...

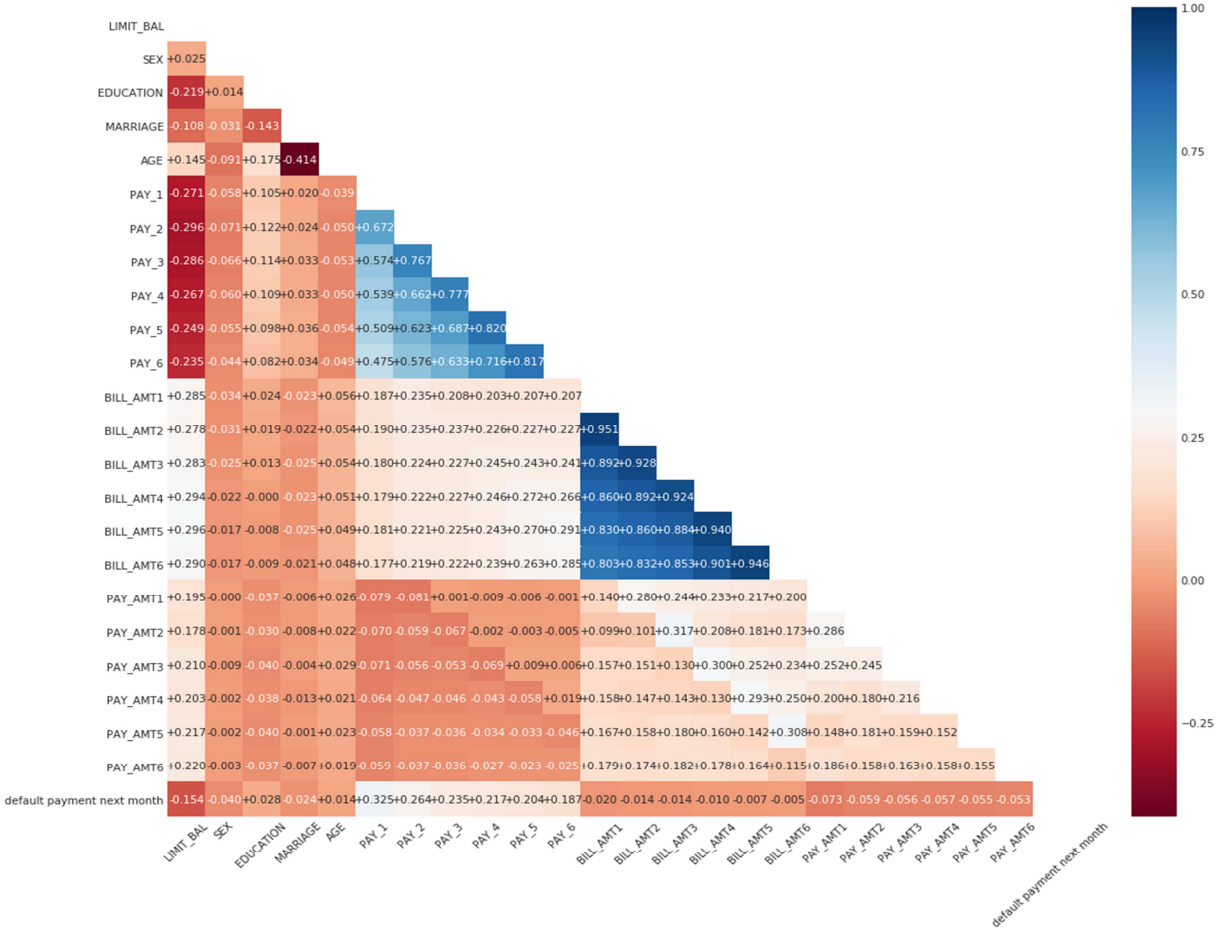
pay_amt6 = amount paid in April, 2005

After looking at data it's clear that there is missing column **PAY_1** wh**ich I replaced whith PAY_0.**

After that I droped the ID column to get in to normalizing the data using mean and sd.

**Exploratory Visualization**

In this EDA I used pair plot on the normalized data and also correlation heatmap.The pair plot shows how each variable behaves. The heat map shows the best that Age and marriage are highly correlated. Also PAY_2 and default payment next month are highly correlated.



**Solution Statement**

Using decision tree classifier I fit the train and test dataset  and it worked great, getting the (0.8196, 0.8196)for th e tain and test data. Then using boosting I got to improve the result to (0.8272, 0.820533333).

**Algorithms and Techniques**

GradientBoostingClassifier(),RandomForestClassifier(), and SVC was the algorithms that I used for this the model. The best  result was for the GradientBoostingClassifier().

**Benchmark**

The best benchmark for this modeling is Decision Tree. Given the result this dataset will best fit in this model. And we can make a better prediction on the data.

**Project Design**

From the plot on our Decision Tree model we can clearly see that it has work really well on our data.

The variables like Age and Marital status are clearly play a big role in data set.

My strategy would be to use Decision Tree modeling on the most important data and use GradientBoostingClassifier() , to make a better prediction.