# Influence Factors of Credit Card Default and Prediction Model

**Authors:**

Tianbo Zhang (Student ID: 1938501)

Peiran Chen (Student ID: 1935420)

**Executive Summary:**

**Central Research Problems:**

1. Determine the influence factors of credit card default.
   1. Limited balance, gender, marital status, gender, education level are all influence factors of credit card.
2. Find the prediction model with the highest accuracy.
   1. The best model for prediction of default customers is the Random Forest model.
3. Determine the effect of hyper-parameter with different values on model accuracy.
   1. Really interesting result!!! Detail explanation in the report.

**Motivation:**

Ever since the invention of credit card in the 1950s, the usage of credit cards has been increasingly popular among all age groups. With the routinize the use of credit card, the problem of credit card default became a severe problem in the financial market. Finding the influence factors of credit card default can help banks to determine their potential clients and the government may create new systems for the credit card markets to prevent credit car default. The discovery of a reliable prediction model will also help banks to action before potential credit card default.

**Dataset:**

The dataset contains the basic credit card client’s information in Taiwan. There are 30000 instances in the dataset. And there are 23 input variables and 1 output variable. The first 5 columns contain the data of amount of given credit, gender, education, marital status, age and the rest of the input variables are credit records. The output variable is the default payment (Yes = 1, No = 0).

URL: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

**Method:**

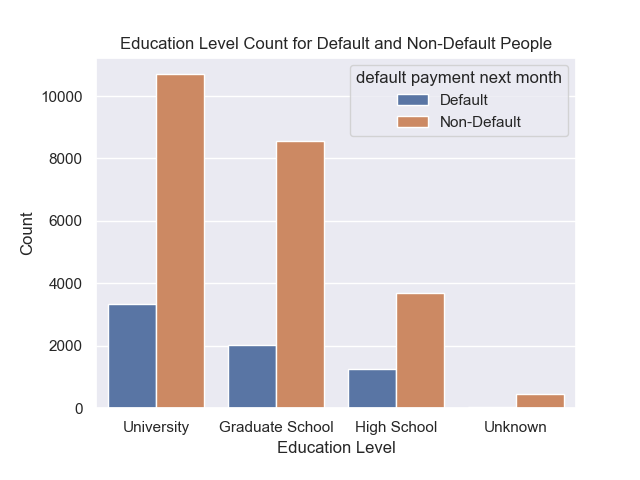
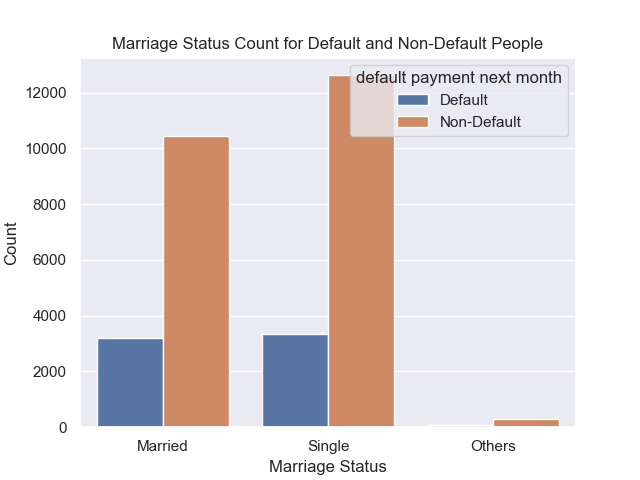
1. For question one, first fit the whole dataset with logistic classifier model. Note that, no need for train and test sets since we are just trying to find the influence factor, not the accuracy of the model prediction. Last calculate the T test score to check for the influence factor and look at the coefficient to see for different kinds of effect.
2. For question two, first separate the train and test data set. Then fit the data for each type of models and then look at the accuracy score of each model for the comparison.
3. For question three, take the amount of neuron from 20 to 190 with the step size of 20. Then calculate the accuracy score for the train set and test set prediction. Lastly graph the accuracy scores for visualization.

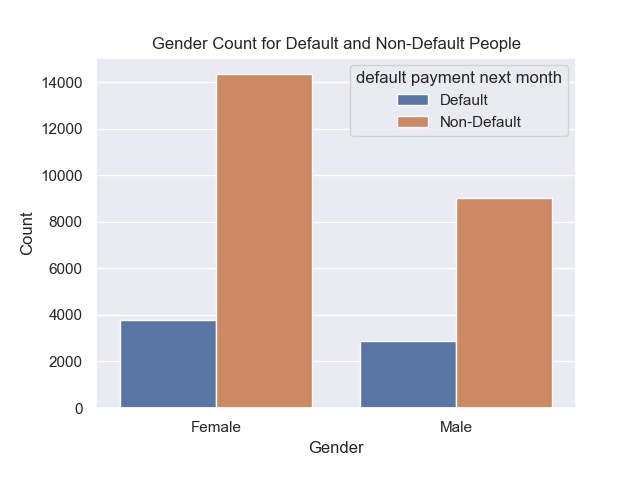
**Data Handling:**

* When we first look at the graph in the csv version, we noticed that the categorical data are easy for us to handle for model building (Male = 1, Female =2, etc.). However, we decided to change the numeric categorical data back into string representation and save it into a new data frame for visualization. We also decided to separate the original data set into default data set and non-default data set which may also help for visualizations. We also noticed that column 12 to 17 is the bill amount from April to September of 2015 and column 18 to 23 is the amount paid from April to September of 2015, so we decided to add the average of bill amount and paid amount into the data frame.

**Graphing**

* **Motivation:** Since we barely know the statistical information about the dataset, we decided to create visualizations for the data. We started with basic count graph for the categorical information. Then we create the joint graph for default and non-default customers. Lastly we create the distribution graph for the
* **Count Graph:**

****

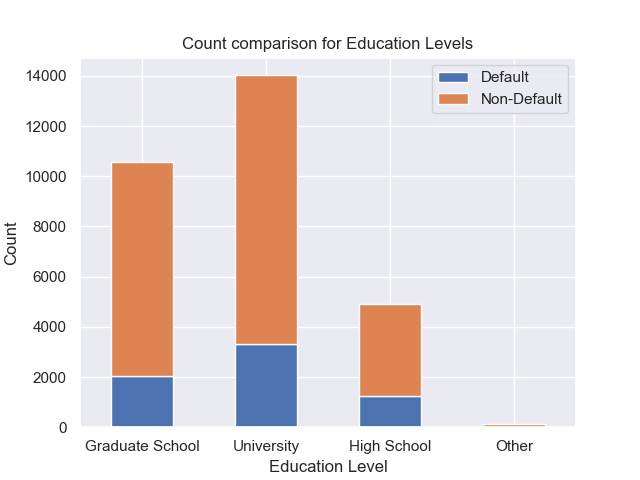
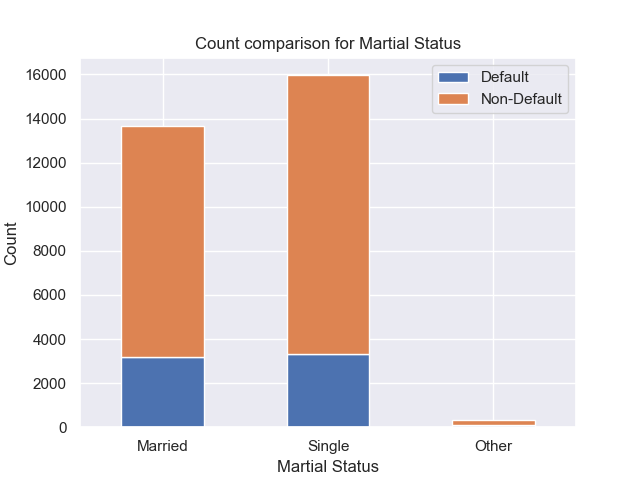
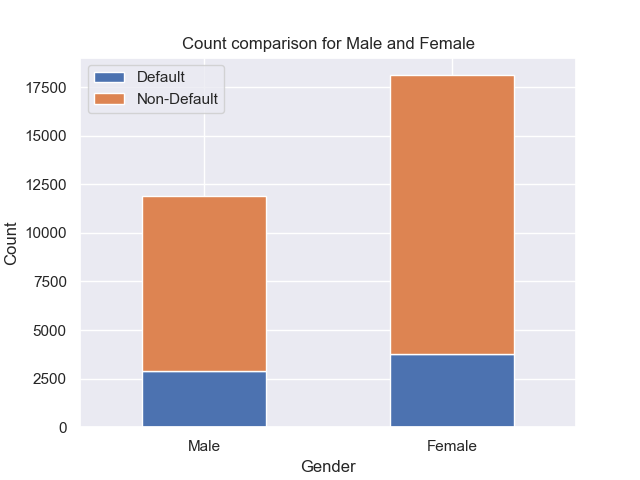
****

**General Interpretations:**

We noticed first that the amount of non-default people is remarkably more than the amount of default people, and we also noticed that the different ratios (gender ratio, education level ratio, marital status ratio) are similar.

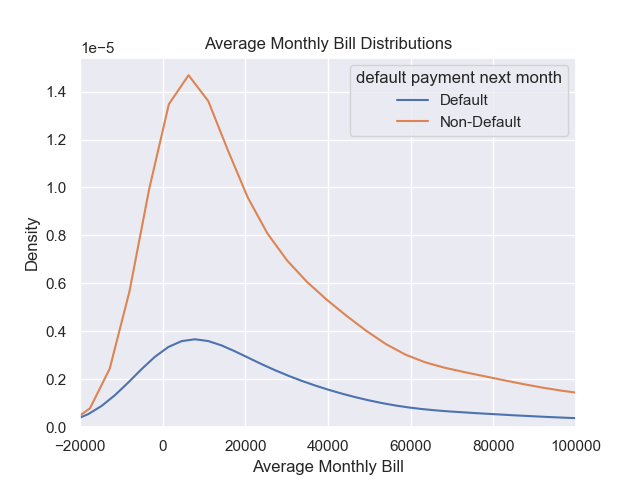
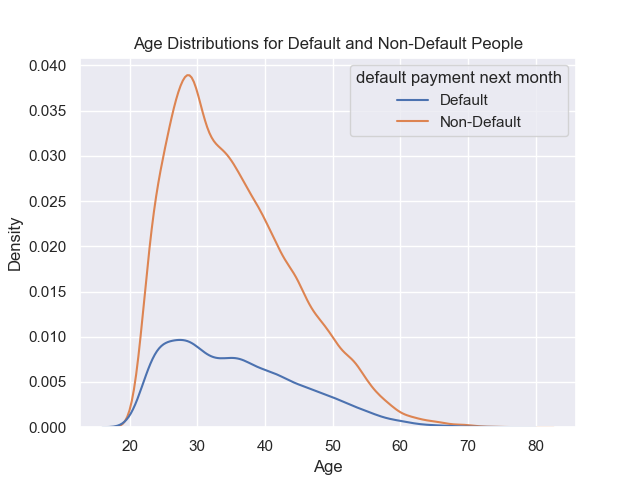
**Detail Interpretation:**

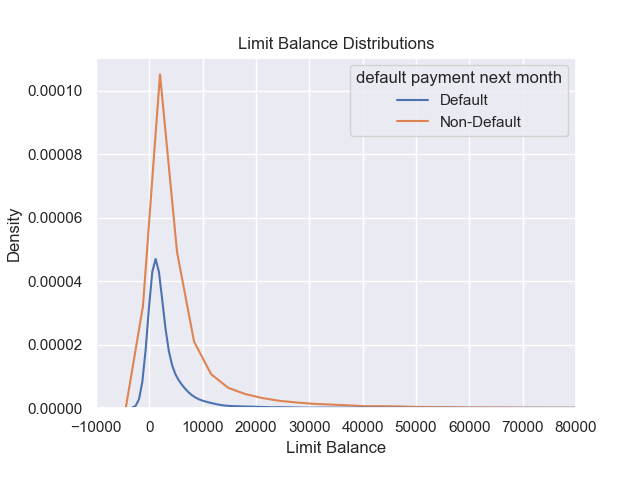
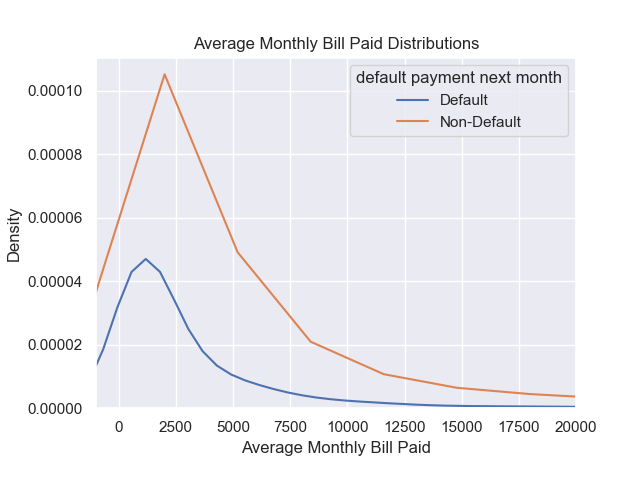
* Marital Status: The amount of single customers has the greatest proportion in the data set.
* Education Status: The amount of people that have been to the University has the greatest proportion in the data set.
* Gender: Female is more than man in the data set.
* **Joint Graph:**

****

**Interpretations:**

* Gender: We noticed that the ratio between different gender in non-default people is different from the ratio in default people.
* Marital Status: Again, the amount of married and single customers are about the same for default people, however single customers for that did not default is more than married customers that did not default.
* Education Level: The ratio is similar for customer who defaulted and those who did not default.
* **Distributions**



****  
**Interpretation:**

* Age distribution: Note most custsmers distributed around the age of 30 (Especially in the non-default people.
* Average Monthly Bill distribution: Note most customers distributed around the average bill of 10000 (Especially in the non-default people).
* Average Monthly Paid distribution: Note most customers distributed around the average bill of 10000 (Especially in the non-default people).
* Limit Balance Distribution: Most customers distributed around the limited balance of 2000.

**Result**

**Influence Factor (Question 1):**

* **Dummies:** Note that the marriage status has 4 levels, although they are all numeric, but the relationship between the numbers do not indicate the relationship between different numeric, so we created dummies for the marriage variable.
* **Hypothesis:**
  + Alpha: 0.05
  + H0:
    - Limit Balance has negative effect on the default result (more limited balance are less likely to default)
    - Age has negative effect on the default result (older people are less likely to default)
    - Education has negative effect on the default result (Since 1 denoted the highest education level, which means lower education level people are less likely to default)
    - Gender do not have effect on the default result
    - Marriage do not have effect on the default result
    - Average bill per month has positive effect on the default result (more amount of average bill indicate more likely to default)
    - Average bill paid per month has negative effect on the default result (more amount of average bill paid indicate less likely to default)
* **Logistic Model Stat Summary:**

**coef std err z P>|z| [0.025 0.975]**

--------------------------------------------------------------------------------------

LIMIT\_BAL -3.148e-06 1.48e-07 -21.285 0.000 -3.44e-06 -2.86e-06

AGE 0.0030 0.002 1.786 0.074 -0.000 0.006

EDUCATION -0.0833 0.019 -4.303 0.000 -0.121 -0.045

SEX -0.1828 0.029 -6.400 0.000 -0.239 -0.127

M\_1 -0.2561 0.092 -2.789 0.005 -0.436 -0.076

M\_2 -0.4832 0.079 -6.129 0.000 -0.638 -0.329

M\_3 -0.3655 0.159 -2.306 0.021 -0.676 -0.055

AVG Bill Per Month 3.525e-06 2.74e-07 12.854 0.000 2.99e-06 4.06e-06

AVG Paid Per Month -5.292e-05 3.62e-06 -14.599 0.000 -6e-05 -4.58e-05

* **Conclusion:**
  + First Note from the P > |z| (T test score), we can see that only Age variable has the value greater than 0.05(Alpha), thus we can conclude that only age do not influence the default status.
  + Variable Influences (From coefficient column):
    - Limit Balance has negative effect on the default result (more limited balance are less likely to default)
    - Education has positive effect on the default result (Since 1 denoted the highest education level and higher number denoted lower education level, which means lower education level people are less likely to default)
    - Gender has negative effect on the default result (Female is more likely to default)
    - Marriage has negative effect on the default result (Single customers are more likely to default)
    - Average bill per month has positive effect on the default result (more amount of average bill indicate more likely to default)
    - Average bill paid per month has negative effect on the default result (more amount of average bill paid indicate less likely to default)
* **Discussion:**
  + Why not age?
    - We think the reason is that default is common among all age group where there isn’t difference in the default status.
    - Another reason might be there isn’t enough data to make conclusions.
    - Last reason is that maybe there is collection bias
  + Education: It’s really interesting to see the result different from our hypothesis. However, relating the result to the central statement from the book *Poor Economics,* there is a really common idea among lower education people is where “bank is not trustworthy.” We concluded the reason is that lower education people only buy thing they need, rather than things they wanted, which means that people are less likely to spend a lot of money using credit card, hence less likely to default. Another reason is that maybe lower level education people are less likely to even have a credit card.
  + Marriage: We first hypothesized that marriage will not affect the default status, however, from the t test and coefficient, we can see that single people are more likely to default. We conclude that single people has less responsibilities, hence they have more freedom to decide where they spend the money, and when they are buying an reasonless product, no one will stop them.
  + Gender: Discussion in the Ethics portion.

**Best Model (Question 2):**

* **Data:** We first separated the data in to train and test data set. And since we are trying to find the accuracy score for different models, there is no need to only use the Average amount for different continuous month variables (Bill amount for April to September, Bill paid amount for April to September).
* **Model:** Since we are just looking at the accuracy of each model, we do not change the default hyper-parameter to make sure the fairness of different models.
* **Result:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Logistic | SVM | DecisionT | RandomF | Neural | K-Nearest | Adaboost |
| 0.779 | 0.779 | 0.73 | 0.819333 | 0.642667 | 0.750167 | 0.817667 |

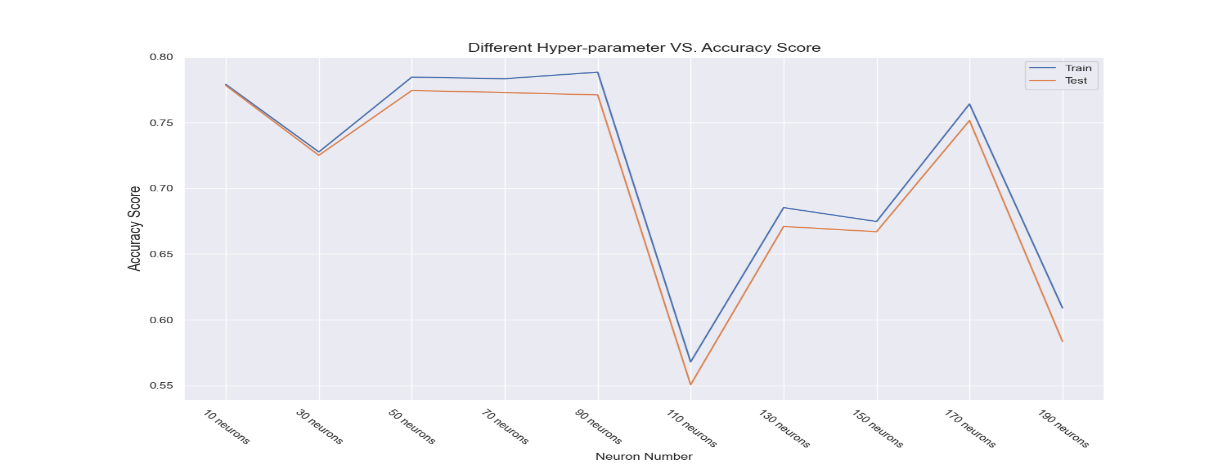
* **Interpretations:** From the accuracy score from different models, we can see that Random Forest model has the highest accuracy score. Thus the random forest model is the “best” prediction model with the default hyper-parameters. After some research into this certain type of model, we found that Random Forest model combines the simpatplicity of decision trees with the flexibility which will result in a vast improvement in the accuracy.

**Hyper-parameters of Neural Network (Question 3):**

* **Data:** We first separated the data into train and test data set. Then save the accuracy scores for train and test set with the different neuron number (Hyper-parameter).
* **Range:** We decided to start from 20 neurons up to 190 neurons by the step size of 20, since the default neuron number is 100.
* **Result:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 10 Neurons | 30 Neurons | 50 Neurons | 70 Neurons | 90 Neurons |
| Train | 0.779167 | 0.72775 | 0.784708 | 0.7835 | 0.7885 |
| Test | 0.778333 | 0.725167 | 0.7745 | 0.773 | 0.771167 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 110 Neurons | 130 Neurons | 150 Neurons | 170 Neurons | 190 Neurons |
| Train | 0.567875 | 0.685333 | 0.674792 | 0.764292 | 0.609 |
| Test | 0.5505 | 0.671 | 0.667 | 0.751667 | 0.5835 |



* **Interpretation:** The result is really interesting. We thought the accuracy score for the train and test group will be somewhat monotonic. However the pattern is not monotone. Our explanation for this pattern is where the neural network has not reach a stable status, hence there’s an unpredictable pattern.

**Challenge Goal:**

* **Messy Data:**
  + We thought that our data was easy to handle, however, it is not!
    - Cateneucoeparagorical data being numeric, hence it’s really difficult to graph at first. We changed the nerumeric data back into normal labels (1 = Male, 2 = Female, etc.)
    - We noticed that we need to create dummies in order to train the model for marriage variable.
    - There are nan in the data set, so we had to remove them.
    - There are data collection error in the data, some data should not be in that variable.
* **Result Validty:**
  + We used a lot of statistical method for our first question. We built a hypothesis testing question and used t-score and coefficient of variables to validate out hypothesis
* **Machine Learning:**
  + This research is mainly built on model prediction, we used several models from the sklearn package. We also used accuracy score and train test data set split function from the package. So a lot of research were needed.
  + We also studied the effect of different hyper-paramereseter values for the neural network model.
* **External Library:**
  + We used statmodels package to get the statistic summary for question 1.

**Work Plan Validation:**

We started from our plan, however, a lot of problems appeared during the research. The plan is some where in the middle from the real research process. More researches were needed before creating the plan. The researches should include the basic informations about all the models, details about the data, what kind of graphs do we need? Etc.

**Testing:**

The testing portion for our research is not like the usual testing. Since the whole project is about model building, the only thing we needed to test is the variable themselves.

* Mean
* Count
* Maximum

**Policy:**

We decided to add this portion into our project to give some possible policies that can be passed in the future.

* Finance classes:
  + Like the result interpretation for question 2, we think that lower educated people should also learn about some basic finance knowledges. Better finance decisions may help them overcome their financial situation.
  + Higher educated people should also learn more about finance. Higher education does not mean better finance decisions. Maybe this will lower the default rate in among the higher educated people.
* New credit card policies for lower income people:
  + Our ideal policy is to increase the time for people to pay their bill with credit card.

**Ethics:**

From question 2, we noticed that there is an influence between different gender, so we would want to discuss the ethic problem if we actually use this information to give policies.

Options:

* Give direct policies
  + Then we will raise social injustice, since this is an prediction againest different gender, hence the government or banks should not give direct policies.
* See other dataset to build a more through model:
  + We believe the problem is not directly caused by different genders, there has to be some other problems which caused the difference.

Hence, we would suggest to look into other data set that are related to credit card record to come up with a policy.

**Collaboration:**

Professor Xiaohang Zhang

Field: Economic and Management

University: Beijing University of Posts and Telecommunications