**Credit Default Risk Analysis**

Created By

Nethra Shivanna

Credit One

Table of Contents

[Project Overview and Business Goal 3](#_Toc61475670)

[Data Exploration 3](#_Toc61475671)

[Data Source 3](#_Toc61475672)

[Dataset Information 3](#_Toc61475673)

[Data Preparation 4](#_Toc61475674)

[Visualization 5](#_Toc61475675)

[Machine Learning Approach 7](#_Toc61475676)

[Confidence 8](#_Toc61475677)

[Conclusion 10](#_Toc61475678)

[Recommendation 10](#_Toc61475679)

# Project Overview and Business Goal

Over the past year or so Credit One has seen an increase in the number of customers who have defaulted on loans they have secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away. They have enlisted the help of our Data Science team to design and implement a creative, empirically sound solution.

So, the problem we are trying to analyze is to identify which customers are at risk to default based on the history of their credit re-payments and helping our banking clients to decide if a customer has the potential to repay the credit of the bank and avoid any defaulters in future, without affecting the customers spending sentiments.

# Data Exploration

## Data Source

Name: I-Cheng Yeh email addresses: (1) icyeh '@' chu.edu.tw (2) 140910 '@' mail.tku.edu.tw institutions: (1) Department of Information Management, Chung Hua University, Taiwan. (2) Department of Civil Engineering, Tamkang University, Taiwan.

## Dataset Information

The dataset is in csv format with following attributes:

**Attributes**

{NOTE: The following is updated information from the source’s author}

This analysis employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. We used the following 23 variables as explanatory variables:

| **Attribute Name** | **Description** | **Categories** |
| --- | --- | --- |
| ID | Row ID |  |
| LIMIT\_BAL | Amount of the given credit (NT dollar) |  |
| SEX | Gender details | 1 = male; 2 = female. |
| EDUCATION | Education details of customer | 1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others. |
| MARRIAGE | Marital status of customer | 1 = married; 2 = single; 3 = divorce; 0=others. |
| PAY\_X [where X is numbers from 0,2,3,4,5,6] | History of past payment status (from  April to September, 2005) as follows: PAY\_0 = 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:  -2: No consumption; -1: Paid in full; 0: The use of revolving credit; 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\_AMTX [where X is numbers 1, 2, 3,4,5, 6] | 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\_AMTX [where X is numbers 1, 2, 3,4,5, 6] | 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. |  |
| default payment next month | client's behavior | Y=0 then not default, Y=1 then default |

## Data Preparation

1. Missing Values- There were no missing values found in the dataset, hence there was no need to remove or impute any data.
2. Renaming attributes-
   1. The PAY\_0 attribute was renamed to PAY\_1 for readability purpose alone.
   2. The default payment next month attribute was renamed to default\_pay.
3. Datatype conversion-
   1. SEX, EDUCATION, MARRIAGE, default\_pay, and PAY\_X attributes were converted from integer to categorical type.
   2. Monetary attributes were converted from integer to float.
4. Categorical attributes-
   1. EDUCATION- merged levels- 0, 4, 5, 6 to 4- Others. Also, renamed levels as-

1 = Graduate School; 2 = University; 3 = High School; 4 = Others

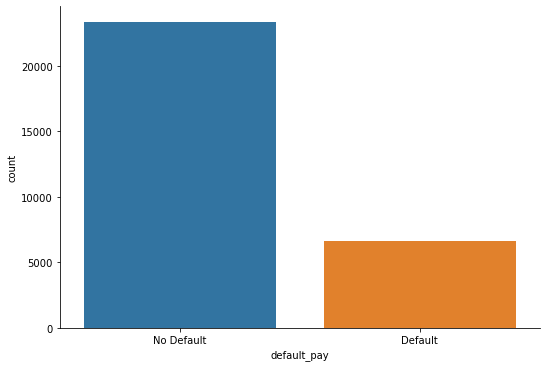
* 1. SEX- Renamed levels in SEX as 1- Male, 2- Female
  2. MARRIAGE- Renamed levels in Marriage as 1- Married, 2- single, 3- Divorce, 0- Others
  3. default\_pay- Renamed levels in default\_pay as 1- Default and 0- No Default

1. Dropping attributes- Dropped attribute ID, as this held only row ID info.
2. Discretize by Binning – Discretized AGE by binning into age groups of 20s,30s, 40s and so on, to understand the credit spending behavior of each age groups.

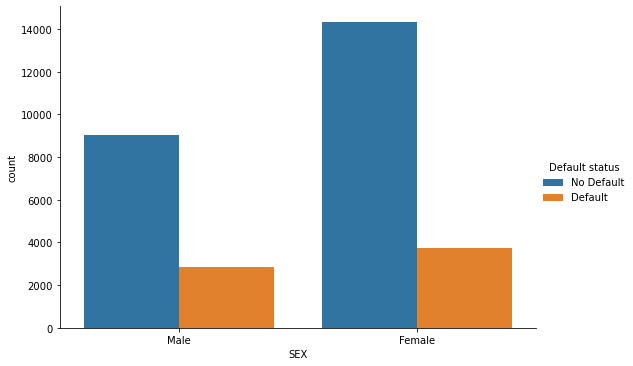
## Visualization

We performed visualization on certain attributes to understand the data and the relationship between each attribute and how each attribute is affecting the credit defaults/ No defaults.

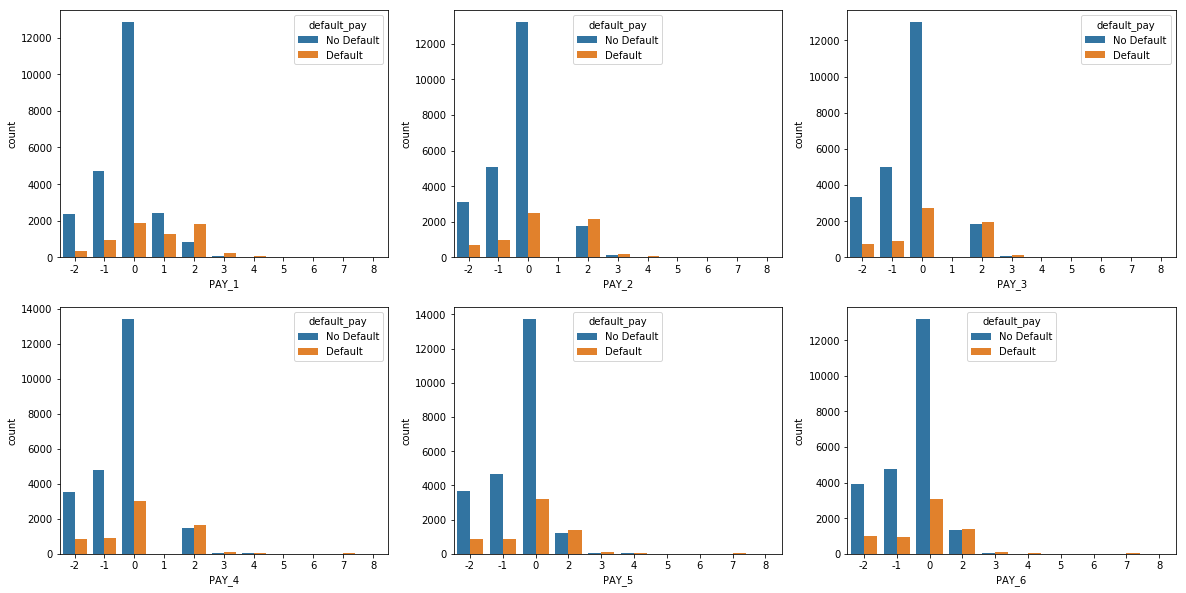
* We checked for the classification distribution in the dataset and found that the data is highly imbalanced with 78% of the population not defaulted and only 22% of population is defaulted.



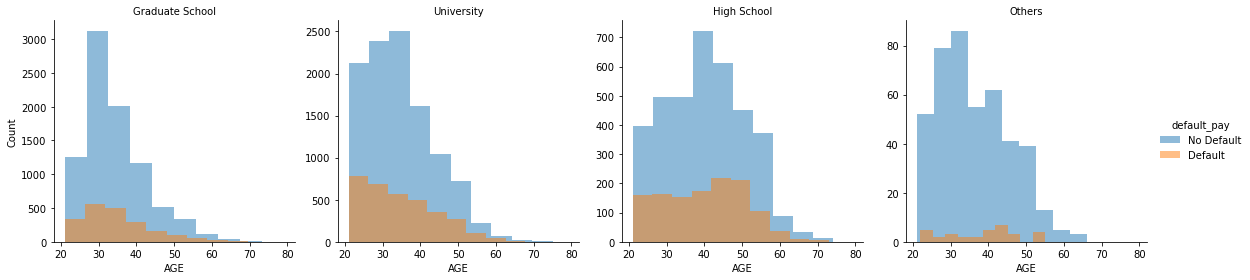
* We later discovered from the data that the majority of defaulters were Female.



* While visualizing the Payment Status attributes (i.e., PAY\_X attributes), we discovered that the majority of the customers are defaulted for delay in payment for 2 months.



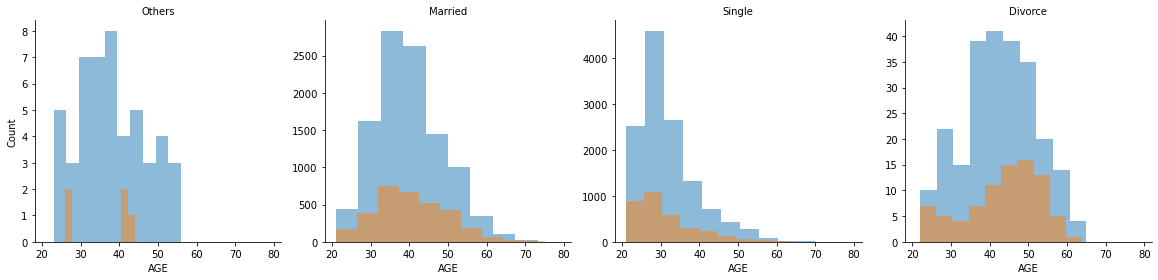
* We also tried to establish relationship between Age, Education, and Defaulters to see how the defaulters are distributed across the Dataset.



As per the above graphs, we can rank in order the defaulters as below:

1. Defaulters with age group 20-50 and with Education from University.
2. Defaulters with age group 20-50 and with Education from Graduate School.
3. Defaulters with age group 20-50 and with Education from High School.
4. Defaulters with age group 20-50 and with other qualifications.

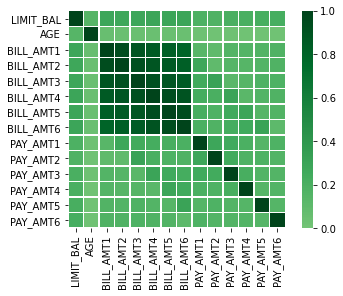
* We further explored the relationship between Age, Marriage, and Defaulters to see the data behavior



As per the above graph, we can rank in order the defaulters based on their marital status and age as below:

1. Defaulters with age group 20-50 and are Single.
2. Defaulters with age group 20-50 and are Married.
3. Defaulters with age group 20-50 and are Divorced.
4. Defaulters with age group 20-40 with other marital status.

* Correlation- As per the correlation heat map below:
  + We see that there is high correlation between the variables BILL\_AMTX and their correlation also increases with time. i.e., the BILL\_AMT2 is highly correlated with BILL\_AMT1 when compared to other BILL\_AMTX variables. This indicates that the Bill amount increases month by month.
  + We also noticed a kind of pattern that exists between BILL\_AMTX and PAY\_AMTX variables. The BILL\_AMT2 is highly correlated to PAY\_AMT1. Which means that the Amount of Bill Statement in August is correlated to Amount paid in September. This is a similar pattern observed between BILL\_AMTX and PAY\_AMTX+1.
  + Credit Limit is correlated to all variables. But, its comparatively high with variables BILL\_AMTX and PAY\_AMTX: we can consider 2 reasons here:
    1. Considering the increase in Previous Payment history, the customer would be provided with higher credit limit.
    2. Since the Credit Limit increases the Bill amount also increases.



# Machine Learning Approach

The following approaches were used in building best fit models for predicting defaulters. For in-dept analysis we used Scikit-Learn package to do the heavy lifting of the analysis.

* **Data Preprocessing**

For data preprocessing we used below listed methodologies to tune the dataset.

* 1. One-Hot encoder for classification attributes
  2. Min-Max scalar for normalization
  3. We used Pipeline to preprocess and create an estimator
  4. Used hyperparameters for model tuning.
* **Modelling**

For modelling, we split the data into 70:30 ratio as Training and Testing sets. The following listed classification models of scikit-learn package were used to train and test the data.

* 1. Logistic Regression
  2. Random Forest
  3. Support Vector Machine (SVM)- in this exercise- We used LinearSVC model to train the data
  4. K-nn

The Support Vector Machine (SVM), Random Forest, and Logistic Regression were the suitable models for the dataset. However, Support Vector Machine was best fitted than the other models

**Performance metrics**

Accuracy, Recall, Precision, F1 Score, ROC, and Confusion Matrix were used as performance metrics to choose best fit model.

# Confidence

As mentioned previously, the performance metrics we used to select the best fit classification models are Accuracy, Recall, Precision, F1 Score, ROC, and Confusion Matrix. All models were trained with GridSearch cross validation with K folds where K= 5.

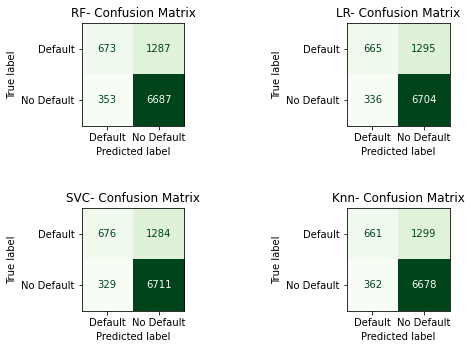
**Performance table**

|  | **Model** | **Model Score** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Execution Time** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Random Forest | 0.819429 | 0.817778 | 0.655945 | 0.343367 | 0.450770 | 2822.008861 |
| **1** | Logistic Regression | 0.820524 | 0.818778 | 0.664336 | 0.339286 | 0.449173 | 71.363686 |
| **2** | Support Vector Machine | 0.821238 | 0.820778 | 0.672637 | 0.344898 | 0.455987 | 27.309233 |
| **3** | Knn | 0.814476 | 0.817667 | 0.655914 | 0.342347 | 0.449883 | 227.069306 |

As per the above table, almost all models performed well above 80% score and 80% Accuracy. Although, since the data is imbalanced in classifying defaults and no defaults, we need to consider Precision and Recall to choose the best fit model.

**Confusion Matrix**

The confusion matrix of all models is shown below:



**Precision and Recall**

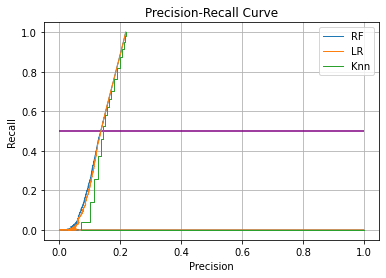
In order to predict defaulters, we need high recall and reasonable precision and since our data is imbalanced, we need to combine all metrics to choose a best fit model. Hence, along with high accuracy, we need high recall and high precision. But sometimes models can make mistakes like:

1. False Positive- predicting a customer will default when he/she will not
2. False Negative- predicting a customer will not default when he/she will.

If False Negative is higher, then the model will not be able to predict actual defaulters. Hence, we need to choose a model with fewer False Negative and higher True Positive.

The below Precision-Recall Curve shows that Random Forest has better Recall than Logistic Regression. However the Precision value of Random Forest is lower compared to Logistic Regression. But as per the Performance table above, Support Vector Machine has higher value of both Precision and Recall compared to all other models.

Note- LinearSVC() in scikit-learn has no predict\_proba() method. Hence, this model could not be used for Precision-Recall curve plotting.



The below plot gives an overview of performance metrics of all models

If we have to rank the models then – 1- Support Vector Machine; 2- Logistic Regression; 3- Random Forest; 4- K-nn.

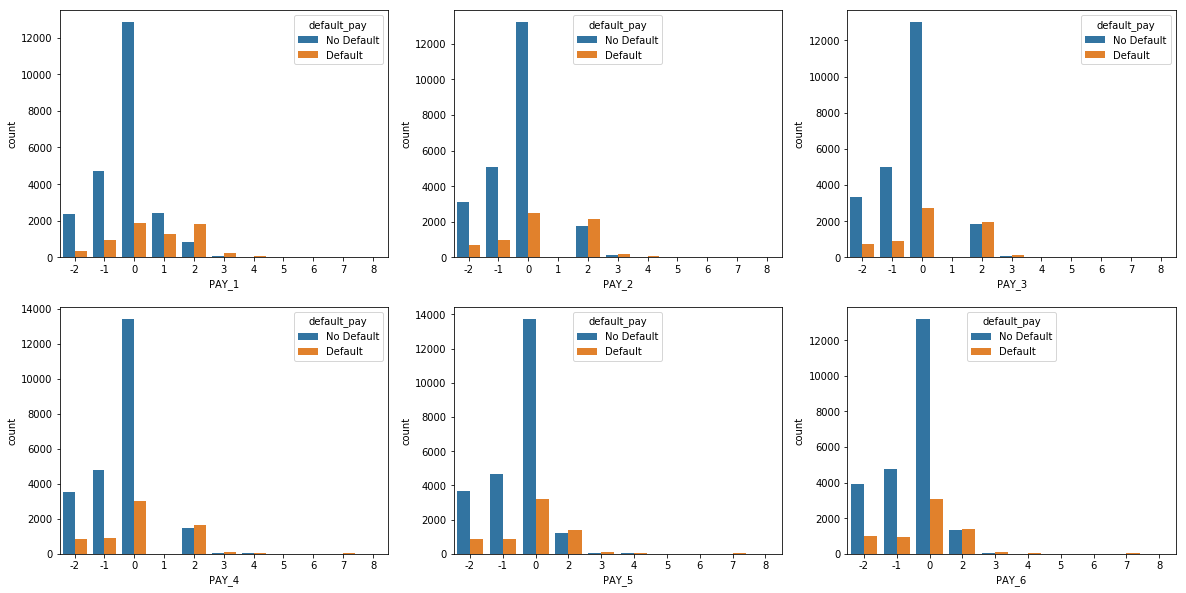
However, considering the metrics- Accuracy, Recall, Precision, F1 score- Support Vector Machine fitted well compared to other models with Accuracy of 82%, Recall- 0.344, Precision- 0.672, and F1 Score- 0.455. Also, the execution time of SVM was faster than other models with k-5 folds cross validation and best parameter: C= 0.1.

# Conclusion

In this project, we analyzed the performance of all models in order to predict the defaulting customers. Considering the customer spending sentiments, we used different performance metrics like- Accuracy, Precision, Recall, F1 Score, ROC curve to identify the best fitted model that has higher True Positive and lower False Negative. Choosing the model solely on accuracy did not make sense in this problem as the data was imbalanced. Hence, Recall and Precision played a major role in choosing the best fit model. The Support Vector Machine (SVM) had a balanced Recall and Precision rates along with an Accuracy of 82% when compared to the other models like- Random Forest, K-nn, and Logistic Regression. In terms of Accuracy all models performed well above 80%. But, since we also aimed for higher True Positive and fewer False Negative count, Support Vector Machine (SVM) performed well in providing the expected results.

# Recommendation

The below graph says that, the majority of the customers are defaulted for delay in payment for 2 months.



We can use this Payment Status has one of the attributes that could identify the potential defaulters in future.

Also, addition of new features with details of customer transaction history can help build the efficient model in predicting the potential customers to default in future.