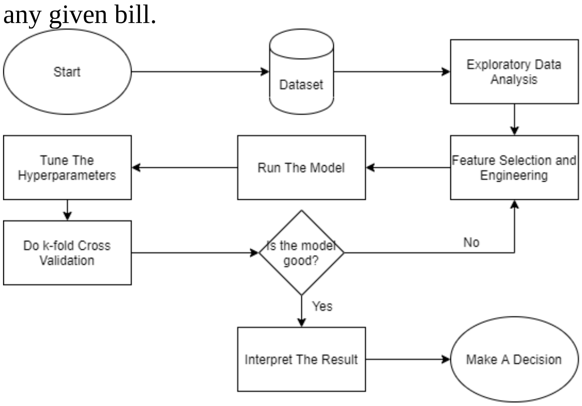
ANALYZING AND PREDICTING CUSTOMER'S DEFAULT AND IT'S CHARACTERISTICS BASED ON TAIWAN DATASET USING EDA AND DECISION TREE

Adjie Buanafijar
IPB University
adjie_buanana@apps.ipb.ac.id

Irfan Alghani Khalid
IPB University
Irfan_khalid@apps.ipb.ac.id

Chapter 1: Introduction to The Dataset

The dataset that will be used for this analysis is "Taiwan" dataset that consists of 30000 observations, 24000 on training data and 6000 on test data, and also has 25 features consisting of 24 predictors and 1 target. Predictor features consist of demographic, amount of payment, amount of bill, payment status, and the target feature which is default payment. Default payment is the condition where the customer fails to meet the obligations of the loan. In other words, the customer fails to pay any given bill.



Implementation design of the analysis

The graph above describes the steps that we will do on this analysis. To begin this analysis, we will use all of these features and make an analysis of it first, also get an insight into what characteristics of the customers that will get a default next month, then we select what feature we will use for predicting default, then we apply machine learning model. After we get a good feature, we will tune the hyperparameter on the model to get a good performance. Finally, we also see feature importance and then make a decision also interpretation from the analysis. The data itself is already clean and we only have to encode the data for visualizing the heatmap and also for doing predictive modelling. Because of that, we don't have to pre-process the data a lot and straight to analyze the data.

Chapter 2: Proposed Method

For doing data analysis, we propose a method using Decision Tree for Predictive Modelling and Exploratory Data Analysis for analyzing the characteristics of the customers. We implement this using R, Python, and it's supporting packages.

Exploratory Data Analysis: Exploratory Data Analysis (EDA) is a method for understanding and analyzing the data by knowing the correlation, distribution, and pattern by visualizing graph, summarization of the data, or by doing statistical summary on the data. This part is really important because by doing that we will know what kind of features that will affect the target.

Decision Tree Algorithm: Decision Tree is a tree-based Machine Learning model and this model works by splitting the data by the most impactful feature that contributes to the prediction. The split will go on until it reaches the end of the node, which is the default feature's value.

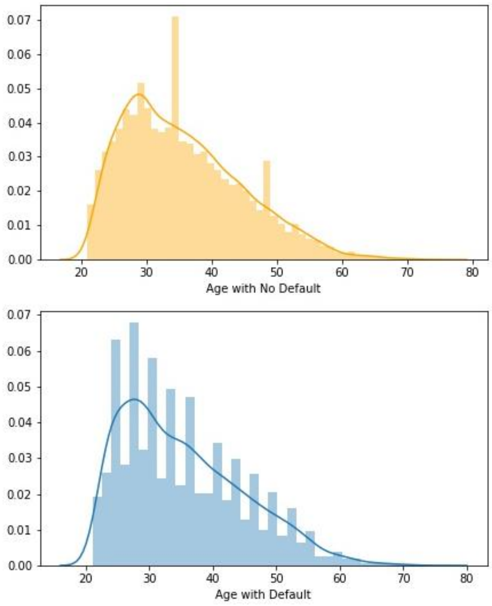
We use this algorithm because the model is capable of interpreting which feature that contributes to the prediction result. This algorithm yields a tree that each node represents which feature that contributes to the prediction so with that reason this model is interpretable. This is different compared to models such as Neural Network, Support Vector Machine, or Random Forest that those being called as the black-box model which cannot be interpreted directly by the data itself.

Chapter 3: Result of Analysis

Exploratory Data Analysis

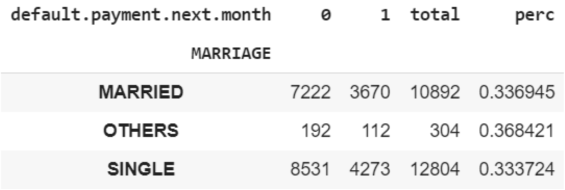
For this analysis, we would like to know what are the characteristics of the customer that will have default next month or not. Then, we make the model for predicting whether the customer does have default next month or not. Finally, we make a conclusion and suggestion based on what we have got. Now, we have to analyze each feature visually.

- AGE feature



Based on visualization above, we know that the distribution on the age based on the target has the same distribution, so there is no significant indicator that Age has contributed to the target.

- MARRIAGE feature



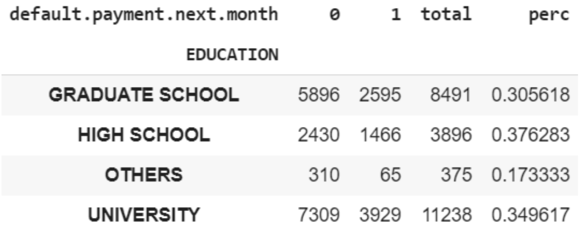
default.payment.next.month	0	1	total	perc
MARRIAGE				
MARRIED	7222	3670	10892	0.336945
OTHERS	192	112	304	0.368421
SINGLE	8531	4273	12804	0.333724

Contingency Table of MARRIAGE feature

Here is the marriage feature that has values of single, married, and others. Here, we know that single is the largest value on this feature. Also, there are ‘others’ values. We assume that this value means for divorced people and also LGBT people because in 2005, Taiwan still did not legalize LGBT marriage so we assume that is the value of others.

Based on the summarization above, we know that this feature does not significantly contribute to the default value. Despite that, we just know here from the summarization that people who married have a huge proportion for getting default.

- EDUCATION feature



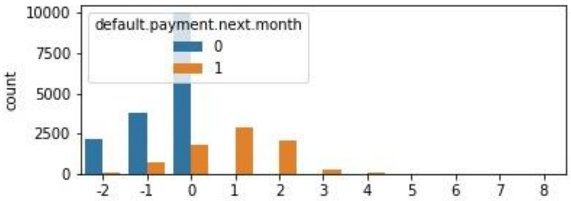
default.payment.next.month	0	1	total	perc
EDUCATION				
GRADUATE SCHOOL	5896	2595	8491	0.305618
HIGH SCHOOL	2430	1466	3896	0.376283
OTHERS	310	65	375	0.173333
UNIVERSITY	7309	3929	11238	0.349617

Contingency Table of EDUCATION feature

Here is the education feature that has values of university, graduate school, high school, and others. Here, we know that people with a university degree are the majority on this feature. Also, there are others' values. We assume that this value means for a person that didn’t have a greater education below high school.

We also make a contingency table of the feature, we infer that there’s no relationship between both features so it’s not useful if this feature is being used to predict the default. Beside that, we also know that the proportion for people with high school education have slightly higher proportion for doing default compared to customers that have higher education.

- PAY_ features



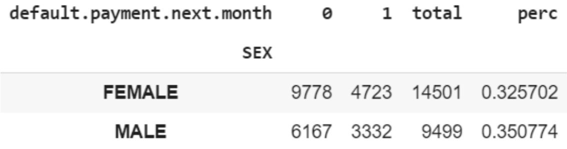
PAY_0 count plot. The other PAY_ feature shows the same distribution

PAY_Num feature describes the customer’s status whether they have paid the bill or not. The value consists of:

-2	Early Payment
-1	Paid on time fully
0	Paid not fully
1	Payment Delay for One Month
2, 3, ... , 9	Payment Delay for Two Months, and so on

Based on the visualization above, we know that the majority of people that will have default payment are the people that delay their payment in months despite there’s also got default and have paid the bill.

- SEX feature

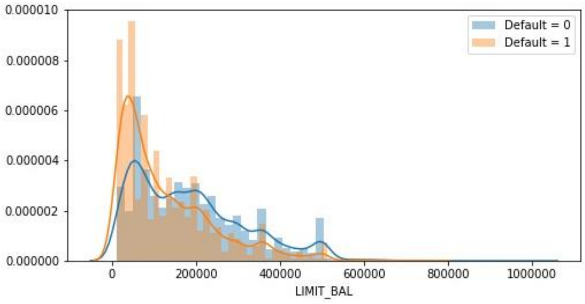


default.payment.next.month	0	1	total	perc
SEX				
FEMALE	9778	4723	14501	0.325702
MALE	6167	3332	9499	0.350774

Contingency Table of SEX feature

In SEX Feature, we know that the female is the majority of the customer in this data with value of 14501. From the summarization itself, we do not find any interesting pattern on it and has no direct impact on predicting default or not. Despite that, when we make a contingency table, we know that the male customer has a slightly higher proportion for doing getting default than female customer.

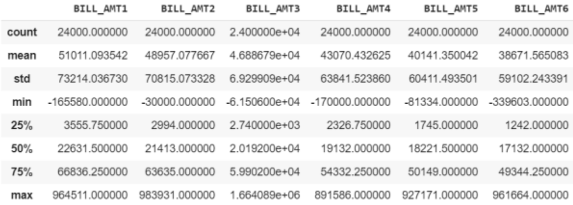
- LIMIT_BAL feature



Distribution Plot of LIMIT_BAL

LIMIT_BAL feature is a feature that tells us the amount of given credit in NT dollars. Based on this visualization, we know that if the LIMIT_BAL value is above 100000 there are a lot of customers that will not get default next month compared to those that have less than 100000. So, LIMIT_BAL has an impact on predicting the target value.

- BILL_AMT features

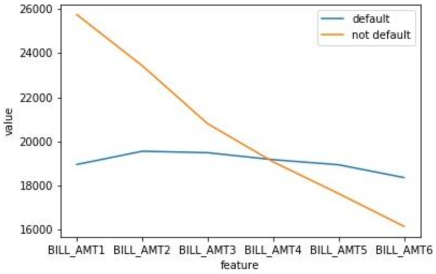


	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6
count	24000.000000	24000.000000	2.400000e+04	24000.000000	24000.000000	24000.000000
mean	51011.093542	48957.077667	4.688679e+04	43070.432625	40141.350042	38671.565083
std	73214.036730	70815.073328	6.929909e+04	63841.523860	60411.493501	59102.243391
min	-165580.000000	-30000.000000	-6.150600e+04	-170000.000000	-81334.000000	-339603.000000
25%	3555.750000	2994.000000	2.740000e+03	2326.750000	1745.000000	1242.000000
50%	22631.500000	21413.000000	2.019200e+04	19132.000000	18221.500000	17132.000000
75%	66836.250000	63635.000000	5.990200e+04	54332.250000	50149.000000	49344.250000
max	964511.000000	983931.000000	1.664089e+06	891586.000000	927171.000000	961664.000000

Statistical summary of BILL_AMT features

BILL_AMT features describe how much the bill is in a given month. Based on the statistical summary above, given that there are also negative value of the bill amount. This value describes the amount of saving money the customer has.

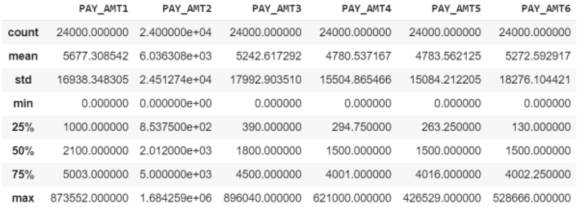
We would like to know how the dynamics of the BILL_AMT feature are given time. To visualize this, we will use median to get the central tendency of the feature. The reason for using median is because it is robust to the outliers so the result using that makes sense.



Line Plot that describe the movement of BILL_AMT features in general based on default feature

Here, we know that people that are not default will have a constantly increasing BILL_AMT features rather than those who have default. The default one has a different pace of BILL_AMT given time and it's nearly constant. It's reasonable because the people with default will not spend a lot of money.

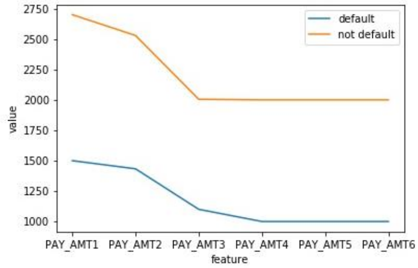
- PAY_AMT features



	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
count	24000.000000	2.400000e+04	24000.000000	24000.000000	24000.000000	24000.000000
mean	5677.308542	6.036308e+03	5242.617292	4780.537167	4783.562125	5272.592917
std	16938.348305	2.451274e+04	17992.903510	15504.865466	15084.212205	18276.104421
min	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000
25%	1000.000000	8.537500e+02	390.000000	294.750000	263.250000	130.000000
50%	2100.000000	2.012000e+03	1800.000000	1500.000000	1500.000000	1500.000000
75%	5003.000000	5.000000e+03	4500.000000	4001.000000	4016.000000	4002.250000
max	873552.000000	1.684259e+06	896040.000000	621000.000000	426529.000000	528666.000000

Statistical Summary of PAY_AMT features

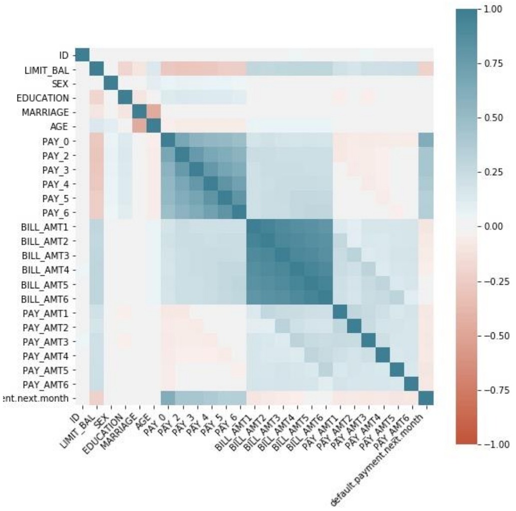
PAY_AMT feature is a feature that describes how much the customer pays in a given month. When we visualize it in given time and take the median value of it, there is no significant dynamics that could describe people with default or not despite the fact that we know that the default has a lower PAY_AMT than the not default.



Line Plot that describe the movement of PAY_AMT features in general based on default feature

Heatmap Analysis and Feature Selection

Beside we analyze single features, we also analyze features that actually have a relationship on it. Given this heatmap, we will work on predictors that have interesting work to do.



Also, based on this heatmap, it shows that there are interesting patterns that we have found here. First, the AGE and MARRIAGE features correlate each other negatively as it shows the color is almost dark red. For doing some kind of modelling, we have to remove one of the predictors that correlate with each other. For this case, we take out the AGE feature based on the exploration that have done before. In addition, the AGE feature doesn't give any meaningful information that could separate the default feature.

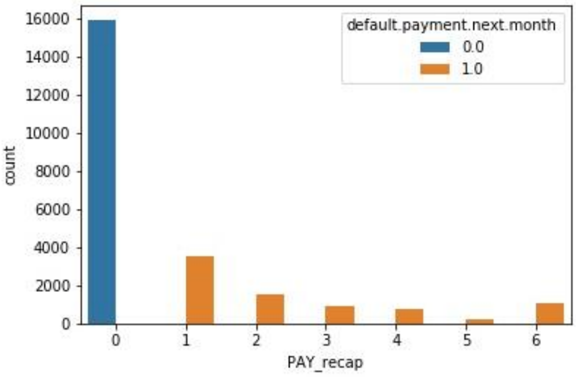
Second, we found an interesting pattern at BILL_AMT features. It shows that these features correlate with each other. It is reasonable because they have time related to it. We know that BILL_AMT1 describes the bill amount in September 2005 and then BILL_AMT6 describes the bill amount in April 2005. As we analyze the line plot from these features, it has shown an interesting pattern on it. So, we will keep this feature.

Third, the LIMIT_BAL is negatively correlated to default feature as it shows light red color on the graph. Also, same as the PAY_AMT and BILL_AMT features that slightly negatively correlated with the default feature.

Last, the PAY_ features have a positive correlation with the default feature and each feature is also correlated with each other. As the analysis before on this feature, it's really great for separating the default feature. But, we have to be careful to use these features because it could make the model overfit to the data.

Feature Engineering

We see that the PAY_ features have a strong correlation with each other and it also has a strong correlation on the default feature. We will create a new feature based on the PAY_ features called PAY_Recap. To make this feature, we have to encode each feature if the value is greater than 0, set to 1 because it indicates the customer late to pay the bill and else the value will be 0. Then, we sum each feature to create the feature and the visualization with the default feature looks like this:

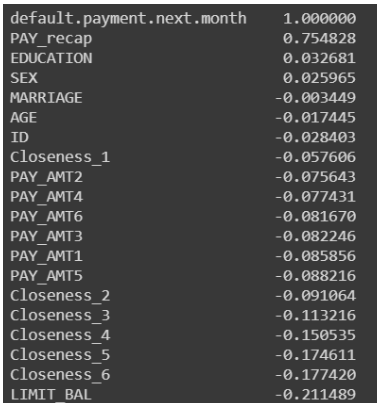


PAY_recap plot is really great for predict the feature

The PAY_recap is really great for separating the default feature, but we will not use this feature because it could make overfitting on the model. For the correlation value, the PAY_recap scores 0.754 which positively correlated with the feature. Because we don't want the model to overfit the data just because of the PAY_ feature, we remove the PAY_ features and not use the PAY_recap feature.

After we engineer the PAY_recap feature, we move on to create a new feature based on the BILL_AMT feature and LIMIT_BAL feature. From the data, we want to know how close the bill amount in the given month to the limit. This feature we name as closeness and there are 6 features of that feature based on the given month. Because of creating this feature, we remove the BILL_AMT features.

After we engineer the features, we come up with this correlation value of each feature with the default feature:

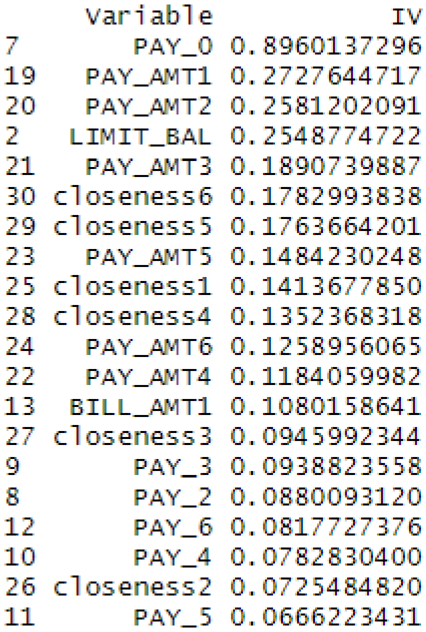


```
default.payment.next.month    1.000000
PAY_recap                     0.754828
EDUCATION                     0.032681
SEX                           0.025965
MARRIAGE                     -0.003449
AGE                          -0.017445
ID                           -0.028403
Closeness_1                  -0.057606
PAY_AMT2                     -0.075643
PAY_AMT4                     -0.077431
PAY_AMT6                     -0.081670
PAY_AMT3                     -0.082246
PAY_AMT1                     -0.085856
PAY_AMT5                     -0.088216
Closeness_2                  -0.091064
Closeness_3                  -0.113216
Closeness_4                  -0.150535
Closeness_5                  -0.174611
Closeness_6                  -0.177420
LIMIT_BAL                    -0.211489
```

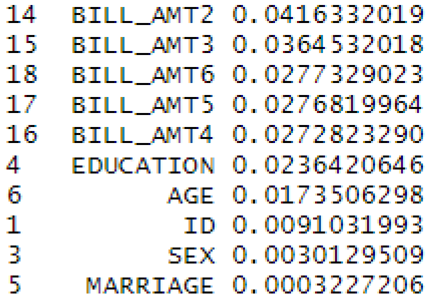
Here it shows the correlation value with the default feature.

Modelling

Now comes to the modelling part, we will be using Decision Tree Classifier for predicting the default feature based on the data that we have processed before. First step in modelling, selecting features that will contribute to the model. To analyze which feature with the most contribution to least contribution, we will use Information Value of each predictor.



```
      Variable            IV
7        PAY_0  0.8960137296
19    PAY_AMT1  0.2727644717
20    PAY_AMT2  0.2581202091
2    LIMIT_BAL  0.2548774722
21    PAY_AMT3  0.1890739887
30  closeness6  0.1782993838
29  closeness5  0.1763664201
23    PAY_AMT5  0.1484230248
25  closeness1  0.1413677850
28  closeness4  0.1352368318
24    PAY_AMT6  0.1258956065
22    PAY_AMT4  0.1184059982
13    BILL_AMT1  0.1080158641
27  closeness3  0.0945992344
9        PAY_3  0.0938823558
8        PAY_2  0.0880093120
12       PAY_6  0.0817727376
10       PAY_4  0.0782830400
26  closeness2  0.0725484820
11       PAY_5  0.0666223431
```

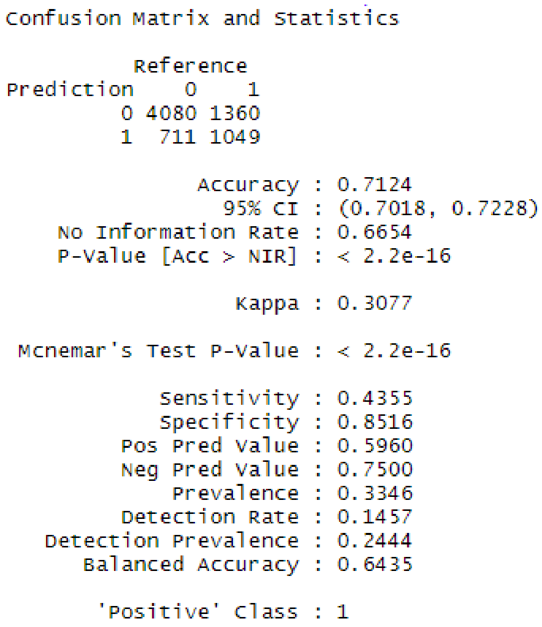


```
14   BILL_AMT2 0.0416332019
15   BILL_AMT3 0.0364532018
18   BILL_AMT6 0.0277329023
17   BILL_AMT5 0.0276819964
16   BILL_AMT4 0.0272823290
4    EDUCATION 0.0236420646
6          AGE 0.0173506298
1           ID 0.0091031993
3          SEX 0.0030129509
5      MARRIAGE 0.0003227206
```

By using Information Value, it helps to rank features on the basis of their importance. The interpretation of the values are :

- Less than 0.02 → not useful
- 0.02 to 0.1 → weak
- 0.1 to 0.3 → medium
- 0.3 to 0.5 → strong
- More than 0.5 → suspicious

Thus, we will drop MARRIAGE, SEX, AGE, and EDUCATION. Due to the feature engineering of Closeness, we will also drop BILL_AMT. As mentioned before, PAY_AMT doesn’t affect target features, based on it’s result on the median plot and correlation analysis. Therefore, the remaining features are ID, LIMIT_BAL, and Closeness. Next, we step into the Model creation. For model creation, we still use the default parameter given the model, so the result looks like below:

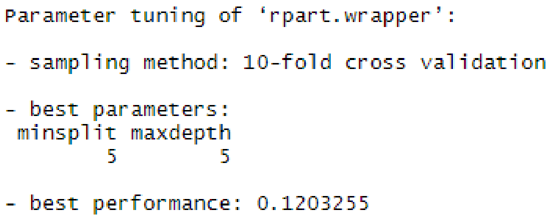


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4080 1360
         1  711 1049
                                          
               Accuracy : 0.7124          
                 95% CI : (0.7018, 0.7228)
    No Information Rate : 0.6654          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.3077          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.4355          
            Specificity : 0.8516          
         Pos Pred Value : 0.5960          
         Neg Pred Value : 0.7500          
             Prevalence : 0.3346          
         Detection Rate : 0.1457          
   Detection Prevalence : 0.2444          
      Balanced Accuracy : 0.6435          
                                          
       'Positive' Class : 1               
```

We know that the accuracy is around 71.24%, Specificity around 85.16%, and sensitivity is around 43.55% which is really bad. For additional information, specificity is also called as precision, it describes as how well the model could predict the default value correctly from all predicted value as default. Then, sensitivity is also known as recall, it describes how well the model could retrieve all default value from the data itself.

Based on the results before, we have to improve the model so it could predict the customer accurately. So, we will tune the parameters, which in this case the parameters are minsplit and maxdepth. Minsplit is a parameter that represents the minimum number of observations that must exist in a node in order for a split to be attempted. And the maxdepth represents how much the split will happen. To tune the parameter, we set the minsplit values around 5, 10, 15, and 20 and the maxdepth values around 1, 3, 5, 10, 15, 20, 25, and 30. With that value, we will do grid search to find the best parameter for the model.



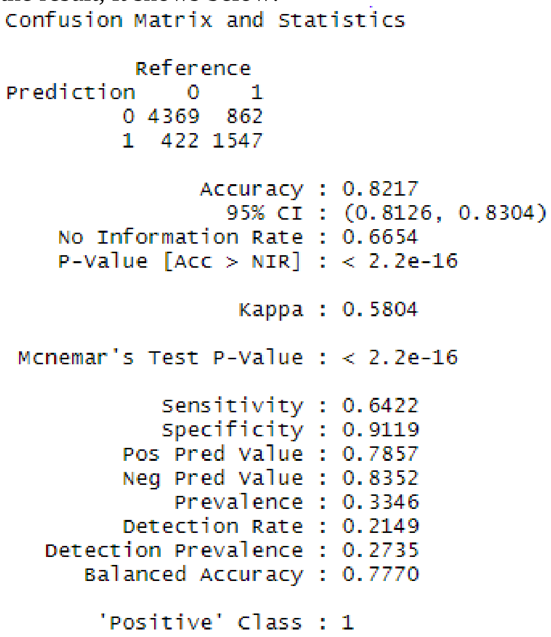
```
Parameter tuning of ‘rpart.wrapper’:

- sampling method: 10-fold cross validation

- best parameters:
 minsplit maxdepth
        5        5

- best performance: 0.1203255
```

As a result, we got the best parameter for minsplit and maxdepth, each value of the parameters are 5 and 5. Beside tuning the hyperparameters, we also bring back the PAY_AMT features, but only PAY_AMT1, PAY_AMT2, and PAY_AMT3. As the result, it shows below:

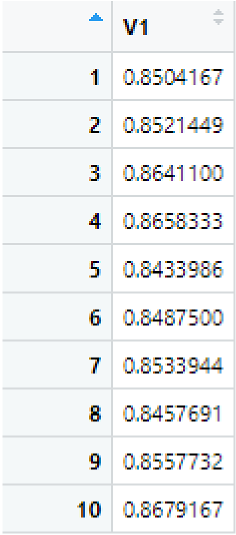


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4369  862
         1  422 1547
                                          
               Accuracy : 0.8217          
                 95% CI : (0.8126, 0.8304)
    No Information Rate : 0.6654          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.5804          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.6422          
            Specificity : 0.9119          
         Pos Pred Value : 0.7857          
         Neg Pred Value : 0.8352          
             Prevalence : 0.3346          
         Detection Rate : 0.2149          
   Detection Prevalence : 0.2735          
      Balanced Accuracy : 0.7770          
                                          
       'Positive' Class : 1               
```

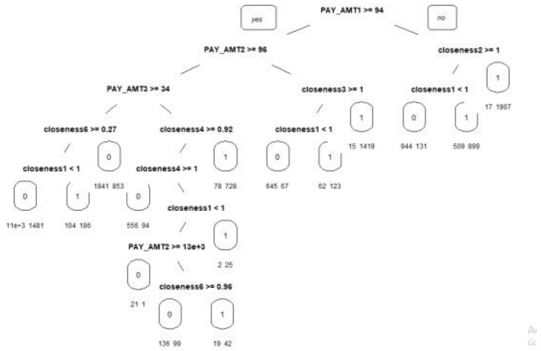
As we can see here, the performance of the model is improving. The accuracy of this model is 82.17%, the sensitivity is 64.22%, and the specificity is 91.19%.

To make sure that our model is great at any condition, we will do k-fold cross validation to make sure our prediction has a great value of accuracy.



	V1
1	0.8504167
2	0.8521449
3	0.8641100
4	0.8658333
5	0.8433986
6	0.8487500
7	0.8533944
8	0.8457691
9	0.8557732
10	0.8679167

Here we have the result of cross validation. The model itself has mean about 85.47% based on the 10 time cross validation. Beside the result of the model, here is the visualization of the tree based on the model:



Based on the tree above, we see the process of how the model predicts the default.

Chapter 4: Conclusion and Suggestion

After we analyze and make model from this data, we know some characteristics of this data, they are:

- Age does not correlate for predict people that have default next month,
- Male customer probably will get default rather than woman,
- People that had married probably have a default next month rather than others,
- Customers that have a high school education tend to get a default rather than the other education,
- The BILL_AMT features that have default value tend have a more constant movement rather than those who don't have a default,
- The PAY_AMT features don't have any single information that could be interpreted from it. People that have default and don't have default has the same distribution in PAY_AMT features,
- The PAY_ features tend to have a significant effect for predicting people with default. By doing feature engineering on it, we increase the correlation with the target and now it has become a huge contributor for predicting the target based on the heatmap.

From modelling itself, we got some interesting insights, such as:

- For doing predictive modelling, we don't use the PAY_ features because it overfits the data,
- Despite the PAY_AMT doesn't explaining the customer will get default, it contributes to the increasing model's accuracy.

For future work, we will recommend using another machine learning method for predicting this data and also seeing the feature importance. We also have read about the performance of Neural Networks and Support Vector Machine for predictive modelling and it has a huge accuracy on it. Beside that, we know that these models are black-box models that cannot be interpreted why we have this accuracy. So, Model-Agnostic interpretation will be great to be implemented if the model is being used further.