Default of Credit Card Clients

By

Fabjola Kasaj

Farzad Emami

Kanmani Natarajan

Neeta Ganamukhi

Uday Patel

Yinjia Liu

CALIFORNIA STATE UNIVERSITY-EAST BAY

California State University, East Bay

Data Mining Project

BAN620

Prof. Balaraman Rajan

17th May 2020

Table of Contents

1. Introduction……3

a. Problem Context……...3

b. Data Source and Description……3

c. Research Questions……..5

2. Project Analysis and Modeling……...6

a. Data Summary and Initial Analysis……6

b. Data Cleaning……..8

i. Removed ID……...8

ii. Check for extreme values……8

c. Data Visualization……...9

i. Payment Status of customers……9

ii. Payment Amount By defaulters……9

iii. Effect of categorical and numeric variables on defaulter/ non-defaulter……10

d. Modeling (Logit, Neural Network, CART)……..13-20

i. Pre-processing

ii. Analysis and Results

iii. Tuning (Improvements)

3. Conclusion……21

a. Best Model……..21

b. Recommendation……21

1. Introduction

a. Problem Context

Our organization True Mining, Inc. headquartered in California, USA has been hired by Gold Credit Bank in Taiwan. The bank would like to introduce a set of new credit cards to improve their bottom-line. The existing customer base that own their gold credit cards is a valuable data source to formulate their new credit card services. More specifically our task is to classify customers as defaulters or non-defaulters.

b. Data Source and Description

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

For this data mining project, we have used a dataset published by UCI Machine Learning (the link attached above). The intention of choosing this dataset is to improve our knowledge on data mining concepts using real time data, building different models to predict the outcome variable and to interpret our result.

We are interested in determining the non-defaulters based on the previous six months data. We used “bill amount”, “Payment amount” and “repayment status” as our predictors to classify the non-defaulters and defaulters.

In addition, there are some categorical variables such as education, gender and marriage.

Number of Variables	25
Number of Rows	30,000
VARIABLES	
LIMIT_BAL	Amount of the given credit (NT dollar). It includes both the individual consumer credit and his/her family (supplementary) credit.
SEX	1 = male, 2 = female
EDUCATION	1 = graduate school; 2 = university; 3 = high school; 4 = others
MARRIAGE	Marital Status (1 = married; 2 = single; 3 = divorce: 4 = others)
AGE	Year
PAY_Sep, PAY_Aug- PAY_Apr	History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: **PAY_Sep** = the repayment status in September, 2005;

	PAY_Aug = the repayment status in August, 2005; . . .; **PAY_Apr** = 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_AMT_Sep - BILL_AMT_Apr	Amount of bill statement (NT dollar). **BILL_AMT_Sep** = amt. of bill statement in September, 2005; **BILL_AMTAug** = amt of bill statement in August, 2005; . . .; **BILL_AMTApr** = amt of bill statement in April, 2005.
PAY_AMT_Sep- PAY_AMT_Apr	Amount of previous payment (NT dollar). **PAY_AMT_Sep** = amount paid in September, 2005; **PAY_AMT_Aug** = amount paid in August, 2005; . . .; **PAY_AMT_Apr** = amount paid in April, 2005.
default.payment.next.month	Missed Payment = 1, Duly Payment = 0 **[Outcome Variable]**
PREDICTORS	
Categorical	**Numeric**
1. Gender 2. Education 3. Marital Status 4. Past Payment (Timeliness of payment)	1. Credit limit 2. Age 3. Credit Card Bill Amount (Monthly) 4. Payment Amount (Monthly)
Classification: Non-defaulters of the next month	

Missing Data	No missing data
Anomalous Data	EDUCATION has category 5 and 6 that are unlabelled, moreover the category 0 is undocumented. MARRIAGE has a label 0 that is undocumented
Negative Values of "Bill_AMT_#"	Can be interpreted as credit

Data Cleaning	The 0 in MARRIAGE can be categorized as 'Other' (thus 3). The 0 (undocumented), 5 and 6 (label unknown) in EDUCATION can also be put in a 'Other' category (thus 4)

c. Research Questions

1. Which attributes are important predictors?

    1.1 Do demographic predictors effect defaulters?

    1.2 Which numeric attributes help predict defaulters?

2. How far in the future can we predict?

3. Which categorical variables help to predict defaulters?

2. Project Analysis and Modeling

a. Data Summary and Initial Analysis

Summary of the entire dataset:

Our dataset consists of 30,000 rows and 25 columns. We have 4 categorical variables with multiple levels and 14 numerical variables. Below is the summary of the dataset under study.

Demographic Variables

```
       ID          LIMIT_BAL        SEX        EDUCATION  MARRIAGE       AGE
Min.   :    1   Min.   :  10000   1:11888   0:   14   0:   54   Min.   :21.00
1st Qu.: 7501   1st Qu.:  50000   2:18112   1:10585   1:13659   1st Qu.:28.00
Median :15000   Median : 140000             2:14030   2:15964   Median :34.00
Mean   :15000   Mean   : 167484             3: 4917   3:  323   Mean   :35.49
3rd Qu.:22500   3rd Qu.: 240000             4:  123             3rd Qu.:41.00
Max.   :30000   Max.   :1000000             5:  280             Max.   :79.00
                                            6:   51
```

	o use
Credit limit	Ranges from $10 thousand to $1 million Taiwanese Dollars On average credit limit is $167,484
Sex	More female (2) credit card holders
Education	Majority belongs to who attended university (2) followed by graduate school (1) and High school (3).
Marriage	Single (2) holds most credit cards, then Married (1) and divorced (3).
Age	Minimum age to own a credit in Taiwan is 21 years of age.

Bill Amount

```
 BILL_AMT_Sep      BILL_AMT_Aug      BILL_AMT_Jul      BILL_AMT_Jun      BILL_AMT_May      BILL_AMT_Apr
Min.   :-165580   Min.   : -69777   Min.   :-157264   Min.   :-170000   Min.   : -81334   Min.   :-339603
1st Qu.:   3559   1st Qu.:   2985   1st Qu.:   2666   1st Qu.:   2327   1st Qu.:   1763   1st Qu.:   1256
Median :  22382   Median :  21200   Median :  20089   Median :  19052   Median :  18105   Median :  17071
Mean   :  51223   Mean   :  49179   Mean   :  47013   Mean   :  43263   Mean   :  40311   Mean   :  38872
3rd Qu.:  67091   3rd Qu.:  64006   3rd Qu.:  60165   3rd Qu.:  54506   3rd Qu.:  50191   3rd Qu.:  49198
Max.   : 964511   Max.   : 983931   Max.   :1664089   Max.   : 891586   Max.   : 927171   Max.   : 961664
```

The bill amount in negative values indicates that the customer has made advance payment.

- Across six months, on average people are spending more and paying less. This average bill amount increases from $38,872 in April to $51,223.

Payment Amount

PAY_AMT_Sep	PAY_AMT_Aug	PAY_AMT_Jul	PAY_AMT_Jun	PAY_AMT_May	PAY_AMT_Apr
Min. : 0	Min. : 0	Min. : 0	Min. : 0	Min. : 0.0	Min. : 0.0
1st Qu.: 1000	1st Qu.: 833	1st Qu.: 390	1st Qu.: 296	1st Qu.: 252.5	1st Qu.: 117.8
Median : 2100	Median : 2009	Median : 1800	Median : 1500	Median : 1500.0	Median : 1500.0
Mean : 5664	Mean : 5921	Mean : 5226	Mean : 4826	Mean : 4799.4	Mean : 5215.5
3rd Qu.: 5006	3rd Qu.: 5000	3rd Qu.: 4505	3rd Qu.: 4013	3rd Qu.: 4031.5	3rd Qu.: 4000.0
Max. :873552	Max. :1684259	Max. :896040	Max. :621000	Max. :426529.0	Max. :528666.0

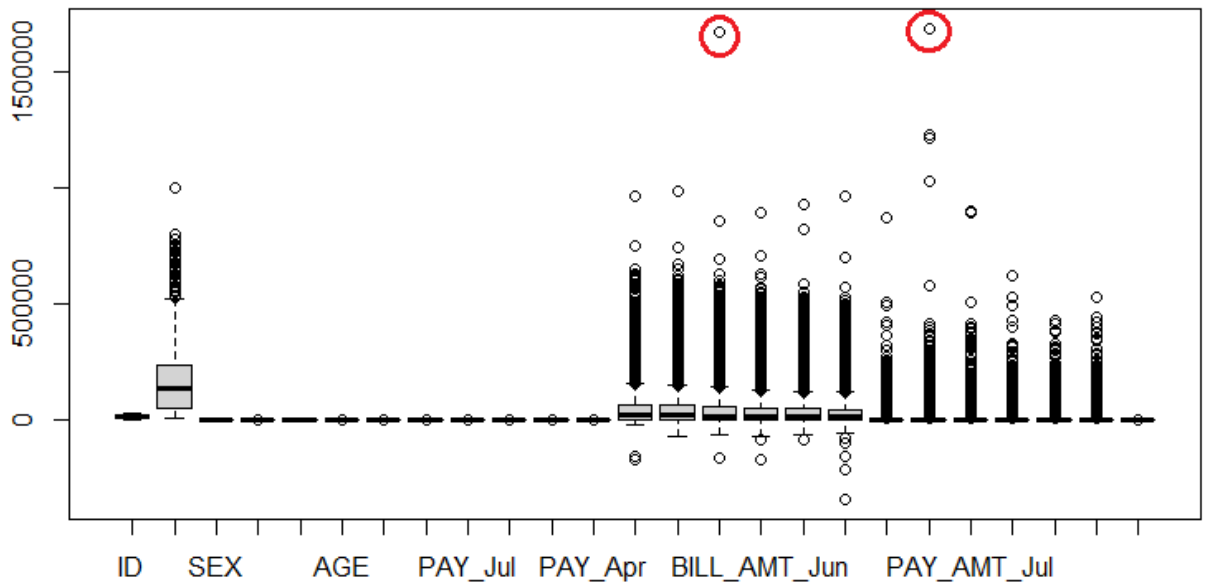
- The average payment staying fairly similar across six months. The highest average is in August pulled up by the max value of 1,684,259.

b. Data Cleaning

i. Removed ID Column

```
> defaulter.df<-defaulter.df[,-1]
> colnames(defaulter.df)
 [1] "LIMIT_BAL"                     "SEX"
 [3] "EDUCATION"                     "MARRIAGE"
 [5] "AGE"                           "PAY_Sep"
 [7] "PAY_Aug"                       "PAY_Jul"
 [9] "PAY_Jun"                       "PAY_May"
[11] "PAY_Apr"                       "BILL_AMT_Sep"
[13] "BILL_AMT_Aug"                  "BILL_AMT_Jul"
[15] "BILL_AMT_Jun"                  "BILL_AMT_May"
[17] "BILL_AMT_Apr"                  "PAY_AMT_Sep"
[19] "PAY_AMT_Aug"                   "PAY_AMT_Jul"
[21] "PAY_AMT_Jun"                   "PAY_AMT_May"
[23] "PAY_AMT_Apr"                   "default.payment.next.month"
>
```

ii. Identified extreme values

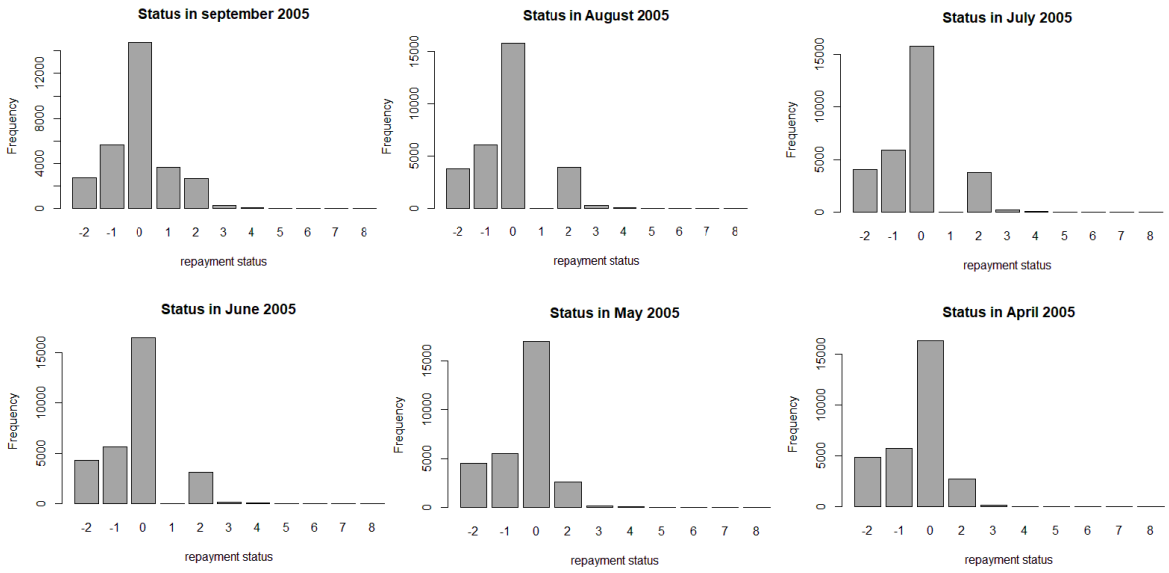


The above boxplot shows two extreme values (marked in red) for the Bill Amount and Payment Amount. We cannot classify them as outliers because the customer's limit balance could be high and hence their bill amount and pay amount might be high. These particular values need special investigation.

c. Data Visualization:

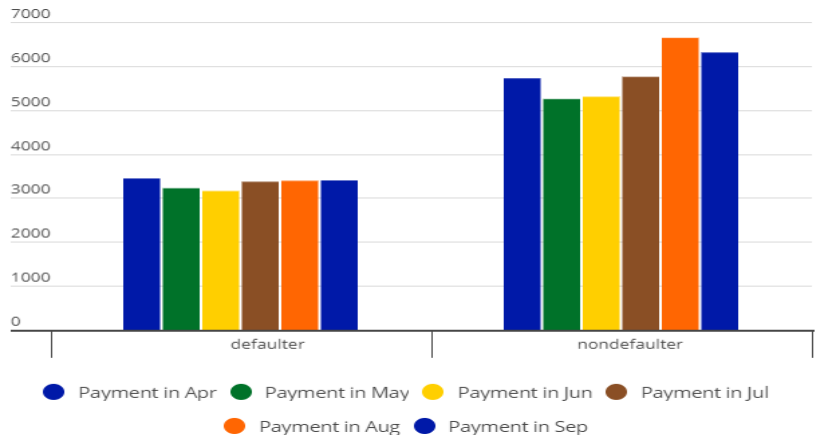
i. Payment Status of customers

According to the dataset description the measurement scale for the repayment status, i,e PAY_X is a set of categorical variables with the levels: -2 = no balance to pay, -1 = pay duly, 0= The use of revolving credit,1 = payment delay for one month, 2 = payment delay for two months, 8 = payment delay for eight months & above



The graphs above shows nearly 50% of the customers had revolving credit from April-05–September-05. Approximately 16.67% customers duly paid credit card bill. Nearly 10 % customers had no balance to pay and almost 16.6% customers had payment delay for two months

ii. Payment Amount By defaulters

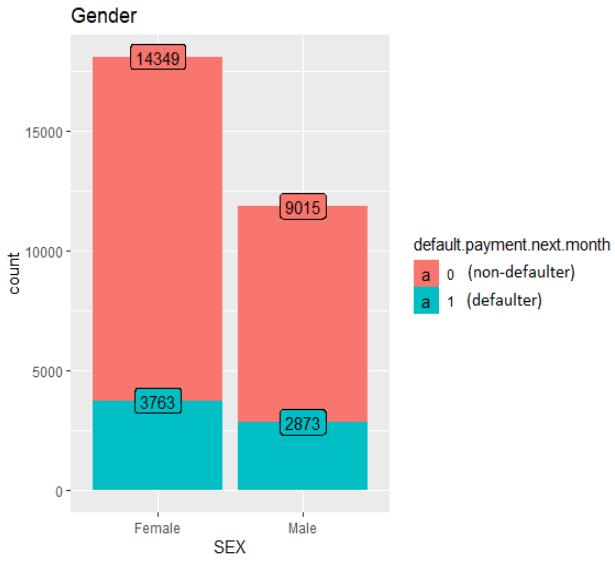


The graph above shows the payment made by defaulters and non-defaulters. It is clear from the graph that payment made by defaulters from April to September is consistent and in lower range whereas it is fluctuating in case of non-defaulters and it's in higher range. Defaulters are paying approx. two thousand less than non-defaulters.

iii. Effect of categorical and numeric variables on defaulter/ non-defaulter

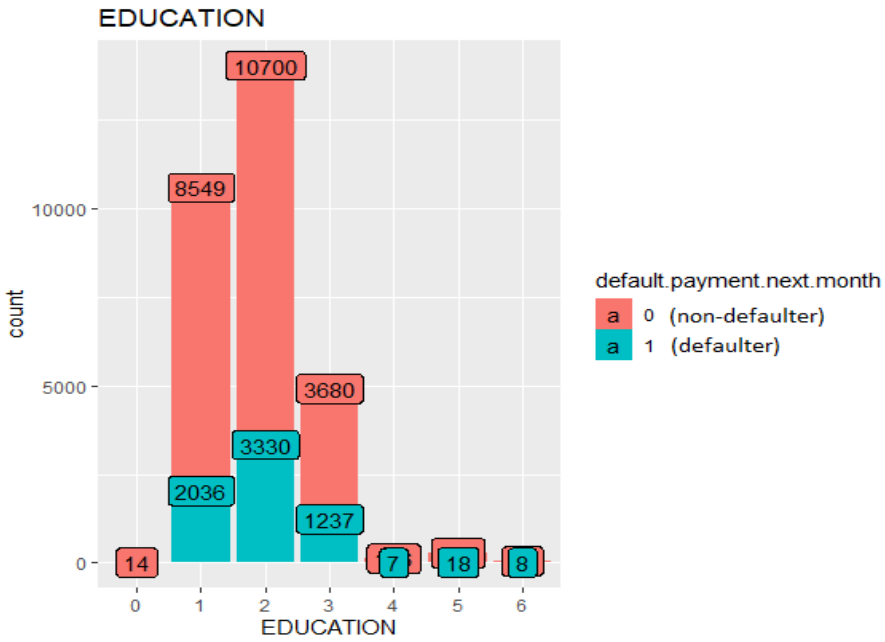
a. Social Status Predictors

Gender is the categorical predictor with 2 levels. 1= Male and 2=Female in our dataset.



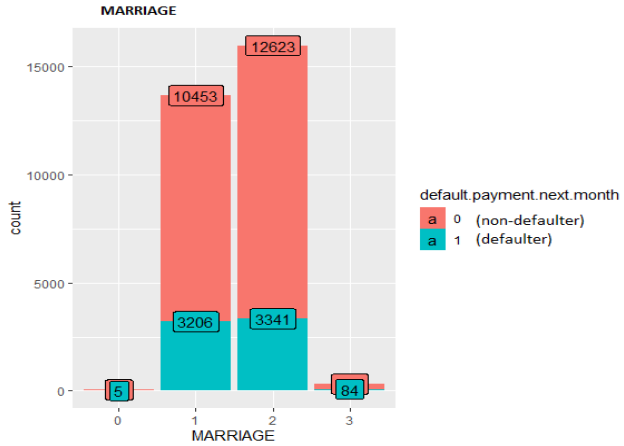
We found that there are 20.77% female default customers and 24.16% male default customers in our study. It is also found that nearly 77.8% of the customers fall under non defaulters' category with female percentage being the highest i,e 61.25.

b. Education is the categorical predictor in our dataset with 6 levels. 1 being at the top for the graduates, 2 for university, 3 for high school students and 4,5,6 are unknown. Level 0 is unclear for us.



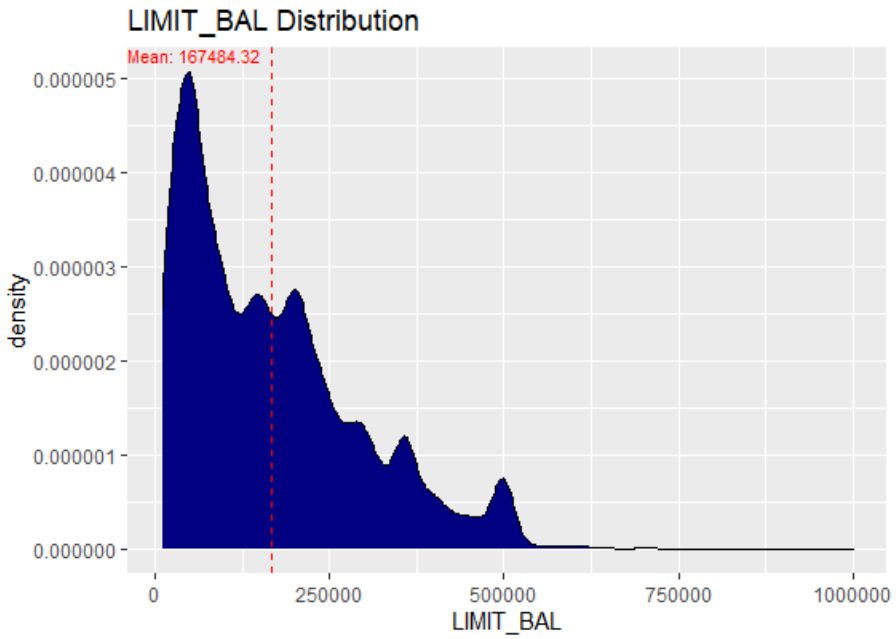
The graph above shows the education level of customers under study. We have approximately 19.2% graduates, 31.1% university and 33.6% high school students who fall under defaulter's category respectively. It is observed that university students are the most who fall under non-defaulter's category with 45.8%

c. Marriage is another categorical predictor in our study with 4 levels.



Above graph shows that we have around 45.5% married, 53.2% single, 1.07% divorced and rest to others. in our project. Among them 23.4% married couple, 20.9% singles and 23.6 % fall under defaulter's category. Singles being the highest who fall under non-defaulter's category with 79%.

d. Limit Balance Distribution:



The Limit Balance density graph above shows the distribution of Limit Balance amount across 30,000 customers. Mean value of Limit Balance amount is 167,484.32.

d. Modeling

Scope:

Even though the dataset is limited to defaulters record for October month, our scope is not just limited to prediction for October (Next month) but up to six months. This was done by limiting the number of earlier months to predict October. For instance, if our model is to predict defaulters for 6th month from current month, we moved our current month to April to predict October. That means our model will include only data up to April to predict 6th month (i.e. October). With this idea six models were developed in total & accuracy is measured.

Logit Model

i. Pre-Processing:

1.Numerical to categorical
Transfer SEX, EDUCATION, MARRIAGE, PAY_Apr, PAY_May, PAY_Jun, PAY_Jul, PAY_Aug, PAY_Sep from numeric data type to categorical data type.

2.Partition: Training(60%) & Validation(40%)
After partition data set, training data has 18000 rows and 24 columns, validation data has 12000 rows and 24 columns.

3.Processing rare levels in validation data
Since level 1 of PAY_Jun has two records (6783 and 11498) in validation data but no in training data, we remove these two records into training data in order to use predict function without error messages.

Similarly, we also find level 8 of Pay_May has one record (8655) in validation data but no in training data, we remove this record in to training data.
Finally, we get 18003 rows in training data and 11997 rows in validation data.

ii. Analysis and results:

Selecting predictors and developing models:
Totally, we develop seven models with different predictors. The most important logic for selecting predictors is by selecting different previous months data. For example, Model 7 only using April data with Pay April status, April Bill Amount, April Pay Amount to predict for the 6th month which is October. Then Model 8 we add data from May along with April to predict for 5th month which is also October. Then Model 5 we add data from Jun,etc. In Model 2, we remove all the payment status. In Model 1, we use all predictors from all six months data.

iii. Comparison for classification performance:

Table 1-1:

Models No	Predictors	Accuracy	Sensitivity (Defaulter Classification)	Specificity (Non-defaulter Classification)	Prediction Month
1	Apr - Sep	0.8219	0.3483	0.9546	1st
2	Apr - Aug	0.8085	0.2619	0.9616	2nd
3	Apr - Jul	0.8012	0.2127	0.9661	3rd
4	Apr – Jun	0.8005	0.1968	0.9698	4th
5	Apr – May	0.7977	0.1910	0.9677	5th
6	Apr alone	0.7928	0.1602	0.9701	6th
7	without Apr to Sep	0.7809	0.0000	0.9998	1st

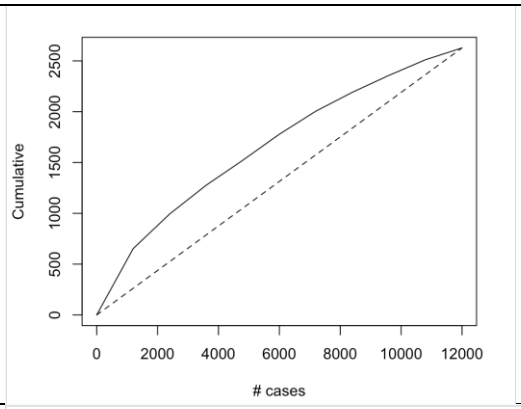
Table 1-1 shows accuracy metrics on validation data for seven models.

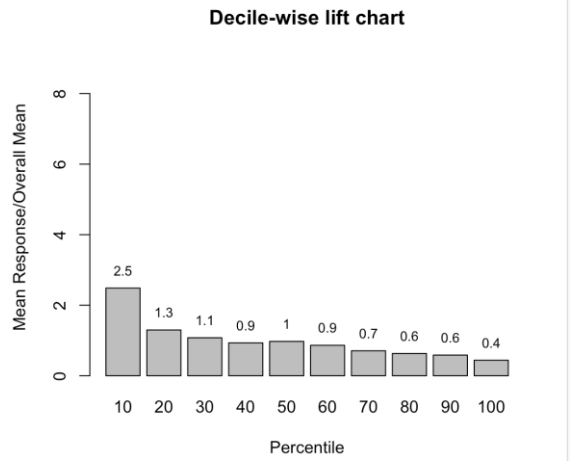
At first glance of this table, we may have such conclusion that Model 1 with all predictors is the best model since it has high prediction accuracy on validation data, also a good specificity and sensitivity is also higher than others which means model 1 can predict defaulters more accurately for bank to avoid lost.

However, from the perspective of credit card company, Model 6 is the most useful one. That is because just using April's available information, the company can predict what is going to happen in October. The company don't need to wait until September to get all six s information. So, even though the accuracy for Model 6 is 3% below than the Model 1, Model 6 is still the most useful one for our case.

Gains table and Decile wise lift chart for Model 6:

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume Pct of Total Resp	Lift Index	Cume Lift	Mean Model Score
10	1199	1199	0.54	0.54	24.9%	249	249	0.52
20	1200	2399	0.28	0.41	37.8%	130	189	0.28
30	1200	3599	0.24	0.35	48.6%	108	162	0.25
40	1199	4798	0.20	0.32	57.9%	93	145	0.23
50	1200	5998	0.21	0.30	67.7%	97	135	0.21
60	1200	7198	0.19	0.28	76.3%	86	127	0.19
70	1199	8397	0.16	0.26	83.4%	71	119	0.18
80	1200	9597	0.14	0.25	89.7%	63	112	0.16
90	1200	10797	0.13	0.23	95.6%	59	106	0.13
100	1200	11997	0.10	0.22	100.0%	44	100	0.08






```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 9090 2206
         1  280  421

               Accuracy : 0.7928
                 95% CI : (0.7854, 0.8)
    No Information Rate : 0.781
    P-Value [Acc > NIR] : 0.0009039

                  Kappa : 0.1771

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.16026
            Specificity : 0.97012
         Pos Pred Value : 0.60057
         Neg Pred Value : 0.80471
             Prevalence : 0.21897
         Detection Rate : 0.03509
   Detection Prevalence : 0.05843
      Balanced Accuracy : 0.56519

       'Positive' Class : 1
```

For the gain table, it shows without our model, there is 22% chance to find a defaulter, after using Model 7, just in first group, our model already can identify 54% defaulter.

For the lift chart, the lift over the base curve indicates for a given number of cases, the additional defaulters that Model 7 can identify.

Neural Net Model:

i. Pre-processing

The data was manipulated to cater the neural net function in R.

1. The demographic categorical variables such as Education, Sex and Marriage were converted to a factor; subsequently, these factors had to convert to dummies.
2. The numeric variables were scaled between 0 and 1.
3. Same data partition & seed was used for Neural Net as Logit and CART.

ii.Analysis and Result

1. Which attributes to use?

	Model No	Predictors	Accuracy	Sensitivity (Defaulter Classification)	Specificity (Non-defaulter Classification)	Prediction Month
Layer = 1 Node = 3	1	Apr-Sep	0.7773	0.4374	0.8726	1st
	2	Apr - Aug	0.7831	0.2958	0.9197	2nd
	3	Apr - Jul	0.7624	0.2809	0.8974	3rd
	4	Apr – Jun	0.7759	0.2809	0.9146	4th
	5	Apr – May	0.7698	0.2250	0.9224	5th
	6	Apr alone	0.7628	0.2048	0.9191	6th
	7	without pay	0.755	0.1724	0.9183	1st

The above table shows the use of different combinations of attributes used to classify defaulters. The original data consists of 77.9% of non-defaulters and 22.1% of defaulters. Model 1 is trained on demographic predictors + bill amount and payment amount for six s. The overall accuracy is lower than the rest. The sensitivity (defaulters classified correctly) is 17.24% and specificity (non-defaulters classified correctly) is 91.83%. Model 1 is better at classifying non-defaulter than defaulters and the same goes for all models. Because majority belong to non-defaulter category (77.9%).

Model 2 slightly improves accuracy compared to Model 1. The accuracy for Model 3 with the defaulter & non-defaulter classification increases to 22.50% & 92.24% respectively compared to Model 2.

Model 5 the accuracy is less compared to Model 4 despite having data of one more. The accuracy defaulter classification is identical for Model 4 & 5. Model 6 has highest overall accuracy will defaulter classification nearly 30%. Model 7 the overall accuracy & non-defaulter reduces somewhat, but the defaulter classification is the highest (43.74%).

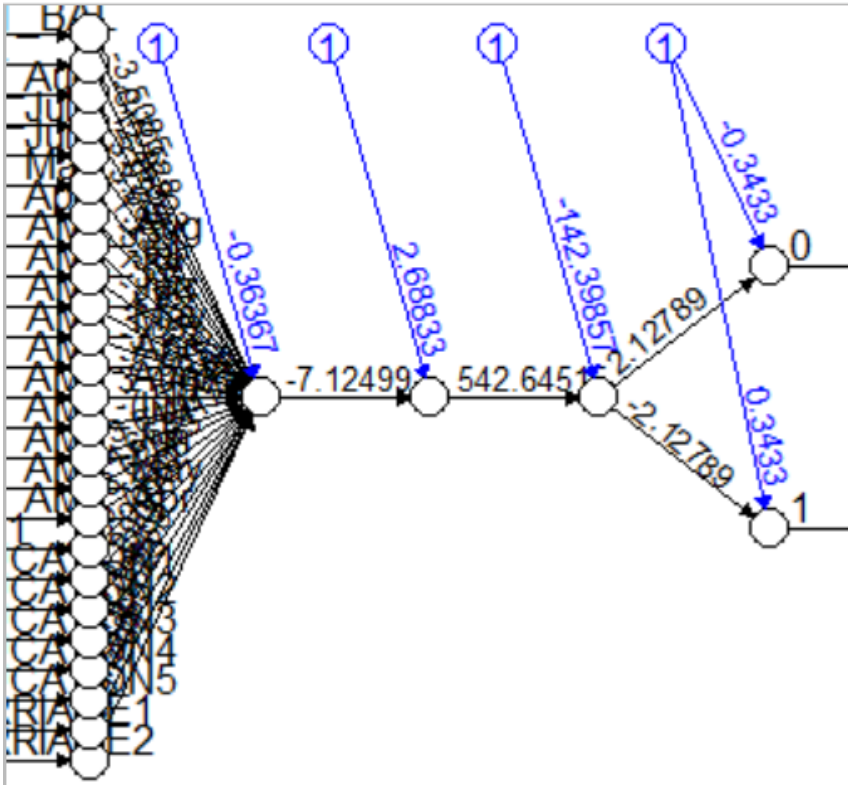
Model 2 predicts for 6 months with low defaulter classification. Model 6 is able to predict 29.58 % of defaulters two months into the future. Model 7 does improve the defaulter classification; however, it also means sacrificing 1 month of future prediction. Model 6 with highest overall accuracy will be used to experiment with different layers.

iii Tuning - How many layers to use?

Layer / Nodes	Model No	Predictors	Accuracy	Sensitivity (Defaulter Classification)	Specificity (Non-defaulter Classification)	Prediction
(1,1,1)	10	Apr to Aug	0.7825	0.3936	8915	2nd
(1,1)	9		0.7889	0.3779	0.9057	
1	8		0.779	0.3623	0.8957	

After experimenting with multiple layer and node choices. The above table shows the layer and node combination that improved the previous result. Model 10 will three layers one node each improved the defaulter classification the most and will be used to compare with best models from other algorithms.

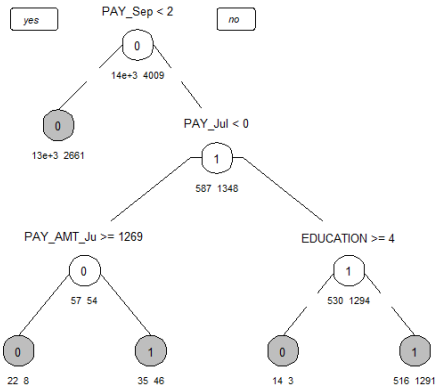
2. Model 10 Visual



CART Model:

i. CART Models Analysis:

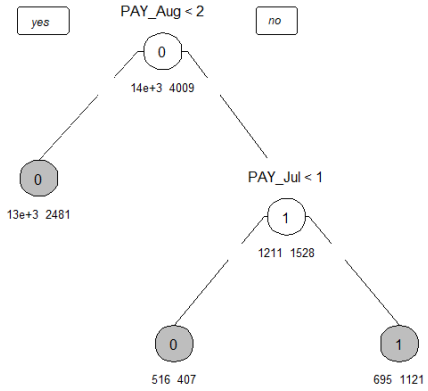
Model 1
(Prediction for 1st month)



71.4% of customers who have at least two-month payment delay in September and not made full payment in July with education level 3 or below are **defaulters** in October.

84% of customers with less than two-month payment delay in September are **non-defaulters** in October.

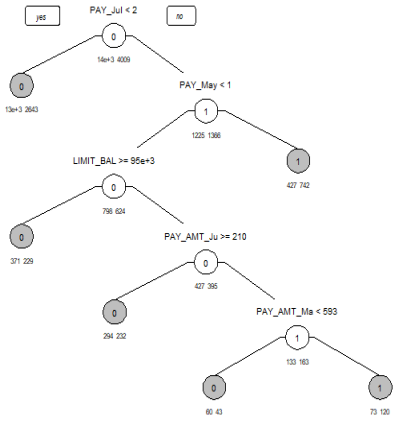
Model 2
(Prediction for 2nd month)



62% of customers who missed payment in July & August are **defaulters** in October.

79.6% of customers with less than two-month payment delay are **non-defaulters** in October.

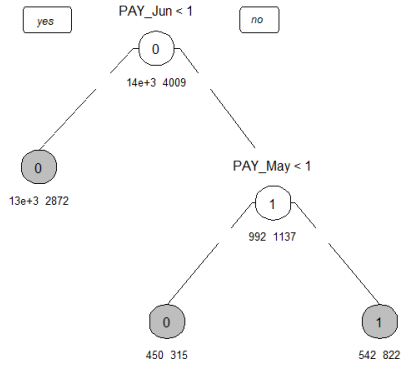
Model 3
(Prediction for 3rd month)



64% of customers who missed payment in May & July are **defaulters** in October.

83% of customers who have less than two-month payment delay in July are **non-defaulters** in October.

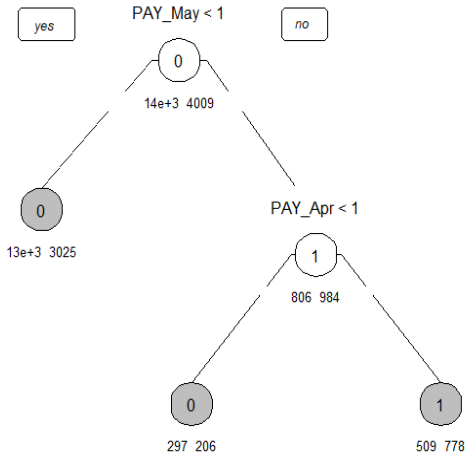
Model 4
(Prediction for 4th month)



60% of customers who missed payment in May & June are **defaulters** in October.

82% of customers who have made at least partial payment in June are **non-defaulters** in October.

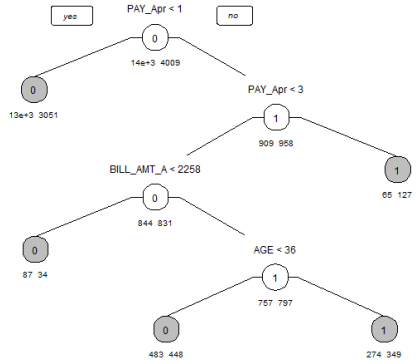
Model 5
(Prediction for 5th month)



60.5% of customers who missed payment in April & May are **defaulters** in October.

81.3% of customers who made at least partial payment in May are **non-defaulters** in October.

Model 6
(Prediction for 6th month)



66% of customers who have at least three-month payment delay in April are **defaulters** in October.

56% of customers who have two-month payment delay, more than 2258 bill amount in April with age above 35 are **defaulters** in October

81% of customers who made at least partial payment in April are **non-defaulters** in October month

ii.Analysis and Results

Comparison of accuracy in different models using CART:

	included in model for prediction	Accuracy			Specificity (Non-defaulter classification)			Sensitivity (Defaulter classification)			Prediction Month
		CV & Pruned	BT	RF	CV & Pruned	BT	RF	CV & Pruned	BT	RF	
1	Apr to Sep	0.821	**0.823**	0.820	0.963	0.955	0.951	0.313	**0.351**	**0.351**	1st
2	Apr to Aug	0.806	**0.808**	0.806	0.957	0.959	0.954	0.269	0.268	**0.277**	2nd
3	Apr to Jul	**0.799**	**0.799**	0.796	0.966	0.960	0.957	0.203	**0.228**	0.221	3rd
4	Apr to Jun	**0.800**	**0.800**	0.794	0.965	0.963	0.959	0.213	**0.218**	0.206	4th
5	Apr to May	**0.798**	**0.798**	0.793	0.967	0.967	0.960	0.194	0.195	**0.198**	5th
6	Apr	0.789	*	0.784	0.977	*	0.952	0.119	*	**0.184**	6th

* R takes longer time to process

Overall accuracy is almost similar when compared to boosted tree & random forest even though random forest & boosted tree have slightly better accuracy in terms of classifying defaulters.

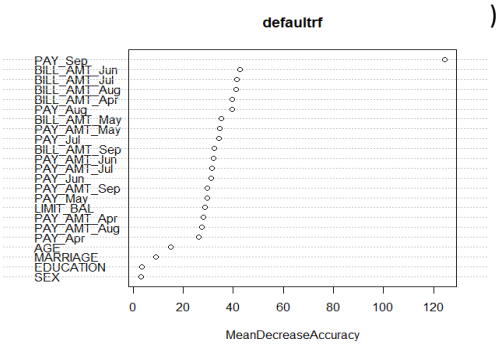
Observation:

Model 1 has better accuracy compared to other models. But Model 6 is the most useful one as it is predicting defaulters for October with just April data.

Variable Importance Score:

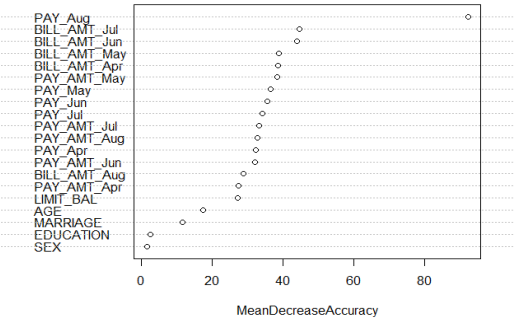
Model 1

)



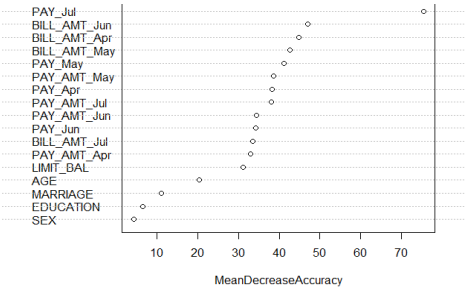
Model 2

(Prediction for 2nd month)



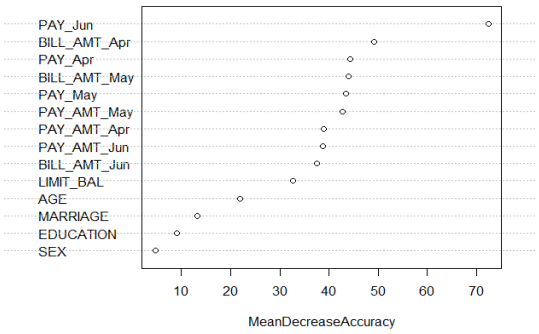
Model 3

(Prediction for 3rd month)

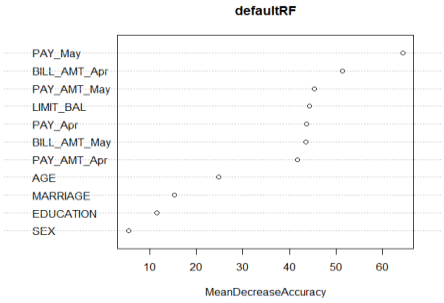


Model 4

(Prediction for 4th month)

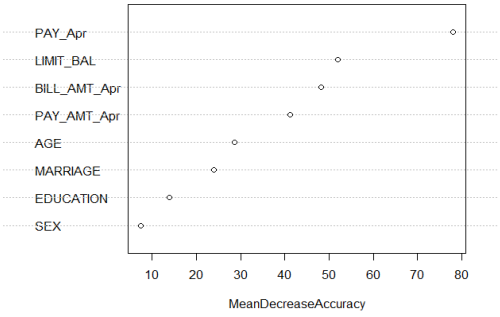


Model 5



Model 6

(Prediction for 6th month)



From variable importance score we observe that the payment status of latest available is crucial in defaulter classification.

Observation: Even though the model 1 which include all the six months data has the overall better accuracy, model 6 is the most useful model for the bank as the model predicts the defaulters in October(6th month) using just April data.

3. Conclusion:

Our dataset has only about 22% defaulters. Our estimate of defaulters will be more reliable if we can include a greater number of defaulter records into our dataset. And in bank perspective, it might be more important to classify defaulters properly than non-defaulters. In that scenario, it can be useful to plot sensitivity and specificity against the cutoff value to find a cutoff value that balances these measures.

a.Best Model

CART & Logit have better overall accuracy compared to NN. Though Logit models are slightly better in terms of overall accuracy & sensitivity, we recommend CART as it gives some useful rules for classifying defaulters & non defaulters.

b. Recommendation

If the bank is interested in identifying defaulters for the following month, we recommend Model 1 as it provides highest accuracy. But there is always a chance for a non-defaulter in one month to become a defaulter in another month. Hence based on our available dataset, it is always useful to predict the defaulters for not just the following month but for up to 6 months.