Logistic Regression and
Default of Credit Cards

A recent study conducted by the financial services and researcher company, The Ascent, was directed to answer the question, "What's the total credit card debt in the U.S.?" The outcome of this research exposed that the total credit card balances in the United States were around $890 billion and that approximately 9.1% of all credit card balances across the nation were 90 days or more delinquent as of the first quarter of 2020 (Frankel & Rosen). From a macro-perspective, the mentioned case illustrates how critical predictability or probability of default could be for financial institutions and their risk management departments when approving or denying credit applications. The following assessment and report aims to explore a customer credit card payments dataset and shows how the data mining technique of logistic regression could benefit the financial sector foreseeing credit card default payments based on customers payment trends and/or activities interrelated to their respective credit balances.

The multivariate dataset comes from the UCI's Center for Machine Learning and Intelligent Systems repository and it consists of 30000 observations and 24 attributes involving customers credit card's payments information. The same was data collected as the result of a research in Taiwan, aiming to forecast default payments based on consumers' history of payments, bill amounts, and/or payment delays among other qualities. The following list contains a brief description of each of the variables in the set:

ID: auto-generated identified **LIMIT_BAL**: amount of given credit **AGE**: years

GENDER: 1 = male; 2 = female **MARRIAGE**: (1 = married; 2 = single; 3 = others)

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

PAY_ 1 to PAY_6: the measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months… and so forth.

BILL_AMT 1 to BILL_AMT 6: amount of bill statement

PAY_AMT 1 to BILL_AMT 6: amount of previous payment

default.payment.next.month: (response variable) default payment yes = 1, no = 0

Exploratory Data Analysis (EDA): During the EDA process, commands like dim( ), head( ), str( ), summary ( ) were employed to better appreciate the data at hand. As displayed in figures 1.0A, and 1.0B, the data includes a variety of attributes, all of these expressed as integers, and it confirms the number of observations and variables.

```
> head(df)     # preview
# A tibble: 6 x 25
     ID LIMIT_BAL   SEX EDUCATION MARRIAGE   AGE PAY_0 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6 BILL_AMT1 BILL_AMT2 BILL_AMT3
  <int>     <int> <int>     <int>    <int> <int> <int> <int> <int> <int> <int> <int>     <int>     <int>     <int>
1     1     20000     2         2        1    24     2     2    -1    -1    -2    -2      3913      3102       689
2     2    120000     2         2        2    26    -1     2     0     0     0     2      2682      1725      2682
3     3     90000     2         2        2    34     0     0     0     0     0     0     29239     14027     13559
4     4     50000     2         2        1    37     0     0     0     0     0     0     46990     48233     49291
5     5     50000     1         2        1    57    -1     0    -1     0     0     0      8617      5670     35835
6     6     50000     1         1        2    37     0     0     0     0     0     0     64400     57069     57608
# … with 10 more variables: BILL_AMT4 <int>, BILL_AMT5 <int>, BILL_AMT6 <int>, PAY_AMT1 <int>, PAY_AMT2 <int>,
#   PAY_AMT3 <int>, PAY_AMT4 <int>, PAY_AMT5 <int>, PAY_AMT6 <int>, default.payment.next.month <int>
> |
```

Figure 1.0A – Dataset initial 6 observations.

```
> str(df)     # structure
tibble [30,000 × 25] (S3: tbl_df/tbl/data.frame)
 $ ID                        : int [1:30000] 1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL                 : int [1:30000] 20000 120000 90000 50000 50000 50000
 $ SEX                       : int [1:30000] 2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : int [1:30000] 2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : int [1:30000] 1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : int [1:30000] 24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0                     : int [1:30000] 2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2                     : int [1:30000] 2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3                     : int [1:30000] -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4                     : int [1:30000] -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5                     : int [1:30000] -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6                     : int [1:30000] -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1                 : int [1:30000] 3913 2682 29239 46990 8617 64400 3679
 $ BILL_AMT2                 : int [1:30000] 3102 1725 14027 48233 5670 57069 4120
 $ BILL_AMT3                 : int [1:30000] 689 2682 13559 49291 35835 57608 4450
 $ BILL_AMT4                 : int [1:30000] 0 3272 14331 28314 20940 19394 542653
 $ BILL_AMT5                 : int [1:30000] 0 3455 14948 28959 19146 19619 483003
 $ BILL_AMT6                 : int [1:30000] 0 3261 15549 29547 19131 20024 473944
 $ PAY_AMT1                  : int [1:30000] 0 0 1518 2000 2000 2500 55000 380 332
 $ PAY_AMT2                  : int [1:30000] 689 1000 1500 2019 36681 1815 40000 6
 $ PAY_AMT3                  : int [1:30000] 0 1000 1000 1200 10000 657 38000 0 43
 $ PAY_AMT4                  : int [1:30000] 0 1000 1000 1100 9000 1000 20239 581
 $ PAY_AMT5                  : int [1:30000] 0 0 1000 1069 689 1000 13750 1687 100
 $ PAY_AMT6                  : int [1:30000] 0 2000 5000 1000 679 800 13770 1542 1
 $ default.payment.next.month: int [1:30000] 1 1 0 0 0 0 0 0 0 0 ...
```

```
> summary(df) # Descriptive statistics
       ID          LIMIT_BAL            SEX          EDUCATION        MARRIAGE          AGE            PAY_0        
 Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :21.00   Min.   :-2.0000  
 1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:28.00   1st Qu.:-1.0000  
 Median :15000   Median : 140000   Median :2.000   Median :2.000   Median :2.000   Median :34.00   Median : 0.0000  
 Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853   Mean   :1.552   Mean   :35.49   Mean   :-0.0167  
 3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:41.00   3rd Qu.: 0.0000  
 Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000   Max.   :3.000   Max.   :79.00   Max.   : 8.0000  
     PAY_2             PAY_3             PAY_4             PAY_5             PAY_6           BILL_AMT1      
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-165580  
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:   3559  
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :  22382  
 Mean   :-0.1338   Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662   Mean   :-0.2911   Mean   :  51223  
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:  67091  
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 964511  
   BILL_AMT2        BILL_AMT3         BILL_AMT4         BILL_AMT5        BILL_AMT6          PAY_AMT1     
 Min.   :-69777   Min.   :-157264   Min.   :-170000   Min.   :-81334   Min.   :-339603   Min.   :     0  
 1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000  
 Median : 21200   Median :  20088   Median :  19052   Median : 18104   Median :  17071   Median :  2100  
 Mean   : 49179   Mean   :  47013   Mean   :  43263   Mean   : 40311   Mean   :  38872   Mean   :  5664  
 3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506   3rd Qu.: 50190   3rd Qu.:  49198   3rd Qu.:  5006  
 Max.   :983931   Max.   :1664089   Max.   : 891586   Max.   :927171   Max.   : 961664   Max.   :873552  
    PAY_AMT2          PAY_AMT3          PAY_AMT4          PAY_AMT5           PAY_AMT6         default.payment.next.month
 Min.   :      0   Min.   :     0   Min.   :     0   Min.   :     0.0   Min.   :     0.0   Min.   :0.0000  
 1st Qu.:    833   1st Qu.:   390   1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8   1st Qu.:0.0000  
 Median :   2009   Median :  1800   Median :  1500   Median :  1500.0   Median :  1500.0   Median :0.0000  
 Mean   :   5921   Mean   :  5226   Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5   Mean   :0.2212  
 3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0   3rd Qu.:0.0000  
 Max.   :1684259   Max.   :896040   Max.   :621000   Max.   :426529.0   Max.   :528666.0   Max.   :1.0000  
```

Figure 1.0B – Structure and descriptive statistics of the dataset.

The set contains a unique identifier named, ID, which could be excluded. The LIMIT_BAL covers a significant scope ranging from 10000 to 1000000. In regards the SEX variable could be converted to a factor as *dummy variables* “male”, “female” and variables PAY_1 to PAY_6 could potentially be converted to factors as payment “on-time”, or payment “delayed”. Furthermore, the BILL_AMT and PAY_AMT attributes have significant figures which could indicates the need to trim outliers and/or scale the distribution—refer to boxplot illustration on figure 1.1 for examples of original distribution.

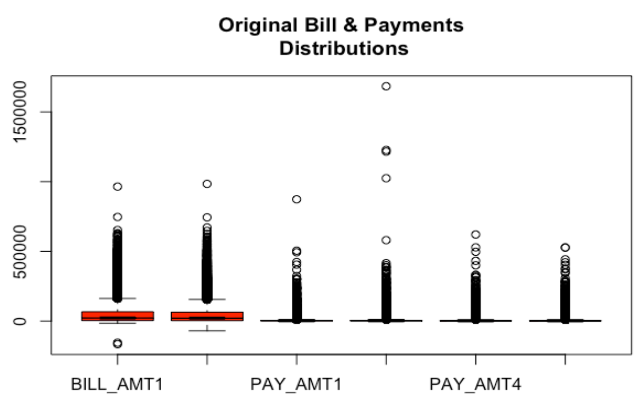


Figure 1.1 – Examples of dispersity across some variables distributions.

Preprocessing: As part of pre-processing, some variables were removed, renamed or transformed. Initially, the ID variable was the only one removed; however, after running multiple iterations, additional variables were removed given their irrelevancy or low influence against the targeted variable.

Attributes like LIMIT_BAL, EDUCATION, PAY_AMT, BILL_AMT and EDUCATION were excluded using the subset( ) command while others were renamed and vectored as factors. As shown in the below code, the response variable had been named "default.payment.next.month" and for simplicity purposes for this assessment was renamed to "PROJECTED_DEFAULT".

```
> # Irrelevant variables to remove given their low influence
> df <- subset(df, select = -c(ID, LIMIT_BAL, EDUCATION, PAY_AMT3, PAY_AMT5, BILL_AMT3, BILL_AMT4,
+                              BILL_AMT5, BILL_AMT6))
> # Convert Marriage status to either married, or not married
> df$MARRIAGE<-factor(df$MARRIAGE == 1, levels = c(FALSE,TRUE), labels = c("NotMarried","Married"))
> #------
> # Convert all PAY variables to OnTime or Delayed
> df$PAY_0<-factor(df$PAY_0 >= 1, levels = c(FALSE,TRUE), labels = c("OnTime","Delayed"))
> df$PAY_2<-factor(df$PAY_2 >= 1, levels = c(FALSE,TRUE), labels = c("OnTime","Delayed"))
> df$PAY_3<-factor(df$PAY_3 >= 1, levels = c(FALSE,TRUE), labels = c("OnTime","Delayed"))
> df$PAY_4<-factor(df$PAY_4 >= 1, levels = c(FALSE,TRUE), labels = c("OnTime","Delayed"))
> df$PAY_5<-factor(df$PAY_5 >= 1, levels = c(FALSE,TRUE), labels = c("OnTime","Delayed"))
> df$PAY_6<-factor(df$PAY_6 >= 1, levels = c(FALSE,TRUE), labels = c("OnTime","Delayed"))
> #------
> colnames(df)[colnames(df)=="default.payment.next.month"]<-"PROJECTED_DEFAULT"  # Rename targeted variable
> df$PROJECTED_DEFAULT<-factor(df$PROJECTED_DEFAULT, levels=0:1, labels=c("No","Yes"))  # Factor targeted variable
> df$SEX<-factor(df$SEX, levels=1:2, labels = c("Male","Female"))
> |
> # Preview Bill and Payment Amount distributions, without outliers, and scaled
> boxplot(df[10:15], col=3:6, outline=FALSE, notch=TRUE, main="Scaled Bill & Payments\n Distribution without Outliers")
> df[10:15] <- rm.outlier(df[10:15], fill = TRUE)
> df[10:15]<-scale(df[10:15], center=FALSE, scale=TRUE)
```

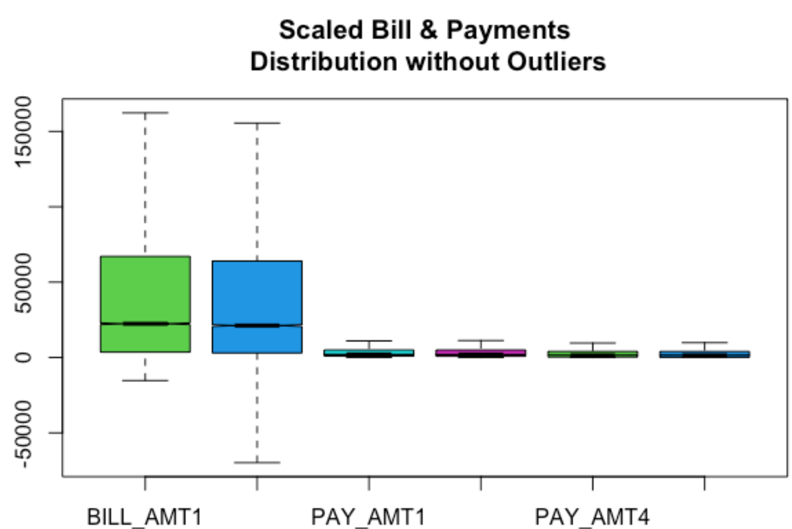


Figure 1.2 – Illustration of variable distributions after scaled and excluding outliers

Moreover, like alluded before, the PAY_0 to PAY_6 variables, which relate to timeliness of credit card payments, were consolidated to indicate if payments were either submitted "OnTime", or "Delayed. Finally, the continuous variables of BILL_AMT and PAY_AMT were scaled down and centered. Once completed these mentioned preprocessing steps the dataframe ended up with a better structured and organized shape, ready for further analysis and modeling. Figure 1.3 shows a summary of the final frame to model using the logistic regression algorithm.

```
> summary(df) # Stats check
     SEX              MARRIAGE           AGE             PAY_0           PAY_2           PAY_3
 Male  :11888    NotMarried:16341   Min.   :21.00    OnTime :23182   OnTime :25562   OnTime :25787
 Female:18112    Married   :13659   1st Qu.:28.00    Delayed: 6818   Delayed: 4438   Delayed: 4213
                                    Median :34.00
                                    Mean   :35.49
                                    3rd Qu.:41.00
                                    Max.   :79.00
     PAY_4           PAY_5           PAY_6           BILL_AMT1            BILL_AMT2
 OnTime :26490   OnTime :27032   OnTime :26921   Min.   :-1.84948   Min.   :-0.80830
 Delayed: 3510   Delayed: 2968   Delayed: 3079   1st Qu.: 0.03975   1st Qu.: 0.03458
                                                 Median : 0.24999   Median : 0.24558
                                                 Mean   : 0.57181   Mean   : 0.56933
                                                 3rd Qu.: 0.74936   3rd Qu.: 0.74140
                                                 Max.   : 8.34170   Max.   : 8.61815
    PAY_AMT1            PAY_AMT2            PAY_AMT4            PAY_AMT6          PROJECTED_DEFAULT
 Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.00000   Min.   : 0.000000   No :23364
 1st Qu.: 0.05966   1st Qu.: 0.03837   1st Qu.: 0.01850   1st Qu.: 0.006444   Yes: 6636
 Median : 0.12528   Median : 0.09253   Median : 0.09377   Median : 0.082086
 Mean   : 0.33615   Mean   : 0.27014   Mean   : 0.30042   Mean   : 0.284457
 3rd Qu.: 0.29864   3rd Qu.: 0.23029   3rd Qu.: 0.25089   3rd Qu.: 0.218895
 Max.   :30.12667   Max.   :56.51795   Max.   :33.06472   Max.   :28.847221
```

Figure 1.3 – Post-EDA and preprocessing dataset of 15 independent variables.

Algorithm Intuition: In simple terms, the logistic regression model forecasts the likelihood of an event to either be one way or the other. In analysis of data, the capacity of foreseen the possibilities of an outcome can help consumers of such data extrapolating assumptions or potential consequences. In supervised learning, logistic regression is one of the fundamental techniques when "finding the best-fitting set of parameters and modeling the relationship between a set of variables". The coefficients generated through this method are key interpreting degrees of probability or conversely levels of uncertainty. In this case, the dataset is comprised of independent categorical and continuous variables (as seen it above) and the response variable dichotomous, the model is handled as a binary logistic regression (Hosmer, Lemeshow, & Sturdivant, 2013).

That said, this logistic regression approach employs the named independent variables of the dataset, which were divided to create two subsets (training and test), to identify how the generated coefficients relate with the dependable variable and what kind of influence these apply to the response variable. As mentioned, the dataframe was divided into two subsets, training and test data. Showing in figure 2.0, 70% of the total dataset was set apart to train the model, while the rest 30%, for testing purposes.

```
# Set the seed, create data subsets
# Set the seed value to ensure that result is reproducible
set.seed(1234)
# Divide the data into train/test subsets
ind<-sample(2, nrow(df), replace=TRUE, prob=c(0.7, 0.3))
train.data<-df [ind == 1,]
test.data<-df [ind == 2,]
```

Figure 2.0 – Seed value set and dataset divided into training and test data.

Followed the separating of the data, the model was created with the glm( ) command, using the train.data as the source and specifying the family as binomial( ). The below code output shows the immediate result including the coefficients and the model's intercept.

```
> # Build and interpret the model. Store the method output in a variable model
> model <- glm(PROJECTED_DEFAULT ~., data = train.data, family = binomial())
> model              # Output the coefficients and intercept

Call:  glm(formula = PROJECTED_DEFAULT ~ ., family = binomial(), data = train.data)

Coefficients:
    (Intercept)      SEXFemale  MARRIAGEMarried            AGE    PAY_0Delayed     PAY_2Delayed
      -1.846898      -0.151393         0.149031       0.002088        1.321684         0.168766
   PAY_3Delayed   PAY_4Delayed     PAY_5Delayed   PAY_6Delayed       BILL_AMT1        BILL_AMT2
       0.411422       0.333658         0.256419       0.451449       -0.253182         0.319032
       PAY_AMT1       PAY_AMT2         PAY_AMT4       PAY_AMT6
      -0.258284      -0.244426        -0.116164      -0.072972

Degrees of Freedom: 20971 Total (i.e. Null);  20956 Residual
Null Deviance:      22130
Residual Deviance: 18740        AIC: 18780
```

Figure 2.1 – Initial model with a residual deviance score of 18740.

Next, the summary( ) command was used to assess the statistics of the model. Illustrated on figure 2.2, the residuals deviance ranged from -1.7754 to 3.4498, with a median of -0.5160. At first sight, the included estimates show positive and negative numbers, underlining the kind of influence these will exert on the targeted variable. Also, most of the coefficients values are under the 0.05 probability rate (see asterisks interpretation), signaling that these values are outside of the standard distribution ranges and should be considered statistically significant. The model produced an Akaike Information Criterion (AIC) rate of 18775 after only 5 iterations.

Finally, based on the illustration of Appendix A., interpretation of the initial model is as follow: Starting from an intercept estimated value of -1.846898, the variable of SEX shows that a female would be -0.151393 or 15% less-probable to default in their credit card payment. Similarly, few attributes below, the PAY_0 variable indicates that when a costumer fails to submit the payment on time, it has a probability greater than 132% of credit cards default.

The model's odds ratio is captured in below figure 2.2. By definition, the odds ratio is interpreted as a measure of association as it approximates how much more likely or unlikely (in terms of odds) it is for the outcome to be present among those subjects (Hosmer, Lemeshow, & Sturdivant, 2013). In simple words, the odds ratio variable measure the changes of odds of one variable to change when another increases one unit. Case in point, the odds of a customer to default on credit cards payments increases over 370% when history shows that the first payment was delayed.

```
> exp(coef(model))    # Odds ratios
  (Intercept)     SEXFemale MARRIAGEMarried          AGE  PAY_0Delayed  PAY_2Delayed  PAY_3Delayed
    0.1577257     0.8595102       1.1607093    1.0020905     3.7497304     1.1838435     1.5089620
 PAY_4Delayed  PAY_5Delayed    PAY_6Delayed    BILL_AMT1     BILL_AMT2      PAY_AMT1      PAY_AMT2
    1.3960656     1.2922935       1.5705869    0.7763267     1.3757956     0.7723757     0.7831538
     PAY_AMT4      PAY_AMT6
    0.8903291     0.9296270
```

Figure 2.2 – Model's odds ratios.

While assessing the accuracy of the model, the *caret* and *e1071* libraries were utilized. Per the graphic bellow, the confusionMatrix( ) command illustrates that within 95% confidence intervals the accuracy of the model is rated at ~ 65%.

```
> confusionMatrix(df$PROJECTED_DEFAULT[1:length(train.data$PROJECTED_DEFAULT)],
+                 train.data$PROJECTED_DEFAULT, dnn=c("Predicted","Actual"))
Confusion Matrix and Statistics

         Actual
Predicted    No   Yes
      No  12583  3603
      Yes  3765  1021

               Accuracy : 0.6487
                 95% CI : (0.6422, 0.6551)
    No Information Rate : 0.7795
    P-Value [Acc > NIR] : 1.0000

                  Kappa : -0.0094

 Mcnemar's Test P-Value : 0.0607

            Sensitivity : 0.7697
            Specificity : 0.2208
         Pos Pred Value : 0.7774
         Neg Pred Value : 0.2133
             Prevalence : 0.7795
         Detection Rate : 0.6000
   Detection Prevalence : 0.7718
      Balanced Accuracy : 0.4953

       'Positive' Class : No
```

Another area evaluated within the model was the precision, recall, and specificity. These metrics are normally used to overall quality of a model. Precision highlights the accuracy of a predicted positive outcome, recall measures the strength of the model, and specificity the model's ability to predict a negative outcome (Bruce & Bruce, 2017). Figure 2.4 shows the receiver operating characteristics (ROC) curve. The closer the generated line gets to the top 1.0 parameter the more accurate the model will be considered.

```
> #==================================================================
> # Build the ROC curve
> par(mfrow=c(1,1))
> ROCRpred<-prediction(mypredictions, test.data$PROJECTED_DEFAULT)
> ROCRperf<-performance(ROCRpred,'tpr','fpr')
> plot(ROCRperf, colorize=TRUE, text.adj=c(-0.2,1.2), lwd=2)
> abline(a=0,b=1,lwd=1,lty=2,col="grey")    #Create a 45 degrees line
```

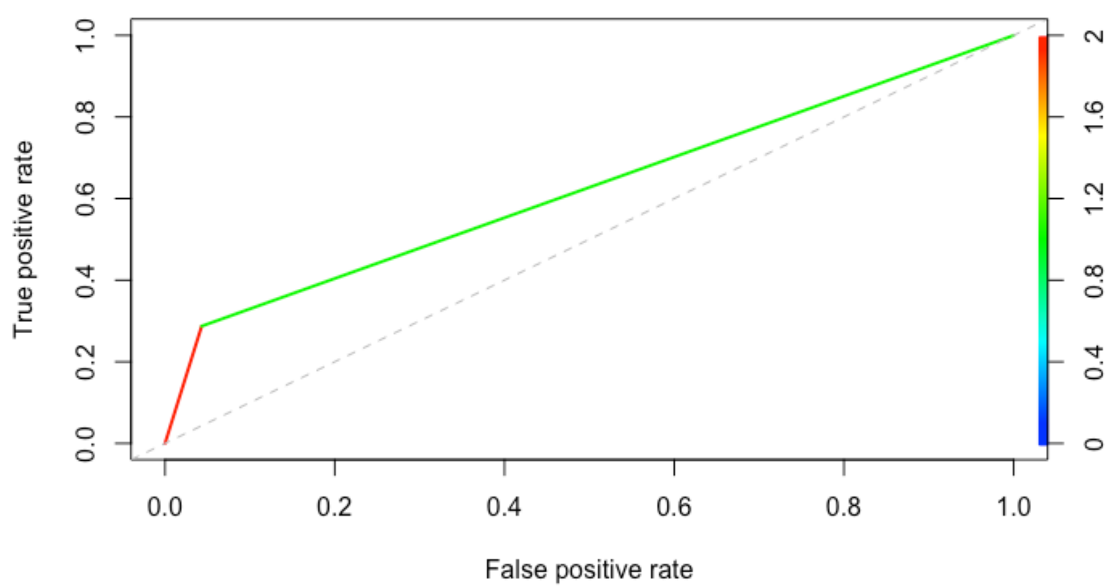


Figure 2.4 – ROC Curve for original model.

Finally, the model was assessed in terms diagnostic analysis of residuals, the delta between what the model predicted versus what was actually observed. Illustrated on figure 2.5, the graphic shows the level of usefulness the model based on the partial residuals out of the model (ranges between As previously mentioned, the angle of the residuals is influence by the dispersed variation of the data values. Of notice, in logistic regression the prime objective is to find that the model that gives you that maximum likelihood estimation (Bruce & Bruce, 2017).

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7754  -0.5665  -0.5160  -0.3587   3.4498
```

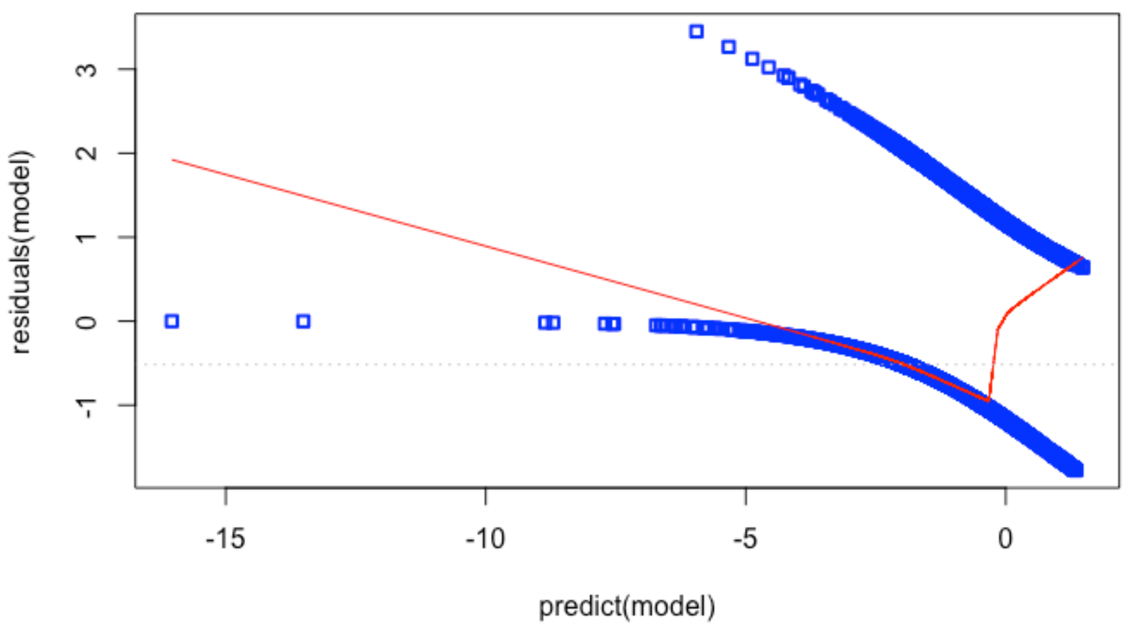


Figure 2.5 – Model residuals graphic.

Summary: Based on the aforesaid evaluation, it is accurate to say that the model yields a level of practicality while identifying potential default customers by considering indicators like previously missed payments, payments amount trend, or account balance (reflected on billed amount). One of the reasons the residual model shows a significant amount of negative values is due to the nature of the PAY_AMT variable, which in essence goes against letting the customer default the credit card account. Nevertheless, the model satisfies the initial intend of this study validating the feasibility for financial institutions to monitor their costumers' payment patters or characteristics.

As intended, prior of creating the model the entire dataset was preprocessed and structured for max efficiency. The training data was allotted with 70% of the original set and the rest 30% was preserved for the test data. In regard to the model, the generated coefficients depicted consistent high z-score values and significantly low probability values. In terms of residual deviance the model produced a residual deviance of 18743 on 20956 degrees of freedom, representing an 18% decrease of residual deviance.

In close, this results accentuate an ample business niche in where data analytics can contributed in a unique way the financial sector among many other branches. This assessment intended to weigh the logistic regression model capacity and efficacy while predicting potential credit card defaulting customer accounts. The elaborated R code, process and graphics displays how the inquiry was conducted. In order to generate a more robust model additional data inputs like income, expenses, household characteristics, job details, additional assets, etc. may be required. Additional recommendations include improving the quality of data and/or consistency of data collected, dynamic updates to a customer's profile based on registered expenses, and synchronizing near-real-time transactions to predict consumers inclinations or future expenses.

From a macro-perspective, the mentioned case illustrates how critical predictability or probability of default could be for financial institutions and their risk management departments when approving or denying credit applications. The following assessment and report aims to explore a customer credit card payments dataset and shows how the data mining technique of logistic regression could benefit the financial sector foreseeing credit card default payments based on customers payment trends and/or activities interrelated to their respective credit balances.

References

Bruce, B., & Bruce, A. (2017). Practical statistics for data science. O'Reilly Media, Inc.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients] Irvine, CA: University of California, School of Information and Computer Science.

Frankel, M. and Rosen, K. (2020). Credit Cards Debt Statistics for 2020. https://www.fool.com/the-ascent/research/credit-card-debt-statistics/

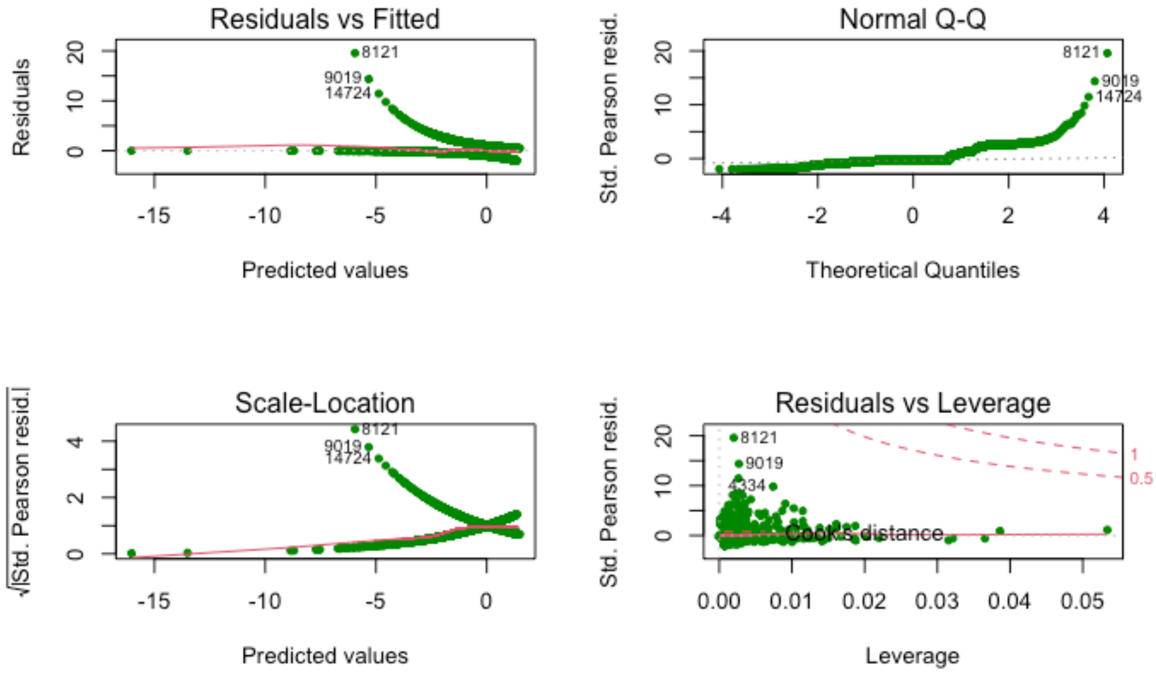
Normalization: https://www.datanovia.com/en/blog/how-to-normalize-and-standardize-data-in-r-for-great-heatmap-visualization

Hosmer, D.W., Lemeshow, S., & Sturdivant, R. (2013). Applied logistic regression. John Wiley & Sons. 3rd ed.

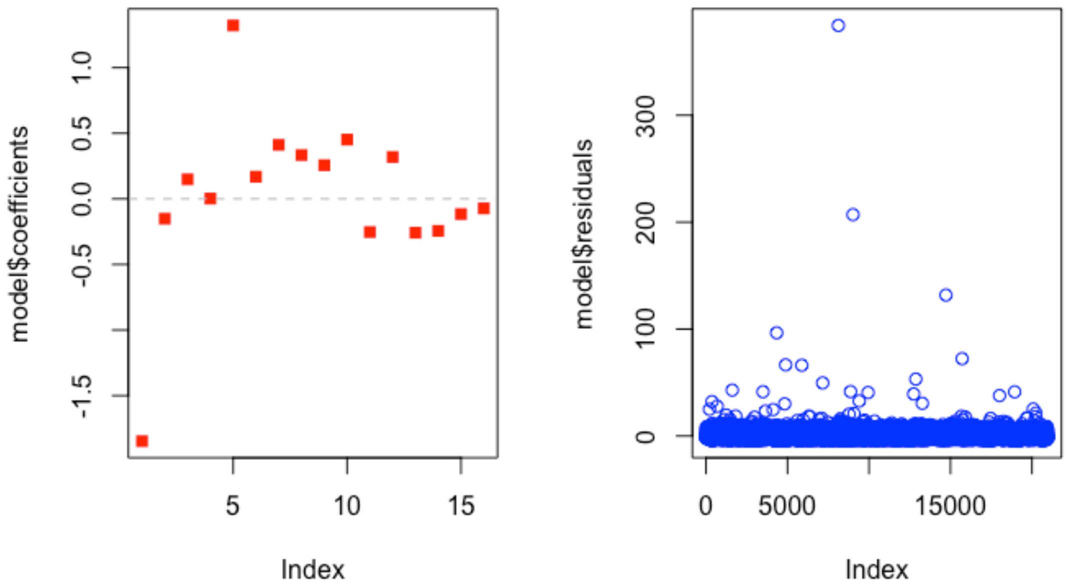
Appendix

```
> summary(model)      # Output the p value for each coefficient

Call:
glm(formula = PROJECTED_DEFAULT ~ ., family = binomial(), data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7754  -0.5665  -0.5160  -0.3587   3.4498

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.846898   0.083275 -22.178  < 2e-16 ***
SEXFemale       -0.151393   0.037697  -4.016 5.92e-05 ***
MARRIAGEMarried  0.149031   0.041621   3.581 0.000343 ***
AGE              0.002088   0.002227   0.938 0.348408
PAY_0Delayed     1.321684   0.049752  26.565  < 2e-16 ***
PAY_2Delayed     0.168766   0.067588   2.497 0.012525 *
PAY_3Delayed     0.411422   0.067690   6.078 1.22e-09 ***
PAY_4Delayed     0.333658   0.074731   4.465 8.01e-06 ***
PAY_5Delayed     0.256419   0.082360   3.113 0.001849 **
PAY_6Delayed     0.451449   0.070667   6.388 1.68e-10 ***
BILL_AMT1       -0.253182   0.109119  -2.320 0.020328 *
BILL_AMT2        0.319032   0.110110   2.897 0.003763 **
PAY_AMT1        -0.258284   0.046768  -5.523 3.34e-08 ***
PAY_AMT2        -0.244426   0.053516  -4.567 4.94e-06 ***
PAY_AMT4        -0.116164   0.033550  -3.462 0.000535 ***
PAY_AMT6        -0.072972   0.028968  -2.519 0.011768 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22126  on 20971  degrees of freedom
Residual deviance: 18743  on 20956  degrees of freedom
AIC: 18775

Number of Fisher Scoring iterations: 5
```

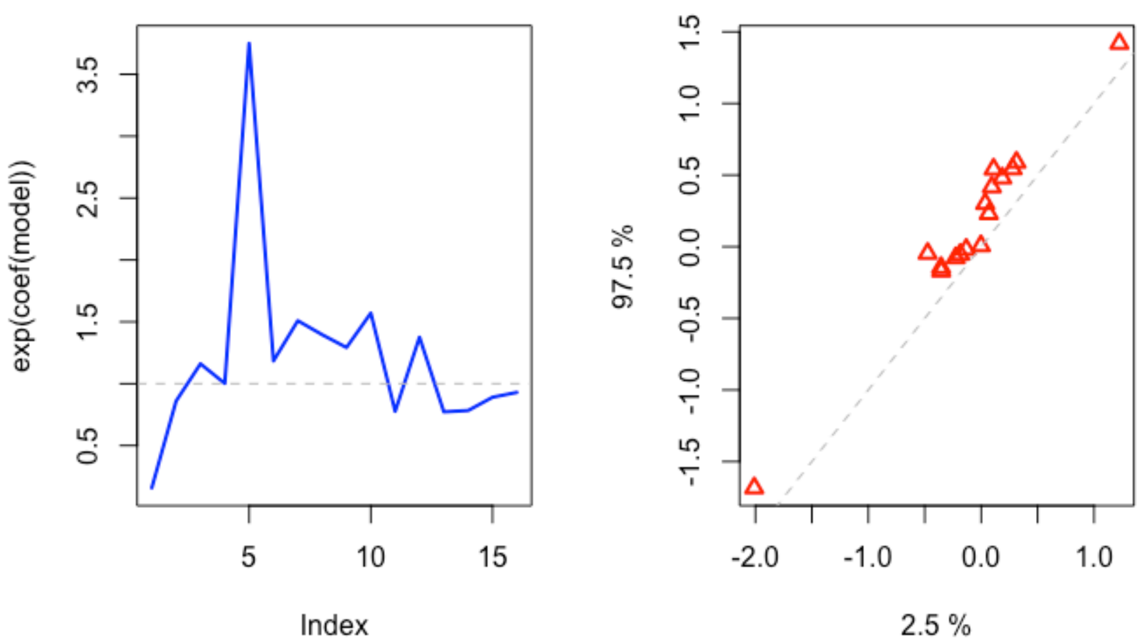
Appendix A. Initial model descriptive statistics.



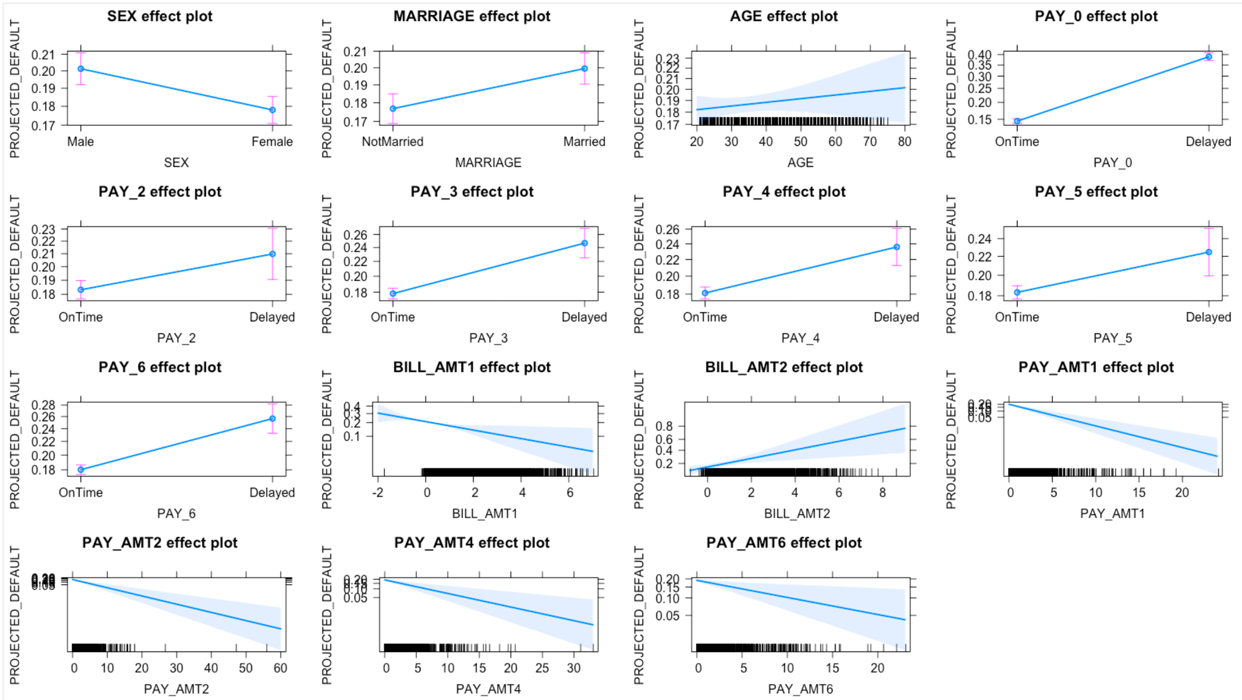
Appendix A1. Training data initial model.



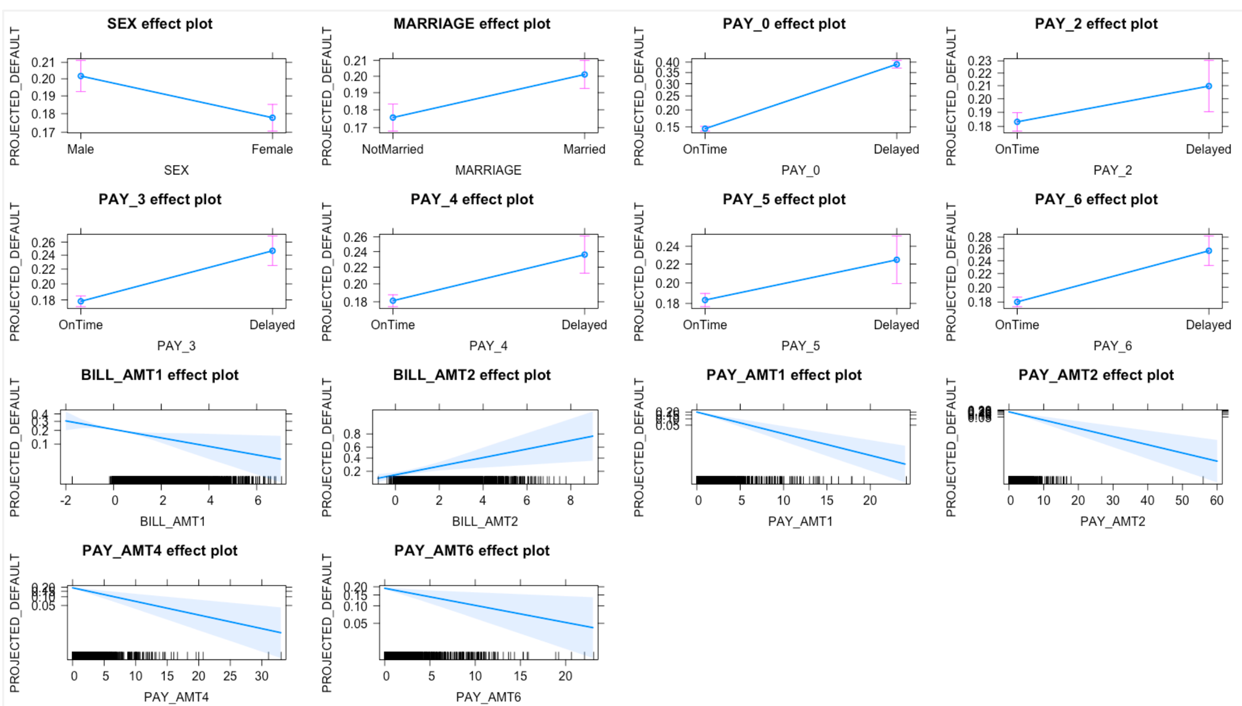
Appendix A2. Training data model coefficients and residuals.



Appendix A2. Training data model Odds ratios and confidence intervals.



Appendix A3 – Model 1, all-Effects plot.



Appendix A4 – Reduced model all-Effects plot.