Model #101: Credit Card Default Model

Model Development Check in #2 – Draft Model Development Guide

Daniel Macdonald, MILR, MHCS
2019 SP_MSDS_498-DL_Sec58
May 12, 2019

Description: Project check in #2 for Model Development Guide for MSDS capstone project. Draft project outline with emphasis on sections 2 to 5: Data Overview & Quality Check, Feature Engineering, Exploratory Data Analysis and Predictive Modelling.

Table of contents:

1. Introduction

TBD | General problem statement and discussion of approach with highlights of results.

2. Data Overview and Quality Check

The data provided for this capstone project are the 'Default of Credit Card Clients Data Set' posted on the UCI Machine Learning Repository.[1] Although the data are available via download from the site, there is a special revised version that is used by our class to ensure consistent application of train, test and validate split. Therefore, in the data dictionary below are five additional fields to facilitate consistent analysis, which are marked.

The description on the website suggests that the original data are a measure of client payment history and additional customer information (e.g. gender, age, marital status, bill history and balance) in relation to customer default. The data were provided by a bank in Taiwan in 2016 for the purposes of this study, and consists of 30,000 individual observations with 30 attributes. A description of the attributes is provided below. Each observation summarizes an individuals' available credit, billing and payment over a six-month period from April to September of 2005. The data are to be used in predictive modeling development to predict an individual's risk of defaulting on their payments prior to providing them the credit.

2.1 Data Dictionary

Below is a dictionary of the data fields. The original title 'PAY_0' has been changed to 'PAY_1' to conform to format established in BILL & PAY_AMT fields.

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

Title	Description	Notes
ID	Unique Identifier by row (1 to 30,000)	
LIMIT_BAL	Amount of the given credit (NT dollar)	individual consumer credit & supplementary credit
SEX	Gender	(1 = male; 2 = female)
EDUCATION	Education	(1 = graduate school; 2 = university 3 = high school; 4 = others)
MARRIAGE	Marital status	(1 = married; 2 = single; 3 = others)
AGE	Age	Years
PAY_1	the repayment status in September, 2005	Repayment Status range -2 to 8: negative numbers represent pre-payment/positive #'s late payment in months
PAY_2	the repayment status in August, 2005	
PAY_3	the repayment status in July, 2005	
PAY_4	the repayment status in June, 2005	
PAY_5	the repayment status in May, 2005	
PAY_6	the repayment status in April, 2005	
BILL_AMT1	the amount of Bill Statement in September, 2005 (NT Dollar)	
BILL_AMT2	the amount of Bill Statement in August 2005 (NT Dollar)	
BILL_AMT3	the amount of Bill Statement in July, 2005 (NT Dollar)	
BILL_AMT4	the amount of Bill Statement in June, 2005 (NT Dollar)	
BILL_AMT5	the amount of Bill Statement in May, 2005 (NT Dollar)	
BILL_AMT6	the amount of Bill Statement in April, 2005 (NT Dollar)	
PAY_AMT1	the amount of Previous Payment in September, 2005 (NT Dollar)	Payment variable lags behind prior month bill. i.e. we assume 'on time' payment September = PAY_AMT1 aligns with August Bill = BILL_AMT2
PAY_AMT2	the amount of Previous Payment in August 2005 (NT Dollar)	
PAY_AMT3	the amount of Previous Payment in July, 2005 (NT Dollar)	
PAY_AMT4	the amount of Previous Payment in June, 2005 (NT Dollar)	
PAY_AMT5	the amount of Previous Payment in May, 2005 (NT Dollar)	
PAY_AMT6	the amount of Previous Payment in April, 2005 (NT Dollar)	
DEFAULT	Classification - Default Y/N (Not clear if 0 or 1 is Y?	Dependent Variable (Y)
u	Random Sort for Train/Test/Validate Split	Not original to data set. Added for purposes of consistent analysis for MSDS 498
train	Dummy variable - 1 = Train data set	
test	Dummy variable - 1 = Test data set	
validate	Dummy variable - 1 = Validate data set	
data.group	Train/Test or Validate Data (1 = Train, 2=Test, 3 = Validate)	

2.2 Observations of original data

All original data are an integer data type, and the data that has been included for the purposes of a train, test, and validate split are a number data type. A table of the summary statistics of the data is found in appendix X.1. I have grouped together data fields by topic. Here are some high-level observations of each data group:

Data Group & Field Title	Notes:
Identification and Sort	
ID, u, train, test, validate, data.group	Data are originally sorted by **ID**, which is 1:30,000. **'u'** is a randomly generated number for the purpose of distributing observations. **'train','test', & 'validate'** are dummy variables where '1' indicates inclusion to group. **'data.group'** is a categorical variable - '1' = train, '2' = test, & '3' = validate.
Limit Balance	
Limit_Bal	Integer variable with minimum of 10,000 and maximum of 1,000,000 Median value is 140,000 with 25% and 75% Quantiles of 50,000:240,000 Histogram of distribution provided in Appendix X.2.1
Demographic Information	
Sex, Education, Marriage, Age	Histogram of distributions provided in Appendix X.2.2 **Sex** – binary value. Suggestion to change to value of '1' if Male & '0' if not.

	Education – Four categories in data description, but observation of additional categories (0,5&6) which = 'others'. Correlation analysis shown in Appendix X.2.2 suggest we bin to 2 categories (Grad & other; High School & University) **Marriage** – Four categorical values (0:3) in the data set. Values 0 & 3 would classify as 'other', but analysis (table X.2.2) suggest binary treatment (Y/N)
Pay Categories (On Time/Delayed)	
PAY_1 : PAY_6	Histogram of distribution provided in Appendix X.2.3
Monthly Bill Amounts	
BILL_AMT_1 : BILL_AMT_6	Histogram of distribution provided in Appendix X.2.4
Monthy Payment Amounts	
PAY_AMT_1 : PAY_AMT_6	Histogram of distribution provided in Appendix X.2.5
Response Variable	
DEFAULT	Histogram of distribution provided in Appendix X.2.6

Below is the observations counts by 'data.group' for the Train, Test & Validate splits:

Table 2.2.1: Train, Test, & Validate Splits

data.group	group	Freq	%
1	Train	15,180	50%
2	Test	7,323	25%
3	Split	7,497	25%

2.2.2 Notes on data discrepancies

Observations for data corrections prior to feature engineering:

- Education categories not in data description (0,5,&6) -> 'Other'
- Marriage category '0' classified as 'Other'
- PAY_1:PAY_6 negative numbers – likely to be reported as 'on time' = 0
- Some PAY_AMT values always above BILL_AMT, but still default?

3. Feature Engineering

Title	Description	Notes
ID	Unique Identifier by row (1 to 30,000)	
LIMIT_BAL	Amount of the given credit (NT dollar)	Range = 10k:1M
SEX_Female	Indicator if Sex = Female	(0 = male; 1 = female)
ED_Grad_Other	Indicator if EDUCATION = Grad or Other	(1 = Grad or Other)
ED_Univ_HS	Indicator if EDUCATION = University or High School	(1 =University or High School)
MARRIAGE_Y	Indicator if MARRIAGE = Yes	(1 = married)
AGE_Below_25	Indictor if AGE bin <= 25	(1 = AGE <= 25)
AGE_25to35	Indicator if AGE bin > 25 and <=35	(1= AGE > 25 & <= 35)
AGE_35to45	Indictor if AGE bin > 35 and <= 45	(1= AGE > 35 & <= 45)
AGE_above+45	Indicator if AGE bin > 45	(1 = AGE > 45)
PAY_X_Sum_6mo_belowZero	Indicator if Sum of 6 month PAY_X value is < 0	(1 = PAY_X 6 month Sum < 0)
PAY_X_Sum_6mo_belowSix	Indicator if Sum of 6 month PAY_X value is <= 6	(1 = PAY_X 6 month Sum <= 6)
PAY_X_Sum_6mo_aboveSix	Indicator if Sum of 6 month PAY_X value is > 6	(1 = PAY_X 6 month Sum > 6)
Avg_Pmt_Amt	Average of the PAY_AMT values across 6 months	(PAY_AMT1:PAY_AMT6)/6
Avg_Util	Average of the monthly utilization rates across 6 months	(Util_Bill_1:Util_Bill_6)/6
Avg_Pay_Ratio	Average of the monthly Pay_ratio across 5 observed payments	(PAY_AMT_X/BILL_AMT_X+1)/5
Balance_Growth_6mo	Difference between Balance and Bill (Δ over 6 months)	LIMIT_BAL - BILL_AMT_X
Util_Growth_6mo	Utilization rate for Bill 1 minus Utilization for Bill 6	Util_Bill_1 – Util_Bill_ 6
Max_Bill_Amt	Maximum amount observed across 6 month billing cycle	Range = -6k:1.6M
Max_Pmt_Amt	Maximum Payment observed across 6 month billing cycle	Range = 0:1.6M
Max_DLQ	Highest observed number in PAY_X variables	Range = 0:8
target	convert 'DEFAULT' variable to response in number format	(1 = DEFAULT)

3.1 SEX Variable

If SEX = Female (2) then 1

3.2 EDUCATION Variable

Review of correlation to response variable indicates that most effective binning using WOE would be to have two groups (table X.2.2.2). Each new variable is a binary indication of inclusion. For example ED_Grad_other = 1 when EDUCATION = 0,1, or 4+, and 0 if not.

- "ED_Grad_other" = ‘Graduate’ (EDUCATION = 1) & ‘Other’ (EDUCATION = 0 OR 4+)
- "ED_Univ_HS" = ‘University (EDUCATION = 2) & ‘High School’ (EDUCATION = 3)

3.3 MARRIAGE Variable

Analysis of the MARRIAGE classification ‘Other’ (table X.2.2.3) suggests that inclusion to this group does not have a significant correlation to the response when separate from the classification ‘single’. Therefore, the recommendation is to make this a binary indicator:

- "MARRIAGE_Y" (TRUE = 1)

3.4 AGE Variable Binning

WOE analysis of AGE data in Appendix X.2.3 indicates that bin distribution to maximize weight of correlation to response would be ‘<=25’, ‘<=35’,’<=45’ & ‘>45’ (table X.2.3.1). Therefore, four variables are created as a binary indication of inclusion to each group.

- "AGE_below_25","AGE_25to35","AGE_35to45","AGE_above_45" (TRUE = 1)

3.5 PAY_X Variable Binning

Analysis of the PAY_X variable in Appendix X.2.4 suggests that we use some sort of summing method across the 6 months to average the totals, and also use binning to optimize the distribution. In the analysis, different methods are compared, and the decision is to sum the PAY_X variables without transformation of PAY_X < 0. Once they are summed, then we split the observations into three groups, with a binary indication of inclusion to the group.

The recommended Groupings:

- "PAY_X_Sum_6mo_belowZero",
- " PAY_X_Sum_6mo_belowSix ",
- " PAY_X_Sum_6mo_aboveSix"

(TRUE = 1)

Table X.2.1b: Correlation of PAY_X 6Mo Sums Binned

	DEFAULT
PAY_X_Sum_6mo_belowZero	-0.36
PAY_X_Sum_6mo_belowSix	0.18
PAY_X_Sum_6mo_aboveSix	0.31

3.6 Avg_Pmt_Amt & Avg_Pay_Ratio

In the original data, we have the fields 'BILL_AMT' and 'PAY_AMT' for all six months tracked. In investigating the data, the assumption is that the 'PAY_AMT1' applied in September is a lagging variable to the 'BILL_AMT2' that was issued in August. Through these fields, we can calculate first the Average Pay Amount (Avg_Pmt_Amt) across the six months as well as the Average Pay Ratio (Avg_Pay_Ratio). The formulas for the calculation are shown below:

- Avg_Pmt_Amt <- SUM(PAY_AMT1:PAY_AMT6)/6
- Avg_Pmt_Ratio <- (PAY_AMT_X/BILL_AMT_X+1)/5 *(*PAY_AMT6 ratio is unknown)*

3.7 Avg_Util

In the original data, we have the fields 'BILL_AMT' and 'LIMIT_BAL'. First we calculate the Utilization for each BILL_X (Util_Bill_X). This variable will not likely make it into the final data set, but it will be used to calculate the Average Utilization (Avg_Util) across the six months.

- Util_Bill_X <- BILL_AMT_X/LIMIT_BAL *(*applied to all 6 months)*
- Avg_Util <- SUM(Util_Bill_1:Util_Bill_6)/6

3.8 Balance_Growth_6mo & Util_Growth_6mo

These variables look at the difference from the most recent Bill to the oldest Bill. The first variable measures the difference between the LIMIT_BAL and BILL_1 & BILL_6. It then takes the difference between the two. The second variable considers the Utilization rates for the same periods, and takes the difference.

- Balance_Growth_6mo <- LIMIT_BAL - BILL_AMT_X *(*applied to months 1 & 6)*

- Util_Growth_6mo <- Util_Bill_1 – Util_Bill_ 6

3.9 Max_Bill_Amt, Max_Pmt_Amt & Max_DLQ

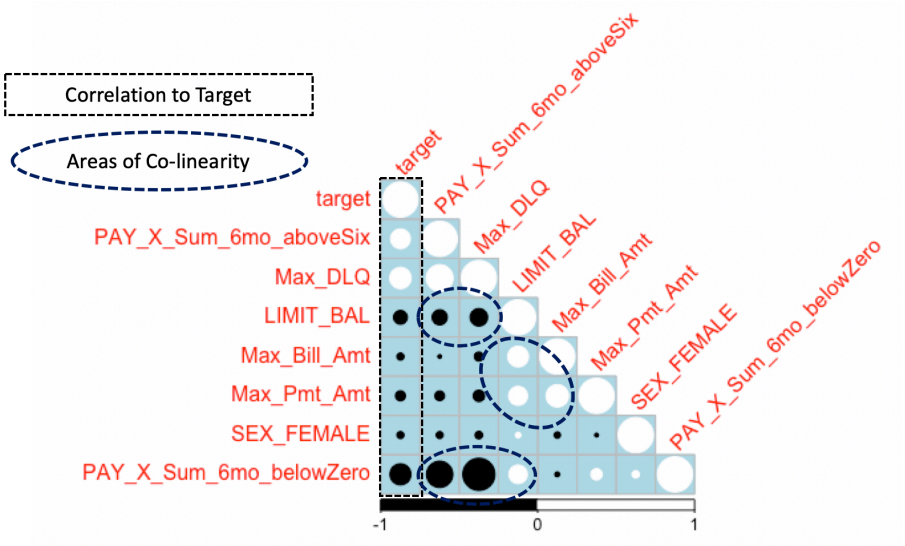
These variables look at the maximum value across the 6 month period for BILL_AMT_X, PAY_AMTX & PAY_X and simply report those as individual indicators of a pattern.

4. Exploratory Data Analysis

4.1 Exploratory Data Analysis – Engineered Features

Initial correlation analysis in appendix X.4.? shows highest correlation to target in variables: 'PAY_X_Sum_6mo' below zero and above 6 and MAX_DLQ as well as some cross correlation.

Figure 4.1.1 – Matrix of highest correlations to target



When we look at the two most influential variables in a mosaic format, the suggestion is that individuals with either a high delinquency rate or a high sum of PAY_X variables is the most highly relevant predictor of default (target = 1)

Figure 4.1.2 –
Mosaic - Max_DLQ to target

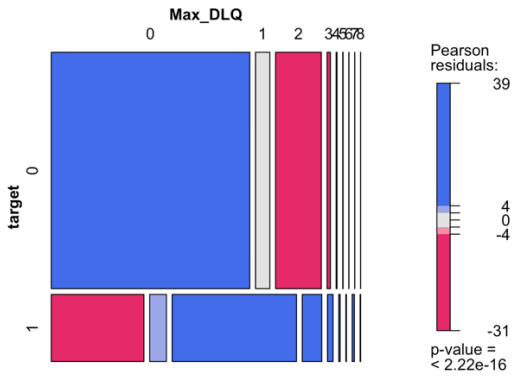
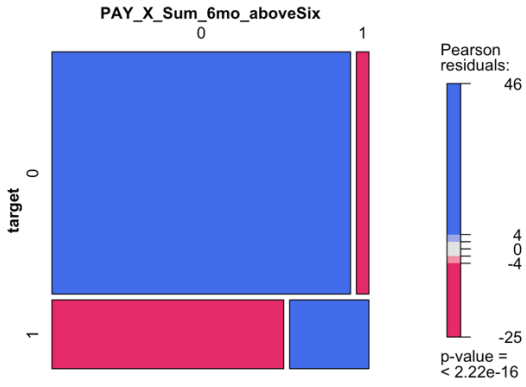


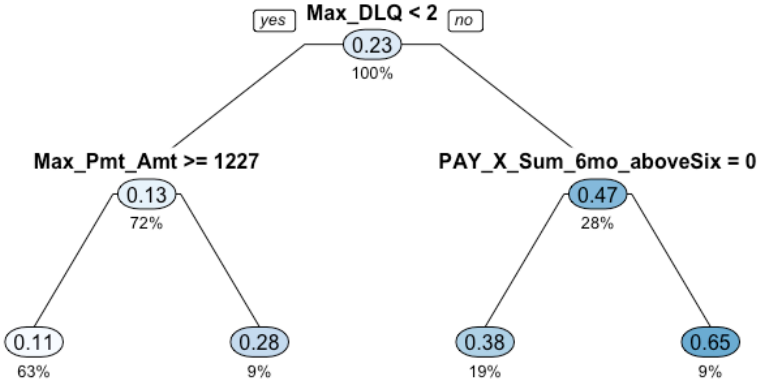
Figure 4.1.3 –
Mosaic - PAY_X_Sum_6mo_aboveSix to target



4.2 Exploratory Data Analysis part 2 – Model Based EDA

Initial GLM model and decision tree analysis in appendix X.4.? show the most relevant variables. During the first transformation many of them were not changed, so decision to re-evaluate binning for highligted variables.

Figure 4.2.1 Decision Tree Analysis



Variables Listed by Importance:

Max_DLQ
PAY_X_Sum_6mo_belowZero
PAY_X_Sum_6mo_belowSix
PAY_X_Sum_6mo_aboveSix
Max_Pmt_Amt
Avg_Pay_Ratio
Avg_Pmt_Amt
Max_Bill_Amt
Avg_Util
LIMIT_BAL

WOE Binning analysis and transformation is shown in appendix X.4.?

4.3 Updated Data Fields following second transformation

Figure 4.3.1

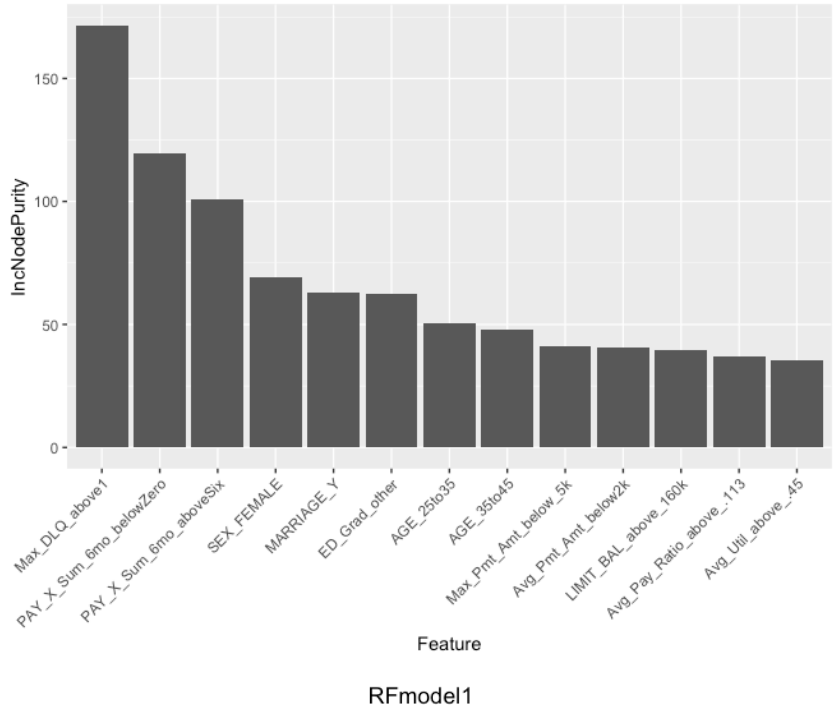
Updated Data Fields following EDA	
SEX_Female	Avg_Pay_Ratio_below_.035
ED_Grad_Other	Avg_Pay_Ratio_above_.113
MARRIAGE_Y	Avg_Pay_Ratio_above_1
AGE_Below_25	Balance_Growth_6mo_below_minus21k
AGE_25to35	Balance_Growth_6mo_below_minus10k
AGE_35to45	Balance_Growth_6mo_above_1k
AGE_above_45	Util_Growth_6mo_below_minus.03
PAY_X_Sum_6mo_belowZero	Util_Growth_6mo_above_0
PAY_X_Sum_6mo_aboveSix	Max_Bill_Amt_below_600
Avg_Pmt_Amt_below2k	Max_Bill_Amt_below_4k
Avg_Pmt_Amt_above12k	Max_Bill_Amt_below_18k
LIMIT_BAL_below_30k	Max_Bill_Amt_below_21k
LIMIT_BAL_above_160k	Max_Bill_Amt_above_52k
Avg_Util_below_.001	Max_Pmt_Amt_below_168
Avg_Util_above_.45	Max_Pmt_Amt_below_5k
Max_DLQ_above1	Max_Pmt_Amt_above_36k
target	

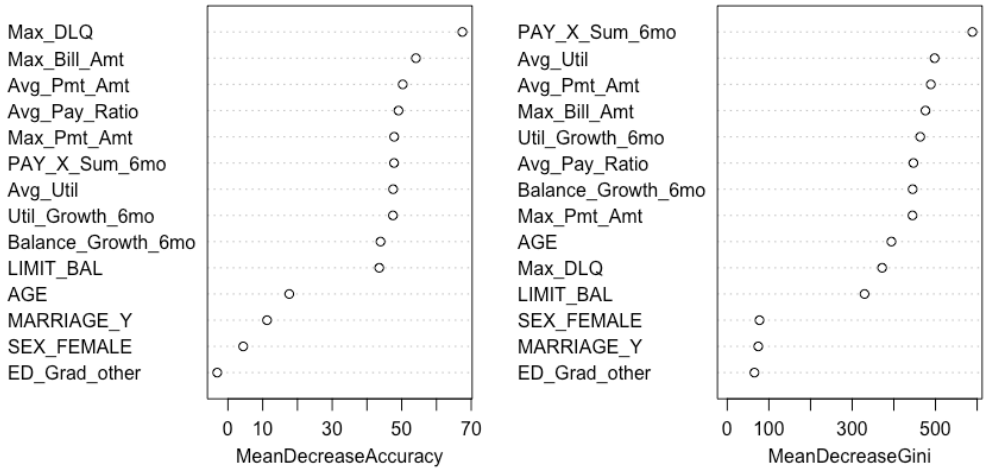
5. Predictive Modeling: Methods and Results

- In this section we will provide the results from four modeling approaches. Three of these modeling approaches are defined for you, and you get to choose the fourth approach from the list of choices.
- For each model provide any relevant or useful model output and a table of the model performance in-sample (i.e. on the training data set) and out-of-sample (i.e. on the test data set).
- The metrics to be measured are: (1) true positive rate or sensitivity, (2) false positive rate, and (3) the accuracy.

5.1 Random Forest Model

Figure 5.1.1 – Importance of Variables in Model





5.1 Random Forest Model – Evaluation

Table xx: Model 1 Training Data Confusion Matrix

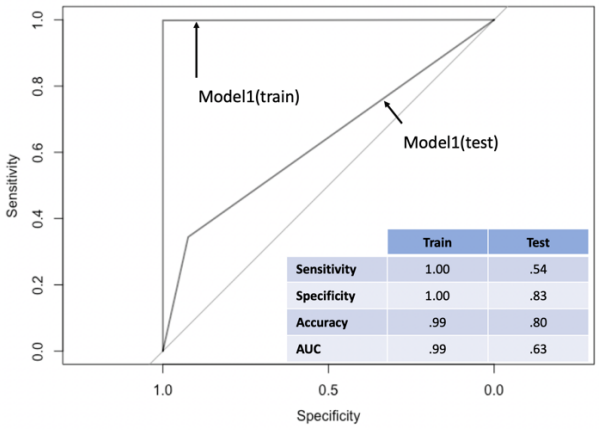
Target	Predicted 0	Predicted 1
0	1.00	0.00
1	0.00	1.00

Table xx: Model 1 Test Data Confusion Matrix

Target	Predicted 0	Predicted 1
0	0.92	0.07
1	0.65	0.34

Model is showing signs of over fitting..

Figure XX: ROC curve for Random Forest train & test



5.2 Gradiant Boosting

TBD

5.3 Logistic Regression with Variable Selection

Table XX: Model #1b

Dependent Variable	Estimate	Std. Error
Constant	-1.757***	0.13
SEX_FEMALE	-0.137**	0.04
MARRIAGE_Y	0.199***	0.04
PAY_X_Sum_6mo_belowZero	0.383***	0.10
PAY_X_Sum_6mo_aboveSix	0.910***	0.07
Avg_Pmt_Amt_below2k	0.317***	0.05
LIMIT_BAL_above_160k	-0.240***	0.05
Avg_Util_above_.45	0.405***	0.06
Avg_Pay_Ratio_above_.113	-0.170**	0.06
Util_Growth_6mo_above_0	-0.091*	0.05
Max_Bill_Amt_below_600	0.886***	0.12
Max_Bill_Amt_below_4k	0.515***	0.08
Max_Pmt_Amt_below_168	0.473***	0.11
Max_DLQ_above1	1.078***	0.09
Observations	15,180	
Log Likelihood	-6,841	
Akaike Inf. Crit.	13,710	
Note:	*p<0.1; **p<0.05; ***p<0.01	

5.3.2 Logistic Regression – Model Evaluation

Table xx: Model 1 **Table xx: Model 1b**

Figure XX: ROC curve for Model 1 & 1b on xTrain data

Training Data Confusion Matrix

Target \ Predicted	0	1
0	0.78	0.38
1	0.22	0.62

Training Data Confusion Matrix

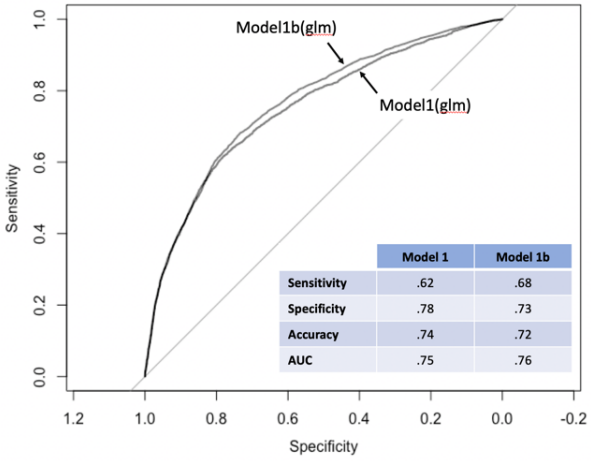
Target \ Predicted	0	1
0	0.73	0.26
1	0.31	0.68

Table xx: Model Test Data Confusion Matrix

Target \ Predicted	0	1
0	0.77	0.23
1	0.37	0.63

Table xx: Model Test Data Confusion Matrix

Target \ Predicted	0	1
0	0.73	0.27
1	0.31	0.68



5.4 Selected Model - TBD

6. Comparison of Results

- Aggregate your results from Section 5 and discuss. All of your model metrics should be presented in a single table for all models for both the training and test data sets. You should be able to compare and contrast the model performance with discussion and easily determine which model performed best.

7. Conclusions

- Conclude your paper. Reiterate your problem and highlight your results.
- How would you characterize the overall quality of your results?
- Do you have any recommendations for approaching the problem in a different manner or with different techniques? Would you recommend any particular avenues for future research?
- A lot of times a good conclusion reads like a good abstract.

8. Bibliography (if needed)

- References can be constructed using any valid style format. However, references can be cited in your paper using a name-year citing or by using the number scheme, e.g. Bhatti [1].
- One type of citing used in the program is the APA style format. However, it can be difficult to use with some types of sources, hence we will allow the number format for convenience.
- When citing references consider using: Bhatti [1] for a single author, Bhatti and Lucas [2] for two authors, and Bhatti et. al. [3] for three or more authors.

X. Appendices

X.1 Summary Statistics for Credit Card Default Original Data

Below are the summary statistics for the original data. I have highlighted some of the observed data discrepancies relative to the proposed data description. These discrepancies are

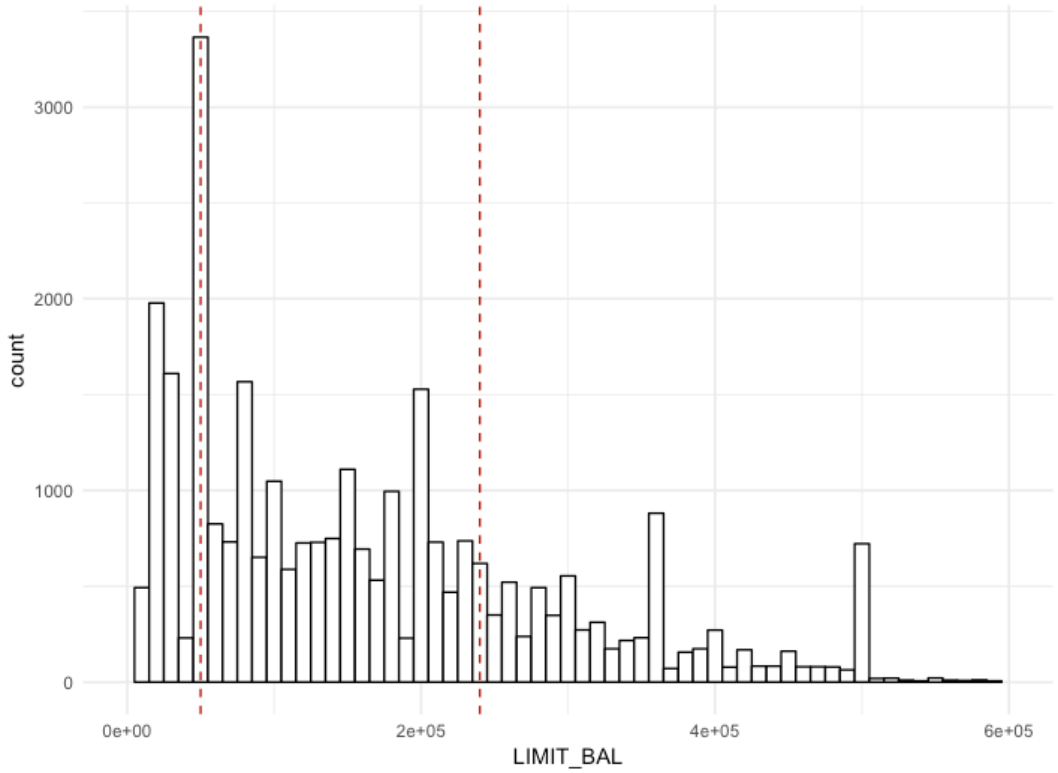
Table X.1.1: Summary Statistics for Credit Card Default

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
ID	30,000	15,000.50	8,660.40	1	7,500.8	15,000.5	22,500.2	30,000
LIMIT_BAL	30,000	167,484.30	129,747.70	10,000	50,000	140,000	240,000	1,000,000
SEX	30,000	1.60	0.49	1	1	2	2	2
EDUCATION	30,000	1.85	0.79	0	1	2	2	6
MARRIAGE	30,000	1.55	0.52	0	1	2	2	3
AGE	30,000	35.49	9.22	21	28	34	41	79
PAY_1	30,000	-0.02	1.12	-2	-1	0	0	8
PAY_2	30,000	-0.13	1.20	-2	-1	0	0	8
PAY_3	30,000	-0.17	1.20	-2	-1	0	0	8
PAY_4	30,000	-0.22	1.17	-2	-1	0	0	8
PAY_5	30,000	-0.27	1.13	-2	-1	0	0	8
PAY_6	30,000	-0.29	1.15	-2	-1	0	0	8
BILL_AMT1	30,000	51,223.33	73,635.86	-165,580	3,558.8	22,381.5	67,091	964,511
BILL_AMT2	30,000	49,179.08	71,173.77	-69,777	2,984.8	21,200	64,006.2	983,931
BILL_AMT3	30,000	47,013.15	69,349.39	-157,264	2,666.2	20,088.5	60,164.8	1,664,089
BILL_AMT4	30,000	43,262.95	64,332.86	-170,000	2,326.8	19,052	54,506	891,586
BILL_AMT5	30,000	40,311.40	60,797.16	-81,334	1,763	18,104.5	50,190.5	927,171
BILL_AMT6	30,000	38,871.76	59,554.11	-339,603	1,256	17,071	49,198.2	961,664
PAY_AMT1	30,000	5,663.58	16,563.28	0	1,000	2,100	5,006	873,552
PAY_AMT2	30,000	5,921.16	23,040.87	0	833	2,009	5,000	1,684,259
PAY_AMT3	30,000	5,225.68	17,606.96	0	390	1,800	4,505	896,040
PAY_AMT4	30,000	4,826.08	15,666.16	0	296	1,500	4,013.2	621,000
PAY_AMT5	30,000	4,799.39	15,278.31	0	252.5	1,500	4,031.5	426,529
PAY_AMT6	30,000	5,215.50	17,777.47	0	117.8	1,500	4,000	528,666
DEFAULT	30,000	0.22	0.42	0	0	0	0	1
u	30,000	0.50	0.29	0.0000	0.25	0.49	0.75	1.00
train	30,000	0.51	0.50	0	0	1	1	1
test	30,000	0.24	0.43	0	0	0	0	1
validate	30,000	0.25	0.43	0	0	0	0	1
data.group	30,000	1.74	0.83	1	1	1	2	3

X.2 Original Data Exploration Histograms and Plots:

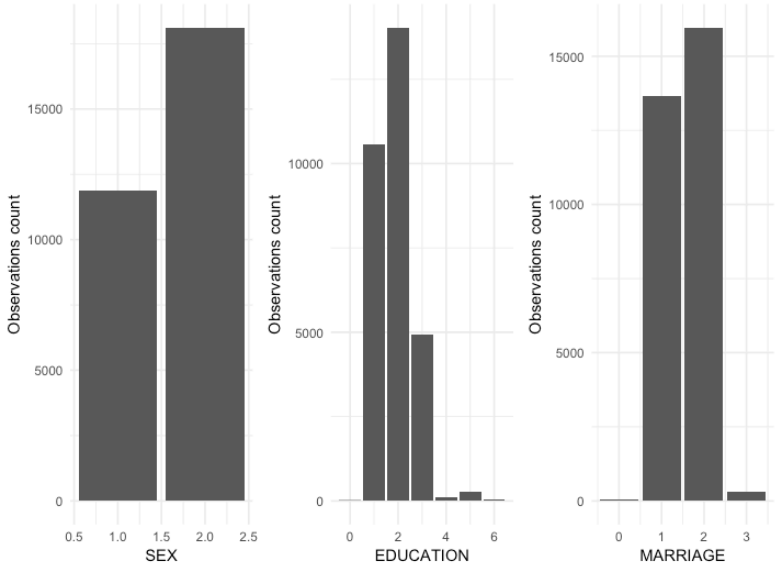
This section is the plots and tables used to support the 'Data Overview and Quality Check' in section 2. Observations of data discrepancies and analysis of approaches to binning will support further data transformation in section 3.

X.2.1 LIMIT_BAL Variable



Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
LIMIT_BAL	30,000	167,484.30	129,747.70	10,000	50,000	140,000	240,000	1,000,000

X.2.2 Demographic data: SEX, EDUCATION, & MARRIAGE



Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
SEX	30,000	1.60	0.49	1	1	2	2	2
EDUCATION	30,000	1.85	0.79	0	1	2	2	6
MARRIAGE	30,000	1.55	0.52	0	1	2	2	3

Analysis of EDUCATION categories and WOE analysis for binning:

Table X.2.2.1a: Summary of EDUCATION

EDUCATION	Freq	PCT
0	14	0%
1	10,585	35.3%
2	14,030	46.8%
3	4,917	16.4%
4	123	0.4%
5	280	0.9%
6	51	0.2%

* group 0,4:6 as 'Other' = 4

Table X.2.2.1b: Education Correlation

Group	DEFAULT
grad	-0.05
univ	0.04
high	0.03
other	-0.05

Correlation analysis suggests binning for Grad/Other and one for High School/University. WOE analysis below supports this idea (in table below 'other' is changed to value of 0)

Table X.2.2.2: Recommended Education Bins

	Final.Bin	Total.Count	Total.Distr.	0.Count	1.Count	0.Distr.	1.Distr.	1.Rate	WOE	IV
1	< = 1	11,053	36.8%	8,984	2,069	38.5%	31.2%	18.7%	21.0	0.015
2	< = Inf	18,947	63.2%	14,380	4,567	61.5%	68.8%	24.1%	-11.2	0.008
4	Total	30,000	100.0%	23,364	6,636	100.0%	100.0%	22.1%	NA	0.023

Analysis of MARRIAGE categories and consideration of dimension reduction

Initially, any '0' values are changed to '3' = Other. But following analysis of correlation to the response variable, the differentiation of this class seems insignificant. Looking at Table X.2.3b, recommendation is to group 'Other' into 'Married_N'.

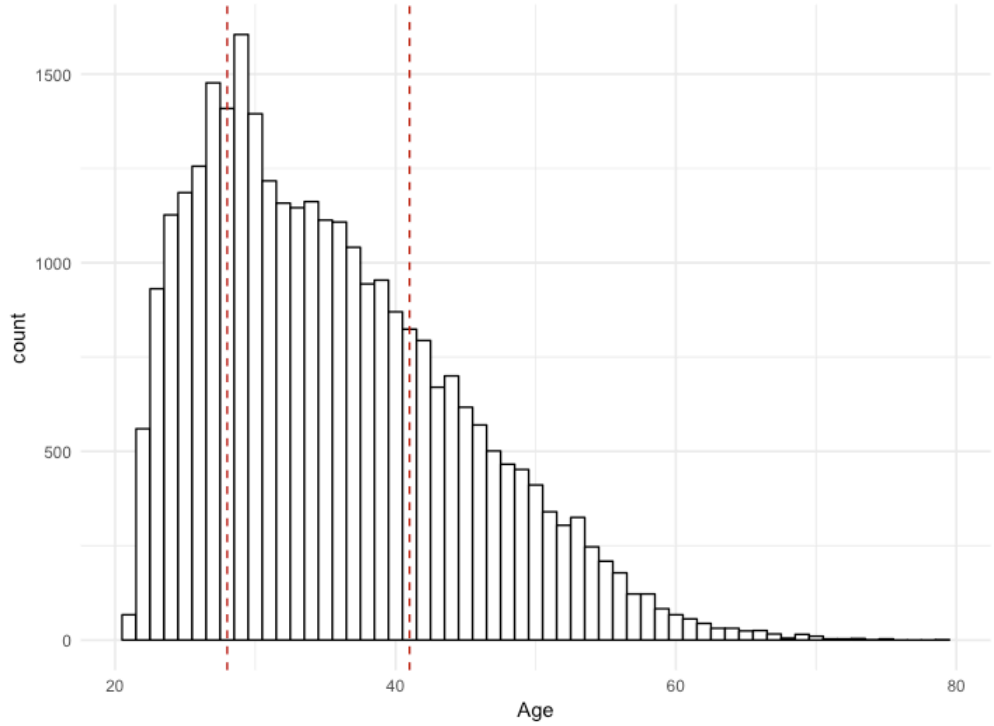
Table X.2.3a: Correlation Matrix (Married Y/N/Other)

	DEFAULT
Married_Y	0.0298
Married_N	-0.0306
Married_Other	0.0040

Table X.2.3b: Correlation Matrix (Married Y/N – drop 'other')

	DEFAULT
Married_Y	0.0298
Married_N	-0.0298

X.2.3 Age data and WOE Binning analysis



Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
AGE	30,000	35.49	9.22	21	28	34	41	79

Table X.3.3.1: Correlation of age bins

	DEFAULT
AGE_below_25	0.04
AGE_25to35	-0.05
AGE_35to45	-0.004
AGE_above_45	0.03

Below is WOE analysis of AGE variable, which suggests grouping the variable into four groups: <=25, 25:35, 35:45, and over 45, which will be the recommended bins for transformation. In the table to the left is an analysis of the correlation to the response variable by group.

Table X.2.3.1: WOE analysis of Age Bins

AGE Bins	Total Count	Total Dist	Count (0)	Count (1)	Age (1) Rate	WOE	IV
< = 25	3,871	12.9%	2,839	1,032	26.7%	-24.7	0.008
< = 35	12,938	43.1%	10,373	2,565	19.8%	13.9	0.008
< = 45	8,522	28.4%	6,661	1,861	21.8%	1.6	0.000
< = Inf	4,669	15.6%	3,491	1,178	25.2%	-17.2	0.005
Total	**30,000**	**100.0%**	**23,364**	**6,636**	**22.1%**	**NA**	**0.021**

X.2.4 PAY_X Variable and WOE Binning Analysis

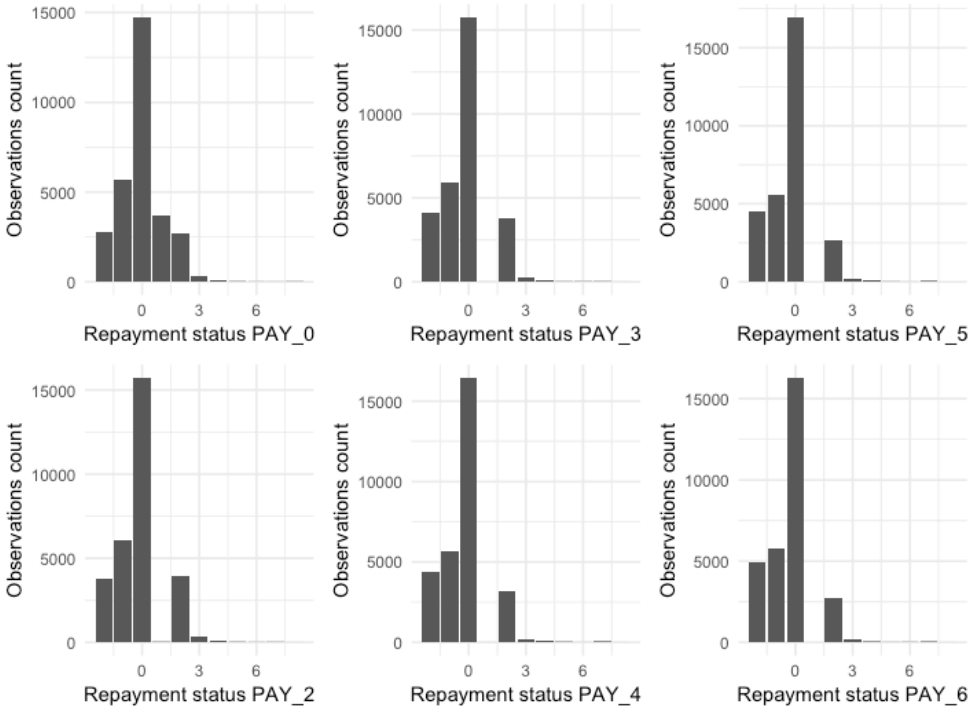


Table X.2.3: Summary of PAY as % of observed in WOE recommended Bins							Pay_1_Correlation to 'DEFAULT' = 1
Bins	**PAY_1**	**PAY_2**	**PAY_3**	**PAY_4**	**PAY_5**	**PAY_6**	
< = 0	77.3%	85.2%	86%	88.3%	90.1%	89.7%	**-.36**
1	12.3%	0.1%	0%	0%	NA%	NA%	**.10**
> =2	10.4%	14.7%	14%	11.7%	9.9%	10.3%	**.39**

PAY_X variables show high levels of cross correlation, which makes sense due to patterns in people's behavior over the months. Below is a summary of the two methods used to summarize the PAY_X data and then bin the results to determine the best way to treat the variable. The recommendation is to bin the sum of the PAY variables and then bin them into 3 groups: <= 0, > 0 & <=6, and > 6 as per WOE analysis in Table X.2.3.2

Table X.2.3.1a: Correlation of PAY Variables

	DEFAULT
PAY_1	0.32
PAY_2	0.26
PAY_3	0.24
PAY_4	0.22
PAY_5	0.20
PAY_6	0.19

Table X.2.1b: Correlation of PAY Sums Binned

	DEFAULT
PAY_X_Sum_6mo_belowZero	-0.36
PAY_X_Sum_6mo_belowSix	0.18
PAY_X_Sum_6mo_aboveSix	0.31

** Greatest differentiation and minimum cross correlation achieved by summing the monthly PAY_X variables and then binning as per WOE analysis (<=0, <= 6, & >6)

Table X.2.3.1c: Correlation of PAY_X Max/Min Binned

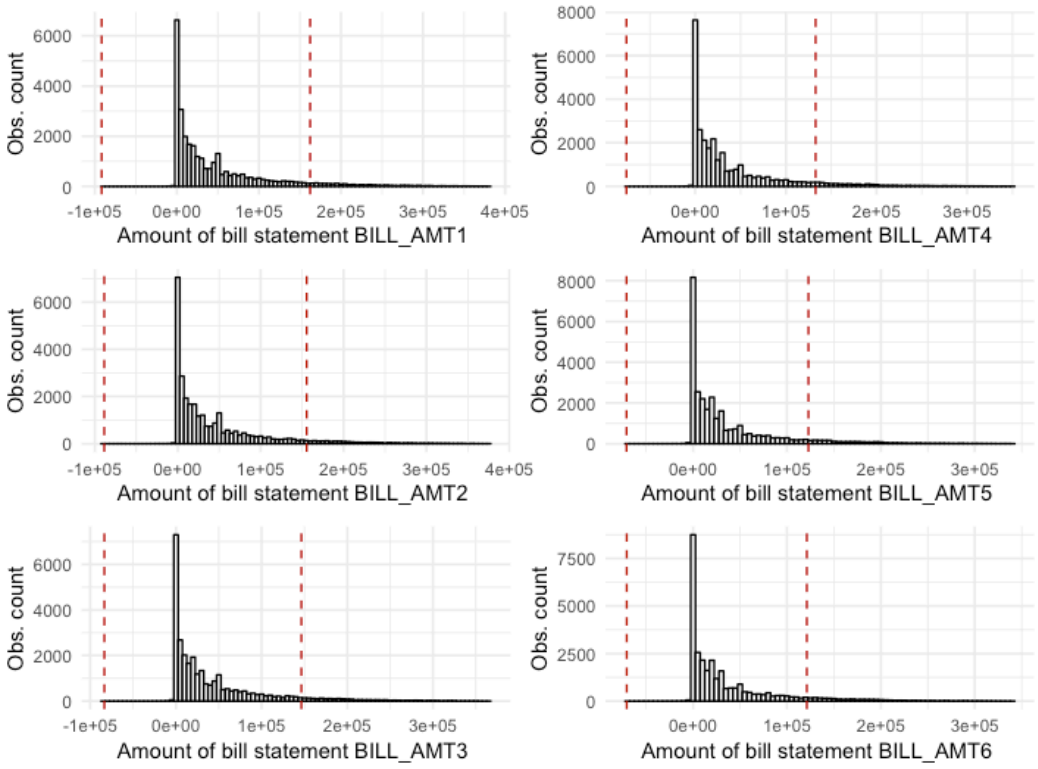
	DEFAULT
Max_belowZero	-0.35
Min_aboveZero	0.25

WOE analysis of 6mo sum method indicates most effective binning would be summed totals <= 0, between 0:6, and >6. Correlation analysis above indicates this is the best approach.

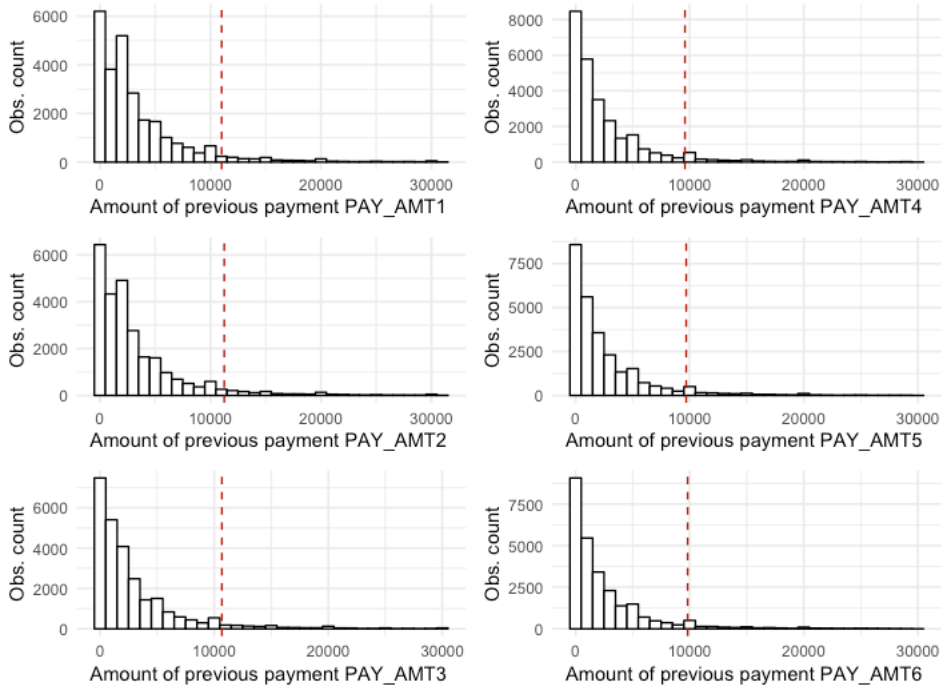
Table X.2.3.2: WOE Analysis of PAY Binning

Final Bin	Total Count	Total Distr.	Rate (1)	WOE	IV
< = 0	22,867	76.2%	13.8%	57.2	0.210
< = 6	4,500	15.0%	39.7%	-83.9	0.128
< = Inf	2,633	8.8%	64.3%	-184.5	0.396
Total	30,000	100.0%	22.1%	NA	0.735

X.2.5 BILL_AMTX


Obs. count
Amount of bill statement BILL_AMT1
Amount of bill statement BILL_AMT4
Amount of bill statement BILL_AMT2
Amount of bill statement BILL_AMT5
Amount of bill statement BILL_AMT3
Amount of bill statement BILL_AMT6

X.2.6 PAY_AMTX

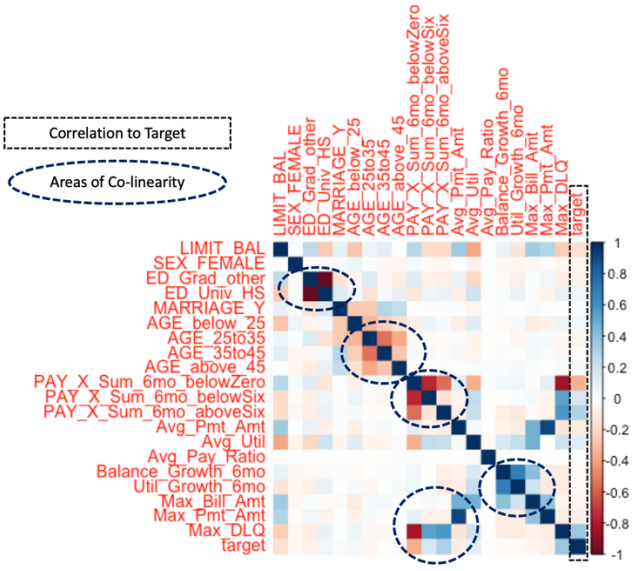


X.4.? Initial correlation of transformed data

Table XX: Correlation of first transformation

	target
LIMIT_BAL	-0.15
SEX_FEMALE	-0.04
ED_Grad_other	-0.06
ED_Univ_HS	0.06
MARRIAGE_Y	0.03
AGE_below_25	0.04
AGE_25to35	-0.05
AGE_35to45	-0.004
AGE_above_45	0.03
PAY_X_Sum_6mo_belowZero	-0.36
PAY_X_Sum_6mo_belowSix	0.18
PAY_X_Sum_6mo_aboveSix	0.31
Avg_Pmt_Amt	-0.10
Avg_Util	0.12
Avg_Pay_Ratio	-0.01
Balance_Growth_6mo	-0.03
Util_Growth_6mo	-0.02
Max_Bill_Amt	-0.04
Max_Pmt_Amt	-0.08
Max_DLQ	0.37
target	1

Figure XX: Correlation of first transformation



Initial correlation analysis suggests some overlap and redundancy in binned data from transformed data set, such as EDUCATION and AGE bins. Most highly correlated fields are 'Max_DLQ' and 'PAY_X_Sum_6mo' above 6 and below zero.

X.5.? Summary of GLM Model 1 – model exploration

Preliminary General Linear Model analysis demonstrates range of variable significance and redundancy in binned data as suggested above. This is due to the fact that all bins are used as variables for consideration. For example, if we consider the 'AGE' variable, and the subsequent binning, suggestion is to split the data into 4 categories. For the purposes of modelling, one of the four categories is redundant (NA below).

Table XX: Model #1

Dependent Variable	Estimate	Std. Error
Constant	-0.18	0.15
LIMIT_BAL	**-0.000001*** **	**0.000000**
SEX_FEMALE	**-0.13*** **	**0.04**
ED_Grad_other	0.01	0.05
ED_Univ_HS	NA	NA
MARRIAGE_Y	**0.17*** **	**0.05**
AGE_below_25	-0.03	0.08
AGE_25to35	**-0.14** **	**0.07**
AGE_35to45	0.01	0.07
AGE_above_45	NA	NA
PAY_X_Sum_6mo_belowZero	**-1.27*** **	**0.11**
PAY_X_Sum_6mo_belowSix	**-0.78*** **	**0.08**
PAY_X_Sum_6mo_aboveSix	NA	NA
Avg_Pmt_Amt	**-0.0001*** **	**0.00001**
Avg_Util	**-0.23** **	**0.10**
Avg_Pay_Ratio	-0.001	0.002
Balance_Growth_6mo	0.000000	0.000001
Util_Growth_6mo	0.04	0.10
Max_Bill_Amt	**0.000002*** **	**0.000001**
Max_Pmt_Amt	0.00001***	0.000003
Max_DLQ	0.43***	0.04
Observations	15,180	
Log Likelihood	-6,964.87	
Akaike Inf. Crit.	13,965.73	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table xx: Model 1 Confusion Matrix

… comments here

		Predicted	
		0	**1**
Target	**0**	0.78	0.38
	1	0.22	0.62

X.5.? Additional variable transformation & exploration

As a result of this analysis, decision to go back and evaluate binning approach for all continuous variables: Max_DLQ, Avg_Pmt_Amt, LIMIT_BAL, Avg_Util, Avg_Pay_Ratio, Balance_Growth_6mo, Util_Growth_6mo, Max_Bill_Amt, Max_Pmt_Amt

Table XX: WOE Analysis of Max_DLQ

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = 1	21,620	72.1%	12.7%	66.5	0.261
< = Inf	8,380	27.9%	46.3%	-111.0	0.435
Total	30,000	100.0%	22.1%	NA	0.696

```
## split Max_DLQ above 1
df_T3$Max_DLQ_above1 <- ifelse(df_T3$Max_DLQ > 1,1,0)
```

Table XX: WOE Analysis of Avg_Pmt_Amt

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = 2045	13,500	45.0%	29.3%	-37.7	0.071
< = 12000	13,500	45.0%	17.7%	27.6	0.032
< = Inf	3,000	10.0%	9.6%	98.0	0.071
Total	30,000	100.0%	22.1%	NA	0.173

```
## split Avg_Pmt_Amt into 3 groups: <= 2045, <= 11958, > 11958
df_T3$Avg_Pmt_Amt_below2k <- ifelse(df_T3$Avg_Pmt_Amt <= 2045,1,0)
df_T3$Avg_Pmt_Amt_above12k <- ifelse(df_T3$Avg_Pmt_Amt > 12000,1,0)
```

Table XX: WOE Analysis of LIMIT_BAL

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = 30000	4,081	13.6%	35.8%	-67.7	0.073
< = 160000	13,013	43.4%	24.5%	-13.1	0.008
< = Inf	12,906	43.0%	15.4%	44.3	0.074
Total	30,000	100.0%	22.1%	NA	0.155

```
## split LIMIT_BAL into 3 groups: <= 30k, <= 160k, > 160k
df_T3$LIMIT_BAL_below_30k <- ifelse(df_T3$LIMIT_BAL <= 30000,1,0)
df_T3$LIMIT_BAL_above_160k <- ifelse(df_T3$LIMIT_BAL > 160000,1,0)
```

Table XX: WOE Analysis of Avg_Util

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = 0.0010	1,500	5.0%	30.0%	-41.1	0.009
< = 0.4517	16,500	55.0%	17.1%	31.8	0.051
< = Inf	12,000	40.0%	28.0%	-31.4	0.043
Total	30,000	100.0%	22.1%	NA	0.103

```
## split Avg_Util into 3 groups: <= .001, <= .45, > .45
df_T3$Avg_Util_below_.001 <- ifelse(df_T3$Avg_Util <= .001,1,0)
df_T3$Avg_Util_above_.45 <- ifelse(df_T3$Avg_Util > .45,1,0)
```

Table XX: WOE Analysis of Avg_Pay_Ratio

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = 0.0352	1,500	5.0%	46.6%	-112.2	0.080
< = 0.1134	12,000	40.0%	25.3%	-17.6	0.013
< = 1	13,395	44.6%	19.0%	18.9	0.015
< = Inf	3,105	10.3%	11.3%	80.1	0.052
Total	30,000	100.0%	22.1%	NA	0.160

```
## split Avg_Pay_Ratio into 3 groups: <= .035, <= .113, <= 1, > 1
df_T3$Avg_Pay_Ratio_below_.035 <- ifelse(df_T3$Avg_Pay_Ratio <= .035,1,0)
df_T3$Avg_Pay_Ratio_above_.113 <- ifelse(df_T3$Avg_Pay_Ratio > .035 &
                                         df_T3$Avg_Pay_Ratio <= .113,1,0)
df_T3$Avg_Pay_Ratio_above_1 <- ifelse(df_T3$Avg_Pay_Ratio > 1,1,0)
```

Table XX: WOE Analysis of Balance_Growth_6mo

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = -21881.5	1,500	5.0%	12.7%	66.6	0.018
< = -10172.8	1,500	5.0%	19.7%	14.9	0.001
< = 923	12,002	40.0%	29.0%	-36.3	0.058
< = Inf	14,998	50.0%	17.8%	27.2	0.034
Total	30,000	100.0%	22.1%	NA	0.111

```
## split Balance_Growth_6mo into 3 groups: <= -21k, <= -10k, <= 1k, > 1k
df_T3$Balance_Growth_6mo_below_minus21k <- ifelse(df_T3$Balance_Growth_6mo <= -21800,1,0)
df_T3$Balance_Growth_6mo_below_minus10k <- ifelse(df_T3$Balance_Growth_6mo > -21800 &
                                                  df_T3$Balance_Growth_6mo <= -10000,1,0)
df_T3$Balance_Growth_6mo_above_1k <- ifelse(df_T3$Balance_Growth_6mo >= 1000,1,0)
```

Table XX: WOE Analysis of Util_Growth_6mo

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = -0.02909	7,500	25.0%	29.2%	-37.4	0.038
< = -0.00301	3,001	10.0%	19.3%	17.0	0.003
< = 0	2,726	9.1%	28.5%	-34.1	0.012
< = Inf	16,773	55.9%	18.4%	23.0	0.028
Total	30,000	100.0%	22.1%	NA	0.081

```
## split Util_Growth_6mo into 4 groups: <= -.03 <= -.003, <= 0, > 0
df_T3$Util_Growth_6mo_below_minus.03 <- ifelse(df_T3$Util_Growth_6mo <= -.03,1,0)
df_T3$Util_Growth_6mo_below_minus.003 <- ifelse(df_T3$Util_Growth_6mo > -.03 &
                                                df_T3$Util_Growth_6mo <= -.003,1,0)
df_T3$Util_Growth_6mo_above_0 <- ifelse(df_T3$Util_Growth_6mo > 0,1,0)
```

Table XX: WOE Analysis of Max_Bill_Amt

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = 600	1,507	5.0%	31.7%	-49.2	0.014
< = 4079	2,994	10.0%	26.5%	-24.0	0.006
< = 18400.65	5,999	20.0%	21.5%	3.4	0.000
< = 21034	1,502	5.0%	26.9%	-25.9	0.004
< = 52496.15	7,498	25.0%	22.4%	-1.4	0.000
< = Inf	10,500	35.0%	19.0%	19.4	0.012
Total	30,000	100.0%	22.1%	NA	0.036

```
## split Max_Bill_Amt into 6 groups: <= 600, <= 4k, <= 18.4k, <=21k, <= 52k, > 52k
df_T3$Max_Bill_Amt_below_600 <- ifelse(df_T3$Max_Bill_Amt <= 600,1,0)
df_T3$Max_Bill_Amt_below_4k <- ifelse(df_T3$Max_Bill_Amt > 600 &
                                      df_T3$Max_Bill_Amt <= 4000,1,0)
df_T3$Max_Bill_Amt_below_18k <- ifelse(df_T3$Max_Bill_Amt > 4000 &
                                       df_T3$Max_Bill_Amt <=18400,1,0)
df_T3$Max_Bill_Amt_below_21k <- ifelse(df_T3$Max_Bill_Amt > 18400 &
                                       df_T3$Max_Bill_Amt <=21000,1,0)
df_T3$Max_Bill_Amt_above_52k <- ifelse(df_T3$Max_Bill_Amt > 52000,1,0)
```

Table XX: WOE Analysis of Max_Pmt_Amt

Final Bin	Total Count	Total Distr.	1.Rate	WOE	IV
< = 168	1,501	5.0%	37.8%	-76.0	0.035
< = 5000	14,002	46.7%	26.3%	-22.9	0.026
< = 36621.4	11,497	38.3%	17.8%	26.8	0.025
< = Inf	3,000	10.0%	11.1%	82.5	0.053
Total	30,000	100.0%	22.1%	NA	0.139

```
## split Max_Bill_Amt into 4 groups: <= 168, <= 5k, <= 36k, > 36k
df_T3$Max_Pmt_Amt_below_168 <- ifelse(df_T3$Max_Pmt_Amt <= 168,1,0)
df_T3$Max_Pmt_Amt_below_5k <- ifelse(df_T3$Max_Pmt_Amt > 168 &
                                     df_T3$Max_Pmt_Amt <= 5000,1,0)
df_T3$Max_Pmt_Amt_above_36k <- ifelse(df_T3$Max_Pmt_Amt > 36000,1,0)
```

X.5.? Summary of GLM Model 1b – reduced variables after transformation

Table XX: Model #1b

Dependent Variable	Estimate	Std. Error
Constant	-1.757***	0.13
SEX_FEMALE	-0.137**	0.04
MARRIAGE_Y	0.199***	0.04
PAY_X_Sum_6mo_belowZero	0.383***	0.10
PAY_X_Sum_6mo_aboveSix	0.910***	0.07
Avg_Pmt_Amt_below2k	0.317***	0.05
LIMIT_BAL_above_160k	-0.240***	0.05
Avg_Util_above_.45	0.405***	0.06

Avg_Pay_Ratio_above_.113	-0.170**	0.06
Util_Growth_6mo_above_0	-0.091*	0.05
Max_Bill_Amt_below_600	0.886***	0.12
Max_Bill_Amt_below_4k	0.515***	0.08
Max_Pmt_Amt_below_168	0.473***	0.11
Max_DLQ_above1	1.078***	0.09
Observations	15,180	
Log Likelihood	-6,841	
Akaike Inf. Crit.	13,710	
Note:	*p<0.1; **p<0.05; ***p<0.01	

X.5.? Comparison of exploratory GLM Model before & after transformation

Table xx: Model 1 Training Data Confusion Matrix

Target \ Predicted	0	1
0	0.78	0.38
1	0.22	0.62

Table xx: Model 1b Training Data Confusion Matrix

Target \ Predicted	0	1
0	0.73	0.26
1	0.31	0.68

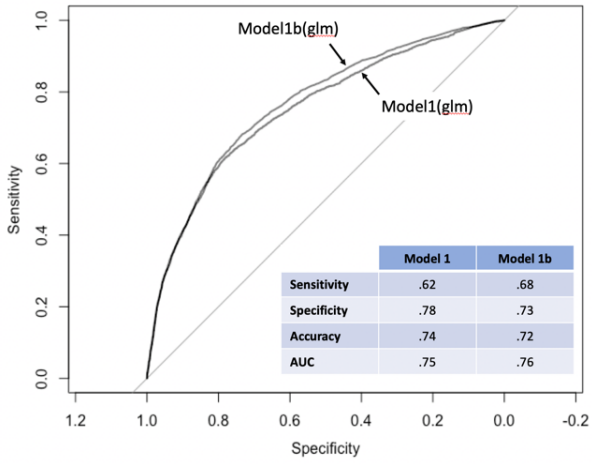
Table xx: Model Test Data Confusion Matrix

Target \ Predicted	0	1
0	0.77	0.23
1	0.37	0.63

Table xx: Model Test Data Confusion Matrix

Target \ Predicted	0	1
0	0.73	0.27
1	0.31	0.68

Figure XX: ROC curve for Model 1 & 1b on xTrain data



Observation: Model Sensitivity and AIC measures improve with due to selection of most highly relevant variables and additional binning techniques. Specificity and Accuracy drop, but if intention is to predict potential default, then recommendation would be to select model that more accurately predicts positive instances, which Model 1b does.