Model#101:Credit Card Default Model

Model Development Guide

Noé Flores

Northwestern University

Introduction:

The problem presented for us is to predict and detect the likelihood of individuals who are likely to default on their credit card payment. Credit card debt in the United States is roughly around one trillion dollars. Given those numbers, the impact on financial institutions from consumers defaulting on their credit card would be substantial. Our goal is to develop a predictive model that is effective at determining whether or not a client will default on their next monthly payment. We began the process by examining and preparing the data and introducing certain aspects of feature engineering to ensure that we are working with optimal and descriptive variables. We followed this process by performing an exploratory analysis of our data through the utilization of histograms, box-plots, and frequency tables set up to gather information on the variables we have available to incorporate into our analysis. Finally, we will build and compare four different classification models to determine which of them has the highest predictive value, interpretability, and practicality with respect to our goal of identifying credit card customers likely to default on their next bill. Results of this analysis should produce models we can present and objectively compare to arrive at the best solution to our credit card default problem.

2. The Data:

Our dataset contains 30,000 observations of twenty-three different explanatory variables and one response variable (DEFAULT). The variables in our dataset include a total of nine Categorical and fourteen Continuous data types. The table below provides a breakdown of those variables aggregated by their types.

Figure .1: Table of Variables by Type

Categorical Variables	Continuous Variables
SEX	AGE
EDUCATION	LIMIT_BAL
MARRIAGE	BILL_AMT1 - BILL_AMT6
PAY_0 - PAY_6	PAY_AMT1 - PAY_AMT6

2(b). Data Description:

The following table dictionary contains a detailed description of each variable. In the case of our PAY_0 through PAY_6 variables, we have categorical data with scaled levels signifying a different grade of payment history. For the purpose of clarity, the detailed description of that scale can be found in the data table below.

Figure.2: Data Dictionary and Description of Variables

Independent Variables	Description		
ID	User Identification		
LIMIT_BAL	Credit Limit (Dollar)		
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 (year)		
PAY_0	History of past payment(Scale -1 to 9)	Sep	2005
PAY_2	History of past payment (Scale -1 to 9)	Aug	2005
PAY_3	History of past payment (Scale -1 to 9)	Jul	2005
PAY_4	History of past payment(Scale -1 to 9)	Jun	2005
PAY_5	History of past payment(Scale -1 to 9)	May	2005
PAY_6	History of past payment(Scale -1 to 9)	Apr	2005
BILL_AMT1	Statement Balance (Dollar)	Sep	2005
BILL_AMT2	Statement Balance (Dollar)	Aug	2005
BILL_AMT3	Statement Balance(Dollar)	Jul	2005
BILL_AMT4	Statement Balance(Dollar)	Jun	2005
BILL_AMT5	Statement Balance (Dollar)	May	2005
BILL_AMT6	Statement Balance(Dollar)	Apr	2005
PAY_AMT1	Amount of Previous Payment(Dollar)	Sep	2005
PAY_AMT2	Amount of Previous Payment (Dollar)	Aug	2005
PAY_AMT3	Amount of Previous Payment (Dollar)	Jul	2005
PAY_AMT4	Amount of Previous Payment(Dollar)	Jun	2005
PAY_AMT5	Amount of Previous Payment(Dollar)	May	2005
PAY_AMT6	Amount of Previous Payment(Dollar)	Apr	2005
Dependent Variables	**Description**		
DEFAULT	Binary Response Variable (1 = Default in October , 0 = No Default)		

Figure.3: Scale for Payment Status

Scale for Payment Status (PAY_0 - PAY_6)	
- 1 = pay duly	5 = payment delay for five months
1 = payment delay for one month	6 = payment delay for six months
2 = payment delay for two months	7 = payment delay for seven months
3 = payment delay for three months	8 = payment delay for eight months
4 = payment delay for four months	9 = payment delay for nine months

2(c). Data Summary:

The summary statistics below serves to provide us with a quick and simple description of the data and include simple measures of central tendency which will allow us to check the quality of the data and observe any apparent discrepancies from the report in the data dictionary.

Figure.4: Data Variable Summary

variable	missing	complete	mean	sd	p0	p25	p50	p75	p100
AGE	0	30000	35	9	21	28	34	41	79
BILL_AMT1	0	30000	51223	73636	-165580	3559	22382	67091	964511
BILL_AMT2	0	30000	49179	71174	-69777	2985	21200	64006	983931
BILL_AMT3	0	30000	47013	69349	-157264	2666	20089	60165	1664089
BILL_AMT4	0	30000	43263	64333	-170000	2327	19052	54506	891586
BILL_AMT5	0	30000	40311	60797	-81334	1763	18105	50191	927171
BILL_AMT6	0	30000	38872	59554	-339603	1256	17071	49198	961664
EDUCATION	0	30000	2	1	0	1	2	2	6
LIMIT_BAL	0	30000	167484	129748	10000	50000	140000	240000	1000000
MARRIAGE	0	30000	2	1	0	1	2	2	3
SEX	0	30000	2	0	1	1	2	2	2
PAY_0	0	30000	0	1	-2	-1	0	0	8
PAY_2	0	30000	0	1	-2	-1	0	0	8
PAY_3	0	30000	0	1	-2	-1	0	0	8
PAY_4	0	30000	0	1	-2	-1	0	0	8
PAY_5	0	30000	0	1	-2	-1	0	0	8
PAY_6	0	30000	0	1	-2	-1	0	0	8
PAY_AMT1	0	30000	5664	16563	0	1000	2100	5006	873552
PAY_AMT2	0	30000	5921	23041	0	833	2009	5000	1684259
PAY_AMT3	0	30000	5226	17607	0	390	1800	4505	900000
PAY_AMT4	0	30000	4826	15666	0	296	1500	4013	621000
PAY_AMT5	0	30000	4799	15278	0	253	1500	4032	426529
PAY_AMT6	0	30000	5216	17777	0	118	1500	4000	528666

Referencing the information above, we can see that the values we are getting for PAY_0 - PAY_6 don't match with the description we were given by that data dictionary. On a basic level, we see that there is a type-o as PAY_0 should probably be labeled PAY_1 in order to be in uniform with the variables in the dataset. We also have some issues with EDUCATION, MARRIAGE, and the scale values for the history of past payment variables (PAY_0 - PAY_6) don't match what we see in the data dictionary. The variables are also in need of renaming and rebranding in order to produce data visuals that are easier to read, interpret, and manipulate in the future. On a positive note, we don't have any missing variables in this dataset to deal with or impute.

2(d). Data Cleaning and Preparation:

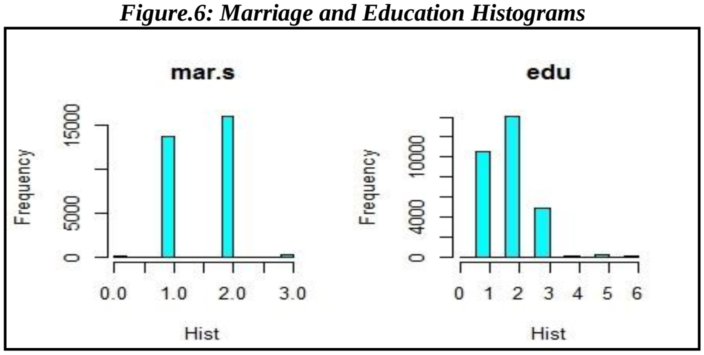
As I previously mentioned, we want to clean up the naming convention for our variables to produce cleaner and more interpretable data visuals which should improve the modeling process. In the table below, we can see the previous names our variables were given and the new names assigned to each variable in bold.

Figure.5: New Variable names

Rename and Rebrand variables for interpretability	
['LIMIT_BAL'] == **'c.limit'**	['BILL_AMT1'] == **'bill.amt1'**
['SEX'] == **'sex'**	['BILL_AMT2'] == **'bill.amt2'**
['EDUCATION'] == **'edu'**	['BILL_AMT3'] == **'bill.amt3'**
['MARRIAGE'] == **'mar.s'**	['BILL_AMT4'] == **'bill.amt4'**
['AGE'] == **'age'**	['BILL_AMT5'] == **'bill.amt5'**
['PAY_0'] == **'pay.1'**	['BILL_AMT6'] == **'bill.amt6'**
['PAY_2'] == **'pay.2'**	['PAY_AMT1'] == **'pay.amt1'**
['PAY_3'] == **'pay.3'**	['PAY_AMT2'] == **'pay.amt2'**
['PAY_4'] == **'pay.4'**	['PAY_AMT3'] == **'pay.amt3'**
['PAY_5'] == **'pay.5'**	['PAY_AMT4'] == **'pay.amt4'**
['PAY_6'] == **'pay.6'**	['PAY_AMT5'] == **'pay.amt5'**
['DEFAULT] == **'default'**	['PAY_AMT6'] == **'pay.amt6'**

According to our data dictionary, Education (edu), should have a range between one and four depending on the customer's education level. Marriage, (mar.s) should have values between one and three. The side-by-side histograms below show the outlying data points for both variables that we noticed in the data summary.

Figure.6: Marriage and Education Histograms



With respect to Marital Status (mar.s) we have values below one and equal to zero. Education (edu) has values that are greater than four and less than one. These numbers are inconsistent with our expectations for these two variables and need to be addressed before moving forward.

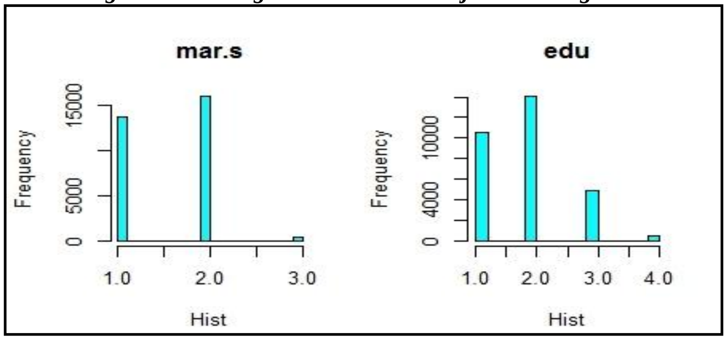
Specific examination of (edu) and (mar.s) in the data set produced results that informed us that there was a relatively small amount of outlying data points in each of the two variables. The most significant portion of outlying points was found in (edu) and compromised slightly more than one percent of the total observations, while (mar.s) only had a fractional percentage of outliers.

Figure.7: Marriage and Education Histograms

Outliers for Marriage and Education	
sum(credit_card_default$edu > 4)	= 331
sum(credit_card_default$edu < 1)	= 14
sum(credit_card_default$mar.s < 1)	= 54

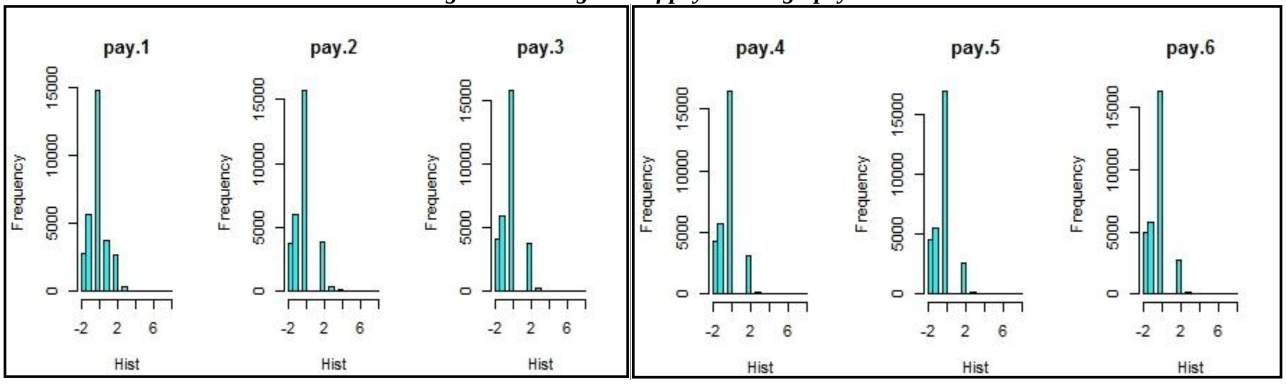
Given the relatively small amount of outlying data and erring on the conservative side, we mapped all the outlying data in both variables to be classified as "other." The new histograms are below.

Figure.8: Marriage and Education Adjusted Histograms



We can produce similar histograms for (pay.1) through (pay.6) to better understand why our values aren't matching the data dictionary and just how far off they are, and why they are misplaced.

Figure.9: Histograms of pay.1 through pay.6



The data dictionary states that the categorical values for (pay.1) through (pay.6) should have a range of -1 to 9. We can observe in the histograms that our values actually range from -2 to 8.

The frequency table below gives us a more detailed view of how the data is distributed and the concentration of the data points at each level.

Figure.10: Frequency Table for pay.1 to pay.6

Frequency Table										
table_pay.1										
-2	-1	0	1	2	3	4	5	6	7	8
2759	5686	14737	3688	2667	322	76	26	11	9	19
table_pay.2										
-2	-1	0	1	2	3	4	5	6	7	8
3782	6050	15730	28	3927	326	99	25	12	20	1
table_pay.3										
-2	-1	0	1	2	3	4	5	6	7	8
4085	5938	15764	4	3819	240	76	21	23	27	3
table_pay.4										
-2	-1	0	1	2	3	4	5	6	7	8
4348	5687	16455	2	3159	180	69	35	5	58	2
table_pay.5										
-2	-1	0	2	3	4	5	6	7	8	
4546	5539	16947	2626	178	84	17	4	58	1	
table_pay.6										
-2	-1	0	2	3	4	5	6	7	8	
4895	5740	16286	2766	184	49	13	19	46	2	

There are several issues here that fail to coincide with the scale we currently have for these variables. Our data dictionary for the history of past payment records (pay.1 to pay.6) has levels that should range from -1 to 9 for a total of 10 different classification bins. This isn't the case.

The issues with the payment status variables were troublesome to resolve, but it's to be expected when working with a large dataset such as this one. After careful manual examination of the data, we came to the following conclusions:

- Values of (-2) indicate that a payment was made on time and for the full balance or there wasn't a bill or payment that cycle. No credit was used.
- Values of (-1) indicate that a payment was made on time and for the full balance.
- Values of (0) indicate that a payment was made on time for partial balance payment.

There is a one month delay with credit cards as the consumer is typically paying the last months bill. The table below has some examples of the research we conducted color coded for clarity.

Figure.11:Examples of coresponding bill.amt, pay.amt and pay status.

ID	**pay.1**	**pay.2**	**pay.3**	**bill.amt2**	**pay.amt1**	**bill.amt3**	**pay.amt2**	**bill.amt4**	**pay.amt3**
30	0	0	0	16,575	1,500	17,496	1,500	17,907	1,000
31	-1	-1	-1	17,265	17,270	13,266	13,281	15,339	15,339
101	-2	-2	-2	10,212	10,212	850	850	415	415
46	-2	-2	-2	0	0	0	0	0	0

This portion of the data preparation also required us to use and make some judgment calls. We want to line up the values as best we can with the data dictionary. There are enough examples in the dataset such as the one we highlighted in *Figure.11* that we feel confident in the assumptions we are making. The payment status variables will be mapped as follows:

For payment status values (pay.1) through (pay.4):

- Values of (-2) will remain the same with the new definition from analysis.
- Values of (-1) will remain the same with the new definition from analysis.
- Values of (0) through (7) will remain the same with new definitions from analysis.
- Values of (8) will be mapped to a value of (7)

For payment status values (pay.5) and (pay.6):

- Values of (-2) will remain the same with the new definition from analysis.
- Values of (-1) will remain the same with the new definition from analysis.
- Values of (0) through (7) will remain the same with the new definitions from analysis.
- Values of (2) through (8) will be mapped as (1) through (7) respectively.

The frequency table for new payment status variables can be found below in Figure.12. This frequency table has some significant changes when compared to the original data dictionary and the scale for payment status. We are confident that the changes line up well with the hard data located in our data set.

Figure.12: New Frequency Table for pay.1 to pay.6

Frequency Table									
table_pay.1									
-2	-1	0	1	2	3	4	5	6	7
2759	5686	14737	3688	2667	322	76	26	11	28
table_pay.2									
-2	-1	0	1	2	3	4	5	6	7
3782	6050	15730	28	3927	326	99	25	12	21
table_pay.3									
-2	-1	0	1	2	3	4	5	6	7
4085	5938	15764	4	3819	240	76	21	23	30
table_pay.4									
-2	-1	0	1	2	3	4	5	6	7
4348	5687	16455	2	3159	180	69	35	5	60
table_pay.5									
-2	-1	0	1	2	3	4	5	6	7
4546	5539	16947	2626	178	84	17	4	58	1
table_pay.6									
-2	-1	0	1	2	3	4	5	6	7
4895	5740	16286	2766	184	49	13	19	46	2

With the changes that we made there was a need to update the payment scale. That information can be found below and reflects a new range for the future data analysis and models we will build.

Figure.13: Scale for Payment Status

New Scale for Payment Status (pay.1 - pay.6)	
- 2 = pay full amount or no usage	3 = payment delay for three months
- 1 = pay full amount	4 = payment delay for four months
0 = pay duly	5 = payment delay for five months
1 = payment delay for one months	6 = payment delay for six months
2 = payment delay for two months	7 = payment delay for seven months

We still have ten different categories for payment status, but now we have a better understanding of what those values in our dataset actually mean and represent. The changes we made to the description of the scale appear significant, but in reality, the essential values, which represent the bulk of delayed or on time payments remain the same.

2(e). Data Split:

The final step in our data preparation is to split our data into training, test, and validation datasets. Our models will be fit using the more extensive training data, and the independent test data will be used to determine how well our models perform and if any over-fitting is present. Finally, the validation set is used as a means for future tuning of the model and determining the models continued performance and usefulness. The table below gives a description of how we have split our data into those three distinct components for our modeling.

Figure.14: Frequency Table for PAY_1 to PAY_6

Data Split		
Training	Testing	Validation
15180	7323	7497

3. Feature Engineering:
Our goal is to provide the best possible predictive model for the likelihood of consumers defaulting on their credit card payments. In order to do that, we need to get the most out of the data that we have in this data set. Feature engineering is used to transform some of the information we have into features (variables) that will better help us understand the problem we are looking to solve. In our case, we want to predict credit card defaults, so we created the following four features to help us and hopefully provide us with improved inputs for our models.

1. Credit card utilization (**cr.usage**): This feature will allow us to track how much of their credit line each consumer is utilizing. The feature is calculated by taking the last know balance (bill.amt) and dividing it by the amount of the given credit limit (c.limit).

   *Formula = (**cr.usage**) = (bill.amt)/ (c.limit)*

2. Actual credit card utilization (**act.usage**): Takes the balance (bill.amt), subtracts the last payment amount(pay.amt) and divides the difference by the given credit card limit (c.limit). This should give us a different view of individuals who actually paid more or less on their last bill although it isn't using any information with respect to the most current bill.

   Formula **= *(act.usage)*** *= (bill.amt - pay.amt)/ (c.limit)*

3. Change in credit usage (**use.change**): This gives us an idea of how the specific user's credit utilization has changed from the beginning of the period we are tracking to the current level. This metric takes the difference between (cr.usage1) and (cr.usage6).

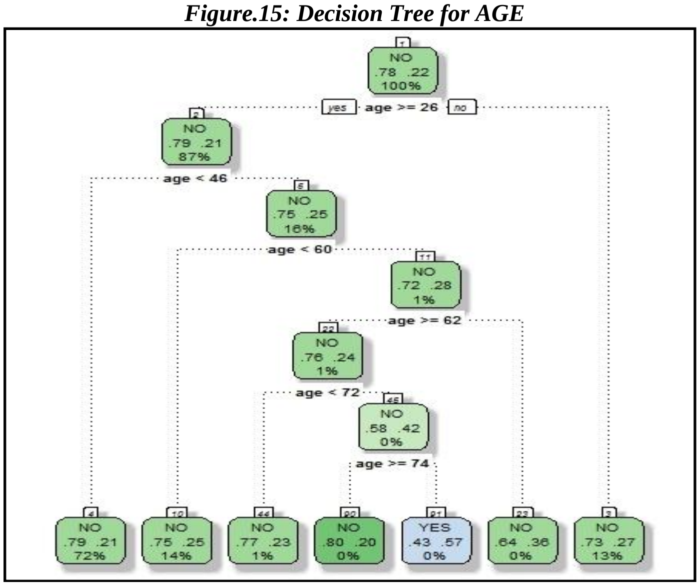
   Formula **= (use.change)** = (cr.usage1 - cr.usage6)

4. Payment ratio (**pay.ratio**): Takes the last months payment (pay.amt) and divides it by the last known balance (bill.amt). Recall that we are dealing with the previous month's bills and the features will line up down the line as follows (pay.amt1/bill.amt2)

   Formula **= (pay.ratio)** = (pay.amt1 / bill.amt2)

3(b). Age Bins
In an effort to get greater insight, we also decided to create age bins for the consumers in the data set. The new age bins should help us observe any significance between certain age groups with respect to the possibility of future credit card defaults. The age bins were created through the use of a decision tree regressing (default ~ age). The results of the decision tree are below.

Figure.15: Decision Tree for AGE



Based on the output from the decision tree, we decided to create the following five age bin variables for our analysis.

1. agegroup.1 **: Age** < = **"26"**
2. agegroup.2 **: Age** > = **"27"** <= **"46"**
3. agegroup.3 **: Age** > = **"47"** <= **"60"**
4. agegroup.4 **: Age** > = **"61"** <= **"72"**
5. agegroup.5 **: Age** > = **"73"**

Generally speaking, feature engineering comes down to the data scientist's knowledge of the domain. There are other possible features we could add to our analysis, but we feel confident that the five elements we generated will help inform us through our exploratory study and the modeling process that will follow.

4. Exploratory Data Analysis:

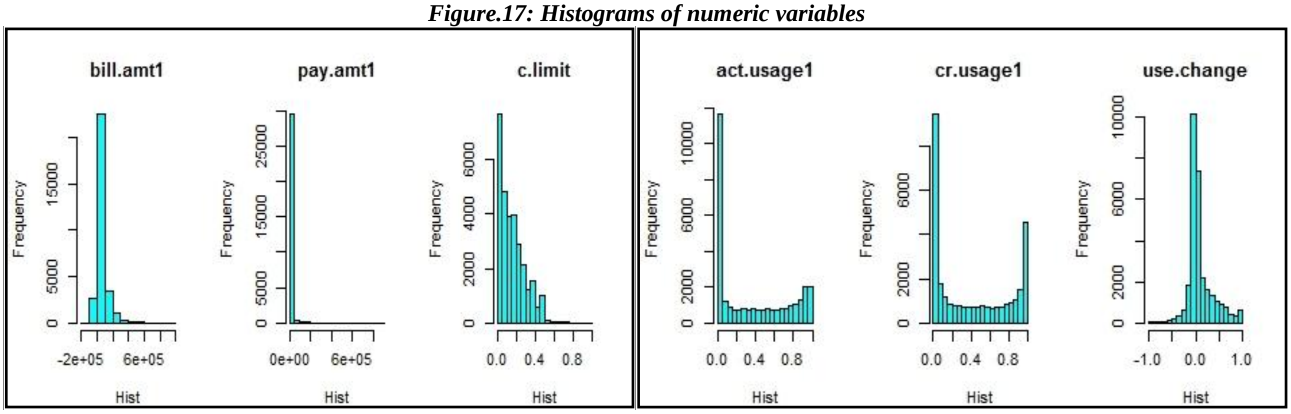
Now that our data is prepared, we can use exploratory data analysis to gain insight into the relationships present between our variables and the likelihood of a consumer defaulting on their credit card payment. We will start in a familiar place and look at the new data summary.

Figure.16: New Data Summary

Variable	Missing	Complete	mean	sd	p0	p25	p50	p75	p100
c.limit	0	30000	167484.32	129747.66	10000	50000	140000	240000	1000000
sex	0	30000	1.60	0.49	1	1	2	2	2
edu	0	30000	1.85	0.79	0	1	2	2	6
mar.s	0	30000	1.55	0.52	0	1	2	2	3
age	0	30000	35.49	9.22	21	28	34	41	79
pay.1	0	30000	-0.02	1.12	-2	-1	0	0	8
pay.2	0	30000	-0.13	1.20	-2	-1	0	0	8
pay.3	0	30000	-0.17	1.20	-2	-1	0	0	8
pay.4	0	30000	-0.22	1.17	-2	-1	0	0	8
pay.5	0	30000	-0.27	1.13	-2	-1	0	0	8
pay.6	0	30000	-0.29	1.15	-2	-1	0	0	8
bill.amt1	0	30000	51223.33	73635.86	-165580	3558.75	22381.5	67091	964511
bill.amt2	0	30000	49179.08	71173.77	-69777	2984.75	21200	64006.25	983931
bill.amt3	0	30000	47013.15	69349.39	-157264	2666.25	20088.5	60164.75	1664089
bill.amt4	0	30000	43262.95	64332.86	-170000	2326.75	19052	54506	891586
bill.amt5	0	30000	40311.4	60797.16	-81334	1763	18104.5	50190.5	927171
bill.amt6	0	30000	38871.76	59554.11	-339603	1256	17071	49198.25	961664
pay.amt1	0	30000	5663.58	16563.28	0	1000	2100	5006	873552
pay.amt2	0	30000	5921.16	23040.87	0	833	2009	5000	1684259
pay.amt3	0	30000	5225.68	17606.96	0	390	1800	4505	900000
pay.amt4	0	30000	4826.08	15666.16	0	296	1500	4013.25	621000
pay.amt5	0	30000	4799.39	15278.31	0	252.5	1500	4031.5	426529
pay.amt6	0	30000	5215.50	17777.47	0	117.75	1500	4000	528666
act.usage1	0	30000	0.36	0.37	0	4.9e-05	0.24	0.74	1
act.usage2	0	30000	0.35	0.36	0	2.9e-05	0.22	0.69	1
act.usage3	0	30000	0.32	0.35	0	1.4e-05	0.19	0.62	1
act.usage4	0	30000	0.3	0.33	0	0	0.16	0.56	1
act.usage5	0	30000	0.29	0.33	0	0	0.14	0.54	1
cr.usage1	0	30000	0.41	0.39	0	0.022	0.31	0.83	1
cr.usage2	0	30000	0.4	0.38	0	0.019	0.3	0.81	1
cr.usage3	0	30000	0.39	0.37	0	0.017	0.27	0.76	1
cr.usage4	0	30000	0.36	0.36	0	0.015	0.24	0.67	1
cr.usage5	0	30000	0.33	0.34	0	0.012	0.21	0.6	1
cr.usage6	0	30000	0.32	0.34	0	0.0086	0.19	0.58	1
pay.ratio1	0	30000	0.43	1.21	0	0.045	0.1	1	1
pay.ratio2	0	30000	0.45	1.34	0	0.045	0.1	1	1
pay.ratio3	0	30000	0.44	1.17	0	0.038	0.084	1	1
pay.ratio4	0	30000	0.44	1.1	0	0.036	0.077	1	1
pay.ratio5	0	30000	0.48	1.68	0	0.038	0.09	1	1
use.change	0	30000	0.098	0.27	-1	-0.028	0.0053	0.17	1

What we can quickly gather from our new data summary is that some of our numeric values don't appear to be normally distributed. While confirmation of this can be obtained from an examination of histograms, we can also take a look at the mean and compare it to the min (p0) and max(p100) points for a particular variable. We will take a closer look at a few variables just be sure.

Figure.17: Histograms of numeric variables



The histograms of (use.change) and (bill.amt1) appear to be the only variables with what seems to be reasonably normal distributions albeit reasonably narrow in both cases. The other variables in Figure.17 all have distributions which lack normality, be it through right sided or positive skewness or considerable front and rear loadings. We are going to standardize all the numeric predictor variables in our data set. Standardization should serve to help introduce stability to the models we will build. The normalization for each variable was implemented using the formula below.

Figure.18:

Standardization Formula
*(bill.amt1 - **mean**(bill.amt1)) / **standard deviation**(bill.amt1)*

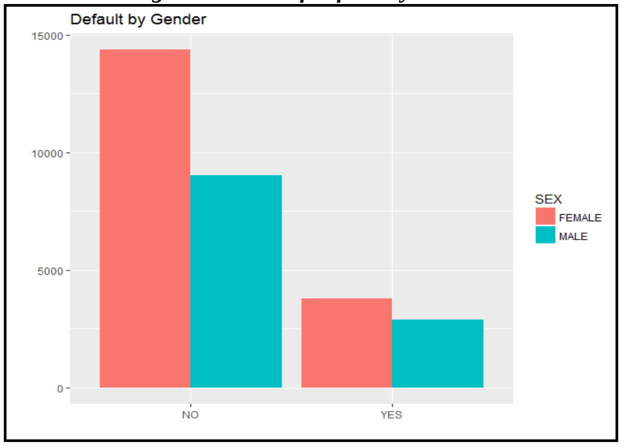
Now that we are fully satisfied with our data, we can proceed with a traditional exploratory data analysis. We will begin at the highest level, an examination of the split in our data between defaults and non-defaults.

Figure.19: Frequency Table for Default

DEFAULT	**NO**	**YES**
	23,364	6,636

There are a total of 6,636 defaults present in our dataset, which amounts to roughly 22% percent of the total observations. We will now look at how those defaults are related to some of our variables, starting with the most basic, which is gender.

Figure.20: Plot of default by Gender



When we look at the bar-plots above, it would appear that women seem more likely to default on their credit card obligations than men would. That isn't giving us the whole picture.

Figure.21: Frequency table of defaults by gender.

	NO	YES	Default Percent of gender	Default Percent of total
Female	14349	3763	26%	13%
Male	9015	2873	32%	10%

The frequency table is able to provide a more detailed view of what the actual amount of defaults with respect to gender really are. Women represent a slightly more significant share than the men, but in terms of percentage, it isn't a huge number.

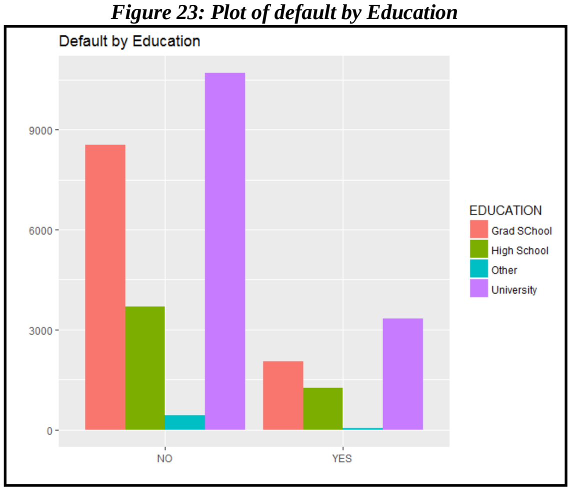
The consumer's education should provide another window with a view of who would be most likely to default. Among defaults, the university students are more prone to default, but they also make up the most prominent group. The frequency table and the bar-plot of the group are below.

Figure. 22: Frequency table of default by education

DEFAULT	NO	YES	Default Percentage of Group	Default Percentage of Total
Grad School	8549	2036	19%	7%
High School	3680	1237	34%	4%
Other	435	33	8%	0%
University	10700	3330	31%	11%

High school education level customers have the most significant in-group default percentage but a small percent of the total defaults in our data. University and Graduate schools customers have similar in group numbers, but they also account for the two most substantial default percentages. The plot is below.

Figure 23: Plot of default by Education

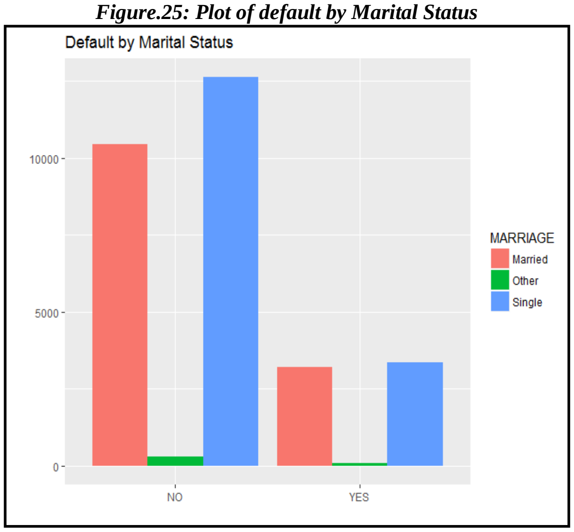


Marital status is another qualitative variable in our data that could present some information. In this case, we see that the data is pretty evenly split in terms of default between Married and Single

Figure .24: Frequency table by education and percentage

DEFAULT	NO	YES	Default Percentage of Group	Default Percentage of Total
Married	10453	3206	31%	11%
Other	288	89	31%	.01%
Single	12623	3341	26%	11%

Figure.25: Plot of default by Marital Status

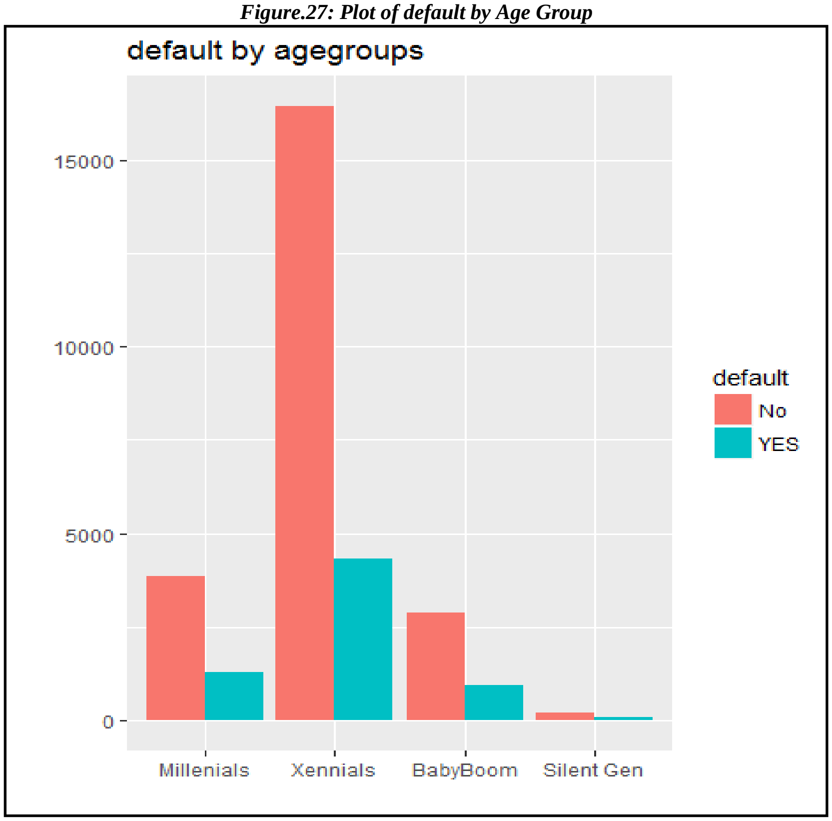


Age should be able to provide more clues for us to examine. We were able to create age bins during our feature engineering process. Our age bins coincided with some commonly used terms to describe different generations. We are able to label those age bins we created earlier as either Millennials, Xennials, Baby Boomers, and Silent Generation.

Figure.26: Frequency table of default by Age Group

DEFAULT	NO	YES	Default Percentage of Group	Default Percentage of Total
Millenials	3842	1285	33%	4%
Xennials	16444	4330	26%	14%
Baby Boomer	2879	948	33%	3%
Silent Gen	199	73	37%	0%

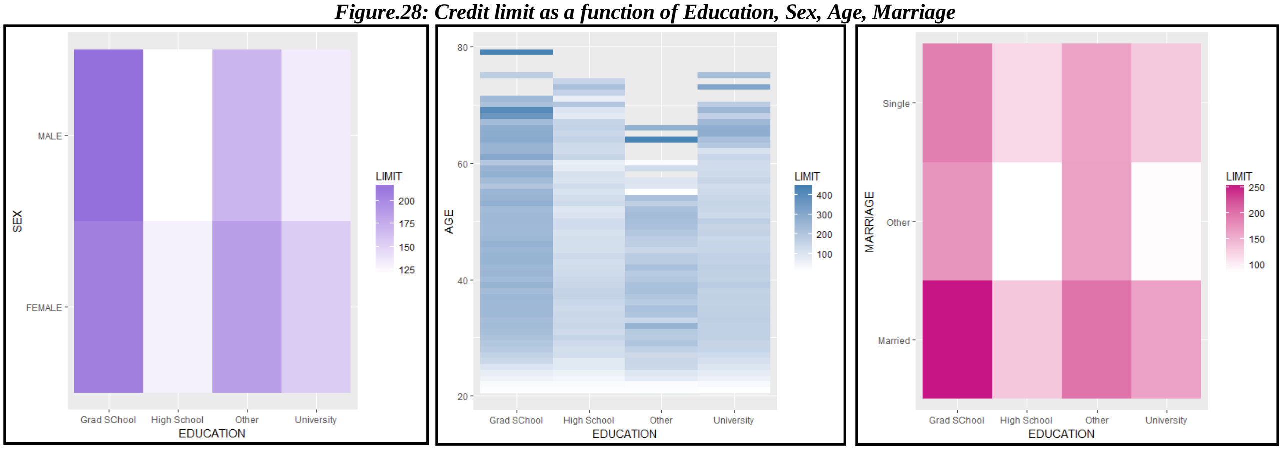
Figure.27: Plot of default by Age Group



The Xennials age group, which according to the bins we created would include individuals who are older than twenty-six years of age and up to forty-six years of age, compromise the most abundant group and therefore have the most default totals. Careful examination would point out that as a percentage, the Millennial and Baby Boomers groups definitely struggle more with paying on time.

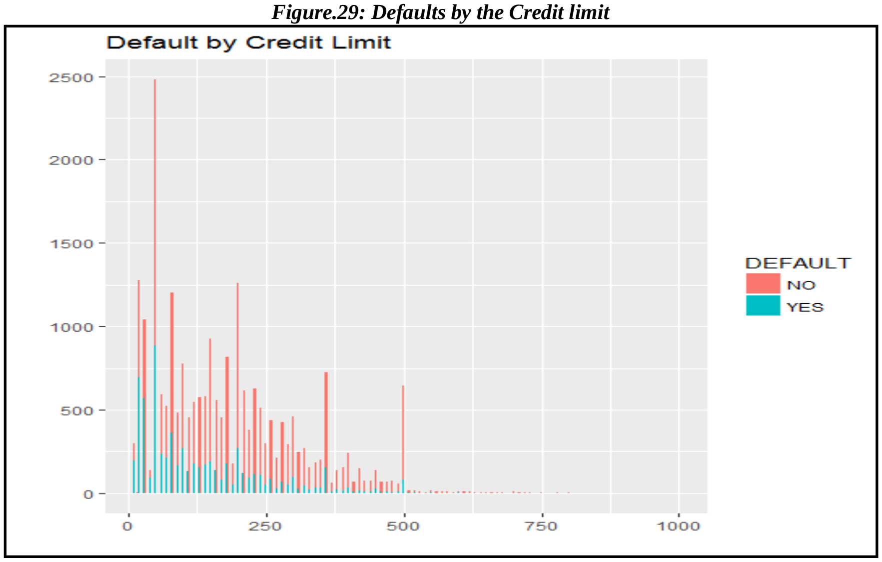
As a final look at these three qualitative variables we just covered, we produce heat maps to compare how those three intersect with one another and credit limit. In the plots below, we are looking at the intersection of credit limit with sex, age, and education. From the maps below, it appears that a consumer who is married has a graduate degree, and is above what we would consider middle-aged, strikes the sweet spot for credit limit.

Figure.28: Credit limit as a function of Education, Sex, Age, Marriage



We can also look at the defaults in our data in terms of the credit limit. The bulk of the default appears to be concentrated to individuals with lower credit limits. That isn't too surprising, given that exceptionally high credit limits aren't typically observed and individuals with lower limits might be new consumers or those deemed unqualified or too risky for upper credit limits.

Figure.29: Defaults by the Credit limit



In order to get a sense of the relationship between our dependent variable (default) and our independent variables, we constructed a correlation table. The results below don't appear to indicate anything extraordinary, but they are indicative of the need for a model outside the realm of linearity.

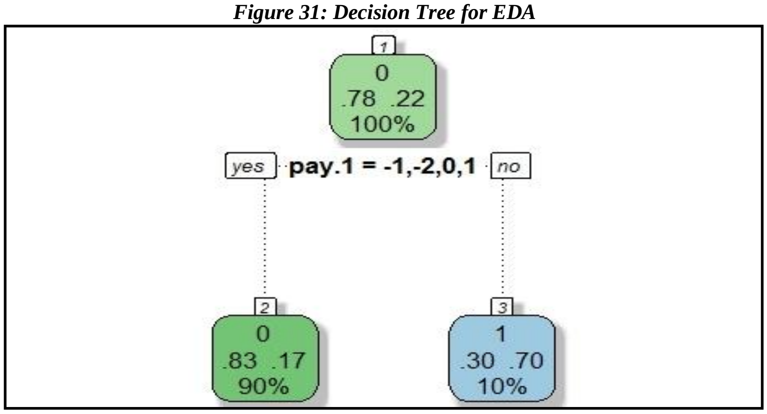
Figure.30: Correlation Table

Correlation							
default	**1**						
pay.1	0.326	act.usage2	0.123	bill.amt5	-0.007	pay.amt3	-0.056
pay.2	0.264	cr.usage5	0.122	bill.amt4	-0.010	pay.amt4	-0.057
pay.3	0.235	cr.usage4	0.120	bill.amt3	-0.014	pay.amt2	-0.059
pay.4	0.217	act.usage1	0.112	bill.amt2	-0.014	pay.amt1	-0.073
pay.5	0.162	cr.usage3	0.109	bill.amt1	-0.020	pay.ratio4	-0.085
pay.6	0.148	cr.usage2	0.102	mar.s	-0.028	pay.ratio5	-0.086
act.usage5	0.134	cr.usage1	0.089	use.change	-0.028	pay.ratio3	-0.098
act.usage4	0.131	edu	0.034	sex	-0.040	pay.ratio2	-0.103
act.usage3	0.129	age	0.014	pay.amt6	-0.053	pay.ratio1	-0.104
cr.usage6	0.126	bill.amt6	-0.005	pay.amt5	-0.055	c.limit	-0.154

4.B. Model-Based Exploratory Data Analysis:

Following the correlation table, we fit a decision tree to our data in order to see which variables appear to give us the best overall view of the likelihood of default. After pruning the tree, which means we removed sections of the tree that provided little classification power, we got the following results.

Figure 31: Decision Tree for EDA



The decision tree appears to zero in on one particular variable,(pay.1) which relates to the standing of the credit card holders payment history. The tree is clear as to which breaking points are significant for that payment status classification variable (-2,-1,0,1).

Let's examine the tree analysis a bit more. The specificity, sensitivity, and accuracy metrics below are based on the training dataset.

Figure .32: Decision Tree for EDA Specificity and Sensitivity

		No	Yes		
Decision	No	11299	2317	**Accuracy**	81.72%
Tree	Yes	458	1106	**Sensitivity**	32.31%
				Specificity	96.10%

In a confusion matrix, the values on the diagonal of the matrix represent individuals whose default statuses were predicted correctly. The off-diagonal elements represent individuals that were misclassified. In our example, the decision tree made incorrect predictions for 458 who did not default, and for 2,317 customers who did happen to default. We also measure how accurate the classification model is. In this case, we were 81.72% accurate with this particular model.

The frequency table of (pay.1) gives us the breakdown between defaults and non-defaults for this particular payment status.

Figure.33: Frequency Table for pay.1

Default Scale	**No**	**Yes**
-2	2394	365
-1	4732	954
0	12849	1888
1	2436	1252
2	823	1844
3	78	244
4	24	52
5	13	13
6	5	6
7	10	8

If we compare the frequency table above to the scale for payment status, it becomes clear why those values, in particular, appear to be a real breaking point for default classification.

Figure.34: Scale for Payment Status

New Scale for Payment Status (pay.1 - pay.6)	
- 2 = pay full amount or no usage	3 = payment delay for three months
- 1 = pay full amount	4 = payment delay for four months
0 = pay duly	5 = payment delay for five months
1 = payment delay for one months	6 = payment delay for six months
2 = payment delay for two months	7 = payment delay for seven months

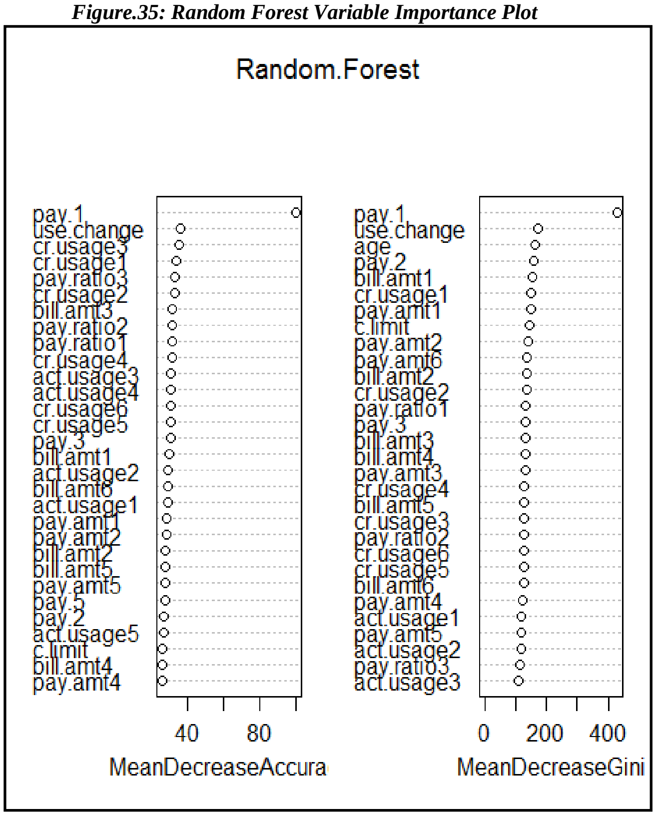
5. Predictive Modeling:

The goal is to fit four distinct models which are suitable for binary classification problems such as the one we have for predicting credit card default. In an effort to be thorough, we will fit four different classification models. We won't devote a lot of time to explanations of the modeling itself, but rather to an overview of each model, with the primary focus on the results each one produces. As you know, our dataset was broken down into training, testing, and validation subsets. The modeling was performed on the training data and tested on the test set.

5.A. Random Forest:

We start our process by building on the decision tree and implementing a Random Forest model which serves as a subset of the Decision Tree we performed during the exploratory phase.

Figure.35: Random Forest Variable Importance Plot



- The Random Forest outputs a variable of importance feature. This feature provides a view of the predictors that drive the model. Once again we see (pay.1) and (use.change) driving the model although a more detailed look is needed.

- The Random Forest Gini feature of importance provides us with another measuring of relevance. Interestingly enough, some of the variables traded places but nothing vastly different.

This feature provides a more detailed view of the predictors that drove the Random Forest model.

You can see just how much influence (pay.1) has on the performance of our model.

Figure.35: Random Forest Variable Importance

Random Forest Variable importance				
	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
c.limit	24.917	-6.063	26.022	143.214
sex	2.218	-2.616	0.477	23.237
edu	1.709	0.862	2.012	52.457
mar.s	9.049	-4.237	5.769	31.514
age	13.806	-0.215	12.417	163.895
pay.1	85.174	34.969	100.342	430.366
pay.2	22.514	11.098	26.960	159.854
pay.3	27.339	4.692	30.127	132.641
pay.4	20.893	10.092	25.355	96.722
pay.5	21.953	8.940	27.480	92.045
pay.6	16.107	7.350	20.830	65.445
bill.amt1	25.415	-2.333	29.420	153.428
bill.amt2	25.880	-8.640	27.662	136.241
bill.amt3	29.095	-10.221	31.540	130.745
bill.amt4	23.203	-5.588	25.973	129.896
bill.amt5	24.066	-5.051	27.632	127.415
bill.amt6	26.236	-5.208	28.872	125.467
pay.amt1	26.192	-5.601	28.296	148.210
pay.amt2	25.627	-2.139	27.811	138.977
pay.amt3	21.569	-0.543	25.747	129.665
pay.amt4	23.314	-5.324	25.903	122.050
pay.amt5	25.675	-6.739	27.567	119.178
pay.amt6	19.617	6.948	24.065	137.175
cr.usage1	29.579	-5.145	33.788	149.485
cr.usage2	30.962	-11.688	32.431	135.550
cr.usage3	34.107	-14.783	35.456	127.156
cr.usage4	28.133	-5.659	31.081	127.921
cr.usage5	28.118	-8.629	30.241	126.088
cr.usage6	28.741	-10.621	30.306	126.284
act.usage1	26.305	-7.918	28.590	119.663
act.usage2	26.984	-4.100	28.916	117.322
act.usage3	28.858	-9.017	30.766	107.831
act.usage4	28.503	-8.988	30.359	103.443
act.usage5	25.435	-8.482	26.355	103.823
use.change	32.090	-3.184	35.651	170.177
pay.ratio1	28.886	-1.975	31.382	132.919
pay.ratio2	27.167	-0.990	31.448	126.854
pay.ratio3	28.843	0.522	32.620	115.896
pay.ratio4	22.697	-4.970	23.227	97.546
pay.ratio5	21.179	0.590	25.179	96.902
agegroup.1	5.365	-0.670	4.906	17.232
agegroup.2	5.576	-3.662	3.191	22.971
agegroup.3	4.177	-6.224	0.584	16.888
agegroup.4	5.164	3.263	6.806	6.040
agegroup.5	0.000	0.000	0.000	0.084

There is a significant improvement in the confusion matrix for the Random Forest Model when compared to the decision tree, but it looks like it may have been over-fitted. We need to run this model on the test data to really gauge how the model will perform as a classifier.

Figure.36: Random Forest Specificity and Sensitivity

Train		No	Yes	**Accuracy**	99.24%
Random	No	11750	109	**Sensitivity**	96.82%
Forest	Yes	7	3314	**Specificity**	99.94%

Figure.37: Random Forest Specificity and Sensitivity Test Set

Test		No	Yes	**Accuracy**	82.00%
Random	No	5428	983	**Sensitivity**	36.87%
Forest	Yes	338	574	**Specificity**	94.14%

Those results fall more in line with what we observed in the Decision Tree. We can now check the model's error rate as a final metric for comparison.

Figure.38: Random Forest Training and Test set Error Rates

Model	**Mean Error**
Decision Tree Mean Error (Train)	0.007
Decision Tree Mean Error (Test)	0.180

There is some explicit over fitting present in the Random Forest model's performance on the training data. This is brought back in line by the validation process on the test data, and the mean error rate there is reasonable.

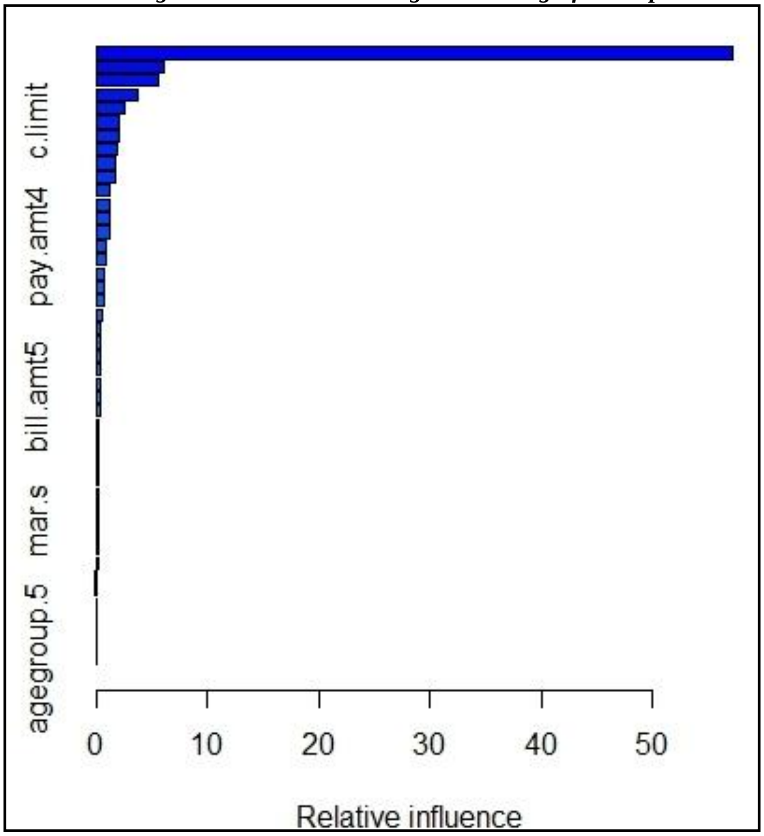
5.B. Gradient Boosting:

Boosting is another method we can use to improve upon our Random Forest Models. It has three tuning parameters. The number of trees, selected through cross-validation, the shrinkage parameter, and the number of splits in each tree. The variable's significance and plot are shown below.

Figure 39: Gradient Boosting Variable Significance

Gradient Boosting Variable significance	
var	rel.inf
pay.1	57.258
pay.3	6.121
pay.2	5.694
pay.5	3.768
pay.amt1	2.612
c.limit	2.091
pay.4	2.085
pay.6	1.931
pay.amt3	1.730
pay.amt2	1.720
use.change	1.309
bill.amt1	1.280
cr.usage1	1.270
pay.amt6	1.189
pay.amt4	0.936
pay.ratio3	0.912
act.usage2	0.782
act.usage1	0.772
cr.usage2	0.766
pay.ratio2	0.494
pay.ratio1	0.432
bill.amt3	0.418
pay.amt5	0.394
age	0.380
pay.ratio5	0.330
bill.amt5	0.326
bill.amt2	0.323
cr.usage5	0.275
cr.usage6	0.273
act.usage3	0.254
bill.amt4	0.230
act.usage5	0.217
act.usage4	0.215
cr.usage3	0.207
mar.s	0.203
cr.usage4	0.197
pay.ratio4	0.168
edu	0.157
bill.amt6	0.134
sex	0.121
agegroup.4	0.015
agegroup.2	0.012
agegroup.3	0.001
agegroup.1	0.000
agegroup.5	0.000

Figure.40:Gradient Boosting Variable Significance plot



The gradient boosting provides us with an intriguing set of variables which influence the model's predictive ability. Still, the further we proceed with this analysis, the more we see that (pay.1) is one of, if not the primary predictor variable in our study.

Figure.40: Gradient Boosting Specificity and Sensitivity

Train		No	Yes	**Accuracy**	82.11%
Gradient	No	11219	2170	**Sensitivity**	36.61%
Boosting	Yes	538	1253	**Specificity**	95.42%

Once again, we see some improvements over the results we achieved using the random forest model. There isn't too much discrepancy between the training and test set confusion matrices for our boosting model.

Figure.41: Gradient Boosting Specificity and Sensitivity Test-Set

Test		No	Yes	**Accuracy**	82.74%
Gradient	No	5473	971	**Sensitivity**	37.64%
Boosting	Yes	293	586	**Specificity**	94.92%

Our test accuracy is excellent with this model, and it doesn't differ too much between the training data and the test data. We could expect that this model would perform pretty well if it was used in a real-world application.

Figure.42: Random Forest Training and Test set Error Rates

Model	**Mean Error**
Gradient Boosting Mean Error (Train)	0.179
Gradient Boosting Mean Error (Test)	0.173

5.C. Logistic Regression With Variable Selection:

Logistic models are not typically built for classification problems, but they represent a powerful tool that is flexible to use. The first logistic model we will run will be based on variables we selected from the results obtained through the Random Forest Model and Gradient Boosting modeling. Those variables and the results of the base logistic model are below.

Figure .43: Logistic Regression Full

Logistic regression Full					
Coefficients:					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.453	0.120	-12.071	< 2e-16	***
c.limit	-0.231	0.037	-6.288	0.000	***
sex2	-0.147	0.043	-3.400	0.001	***
edu2	-0.059	0.050	-1.190	0.234	
edu3	-0.089	0.067	-1.320	0.187	
edu4	-0.986	0.244	-4.036	0.000	***
mar.s2	-0.172	0.049	-3.509	0.000	***
mar.s3	-0.317	0.185	-1.720	0.086	.
age	0.006	0.003	2.434	0.015	*
pay.1	0.589	0.025	23.805	< 2e-16	***
pay.2	0.086	0.029	2.986	0.003	**
pay.3	0.137	0.031	4.349	0.000	***
pay.4	0.140	0.035	4.043	0.000	***
pay.5	-0.145	0.052	-2.766	0.006	**
pay.6	-0.117	0.044	-2.663	0.008	**
bill.amt1	0.197	0.133	1.479	0.139	
bill.amt2	-0.145	0.169	-0.856	0.392	
bill.amt3	0.175	0.150	1.169	0.242	
bill.amt4	-0.167	0.143	-1.167	0.243	
bill.amt5	0.050	0.139	0.360	0.719	
bill.amt6	-0.019	0.098	-0.194	0.846	
pay.amt1	-0.223	0.055	-4.036	0.000	***
pay.amt2	-0.196	0.066	-2.984	0.003	**
pay.amt3	-0.002	0.041	-0.041	0.967	
pay.amt4	-0.034	0.035	-0.981	0.327	
pay.amt5	-0.081	0.042	-1.951	0.051	.
pay.amt6	-0.068	0.033	-2.069	0.039	*
cr.usage1	-0.512	0.104	-4.917	0.000	***
cr.usage2	0.226	0.124	1.817	0.069	.
cr.usage3	-0.040	0.106	-0.379	0.705	
cr.usage4	0.149	0.096	1.562	0.118	
cr.usage5	0.029	0.071	0.411	0.681	

After running the full logistic model, we can move forward with an automated variable selection process. In this case, we will use stepwise variable selection to obtain the optimal logistic model.

Figure.44: Logistic Regression Stepwise selection results

Stepwise Model Path

Initial Model:
default ~ c.limit + sex + edu + mar.s + age + pay.1 + pay.2 + pay.3 + pay.5 + bill.amt1 + pay.amt1 + pay.amt2 + pay.amt3 + pay.amt4 + pay.amt5 + pay.amt6 + cr.usage1 + cr.usage2 + cr.usage3 + cr.usage4 + cr.usage5 + cr.usage6 + act.usage1 + act.usage2 + pay.ratio1 + pay.ratio2 + pay.ratio3 + pay.ratio4 + pay.ratio5

Final Model:
default ~ c.limit + sex + edu + mar.s + age + pay.1 + pay.2 + pay.3 + pay.5 + bill.amt1 + pay.amt1 + pay.amt2 + pay.amt4 + pay.amt5 + pay.amt6 + cr.usage1 + cr.usage3 + cr.usage6 + act.usage1 + act.usage2 + pay.ratio1 + pay.ratio2 + pay.ratio4 +pay.ratio5

The results in Figure.44 show the full model we started with, which includes all the variables we selected based on the Decision Tree and Gradient Boosting. It also shows the final model based on a stepwise variable selection process we implemented. We can see the coefficients below.

Figure.45: Logistic Regression Stepwise selection coefficients

Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.318	0.122	-10.838	< 2e-16	***
c.limit	-0.249	0.038	-6.614	0.000	***
sex2	-0.143	0.044	-3.269	0.001	**
edu2	-0.013	0.050	-0.262	0.794	
edu3	-0.049	0.068	-0.720	0.472	
edu4	-0.950	0.245	-3.878	0.000	***
mar.s2	-0.157	0.050	-3.161	0.002	**
mar.s3	-0.293	0.187	-1.564	0.118	
age	0.005	0.003	2.040	0.041	*
pay.1	0.564	0.025	22.930	< 2e-16	***
pay.2	0.169	0.033	5.098	0.000	***
pay.3	0.239	0.033	7.339	0.000	***
pay.5	0.123	0.043	2.863	0.004	**
bill.amt1	0.168	0.039	4.351	0.000	***
pay.amt1	-0.211	0.065	-3.230	0.001	**
pay.amt2	-0.116	0.067	-1.739	0.082	.
pay.amt3	-0.041	0.042	-0.979	0.327	
pay.amt4	-0.050	0.036	-1.364	0.173	
pay.amt5	-0.146	0.044	-3.304	0.001	***
pay.amt6	-0.071	0.033	-2.129	0.033	*
cr.usage1	-0.462	0.091	-5.106	0.000	***
cr.usage2	-0.084	0.160	-0.524	0.600	
cr.usage3	-0.295	0.146	-2.016	0.044	*
cr.usage4	0.014	0.076	0.177	0.859	
cr.usage5	-0.065	0.082	-0.798	0.425	
cr.usage6	0.166	0.067	2.464	0.014	*
act.usage1	0.350	0.160	2.190	0.029	*
act.usage2	0.514	0.144	3.558	0.000	***
pay.ratio1	0.297	0.057	5.228	0.000	***
pay.ratio2	0.221	0.060	3.675	0.000	***
pay.ratio3	-0.008	0.050	-0.168	0.867	
pay.ratio4	0.136	0.049	2.759	0.006	**
pay.ratio5	0.141	0.043	3.243	0.001	**

We can now check the confusion matrix and error rates for the final version of the logistic model.

Figure.46: Logistic Regression Specificity and Sensitivity

Train		No	Yes	**Accuracy**	80.60%
Logistic	No	11302	2493	**Sensitivity**	27.17%
Regression	Yes	455	930	**Specificity**	96.13%

Figure.47: Logistic Regression Specificity and Sensitivity Test-Set

Test		No	Yes	**Accuracy**	81.43%
Logistic	No	5573	1167	**Sensitivity**	25.05%
Regression	Yes	193	390	**Specificity**	96.65%

There is a slight decline present in the confusion matrix, based on the training data when compared to the previous two models. Both the training and test datasets produced similar confusion matrix results.

Figure.48: Logistic Regression Training and Test set Error Rates

Model	**Mean Error**
Logistic Regression Mean Error (Train)	0.194
Logistic Regression Mean Error (Test)	0.186

Our test accuracy is satisfactory with this model as well. While there is a slight decline if we compare it to the Boosting model, the results are more than reasonable. This too could prove to be a useful model in production.

5.D. Linear Discriminate Analysis(LDA):

Linear Discriminate Analysis (LDA) is closely related to logistic regression in that they both work well with classification problems. However, LDA typically works better with larger classification problems, and therefore it might represent a more stable alternative to Logistic Regression.

Figure.49: Linear Discriminate Analysis

Linear Discriminate Analysis		
Prior probabilities of groups:		
	0	1
	0.787	0.213
Coefficients		
		LD1
pay.1		0.652
act.usage2		0.641
pay.2		0.389
pay.ratio1		0.375
act.usage1		0.347
pay.ratio2		0.176
cr.usage6		0.147
pay.ratio4		0.127
pay.3		0.107
edu3		0.061
pay.ratio5		0.057
pay.5		0.032
pay.amt2		0.007
edu2		0.007
age		0.002
pay.amt6		-0.009
bill.amt1		-0.019
pay.amt5		-0.045
pay.amt4		-0.065
sex2		-0.102
pay.amt1		-0.115
mar.s2		-0.160
c.limit		-0.167
mar.s3		-0.226
cr.usage3		-0.476
cr.usage1		-0.534
edu4		-0.727

The Prior probabilities of the group are the ones that already exist in your training data. In this case, 78.7% of our training data corresponds to default risk evaluated as (0), and 23.3% of our training data corresponds to default risk assessed as (1). The second thing you see is the group means, which are the average of each predictor within each class. These values suggest that (pay.1), (act.usage2), and (pay.2) have the most significant influence on predicting future defaults.

We can now check the LDA confusion matrix and error rates across the training and test sets.

Figure.50: LDA Specificity and Sensitivity

Train		No	Yes	**Accuracy**	80.64%
Random	No	11299	2484	**Sensitivity**	27.52%
Forest	Yes	458	942	**Specificity**	96.01%

Figure.51: LDA Specificity and Sensitivity Test-Set

Test		No	Yes	**Accuracy**	81.70%
Random	No	5541	1115	**Sensitivity**	28.39%
Forest	Yes	225	442	**Specificity**	96.10%

The results of the LDA are a slight improvement over the ones observed in the Logistic regression models but nothing that one would call eye-catching. In terms of the error rates, almost identical results between LDA and logistic models.

Figure.52: Logistic Regression Training and Test set Error Rates

Model	**Mean Error**
LDA Mean Error (Train)	0.194
LDA Mean Error (Test)	0.183

6.Comparison of results:

Our final model selection criteria will be based on the error rates in the Training and Test set, and we will also consider the results from the confusion matrices. The table below provides a comparison of the classification models and their mean prediction error rates.

Figure.53: Comparison of Model Errors

Model	Train Error	Test Error	Test Sensitivity	Test Specificity	Test Accuracy
Gradient Boosting	***0.179***	***0.173***	***37.64%***	94.92%	**82.74%**
Random Forest	0.007	0.180	36.87%	94.14%	82.00%
Linear Discriminate	0.194	0.183	28.40%	96.10%	81.70%
Logistic Regression	0.194	0.186	25.05%	96.65%	81.43%

Looking at the table above, the Gradient Boosting Model performed better than the others with respect to the model's error rates across the training and test sets. The model also exhibited the best test set sensitivity, accuracy, and the specificity was within striking distance of the other models we utilized. This model also provides an output that is easily interpreted. It is our recommendation that we should proceed with the Boosting model for our prediction of future customer defaults.

7.Conclusion:

This analysis was designed to develop a predictive model to improve the detection of consumers who were most likely to default on their next credit card payment. We fit four different classification models to find the best performer in terms of predictive accuracy, based on the model's true positive rate (sensitivity), true negative rate (specificity), and the overall accuracy of the model. In this analysis, we used Random Forest, Gradient Boosting, Logistic Regression and Linear Discriminate Analysis as our classification modeling techniques. The classification model which incorporated boosting was the best performer for predicting future defaults, with the lowest error rate and the highest sensitivity and accuracy. In all of these models, it was clear that the payment status (pay._) of the consumer was either the main or one of the main determinates in identifying those most likely to default on their next payment. While the error rate on our best model was low, the sensitivity was also lower than we would like from a classification model. The quality of our results weren't as strong as we would have wanted to observe, but there is a chance that this could be related to the values we amended within the data set to align with the given data dictionary. There is room to incorporate different variable transformations and expand the use of feature engineer to produce more dynamic variables that can assist in building our models. It might be beneficial to experiment with different classification modeling techniques as well, such as KNN, Structured Vector Machines, and Ridge Regression. Even without the use of those modeling techniques, it's clear that we should create and build many models and maintain the set criteria for selection when the analysis is complete. On a final note, we learned that credit card modeling is a tough task. Even though we aren't completely satisfied with the results, this drawback is merely a limitation of the modeling techniques we incorporated during this specific analysis, and it is something that we can improve in the future with different data preparation and modeling methods.