Model #101:
Credit Card Default Model

Tannia Dubon
MSDS 498

1. INTRODUCTION

This project addresses a binary classification problem using the "default of credit card clients" data set from the UCI Machine Learning Repository. The data contained demographic information as well as the payment history of 30,000 individuals, from April to September 2005. Using this information, five different types of models were developed: Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine and Principal Component Analysis.

Features were engineered from the original variables, such as average utility, average bill amount, maximum utility, maximum delinquency, etc., to boost predictive capacity.

An Extreme Gradient Boosting model was trained using variables identified as important by the results of a Random Forest model that assessed the complete suite of variables, and it produced the best results. It had a Type I Error of .16, an AUC of .76, and a Sensitivity of .68. Overall these results were satisfactory, depending on the specific use of the model (exploratory or to release into production) and the acceptable accuracy threshold for Type I and Type II errors.

Considering that the data used was limited to demographic and 6 month payment history for 30,000 individuals, these results are acceptable and they help to inform the business of the risks they would undertake if they would use the highest performing model.

In the future, it would be beneficial to use additional information to help improve prediction capacity. In particular, bureau data, application data and longer payment histories are typically used in the industry. Another strategy that should be considered is to populate the binned variables using the weights of evidence to allow for a more nuanced approach to capturing the differences between the records. Given the wide ranges encompassed by the variables, stratifying the groups by relevant categories, such as credit limit, may also prove beneficial in capturing and working with the distributions of each group.

2. MODEL DEVELOPMENT

2.1 DESCRIPTION OF THE DATA

The data contain observations for 30,000 credit card customers in Taiwan. They include 23 variables for characteristics of the customer, such as SEX, EDUCATION, MARRIAGE, and various credit and payment attributes, such as BILL_AMT and PMT_AMT. There are no missing entries.

The predictor variable is binary; 1 is equal to default and 0 is equal to non-default.

2.2 DATA DICTIONARY

Table 1: Data Dictionary for the Raw Variables

	Variable Name	Type	Description
1	ID	Continuous	Customer identification number – not a prediction variable
2	LIMIT_BAL	Continuous	Credit given in New Taiwan (NT) dollars, to individual and his/her family
3	SEX	Nominal	Gender. Coded as 1 = male, 2 = female
4	EDUCATION	Ordinal	Education. Coded as 1 = graduate school, 2 = university, 3 = high school, 4 = others.
5	MARRIAGE	Nominal	Marital status. Coded as 1 = married, 2 = single, 3 = others
6	AGE	Continuous	Age in years
7	PAY_1	Ordinal	September 2005 payment record. Coded as follows: -1= duly paid, 1=payment delay for one month, 2= payment delay for two months⋯8=payment delay for 8 months, 9= payment delay for 9 months or more
8	PAY_2	Ordinal	August 2005 payment record. Coded as follows: -1= duly paid, 1=payment delay for one month, 2= payment delay for two months⋯8=payment delay for 8 months, 9= payment delay for 9 months or more
9	PAY_3	Ordinal	July 2005 payment record. Coded as follows: -1= duly paid, 1=payment delay for one month, 2= payment delay for two months⋯8=payment delay for 8 months, 9= payment delay for 9 months or more
10	PAY_4	Ordinal	June 2005 payment record. Coded as follows: -1= duly paid, 1=payment delay for one month, 2= payment delay for two months⋯8=payment delay for 8 months, 9= payment delay for 9 months or more
11	PAY_5	Ordinal	May 2005 payment record. Coded as follows: -1= duly paid, 1=payment delay for one month, 2= payment delay for two months⋯8=payment delay for 8 months, 9= payment delay for 9 months or more
12	PAY_6	Ordinal	April 2005 payment record. Coded as follows: -1= duly paid, 1=payment delay for one month, 2= payment delay for two months⋯8=payment delay for 8 months, 9= payment delay for 9 months or more
13	BILL_AMT1	Continuous	Amount (NT) of statement billed in September 2005.
14	BILL_AMT2	Continuous	Amount (NT) of statement billed in August 2005.

Table 1: Data Dictionary for the Raw Variables

	Variable Name	Type	Description
15	BILL_AMT3	Continuous	Amount (NT) of statement billed in July 2005.
16	BILL_AMT4	Continuous	Amount (NT) of statement billed in June 2005.
17	BILL_AMT5	Continuous	Amount (NT) of statement billed in May 2005.
18	BILL_AMT6	Continuous	Amount (NT) of statement billed in April 2005.
19	PAY_AMT1	Continuous	Amount (NT) paid in September 2005.
20	PAY_AMT2	Continuous	Amount (NT) paid in August 2005.
21	PAY_AMT3	Continuous	Amount (NT) paid in July 2005.
22	PAY_AMT4	Continuous	Amount (NT) paid in June 2005.
23	PAY_AMT5	Continuous	Amount (NT) paid in May 2005.
24	PAY_AMT6	Continuous	Amount (NT) paid in April 2005.
25	DEFAULT	Nominal	Dependent variable; default history. 1 = Yes, 0 = No

2.3 DATA QUALITY CHECK

Several issues were identified during the course of the data quality check. Multiple discrepancies are present between the data and the data dictionary, thus making it difficult to interpret the undefined values observed. These are described in the subsections that follow.

Most of the individuals reflected in the data set have a university to graduate level education (24,615 out of 30,000). Also, a significant class imbalance was not observed; the distribution of classes for the variables SEX (60% female) and MARRIAGE (53% single) were relatively evenly distributed.

2.3.a PAY_1 – PAY_6

The variables PAY_1 – PAY_6 were problematic. More than half of the records were coded as 0 and -2, and these levels were not captured in the data dictionary. Further, the use of these codes was inconsistent.

For example, PAY_1 for records 24, 34, 35, and 2671, 2688 was coded as -2. See Table 2 below. The amounts payed (Pmt_Amt_1) were more than or equal to the amounts billed (BILL_AMT_2). For records 10, and 46, PAY_1 was coded as -2 but the billed and paid amounts were zero. Record 66 shows an instance where PAY_1 was coded as -2 even there was no payment towards a billed amount of 148,751.

Table 2: Highlighted Records Coded -2

ID	PAY_1	PAY_2	BILL_AMT2	PAY_AMT1
10	-2	-2	0	0
24	-2	-2	19420	19428
34	-2	-2	4152	4152
35	-2	-2	5006	5006
46	-2	-2	0	0
53	-2	-2	7867	7875
66	-2	-2	148751	0
2671	-2	-2	6305	6336
2688	-2	-1	367979	368199

It appears that the records with payment amounts less than the amount billed, but not equal to 0, were coded as 0. For instance, see records 3, 4, 6, 8, 9 listed in Table 3 below. However, exceptions in the use of this code level have been observed. PAY_1, record 5 was coded as -1; the amount paid was less than the amount billed. PAY_1, record 7 was coded as 0; in this case the amount paid was more than the amount billed. Given these observations, -2 and 0 were recoded to -1. But in practice, it would be necessary to confirm the business rule for determining whether a loan is considered delayed when a partial payment is made.

Table 3: Highlighted Records Coded 0

ID	PAY_1	PAY_2	BILL_AMT2	PAY_AMT1
3	0	0	14027	1518
4	0	0	48233	2000
5	-1	0	5670	2000
6	0	0	57069	2500
7	0	0	412023	55000
8	0	-1	380	380
9	0	0	14096	3329

2.3.b EDUCATION

This variable contained the values 0, 5, 6 which were not contained in the data dictionary. The total entries containing these values were 345, or 1% of the data. As such, the observations containing these values were recoded to 4: Others.

2.3.c MARRIAGE

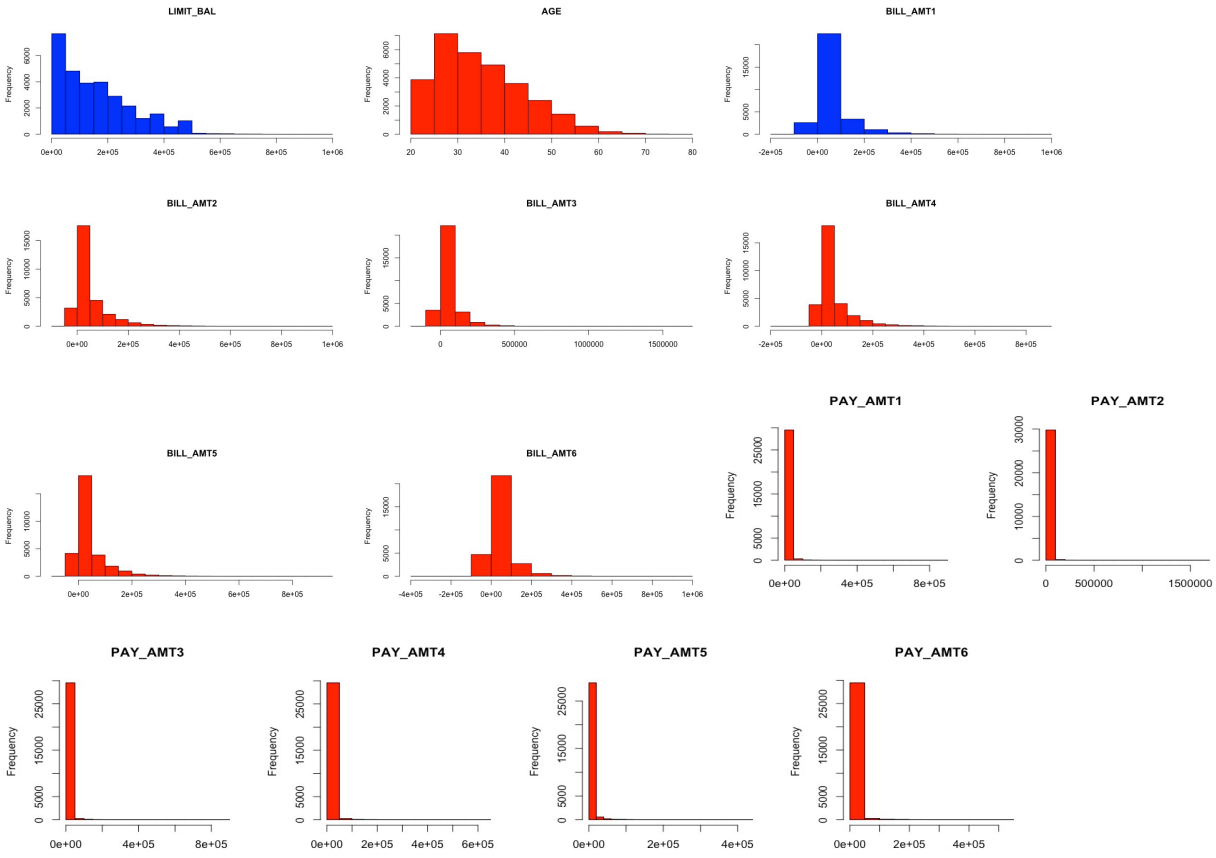
The variable MARRIAGE contained 54 entries with the value 0, which was not in the data dictionary. These values were recoded to 3: Others.

2.3.d CONTINUOUS VARIABLES

The raw variables had very wide and skewed distributions. See the thumbnails for the corresponding variables below. Summary statistics are also available to illustrate this in Index 1. As a result, this limits the statistical assumptions that we can make. Given that there are many outliers in each variable, and median and quartile information will better inform us of the dispersion. Also, transformations or feature engineering will be an important way of addressing some of the irregularities that we are seeing.

Lastly, variables BILL_AMT1-6 contained large negative values. These may be valid entries as some individuals may have issued large overpayments, or they may be errors in the data.

Image 1: Histograms of the continuous variables



2.3.e DATA SPLIT

The data set containing 30,000 records was divided into training, testing and validation sets. The counts of these are reflected in Table 4.

Table 4: Count of Observations by Split

	Total	%
Training	15,180	51%
Testing	7,323	24.41%
Validation	7,497	24.99%

3. FEATURE ENGINEERING

A total of 12 features were engineered. These are summarized in Table 5 below.

Table 5: Summary Statistics for Default of Credit Card Clients Data

Limit_Bal	Woe binning was used to determine the intervals: (0, 30000] (30000,160000] (160000, 1000000]
Age	Woe binning was used to determine the intervals: 0-24, 25-32, 33-80.
Avg_Bill_Amt	Average of Bill_Amt1 through Bill_Amt6. Woe binning was used to determine intervals: (-Inf, 189] (189, 2847] (2847, 7489] (7489, 31916] (31916, 900000]
Avg_Pmt_Amt	Average of Pay_Amt1 through Pay_Amt6. Woe binning was used to determine intervals: (-Inf, 28323] (2833, 12093] (12093, 627345]
Pmt_Ratio1	Ratio formula is Pay_Amt1/Bill_Amt2. If payment and bill variables equal zero, the ratio will equal 100. This was done to distinguish from the scenario where a payment was made to a bill with a zero balance due. If payment variable equals zero but bill variable does not equal zero, then ratio will equal zero. If payment variable does not equal zero but bill variable does, then ratio equals one given that a payment was made although a balance was not due.
Pmt_Ratio2	Ratio formula is Pay_Amt2/Bill_Amt3. Same conditions detailed above.
Pmt_Ratio3	Ratio formula is Pay_Amt3/Bill_Amt4. Same conditions detailed above.
Pmt_Ratio4	Ratio formula is Pay_Amt4/Bill_Amt5. Same conditions detailed above.
Pmt_Ratio5	Ratio formula is Pay_Amt5/Bill_Amt6. Same conditions detailed above.
Avg_Pmt_Ratio	Average of Pmt_Ratio1 through Pmt_Ratio5. Woe binning was used to determine intervals: (-Inf, 0.03406552305] (0.03406552305, 0.1583070591] (0.1583070591, 1.000163684] (1.000163684, 1.17911605] (1.17911605, 2688]
Util1	Utilization formula is Bill_Amt1/Limit_Bal. If Bill_Amt equals zero, then Util variable will equal zero as this scenario indicates no utilization of credit limit.
Util2	Bill_Amt2/Limit_Bal. Condition same as above.
Util3	Bill_Amt3/Limit_Bal. Condition same as above.
Util4	Bill_Amt4/Limit_Bal. Condition same as above.
Util5	Bill_Amt5/Limit_Bal. Condition same as above.
Util6	Bill_Amt6/Limit_Bal. Condition same as above.

Table 5: Summary Statistics for Default of Credit Card Clients Data

Avg_Util	Average of Util1 through Util6. Woe binning was used to determine intervals: (-Inf, 0.0009562517806] (0.0009562517806, 0.009194422492] (0.009194422492, 0.3673418095] (0.3673418095, 0.8231566667] (0.8231566667, 6]
Bal_Chx1	Bill_Amt2 - Bill_Amt1. This calculation captures the change from one month to the next.
Bal_Chx2	Bill_Amt3 - Bill_Amt2
Bal_Chx3	Bill_Amt4 - Bill_Amt3
Bal_Chx4	Bill_Amt5 - Bill_Amt4
Bal_Chx5	Bill_Amt6 - Bill_Amt5
Bal_Growth_6mo	Sum of Bal_Chx1 through Bal_Chx5. This calculation aggregates the changes throughout the six month period. Woe binning was used to determine intervals: (-Inf, -123] (-123, 21389.95] (21389.95, 428792]
Util_Chx1	Util2 - Util1. This calculation captures the change from one month to the next.
Util_Chx2	Util3 - Util2
Util_Chx3	Util4 - Util3
Util_Chx4	Util5 - Util4
Util_Chx5	Util6 - Util5
Util_Growth_6mo	Sum of Util_Chx1 through Util_Chx5. This calculation aggregates the changes throughout the six month period. Woe binning was used to determine intervals: (-Inf,-0.72262] (-0.72262,-0.00075] (-0.00075,0] (0,0.02930625] (0.02930625, Inf]
Max_Bill_Amt	Maximum amount billed for Bill_Amt1 through Bill_Amt6
Max_Pmt_Amt	Maximum amount paid for Pay_Amt1 through Pay_Amt6
Max_DLQ	Maximum amount paid for Pay_1 through Pay_6. Given the results from variable importance tests, Pay_1 through Pay_6 were kept in the modeling suite and Max_DLQ was left out.
Max_Util	Maximum utility, take from Util1 through Util6

4. EXPLORATORY DATA ANALYSIS

Exploratory data analysis was performed for the variables that will be used for modeling. The training data set was used for this section. The statistics for the key engineered variables are captured below.

Note that the summary statistics for the complete list of the engineered variables are contained in Index 2.

4.a TRADITIONAL EDA

Table 6: Quantile, Skew and Kurtosis Information for Key Engineered Variables

	0.00	0.25	0.50	0.75	0.95	0.97	0.99	1.00	Skew	Kurtosis
Variables included in modeling suite as binned										
Avg_Bill_Amt	-56043	4789	21198	56880	174643	209747	302473	592432	3	13
Avg_Pmt_Amt	0	1112	2389	5554	19416	26862	45205	627344	19	955
Avg_Pmt_Ratio	-605.46	0.05	0.16	1.00	80.20	100.00	100.00	2687.00	31.50	2297.84
Avg_Util	-0.23	0.03	0.29	0.69	0.96	0.99	1.06	5.36	0.72	4.86
Variables included in modeling suite										
Bal_Growth_6mo	-497231	-20094	-919	2962	21390	34793	84220	399983	-2	24
Util_Growth_6mo	-5.31	-0.18	-0.01	0.03	0.20	0.29	0.56	1.83	-1.71	14.32
Max_Bill_Amt	-2900	10051	31588	79120	218117	270420	374259	823540	2.54	11.67
Max_Pmt_Amt	0	2196	5000	12201	72074	100647	177000	1215471	7.29	123.45

As noted previously, the dispersion of the numerical variables is very wide. Table 4 provides more clarity on the extent of the dispersion and the leaps between quantiles. Many outliers were identified for each variable. Avg_Bill_Amt more than triples from KRW 56,880 in the 75th percentile to KRW 174,643 in the 95th percentile. The graph of this variable reflects a markedly asymmetrical distribution, with a positive skew of 3, kurtosis of 13, and heavy tails. The boxplot depicts numerous outliers. All the other variables in Table 4 reflect a similar asymmetry and non-normal distributions.

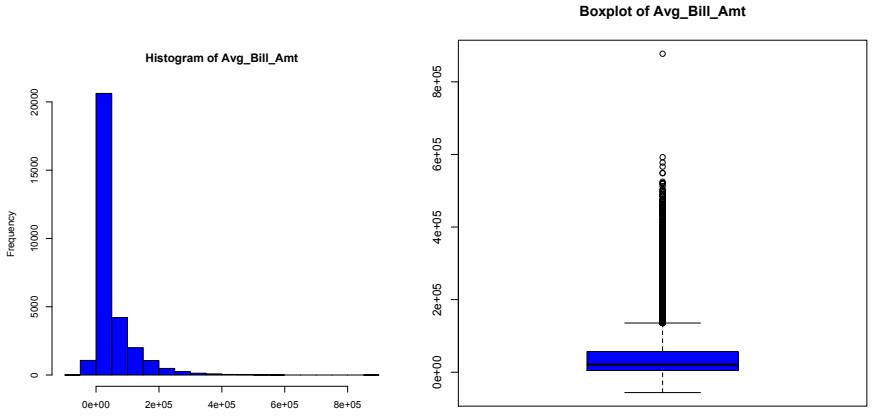


Image 2: Plots of Avg_Bill_Amt

The Avg_Pmt_Amt for the 50th percentile is up to KRW 2,389. This figure not only increases by more than double between the 50th and the 75th percentiles, it nearly quadruples between the 75th and the 95th percentiles. This suggests that we may be able to stratify segments of the sample population based on their ability to pay.

It is interesting to note that the Avg_Util jumps from .29 in the 50th percentile to .69 in the 75th percentile. This means that between those two percentiles, we see the utility more than double. Also, we can infer that the utility of the 50% individuals only use up to 29% of their credit limit, which is relatively low. This variable is among those that are binned. This is done to take a closer look at the interactions between the segmented Avg_Util and other variables.

The large negative numbers observed for variables Avg_Bill_Amt, Avg_Pmt_Ratio, and Bal_Growth_6mo are difficult to rationalize. For example, the quantiles for Avg_Pmt_Ratio indicate that 50% of the sample population have paid only up to 16% of their balance. But somewhere between the 50th and the 75th percentiles, the payment amounts are equivalent to the billed amounts. The negative values and extremely large positive values in this variable have a great impact on the distribution.

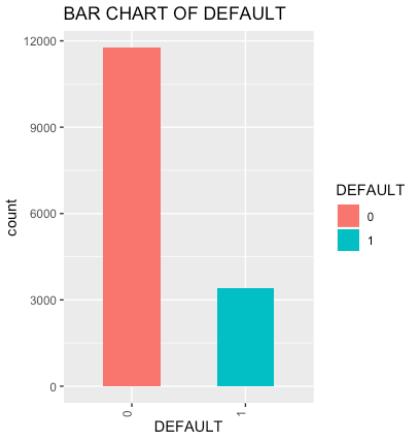


Image 3: DEFAULT BAR CHART

The target variable DEFAULT reflects a default percentage of 22.5%, and an imbalance between the classes.

The results of binning the AGE variable are reported in Table 7. Based on the count of the training data, we see that 50% of the population is 33 to 80 years old. This is a wide range considering the other half of the sample population represents 21-32 year olds. That's only an eleven year difference.

Finally, the correlation matrix captured in Index 5 reflects only a strong correlation between Max_Bill_Amt and Avg_Bill_Amt and between Max_Pmt_Amt and Avg_Pmt_Amt. This raises the question, just how much influence are the large bill and payment values having on the average. In the future, it may be good approach to calculate the Cook's D measure for these variables and to remap the extreme outliers.

Table 7: Age Bins

Age Range	Count	Counts of Default = 1	% Default of Total Training	% Default of Total in Bin Group
0_24	1,357	379	28%	2%
25_32	5,398	1097	20%	7%
33_80	8,425	1947	23%	13%

4.b. MODEL BASED EDA

In this section we consider the results of running a decision tree, the One Rule Machine Learning Algorithm OneR, and a simple logistic regression model. We do this while keeping in mind that we want to reduce our type II errors, the specificity, while maintaining an understanding of the business's tolerance for type 1 errors as these would represent an opportunity loss, or individuals who do not receive loan approvals, but whom don't actually default.

4.b.1 RPART DECISION TREE

Three decision trees were created. The first was built using all of the variables in the modeling suite, including PAY_1-6 and Max_DLQ. The second was built using all of the variables except PAY_1-6. The third was built using only those variables that were identified as significant by the simple logistic regression model:

Formula 2a =
SEX + EDUCATION + MARRIAGE +
PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
LIMIT_BAL_Neg_29999 + LIMIT_BAL_30000_159999 +
Avg_Bill_Amt_Neg_56050_188 + Avg_Bill_Amt_189_2846 + Avg_Bill_Amt_2847_7488 + Avg_Bill_Amt_7489_31915 +
Avg_Pmt_Amt_0_2832 + Avg_Pmt_Amt_2833_12092 + Max_Bill_Amt + Max_DLQ

The first tree captures the importance of PAY_1: 10% of individuals do not make their first payment. Of these, 71% will default. This is an important finding, as missing the first payment may be a precursor to defaulting on a loan. The output indicates that only PAY_1, PAY_3, PAY_4, PAY_5, PAY_6, and Max_DLQ were used to build the tree.

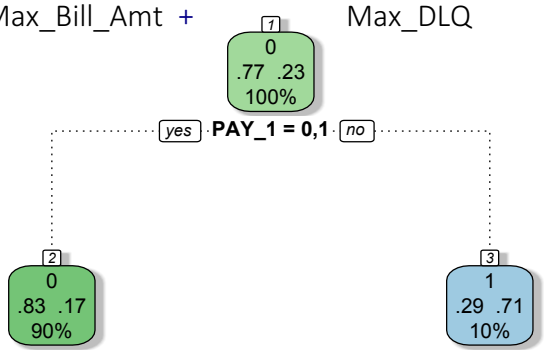


Image 4: Tree from First Model Containing All Variables

The second tree appears in Image 5. This tree indicates that of those with a maximum delinquency of more than or equal to 2.5 months, 65% of them defaulted on their loan. The results also indicate that only Max_DLQ and Avg_Pmt_Ratio_Neg16429_pt02 were used to build the model.

The tree that was output from the third model was the same as the tree for the first, as depicted in Image 4. The AUC remained at .64 on the training and testing data for the first and last models, the Type I error rate (.16) and the Sensitivity (.67) did not vary either.

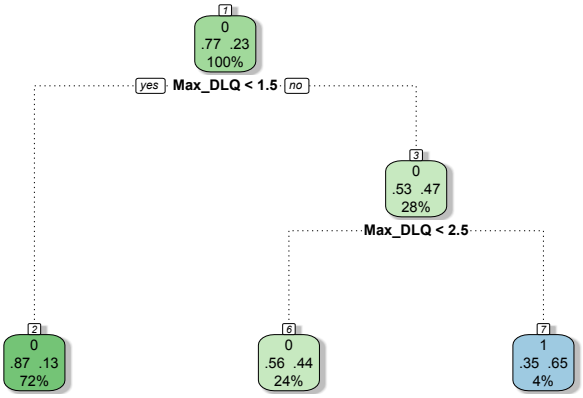


Image 5: Tree from Model Containing All Variables Except PAY

The second model that did not contain the PAY variables performed the worst (test data, AUC=.55, Type I Error = .20, Sensitivity=.60).

4.b.2 OneR

A model was built using only the variables that were considered important by the simple logistic regression model, formula 2a listed in the previous section. The results output for this model are listed in Table 8. The performance metrics indicate that this model had the same scores as the OneR model that was built using all the variables, including Pay_1:5. It is worth noting that it performed better than the OneR model that was built using all the variables, excluding Pay_1:5.

The output suggests a pattern has been identified where individuals that are behind on their payments two months or more, are likely to default, except those that are 5 months behind. To understand why the output rules above indicate DEFAULT = 0 for PAY_1 = 5, let's review Table 9 which shows the default/non-default for this level.

There is a 50/50 split between those that defaulted and did not default, despite being 5 months late on their payments. This is not very predictive; the training data set similarly reflects a 60/40 split. The likelihood ratios reflect the same: the probability that a client will default give that he/she paid their September bill 5 months late is .0017528.

The chart depicted in Image 6 reflects just how close to 50/50 default/non-default that level "delay of 5 months" really is. The other levels do not appear to be as close to this split.

Table 8: OneR Output

	Attribute	Accuracy
1	*PAY_1	81.73%
2	PAY_2	79.19%
3	PAY_5	78.81%
4	Max_DLQ	78.62%
5	PAY_4	78.56%
6	PAY_3	78.52%
7	SEX	77.45%
7	EDUCATION	77.45%
7	MARRIAGE	77.45%
7	LIMIT_BAL_Neg_29999	77.45%
7	LIMIT_BAL_30000_159999	77.45%
7	Avg_Bill_Amt_Neg_56050_188	77.45%
7	Avg_Bill_Amt_189_2846	77.45%
7	Avg_Bill_Amt_2847_7488	77.45%
7	Avg_Bill_Amt_7489_31915	77.45%
7	Avg_Pmt_Amt_0_2832	77.45%
7	Avg_Pmt_Amt_2833_12092	77.45%
7	Max_Bill_Amt	77.45%

Rules:

If PAY_1 = -1 then DEFAULT = 0
If PAY_1 = 1 then DEFAULT = 0
If PAY_1 = 2 then DEFAULT = 1
If PAY_1 = 3 then DEFAULT = 1
If PAY_1 = 4 then DEFAULT = 1
If PAY_1 = 5 then DEFAULT = 0
If PAY_1 = 6 then DEFAULT = 1
If PAY_1 = 7 then DEFAULT = 1
If PAY_1 = 8 then DEFAULT = 1

Table 9: DEFAULT Totals for PAY_1=5

	Count of Default Category	
PAY Value	0	1
Training Data		
5	7	6
All Data		
5	13	13

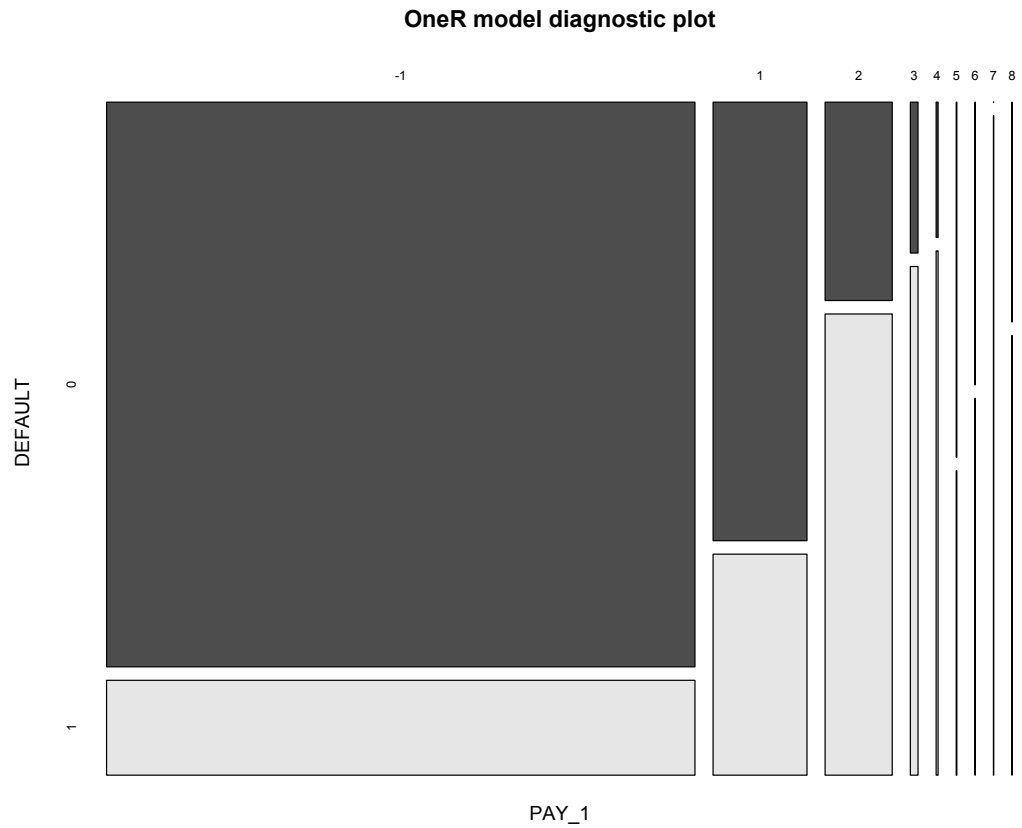


Image 4: Plot of One R model

4.b.3 SIMPLE LOGISTIC REGRESSION

The first logistic regression model was built using all of the variables from the training data. The results from this indicated that only a subset of the variables were significant. The model on the test data had an AUC value of .65, a sensitivity of .66, and a .15 Type I Error.

The second logistic regression model was constructed using only those variables that were identified as having p values of at least 0.01 (form2a in Index 7). They were as follows:

```
SEX + EDUCATION + MARRIAGE +
PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
LIMIT_BAL_Neg_29999 + LIMIT_BAL_30000_159999 + Avg_Bill_Amt_Neg_56050_188
 + Avg_Bill_Amt_189_2846 +Avg_Bill_Amt_2847_7488 + Avg_Bill_Amt_7489_31915
 + Avg_Pmt_Amt_0_2832 + Avg_Pmt_Amt_2833_12092 + Max_Bill_Amt + Max_DLQ
```

The AUC was .65, the sensitivity was .65 and the Type I Error was .15. The first model performed only slightly better on the test data.

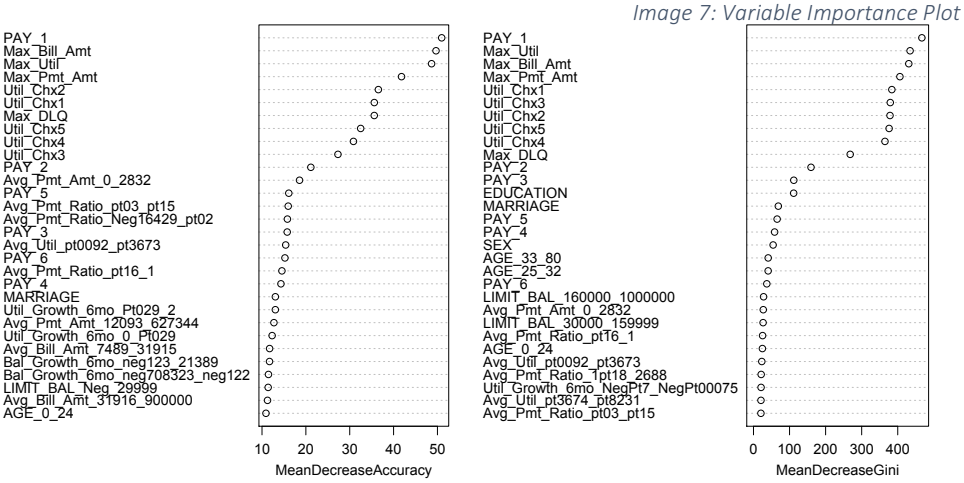
5. PREDICTIVE MODELING: METHODS & RESULTS

5.a RANDOM FOREST

Two models were developed using Random Forest. The first was trained using all of the variables in the training data.

The results of the randomForest model appear in Image 7. The Gini index identifies the variables that had the most impact during tree pruning, listed in Table 8.

Image 7: Variable Importance Plot



As we review these results, it's important to keep in mind that random forests remove some correlation effects among the trees by only considering a subset of the variables at a time (James, Witten, Hastie, Tibshirani, pg 320). This may be the reason why the variables identified in Table 8 vary from those we saw in the previous sections. The removal of the variables identified has the greatest impact on node purity while building the tree. This observation also aligns with the results from the Principal Component Analysis discussed in a later.

Table 10: Comparison of top 10 variables

	MeanDecreaseAccuracy	MeanDecreaseGini
PAY_1	50.968488	466.904341
Max_Util	48.662877	434.2111
Max_Bill_Amt	49.673479	430.603273
Max_Pmt_Amt	41.807768	406.007953
Util_Chx1	35.602539	383.644177
Util_Chx3	27.313471	379.30641
Table 10: Comparison of top 10 variables		
Util_Chx2	36.529396	378.53332
Util_Chx5	32.496147	376.075247
Util_Chx4	30.845884	364.58566
Max_DLQ	35.593939	268.157052

The results for this model, on the training data, were spectacular (AUC=.98, T1E=.00, Sensitivity=.97). However, the model did not perform as well on the test data (AUC=.66, T1E=.06, Sensitivity=.38), which is indicative of overfitting.

The second model was built using the ten variables identified in Table 8. The prediction results for this model were better than the first model (Test data: AUC=.66, T1E=.15, Sensitivity=.63). This suggests that these variables are highly predictive.

5.b GRADIENT BOOSTING

The XGBoost package was used to perform extreme gradient boosting. Three models were built: the first used all of the variables in the modeling suite, the second took the top 10 variables identified as important by the Random Forest results, and the third took all the variables except Pay_1:6. The performance of the third model was inferior than the first two and thus results are not included below.

The following parameters were used for all three models:

- `eta= .1, gamma = 3, max_depth=15, subsample=0.5, seed=1, eval_metric = "error", nthread = 2, objective="binary:logistic", booster = "gblinear")`
- `my_etas <- list(eta=c(.01, .05, 0.1, 0.5, 0.9))`

The first gradient boosting model identifies the variables listed in Image 8 as the most important variables. The key takeaway from this plot is that **PAY_1, Max_Util, Util_Chx1,** and **Max_DLQ** are considered the most important. These have already been identified as important by the Random Forest models.

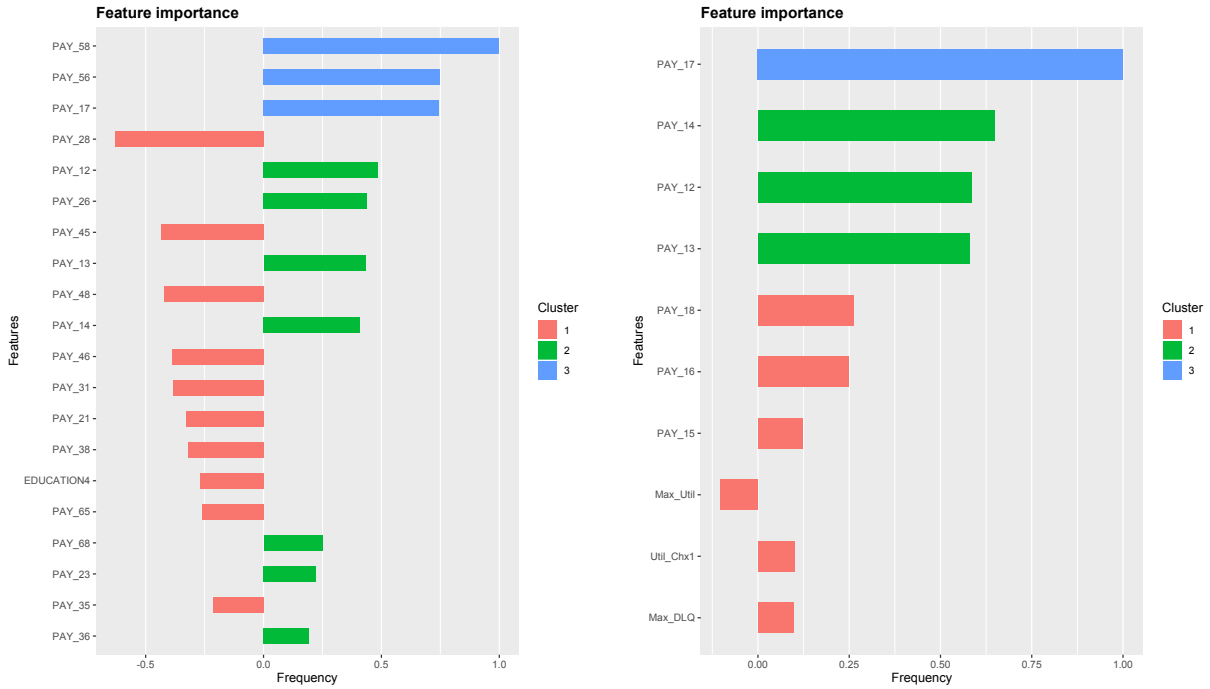


Image 5: Importance for Model Taking All Variables vs. Model Taking Only Variables from Random Forest

The first and second models produce results that are only marginally different. In this case, the second model performs very well, and as such, the 10 variables identified by the Random Forest Model will continue to be used.

- Model 1, test data : AUC = .74, Sensitivity = .67, Type I Error = .16
- Model 2, test data : AUC = .76, Sensitivity = .68, Type I Error = .16

5.c LOGISTIC REGRESSION WITH VARIABLE SELECTION

Variable selection was performed using regsubsets. Given that we want to determine optimal variable selection, all of the variables were included in the formula. Forward, backward and stepwise variable selection methods were used.

The model with the lowest Bayesian Information Criterion (BIC) value contained 11 values. As such, the coefficients were extracted and are depicted in Table 11 below. The following variables were not statistically significant:

- Avg_Bill_Amt_Neg_56050_188
- Avg_Pmt_Ratio_Neg16429_pt02
- Avg_Util_negpt2326_pt00095

The results lend insight into those variables that are likely to be the least significant, and we would not include them in the final model. For example, Avg_Pmt_Ratio_Neg16429_pt02 had a p-value of .995. We would not reject the null hypothesis for these variables. However, it is noted that these variables are having some effect on the BIC of model that was chosen using this approach as it was included in the results below.

Table 11: Logistic Regression 11 Variable Model

Coefficients:					
	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-1.74E+00	3.90E-02	-44.659	< 2e-16	***
PAY_11	6.25E-01	6.39E-02	9.779	< 2e-16	***
PAY_12	2.18E+00	7.12E-02	30.64	< 2e-16	***
PAY_13	2.22E+00	2.14E-01	10.378	< 2e-16	***
PAY_14	2.14E+00	4.14E-01	5.165	2.41E-07	***
PAY_15	7.37E-01	6.08E-01	1.211	0.22582	
PAY_16	1.01E+00	8.83E-01	1.14	0.254099	
PAY_17	1.56E+01	2.19E+02	0.071	0.943032	
PAY_18	-1.24E+01	5.35E+02	-0.023	0.981509	
PAY_31	-1.21E+01	3.78E+02	-0.032	0.974364	
PAY_32	7.22E-01	6.68E-02	10.814	< 2e-16	***
PAY_33	7.01E-01	2.14E-01	3.274	0.001062	**
PAY_34	3.56E-01	4.42E-01	0.804	0.421204	
PAY_35	-7.06E-01	1.12E+00	-0.632	0.527583	
PAY_36	1.41E+01	5.35E+02	0.026	0.979031	
PAY_37	8.06E-02	1.05E+00	0.077	0.938905	

Table 11: Logistic Regression 11 Variable Model

PAY_38	-1.33E+01	5.35E+02	-0.025	0.98024	
PAY_52	7.34E-01	7.53E-02	9.747	< 2e-16	***
PAY_53	4.58E-01	2.68E-01	1.709	0.087541	.
PAY_54	7.71E-01	4.85E-01	1.588	0.112209	
PAY_55	-6.27E-01	8.73E-01	-0.719	0.472411	
PAY_56	2.64E+01	7.57E+02	0.035	0.972242	
PAY_57	1.13E+00	8.10E-01	1.397	0.162264	
PAY_58	2.64E+01	7.57E+02	0.035	0.972199	
AGE_25_32	-1.61E-01	4.64E-02	-3.466	0.000528	***
Avg_Bill_Amt_Neg_56050_188	2.53E-01	2.24E-01	1.132	0.257509	
Avg_Pmt_Amt_2833_12092	-3.21E-01	4.99E-02	-6.424	1.32E-10	***
Avg_Pmt_Ratio_Neg16429_pt02	5.97E-04	9.98E-02	0.006	0.995232	
Avg_Util_negpt2326_pt00095	3.06E-01	2.23E-01	1.374	0.169502	
Avg_Pmt_Ratio_1pt18_2688	1.38E-01	6.00E-02	2.296	0.021662	*
Bal_Growth_6mo_21390_428792	-4.64E-01	1.18E-01	-3.93	8.50E-05	***

As the next step, a second logistic regression model was build using the variables identified as important in the Random Forest output. The values for both models are surprisingly very close:

- Regsubset Variable Model 1: AUC = .65, Sensitivity = .66, Type I Error = .15
- Random Forest Variable Model 2: AUC = .65, Sensitivity = .67, Type I Error = .16

5.c SUPPORT VECTOR MACHINE

Four models were developed and tested using this approach.

The first model was trained using all of the variables in the modeling suite. The parameters for the first and second models were set as noted below and the results of the tuning parameters from the cross-validation are listed in Table 10:

- method= "repeatedcv", number = 5, summaryFunction=twoClassSummary, classProbs=TRUE
- method= "svmRadial, tuneLength=10, preProc=c("center", "scale"), metric= "ROC"

The results for this model indicated that the best tuning parameters are C= 0.25 and σ = 1.119692. With this model, a sensitivity of .97 is expected, which is the highest that we have seen. However, this comes at a cost; the specificity of the model drops to only .23.

Table 12: Results reported for Tuning Parameters

	sigma	C	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
1	1.119692	0.25	0.7094106	0.9700602	0.2322521	0.0254288	0.00205673	0.01385188
2	1.119692	0.5	0.7033392	0.9650419	0.26321851	0.02263968	0.003715118	0.01910575
3	1.119692	1	0.7032039	0.9591732	0.2795757	0.02182591	0.004316414	0.02309177
4	1.119692	2	0.6947753	0.9590029	0.27577923	0.02166133	0.003837763	0.0246985

Table 12: Results reported for Tuning Parameters

5	1.119692	4	0.6539823	0.9577927	0.27026615	0.08946009	0.004315988	0.02362903
6	1.119692	8	0.6857179	0.958833	0.26438596	0.01553916	0.003433002	0.02358957
7	1.119692	16	0.6863271	0.9604491	0.25620225	0.01536072	0.002722446	0.02070043
8	1.119692	32	0.6835784	0.9658075	0.22698724	0.01825596	0.004301164	0.01762392
9	1.119692	64	0.6790596	0.9721864	0.17028898	0.0174045	0.006695337	0.05597114
10	1.119692	128	0.678052	0.9834146	0.07186708	0.02233504	0.003618006	0.0181078

The second model was developed to fine tune the vectors separating the classes. It contained all the variables. The following parameters were used to train it:

- sigma = 0.001, 0.01, .015, .02
- C = c(0.1, 0.2, 0.25, 0.26, 0.27))

The third support vector machine was modeled using the same parameters and the following variables:

```
SEX + EDUCATION + MARRIAGE + PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5
+  LIMIT_BAL_Neg_29999 + LIMIT_BAL_30000_159999 + AGE_25_32 +
Avg_Bill_Amt_Neg_56050_188 + Avg_Bill_Amt_189_2846 +
Avg_Bill_Amt_2847_7488 + Avg_Bill_Amt_7489_31915 + Avg_Pmt_Amt_0_2832 +
Avg_Pmt_Amt_2833_12092 + Avg_Pmt_Ratio_1_1pt17 + Avg_Util_pt0092_pt3673 +
Util_Growth_6mo_Neg5_NegPt7 + Util_Growth_6mo_NegPt7_NegPt00075 +
Util_Growth_6mo_NegPt00075_0 + Util_Growth_6mo_0_Pt029 +
Util_Growth_6mo_Pt029_2 + Max_Bill_Amt + Max_DLQ +
Bal_Growth_6mo_neg708323_neg122
```

The results for this model were nearly the same as the first model, when comparing the predictions on the test data (AUC = .66, Sensitivity = .66, Type I Error = .16). The AUC was .01 higher.

The fourth model was trained on the variables identified by the Random Forest model, using the fine tuned parameters:

```
PAY_1 + Max_Util +
Max_Bill_Amt +
Max_Pmt_Amt + Util_Chx1 +
Util_Chx3 + Util_Chx2 +
Util_Chx5 + Util_Chx4 +
Max_DLQ
```

Despite having the least number of variables, had the highest sensitivity of all the SVMs (AUC = .64, Sensitivity = .67, Type I Error = .16).

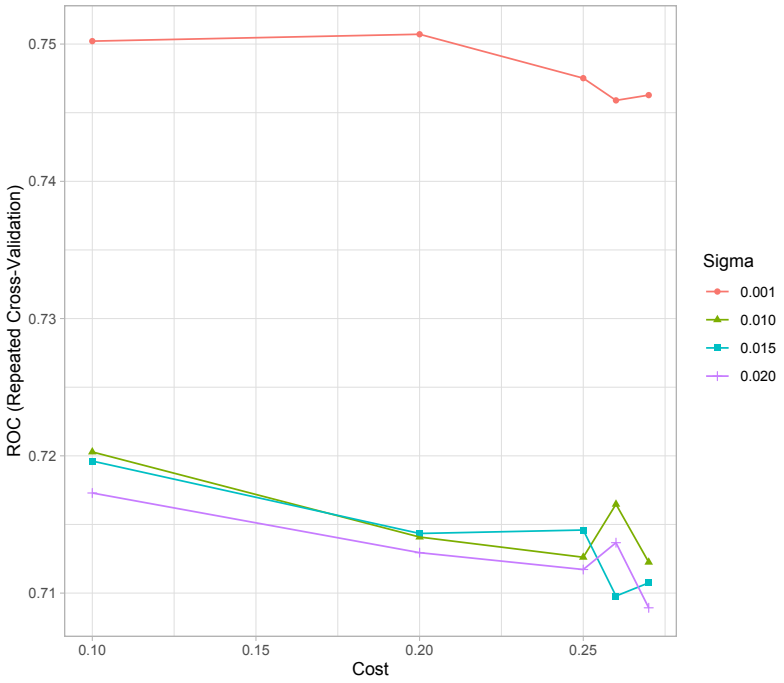


Image 9: Tuning Parameter Performance for 4th SVM

5.c PRINCIPAL COMPONENT ANALYSIS

Given that the models discussed above have produced average results, a principal component analysis was performed on all the continuous variables in our data. See Index 7 for a list of the variables used. Also, the procedure would allow for the investigation of dependencies between the variables. The correlations listed in Table 13 illustrate the variables and the extent of their correlations for perspective.

Table 13: Correlations of .50 or more

	Var1	Var2	Freq
1	Avg_Bill_Amt	Max_Bill_Amt	0.943825984
2	Avg_Util	Max_Util	0.924110832
3	Avg_Pmt_Amt	Max_Pmt_Amt	0.910191244
4	Bal_Growth_6mo	Util_Growth_6mo	0.682274858
5	Avg_Bill_Amt	Avg_Util	0.547751392
6	Max_Bill_Amt	Max_Util	0.50258326
7	Avg_Bill_Amt	Max_Util	0.486811588
8	Avg_Pmt_Amt	Max_Bill_Amt	0.47271112
9	Avg_Util	Max_Bill_Amt	0.471248664

Taking a look at the scree plot in Image 10, we see that approximately 75% of the variance is explained in 6 dimensions. Further the chart in Image 9 reflects the groupings based on the distances between the variables. To summarize the takeaways, the following general groupings are observed:

- PAY 1:6 & LIMIT_BAL
- Max_Pmt_Amt & Avg_Pmt_Amt
- Util_Growth_6mo & Bal_Growth_6mo
- Avg_Pmt_Ratio
- Age
- DEFAULT & Max_DLQ
- Max_Util & Avg_Util
- BILL_AMT1:6, Max_Bill_Amt, Avg_Bill_Amt

This information allows us to have a closer look at our data and some of the limitations or opportunities for improvement. It is worth noting that the variances between DEFAULT and Max_DLQ are very similar, and indicative the predictive capacity of this variable. This may have to do with the fact the this variable is likely to be tied to the business rule that defines at which point the loan is considered to be in default status.

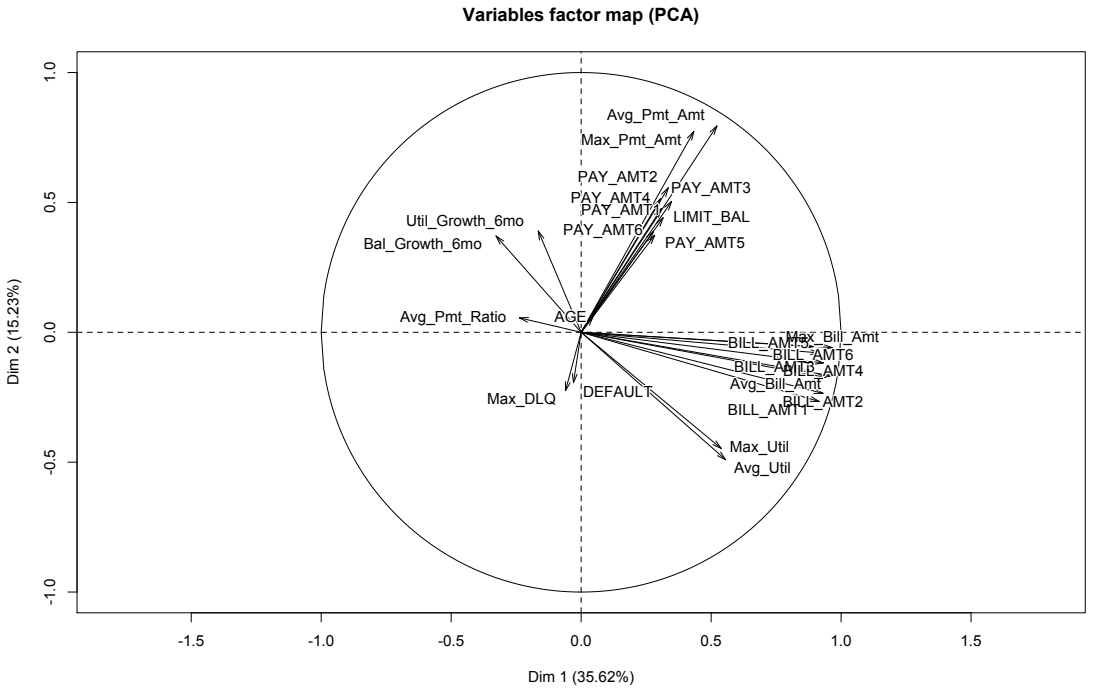


Image 10: Variable Factor Map

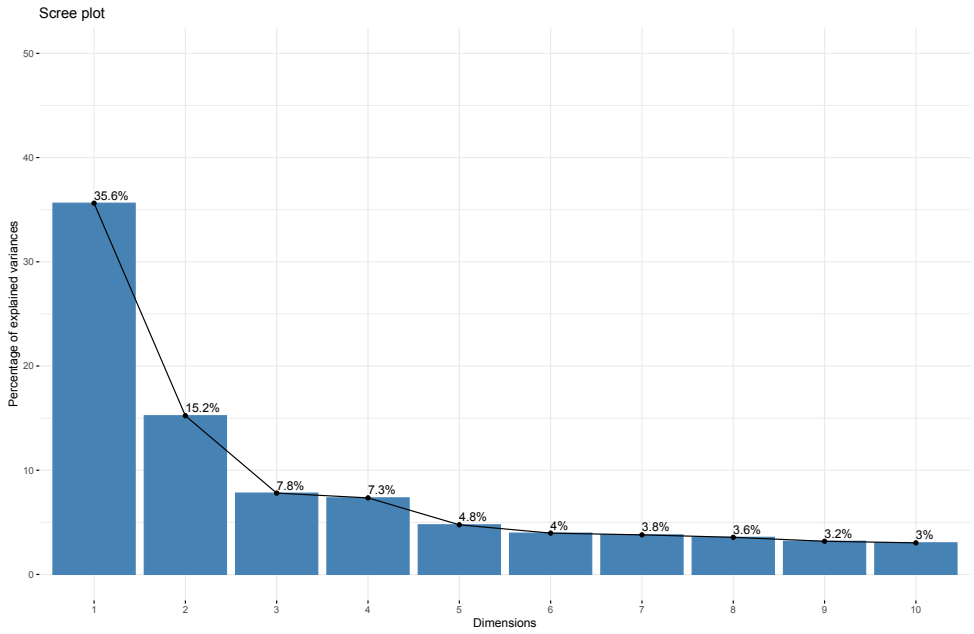


Image 6: Scree Plot

6. COMPARISON OF RESULTS

In this section, the results of all the models are reviewed, but the best performing model in each category are discussed in detail. See Index 7: Confusion Matrices for all Models.

6.a. RANDOM FOREST

As previously mentioned, two random forest models were built. One was trained on the whole suite of variables, as defined in Index 3. The first model displays signs of severe overfitting on the training data (Type I Error = .00, AUC =.98, and Sensitivity = .97). The results from the test data were significantly lower (Type I Error = .06, AUC =.66, and Sensitivity = .38). The sensitivity is very low.

The second was built using the top ten variables that were identified in the previous step.
`PAY_1 + Max_Util + Max_Bill_Amt + Max_Pmt_Amt + Util_Chx1 + Util_Chx3 + Util_Chx2 + Util_Chx5 + Util_Chx4 + Max_DLQ`

Here we do not observe such a difference between the results from the training (Type I Error = .01, AUC =.77, and Sensitivity = .54) and the testing data (Type I Error = .06, AUC =.65, and Sensitivity = .36). This suggest that there is less overfitting and that the second model is more stable. This is the better model in this set. The sensitivity levels for these two models is very low and potentially unacceptable, based on the business' risk tolerance.

6.b. GRADIENT BOOSTING

While a total of 3 models were built using Gradient Boosting, we will focus on the results of the best performing model that was built using the following variables (identified as the most important in Random Forest):

```
PAY_1 + Max_Util + Max_Bill_Amt + Max_Pmt_Amt + Util_Chx1 + Util_Chx3 +
Util_Chx2 + Util_Chx5 + Util_Chx4 + Max_DLQ
```

This model had very similar results for the predictions made on training (Type I Error = .17, AUC =.64, and Sensitivity = .70) and the testing data (Type I Error = .16, AUC =.76, and Sensitivity = .68).

In comparing the results on the test data for this model with the best Random Forest model, we see that the AUC percentage is .05 lower and the Type 1 Error for this model .06 higher. However, the sensitivity is significantly higher in this model; it increased from .36 to .68. This is a more acceptable percentage for accurately predicted default loans.

6.c. LOGISTIC REGRESSION

More than 5 logistic regression models were tested. However, the focus will be on the baseline model that was trained using all the variables (Index 3), and the final model that was selected for scoring purposes given that the results produced by different variable groupings produced only inferior results.

The results of the baseline model were within the range of the best performing models discussed up to this point (Testing data: Type I Error = .15, AUC =.65, and Sensitivity = .66). It is interesting to note that although all of the variables were used for the model, the results predicting using the training data did not show signs of overfitting, as was the case in the first Random Forest model discussed.

The final logistic regression model that was selected included those variables that were identified as important by the Random Forest model:

```
PAY_1 + Max_Util + Max_Bill_Amt + Max_Pmt_Amt + Util_Chx1 + Util_Chx3 +
Util_Chx2 + Util_Chx5 + Util_Chx4 + Max_DLQ
```

This set of variables was chosen because Random Forests are expected to adjust for some of correlation going on between many of the variables. By performing a principal component analysis, it was possible to confirm the groupings with very similar distances between their variance, and the importance of excluding the variables that have high correlations.

The measures for the final model were as follows: Type I Error = .16, AUC =.65, and Sensitivity = .67. This model is the best among the logistic regression models given its higher sensitivity percentage and the limited number of variables that produced these results. However, the best performing gradient boosting model is better than this model given that the AUC is .11 higher.

6.c. SUPPORT VECTOR MACHINE (SVM)

Four SVMs were built, the first was trained using all of the variables in the modeling suite. The results for this model were as follows:

- Training data: Type I Error = .17, AUC =.65, and Sensitivity = .66
- Testing data: Type I Error = .16, AUC =.65, and Sensitivity = .66

These results are lower than the results we see for the logistic and best gradient boosting models.

The second model was also trained using all of the variables, but it used refined values for the cost of misclassification C and the kernel parameter gamma. This model had a lower sensitivity percentage than the first SVM model but a lower Type I Error and a higher AUC, when predicting on the test data (Type I Error = .15, AUC =.66, and Sensitivity = .64).

The third model was developed using a subset of the variables, those identified as significant by the logistic regression model:

```
SEX + EDUCATION + MARRIAGE + PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
LIMIT_BAL_Neg_29999 + LIMIT_BAL_30000_159999 + AGE_25_32 +
Avg_Bill_Amt_Neg_56050_188 + Avg_Bill_Amt_189_2846 +
Avg_Bill_Amt_2847_7488 + Avg_Bill_Amt_7489_31915 + Avg_Pmt_Amt_0_2832 +
Avg_Pmt_Amt_2833_12092 + Avg_Pmt_Ratio_1_1pt17 + Avg_Util_pt0092_pt3673 +
Util_Growth_6mo + Max_Bill_Amt + Max_DLQ + Bal_Growth_6mo_neg708323_neg122
```

The third SVM model performed marginally better than the previous two (Type I Error = .16, AUC =.66, and Sensitivity = .66). It was superior given that is produced similar results using fewer variables. The Type I Error was not the lowest for this model, but it did have a higher AUC and Sensitivity.

The fourth model was trained using the following variables:

```
PAY_1 + Max_Util + Max_Bill_Amt + Max_Pmt_Amt + Util_Chx1 + Util_Chx3 +
Util_Chx2 + Util_Chx5 + Util_Chx4 + Max_DLQ
```

This model had the highest sensitivity out of all the SVMs, but also the lowest AUC (AUC = .64, Sensitivity = .67, Type I Error = .16).

6.d. BEST MODEL

The best model was developed using Gradient Boosting and trained on the following variables:
`PAY_1 + Max_Util + Max_Bill_Amt + Max_Pmt_Amt + Util_Chx1 + Util_Chx3 + Util_Chx2 + Util_Chx5 + Util_Chx4 + Max_DLQ.`

These were the variables that were identified using Random Forest. Image 8 captures an area under the curve of .76, which is significantly higher than the other models. The Type I Error was .16; this indicates the percentage of individuals that were incorrectly predicted to default, but who

did not actually default. This misclassification is problematic to the business as it reflects opportunity costs. If we were to put this model into production, approximately 1,000 out of 6,500 individuals would not be approved for a loan, but should have been if the model were more robust.

The sensitivity of the model reflects the opposite – the percentage represents the amount of the total which was correctly predicted. In this case, 68% of the 767 actual defaults were correctly predicted. While this is the best performance that we've seen here, it indicates the other 32% were approved for a loan but who defaulted, which means losses to the business.

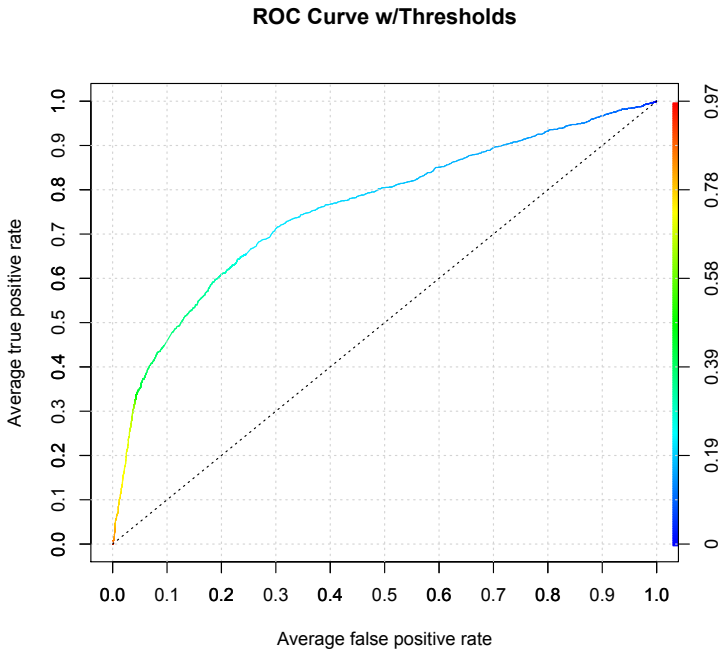


Image 7: ROC Curve for Best Model

Conclusion

Five different approaches were used to predict the occurrence of default. The following types of models were used to address this binary classification problem: Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine and Principal Component Analysis. Various features were also engineered from the original variables, such as average utility, average bill amount, max utility. Using these approaches, the best results were achieved using Extreme Gradient Boosting with a Type I Error of .16, an AUC of .76, and a Sensitivity of .68. The final set of variables associated with this model were identified using the Random Forest variable importance plot.

Considering that the data used was limited to demographic and 6 month payment history for 30,000 individuals, these results are acceptable and they help to inform the business of the risks they would undertake if they would use the highest performing model.

The Type I Error indicates the percentage of individuals that were incorrectly predicted to default, but who did not actually default. As such, the rate of misclassification is problematic to the business as it reflects opportunity costs.

The sensitivity of the model reflects the opposite – the percentage represents the amount of the total which was correctly predicted. While the best model is expected to predict the true positives at a rate of 68%, this means that the other 32% were approved for a loan but who defaulted, which results in losses to the business.

Reflecting on improvement opportunities, additional information would be expected to help improve prediction capacity. Particularly considering that a more traditional approach would have included bureau data, application data and longer payment histories. Another strategy that should be considered in the future would be to populate the binned variables using the weights of evidence and to modify the variables to reduce the impact of their dispersion. This would have still allowed for capturing the nuances of the different records. Lastly, it would have been a worthy exercise to study the differences between the classes and to create models for different profiles, such as by credit limit groups. This is a common approach that has the potential of refining the predictability of models.

Index 1 Summary Statistics for RAW Credit Card Clients Data

Table 4: Summary Statistics for RAW Credit Card Clients Data – FULL DATA SET

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max	Range
LIMIT_BAL	30,000	167,484.30	129,747.70	10,000	50,000	140,000	240,000	1,000,000	990,000
AGE	30,000	35.49	9.22	21	28	34	41	79	58
BILL_AMT1	30,000	51,223.33	73,635.86	-165,580	3,558.80	22,381.50	67,091	964,511	1,130,091
BILL_AMT2	30,000	49,179.08	71,173.77	-69,777	2,984.80	21,200	64,006.20	983,931	1,053,708
BILL_AMT3	30,000	47,013.15	69,349.39	-157,264	2,666.20	20,088.50	60,164.80	1,664,089	1,821,353
BILL_AMT4	30,000	43,262.95	64,332.86	-170,000	2,326.80	19,052	54,506	891,586	1,061,586
BILL_AMT5	30,000	40,311.40	60,797.16	-81,334	1,763	18,104.50	50,190.50	927,171	1,008,505
BILL_AMT6	30,000	38,871.76	59,554.11	-339,603	1,256	17,071	49,198.20	961,664	1,301,267
PAY_AMT1	30,000	5,663.58	16,563.28	0	1,000	2,100	5,006	873,552	873,552
PAY_AMT2	30,000	5,921.16	23,040.87	0	833	2,009	5,000	1,684,259	1,684,259
PAY_AMT3	30,000	5,225.68	17,606.96	0	390	1,800	4,505	896,040	896,040
PAY_AMT4	30,000	4,826.08	15,666.16	0	296	1,500	4,013.20	621,000	621,000
PAY_AMT5	30,000	4,799.39	15,278.31	0	252.5	1,500	4,031.50	426,529	426,529
PAY_AMT6	30,000	5,215.50	17,777.47	0	117.8	1,500	4,000	528,666	528,666

Index 2 Summary Statistics for Variables Engineered using Credit Card Clients Data

Table 2: Summary Statistics for Variables Engineered using Credit Card Clients Data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Avg_Bill_Amt	15,180	44,934.91	63,150.01	-56,043	4,789.1	21,198.4	56,880.5	592,432
Avg_Pmt_Amt	15,180	5,255.11	10,150.14	0.00	1,111.75	2,389.42	5,554.42	627,344.30
Pmt_Ratio1	15,180	8.84	45.48	-498	0.04	0.1	1	4,444
Pmt_Ratio2	15,180	10.42	61.61	-41	0.04	0.1	1	5,001
Pmt_Ratio3	15,180	11.10	47.45	-500	0.04	0.1	1	4,444
Pmt_Ratio4	15,180	11.55	41.37	-3,030	0.04	0.1	1	100
Pmt_Ratio5	15,180	13.46	33.87	-31	0.04	0.1	1	448
Avg_Pmt_Ratio	15,180	11.07	35.02	-605.46	0.05	0.16	1.00	2,687.00
Util1	15,180	0.42	0.41	-0.62	0.02	0.32	0.83	6.46
Util2	15,180	0.41	0.41	-1.40	0.02	0.30	0.81	6.38
Util3	15,180	0.39	0.39	-1.03	0.02	0.28	0.75	5.39
Util4	15,180	0.36	0.37	-1	0.02	0.2	0.7	5
Util5	15,180	0.33	0.35	-1	0.01	0.2	0.6	5
Util6	15,180	0.32	0.34	-1	0.01	0.2	0.6	4
Avg_Util	15,180	0.37	0.35	-0.23	0.03	0.29	0.69	5.36
Bal_Chx1	15,180	-1,920.47	22,004.65	-384,675	-2,145	0	1,571.5	489,972
Bal_Chx2	15,180	-2,276.50	25,084.75	-512,650	-2,596.2	0	1,389	391,348
Bal_Chx3	15,180	-3,823.40	24,930.66	-418,926	-3,434	0	1,022.2	429,981
Bal_Chx4	15,180	-3,043.44	22,098.98	-432,730	-2,705	0	996	341,696
Bal_Chx5	15,180	-1,431.53	19,142.81	-400,000	-1,625.2	0	1,184	381,629
Bal_Growth_6mo	15,180	-12,495.35	44,045.77	-497,231	-20,094.2	-918.5	2,962	399,983
Util_Chx1	15,180	-0.01	0.15	-2.63	-0.02	0.00	0.02	1.63
Util_Chx2	15,180	-0.02	0.17	-4.91	-0.02	0.00	0.01	1.99
Util_Chx3	15,180	-0.03	0.17	-3.02	-0.03	0.00	0.01	1.68
Util_Chx4	15,180	-0.03	0.16	-2	-0.02	0	0.01	2
Util_Chx5	15,180	-0.01	0.13	-2	-0.01	0	0.01	2

Table 2: Summary Statistics for Variables Engineered using Credit Card Clients Data

Util_Growth_6mo	15,180	-0.11	0.30	-5.31	-0.18	-0.01	0.03	1.83
Max_Bill_Amt	15,180	60,425.45	77,746.88	-2,900	10,050.8	31,587.5	79,119.5	823,540
Max_Pmt_Amt	15,180	15,620.73	35,279.24	0	2,195.8	5,000	12,200.8	1,215,471
Max_DLQ	15,180	0.68	1.07	0	0	0	2	8

Index 3: Complete Suite of Modeling Variables

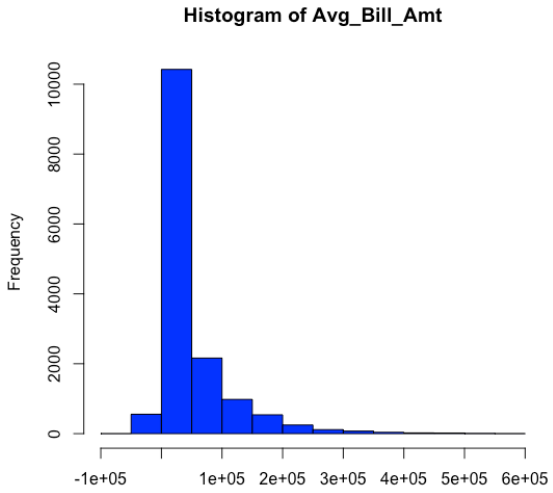
Table 3: Listing of variables 1-94 contained in dataframe cc_default_df.
Complete modeling suite in dataframe train_df (46 total) are highlighted in blue

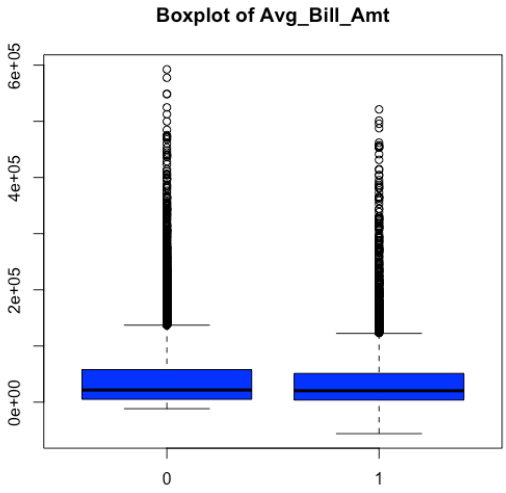
1	ID	int
2	LIMIT_BAL	int
3	SEX	Factor
4	EDUCATION	Factor
5	MARRIAGE	Factor
6	AGE	int
7	PAY_1	Factor
8	PAY_2	Factor
9	PAY_3	Factor
10	PAY_4	Factor
11	PAY_5	Factor
12	PAY_6	Factor
13	BILL_AMT1	int
14	BILL_AMT2	int
15	BILL_AMT3	int
16	BILL_AMT4	int
17	BILL_AMT5	int
18	BILL_AMT6	int
19	PAY_AMT1	int
20	PAY_AMT2	int
21	PAY_AMT3	int
22	PAY_AMT4	int
23	PAY_AMT5	int
24	PAY_AMT6	int
25	DEFAULT	Factor
26	u	num
27	train	Factor
28	test	Factor
29	validate	Factor
30	data.group	Factor
31	LIMIT_BAL_Neg_29999	Factor
32	LIMIT_BAL_30000_159999	Factor
33	LIMIT_BAL_160000_1000000	Factor
34	AGE_0_24	Factor
35	AGE_25_32	Factor
36	AGE_33_80	Factor
37	Avg_Bill_Amt	num

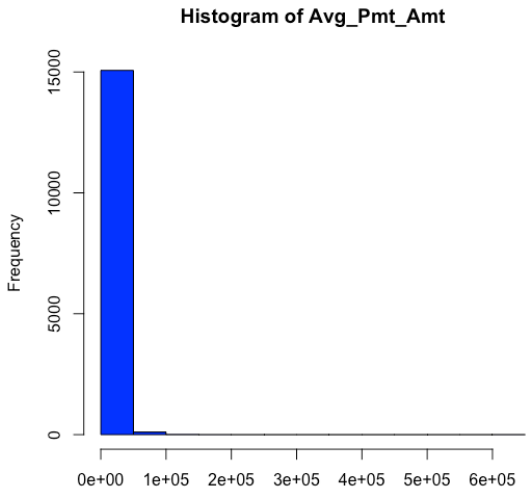
38	Avg_Bill_Amt_Neg_56050_188	Factor
39	Avg_Bill_Amt_189_2846	Factor
40	Avg_Bill_Amt_2847_7488	Factor
41	Avg_Bill_Amt_7489_31915	Factor
42	Avg_Bill_Amt_31916_900000	Factor
43	Avg_Pmt_Amt	num
44	Avg_Pmt_Amt_0_2832	Factor
45	Avg_Pmt_Amt_2833_12092	Factor
46	Avg_Pmt_Amt_12093_627344	Factor
47	Pmt_Ratio1	num
48	Pmt_Ratio2	num
49	Pmt_Ratio3	num
50	Pmt_Ratio4	num
51	Pmt_Ratio5	num
52	Avg_Pmt_Ratio	num
53	Avg_Pmt_Ratio_Neg16429_pt02	Factor
54	Avg_Pmt_Ratio_pt03_pt15	Factor
55	Avg_Pmt_Ratio_pt16_1	Factor
56	Avg_Pmt_Ratio_1_1pt17	Factor
57	Avg_Pmt_Ratio_1pt18_2688	Factor
58	Util1	num
59	Util2	num
60	Util3	num
61	Util4	num
62	Util5	num
63	Util6	num
64	Avg_Util	num
65	Avg_Util_negpt2326_pt00095	Factor
66	Avg_Util_pt00096_pt0091	Factor
67	Avg_Util_pt0092_pt3673	Factor
68	Avg_Util_pt3674_pt8231	Factor
69	Avg_Util_pt8232_6	Factor
70	Bal_Chx1	int
71	Bal_Chx2	int
72	Bal_Chx3	int
73	Bal_Chx4	int
74	Bal_Chx5	int
75	Bal_Growth_6mo	int
76	Bal_Growth_6mo_neg708323_neg122	Factor
77	Bal_Growth_6mo_neg123_21389	Factor

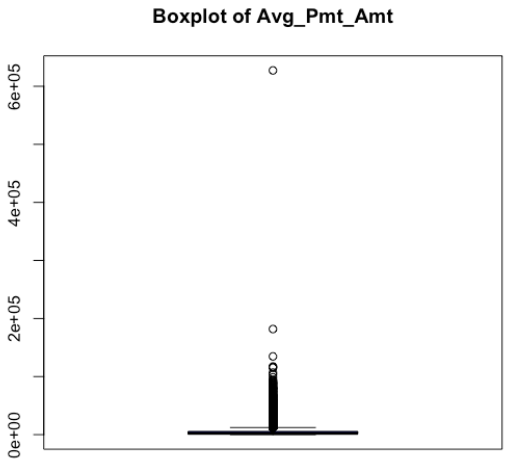
78	Bal_Growth_6mo_21390_428792	Factor
79	Util_Chx1	num
80	Util_Chx2	num
81	Util_Chx3	num
82	Util_Chx4	num
83	Util_Chx5	num
84	Util_Growth_6mo	num
85	Util_Growth_6mo_Neg5_NegPt7	chr
86	Util_Growth_6mo_NegPt7_NegPt00075:	1
87	Util_Growth_6mo_NegPt00075_0	chr
88	Util_Growth_6mo_0_Pt029	chr
89	Util_Growth_6mo_Pt029_2	chr
90	Max_Bill_Amt	int
91	Max_Pmt_Amt	int
92	Max_DLQ	num
93	Max_Util	num

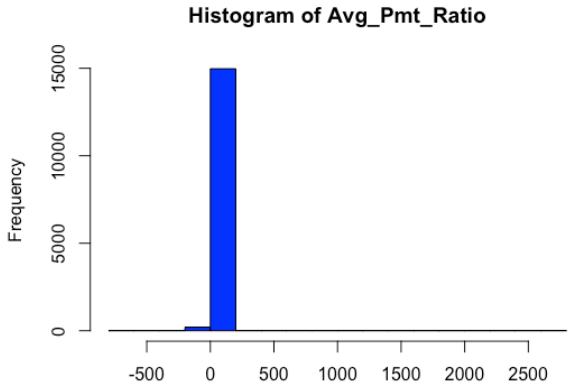
Index 4: EDA Graphics

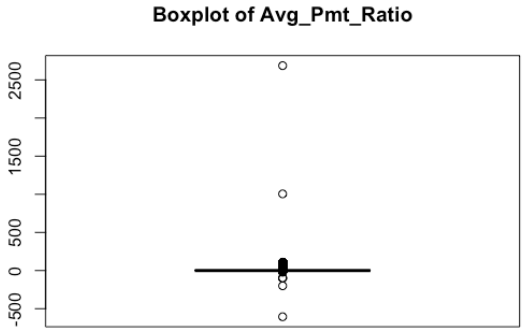


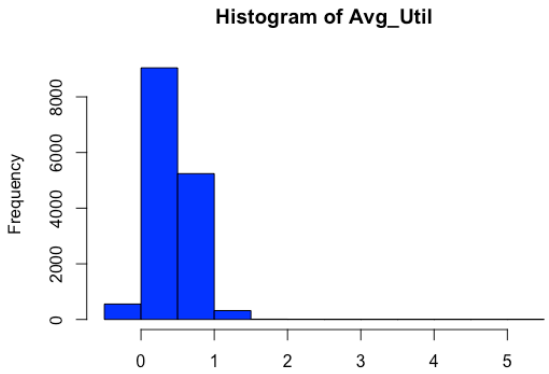


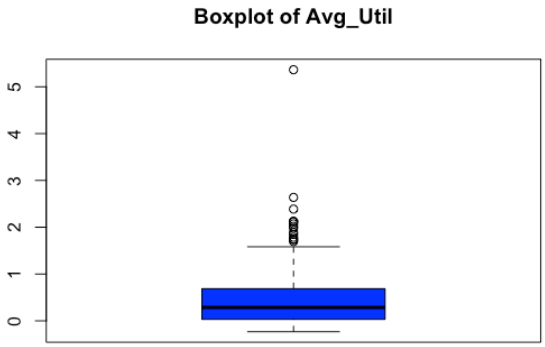


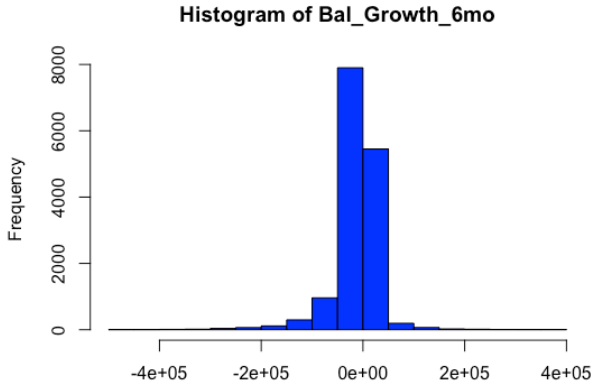


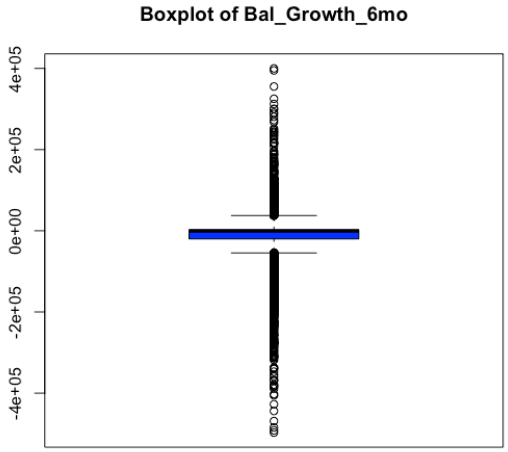


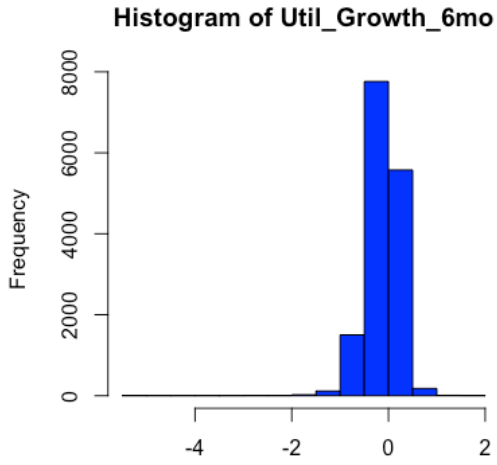


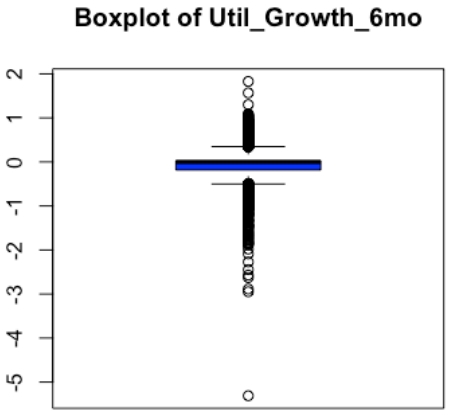

Histogram of Avg_Util
Frequency
0
2000
4000
6000
8000
0
1
2
3
4
5

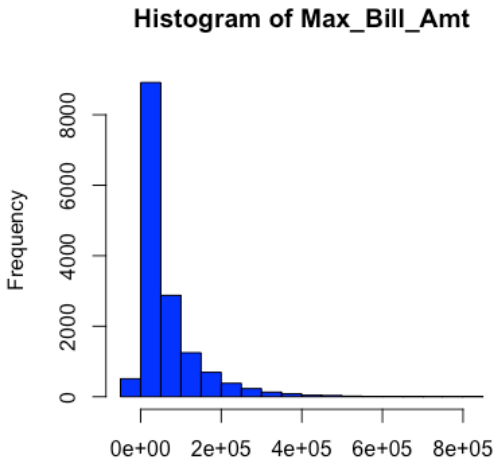

Boxplot of Avg_Util
0
1
2
3
4
5

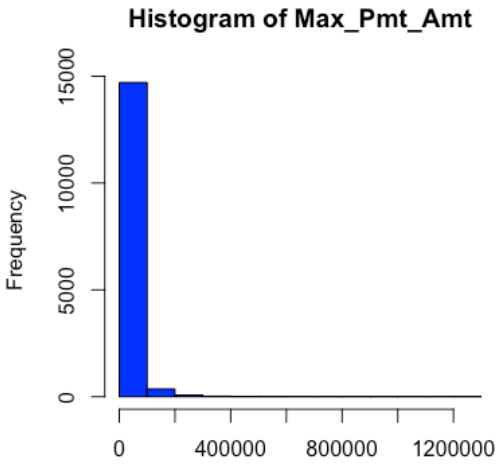

Histogram of Bal_Growth_6mo
Frequency
0
2000
4000
6000
8000
-4e+05
-2e+05
0e+00
2e+05
4e+05

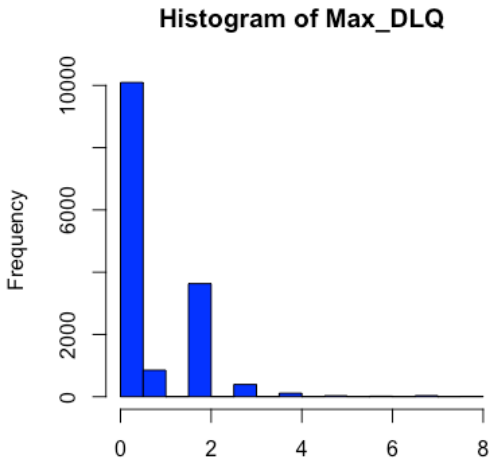

Boxplot of Bal_Growth_6mo
-4e+05
-2e+05
0e+00
2e+05
4e+05

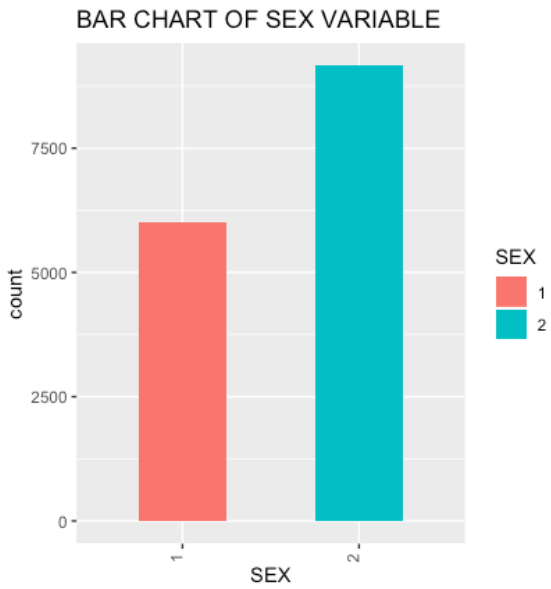

Histogram of Util_Growth_6mo
Frequency
0
2000
4000
6000
8000
-4
-2
0
2

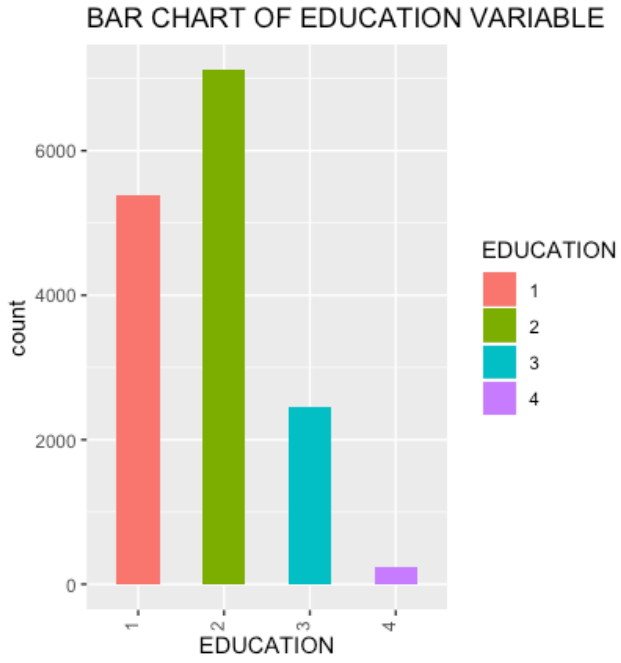

Boxplot of Util_Growth_6mo
-5
-4
-3
-2
-1
0
1
2

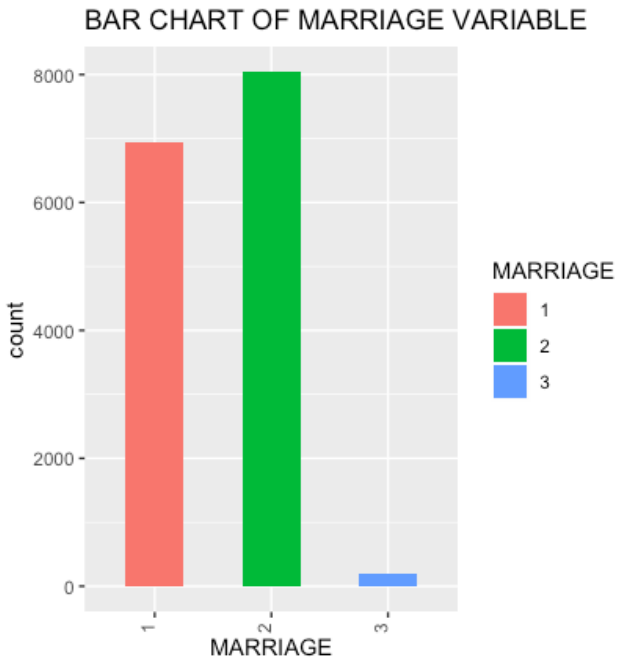

Histogram of Max_Bill_Amt
Frequency
0
2000
4000
6000
8000
0e+00
2e+05
4e+05
6e+05
8e+05

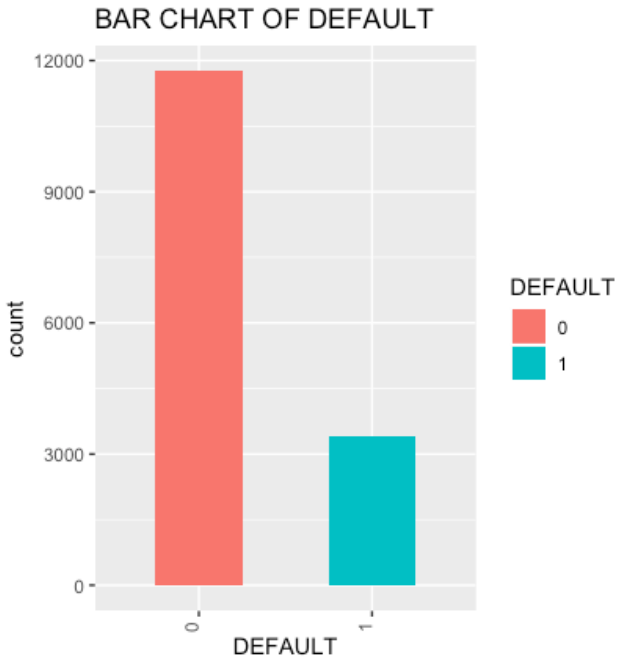

Histogram of Max_Pmt_Amt
Frequency
0
5000
10000
15000
0
400000
800000
1200000


Histogram of Max_DLQ
Frequency
0
2000
6000
10000
0
2
4
6
8


BAR CHART OF SEX VARIABLE
7500
5000
2500
0
count
1
2
SEX
SEX
1
2


BAR CHART OF EDUCATION VARIABLE
6000
4000
2000
0
count
1
2
3
4
EDUCATION
EDUCATION
1
2
3
4


BAR CHART OF MARRIAGE VARIABLE
8000
6000
4000
2000
0
count
1
2
3
MARRIAGE
MARRIAGE
1
2
3


BAR CHART OF DEFAULT
12000
9000
6000
3000
0
count
0
1
DEFAULT
DEFAULT
0
1

Index 5: Correlation Matrix for Engineered Variables

	LIMIT_BAL	AGE	Avg_Bill_Amt	Avg_Pmt_Amt	Avg_Pmt_Ratio	Avg_Util	Bal_Growth_6mo	Util_Growth_6mo	Max_Bill_Amt	Max_Pmt_Amt	Max_DLQ	Max_Util
LIMIT_BAL	1	0.14	0.3	0.35	0.01	-0.38	-0.08	0.13	0.35	0.29	-0.24	-0.4
AGE	0.14	1	0.05	0.04	0.01	-0.04	-0.03	-0.02	0.06	0.02	-0.02	-0.04
Avg_Bill_Amt	0.3	0.05	1	0.34	-0.06	0.55	-0.32	-0.13	0.94	0.25	-0.03	0.49
Avg_Pmt_Amt	0.35	0.04	0.34	1	-0.04	0.03	0.06	0.09	0.47	0.91	-0.14	0.11
Avg_Pmt_Ratio	0.01	0.01	-0.06	-0.04	1	-0.09	0	-0.01	-0.06	-0.03	-0.01	-0.08
Avg_Util	-0.38	-0.04	0.55	0.03	-0.09	1	-0.17	-0.23	0.47	0	0.19	0.92
Bal_Growth_6mo	-0.08	-0.03	-0.32	0.06	0	-0.17	1	0.68	-0.42	0.11	0.1	-0.27
Util_Growth_6mo	0.13	-0.02	-0.13	0.09	-0.01	-0.23	0.68	1	-0.21	0.11	0.11	-0.41
Max_Bill_Amt	0.35	0.06	0.94	0.47	-0.06	0.47	-0.42	-0.21	1	0.41	-0.08	0.5
Max_Pmt_Amt	0.29	0.02	0.25	0.91	-0.03	0	0.11	0.11	0.41	1	-0.11	0.1
Max_DLQ	-0.24	-0.02	-0.03	-0.14	-0.01	0.19	0.1	0.11	-0.08	-0.11	1	0.12
Max_Util	-0.4	-0.04	0.49	0.11	-0.08	0.92	-0.27	-0.41	0.5	0.1	0.12	1

Index 6: Importance Chart for Random Forest Using Subset of Variables (form 2a)

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
SEX	9.095441	8.5483042	12.466403	160.66116
EDUCATION	8.491273	-2.4481932	6.4245928	263.64843
MARRIAGE	15.835839	-0.4626992	13.1207474	188.27745
PAY_1	44.78228	74.8766638	83.5806748	756.46894
PAY_2	1.966628	16.3282007	14.5373777	92.93125
PAY_3	-2.835189	13.268845	6.8030292	76.39696
PAY_4	-11.54529	19.7906141	17.6160086	63.68552
PAY_5	-23.783377	42.5084938	19.9130459	60.85546
LIMIT_BAL_Neg_29999	2.896881	11.9650974	7.4640264	51.8752
LIMIT_BAL_30000_159999	-2.036914	21.322773	7.3271492	98.98809
Avg_Bill_Amt_Neg_56050_188	13.058755	-6.3106067	15.35043	12.05578
Avg_Bill_Amt_189_2846	-2.011258	4.2672019	-0.7034925	21.45897
Avg_Bill_Amt_2847_7488	7.889474	10.2860609	12.4506081	32.98116
Avg_Bill_Amt_7489_31915	20.881326	-3.6562371	20.5994671	54.33202
Avg_Pmt_Amt_0_2832	24.089653	-4.7371078	26.0126324	53.54871
Avg_Pmt_Amt_2833_12092	-13.96649	13.9585275	-7.8308548	50.48338
Max_Bill_Amt	32.528973	13.821541	42.3272584	2658.27138
Max_DLQ	4.382268	23.3380463	31.425423	294.90739

Index 7: Variable Formulations Used for Models

```
#all variables
form <- as.formula(DEFAULT ~ .)

#all variables minus PAY1:6
form2 <- as.formula(DEFAULT ~ SEX + EDUCATION + MARRIAGE +
   LIMIT_BAL_Neg_29999 + LIMIT_BAL_30000_159999 +
   LIMIT_BAL_160000_1000000 + AGE_0_24 + AGE_25_32 + AGE_33_80 +
   Avg_Bill_Amt_Neg_56050_188 + Avg_Bill_Amt_189_2846 +
   Avg_Bill_Amt_2847_7488 + Avg_Bill_Amt_7489_31915 +
   Avg_Bill_Amt_31916_900000 + Avg_Pmt_Amt_0_2832 +
   Avg_Pmt_Amt_2833_12092 + Avg_Pmt_Amt_12093_627344 +
   Avg_Pmt_Ratio_Neg16429_pt02 + Avg_Pmt_Ratio_pt03_pt15 +
   Avg_Pmt_Ratio_pt16_1 + Avg_Pmt_Ratio_1_1pt17 + Avg_Pmt_Ratio_1pt18_2688
   + Avg_Util_negpt2326_pt00095 + Avg_Util_pt00096_pt0091 +
   Avg_Util_pt0092_pt3673 + Avg_Util_pt3674_pt8231 + Avg_Util_pt8232_6 +
   Bal_Growth_6mo_neg708323_neg122 + Bal_Growth_6mo_neg123_21389 +
   Bal_Growth_6mo_21390_428792 + Util_Growth_6mo_Neg5_NegPt7 +
   Util_Growth_6mo_NegPt7_NegPt00075 + Util_Growth_6mo_NegPt00075_0 +
   Util_Growth_6mo_0_Pt029 + Util_Growth_6mo_Pt029_2 + Max_Bill_Amt +
   Max_Pmt_Amt + Max_DLQ + Util_Chx1 + Util_Chx2 + Util_Chx3 + Util_Chx4 +
   Util_Chx5)

#variables identified as important by simple logistic regression
form2a <- as.formula(DEFAULT ~ SEX + EDUCATION +
                       MARRIAGE + PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
                       LIMIT_BAL_Neg_29999 + LIMIT_BAL_30000_159999 +
                       Avg_Bill_Amt_Neg_56050_188 +
                       Avg_Bill_Amt_189_2846 + Avg_Bill_Amt_2847_7488 +
                       Avg_Bill_Amt_7489_31915 + Avg_Pmt_Amt_0_2832 +
                       Avg_Pmt_Amt_2833_12092 + Max_Bill_Amt + Max_DLQ)

#with pay variables and only variables used to build the tree
form3 <- as.formula(DEFAULT ~ PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
      PAY_6 + Avg_Pmt_Ratio_Neg16429_pt02 + Max_DLQ)

#removed Max_DLQ
form4 <- as.formula(DEFAULT ~ PAY_1 + PAY_2 + PAY_3 + PAY_4 + PAY_5 +
      PAY_6 + Avg_Pmt_Ratio_Neg16429_pt02)

#top 10 variables from Random Forest full run
form5 <- as.formula(DEFAULT ~ PAY_1 + Max_Util + Max_Bill_Amt +
     Max_Pmt_Amt + Util_Chx1 + Util_Chx3 + Util_Chx2 + Util_Chx5 +
     Util_Chx4 + Max_DLQ)
```

```
#handpicked for SVM
form6 <- as.formula(DEFAULT ~ SEX + EDUCATION + MARRIAGE + PAY_1 + PAY_2 +
      PAY_3 + PAY_4 + PAY_5 +  LIMIT_BAL_Neg_29999 +
      LIMIT_BAL_30000_159999 + AGE_25_32 + Avg_Bill_Amt_Neg_56050_188 +
      Avg_Bill_Amt_189_2846 + Avg_Bill_Amt_2847_7488 +
      Avg_Bill_Amt_7489_31915 + Avg_Pmt_Amt_0_2832 +
      Avg_Pmt_Amt_2833_12092 + Avg_Pmt_Ratio_1_1pt17 +
      Avg_Util_pt0092_pt3673 + Util_Growth_6mo + Max_Bill_Amt + Max_DLQ +
      Bal_Growth_6mo_neg708323_neg122)


#variables identified as important by regsubsets
form7 <- as.formula(DEFAULT ~ PAY_1 + PAY_3 + PAY_5 + AGE_25_32 +
      Avg_Bill_Amt_Neg_56050_188 + Avg_Pmt_Amt_2833_12092 +
      Avg_Pmt_Ratio_Neg16429_pt02 + Avg_Util_negpt2326_pt00095 +
      Avg_Pmt_Ratio_1pt18_2688 + Bal_Growth_6mo_21390_428792)
```

Index 7: Confusion Matrices for all Models

RF Model 1

Model Random Forest
form, train data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1	TP	0.97	TP+TN	1.97	AUC	0.98
0	11,722	35	11,757		0	1.00	0.00	TN	1.00	Precision	0.99	Sensitivity	0.97
1	106	3,317	3,423		1	0.03	0.97	Type I Error	0.00	Recall	0.97	Specificity	1.00
								Type II Error	0.03	F1	0.98		

Model Random Forest
form, test data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1	TP	0.38	TP+TN	1.32	AUC	0.66
0	5,407	359	5,766		0	0.94	0.06	TN	0.94	Precision	0.63	Sensitivity	0.38
1	958	599	1,557		1	0.62	0.38	Type I Error	0.06	Recall	0.38	Specificity	0.94
								Type II Error	0.62	F1	0.53		

RF Model 2

Model Random Forest
form5, train data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1	TP	0.54	TP+TN	1.53	AUC	0.77
0	11,627	130	11,757		0	0.99	0.01	TN	0.99	Precision	0.93	Sensitivity	0.54
1	1,561	1,862	3,423		1	0.46	0.54	Type I Error	0.01	Recall	0.54	Specificity	0.99
								Type II Error	0.46	F1	0.70		

Model Random Forest
form5, test data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1	TP	0.36	TP+TN	1.31	AUC	0.65
0	5,447	319	5,766		0	0.94	0.06	TN	0.94	Precision	0.64	Sensitivity	0.36
1	992	565	1,557		1	0.64	0.36	Type I Error	0.06	Recall	0.36	Specificity	0.94
								Type II Error	0.64	F1	0.51		

Key: Cells in Yellow are metrics for models trained on training data. Cells in Green are metrics for models trained on testing data. The model formulations (form, form5) indicated here are defined in Index 7.

GB Model 1

Model Gradient Boost
form, train data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1						
0	11,248	2,260	13,508		0	0.83	0.17	Type I Error	0.17	Recall	0.70	Specificity	0.83
1	509	1,163	1,672		1	0.30	0.70	Type II Error	0.30	F1	0.75		

							TP	0.70	TP+TN	1.53	AUC	0.74
							TN	0.83	Precision	0.34	Sensitivity	0.70

Model Gradient Boost
form, test data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1						
0	5,494	1,012	6,506		0	0.84	0.16	Type I Error	0.16	Recall	0.67	Specificity	0.84
1	272	545	817		1	0.33	0.67	Type II Error	0.33	F1	0.73		

							TP	0.67	TP+TN	1.51	AUC	0.76
							TN	0.84	Precision	0.35	Sensitivity	0.67

GB Model 2

Model Gradient Boost
form5, training data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1						
0	11,294	2,318	13,612		0	0.83	0.17	Type I Error	0.17	Recall	0.70	Specificity	0.83
1	463	1,105	1,568		1	0.30	0.70	Type II Error	0.30	F1	0.75		

							TP	0.70	TP+TN	1.53	AUC	0.74
							TN	0.83	Precision	0.32	Sensitivity	0.70

Model Gradient Boost
form5, test data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1						
0	5,518	1,038	6,556		0	0.84	0.16	Type I Error	0.16	Recall	0.68	Specificity	0.84
1	248	519	767		1	0.32	0.68	Type II Error	0.32	F1	0.74		

							TP	0.68	TP+TN	1.52	AUC	0.76
							TN	0.84	Precision	0.33	Sensitivity	0.68

LR Model 1

Model Logistic Regression
form7, training data

Actual Class	Predicted Class 0	Predicted Class 1	Totals
0	11,220	2,246	13,466
1	537	1,177	1,714

Actual Class	Predicted Class 0	Predicted Class 1							
0	0.83	0.17	Type I Error	0.17	Recall	0.69	Specificity	0.83	
1	0.31	0.69	Type II Error	0.31	F1	0.74			

TP	0.69	TP+TN	1.52	AUC	0.65
TN	0.83	Precision	0.34	Sensitivity	0.69

Model Logistic Regression
form7, test data

Actual Class	Predicted Class 0	Predicted Class 1	Totals
0	5,477	1,004	6,481
1	289	553	842

Actual Class	Predicted Class 0	Predicted Class 1							
0	0.85	0.15	Type I Error	0.15	Recall	0.66	Specificity	0.85	
1	0.34	0.66	Type II Error	0.34	F1	0.73			

TP	0.66	TP+TN	1.50	AUC	0.65
TN	0.85	Precision	0.36	Sensitivity	0.66

LR Model 2

Model Logistic Regression
form5, train data

Actual Class	Predicted Class 0	Predicted Class 1	Totals
0	11,288	2,309	13,597
1	469	1,114	1,583

Actual Class	Predicted Class 0	Predicted Class 1							
0	0.83	0.17	Type I Error	0.17	Recall	0.70	Specificity	0.83	
1	0.30	0.70	Type II Error	0.30	F1	0.75			

TP	0.70	TP+TN	1.53	AUC	0.64
TN	0.83	Precision	0.33	Sensitivity	0.70

Model Logistic Regression
form5, test data

Actual Class	Predicted Class 0	Predicted Class 1	Totals
0	5,513	1,032	6,545
1	253	525	778

Actual Class	Predicted Class 0	Predicted Class 1							
0	0.84	0.16	Type I Error	0.16	Recall	0.67	Specificity	0.84	
1	0.33	0.67	Type II Error	0.33	F1	0.74			

TP	0.67	TP+TN	1.52	AUC	0.65
TN	0.84	Precision	0.34	Sensitivity	0.67

SVM Model 1:

Model SVM
form, train data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1	TP	0.74	TP+TN	1.57	AUC	0.65
0	11,332	2,242	13,574		0	0.83	0.17	TN	0.83	Precision	0.35	Sensitivity	0.74
1	425	1,181	1,606		1	0.26	0.74	Type I Error	0.17	Recall	0.74	Specificity	0.83
								Type II Error	0.26	F1	0.77		

Model SVM
form, train data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1	TP	0.66	TP+TN	1.50	AUC	0.65
0	5,497	1,029	6,526		0	0.84	0.16	TN	0.84	Precision	0.34	Sensitivity	0.66
1	269	528	797		1	0.34	0.66	Type I Error	0.16	Recall	0.66	Specificity	0.84
								Type II Error	0.34	F1	0.73		

SVM Model 2

Tuned Model SVM
form, train data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1	TP	0.68	TP+TN	1.51	AUC	0.66
0	11,151	2,162	13,313		0	0.84	0.16	TN	0.84	Precision	0.37	Sensitivity	0.68
1	606	1,261	1,867		1	0.32	0.68	Type I Error	0.16	Recall	0.68	Specificity	0.84
								Type II Error	0.32	F1	0.74		

Tuned Model SVM
form, test data

Actual Class	Predicted Class		Totals		Actual Class	Predicted Class							
	0	1				0	1	TP	0.64	TP+TN	1.49	AUC	0.66
0	5,444	966	6,410		0	0.85	0.15	TN	0.85	Precision	0.38	Sensitivity	0.64
1	332	591	923		1	0.36	0.64	Type I Error	0.15	Recall	0.64	Specificity	0.85
								Type II Error	0.36	F1	0.72		

SVM Model 3

Model #4: SVM
form6, train data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1						
0	11,241	2,253	13,494		0	0.83	0.17	TP	0.69	TP+TN	1.53	AUC	0.65
1	516	1,170	1,686		1	0.31	0.69	TN	0.83	Precision	0.34	Sensitivity	0.69
								Type I Error	0.17	Recall	0.69	Specificity	0.83
								Type II Error	0.31	F1	0.75		

Model #4: SVM
form6, test data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1						
0	5,485	1,009	6,494		0	0.84	0.16	TP	0.66	TP+TN	1.51	AUC	0.66
1	281	548	829		1	0.34	0.66	TN	0.84	Precision	0.35	Sensitivity	0.66
								Type I Error	0.16	Recall	0.66	Specificity	0.84
								Type II Error	0.34	F1	0.73		

SVM Model 4

Tuned Model SVM
form5, train data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1						
0	11,309	2,327	13,636		0	0.83	0.17	TP	0.71	TP+TN	1.54	AUC	0.64
1	448	1,096	1,544		1	0.29	0.71	TN	0.83	Precision	0.32	Sensitivity	0.71
								Type I Error	0.17	Recall	0.71	Specificity	0.83
								Type II Error	0.29	F1	0.75		

Tuned Model SVM
form5, test data

Actual Class	Predicted Class 0	Predicted Class 1	Totals		Actual Class	Predicted Class 0	Predicted Class 1						
0	5,517	1,043	6,560		0	0.84	0.16	TP	0.67	TP+TN	1.51	AUC	0.64
1	249	514	763		1	0.33	0.67	TN	0.84	Precision	0.33	Sensitivity	0.67
								Type I Error	0.16	Recall	0.67	Specificity	0.84
								Type II Error	0.33	F1	0.74		

Index 10: List of Continuous Variables Used for Principal Component Analysis

```
> str(sub_cont_var)
'data.frame'        :15180 obs. of  25 variables:
 $ LIMIT_BAL        : int  20000 90000 50000 20000 260000 250000 20000 320000 360000 50000 ...
 $ AGE              : int  24 34 37 35 51 29 24 49 49 47 ...
 $ BILL_AMT1        : int  3913 29239 64400 0 12261 70887 15376 253286 0 650 ...
 $ BILL_AMT2        : int  3102 14027 57069 0 21670 67060 18010 246536 0 3415 ...
 $ BILL_AMT3        : int  689 13559 57608 0 9966 63561 17428 194663 0 3416 ...
 $ BILL_AMT4        : int  0 14331 19394 0 8517 59696 18338 70074 0 2040 ...
 $ BILL_AMT5        : int  0 14948 19619 13007 22287 56875 17905 5856 0 30430 ...
 $ BILL_AMT6        : int  0 15549 20024 13912 13668 55512 19104 195599 0 257 ...
 $ PAY_AMT1         : int  0 1518 2500 0 21818 3000 3200 10358 0 3415 ...
 $ PAY_AMT2         : int  689 1500 1815 0 9966 3000 0 10000 0 3421 ...
 $ PAY_AMT3         : int  0 1000 657 0 8583 3000 1500 75940 0 2044 ...
 $ PAY_AMT4         : int  0 1000 1000 13007 22301 3000 0 20000 0 30430 ...
 $ PAY_AMT5         : int  0 1000 1000 1122 0 3000 1650 195599 0 257 ...
 $ PAY_AMT6         : int  0 5000 800 0 3640 3000 0 50000 0 0 ...
 $ DEFAULT          : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 2 1 1 1 ...
 $ Avg_Bill_Amt             : num  1284 16942 39686 4486 14728 ...
 $ Avg_Pmt_Amt              : num  115 1836 1295 2355 11051 ...
 $ Avg_Pmt_Ratio            : num  60.2 0.084 0.042 60.216 0.803 ...
 $ Avg_Util                 : num  0.0642 0.1882 0.7937 0.2243 0.0566 ...
 $ Bal_Growth_6mo           : int  -3913 -13690 -44376 13912 1407 -15375 3728 -57687 0 -393 ...
 $ Util_Growth_6mo          : num  -0.19565 -0.15211 -0.88752 0.6956 0.00541 ...
 $ Max_Bill_Amt             : int  3913 29239 64400 13912 22287 70887 19104 253286 0 30430 ...
 $ Max_Pmt_Amt              : int  689 5000 2500 13007 22301 3000 3200 195599 0 30430 ...
 $ Max_DLQ                  : num  2 -1 -1 -1 2 -1 2 -1 1 -1 ...
 $ Max_Util                 : num  0.1956 0.3249 1.288 0.6956 0.0857 ...
```