PROJECT

PREDICTING BAD CUSTOMERS IN CREDIT DATASET

WISDOM TAKUMAH

Table of Contents:

Tables:

Figures:

Description of Problem

The goal of this project was to predict "Bad" customers. The dataset consisted of 91,502 observations (for 9,997 customers) and 26 variables. We were provided up to 10 rows per customer: one per month, most beginning in February 2010 and ending November 2010. Over that time, the customer would incur a balance and either make payments or fail to make payments. If a customer failed to make payments, their account would become delinquent. Most of the delinquency were in increments of 30 days. Often, accounts that became 90 days or more delinquent would be given one of the following external statuses: E for revoked, F for Frozen, I for Interest Prohibited, and Z for Charged Off. Some accounts do have a C for Closed. For purposes of our analysis, closed was not considered a "Bad" account as the customer may have closed without becoming delinquent.

We were asked to determine who was a good or bad customer (by the behavior shown in the time frame given) and learn how to predict when bad behavior was forthcoming. We began by exploring the data.

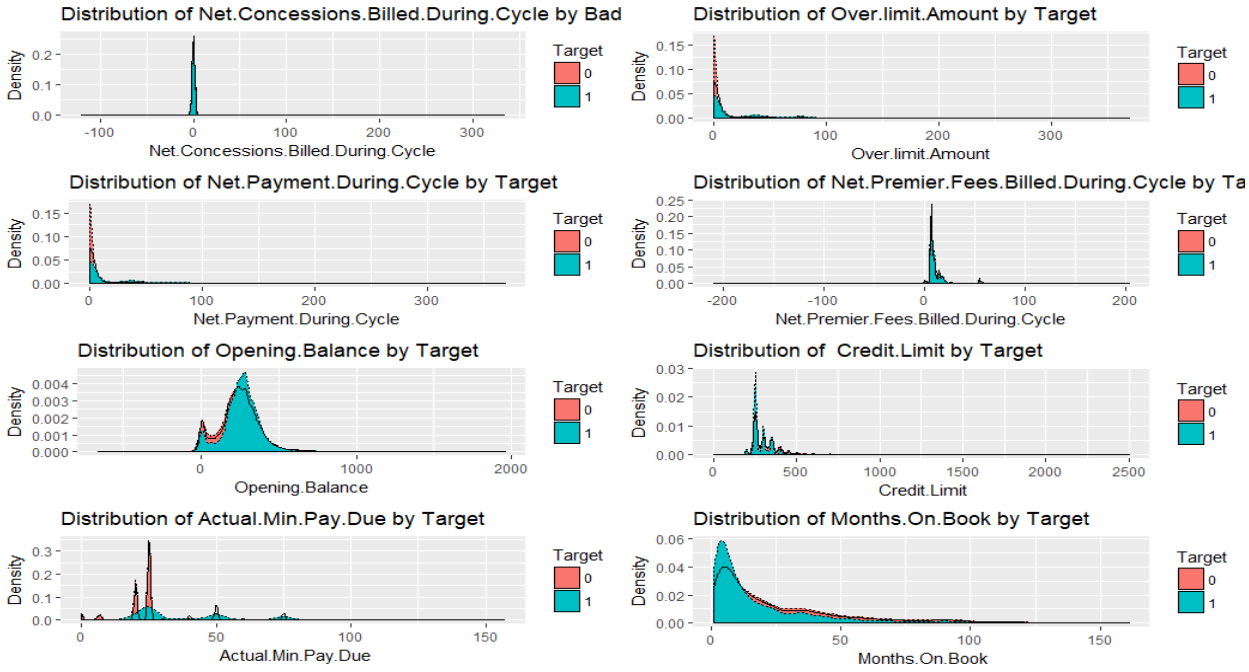
Descriptive Statistics of Numerical Variables

Table 1 contains summary of several notable numerical variables included in the model. These variables are selected based on minimum BIC values. This data includes information such as payments made during the previous cycle.

Table 1: Descriptive Statistics of Numerical Variable

Variables	Min	1st Quart	Median	Mean	3rd Quart	Max
Opening Balance	-665.6	179.00	246.30	247.70	320.20	1959.30
Credit Limit	0.00	250.00	300.00	321.50	350.00	2500.00
Over Limit Amount	0.00	0.00	0.00	11.11	0.00	368.28
Actual Minimum Pay Due	0.00	20.00	25.00	27.75	25.00	157.00
Total Minimum Pay Due	0.00	20.00	25.00	38.84	30.86	443.28
Net Purchases During Cycle	106.82	0.00	5.310	46.86	59.47	767.11
Net Payment During Cycle	-665.60	20.00	36.02	71.21	91.00	1077.80
Net Premier Fees Billed During Cycle	-209.89	7.00	7.00	10.58	12.00	203.00
Net Behavior Fees Billed During Cycle	-60.940	1.250	3.710	8.408	6.800	104.490
Net Cash Advances During Cycle	0.00	00.00	00.00	0.826	00.00	350.00
Net Concessions Billed During Cycle	-119.92	00.00	00.00	1.290	00.00	334.27

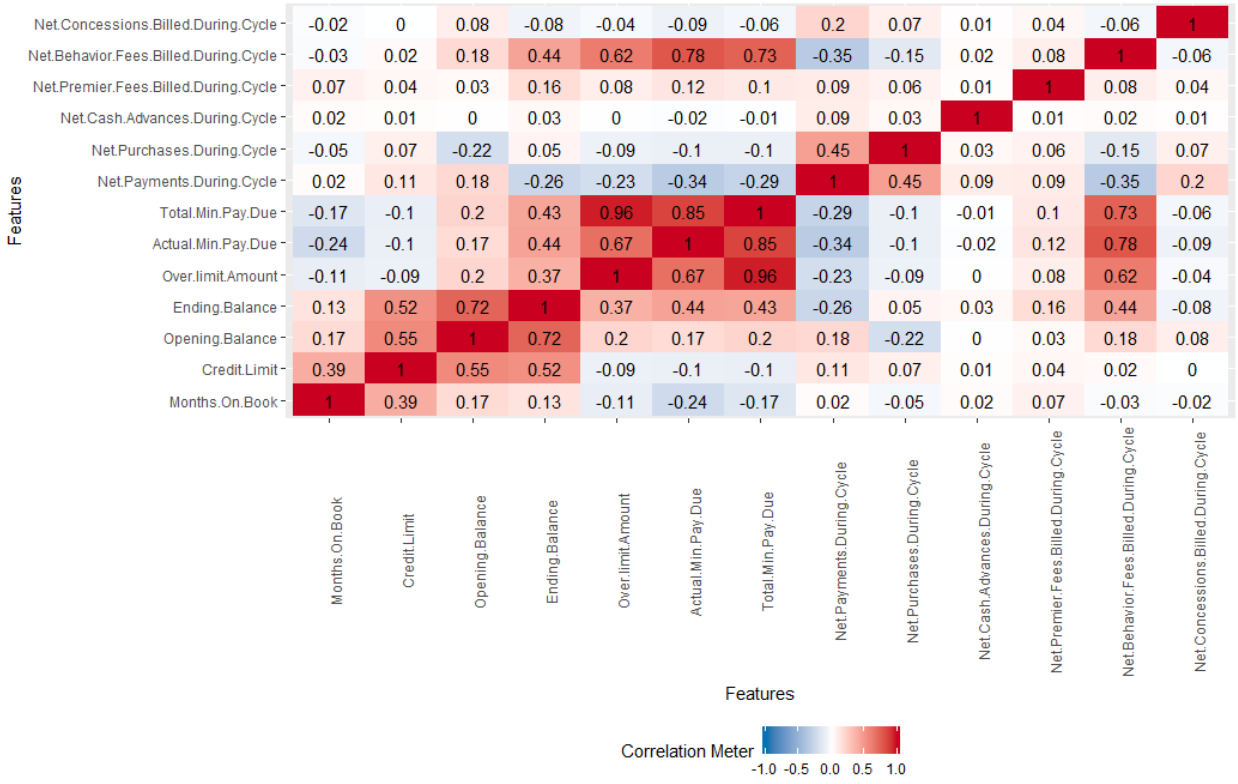
Figure 1: Density Plot of Continuous Variables VS Target



The Distribution of *Net Payments During Cycle* and *Over Limit* Amount by *Target* seem to be right skewed. This is an indication that there is a poor fit in these two variables against our target variable.

A fundamental method of exploratory data analysis is to find a relationship between different attributes in a dataset. Based on this, we also produce the correlation matrix of our data. Figure 3 illustrates this.

Figure 2: Correlation Matrix of selected variables



From the correlation matrix, we observe some high and positive as well as negative correlation among some variables. When we look at *Actual Minimum Pay Due vs Total Minimum Pay Due*, there is a very strong positive correlation between them with a correlation coefficient of 0.85; this tells us that as *Total Minimum Pay Due* increases, *Actual Minimum Pay Due* also increases. Also, we observed very high correlation between *Total Minimum Pay Due* and *Over Limit Amount*, with a coefficient of 0.96. We also observe a strong and positive correlation with *Net Behavior Fees Billed During Cycle vs Days Delinquent* (a correlation coefficient of 0.82) and with *Net Behavior*

Fees Billed During Cycle vs Actual Minimum Pay Due (0.78). This gives us indication that we probably do not need to include each of these variables in our model.

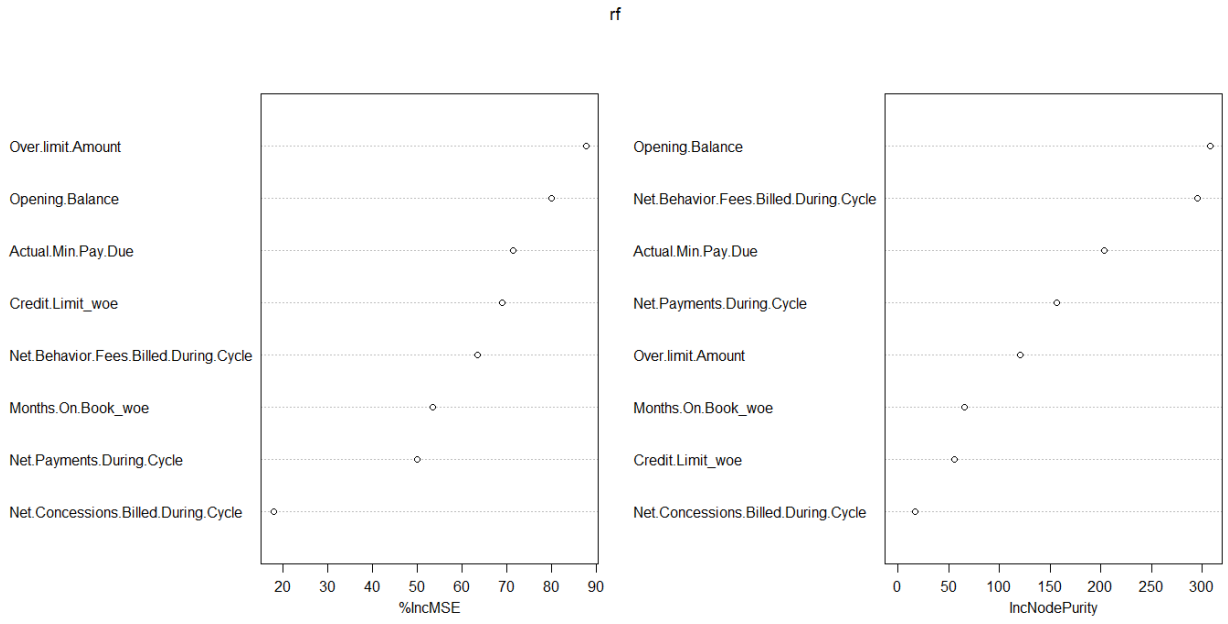
Model Building

Two classifiers were used in this project: logistic model and the random forest model. We compared these models to each other to choose a preferred method.

Random Forest Model

I first split our data into training and testing, with training data containing 60% of customers. Then, I performed validation set approach on the training and testing data. Figure 3 shows the results of variable importance selected by the random forest model. It is observed from the tree diagram that *Over Limit Amount* is the most important variable in predicting *Bad* in terms of percentage MSE. But considering Node Impurity *Opening Balance* appears to be the most important variable in predicting a bad customer.

Figure 3: Random Forest Variable Importance



Evaluating Random Forest Model

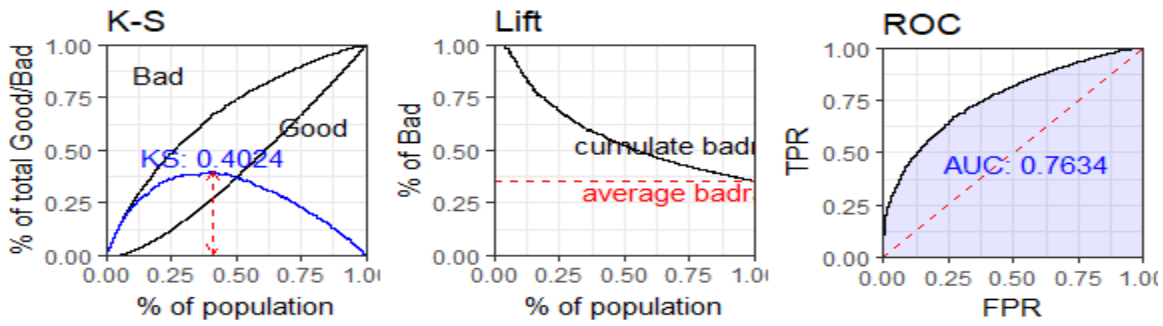
After fitting the random forest model, we decided to evaluate it to check whether the model is working well or not. The following approaches were used in evaluating it (tree model):

- ROC Curve
- KS Statistics
- Lift Plot and Gains table

ROC Curve

An ROC curve, which stands for Receiver Operating Characteristic curve, is a graph that shows the tradeoff between the sensitivity and specificity in any given model. A bad graph would follow that diagonal line and would show that the more specific our model is, the less sensitivity and thus, less effective it is.

Figure 4: KS plot, ROC plots and Lift Plot for Random Forest Model



The plot of the ROC curve shows the false positive rate compared with the true positive rate.

It is observed that the ROC for testing model is 0.7634 and the KS Statistics is given as 0.4024.

Table 2: Gains Table for Random Forest Model

Depth of file	N	Cum N	Mean Response	Cum Mean Response	Cum Pct on total Response	Lift Index	Cum Lift	Mean Model Score
10	399	399	0.89	0.89	25.10%	252	252	0.93
20	400	799	0.6	0.74	42.10%	170	211	0.68
30	400	1199	0.46	0.65	55.20%	130	184	0.51
40	400	1599	0.39	0.58	66.20%	110	165	0.39
50	400	1999	0.31	0.53	75.00%	88	150	0.31
60	399	2398	0.24	0.48	81.80%	69	136	0.24
70	400	2798	0.18	0.44	87.00%	51	124	0.19
80	400	3198	0.22	0.41	93.10%	62	116	0.14
90	400	3598	0.14	0.38	97.00%	39	108	0.09
100	400	3998	0.1	0.35	100.00%	30	100	0.04

The gains table for the random forest model (Table 2) highly reflects the structure of the tree. We are able to group the bottom 10% (first row) in a group which is 89% Target (Bad customer). The mean response shows that each increasing group has a lower percentage of the bad customers. The lift index of 252 for the first group means that our tree gives us a lift of 2.52 for our first group, meaning that our model discovers 2.52 times as many Bad customers in the first group than if we randomly selected 399 of them.

Logistic Regression Model

Before building the logistic regression model, we did variable selection based on minimum BIC value and the selected variables were then used in the logistic regression model. Weight of Evidence was performed on two predictors (*Months on Book and Credit Limit*) and later split our data into training and testing, with training data containing 60% of customers. A validation Set Approach was adopted by fitting the model on the training data and validation performed on the testing data. This summary of the model fit on training data is presented below.

Table 3: Summary of Logistics Model

Parameters	Estimate	Std. Error	Z value	Prob. Value	
(Intercept)	-2.69727	0.12696	-21.245	0.0000	***
Opening Balance	0.002377	0.000299	7.949	0.0000	***
Over limit Amount	0.012541	0.001912	6.559	0.0000	***
Actual Min Pay Due	0.039424	0.00448	8.801	0.0000	***
Net Payments During Cycle	0.00129	0.000353	3.653	0.0003	***
Net Behavior Fees Billed During Cycle	0.017272	0.004242	4.072	0.0000	***
Net Concessions Billed During Cycle	0.009181	0.003119	2.944	0.0032	**
Months on Book_(woe)	-0.562438	0.070123	-8.021	0.0000	***
Credit Limit_(woe)	-1.07066	0.085944	-12.458	0.0000	***

It can be observed that, not all variables in the table are significant at 5% level of significance. The significant variables include Opening Balance, Over Limit Amount, Actual Min Pay Due, Net Payments During Cycle, Net Behavior Fees Billed During Cycle, Net Concessions Billed During Cycle, Credit Limit (woe) and Months on Book (woe). It was found that a unit increase in Net Payments During Cycle will increase the expected log odds of a being a bad customer by 0.001.

Also, a unit increase in Net Behavior Fees Billed During Cycle would increase the expected log odds of being a bad customer by 0.017. Furthermore, a unit increase in Net Concessions Billed During Cycle would increase the expected log odds of being a bad customer by 0.009. For Opening Balance and Actual Min Pay Due, a unit increase in both will increase the expected odd of being a bad customer by 0.002 and 0.013 respectively. The expected log odds of being a bad customer decrease by 0.56 and 1.07 respectively when there is a unit increase in the weight of evidence of both Months on Book and Credit Limit.

Evaluating Logistic Model

The logistic model was evaluated using KS statistics, ROC, Lift plot and Gains plot.

Figure 5: KS plot, ROC Curve and Lift plot for Logistics Model (Train)

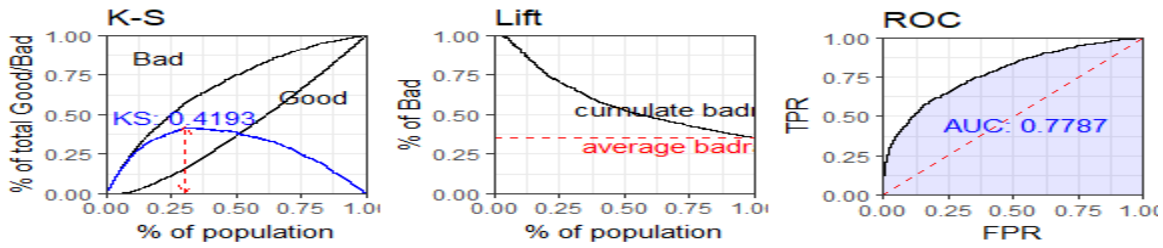
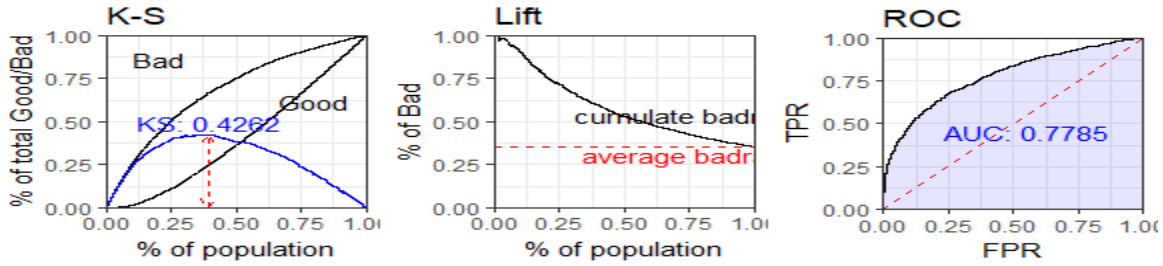


Figure 6: KS plot, ROC Curve and Lift plot for Logistics Model (test)



The graphs above shows there is no difference between ROC for training and testing data sets. The AUC values for training and testing model are basically the same (0.7787 for training and 0.7785 for testing), implying the model is performing well. The KS statistics are also close for training and testing data (0.4163 for training and 0.4262 for testing)

Table 4: Gains Table for Logistic Model

Depth of file	N	Cum N	Mean Response	Cum Mean Response	Cum Pct on total Response	Lift Index	Cum Lift	Mean Model Score
10	399	399	0.9	0.9	25.50%	255	255	0.89
20	400	799	0.61	0.76	42.90%	174	215	0.61
30	400	1199	0.48	0.67	56.70%	138	189	0.43
40	400	1599	0.34	0.58	66.20%	96	166	0.36
50	400	1999	0.3	0.53	74.90%	86	150	0.31
60	399	2398	0.27	0.49	82.60%	77	138	0.27
70	400	2798	0.21	0.45	88.70%	60	127	0.22
80	400	3198	0.18	0.41	93.80%	52	117	0.18
90	400	3598	0.12	0.38	97.40%	35	108	0.13
100	400	3998	0.09	0.35	100.00%	26	100	0.07

The gains table for the logistic model (Table 4) is presented above. I grouped the bottom 10% (first row) in a group which is 90% Target (Bad customer). The mean response shows that each increasing group has a lower percentage of the bad customers. The lift index of 255 for the first group means that our tree gives us a lift of 2.55 for our first group, meaning that our model discovers 2.55 times as many Bad customers in the first group than if we randomly selected 399 of them.

Table 5: Good Customer Score Model

Parameters	Estimate	Std. Error	t value	Prob	
(Intercept)	4.780	9.10E-02	52.542	0.000	***
Good Customer Score	-0.0002	4.15E-05	-4.795	0.000	***
Behavior Score	-0.00667	1.57E-04	-42.404	0.000	***

From the Table 5 above, both Good Customer Score and Behavior Score are significant in the m odel at 0.05 level of significance. A unit increase in Good Customer Score will decrease the odds of being a bad customer by 0.0002, while an increase in behavior score will also decrease the expected log odds for being a bad customer by 0.00667

Figure 7: KS, ROC and Lift plot for Good Customer Score and Behavior Score Model(Train)

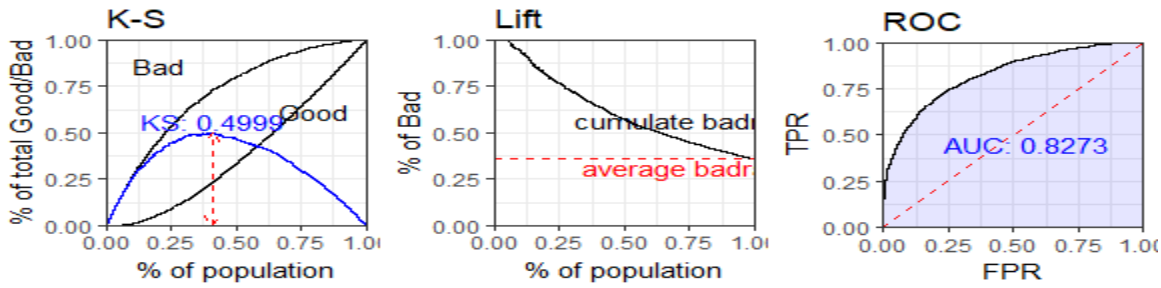
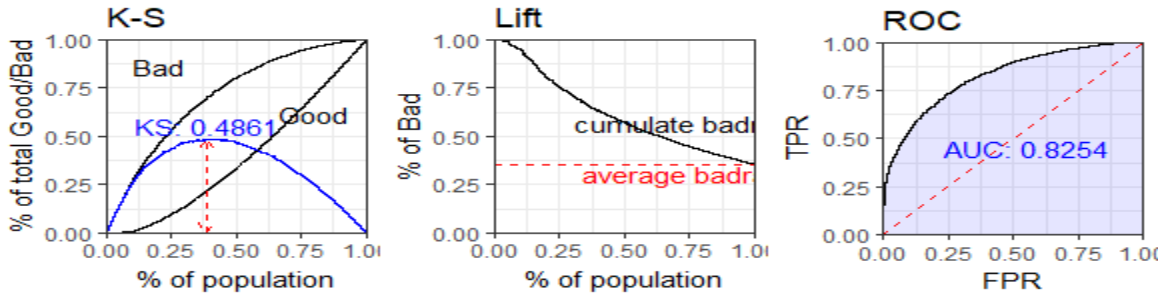


Figure 8: KS, ROC and Lift plot for Good Customer Score and Behavior Score Model(Test)



Based on the values of both KS and AUC, we can conclude that the model is working well. This is because there is no difference between the KS statistics for training model (0.499) and testing (0.4861). Also, the AUC for training (0.8273) is close to that of testing (0.8254).

Table 6: Gains Table for Good Customer Score and Behavior Score Model

Depth of file	N	Cum N	Mean Response	Cum Mean Response	Cum Pct on total Response	Lift Index	Cum Lift	Mean Model Score
10	399	399	0.63	0.63	17.90%	180	180	NA
20	400	799	0.78	0.71	40.10%	222	201	0.71
30	400	1199	0.6	0.67	57.20%	170	191	0.54
40	400	1599	0.44	0.61	69.70%	125	174	0.43
50	400	1999	0.32	0.56	78.80%	91	158	0.36
60	399	2398	0.26	0.51	86.00%	72	143	0.29
70	400	2798	0.2	0.46	91.80%	58	131	0.23
80	400	3198	0.15	0.42	96.10%	43	120	0.17
90	400	3598	0.09	0.39	98.70%	26	110	0.09
100	400	3998	0.04	0.35	100.00%	13	100	-0.03

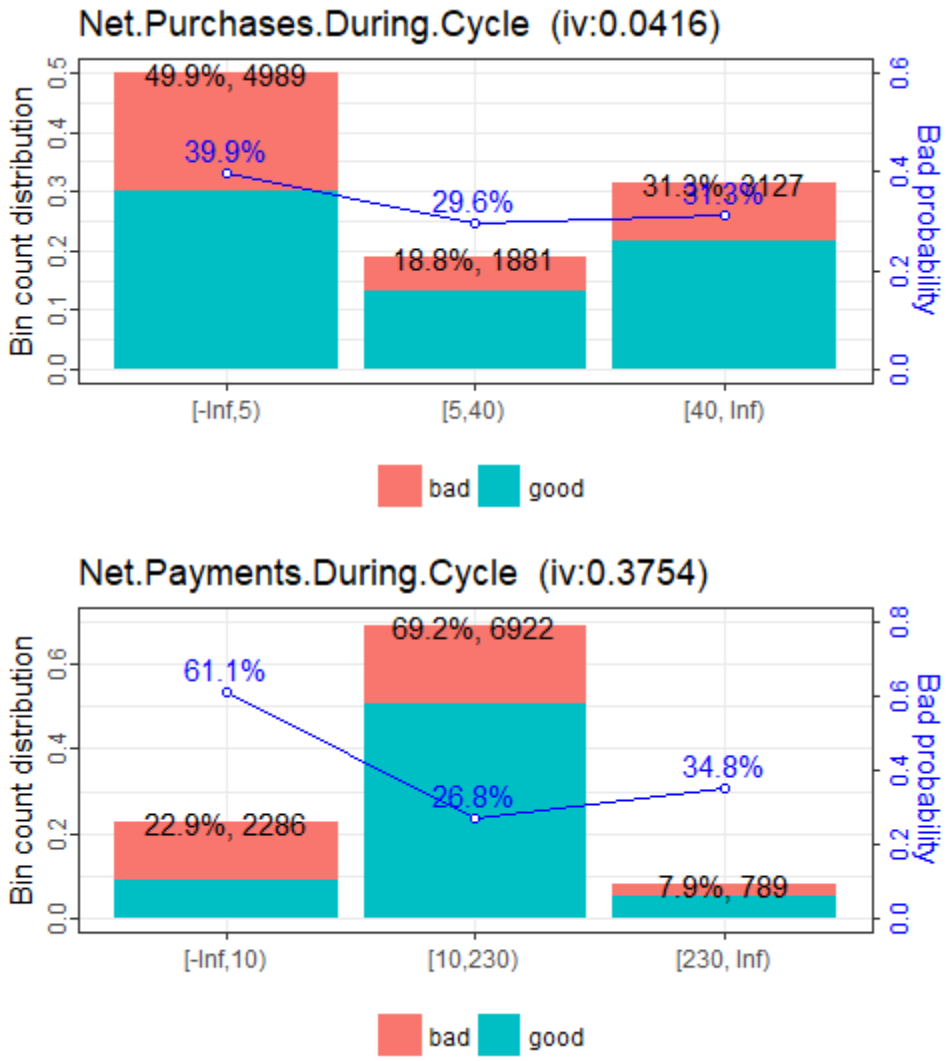
The gains table for the Good Customer Score and Behavior Score Model (Table 6) shows that the bottom 10% (first row) in a group is 63% Target (Bad customer). The lift index of 180 for the first group means that our tree gives us a lift of 1.80 for our first group, meaning that our model discovers 1.80 times as many Bad customers in the first group than if we randomly selected 399 of them.

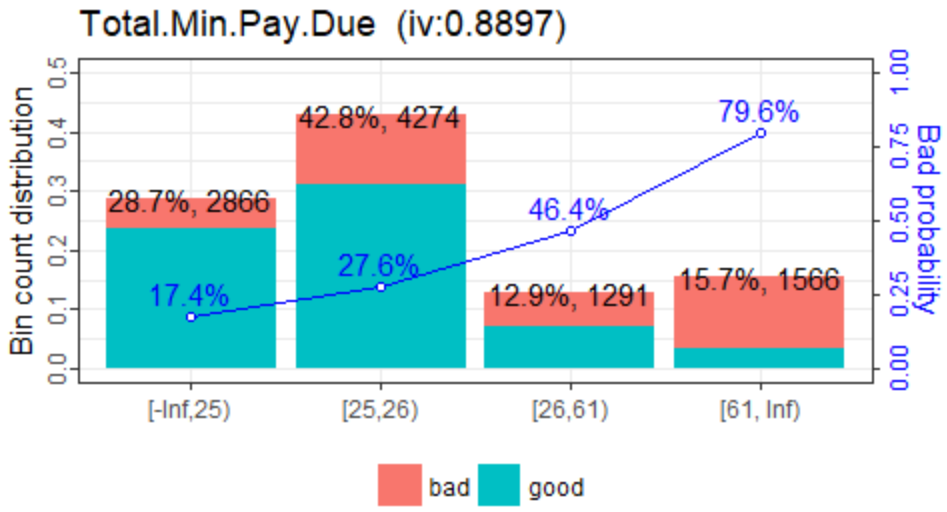
Conclusion

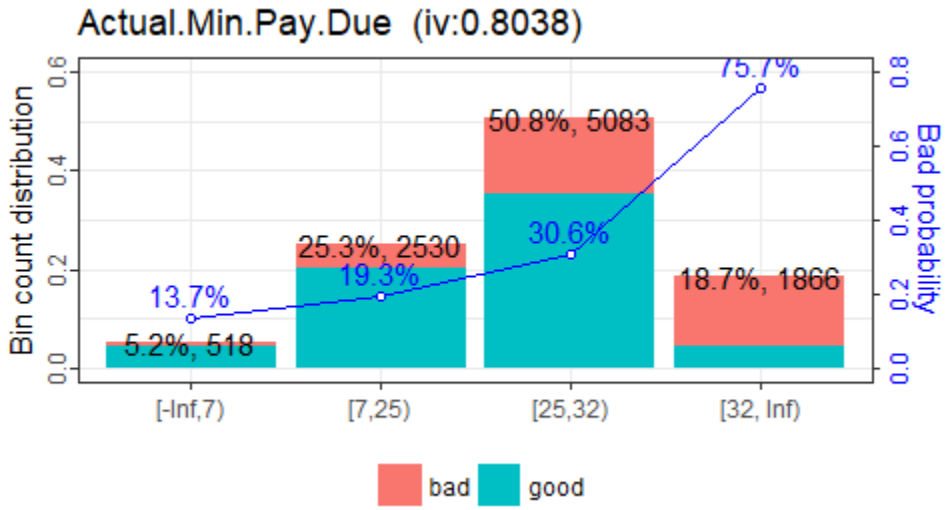
In conclusion, the random forest model performs better at predicting bad customers than the logistics model. The significant variables for the logistic model are Opening Balance, Over Limit Amount, Actual Min Pay Due, Net Payments During Cycle, Net Behavior Fees Billed During Cycle, Net Concessions Billed During Cycle and Credit Limit_woe and Amounts on Book_woe. Higher Opening Balance, Over Limit Amount, Actual Min Pay Due, Net Payments During Cycle, Net Behavior Fees Billed During Cycle, and Net Concessions Billed During Cycle indicate a

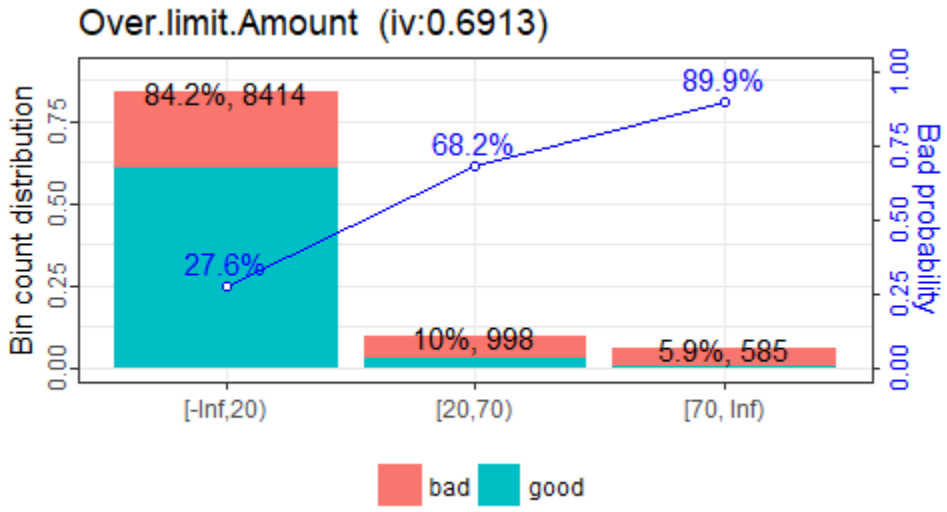
higher probability of an account becoming “Bad”. Good Customer Score model perform well than the Logistics and the Rando Forest Model

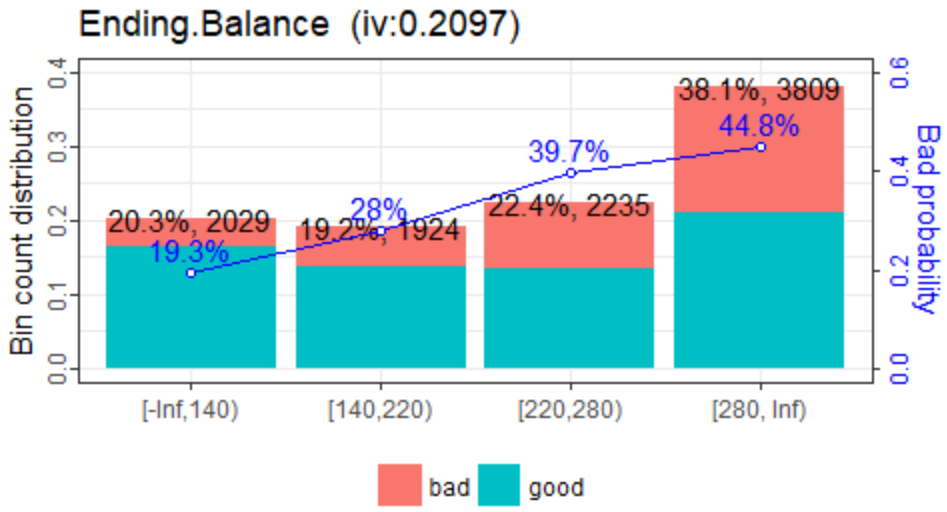
Figure 9: Bad Rate Plots of Bin Variables

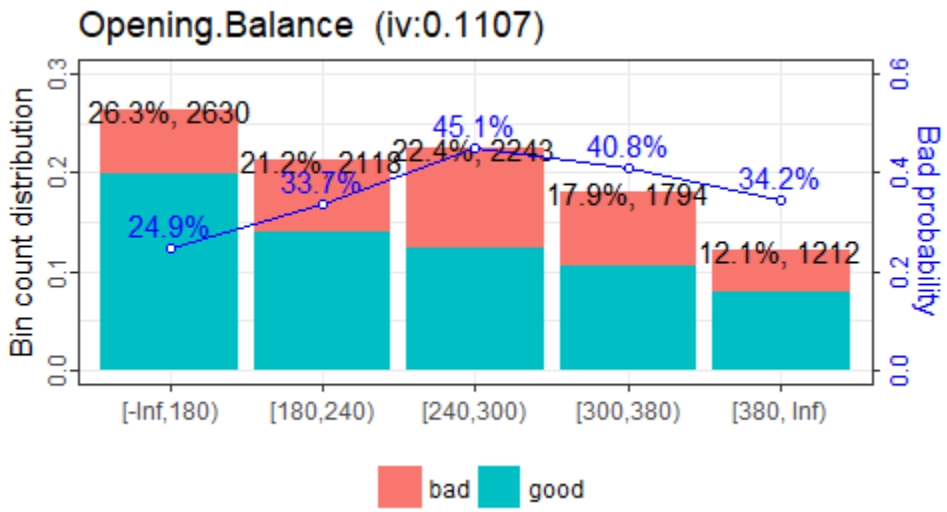


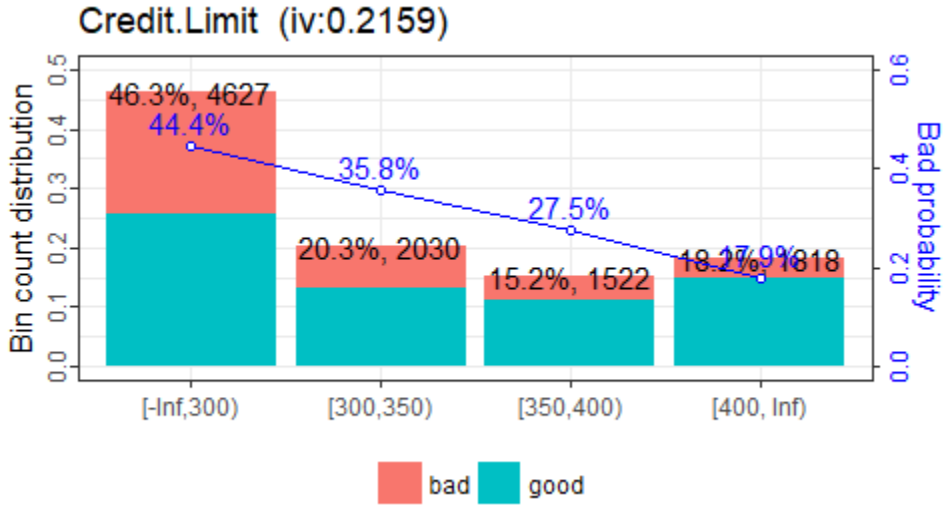

Total.Min.Pay.Due (iv:0.8897)
Bin count distribution
Bad probability
28.7%, 2866
42.8%, 4274
12.9%, 1291
15.7%, 1566
17.4%
27.6%
46.4%
79.6%
[-Inf,25)
[25,26)
[26,61)
[61, Inf)
bad
good

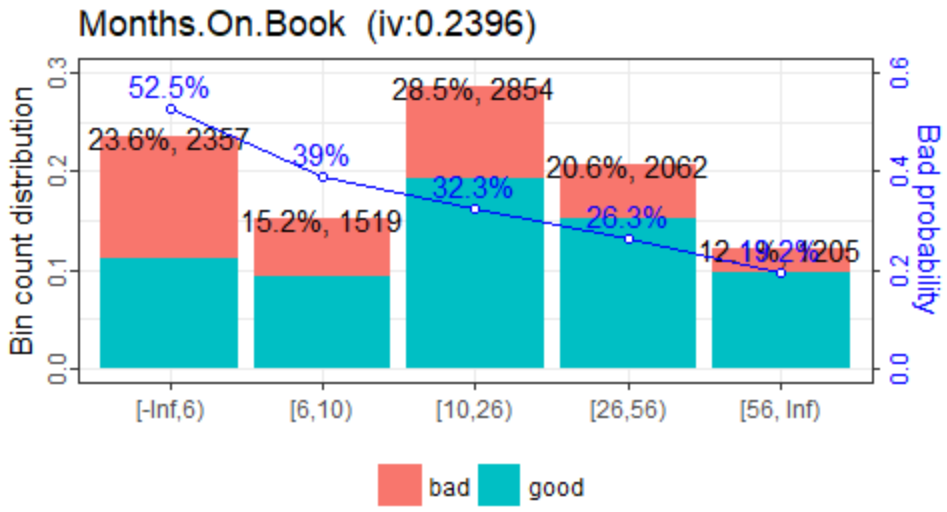

Actual.Min.Pay.Due (iv:0.8038)
Bin count distribution
Bad probability
5.2%, 518
25.3%, 2530
50.8%, 5083
18.7%, 1866
13.7%
19.3%
30.6%
75.7%
[-Inf,7)
[7,25)
[25,32)
[32, Inf)
bad
good

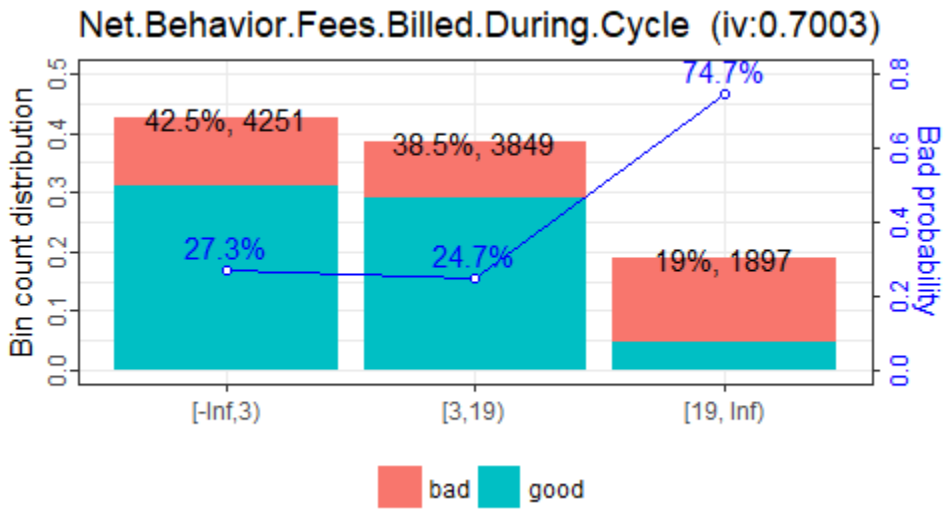

Over.limit.Amount (iv:0.6913)
Bin count distribution
Bad probability
84.2%, 8414
10%, 998
5.9%, 585
27.6%
68.2%
89.9%
[-Inf,20)
[20,70)
[70, Inf)
bad
good


Ending.Balance (iv:0.2097)
Bin count distribution
Bad probability
20.3%, 2029
19.3%
19.2%, 1924
28%
22.4%, 2235
39.7%
38.1%, 3809
44.8%
[-Inf,140)
[140,220)
[220,280)
[280, Inf)
bad
good


Opening.Balance (iv:0.1107)
Bin count distribution
Bad probability
26.3%, 2630
24.9%
21.2%, 2118
33.7%
22.4%, 2243
45.1%
17.9%, 1794
40.8%
12.1%, 1212
34.2%
[-Inf,180)
[180,240)
[240,300)
[300,380)
[380, Inf)
bad
good


Credit.Limit (iv:0.2159)
Bin count distribution
Bad probability
46.3%, 4627
44.4%
20.3%, 2030
35.8%
15.2%, 1522
27.5%
18.2%, 1818
17.9%
[-Inf,300)
[300,350)
[350,400)
[400, Inf)
bad
good


Months.On.Book (iv:0.2396)
Bin count distribution
Bad probability
52.5%
23.6%, 2357
39%
15.2%, 1519
28.5%, 2854
32.3%
20.6%, 2062
26.3%
12.1%, 1205
[-Inf,6)
[6,10)
[10,26)
[26,56)
[56, Inf)
bad
good


Net.Behavior.Fees.Billed.During.Cycle (iv:0.7003)
Bin count distribution
Bad probability
42.5%, 4251
27.3%
38.5%, 3849
24.7%
74.7%
19%, 1897
[-Inf,3)
[3,19)
[19, Inf)
bad
good