Analysis of Credit Risk Data to predict if the customer will go ‘Bad’ or Not!

Pramisha Thapaliya
Md Mominul Islam
Sakib Faisal
Tahmid Alam

Dataset and Introduction

- Initial data dimension: 18,987 observations with 30 variables
- Final data dimension: 14,289 observations with 15 variables, including 3 categorical, 11 numeric and one target.

Field Name	Usage	Description
Debt to Income Ratio	Input	Total monthly debt payments divided by monthly income
Is Borrower Homeowner	Input	Is the Borrower a Homeowner?
Amount Borrowed	Information Only	Loan Amount
Current Delinquencies, Delinquencies last 7 years	Input	Number of accounts delinquent at time of loan application
Revolving Credit Balance	Input	Revolving credit is credit card debt.
Bank Card Utilization	Input	Total credit card balance on all cards divided by total credit line on all cards
Employment Status	Input	
Income	Input	
Bad	Target	1= Bad and 0=Good

Data Reformatting

- -Converted missing values into NA and visualized
- -Removed performance variables, ID column variables and unneeded variables

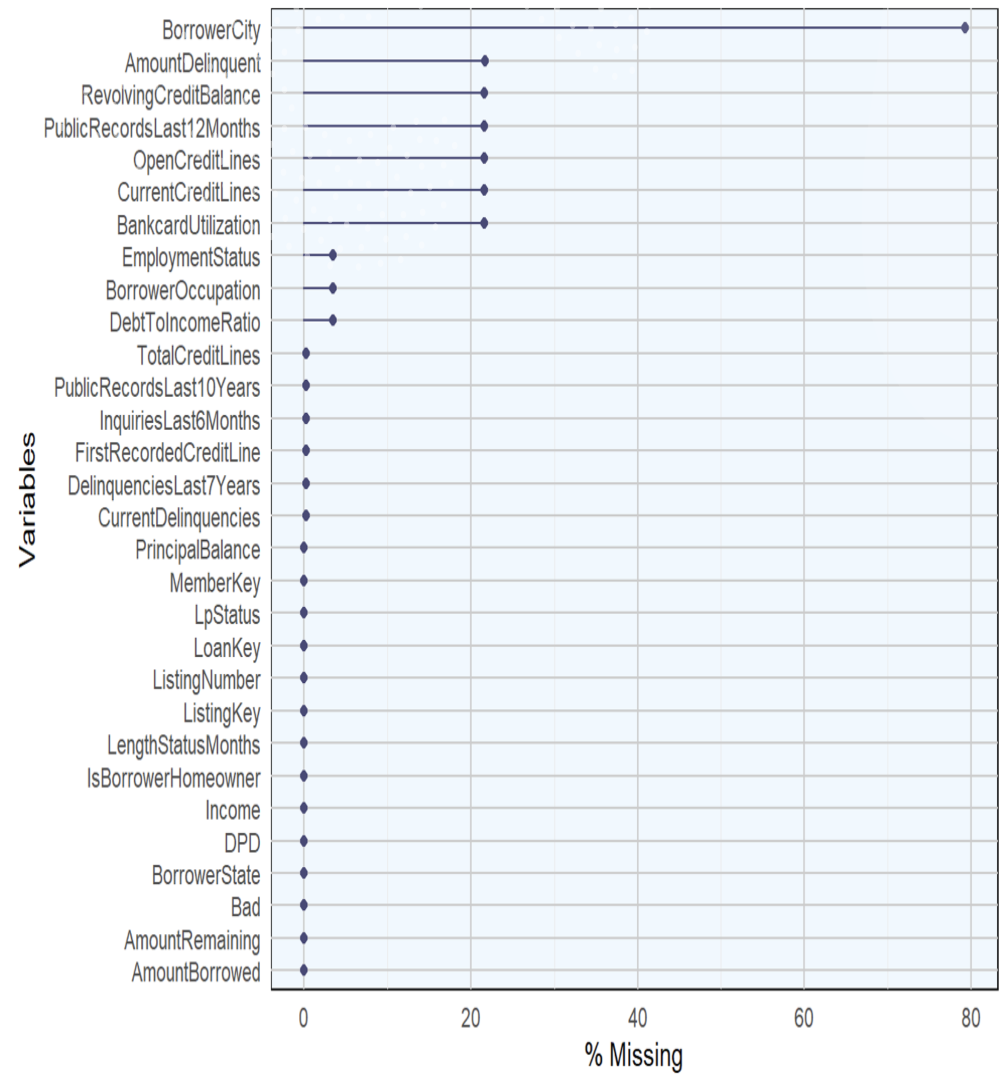


Figure 1. Percentage of missing data present in the variables

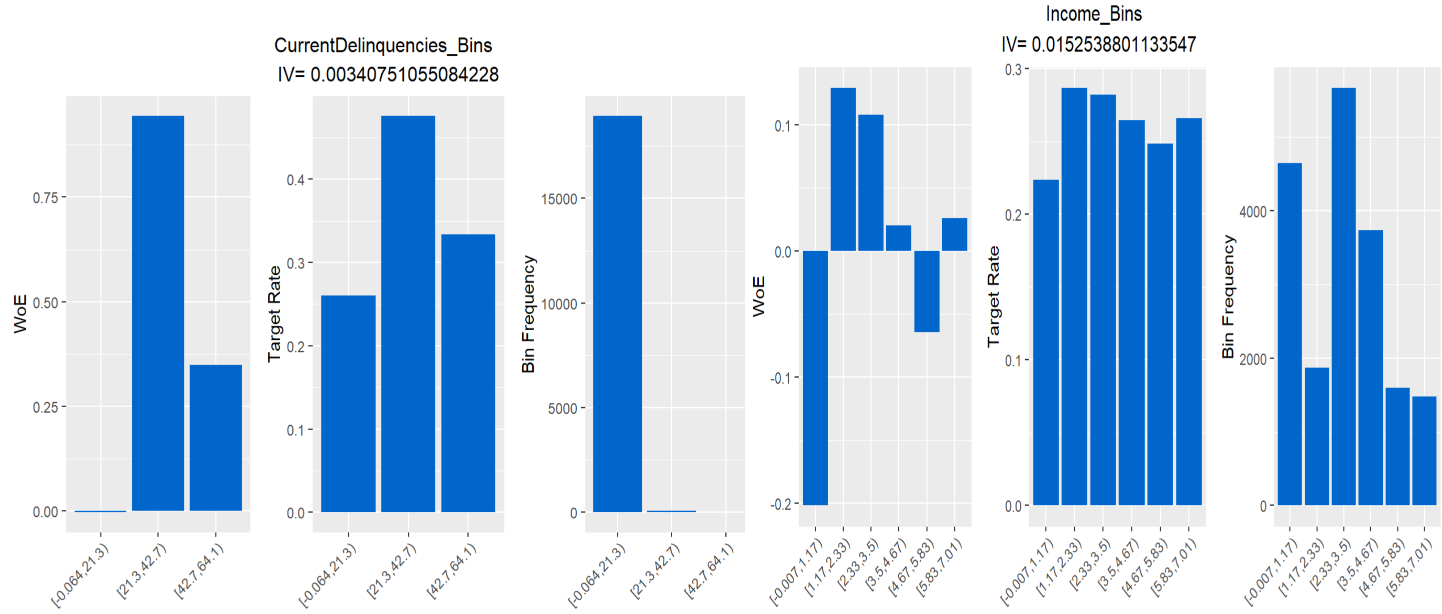
Table 1. Summary statistics of continuous variables

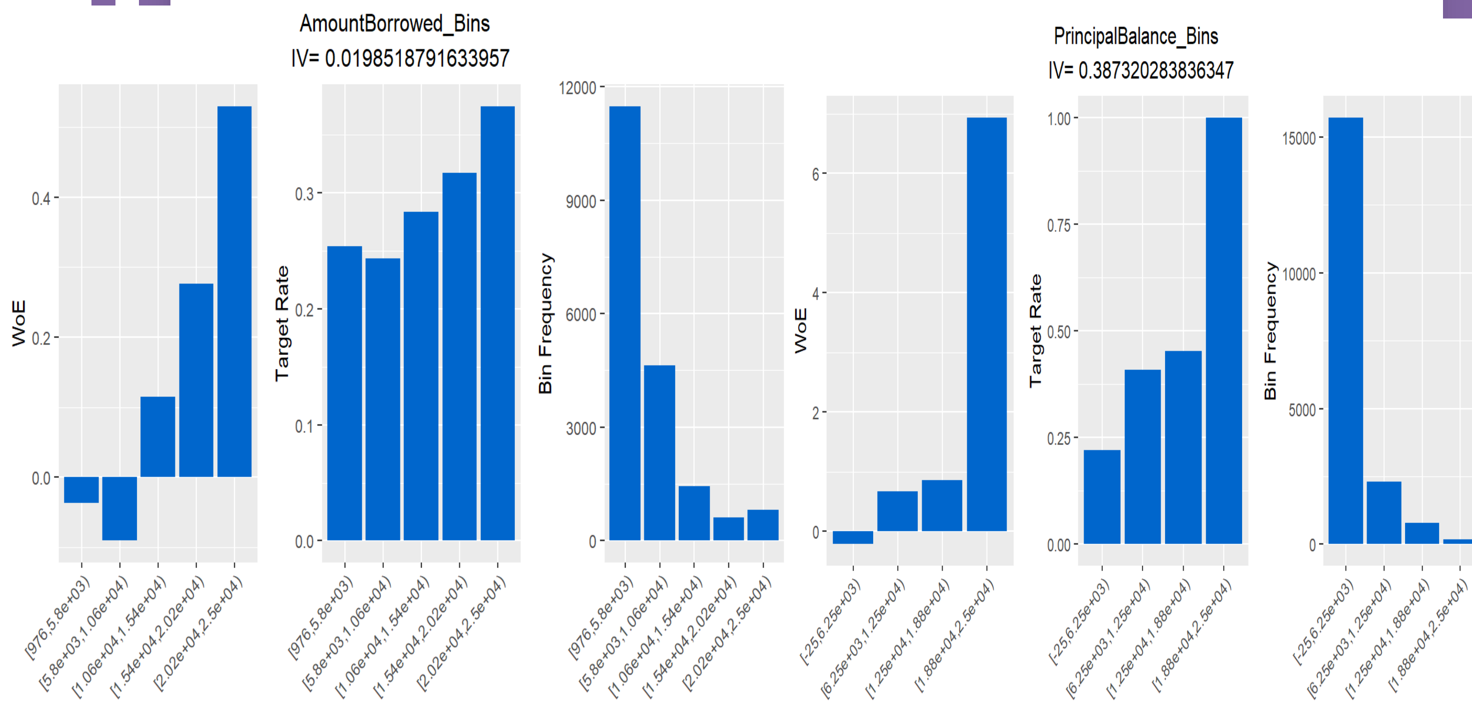
Variables	Target (Frequency)	
	Good (0)	Bad (1)
Borrower State		
AA	5	0
AE	8	0
West	4380	1641
South	4361	1562
AP	7	1
Northeast	1273	322
IA	128	37
ID	108	46
Midwest	3216	1214
IN	278	71
ME	69	11
ND	30	6
NE	76	11
TN	96	30
Is Borrower Homeowner		
FALSE	7860	2700
TRUE	6175	2252
Employment Status		
Full-time	9304	3446
Not available	2670	770
Not employed	91	26
Part-time	452	134
Retired	209	103
Self-employed	749	368
Income		
Level 0	3308	894
Level 1	296	145
Level 2	1335	536
Level 3	4064	1598
Level 4	2744	988
Level 5	1203	398
Level 6	1012	373
Level 7	73	20

Table 2. Summary statistics of categorical variables

Target	Variables	Min.	1st Quantile	Median	Mean	3rd Quantile	Max.	N
0	Debt to Income Ratio	0.00	0.13	0.20	0.32	0.31	10.01	14035
(Good)	Amount Borrowed	1000	2550	5000	6285	8000	25000	14035
	Current Delinquencies	0.00	0.00	0.00	1.20	1.00	50.00	14035
	Delinquencies Last 7 years	0.00	0.00	0.00	5.80	6.00	99.00	14035
	Public Records Last 10 years	0.00	0.00	0.00	0.39	1.00	21.00	14035
	Total Credit Lines	2.00	13.00	22.00	23.74	32.00	108.00	14035
	Inquiries Last 6 Months	0.00	0.00	1.00	2.44	3.00	46.00	14035
	Amount Delinquent	0.00	0.00	0.00	1068	20	190585	14035
	Public Records Last 12 Months	0.00	0.00	0.00	0.04	0.00	7.00	14035
	Current Credit Lines	0.00	5.00	9.00	9.49	13.00	46.00	14035
	Open Credit Lines	0.00	4.00	7.00	8.14	11.00	43.00	14035
	Revolving Credit Balance	0.00	1338.00	5411.0	15570	15213	1435667	14035
	Employment Status	9304	2670	91	452	209	749	14035
	Income	4064	3308	2744	1335	1203	1012	14035
	Principal Balance	0	1357	2529	2254	4312	16755	14035
1	Debt to Income Ratio	0.00	0.14	0.22	0.40	0.34	10.01	4952
(Bad)	Amount Borrowed	1000	2600	5000	7019	9500	25000	4952
	Current Delinquencies	0.00	0.00	0.00	2.07	2.00	64.00	4952
	Delinquencies Last 7 years	0.00	0.00	1.00	7.32	9.00	99.00	4952
	Public Records Last 10 years	0.00	0.00	0.00	0.55	1.00	30.00	4952
	Total Credit Lines	2.00	14.00	23.00	25.45	34.00	129.00	4952
	Inquiries Last 6 Months	0.00	1.00	3.00	4.17	6.00	105.00	4952
	Amount Delinquent	0.00	0.00	0.00	1847.2	590.5	444745.0	4952
	Public Records Last 12 Months	0.00	0.00	0.00	0.05	0.00	7.00	4952
	Current Credit Lines	0.00	5.00	8.00	9.34	13.00	52.00	4952
	Open Credit Lines	0.00	4.00	7.00	8.00	11.00	48.00	4952
	Revolving Credit Balance	0.00	769	3992	16827	14851	493300	4952
	Employment Status	3446	770	26	134	103	368	4952
	Income	1598	988	894	536	398	373	4952
	Principal Balance	0.00	2108	3753	5514	7119	25000	4952

Binning of Variables





Results of MARS model and Logistic Regression Model

Table 3. Results of multivariate adaptive regression splines (MARS) model

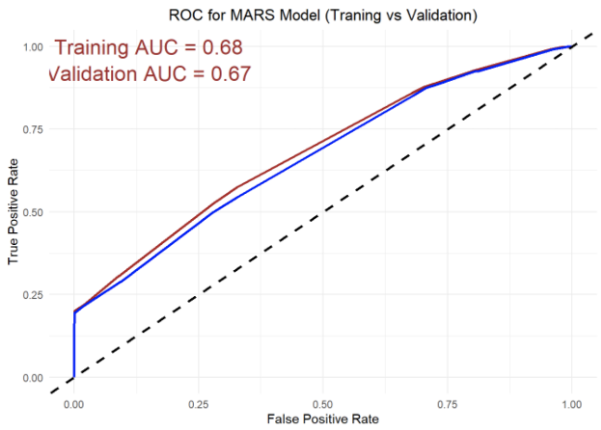
Variable	Bad
(Intercept)	-1.2059975
PrincipalBalance_Bins(6.25e+03,1.25e+04)	5.0899534
PrincipalBalance_Bins(1.25e+04,1.88e+04)	10.6707240
PrincipalBalance_Bins(1.88e+03,2.5e+04)	27.3337580
AmountBorrowed_Bins(5.8e+03,1.06e+04)	-0.7233100
AmountBorrowed_Bins(1.06e+04,1.54e+04)	-4.9618021
AmountBorrowed_Bins(1.54e+04,2.02e+04)	-5.9901291
AmountBorrowed_Bins(2.02e+04,2.5e+04)	-10.6746360
Income_Bins [1.17, 2.33)	0.2896362
Income_Bins [2.33, 3.5)	0.2709326
GCV	0.1656817
RSS	3139.506
Generalized R^2	0.1406927
R^2	0.1423213

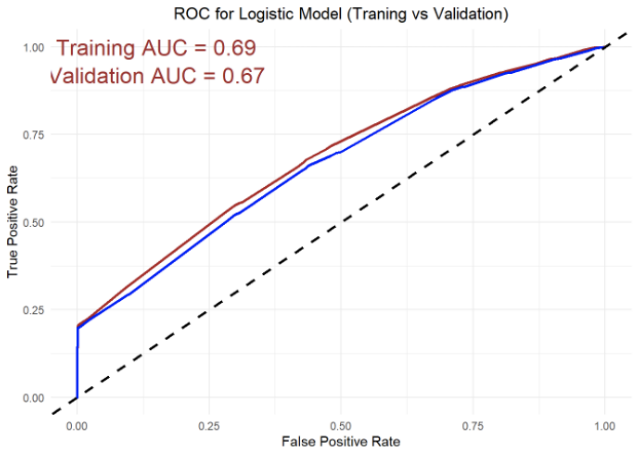
Table 4. Results of logistic regression model

	Coefficient Estimate
(Intercept)***	-1.29473
PrincipalBalance_Bins(6.25e+03,1.25e+04)***	5.25077
PrincipalBalance_Bins(1.25e+04,1.88e+04)***	10.71944
PrincipalBalance_Bins(1.88e+03,2.5e+04)	27.41586
AmountBorrowed_Bins(5.8e+03,1.06e+04)***	-0.74545
AmountBorrowed_Bins(1.06e+04,1.54e+04)***	-5.09771
AmountBorrowed_Bins(1.54e+04,2.02e+04)***	-5.95064
AmountBorrowed_Bins(2.02e+04,2.5e+04)***	-10.71883
CurrentDelinquencies_Bins[21.3,42.7)***	1.18761
CurrentDelingencies_Bins[42.7,64.1)	-14.64085
Income_Bins [1.17, 2.33)***	0.45629
Income_Bins [2.33, 3.5)***	0.36951
Income_Bins [3.5, 4.67)*	0.16484
Income_Bins [4.67, 5.83)	0.14268
Income_Bins [5.83, 7.01)	-0.02826

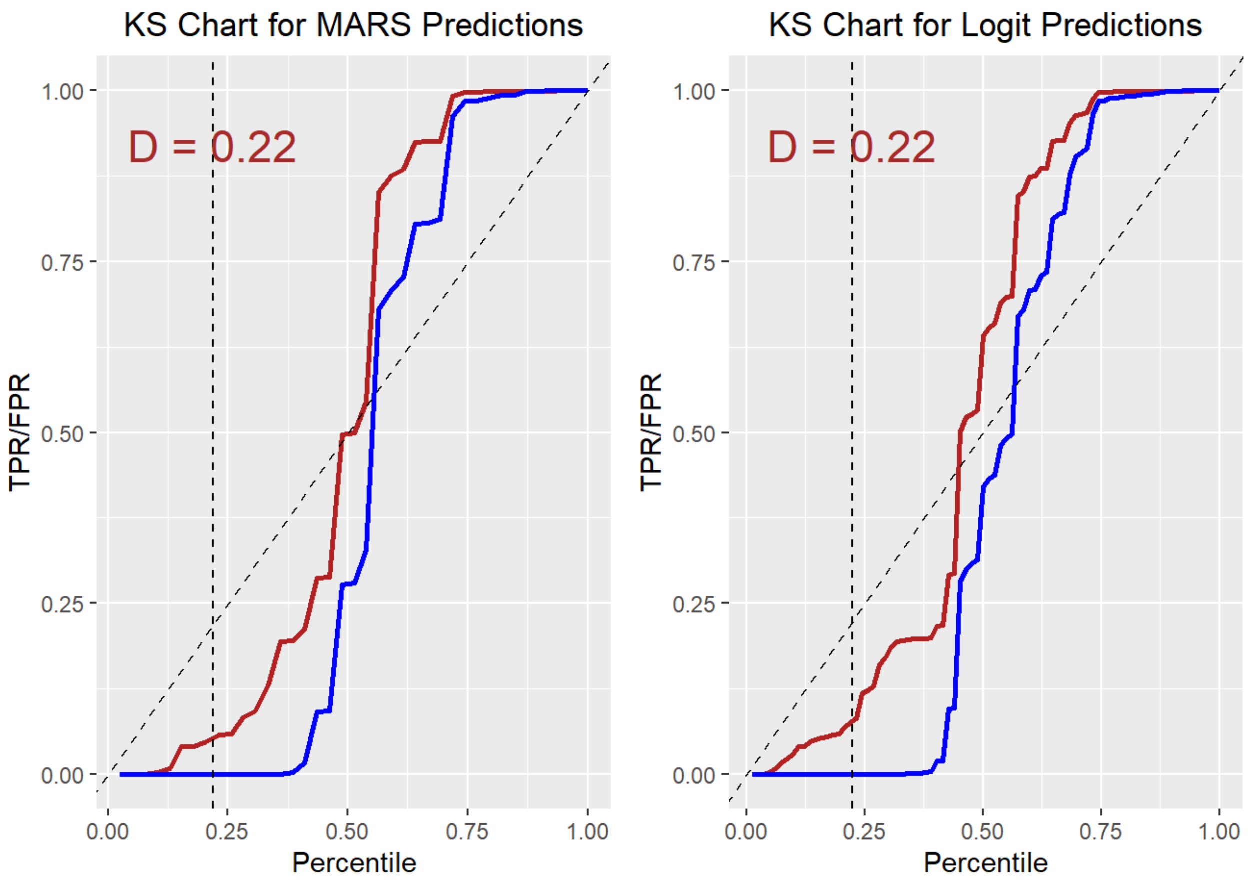
Note: *, **, *** represents significance at 5%, 1% and 0.1% level

ROC Curve


ROC for MARS Model (Traning vs Validation)
Training AUC = 0.68
Validation AUC = 0.67
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00


ROC for Logistic Model (Traning vs Validation)
Training AUC = 0.69
Validation AUC = 0.67
True Positive Rate
False Positive Rate
0.00
0.25
0.50
0.75
1.00

KS Statistics



Conclusion

- MARS and Logistic models- Two good models in predicting if the customer goes bad or not.
- Even though the KS Statistics of both models on the validation dataset are equal, based on the AUC value, the logistic model outperforms the MARS model since it has the highest value.