Loan-level Credit Risk Assessment for P2P Lending

By Yuqing (Elisa) Yao

I. Goals of Analysis

This report aims to illustrate types of loans that tend to have "bad" performance ("Charged Off", "Default", or "Does not meet the credit policy - Charged Off") and to create models to predict individual loan's performance ("good" or "bad") in order to help investors make investment decisions.

II. Recommendations

Based on descriptive statistics and statistical significance tests, the research suggests that:

- Higher interest rate reflected risk premium for exposing to higher credit risk which is consistent to the common sense. Individual investors should be aware of the risk-return relationship and choose reasonable return rate that corresponds to their risk capacity.

- Although shorter term loans (36 months) may be thought to have higher liquidity risk, it appears to have higher credit risk.

- Loans that finance for small businesses, car purchase, credit card payment, debt consolidation, and house purchase have higher risk than other purposes.

- Beware of states that tend to have a higher bad loan rates, such as AL, AR, FL, IA, IN, KY, MO, MS, NV, TN.

- Choosing borrowers that has higher annual income substantially reduces risk exposure to defaults.

- Investors should pay attention to signals that shows the level of eagerness borrow, e.g. inquiries within last 6 months, and number of delinquencies within last 2 years.

III. Expected Effects of Actions

If the investors make decisions according to the recommendations above, they are expected to:

- Reduce their unwanted risk exposure

- Be able to more accurately construct loan portfolios with desired risk-return profile

- More efficiently and scientifically assess a borrower's default probability

Besides, LendingClub can refer to this analysis to:

- Adjust interest rates of individual loans to reflect default probability

- Assess overall credit risk exposure of the platform

- Enhance its current credit rating system

IV. Conclusions from Data Analysis

The following conclusions can be drawn from data analysis. Please see technical discussion in **Appendix**.

- Bootstrap Forest has the highest predictive power among models used.

ROC curve, Lift curve, and Confusion Matrix indicates that the models have satisfactory accuracy. Bootstrap Forest has the highest R-Square among the models.

- Interest rate has the highest indication of defaulting ("Bad" performance).

All of the models used suggest that interest rate has the highest explanatory power among the predictors used. The higher the interest rate, the more likely the loan end up being "Bad".

- Shorter-term (36 months) tend to have a higher rate of bad performance than longer-term (60 months) loans

In all of the prediction models, term ranks very high level among predictors and is significant at 99.9% confidence level.

- Annual income is a significant indicator for loan performance

Annual income usually has large contribution to explanatory power in the models that this analysis used.

- Loan purpose that are small businesses, car purchase, credit card payment, debt consolidation, and house purchase tend to have worse performance

These categories are 99.9% significant in the logistic regression.

- Certain states tend to have higher default rates. This may be correlated to economic status of the states.

By plotting the heat map, we can intuitively find out that AL, AR, FL, IA, IN, KY, MO, MS, NV, TN have more than 20% "bad" rate. However, in the Chi-Square test, the difference of patterns is not statistically significant.

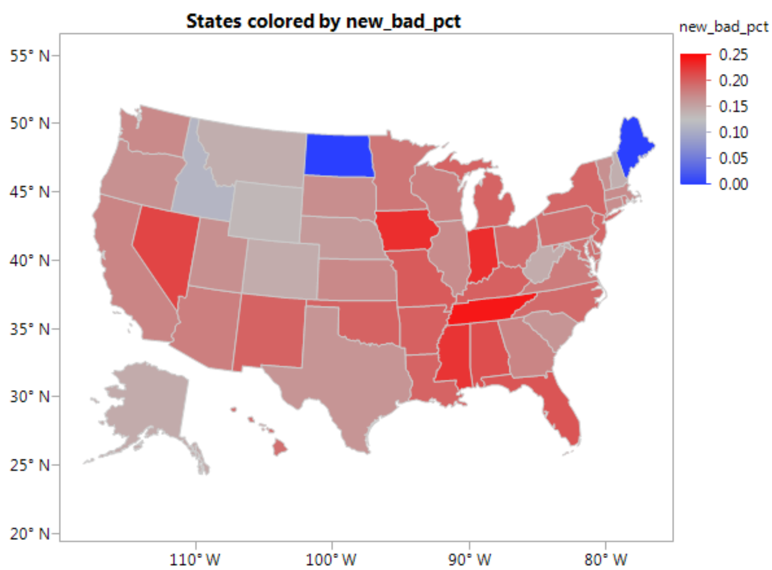


Figure 1. Heatmap of "bad" loan percentage

V. Methodologies

My analysis is based on the fact that good loans and bad loans are qualitative assessment. Popular models that deal with categorical response variable are Logistic Regression and Classification Tree. Therefore, I used both models to conduct analysis. Model set-up and results are listed follows.

new_loan_status = f(loan_amnt, term, int_rate, installment, verification_status, purpose, dti, delinq_2yrs, inq_last_6mths, mths_since_last_delinq, open_acc, pub_rec, revol_util, total_acc, new_home_ownership, new_annual_inc, new_desc_length3, new_cr_sr, new_emp_length)

1. Logistic Regression

Table 1. Parameters estimation and corresponding significance

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	3.74843563	0.2492423	226.18	<.0001*
loan_amnt	7.92115e-6	7.013e-6	1.28	0.2587
term[36 months]	-0.229923	0.0221662	107.59	<.0001*
int_rate	0.08035571	0.0029019	766.80	<.0001*
installment	0.00047474	0.0002163	4.82	0.0282*
verification_status[Not Verified]	-0.0437696	0.0131188	11.13	0.0008*
verification_status[Source Verified]	0.04074365	0.0119552	11.61	0.0007*
purpose[car]	-0.2763171	0.0823808	11.25	0.0008*
purpose[credit_card]	-0.223115	0.0345416	41.72	<.0001*
purpose[debt_consolidation]	-0.1062622	0.0300939	12.47	0.0004*

2. Decision Tree

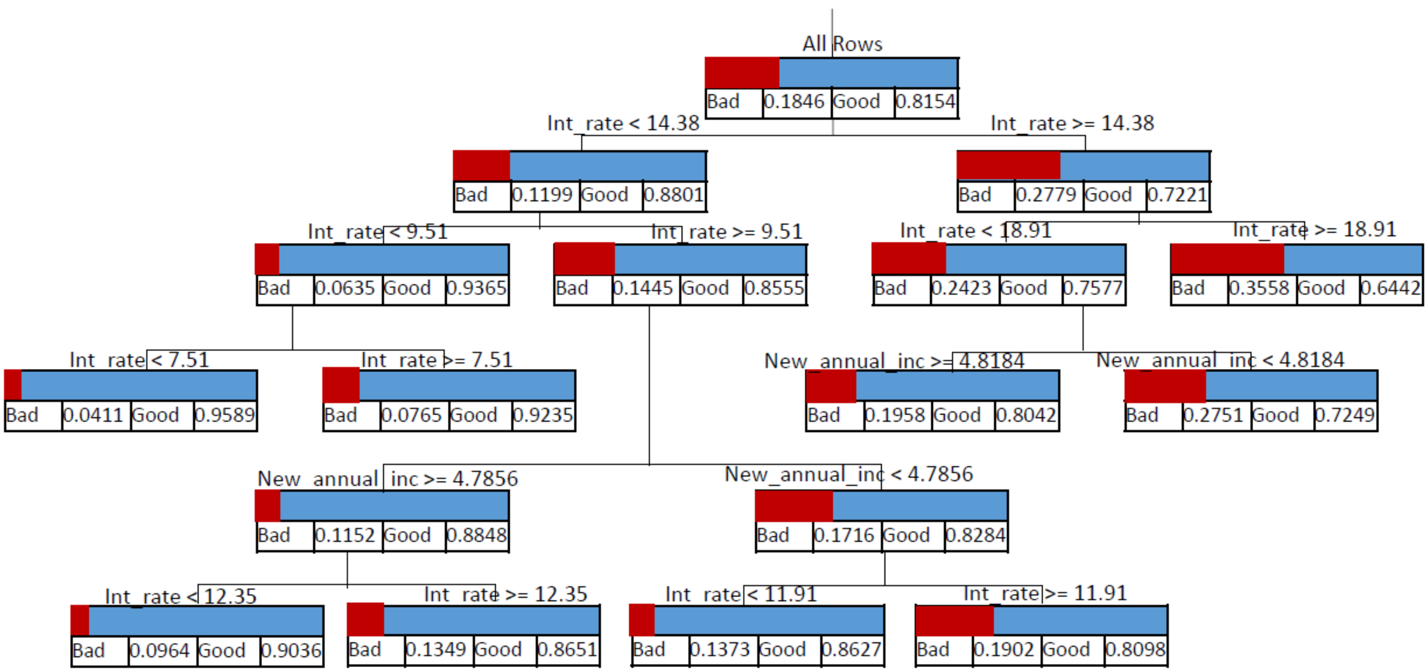


Figure 2. Top portion of Decision Tree

VI. Research Outlook

Further efforts could be made on continuing the research, such as considering structural change, time dependence, and applying ensemble learning. Extended topic using the same dataset could be constructing P2P loan portfolios, Copula of individual loans, and effects of economic factors on loan performance.

Reference

Tsai, Ramiah, Singh (2014). Peer Lending Risk Predictor. Stanford University. Retrieved from: http://cs229.stanford.edu/proj2014/Kevin%20Tsai,Sivagami%20Ramiah,Sudhanshu%20Singh,Peer%20Lending%20Risk%20Predictor.pdf

Davenport, K. (2013). Gradient Boosting: Analysis of LendingClub's Data. Retrieved from: https://www.r-bloggers.com/gradient-boosting-analysis-of-lendingclubs-data/

Walczak, E. (2016). Initial loan book analysis. Kaggle. Retrieved from: https://www.kaggle.com/erykwalczak/d/wendykan/lending-club-loan-data/initial-loan-book-analysis/comments

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PLoS One*, 10(10) doi:http://dx.doi.org/10.1371/journal.pone.0139427

Milad Malekipirbazari, Vural Aksakalli. Risk assessment in social lending via random forests, *Expert Systems with Applications*, Volume 42, Issue 10, 15 June 2015, Pages 4621-4631, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2015.02.001.

Guo, Y., Zhou, W. et al. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*. 249: 417–426

Appendix I: Data Dictionary

LoanStatNew	Description
addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
open_acc	The number of open credit lines in the borrower's credit file.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
pub_rec	Number of derogatory public records
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
total_acc	The total number of credit lines currently in the borrower's credit file
mths_since_rcnt_il	Months since most recent installment accounts opened
total_bal_il	Total current balance of all installment accounts

Appendix II: Fixing Data Issues

1. Missing Data

Solution:

For *annual_inc, delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_util*, and *total_acc*, the missing values are few. Therefore, I will leave them alone by now and take advantage of JMP Informative Missing Partition to optimize the results.

For *emp_title* and *desc*, the missing values are more, I assigned “NA” value to the missing lines since it may have some influence on the “bad” status.

2. Creating New Variable Using Industry Knowledge

There are several variables that does not make much sense when they are used individually. For example, *earliest_cr_line*. These kind of variables need to be combined with other variables to create meaningful predictor.

Other variables may contain useful information when transformed into other format. For example, *desc*. Although Sentiment Analysis in Text Mining is a powerful tool, it will not be touched on in this research because of the time constraint. Instead, it is converted to length of description, which also has some explanatory power to the model.

Solution:

The following new variables are created:

new_loan_status = $\begin{cases} Bad, & loan_{status} == Charged\ Off \mid Default \mid Does\ not\ meet\ the\ credit\ policy \\ Good, & otherwise \end{cases}$

new_home_ownership = $\begin{cases} Rent \\ Own \\ Mortgage \\ Other \end{cases}$

new_desc_length = $\begin{cases} \ln\big(\text{Length}(desc)\big), & desc \neq \text{null} \\ 0, & desc = null \end{cases}$

new_cr_sr = $Date\ Difference(earliest\_cr\_line, issue\_d, "Day")$

new_emp_lenth = group mean of categorical level of *emp_length*

3. Skewness of Data

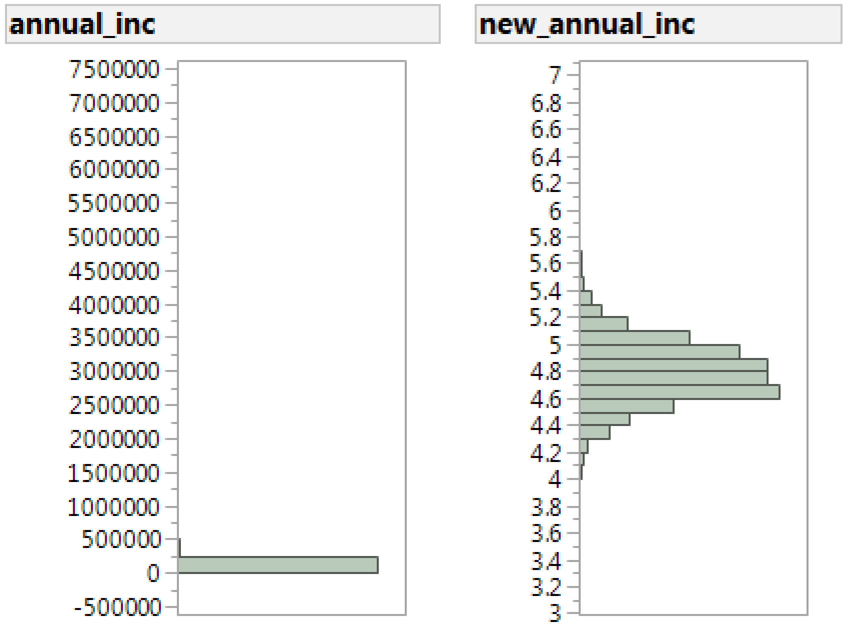
The *annual_inc* variable is highly skewed; therefore, logarithm transformation is needed to reduce the influence of skewness.

Solution:

Use log base 10 transformation to convert the distribution to roughly normally distributed.

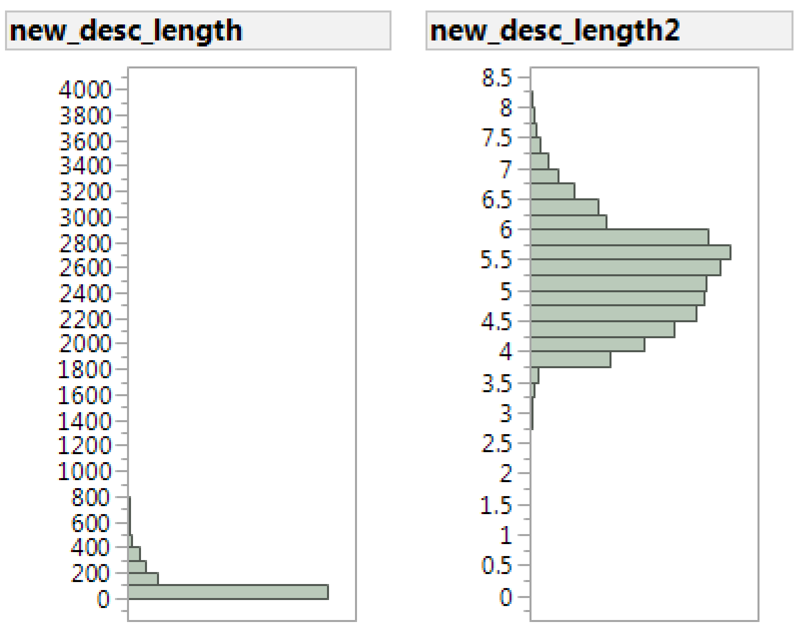
new_annual_inc = $Log10(annual\_inc)$

The distribution of annual income before and after the processing are illustrated below.



Use natural log transformation to convert the distribution to roughly normally distributed.

new_desc_length = $\ln(\mathrm{Length}(desc))$



Appendix III: Models and Results

1. ANOVA

1.1 Credit Line Seniority

Oneway Anova

Summary of Fit

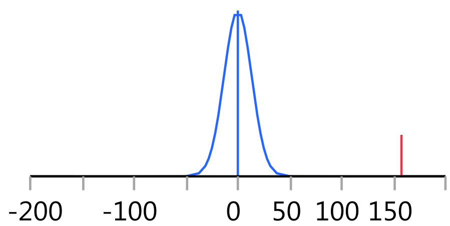
Rsquare	0.000567
Adj Rsquare	0.000563
Root Mean Square Error	2563.168
Mean of Response	5545.602
Observations (or Sum Wgts)	247772

t Test

Good-Bad

Assuming equal variances

Difference	157.618	t Ratio	11.86117
Std Err Dif	13.289	DF	247770
Upper CL Dif	183.663	Prob > |t|	<.0001*
Lower CL Dif	131.573	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



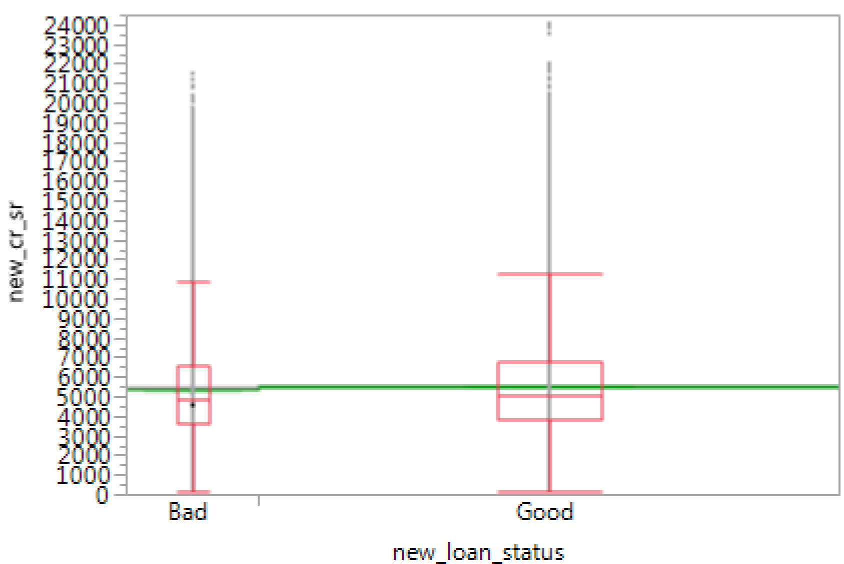
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
new_loan_status	1	924291898	924291898	140.6873	<.0001*
Error	247770	1.6278e+12	6569831.2		
C. Total	247771	1.6287e+12			

Means for Oneway Anova

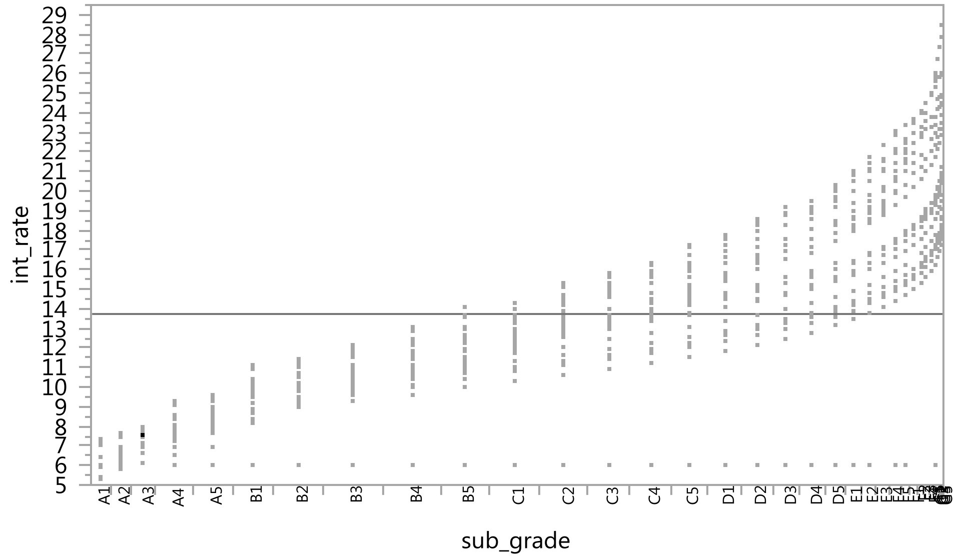
Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Bad	45595	5416.99	12.004	5393.5	5440.5
Good	202177	5574.61	5.700	5563.4	5585.8

Std Error uses a pooled estimate of error variance



1.2 Collinearity of Interest Rate and Sub-grade

Oneway Analysis of int_rate By sub_grade



2 Chi-Square Analysis

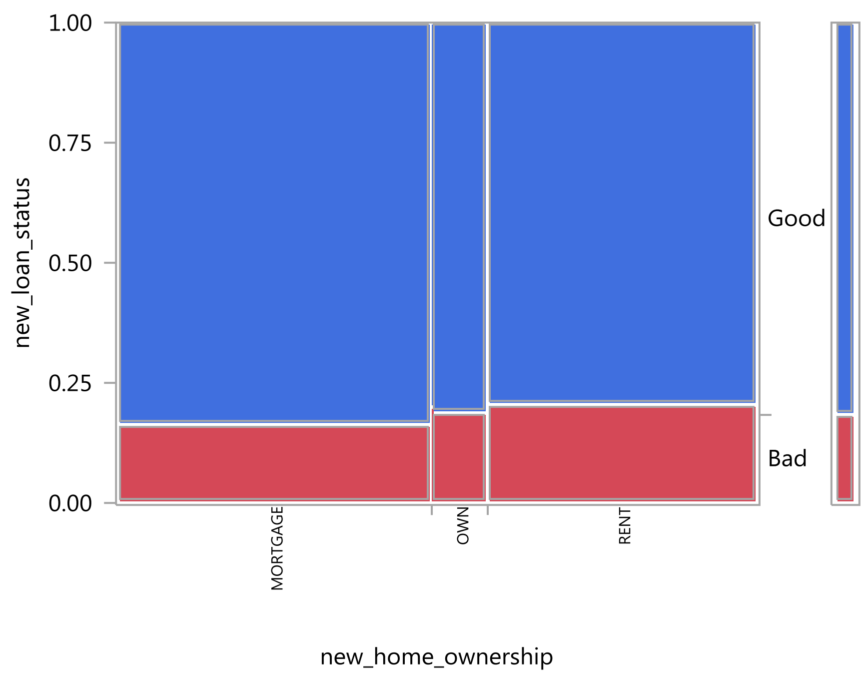
2.1 Home Ownership

Contingency Table

new_home_ownership By new_loan_status

Count Total % Col % Row %	Bad	Good	Total
MORTGAGE	20079 8.10 44.03 16.44	102042 41.18 50.47 83.56	122121 49.28
OTHER	46 0.02 0.10 20.18	182 0.07 0.09 79.82	228 0.09
OWN	4024 1.62 8.82 18.88	17289 6.98 8.55 81.12	21313 8.60
RENT	21449 8.66 47.04 20.60	82690 33.37 40.89 79.40	104139 42.03
Total	45598 18.40	202203 81.60	247801

Mosaic Plot



Tests

N	DF	-LogLike	RSquare (U)
247801	3	324.80134	0.0027

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	649.603	<.0001*
Pearson	650.220	<.0001*

3 Logistic Regression

3.1 All Relevant Variables

new_loan_status = f(loan_amnt, term, int_rate, installment, verification_status, purpose, dti, delinq_2yrs, inq_last_6mths, mths_since_last_delinq, open_acc, pub_rec, revol_util, total_acc, new_home_ownership, new_annual_inc, new_desc_length3, new_cr_sr, new_emp_length)

Nominal Logistic Fit for new_loan_status

Effect Summary

Source	LogWorth	PValue
new_annual_inc	183.345	0.00000
int_rate	168.049	0.00000
dti	57.890	0.00000
purpose	53.490	0.00000
inq_last_6mths	46.092	0.00000
term	24.481	0.00000
revol_util	23.686	0.00000
total_acc	21.372	0.00000
new_home_ownership	17.320	0.00000
open_acc	16.999	0.00000
new_emp_length	11.987	0.00000
delinq_2yrs	9.435	0.00000
verification_status	3.162	0.00069
mths_since_last_delinq	2.595	0.00254
installment	1.550	0.02817
pub_rec	0.954	0.11113
loan_amnt	0.587	0.25869
new_desc_length3	0.447	0.35726
new_cr_sr	0.006	0.98565

Converged in Gradient, 5 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	3908.905	34	7817.809	<.0001*
Full	48792.944			
Reduced	52701.849			

RSquare (U)	0.0742
AICc	97655.9
BIC	97992
Observations (or Sum Wgts)	109562

Measure	Training	Definition
Entropy RSquare	0.0742	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1115	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4453	∑ -Log(ρ[j])/n
RMSE	0.3748	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2808	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1860	∑ (ρ[j]≠ρMax)/n
N	109562	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	109527	48792.944	97585.89

Source	DF	-LogLikelihood	ChiSquare
Saturated	109561	0.000	**Prob>ChiSq**
Fitted	34	48792.944	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	3.74843563	0.2492423	226.18	<.0001*
loan_amnt	7.92115e-6	7.013e-6	1.28	0.2587
term[36 months]	-0.229923	0.0221662	107.59	<.0001*
int_rate	0.08035571	0.0029019	766.80	<.0001*
installment	0.00047474	0.0002163	4.82	0.0282*
verification_status[Not Verified]	-0.0437696	0.0131188	11.13	0.0008*
verification_status[Source Verified]	0.04074365	0.0119552	11.61	0.0007*
purpose[car]	-0.2763171	0.0823808	11.25	0.0008*
purpose[credit_card]	-0.223115	0.0345416	41.72	<.0001*
purpose[debt_consolidation]	-0.1062622	0.0300939	12.47	0.0004*
purpose[educational]	0.02542955	0.18134	0.02	0.8885
purpose[home_improvement]	0.006512	0.0419184	0.02	0.8765
purpose[house]	-0.2573734	0.0974011	6.98	0.0082*
purpose[major_purchase]	-0.0979979	0.0598411	2.68	0.1015
purpose[medical]	0.14244699	0.0710912	4.01	0.0451*
purpose[moving]	0.13416724	0.082748	2.63	0.1049
purpose[other]	0.04820768	0.0404549	1.42	0.2334
purpose[renewable_energy]	0.16952866	0.2212734	0.59	0.4436
purpose[small_business]	0.61801401	0.0540275	130.85	<.0001*
purpose[vacation]	0.08447457	0.0984163	0.74	0.3907
dti	0.01932865	0.0011974	260.57	<.0001*
delinq_2yrs	0.05740388	0.0091592	39.28	<.0001*
inq_last_6mths	0.09450842	0.0065772	206.47	<.0001*
mths_since_last_delinq	-0.0013536	0.0004485	9.11	0.0025*
open_acc	0.01942268	0.0022654	73.51	<.0001*
pub_rec	0.02738942	0.0171923	2.54	0.1111
revol_util	0.00393081	0.0003855	103.96	<.0001*
total_acc	-0.0097472	0.0010085	93.41	<.0001*
new_home_ownership[MORTGAGE]	-0.1071607	0.0605607	3.13	0.0768
new_home_ownership[OTHER]	0.1127043	0.1781863	0.40	0.5271
new_home_ownership[OWN]	-0.0639302	0.0629431	1.03	0.3098
new_annual_inc	-1.4645484	0.0506176	837.15	<.0001*
new_desc_length3	0.0030812	0.003347	0.85	0.3573
new_cr_sr	6.22815e-8	3.4627e-6	0.00	0.9856
new_emp_length	-0.0107505	0.0015086	50.78	<.0001*

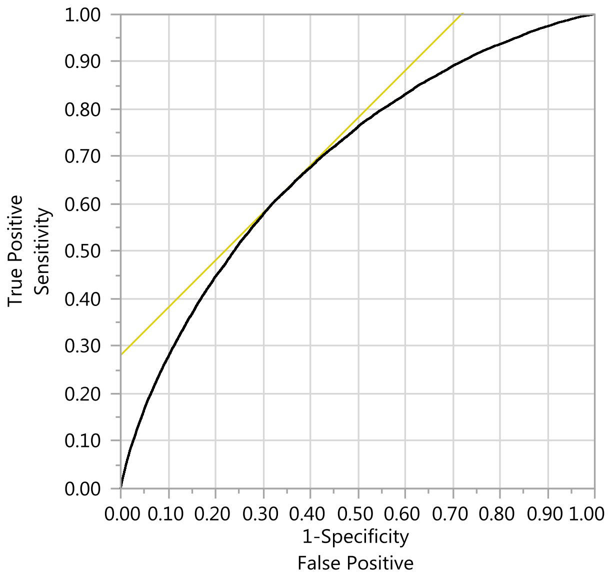
For log odds of Bad/Good

Effect Wald Tests

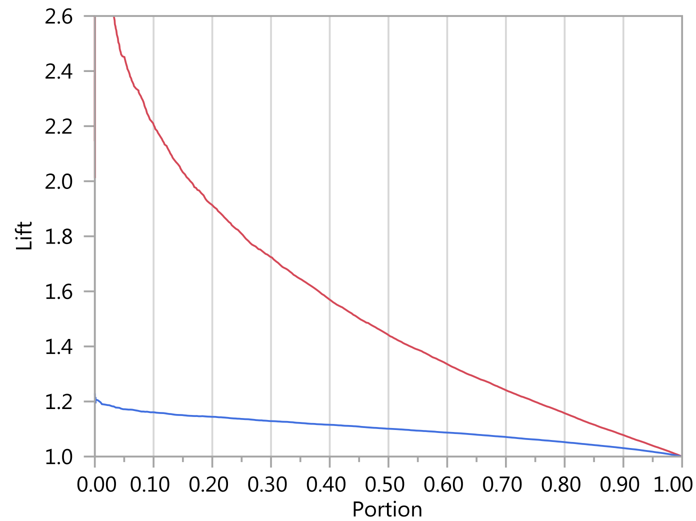
Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
loan_amnt	1	1	1.27576836	0.2587
term	1	1	107.592769	<.0001*
int_rate	1	1	766.796658	<.0001*
installment	1	1	4.81755146	0.0282*
verification_status	2	2	14.5598636	0.0007*
purpose	13	13	289.818102	<.0001*
dti	1	1	260.572221	<.0001*
delinq_2yrs	1	1	39.2796959	<.0001*
inq_last_6mths	1	1	206.469014	<.0001*
mths_since_last_delinq	1	1	9.1108277	0.0025*
open_acc	1	1	73.5078644	<.0001*

Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
pub_rec	1	1	2.53805045	0.1111
revol_util	1	1	103.961894	<.0001*
total_acc	1	1	93.4131475	<.0001*
new_home_ownership	3	3	83.7593731	<.0001*
new_annual_inc	1	1	837.150716	<.0001*
new_desc_length3	1	1	0.84748572	0.3573
new_cr_sr	1	1	0.0003235	0.9856
new_emp_length	1	1	50.7837102	<.0001*

Receiver Operating Characteristic



Lift Curve



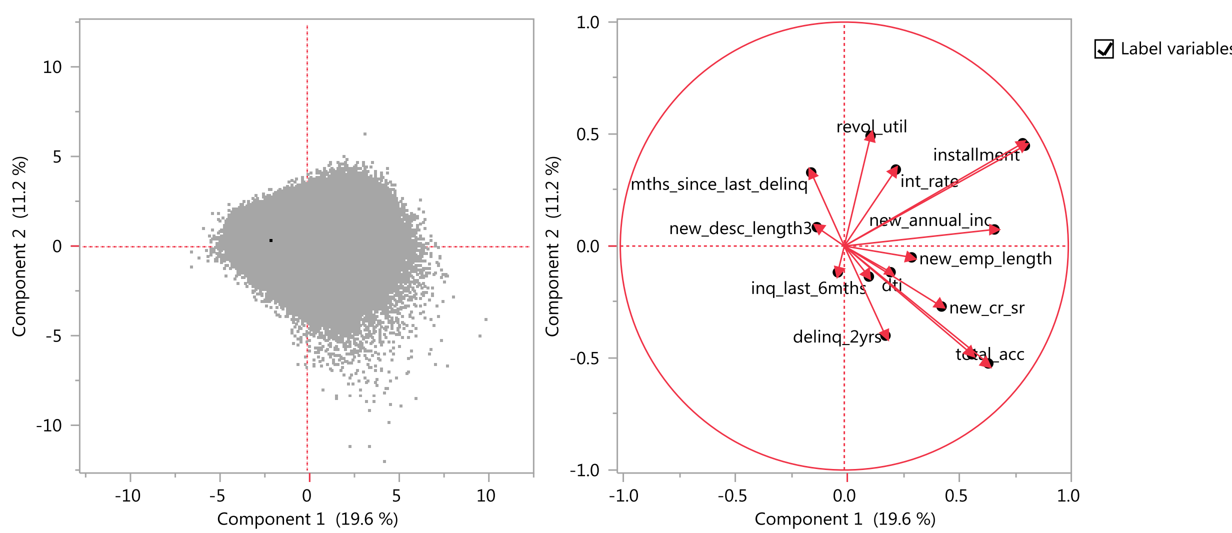
Using new_loan_status='Bad' to be the positive level

AUC
0.69135

new_loan_status
Bad
Good

3.2 Principle Components

Summary Plots



Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
loan_amnt	0.46993	0.34579	-0.00318	-0.13839	-0.08102	0.03490	-0.15854	0.06940
int_rate	0.13408	0.26348	-0.31829	0.43538	0.14084	0.30226	0.04587	-0.02355
installment	0.46421	0.35412	-0.01665	-0.13156	-0.08556	0.04354	-0.15572	0.08347
dti	0.12047	-0.09009	0.00216	0.60099	-0.18402	-0.27604	-0.01167	0.08913
delinq_2yrs	0.10718	-0.30857	-0.60397	-0.11022	0.04127	-0.02755	-0.01155	0.10779
inq_last_6mths	0.06392	-0.10513	-0.02247	0.13232	-0.03777	0.73953	0.31093	-0.30761
mths_since_last_delinq	-0.08617	0.25337	0.62106	0.17028	0.08521	0.04878	0.02600	-0.03607
open_acc	0.33206	-0.37271	0.19238	0.23904	-0.24338	0.03562	-0.10702	0.01246
pub_rec	-0.01681	-0.09041	0.09917	0.09661	0.61463	0.28535	-0.10936	0.59591
revol_util	0.06832	0.37951	-0.21646	0.35786	0.06646	-0.25508	0.27681	0.10130
total_acc	0.37472	-0.40436	0.18389	0.13937	-0.10322	0.01224	0.02693	0.04284
new_annual_inc	0.39074	0.05714	0.06189	-0.32372	0.06390	0.07629	0.06939	-0.06094
new_desc_length3	-0.07140	0.06502	0.04904	-0.16621	-0.43657	0.07762	0.65999	0.51576
new_cr_sr	0.25319	-0.20832	0.11150	-0.08172	0.33508	-0.22040	0.24975	0.18889
new_emp_length	0.17517	-0.03905	0.03474	-0.03282	0.40373	-0.26637	0.49734	-0.44656

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	2.9423	19.615	19.615
2	1.6797	11.198	30.813
3	1.5495	10.330	41.143
4	1.5204	10.136	51.279
5	1.1844	7.896	59.175
6	1.1546	7.698	66.873
7	0.9812	6.541	73.414
8	0.8644	5.763	79.177
9	0.7606	5.071	84.248
10	0.7082	4.721	88.969
11	0.5411	3.608	92.577
12	0.4214	2.810	95.386
13	0.3577	2.385	97.771
14	0.2918	1.945	99.716
15	0.0425	0.284	100.000

new_loan_status = f(Prin1, Prin2, Prin3, Prin4, Prin5)

Nominal Logistic Fit for new_loan_status

Effect Summary

Source	LogWorth	PValue
Prin4	748.387	0.00000
Prin3	338.361	0.00000
Prin2	202.640	0.00000
Prin1	28.755	0.00000
Prin5	3.706	0.00020

Converged in Gradient, 5 iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	2617.963	5	5235.927	<.0001*

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Full	50083.885			
Reduced	52701.849			

RSquare (U)	0.0497
AICc	100180
BIC	100237
Observations (or Sum Wgts)	109562

Measure	Training	Definition
Entropy RSquare	0.0497	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.0755	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4571	∑ -Log(ρ[j])/n
RMSE	0.3797	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.2885	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1865	∑ (ρ[j]≠ρMax)/n
N	109562	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	109556	50083.885	100167.8
Saturated	109561	0.000	**Prob>ChiSq**
Fitted	5	50083.885	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-1.6060548	0.0089282	32359	<.0001*
Prin1	0.05352631	0.0047368	127.69	<.0001*
Prin2	0.18188675	0.0060484	904.32	<.0001*
Prin3	-0.2279293	0.005755	1568.6	<.0001*
Prin4	0.40115165	0.0070334	3253.0	<.0001*
Prin5	-0.0274988	0.0074184	13.74	0.0002*

For log odds of Bad/Good

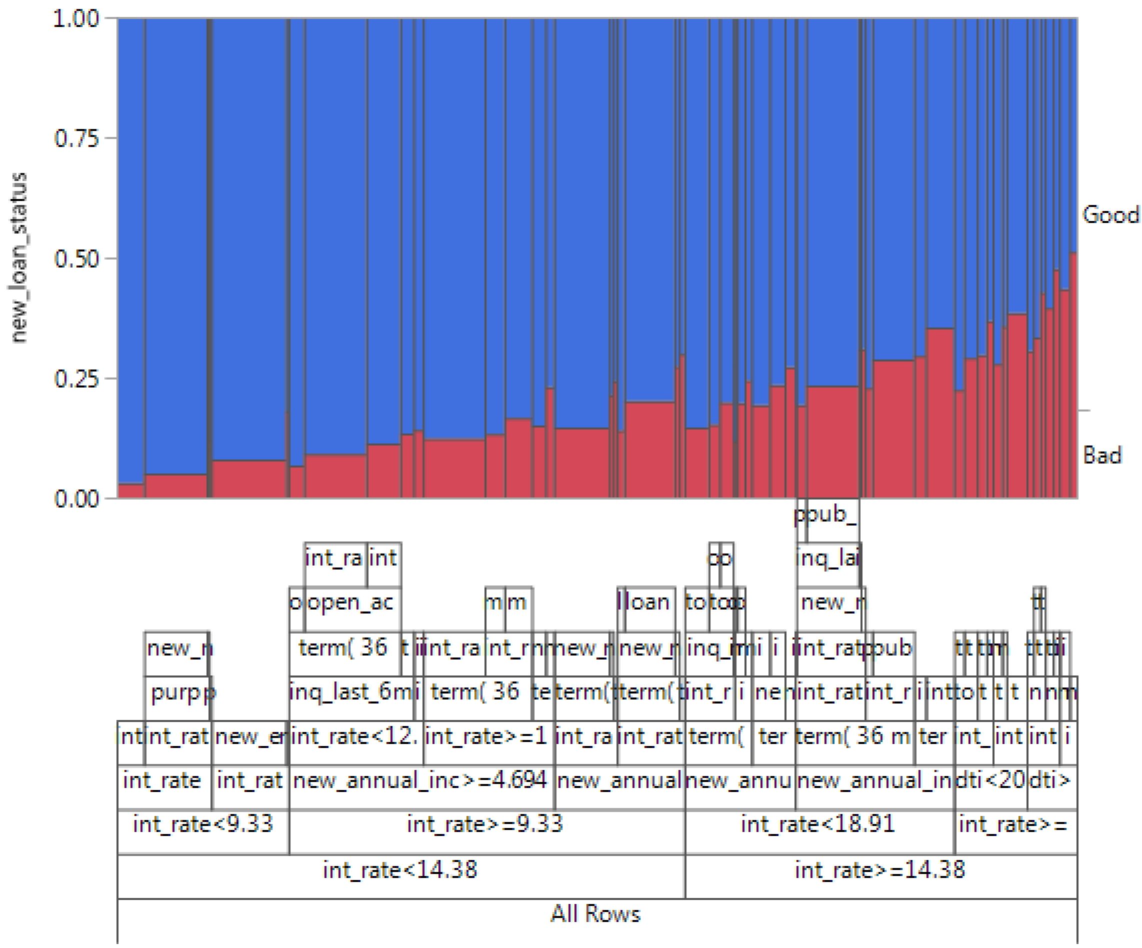
Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Prin1	1	1	127.109182	<.0001*
Prin2	1	1	925.906656	<.0001*
Prin3	1	1	1550.41309	<.0001*
Prin4	1	1	3437.85548	<.0001*
Prin5	1	1	13.8602169	0.0002*

4 Decision Tree Models

4.1 Single Classification Tree

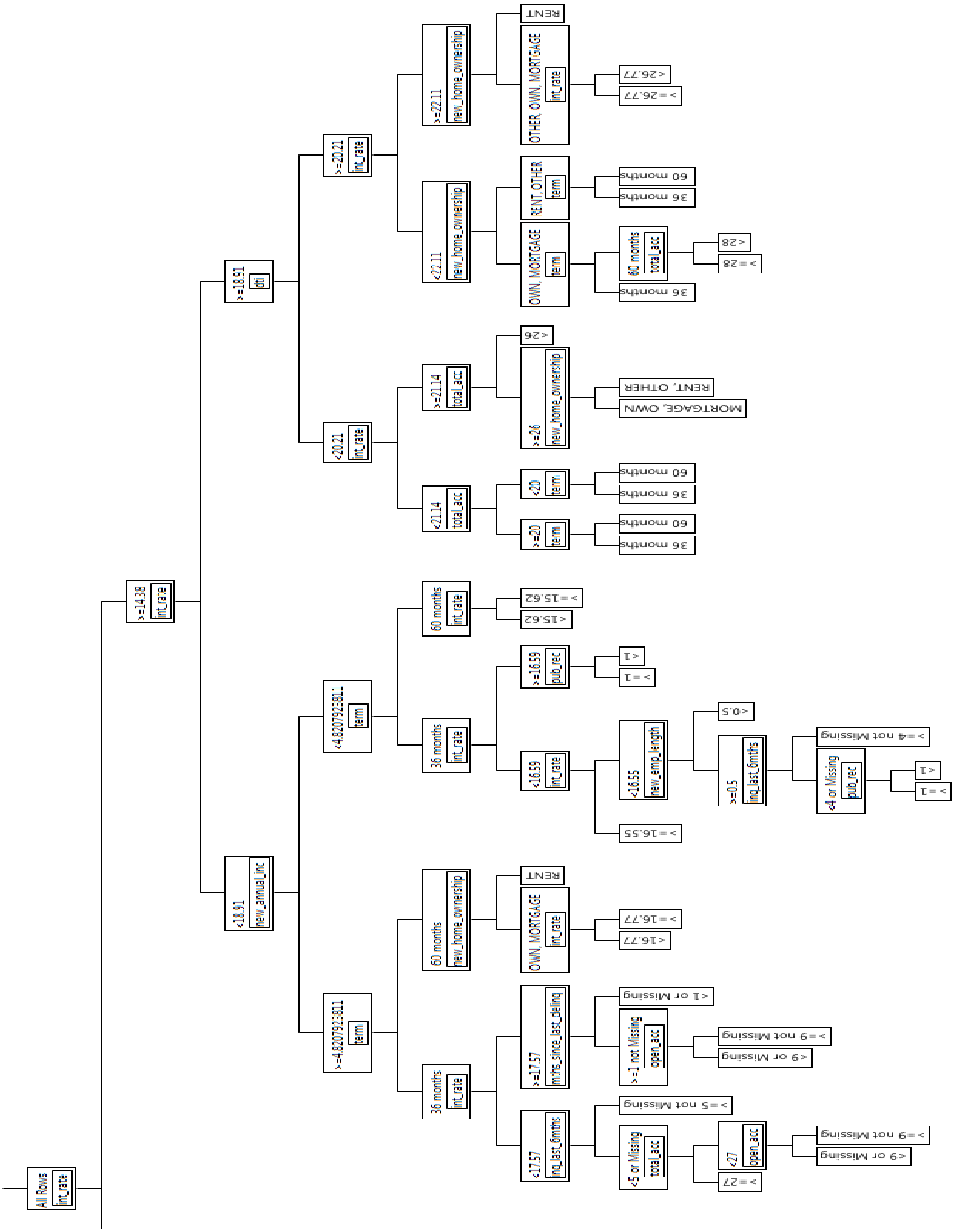
new_loan_status = f(loan_amnt, term, int_rate, installment, verification_status, purpose, dti, delinq_2yrs, inq_last_6mths, mths_since_last_delinq, open_acc, pub_rec, revol_util, total_acc, new_home_ownership, new_annual_inc, new_desc_length3, new_cr_sr, new_emp_length)

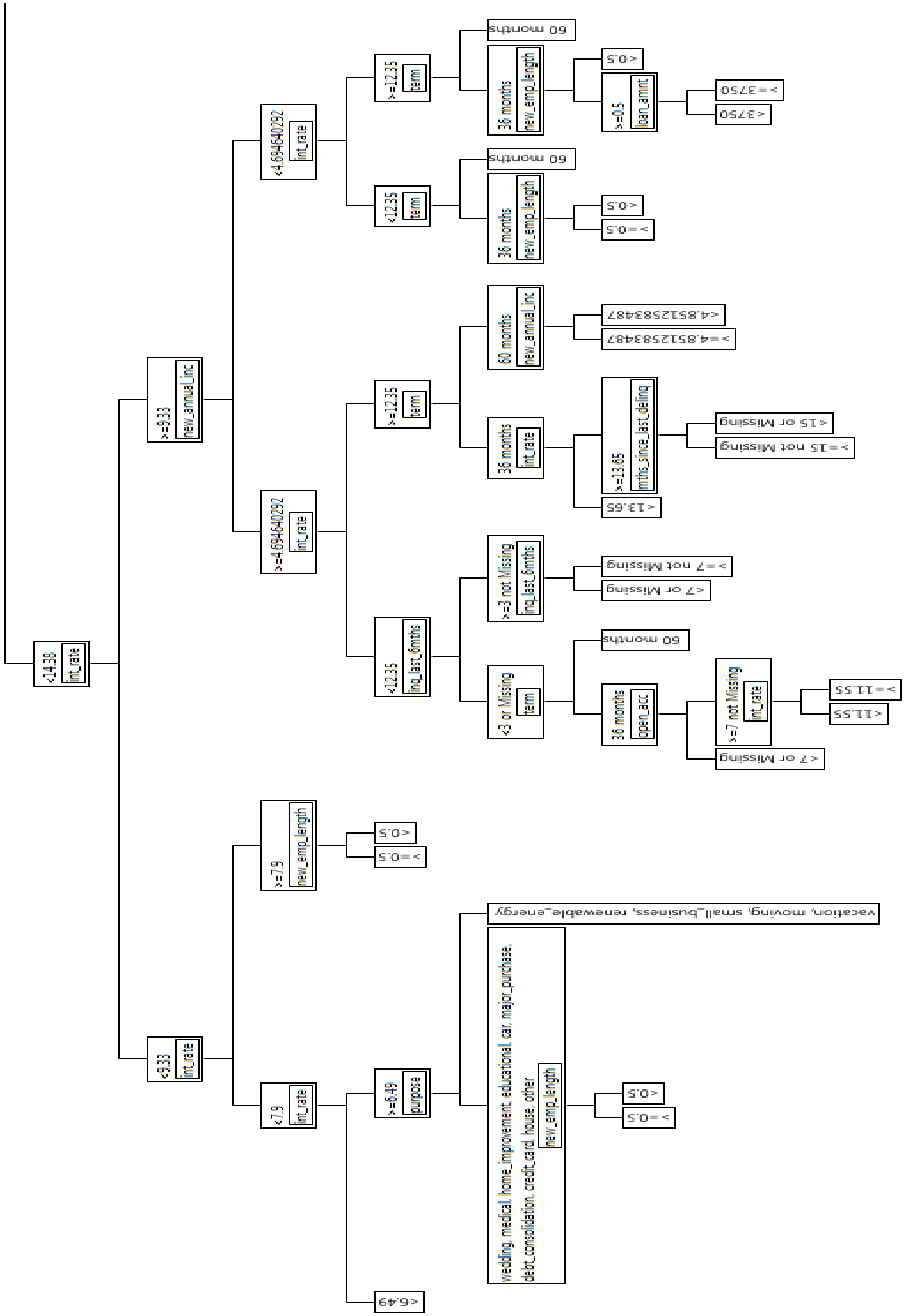


Column Contributions

Term	Number of Splits	G^2		Portion
int_rate	19	10972.696		0.7951
new_annual_inc	6	1174.33932		0.0851
term	11	519.517262		0.0376
dti	1	223.136846		0.0162
loan_amnt	2	219.677221		0.0159
total_acc	6	136.350121		0.0099
new_emp_length	4	123.893821		0.0090
new_home_ownership	3	107.716044		0.0078
inq_last_6mths	4	94.7964813		0.0069
purpose	2	71.1140597		0.0052
mths_since_last_delinq	3	51.3302032		0.0037
pub_rec	2	36.8369407		0.0027
new_desc_length3	1	27.2680958		0.0020
open_acc	1	14.6592954		0.0011
delinq_2yrs	1	13.8445855		0.0010
verification_status	1	12.7726818		0.0009
installment	0	0		0.0000
addr_state	0	0		0.0000
revol_util	0	0		0.0000
new_cr_sr	0	0		0.0000

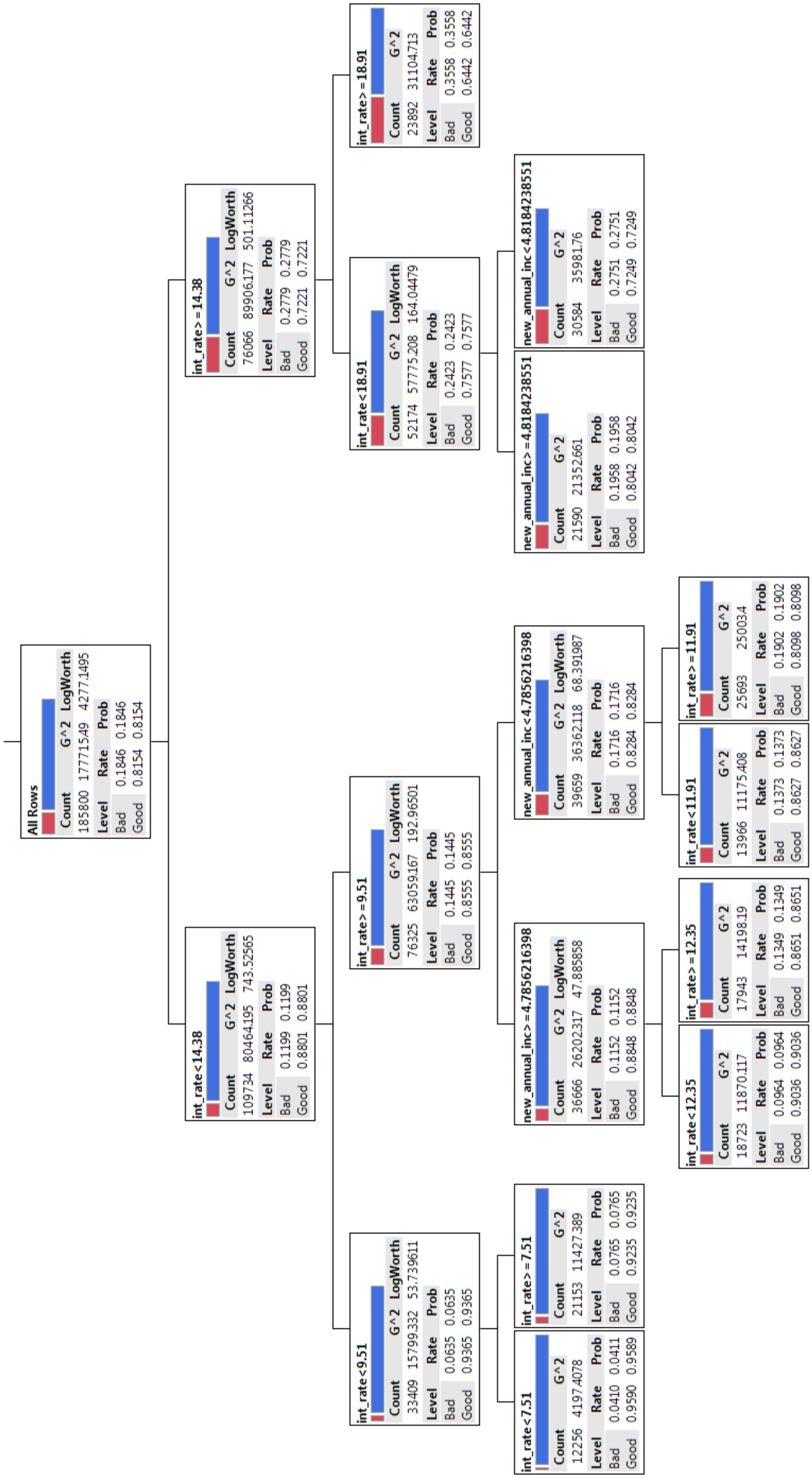
Tree Structure -- Full



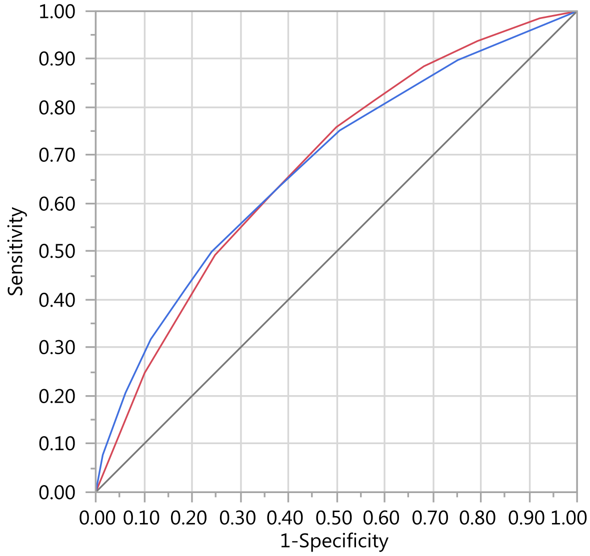


<14.38 int_rate
>=9.33 new_annual_inc
<9.33 int_rate
<4.694640292 int_rate
>=4.694640292 int_rate
>=12.35 term
<12.35 term
60 months
36 months new_emp_length
<0.5
>=0.5 loan_amnt
>=3750
<3750
60 months
36 months new_emp_length
<0.5
>=0.5
>=12.35 term
60 months new_annual_inc
<4.8512583487
>=4.8512583487
36 months int_rate
>=13.65 mths_since_last_delinq
<15 or Missing
>=15 not Missing
<13.65
<12.35 inq_last_6mths
>=3 not Missing inq_last_6mths
>=7 not Missing
<7 or Missing
<3 or Missing term
60 months
36 months open_acc
>=7 not Missing int_rate
>=11.55
<11.55
<7 or Missing
>=7.9 new_emp_length
<0.5
>=0.5
<7.9 int_rate
>=6.49 purpose
vacation, moving, small_business, renewable_energy
wedding, medical, home_improvement, educational, car, major_purchase, debt_consolidation, credit_card, house, other new_emp_length
<0.5
>=0.5
<6.49

Tree Structure -- Top

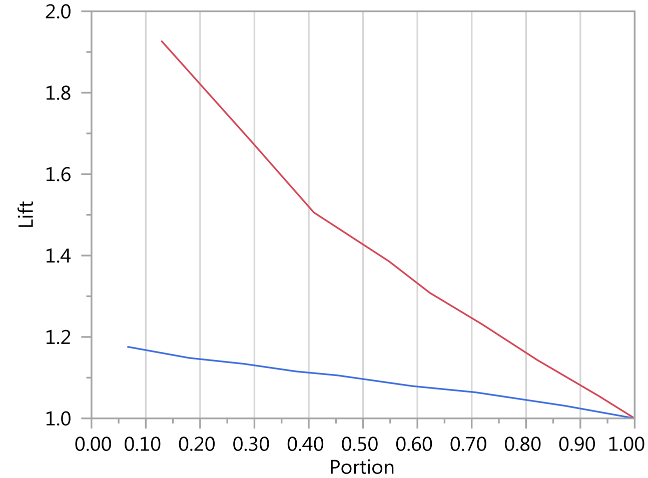


Receiver Operating Characteristic



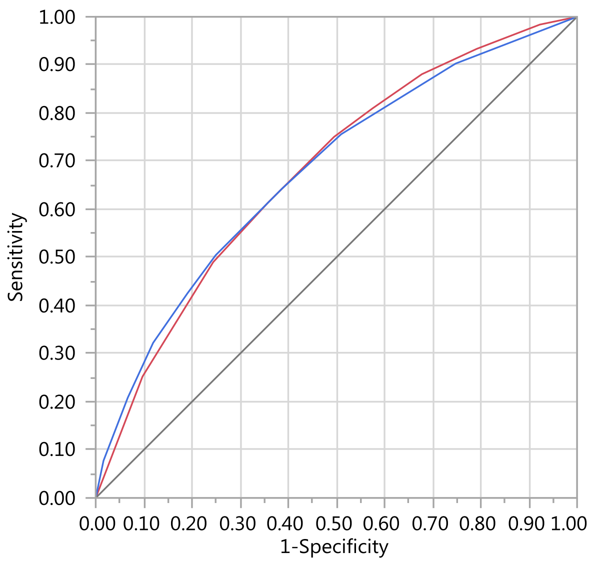
new_loan_status	Area
Bad	0.6774
Good	0.6774

Lift Curve



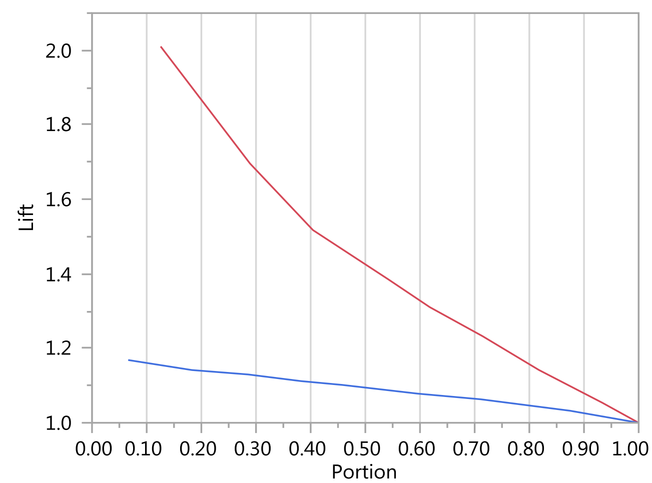
new_loan_status
Bad
Good

Receiver Operating Characteristic on Validation Data



new_loan_status	Area
Bad	0.6780
Good	0.6780

Lift Curve on Validation Data



new_loan_status
Bad
Good

4.2 Random Forest

Bootstrap Forest for new_loan_status

Specifications

Target Column: new_loan_status

Number of trees in the forest:	68
Number of terms sampled per split:	20
Training rows:	185695
Validation rows:	62106
Test rows:	0
Number of terms:	20
Bootstrap samples:	185695
Minimum Splits Per Tree:	10
Minimum Size Split:	247

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.0979	0.0789	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1450	0.1180	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4305	0.4405	$\sum$ -Log(ρ[j])/n
RMSE	0.3688	0.3728	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.2751	0.2780	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.1826	0.1844	$\sum$ (ρ[j]≠ρMax)/n
N	185695	62106	n

Confusion Matrix

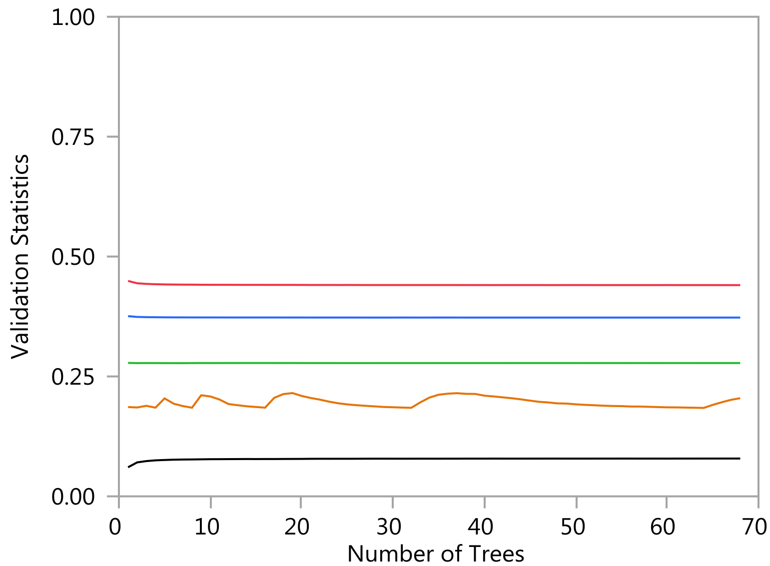
Training

Actual	Predicted	
new_loan_status	Bad	Good
Bad	557	33581
Good	336	151221

Validation

Actual	Predicted	
new_loan_status	Bad	Good
Bad	133	11327
Good	126	50520

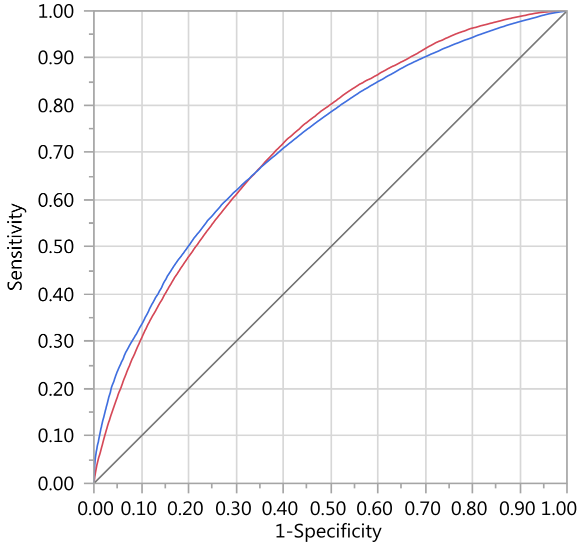
Cumulative Validation



Rsquare

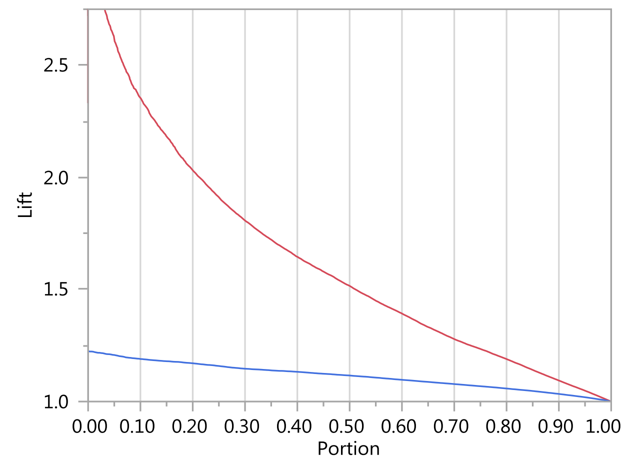
Avg -Log p
RMS Error
Avg Abs Error
MR

Receiver Operating Characteristic



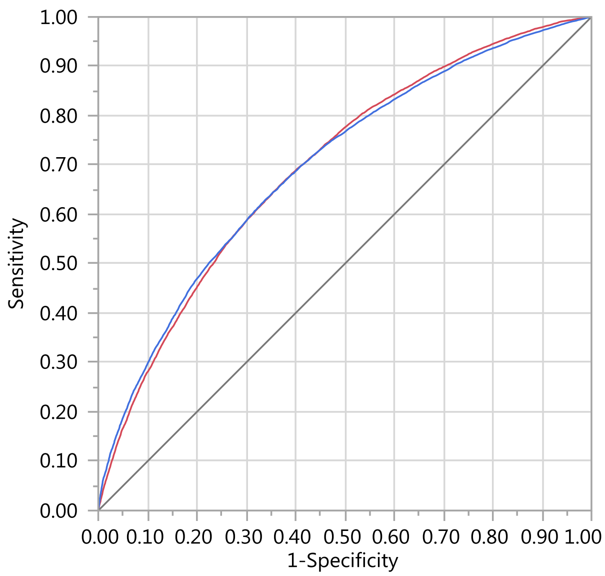
new_loan_status	Area
Bad	0.7196
Good	0.7196

Lift Curve



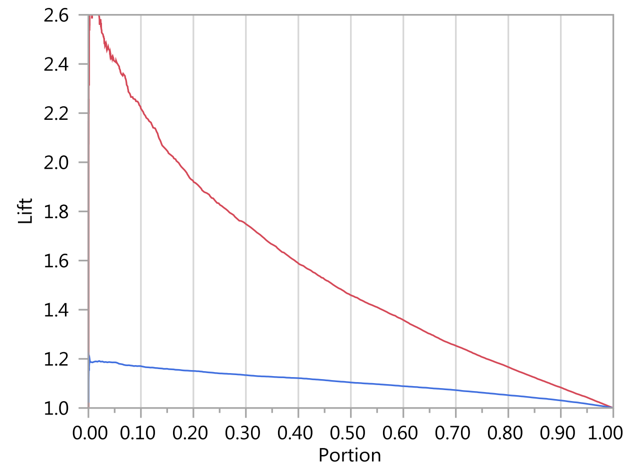
new_loan_status
Bad
Good

Receiver Operating Characteristic on Validation Data



new_loan_status	Area
Bad	0.6984
Good	0.6984

Lift Curve on Validation Data



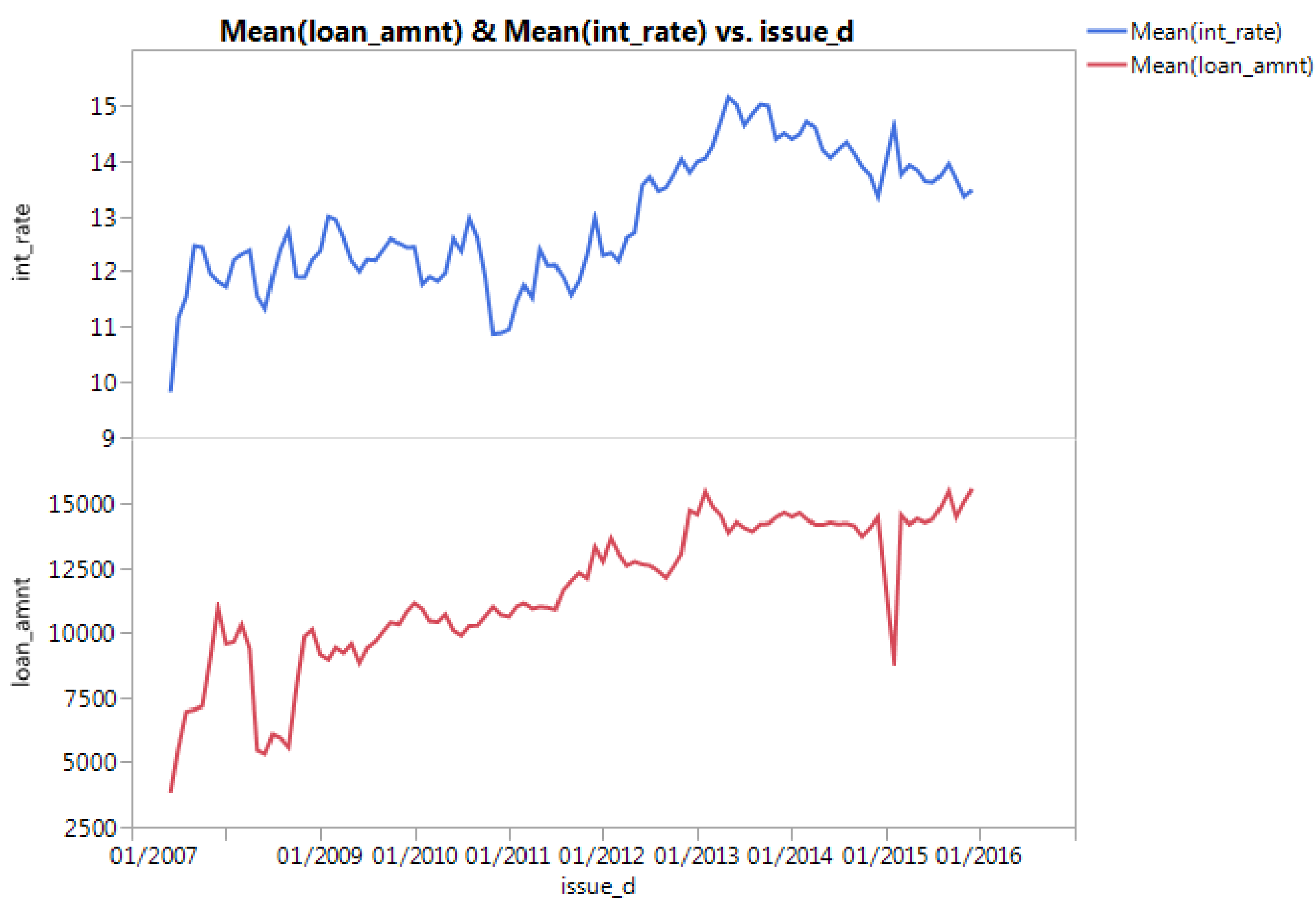
new_loan_status
Bad
Good

Column Contributions

Term	Number of Splits	G^2	Portion
int_rate	3026	7044.25089	0.6724
new_annual_inc	2142	1150.89978	0.1099
term	596	405.923429	0.0387
new_emp_length	3002	278.090206	0.0265
open_acc	2522	223.02119	0.0213
loan_amnt	1210	219.258227	0.0209
total_acc	2108	199.985721	0.0191
new_home_ownership	1839	178.709418	0.0171
purpose	646	143.544778	0.0137
mths_since_last_delinq	1282	139.989527	0.0134
inq_last_6mths	1887	139.364982	0.0133
verification_status	2761	130.723932	0.0125
dti	245	74.3347963	0.0071
new_desc_length3	272	50.3067261	0.0048
pub_rec	299	35.2605321	0.0034
revol_util	190	35.2032815	0.0034
new_cr_sr	104	14.672964	0.0014
delinq_2yrs	116	9.43221194	0.0009
installment	27	2.09946771	0.0002
addr_state	1	0.77854513	0.0001

5 Descriptive Analysis Visualization

5.1 Loan Amount and Interest Rate over Time



5.2 Geographic Indication of Loan Performance

	new_loan_status	
	Bad	Good
addr_state	Row %	Row %
AK	14.53%	85.47%
AL	**21.21%**	**78.79%**
AR	**20.01%**	**79.99%**
AZ	17.99%	82.01%
CA	17.56%	82.44%
CO	14.44%	85.56%
CT	16.89%	83.11%
DC	10.35%	89.65%
DE	18.88%	81.12%
FL	**20.77%**	**79.23%**
GA	17.58%	82.42%
HI	19.04%	80.96%
IA	**23.08%**	**76.92%**
ID	11.11%	88.89%
IL	17.03%	82.97%
IN	**23.15%**	**76.85%**
KS	17.23%	82.77%
KY	**20.03%**	**79.97%**
LA	19.81%	80.19%
MA	17.06%	82.94%
MD	18.87%	81.13%
ME	0.00%	100.00%
MI	19.84%	80.16%
MN	18.58%	81.42%
MO	**20.39%**	**79.61%**
MS	**22.74%**	**77.26%**
MT	14.27%	85.73%
NC	19.39%	80.61%
ND	0.00%	100.00%
NE	15.91%	84.09%
NH	14.24%	85.76%
NJ	19.65%	80.35%
NM	19.93%	80.07%
NV	**21.77%**	**78.23%**
NY	19.65%	80.35%
OH	19.32%	80.68%
OK	19.89%	80.11%
OR	16.65%	83.35%
PA	19.17%	80.83%
RI	18.17%	81.83%
SC	16.30%	83.70%
SD	16.89%	83.11%
TN	**24.35%**	**75.65%**
TX	16.30%	83.70%
UT	16.76%	83.24%
VA	18.22%	81.78%
VT	16.71%	83.29%
WA	17.22%	82.78%
WI	18.00%	82.00%
WV	14.40%	85.60%
WY	13.38%	86.62%