

Credit Risk Modeling

Introduction¹

Many argue that the United States is in the middle of a housing Market correction caused by drastically low housing inventory and historically low interest rates. Others point out that, unlike the previous Housing cycle, we have 17% more Total Payroll Jobs, we virtually have no subprime loans (a major issue with the previous housing cycle), and our delinquency or foreclosure rates remain seriously low.² Financial Institutions had quite a time determine who was a good customer and who was more likely to default on their mortgage loans. This project will focus on historical data, some of which might be on a credit report, provided from a financial institution in order to classify a customer into a “good” or “bad” credit risk for the financial institution.

Our given data set has thirteen variables and 5,960 undated observations or rows. For our data set, we are looking at ten

continuous variables of credit factors:

How much was the loan (**LOAN**), How

much they need to pay on their

mortgage (**MORTDUE**), Assessed

valuation (**VALUE**), Years on the Job

(**YOJ**), Number of Derogatory Reports

(**DEROG**), Number of Delinquent

Trade Lines (**DELINQ**), Age of Oldest

Trade Line (**CLAGE**), Number of recent credit inquiries (**NINQ**), Number of trade lines (**CLNO**),

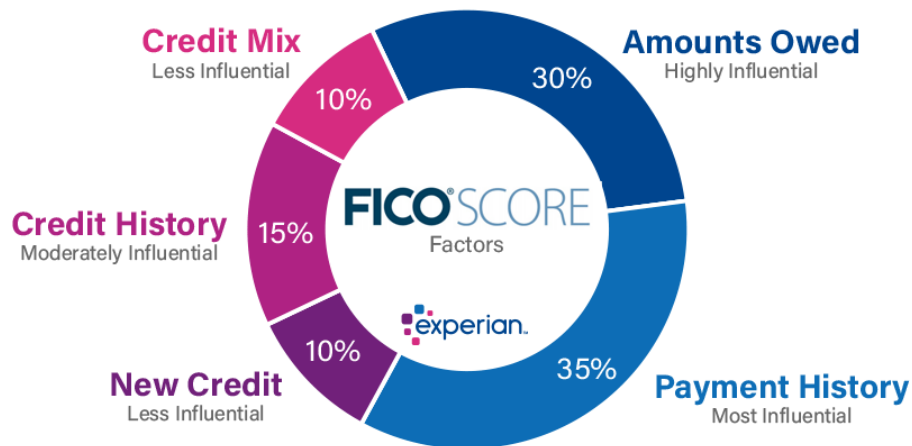
and Debt to income as a percentage (**DEBTINC**). The data set also has three nominal or

categorical descriptor variables: Reason for Loan (**REASON**) and Broad Job Category (**JOB**) in

order to predict the credit worthiness of an individual customer (**BAD**). For the analysis of

determining which of the variables predict the **BAD** credit risk, we will use all the given variables.

Following the risk factors as shown in the graphic above, it would be reasonable to expect that **JOB** category would not heavily weigh on Credit risk as tenure or experience in a job would matter more than the job title itself. It would also be reasonable that **DEROG** would affect the credit risks the most as it is a pattern of either good or poor financial behavior. Likewise, **DEBTINC** would increase in a direct relationship to **BAD** credit risk because the more you owe, the riskier you are.



¹ <https://www.experian.com/blogs/ask-experian/how-is-your-credit-score-determined/>

² <https://cdn.nar.realtor/sites/default/files/documents/2022-12-01-market-update-and-outlook-lawrence-yun-presentation-slides-2022-realtor-party-training-conference-12-05-2022.pdf>

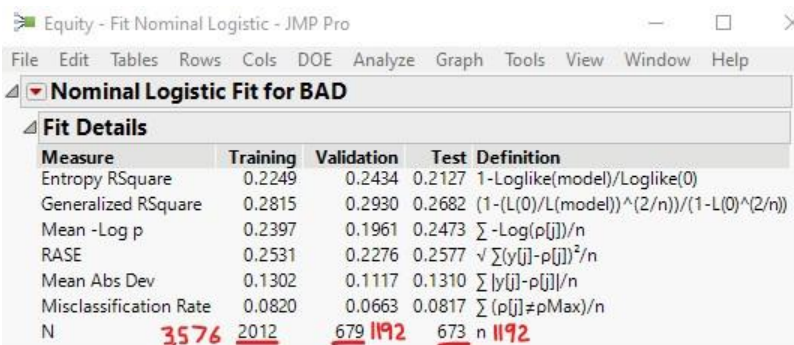
Credit Risk Modeling

Analysis and Model Comparison

For the data set, a Validation column will be created to allow cross-validation to occur within the modeling training phase at the same time estimation of the parameters of variables occurs. The training phase will contain **3,576** observations/rows or **60%**. The Validation data subset which works to determine when to stop when it has found that the model is no longer improving, has **1,192** observations or **20%**. Lastly, the Testing data subset, also **20%**, is used to determine how good the final model is once the Training phase is complete (assessing the generalization of error)³.

The Nominal Logistic Regression Model in JMP is chosen first because the predictor variable is binary, and an investigation is necessary to determine whether or not there is missing data. Often

times, the customer neglects to fill out every line on the form because they think it will negatively impact their ability to obtain the loan. As seen in the graphic, the Nominal method left out **1,564** or **43.7%** of the observations in the Training data because of missing information. Likewise, it left out **513** or **43.0%** of the



Measure	Training	Validation	Test	Definition
Entropy RSquare	0.2249	0.2434	0.2127	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.2815	0.2930	0.2682	$(1 - (L(0)/L(\text{model}))^{2/n}) / (1 - L(0)^{2/n})$
Mean -Log p	0.2397	0.1961	0.2473	$\sum -\text{Log}(p[j])/n$
RASE	0.2531	0.2276	0.2577	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.1302	0.1117	0.1310	$\sum y[j] - p[j] /n$
Misclassification Rate	0.0820	0.0663	0.0817	$\sum (p[j] \neq pMax)/n$
N	3576	2012	679	1192

Validation and of the Testing subsets. Since we are using JMP software, if we do not check the “Informative Missing” option, “each row with a missing value on that predictor is randomly assigned to one of the two sides of the split”⁴. By checking the “Informative Missing” option, the values are sorted, added to the low end of sorted values, and the splits are conducted. Then the process is repeated putting them on the high end of the sorted values and again the splits are constructed. Ultimately, these values are used in the calculations of the final model rather than being left out. For each of the models used, a model with “informative Missing” will be checked and another model without the option will be applied.

From the initial Nominal Logistic Regression, a model using the Adaptive Elastic Net will be applied. This model combines the L1 penalty of the Lasso Regression and the L2 penalty of the Ridge Regression and keeps correlated variables rather than zeroing them out. Unlike the tree ensemble models, the Adaptive Elastic Net will allow a view into Parameter Estimates which can be useful in the final analysis. Both the 1st two models perform better with “Informative Missing” checked.

³ <https://community.jmp.com/t5/German-Swiss-Austrian-JMP-Users/Advanced-and-Predictive-Analytics-with-JMP-PRO-Silvio-Miccio-pdf/m-p/23825>

⁴ <https://www.jmp.com/support/help/en/17.0/index.shtml#page/jmp/informative-missing-2.shtml>

Credit Risk Modeling

The next models to be used is the Bootstrap Forest (aka Random Forest) models and finally the Boosted Tree models (known as gradient boosting). Both are completely capable of handling missing data on their own and are very good at capturing non-linear data such as outliers. Interesting enough, it doesn't even matter if you check the "Informative Missing" option for these models (as shown on the far right in the Bootstrap model comparison picture). Both models are ensembles of decision trees but are different in the training process and how they combine the individual tree's outputs: Bootstrap Forests are collections that have independent trees averaged together, whereas Boosted Trees are built on top of each other in a self-correcting way.⁵

Validation	Creator	.2	.4	.6	.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N
Training	Bootstrap Forest					0.6083	0.7229	0.1994	0.2314	0.1579	0.0705	3576
Training	Bootstrap Forest					0.7413	0.8295	0.1317	0.1854	0.1051	0.0439	3576
Training	Bootstrap Forest					0.6228	0.7352	0.1921	0.2277	0.1518	0.0671	3576
Training	Bootstrap Forest					0.7512	0.8370	0.1266	0.1815	0.1014	0.0428	3576
Validation	Bootstrap Forest					0.4037	0.5217	0.2888	0.2923	0.1967	0.1191	1192
Validation	Bootstrap Forest					0.5527	0.6682	0.2167	0.2553	0.1400	0.0990	1192
Validation	Bootstrap Forest					0.4104	0.5288	0.2856	0.2917	0.1928	0.1200	1192
Validation	Bootstrap Forest					0.5421	0.6585	0.2218	0.2577	0.1386	0.0931	1192
Test	Bootstrap Forest					0.3979	0.5159	0.2924	0.2919	0.1993	0.1099	1192
Test	Bootstrap Forest					0.4918	0.6111	0.2468	0.2722	0.1531	0.1057	1192
Test	Bootstrap Forest					0.4112	0.5299	0.2859	0.2897	0.1935	0.1107	1192
Test	Bootstrap Forest					0.5010	0.6201	0.2423	0.2703	0.1490	0.1065	1192

JMP Alert



The starting number of terms to select, 9, is greater than the maximum terms to select, 6, so it will not do multiple fits.

OK

✕ a final result. However, JMP sent out an error message (seen on the left) that resulted in some of that randomness not being used. So, there were 4 Bootstrap Forest models: 1- "Informative Missing" checked + Default setup with 9 terms, "Informative

Missing" unchecked + Default setup with 9 terms, "Informative Missing" checked + suggested setup with 6 terms, and "Informative Missing" unchecked + suggested setup with 6 terms. All of these were compared to each other in a Model Comparison Tool with the "Informative Missing" checked + Default setup with 9 terms being the best Bootstrap Forest Model.

After the best of each of the 4 models there was a clear winner: the Boosted Tree which didn't care if the "Informative Missing" option was checked or not (as seen on the top of the next page). You can see that the best model of all 5 types included 3,576 Training observations as well as 1,192

⁵ <https://www.baeldung.com/cs/gradient-boosting-trees-vs-random-forests>

Credit Risk Modeling

Training observations and Testing observations. The Random Forest and the Boosted Tree come in close competition with each other but both the estimated R^2 values are higher for the Boosted Tree and the Root Mean 2 Error (RASE) was 2.85% lower for the Boosted Tree. The Misclassification rate showed that the Nominal Logistic, the Adapted Elastic Net, and the Bootstrap Forest had similar errors but the Boosted Tree walked away with only 8.05% which was pretty low. The Bootstrap Forest Final model had 189 Layers, 15 Splits, a learning rate of 12.8% and an overfit Penalty of 0.0001.

Equity - Model Comparison - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

Model Comparison

Target BAD missing a predictor for category Good Risk
Target BAD missing a predictor for category Good Risk
Target BAD missing a predictor for category Good Risk
Target BAD missing a predictor for category Good Risk

Predictors

Measures of Fit for BAD

Validation	Creator	.2	.4	.6	.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N
Training	Fit Generalized Adaptive Elastic Net					0.4590	0.5845	0.2754	0.2836	0.1630	0.1057	3576
Training	Bootstrap Forest					0.7512	0.8370	0.1266	0.1815	0.1014	0.0428	3576
Training	Fit Nominal Logistic					0.4590	0.5845	0.2754	0.2837	0.1628	0.1057	3576
Training	Boosted Tree					0.8443	0.9028	0.0793	0.1324	0.0642	0.0218	3576
Validation	Fit Generalized Adaptive Elastic Net					0.4812	0.6005	0.2513	0.2730	0.1537	0.1074	1192
Validation	Bootstrap Forest					0.5421	0.6585	0.2218	0.2577	0.1386	0.0931	1192
Validation	Fit Nominal Logistic					0.4817	0.6010	0.2511	0.2728	0.1535	0.1074	1192
Validation	Boosted Tree					0.6020	0.7122	0.1928	0.2326	0.1087	0.0721	1192
Test	Fit Generalized Adaptive Elastic Net					0.4134	0.5322	0.2849	0.2878	0.1640	0.1141	1192
Test	Bootstrap Forest					0.5010	0.6201	0.2423	0.2703	0.1490	0.1065	1192
Test	Fit Nominal Logistic					0.4121	0.5308	0.2855	0.2882	0.1640	0.1166	1192
Test	Boosted Tree					0.5815	0.6944	0.2032	0.2418	0.1144	0.0805	1192

Interpretation

Equity - Boosted Tree of BAD - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help

Boosted Tree for BAD

Column Contributions

Term	Number of Splits	G^2	Portion
REASON	369	265737.113	0.3014
NINQ	491	115668.39	0.1312
DELINQ	412	100634.749	0.1142
YOJ	271	92497.0839	0.1049
DEROG	332	89832.7985	0.1019
JOB	293	83963.0059	0.0952
CLNO	226	77012.2133	0.0874
LOAN	90	23484.2404	0.0266
DEBTINC	90	12110.2979	0.0137
CLAGE	105	9623.50669	0.0109
VALUE	74	6508.60665	0.0074
MORTDUE	82	4500.13786	0.0051

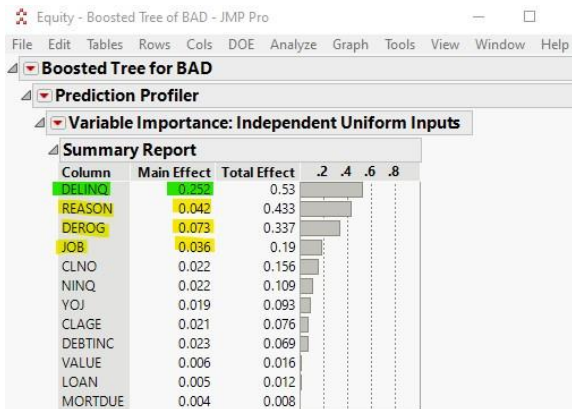
After selecting the best estimating model, the untouched testing subset can then be used to estimate variable parameters. This advanced type of modeling has a built-in “Column Contribution” tool (as compared to the Nominal Logistic Regression which would have to use the Profile “Variable Importance” tool). This tool shows that the REASON variable (Reason for the Loan) is the most important or influential variable at 30.14%. So, a financial institution should pay careful

attention to whether they are working on Home Improvements or not. However, the next four variables (NINQ – Number of recent credit inquiries, DELINQ – number of delinquent trade lines, YOJ – Years on the Job, and DEROG – Number of Derogatory Reports) are similar in influence and altogether are more important the REASON. However, since MORTDUE (how much money is due on the mortgage) is at the bottom of our chart, it is the least important variable.

Our initial assumptions about JOB (Job Category) being a heavy influencer were correct, but the variable did wind up in the middle of our “Column Contribution” table. The prediction the DEROG

Credit Risk Modeling

would be at the top of our table is incorrect as it is directly above JOB. Our assumption that DEBTINC (Debt to income ratio) would be more important was also shown false. Overall, the model showed different results more often than was expected.



To compare, the “Variable Importance” tool (which is created from the Profiles section) shows DELINQ (Number of delinquent trade lines) as the most important variable with Reason coming in 2nd. However, DELINQ has a significant difference between the Main effect and the Total effect, as does DEROG and JOB. This demonstrates variable interactions between at least 2 variables. This time, the “Variable Importance” Tool is less effective at determining

relationships between the predictor variable “BAD” and the other variables. However, again MORTDUE (how much money is due on the mortgage) is at the bottom of our chart, and it is the least important variable.

Even though financial advisor, Dave Ramsey completely advises against any kind of debt⁶, he has been heard to say if you have to have debt, only have debt connected to a tangible asset, such as a house; don’t get into debt without backing of an asset such as in a credit card. So, he might agree with homeowners getting home equity loans for a good REASON. Some in opposition to Dave Ramsey’s philosophy explain that an Equity Loan or Home Equity Line of Credit (HELOC) can be useful for its lower rates⁷ (compared to unsecured debt) and it can be a tax write-off. Either way, Home Equity Loans can have good uses like debt consolidation, home improvements, and tuition, as well as not so good uses like going on vacation, spending it on one-time events, and luxury items.

⁶ <https://www.ramseysolutions.com/real-estate/what-is-a-home-equity-loan>

⁷ <https://lendedu.com/blog/what-does-dave-ramsey-think-about-home-equity-loans/>