Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

Introduction

For assignment seven - Equity, we examine the variables that effect whether a person will have a good or bad credit risk. The possible determinants of BAD are LOAN, MORTDUE, VALUE, REASON, JOB, YOJ, DEROG, DELINQ, CLAGE, NINQ, and DEBTINC. The Y (response variable) is Bad, which is a categorical variable. For the assignment we want to determine how these variables effect the equity loan and whether a person is a good or bad credit risk.

The data set consists of the following variables:

BAD	Whether the customer is a good or bad credit risk. 1 is a bad risk.
LOAN	Value of loan taken out
MORTDUE	How much they need to pay on their mortgage
VALUE	Assessed valuation
REASON	Reason of loan
JOB	Broad job category
YOJ	Years on job
DEROG	Number of Derogatory Reports
DELINQ	Number of Delinquent Trade Lines
CLAGE	Age of Oldest Trade Line
NINQ	Number of recent credit inquiries.
CLNO	Number of trade lines
DEBTINC	Debt to income as a percentage

We used the following 26 methods to model the data set:

1. Nominal Regression
2. Ridge
3. Lasso
4. Adaptive Lasso
5. Elastic Net
6. Adaptive Elastic Net
7. Random Forest (Bootstrap Method)
8. Random Forest with no informative

9. Boosted Tree
10. Booster Tree with no informative
11. Neutral Net (informative) TanH3 (1 Level) and 20 tours
12. Neutral Net (informative) TanH3 (1 Level) and 20 tours with no informative
13. Neutral Net (with no informative) TanH3 (1 Level) and 20 tours
14. Neutral Net (with no informative) TanH3 (1 Level) and 20 tours with no informative
15. Neutral Net (informative) TanH3 (2 Level) and 20 tours
16. Neutral Net (informative) TanH3 (2 Level) and 20 tours with no informative
17. Neutral Net (informative) TanH3 (2 Level), 20 tours, and Booster 40
18. Neutral Net (informative) TanH3 (2 Level), 20 tours, and Booster 40 with no informative
19. Neutral Net (informative) TanH3 (2 Level), Linear (1 Level) and 20 tours
20. Neutral Net (informative) TanH3 (2 Level), Linear (1 Level) and 20 tours with no inform
21. Neutral Net (informative) All Three Activation Methods (2 Level)
22. Neutral Net (informative) All Three Activation Methods (2 Level) with no informative
23. Neutral Net (informative) TanH1 (1 Level), 20 tours, and Booster 40
24. Neutral Net (informative) TanH1 (1 Level), 20 tours, and Booster 40 with no informative
25. Neutral Net (informative) TanH1 (2 Level), and 20 tours
26. Neutral Net (informative) TanH1 (2 Level), and 20 tours with no informative

In all cases, we used random seed 123 for replication, Tours set to 20 and all cases were tested with and without the informative. By selecting Informative, JPM will use the data available to fill in any missing data in the dataset.

In the beginning of the process, I planned to only use Random Tree, Neural Net, and Boosted Tree. I felt these methods would be the best methods for classification prediction variable, because these methods are less likely to overfit the testing data set. In the end, I elected to use all methods except for Ordinary Least Squares, I decided to use Nominal Regression which is a better method for categorical values. The main reason for using all methods, I wanted to better understand how they perform on the dataset.

Analysis and Model Comparison

As mentioned above, we are using twenty-six methods to test our data set.

1. Nominal Regression is a linear regression model that uses the principle of least squares for estimating the unknown variables in the data set.

   Disadvantage - In the case of larger models, this method can overfit the model. When a model overfits, it fits the training data well, but provides poor results for the unbiased test data.

2. Ridge is a penalize regression method that use variable selection and shrinkage. This method imposes a constraint on the sum of the squared value of the β in the model, which causes some variables to shrink towards zero, but no all the way to zero. This method reduces the impact of the variable.

3. Lasso (Least Absolute Selection Shrinkage Operator) is a penalize regression method that use variable selection and shrinkage. The Lasso method imposes a constraint on the sum of the absolute value of the β in the model, which causes some variables to shrink all the way to zero. Thus, eliminating their impact. The model uses Lambda (λ) is considered a tuning parameter to control the strength of the penalty. As lambda increases, more variables are reduced to zero.

   Advantages – It identifies the most important variables that are linked to the response variable, RSVL for this project. It provides greater accuracy and increases model interpretability.
4. Elastic Net is another penalize regression method that uses variable selection and shrinkage. However, this method uses a combination of ridge and lasso regression. The Ridge method imposes a constraint on the square of the β of the model and shrinks variables towards zero. Elastic Net combines the two penalties created by the Ridge and Lasso method and applies a weight to those penalties. Both methods are used to cause variable shrinkage to zero.

   Advantages – Same as Lasso

5. Adaptive Lasso\Elastic Net. By selecting Adaptive in JMP, the OLS model is ran on the data set to create an estimate which will give us an idea of the likely

important variables. Once these likely variables are determined, they will be penalized less in the models – Ridge, Lasso and elastic net. Typically, the adaptive models perform much better.

6. The Random Forest method forms successive independent decision trees using a bootstrap sample method.

   Advantage: Random forest method is unlikely to have a problem with overfitting. Since each tree is constructed independently, the method is not affected by highly correlated variables.

7. The Neutral Net models complex relationship between inputs and outputs. It uses an input layer, hidden layer(s), and an outer layer. The hidden layer contains one or more nodes. The Neutral Net used three types of transformation functions – TanH, Linear, and Gaussian.

   Advantage – Very flexible and models complex relations.

   Disadvantage – The Neutral Net model can seem like a black box, since you are unable to view what is happening in the hidden layers.

8. Boosted tree is built by creating smaller decision trees, which predict the smaller residuals from the previous decision trees. The smaller trees are combined to form a larger tree.

The cross-validation procedure consists of creating a valuation column. This consists of dividing the data into training, validation, and test sets. Since this data set is cross validation, we will use JMP Pro 15 (JMP) to create our validation column. The column consists of a random selection of Training, Validation, and Test using a 60%/20%/20% split with a random seed of 123.

We used the statistical software JMP to provide our analysis of the Equity data set. In JMP, we selected *Analyze*, then *Fit Model*. In the next window, we set up our analysis of the data set. Under *Pick Role Variables*, we selected BAD as our *Y* variable, Validation as our *Validation* variable and added all the variables under *Construct Model Effects*. Next, we updated the *Personality* by selecting *General Regression*, and then *we Run* the model. In the next window, we selected either Ridge, Lasso, Elastic Net, and Go to create the model

For the Neutral Net models, we selected *Analyze*, Predictive *Model and then Neutral*. In the next window, we use Equity set up our analysis of the data set. Under *Pick Role Variables*, we selected BAD as our *Y* variable, Validation as our *Validation* variable and added the predictor variables under *Construct Model Effects* and press OK. In the next window called Model launch, we kept the default and pressed Go. The process above is repeated for the complex Neutral Net. On the Model Launch page, we update the fields for the new model and Go to create the model. We saved the scripts and formula data for each model.

For the Booster models, we selected *Analyze*, Predictive *Model and then Booster Tree*. In the next window, we use Equity set up our analysis of the data set. Under *Pick Role Variables*, we selected BAD as our *Y* variable, Validation as our *Validation* variable and added the predictor variables under *Construct Model Effects* and press OK. In the next window called Model launch, we kept the default, added 123 as the random seed and pressed Go. We also updated Tours and Booster as needed. We saved the scripts and formula data for each model.

In all of models created I used Tour 20. By using Tour in JPM, the program will run the model twenty times and the final result is the average.

Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

Target Predictors (BAD)

Category	Probability Column	
Good Risk	Prob[Good Risk] - Nominal	Fit Nominal Logistic
Good Risk	Probability(BAD=Good Risk) Ridge	Fit Generalized Ridge
Good Risk	Probability(BAD=Good Risk) 2 Lasso	Fit Generalized Lasso
Good Risk	Probability(BAD=Good Risk) 3 Adp Lasso	Fit Generalized Adaptive Lasso
Good Risk	Probability(BAD=Good Risk) 4 Elastic Net	Fit Generalized Elastic Net
Good Risk	Probability(BAD=Good Risk) 5 Adp Elastic Net	Fit Generalized Elastic Net
Good Risk	Prob(BAD==Good Risk) Random Tree	Bootstrap Forest
Good Risk	Prob(BAD==Good Risk) 3 Random Forest no informative	Bootstrap Forest
Good Risk	Prob(BAD==Good Risk) 2 Boosted Tree	Boosted Tree
Good Risk	Prob(BAD==Good Risk) 4 Boosted Tree no informative	Boosted Tree
Good Risk	Probability(BAD=Good Risk) NN T3 L1 T20 inform on	Neural
Good Risk	Probability(BAD=Good Risk) NN T3 L1 T20 inform off	Neural
Good Risk	Probability(BAD=Good Risk) NN T3 L2 T20 inform on	Neural
Good Risk	Probability(BAD=Good Risk) T3 L2 T20 OFF	Neural
Good Risk	Probability(BAD=Good Risk) NN T3 L1 T20 B40	Neural
Good Risk	Probability(BAD=Good Risk) T3 L1 T20 Off B40	Neural
Good Risk	Probability(BAD=Good Risk) NN T3 L2 G1 T20	Neural
Good Risk	Probability(BAD=Good Risk) NN T3 L2 G1 T20 OFF	Neural
Good Risk	Probability(BAD=Good Risk) NN All3 T20	Neural
Good Risk	Probability(BAD=Good Risk) NN All T20 OFF	Neural
Good Risk	Probability(BAD=Good Risk) NN T1 L1 T20 B40	Neural
Good Risk	Probability(BAD=Good Risk) NN T1 L1 T20 OFF	Neural
Good Risk	Probability(BAD=Good Risk) T1 L2 T20	Neural

Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

The best method is the Neutral Net (NN) method TanH, 1 layer, Tour 20, and 40 Boosted with an AUC of 94.54%. This means this model predicted 94.54% of the testing data.

This model also has a good Confusion matrix. Of the Good predictions only 31 were misclassified. Meaning they were predicted as Good but should have been labeled Bad. Of the Bad predictions 70 were misclassified as Good. Overall, this method had an 84.73% misclassification rate.

Neural

Validation Column: Validation
Informative Missing

Model Launch

Model NTanH(3)NBoost(37)

Training

BAD

Measures	Value
Generalized RSquare	0.7910087
Entropy RSquare	0.6911429
RMSE	0.2115937
Mean Abs Dev	0.1068805
Misclassification Rate	0.0584452
-LogLikelihood	562.30397
Sum Freq	3576

Confusion Matrix

Actual	Predicted Count	
BAD	Good Risk	Bad Risk
Good Risk	2779	59
Bad Risk	150	588

Confusion Rates

Actual	Predicted Rate	
BAD	Good Risk	Bad Risk
Good Risk	0.979	0.021
Bad Risk	0.203	0.797

Validation

BAD

Measures	Value
Generalized RSquare	0.7469943
Entropy RSquare	0.6427024
RMSE	0.2251451
Mean Abs Dev	0.1112947
Misclassification Rate	0.0746644
-LogLikelihood	206.31286
Sum Freq	1192

Confusion Matrix

Actual	Predicted Count	
BAD	Good Risk	Bad Risk
Good Risk	939	28
Bad Risk	61	164

Confusion Rates

Actual	Predicted Rate	
BAD	Good Risk	Bad Risk
Good Risk	0.971	0.029
Bad Risk	0.271	0.729

Test

BAD

Measures	Value
Generalized RSquare	0.6793981
Entropy RSquare	0.5647104
RMSE	0.2462626
Mean Abs Dev	0.1252324
Misclassification Rate	0.0847315
-LogLikelihood	251.98093
Sum Freq	1192

Confusion Matrix

Actual	Predicted Count	
BAD	Good Risk	Bad Risk
Good Risk	935	31
Bad Risk	70	156

Confusion Rates

Actual	Predicted Rate	
BAD	Good Risk	Bad Risk
Good Risk	0.968	0.032
Bad Risk	0.310	0.690

On the next page is a comparison of the twenty-six methods used on the data. To determine the best method, it is important to only look at the results for the test set. The best method is the Neutral Net (NN) method TanH, 1 layer, Tour 20, and 40 Boosted with an AUC of 94.54%.

Also note how much using Informative in JPM improves the results of the model. The models without Informative are almost 10% less. It just confirms how important it is to have complete data in the beginning or use JPM.

Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

Measures of Fit for BAD

Validation	Creator	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Nominal Logistic	0.2249	0.2815	0.2397	0.2531	0.1302	0.0820	2012	0.8027
Training	Fit Generalized Ridge	0.2249	0.2815	0.2397	0.2532	0.1304	0.0825	2012	0.8029
Training	Fit Generalized Lasso	0.2249	0.2815	0.2397	0.2531	0.1302	0.0820	2012	0.8027
Training	Fit Generalized Adaptive Lasso	0.2249	0.2815	0.2397	0.2531	0.1303	0.0820	2012	0.8027
Training	Fit Generalized Elastic Net	0.2249	0.2815	0.2397	0.2531	0.1302	0.0820	2012	0.8027
Training	Fit Generalized Adaptive Elastic Net	0.2249	0.2815	0.2397	0.2531	0.1303	0.0820	2012	0.8027
Training	Bootstrap Forest	0.7207	0.8140	0.1422	0.1927	0.1133	0.0467	3576	0.9933
Training	Bootstrap Forest	0.5808	0.6989	0.2134	0.2421	0.1651	0.0682	3576	0.9835
Training	Boosted Tree	0.4231	0.5479	0.2937	0.2907	0.2032	0.1163	3576	0.9272
Training	Boosted Tree	0.2776	0.3855	0.3678	0.3269	0.2732	0.1367	3576	0.8803
Training	Neural	0.5505	0.6717	0.2289	0.2574	0.1385	0.0886	3576	0.9387
Training	Neural	0.4402	0.5168	0.1731	0.2146	0.0965	0.0537	2012	0.8993
Training	Neural	0.5510	0.6722	0.2286	0.2582	0.1376	0.0853	3576	0.9401
Training	Neural	0.5023	0.5790	0.1539	0.2038	0.0858	0.0487	2012	0.9226
Training	Neural	0.6911	0.7910	0.1572	0.2116	0.1069	0.0584	3576	0.9778
Training	Neural	0.6055	0.6772	0.122	0.1808	0.0822	0.0422	2012	0.9738
Training	Neural	0.5516	0.6727	0.2283	0.2579	0.1375	0.0878	3576	0.9403
Training	Neural	0.3920	0.4668	0.188	0.2235	0.1014	0.0611	2012	0.8776
Training	Neural	0.6637	0.7691	0.1712	0.2239	0.1047	0.0691	3576	0.9675
Training	Neural	0.6252	0.6953	0.1159	0.1784	0.0704	0.0417	2012	0.9645
Training	Neural	0.4917	0.6166	0.2588	0.2753	0.1580	0.0951	3576	0.9251
Training	Neural	0.2638	0.3265	0.2277	0.2439	0.1197	0.0711	2012	0.7685
Training	Neural	0.4591	0.5846	0.2754	0.2837	0.1631	0.1065	3576	0.9116
Validation	Fit Nominal Logistic	0.2434	0.2930	0.1961	0.2276	0.1117	0.0663	679	0.8166
Validation	Fit Generalized Ridge	0.2427	0.2922	0.1963	0.2278	0.1119	0.0663	679	0.8165
Validation	Fit Generalized Lasso	0.2433	0.2929	0.1961	0.2277	0.1117	0.0663	679	0.8165
Validation	Fit Generalized Adaptive Lasso	0.2432	0.2928	0.1962	0.2277	0.1117	0.0663	679	0.8164
Validation	Fit Generalized Elastic Net	0.2433	0.2929	0.1961	0.2277	0.1117	0.0663	679	0.8165
Validation	Fit Generalized Adaptive Elastic Net	0.2432	0.2928	0.1962	0.2277	0.1117	0.0663	679	0.8164
Validation	Bootstrap Forest	0.5531	0.6686	0.2165	0.2516	0.1437	0.0864	1192	0.9522

Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

Validation	Creator	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Validation	Bootstrap Forest	0.3819	0.4985	0.2994	0.2934	0.2004	0.1057	1192	0.8992
Validation	Boosted Tree	0.4256	0.5446	0.2783	0.2787	0.1937	0.0948	1192	0.9241
Validation	Boosted Tree	0.2711	0.3723	0.3531	0.3167	0.2670	0.1233	1192	0.8858
Validation	Neural	0.5298	0.6470	0.2278	0.2557	0.1371	0.0889	1192	0.9317
Validation	Neural	0.3729	0.4345	0.1626	0.1979	0.0866	0.0442	679	0.8502
Validation	Neural	0.5271	0.6445	0.2291	0.2597	0.1349	0.0973	1192	0.9379
Validation	Neural	0.4145	0.4779	0.1518	0.1899	0.0756	0.0412	679	0.8713
Validation	Neural	0.6427	0.7470	0.1731	0.2251	0.1113	0.0747	1192	0.9672
Validation	Neural	0.5684	0.6309	0.1119	0.1689	0.0733	0.0353	679	0.9607
Validation	Neural	0.5236	0.6412	0.2308	0.2609	0.1368	0.0948	1192	0.9346
Validation	Neural	0.3776	0.4395	0.1613	0.2030	0.0897	0.0457	679	0.8732
Validation	Neural	0.5759	0.6892	0.2054	0.2486	0.1142	0.0856	1192	0.9541
Validation	Neural	0.5018	0.5662	0.1291	0.1818	0.0704	0.0324	679	0.9239
Validation	Neural	0.5074	0.6259	0.2386	0.2652	0.1493	0.0998	1192	0.9349
Validation	Neural	0.2681	0.3208	0.1897	0.2166	0.1038	0.0530	679	0.7884
Validation	Neural	0.4819	0.6013	0.251	0.2727	0.1537	0.1065	1192	0.9220
Test	Fit Nominal Logistic	0.2127	0.2682	0.2473	0.2577	0.1310	0.0817	673	0.7861
Test	Fit Generalized Ridge	0.2124	0.2678	0.2474	0.2578	0.1312	0.0817	673	0.7860
Test	Fit Generalized Lasso	0.2127	0.2681	0.2474	0.2577	0.1310	0.0817	673	0.7860
Test	Fit Generalized Adaptive Lasso	0.2128	0.2683	0.2473	0.2577	0.1310	0.0817	673	0.7862
Test	Fit Generalized Elastic Net	0.2127	0.2681	0.2474	0.2577	0.1310	0.0817	673	0.7860
Test	Fit Generalized Adaptive Elastic Net	0.2128	0.2683	0.2473	0.2577	0.1310	0.0817	673	0.7862
Test	Bootstrap Forest	0.4937	0.6130	0.2459	0.2710	0.1578	0.1015	1192	0.9355
Test	Bootstrap Forest	0.3705	0.4864	0.3057	0.2999	0.2056	0.1174	1192	0.8912
Test	Boosted Tree	0.3814	0.4982	0.3004	0.2936	0.2038	0.1133	1192	0.8993
Test	Boosted Tree	0.2441	0.3396	0.3671	0.3262	0.2727	0.1393	1192	0.8689
Test	Neural	0.4712	0.5910	0.2568	0.2715	0.1460	0.0982	1192	0.9129
Test	Neural	0.2768	0.3422	0.2272	0.2479	0.1137	0.0684	673	0.8257
Test	Neural	0.4629	0.5827	0.2608	0.2734	0.1453	0.0973	1192	0.9075
Test	Neural	0.2161	0.2721	0.2463	0.2399	0.1013	0.0654	673	0.8162
Test	Neural	0.5647	0.6794	0.2114	0.2463	0.1252	0.0847	1192	0.9454
Test	Neural	0.4089	0.4857	0.1857	0.2179	0.0960	0.0550	673	0.8790

Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

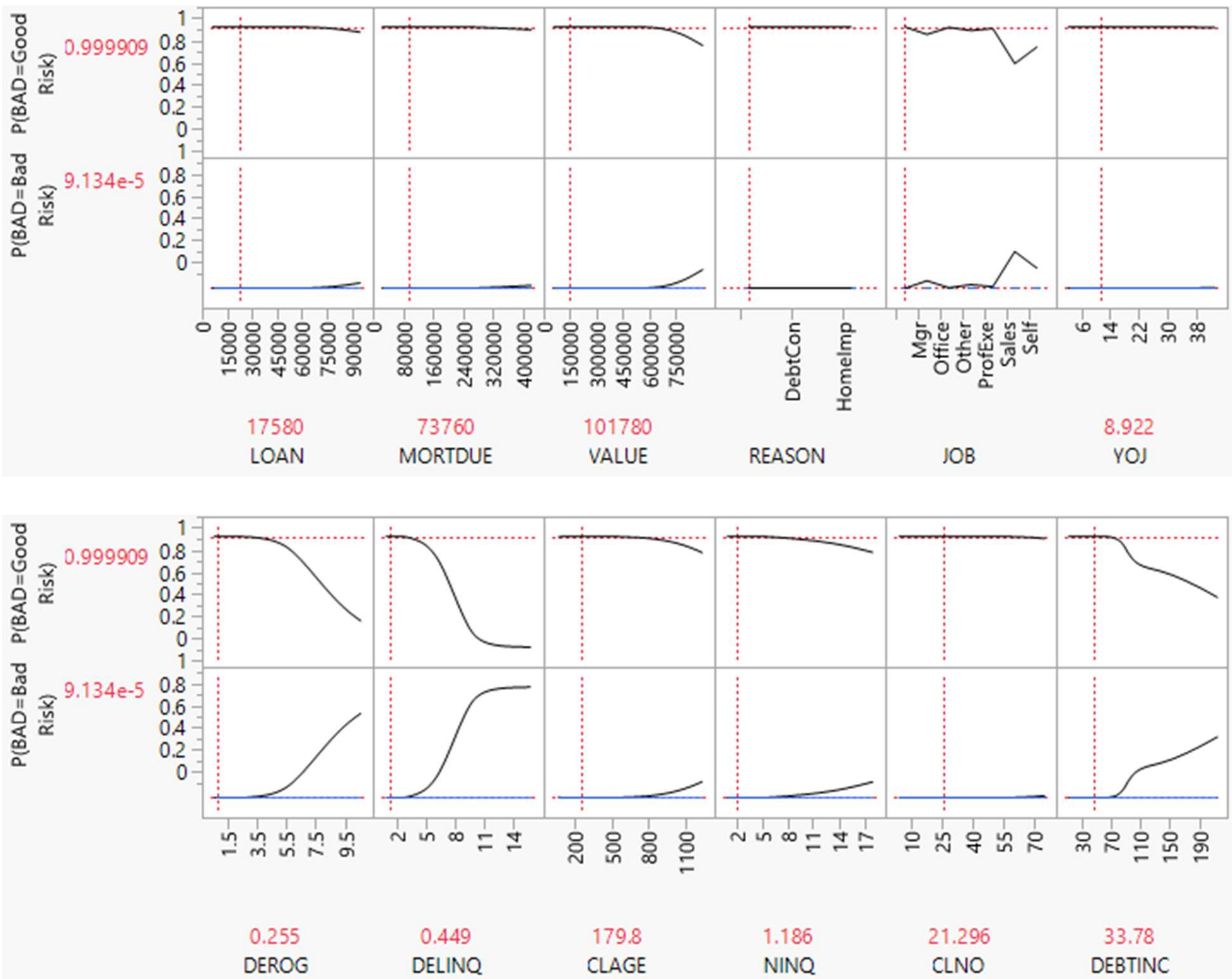
Validation	Creator	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Test	Neural	0.4791	0.5987	0.253	0.2703	0.1435	0.0956	1192	0.9179
Test	Neural	0.2287	0.2869	0.2423	0.2501	0.1134	0.0728	673	0.7952
Test	Neural	0.4609	0.5807	0.2618	0.2706	0.1267	0.0931	1192	0.9203
Test	Neural	0.3116	0.3812	0.2163	0.2267	0.0925	0.0550	673	0.8366
Test	Neural	0.4377	0.5573	0.2731	0.2827	0.1615	0.1116	1192	0.9021
Test	Neural	0.1962	0.2487	0.2525	0.2592	0.1266	0.0802	673	0.7536
Test	Neural	0.4121	0.5308	0.2855	0.2882	0.1642	0.1174	1192	0.8929

Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

Prediction Profiler



Loan, Mortdue, Value, Reason, and Job have little effect on whether a person will be a good or bad credit risk.

In the broad category of Job, if you are in sales you are more likely to be a bad credit. This may be due to seasonal effects of some positions.

The DEROG has a positive relationship with BAD. As DEROG increases, the likelihood that a person will be a bad credit risk increases.

Paula McCree Bailey

Module 7: Assignment 1 – Equity

October 11, 2020

The DEBTINC has a positive relationship with BAD. As DEBTINC increases, the likelihood that a person will be a bad credit risk increases.

The CLAGE and NINQ both have a slightly positive relationship with BAD. As they increase, the likelihood that a person will be a bad credit risk increases.

The CLNO has no effect on a person's credit risk.

Variable Importance: Independent Uniform Inputs
Summary Report
Overall

Column	Main Effect	Total Effect
DEBTINC	0.276	0.628
CLAGE	0.085	0.419
DELINQ	0.066	0.31
DEROG	0.024	0.187
VALUE	0.014	0.152
NINQ	0.012	0.149
JOB	0.02	0.147
MORTDUE	0.013	0.129
CLNO	0.013	0.079
YOJ	0.013	0.078
REASON	0.012	0.067
LOAN	0.01	0.065

Looking at the prediction variables with a total effect greater than 15%:
DEBTINC, the debt to income as a percentage is the most significant variable. It predicts 62.8% whether a person is a good or bad credit risk.

CLAGE predicts 41.9%.

DELINQ predicts 31%.

VALUE predicts 15.2%.