**MG 222, 2018**

**Assignment**



**Variable Selection and Profile Analysis**

From the given set of variables, we can identify (in a naive manner) which of the predictors(X) will affect the Credit Score(BAD) in which manner.

An intuitive list of such parameters would be:

**DOB**

We can calculate the AGE of a person, given their date of Birth. For the rest of this analysis, we will be considering AGE = 99-DOB for our analysis. (AGE = 0 indicates no data given)

Prior knowledge indicates that good credit scores are given to people in their mid-lives (25-45). They have a high propensity to earn and to spend. People aged <25 and >45 are usually considered as “non-earning” and have low credit ratings.

**NKID**

Number of children inversely affects credit rating. Having children is an expensive affair involving heavy expenditure. One could argue that being a parent makes one a cautious spender, so this assumption needs validation.

**DEP**

Increased number of Dependents affects credit rating inversely. Credit score goes down as dependents increase.

**PHON**

In 1999, having a phone was considered a rarity so owning one should indicate good credit.

**SINC and DAINC**

This gives the total income in the Household and a greater value gives a better indication of good credit.

**AES**

Type of employment (especially ones that indicate whether employed or unemployed) is a good indicator of credit.

**RES**

Homeowners tend to have good credit, provided their mortgage payments aren’t very high (DMORT). This term most likely involves interaction with DMORT.

**DHVAL**

Value of Home Asset, along with a low DMORT would indicate good credit. These terms may interact and we need to consider them in our Model later on.

**DOUTM/DOUTL/DOUTCC/DOUTHP**

These indicate the Outgoing expenditure for a person and inversely affect credit scores.

**BAD**

This is the response variable we need to predict. BAD = 1 indicates a bad credit score and BAD = 0 indicates a good credit score. (Technically, NOT BAD credit score)

Overall, each of the above variables seem to have an effect on Credit Scores(BAD) in one way or another. In order to determine which the important variables are, we can individually check for each of them as shown in 1.

**Variable Selection**

Among continuous variables:

DAINC, DOUTCC, DOUTHP, AGE, SINC, DOUTM are significant

Among categorical variables:

AES, RES are significant

**Profiling**

Here, we try to figure out if there is a significant difference in the values that fall in a particular category and their distribution of good or bad scores.

For categorical variables, we test the mean for each group(level) against the global mean to see if there is any significant difference.

As per the analysis (shown in 2), we see that there seems to be no significant differences in means except for the case of AES.

(We could discretize the continuous variables by binning them and performing the same tests here. After attempting that, I found no difference in the variables I found from Variable Selection 1 performed in 1. So I am omitting that result here.)

**Variable Selection 2 – Information Value(IV) and Weight of Evidence(WoE)**

Based on the Information Value for each of the given Predictors, we can sort them in the following order:

**Variable IV**

DAINC 0.228686369

AGE 0.197973342

AES 0.176406850

DOUTM 0.067657130

DMORT 0.060520102

RES 0.059269664

DHVAL 0.049779687

DOUTCC 0.031587658

DOUTL 0.026662236

SINC 0.019189252

NKID 0.017447368

DOUTHP 0.011486183

PHON 0.009115428

DEP 0.003923339

As a rule-of-thumb, we can eliminate those Variables from our model that has a IV < 0.05.

We get the following variables that can be used:

**DAINC, AGE, AES, DOUTM, DMORT, RES**



**Logistic Regression – All available Predictors**

The Below ROC curves and AUC values are obtained for the Model (during 10-cross



Mean Accuracy: 72.54658

Mean Sensitivity: 11.77288

Mean Specificity: 94.39425

False Positive Rate: 5.605752

False Negative Rate: 88.22712

ROC and AUC after Training (75%) predicting on Test(25%)



AUC – 62.7%

As can be seen, the Model doesn’t give very good results. We can improve upon it.

**Logistic Regression with Important Predictors**

We can use the Important Variables obtained based on the Information Value analysis done above and build a Logistic Regression Model. We can include the Interaction terms and perform a step-selection of variables using stepAIC() based on AIC values.

Based on AIC values along with significance of terms, we get the following Logistic Regression Model

**BAD ~ DAINC + RES + DMORT + DAINC:DMORT + DAINC:AGE + DMORT:AGE**



Mean Accuracy: 73.95485

Mean Sensitivity: 5.739562

Mean Specificity: 98.38689

False Positive Rate: 1.613115

False Negative Rate: 94.26044

AUC – 59.6%



**Transformation of Variables – Pooling Approach**

I considered multiple ways to transform the variables and built logistic regression Models and Decision Tree Models on them. Most of the transformations provided no significant improvement in terms of Model performance. For the sake of brevity, I am displaying here only those transformations that made logical sense.

Create new columns indicating the Net Inflow of Money into a Household and the Net Outflow of Money on various payments:

NETIN = SINC + DAINC + DHVAL

NETOUT = DMORT + DOUTM + DOUTCC + DOUTHP + DOUTL

Based on StepAIC and Logistic Regression, we get the following Model (including only the significant terms) :

**BAD ~ NKID + RES + NETIN + AES:NETIN + AES:NETOUT + RES:AGE + AGE:NETIN + AGE:NETOUT**

AIC : 1023.4



Mean Accuracy: 72.75442

Mean Sensitivity: 12.55209

Mean Specificity: 94.1826

False Positive Rate: 5.817399

False Negative Rate: 87.44791

Mean AUC – 61.8%

**Note:** I tried several other transformations like log transformations, scaling of values, taking ratios like (NETIN/NETOUT), etc. None of the models gave any significant improvement in the above mentioned Metrics. Therefore, I am omitting those Models.

**Variable Transformation – Binning Approach**

IV and WoE values for the categorical variables and binned continuous variables:

NKID, AGE, DAINC, AES, RES, DOUTM, DOUTCC, PHON



A StepAIC for Logistic Interaction Model with 2-way Interactions yields the following model:

**BAD ~ DAINC + AGE**

For 10-cross validation:

Mean Accuracy: 74.83278

Mean Sensitivity: 11.92578

Mean Specificity: 97.22115

False Positive Rate: 2.778845

False Negative Rate: 88.07422

Mean AUC – 60.408%

We have observed a slight improvement in the Model performance after Binning.

**Variable Transformation – Weight of Evidence Approach**

In this transformation:

 Create bins for Continuous Variables

 Wherever binning was possible, replace the bins with their corresponding calculated Weight of Evidence values.

 Perform stepAIC on Logistic Regression to obtain the significant terms:

AIC: 1006.6

**BAD ~ DAINC + RES + DOUTM + DOUTCC + AGE + DAINC:DOUTM + RES:DOUTCC + DOUTCC:AGE**

Used the above Model for prediction (along with 10-cross Validations) as follows



AUC – 65.1% for 75%-25% Train Test Split

For 10-cross Validation:



Mean AUC – 65.39%

Mean Accuracy: 74.94386

Mean Sensitivity: 16.79624

Mean Specificity: 95.88764

False Positive Rate: 4.112356

False Negative Rate: 83.20376

While interpretability has been lost to an extent, this Model has a better performance than all previous Models.

**Optimal Cut-off Calculations**

One issue noted in the dataset is how unbalanced the data appears to be.

There are more GOOD credit scores in the data compared to BAD credit scores. This has to be kept in mind when selecting a cut-off probability value for the Logistic Regression Model.

For the Model built in the previous step[ 8], I have created a plot that shows the Model accuracy for differing levels of cut-offs.



We can find an Optimal Cut-off Value that can be used, by checking the Accuracy of the model by checking in the range of [0.4,0.8] in increments of 0.05.

As per the above chart, our choice of cut-off at 0.5 seems to be optimal for the given data.

**Optimal Cut-off Calculation for the Logistic Regression Model**



Optimal Cut-off appears to be at 0.45

Running 10-cross validation with the new cut-off value:

Mean Accuracy: 72.66006

Mean Sensitivity: 9.401855

Mean Specificity: 95.47642

False Positive Rate: 4.523576

False Negative Rate: 90.59815

Mean AUC – 64.29%

Changing the Cut-off for Model in 5 doesn’t improve the Model performance significantly. Sensitivity has improved though for a finely tuned cut-off value, as can be expected.

**Recursive Partitioning Trees – Variables Selection Based on IV**

We usually don’t plot ROC curve or calculate AUC for Decision Trees, but for the sake of comparison, we could attempt that as follows.



UC – 57.87%

Performing 10-cross validations:

Mean Accuracy: 72.21811

Mean Sensitivity: 8.708462

Mean Specificity: 94.99527

False Positive Rate: 5.004733

False Negative Rate: 91.29154

**Recursive Partitioning Trees based on Weight of Evidence Transformations**

If we take the dataset that was prepared earlier after binning and replacing the bins with WoE values, we can proceed to build a decision Tree. This would be a loss of interpretability but the Model related performance measures are as follows:

AUC – 74.1%

Mean Accuracy: 75.05136

Mean Sensitivity: 14.57228

Mean Specificity: 96.90624

False Positive Rate: 3.093761

False Negative Rate: 85.42772

This Model is giving a superior performance on all Metrics in terms of predictive power.

**Random Forest Approach for WoE-updated Data**

Finally, I was curious to know if using Random Forest or other such computationally intensive algorithms would give me results that were better than Logistic Regression Models or Simple Decision Trees.

I fit a Random Forest to the Binned Data prepared earlier which had been updated with their weights of evidence.

Variable Importance Plot

**Importance Overall**

PHON 4.776573

AES 17.377710

DAINC 10.368260

RES 14.208134

DOUTM 6.073237

DOUTCC 4.325596

AGE 11.131813

Performance Measures:

Mean Accuracy: 74.51321

Mean Sensitivity: 9.55

Mean Specificity: 97.77656

False Positive Rate: 2.223442

False Negative Rate: 90.45

**Random Forest Approach for Important Variables**

Same approach as earlier, instead I am using the important variables we had selected earlier based on Information Value.

**Variable Importance Plot**



**Overall**

AES 35.66855

DAINC 79.90280

RES 24.15490

DMORT 33.82487

DOUTM 49.69102

AGE 79.59964

Performance Measures:

Mean Accuracy: 71.78866

Mean Sensitivity: 16.13333

Mean Specificity: 91.72959

False Positive Rate: 8.270413

False Negative Rate: 83.86667

**Support Vector Machines – based on WoE based Binned Data**

Again, I tried this method out of pure curiosity (Note, I didn’t tune the SVM for each of the train-test split).

Performance Measures:

Mean Accuracy: 74.07543

Mean Sensitivity: 12.8

Mean Specificity: 96.00527

False Positive Rate: 3.994732

False Negative Rate: 87.2

As can be observed, the performance hasn’t improved to a great extent.

**Conclusion**

If model performance is the main aim, replace each category in the data(bins if continuous data) with their corresponding WoE values and the model obtained (either Logistic Regression or Partition Trees) gives a better predictive performance across the board.

Using newer algorithms(that are computationally intensive like SVM and Random Forests) doesn’t seem to improve predictive power significantly. Perhaps Discriminant Analysis or Principal Component Analysis along with the afore-mentioned approaches would yield better predictive performance, at the cost of easy interpretability.

**References**

<https://multithreaded.stitchfix.com/blog/2015/08/13/weight-of-evidence/>

<http://mlwiki.org/index.php/ROC_Analysis>

<http://ethen8181.github.io/machine-learning/unbalanced/unbalanced.html>

An Introduction to Statistical Learning: With Applications in R

Book by Daniela Witten, Gareth James, Robert Tibshirani, and Trevor Hastie

Information Value Statistic, Bruce Lund,David Brotherton

<https://www.mwsug.org/proceedings/2013/AA/MWSUG-2013-AA14.pdf>

**Acknowledgement**

I have discussed this assignment with Rohith polavarapu. We tried our core work of the analysis over here.

**Difficulties faced**

Unbalanced data & linearly seperable

**Submission by:**

**Revanth Guthala**

**14333**

**MGMT**

**IISC Bangalore**