CREDIT APPROVAL DATA ANALYSIS USING CLASSIFICATION AND REGRESSION MODELS

Ms.D.Jayanthi, Assistant Professor, Department of Information Technology,
Sri Venkateswara College of Engineering. Sriperumbudur

Abstract : Algorithms that are used to decide the outcome of credit applications vary from one provider to another and across sectors and geographies. There are however, high degrees of similarity in the attributes used to generate those algorithms. The differences may be in the weights applied to individual attributes. This paper analyses the credit application data taken from the UCI machine learning repository. Various preprocessing techniques like data exploratory analysis and data transformations like handling missing values, continuous values, and categorical values are computed on the data. Several data visualization techniques are adopted to understand the data. The analytical models like regression and classification techniques are generated and implemented on the data. The results are evaluated using various performance metrics.

Keywords: Machine Learning, Classification, Regression.

A. Introduction

The accurate assessment of consumer credit risk is of uttermost importance for lending organizations. Credit scoring is a widely used technique that helps financial institutions evaluates the likelihood for a credit applicant to default on the financial obligation and decide whether to grant credit or not. The precise judgment of the creditworthiness of applicants allows financial institutions to increase the volume of granted credit while minimizing possible losses. The credit industry has experienced a tremendous growth in the past few decades (Crook et al., 2007). The increased number of potential applicants impelled the development of sophisticated techniques that automate the credit approval procedure and supervise the financial health of the borrower. The large volume of loan portfolios also imply that modest improvements in scoring accuracy may result in significant savings for financial institutions (West, 2000). The goal of a credit scoring model is to classify credit applicants into two classes: the "good credit" class that is liable to reimburse the financial obligation and the "bad credit" class that should be denied credit due to the high probability of defaulting on the financial obligation. The classification is contingent on sociodemographic characteristics of the borrower (such as age, education level, occupation and income), the repayment Performance on previous loans and the type of loan. These models are also applicable to small businesses since these may be regarded as extensions of an individual costumer. In the last few decades, various quantitative methods were proposed in the literature to evaluate consumer loans and improve the credit scoring accuracy (for a review, see e.g. Crook et al., 2007). These models can be grouped into parametric and non-parametric or data mining models. The most popular parametric models are the linear discriminant analysis and the logistic regression. Linear discriminant analysis was the first parametric technique suggested for credit scoring purposes (Reichert et al., 1983). This approach has attracted criticism due to the categorical nature of the data and the fact that the covariance matrices of the good credit and bad credit groups are typically distinct. The logistic regression (Wiginton, 1980) allows overcoming these deficiencies and became a common credit scoring tool of practitioners in financial institutions. Non-parametric techniques applied to credit scoring include the k-nearest neighbor (Henley and Hand, 1996), decision trees (Frydman et al., 1985; Davis et al., 1992), artificial neural networks (Jensen, 1992), genetic programming (Ong et al., 2005) and support vector machines (Baesens et al., 2003). More recently, research on hybrid data mining approaches has shown promising results (Lee et al., 2002; Hsieh, 2005; Lee and Chen, 2002). While the pursuit of better classifiers for credit scoring applications is a crucial research effort, improved accuracies can be easily achieved by aggregating scores predicted by an ensemble of individual classifiers. West et al. (2005) found that the accuracy of an ensemble of neural networks is superior to that of a single neural network in credit scoring and bankruptcy prediction applications. This paper proposes a credit scoring model of consumer loans based on various analytical models. The rest of this paper is organized as follows. In the next section, exploratory analysis and data transformation is presented. This is followed by a description of the data sets and a comparison of the predictive accuracy of the models. A discussion of the relative contribution of the attributes to separate the good credit and bad credit classes is also given. Section 4 concludes the paper.

B. Exploratory Data analysis

In statistics, **exploratory data analysis** (**EDA**) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

EDA tackle specific tasks such as:

- Spotting mistakes and missing data;
- Mapping out the underlying structure of the data;
- Identifying the most important variables;
- Listing anomalies and outliers;
- Testing a hypotheses / checking assumptions related to a specific model;
- Establishing a parsimonious model (one that can be used to explain the data with minimal predictor variables);
- Estimating parameters and figuring out the associated confidence intervals or margins of error.
  Specific statistical functions and techniques you can perform with these tools include:
- Clustering and dimension reduction techniques, which help you to create graphical displays of high-dimensional data containing many variables;
- Univariate visualization of each field in the raw dataset, with summary statistics;
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you're looking at;
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data;
- K-Means Clustering (creating "centres" for each cluster, based on the nearest mean); Predictive models, e.g. linear regression.

The first step in any analysis is to obtain the dataset and codebook. Both the dataset and the codebook can be downloaded for free from the UCI website. A quick review of the codebook shows that all of the values in the dataset have been converted to meaningless symbols to protect the confidentiality of the data. This will still suit our purposes as a demonstration dataset since we are not using the data to develop actual credit screening criteria. Once the dataset is loaded, we'll use the str() function to quickly understand the type of data in the dataset. This function only shows the first few values for each column so there may be surprises deeper in the data but it's a good start. Here you can see the names assigned to the variables. The first 15 variables are the credit application attributes. The Approved variable is the credit approval status and target value. Using the output below, we can see that the outcome values in Approved are '+' or '-' for whether credit had been granted or not. These character symbols aren't meaningful as is so will need to be transformed. Turning the '+' to a '1' and the '-' to a '0' will help with classification and logistic regression models later in the analysis.

Table 1: Dataset and codebook for credit approval data

```
data.frame':   689 obs. of  16 variables:
 $ Male          : num  1 1 0 0 0 0 1 0 0 0 ...
 $ Age           : chr  "58.67" "24.50" "27.83" "20.17" ...
 $ Debt          : num  4.46 0.5 1.54 5.62 4 ...
 $ Married       : chr  "u" "u" "u" "u" ...
 $ BankCustomer  : chr  "g" "g" "g" "g" ...
 $ EducationLevel: chr  "q" "q" "w" "w" ...
 $ Ethnicity     : chr  "h" "h" "v" "v" ...
 $ YearsEmployed : num  3.04 1.5 3.75 1.71 2.5 ...
 $ PriorDefault  : num  1 1 1 1 1 1 1 1 1 0 ...
 $ Employed      : num  1 0 1 0 0 0 0 0 0 0 ...
 $ CreditScore   : num  6 0 5 0 0 0 0 0 0 0 ...
 $ DriversLicense: chr  "f" "f" "t" "f" ...
 $ Citizen       : chr  "g" "g" "g" "s" ...
 $ ZipCode       : chr  "00043" "00280" "00100" "00120" ...
 $ Income        : num  560 824 3 0 0 ...
 $ Approved      : chr  "+" "+" "+" "+" ...
```

C. Data Transformations

1. Continuous Values

To start with, we will use the summary() function to see the descriptive statistics of the numeric values such as min, max, mean, and median. The range is the difference between the minimum and maximum values and can be calculated from the summary() output. For the B variable, the range is 66.5 and the standard deviation is 11.9667.

```
Age  Debt  YearsEmployed  CreditScore  Income    Min.  :13.75  Min.  : 0.000  Min.  : 0.000  Min.  : 0.000  Min :  0
1st Qu.:22.58  1st Qu.: 1.000  1st Qu.: 0.165  1st Qu.: 0.000  1st Qu.:  0  Median :28.42  Median : 2.750  Median : 1.000  Median : 0.000  Median:  5  Mean  :31.57  Mean  : 4.766  Mean  : 2.225  Mean  : 2.402  Mean: 1019
3rd Qu.:38.25  3rd Qu.: 7.250  3rd Qu.: 2.625  3rd Qu.: 3.000  3rd Qu.: 396  Max.  :80.25  Max.  :28.000  Max.  :28.500  Max.  :67.000  Max.:100000  NA's  :12
```

```
[1] 11.9667
```

2. Missing Values

We can see from the summary output that the Debt variable has missing values that we'll have to fill in. We could simply use the mean of all the existing values to do so. Another method would be to check the relationship among the numeric values and use a linear regression to fill them in. The table below shows the correlation between all of the variables. The diagonal correlation values equal 1.000 because each variable is perfectly correlated with itself. To read the table, we will look at the data by rows. The largest value in the first row is 0.396 meaning age is most closely correlated with YearsEmployed. Similarly, Debt is mostly correlated with YearsEmployed

```
             Age  Debt YearsEmployed CreditScore Income
Age          1.000 0.202       0.396      0.186  0.019
Debt         0.202 1.000       0.301      0.271  0.122
YearsEmployed 0.396 0.301      1.000      0.327  0.053
CreditScore  0.186 0.271       0.327      1.000  0.063
Income       0.019 0.122       0.053      0.063  1.000
```

We can use this information to create a linear regression model between the two variables. The model produces the two coefficients below: Intercept and YearsEmployed. These coefficients are used to predict future values. The YearsEmployed coefficients is multiplied by the value for YearsEmployed and the intercept is added.

```
(Intercept) YearsEmployed
  28.446953      1.412399
```

2. Descriptive Statistics

The next step of working with continuous variables is to standardize them or calculate the z-score. First, we use the mean and standard deviation calculated above. Then, subtract the mean from each value and, finally, divide by the standard deviation. The end result is the z-score. When we plot the histograms, the distribution looks the same but the z-scores are easier to work with because the values are measured in standard deviations instead of raw values. One thing to note is that the data is skewed to the right because the tail is longer.

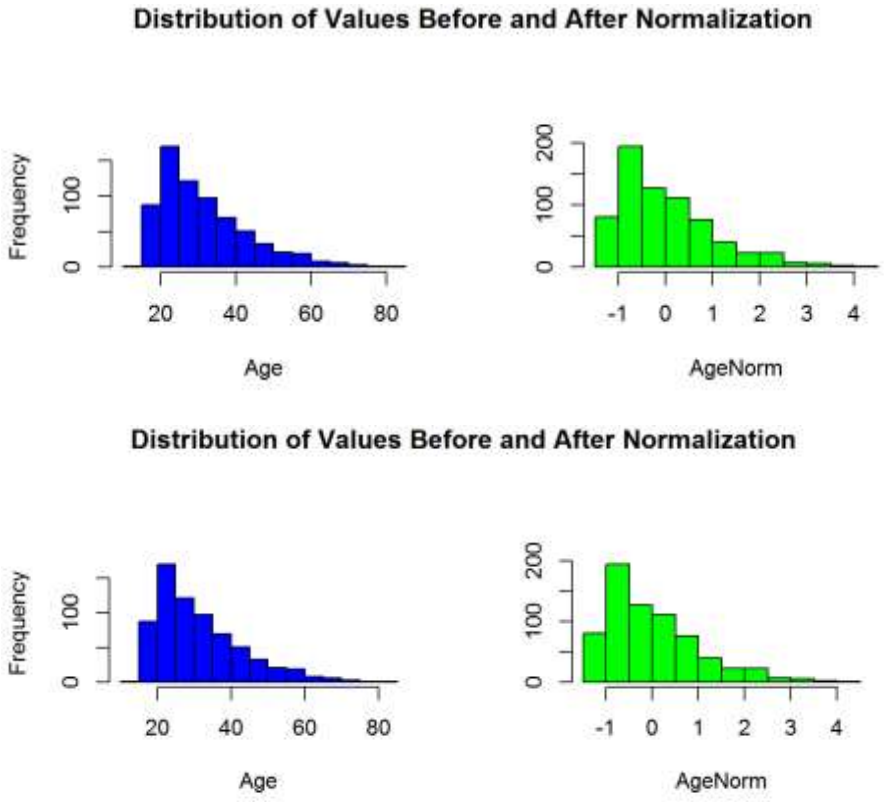
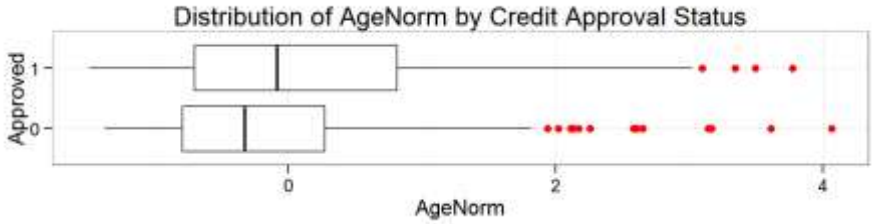


Fig.1 Descriptive statistics of values using normalization



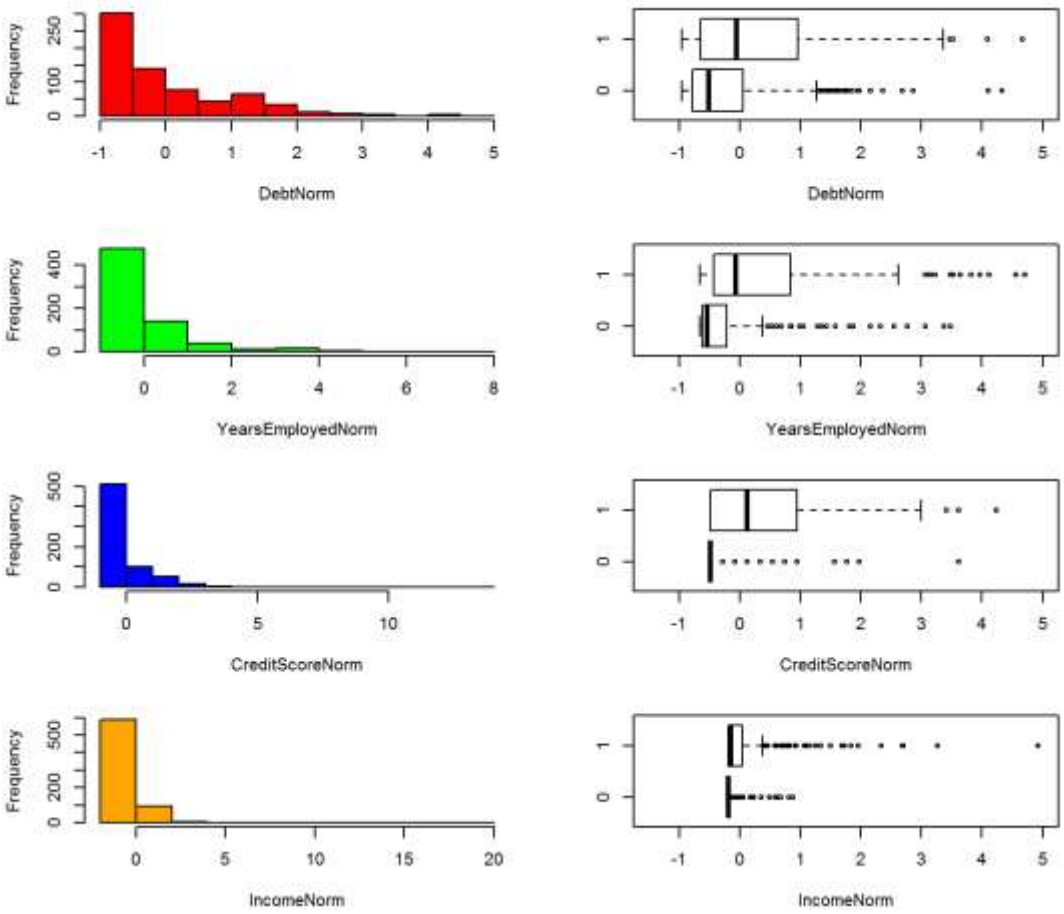


Fig.2. Data Visualization methods for credit approval data

4. C**ategorical Variables (Association Rules)**

We will now work with categorical values in column Male. The data is distributed across factors '1' and '0' plus 12 of them are missing values. Again, the missing values will not work well in classifier models so we'll need to fill in them in. The simplest way to do so is to use the most common value. For example, since the '0' factor is the most common, we could replace all missing values with 'o'.

```
 0  1
479 210
```

D. Generate Analytic Models

In order to prepare and apply a model to this dataset, we'll first have to break it into two subsets. The first will be the training set on which we will develop the model. The second will be the test dataset which we will use to test the accuracy of our model. We will allocate 75% of the items to Training and 25% items to the Test set.

Once our dataset has been split, we can establish a baseline model for predicting whether a credit application will be approved. This baseline model will be used as a benchmark to determine how effective the models are. First, we determine the percentage of credit card applications that were approved in the training set: There are 517 applications and 287 or 56% of which were denied. Since more applications were denied than were approved, our baseline model will predict that all applications were denied. This simple model would be correct 56% of the time. Our models have to be more accurate than 56% to add value to the business.

```
 0  1
```

```
287 230
```

1. **Logistic Regression -Create the Model**

Regression models are useful for predicting continuous (numeric) variables. However, the target value in Approved is binary and can only be values of 1 or 0. The applicant can either be issued a credit card or denied- they cannot receive a partial credit card. We could use linear regression to predict the approval decision using threshold and anything below assigned to 0 and anything above is assigned to 1. Unfortunately, the predicted values could be well outside of the 0 to 1 expected range. Therefore, linear or multivariate regression will not be effective for predicting the values. Instead, logistic regression will be more useful because it will produce probability that the target value is 1. Probabilities are always between 0 and 1 so the output will more closely match the target value range than linear regression.

The model summary shows that the p-values for each coefficient. Alongside these coefficients, the summary gives R's usual at-a-glance scale of asterisks for significance. Using this scale, we can see that the coefficients for AgeNorm and Debt3 are not significant. We can likely simplify the model by removing these two variables and get nearly the same accuracy.

```
Call:
glm(formula = Approved ~ AgeNorm + DebtLog + YearsEmployedLog +
    CreditScoreLog + IncomeLog, family = binomial, data = Train)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4345  -0.7844  -0.4906   0.7212   2.1822
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.13120    0.11197  -1.172 0.241315
AgeNorm          0.01151    0.11721   0.098 0.921797
DebtLog          0.10338    0.11517   0.898 0.369364
YearsEmployedLog  0.70361    0.12782   5.505 3.70e-08 ***
CreditScoreLog    1.03286    0.13884   7.439 1.01e-13 ***
IncomeLog         0.46008    0.11970   3.844 0.000121 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 710.42  on 516  degrees of freedom
Residual deviance: 508.93  on 511  degrees of freedom
AIC: 520.93

Number of Fisher Scoring iterations: 5
```

The confusion matrix shows the distribution of actual values and predicted values. The top left value is the number of observations correctly predicted as denied credit and the bottom right is the number of observations correctly predicted as credit granted. The other values are the false positive and false negative values. Of the 517 observations, the model correctly predicted 398 approval decisions (249 + 149) or about 77% accuracy. Already, we can see that we have improved on the baseline model and improved our accuracy by 21%. We can use this matrix to compare the results of the model after removing the non-significant variables.

```
   FALSE TRUE
0   249   38
1    81  149
```

2. **Classification and Regression Tree - Create the Model**

Classification and Regression Trees (CART) can be used for similar purposes as logistic regression. They both can be used to classify items in a dataset to a binary class attribute. The trees work by splitting the dataset at series of nodes that eventually segregates the data into the target variable. The models are sometimes referred to as decision trees because at each node the model determines which path the item should take. They have an advantage over logarithmic regression models in that the splits or decision are more easily interpreted than a collection of numerical coefficients and logarithmic scores.

The model split the training dataset at PriorDefault variable. If the value in PriorDefault is f or false, then the target value will most likely be 0. If the value is true, then the target will most likely be 1.

```
n= 517
node), split, n, loss, yval, (yprob)
      * denotes terminal node
1) root 517 230 0 (0.55512573 0.44487427)
  2) PriorDefault=0 247  16 0 (0.93522267 0.06477733) *
  3) PriorDefault=1 270  56 1 (0.20740741 0.79259259) *
```

The confusion matrix resulting from this CART model shows that we correctly classified 231 denied credit applications and 214 approved applications. The accuracy score for this model is 86.1% which is better than the 75% accuracy the logistic regression model scored and significantly better than the baseline model.

```
 FALSE TRUE
0   231   56
1    16  214
```

3.Apply the Model We'll now apply our classifier model to the test dataset and determine how effective it is. Our confusion matrix shows 144 items were correctly predicted for 83% accuracy. We can see that this model is both more effective and easier to interpret than the logistic regression model.

```
 FALSE TRUE
0    75   21
1     7   69
```

E. Conclusion and Future enhancement

In this paper, data preprocessing and transformation techniques are applied and results are generated by implementing analytical models. The performance is analyzed using the confusion matrix table. We can also use this model to make detail testing selections. Any credit application that does not have the same outcome as predicted by the model is potential audit exception. The inherent risk is that a credit card was issued to someone that should have been denied. This account is more likely to default than a properly approved account which, in turn, exposes the company to loss. The different machine learning models can be implemented and the performance can be compared.

References

1.Wuning Tong,Yuping Wang, Junkun Zhong, Wei Yan," A New Weight Based Density Peaks Clustering Algorithm for Numerical and Categorical Data", CIS (2017), IEEE 10.1109/CIS.2017.00044.

2. Ashlesha Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval", ICCNT(2017), 10.1109/ICCCNT.2017.8203946

3. LaiHui, ShuaiLiZhou, Zongfang, "The Model and Empirical Research of Application Scoring based on Data Mining Methods",Elseiver, Volume 17, 2013, Pages 911-918.

4. Hongmei, ChenaYaoxin, Xianga,"The Study of Credit Scoring Model Based on Group Lasso", Elseiver, Volume 122, 2017, Pages 677-684.

5. FlorentinButarua,Qingqing,ChenaBrian, ClarkaeSanmay,DasbAndrew, W.LocdAkhtarSiddique, "Risk and risk management in the credit card industry", Elseiver, Volume 72, November 2016, Pages 218-239.