DATA621 Final Project: Credit Risk Analysis

Javern Wilson and Jack Russo

5/22/2020

Contents

Abstract

The purpose of this paper is to provide researchers and readers a comprehensive review of scholarly research on credit risk analysis using various classification models to predict whether a customer may get approved for a loan based on the features of their background similar to what is in the dataset used. Credit risk is usually the result of whether to approve loans especially in financial institutions. It often refers to the uncertainty of whether clients can keep up with servicing loans which consequently can cause bank crisis if they do not. Tables, figures and illustrations of data and models are included to comprehensively assess the results and assumptions to be verified. Online research journals and texts books demonstrating the use of the models used were reviewed for proper use and reporting on these models. The illustrations of the models are applied to the dataset to test the research hypothesis. In addition to this, relationships between the independent variables and a dichotomous dependent variable are also explored. As a result, recommendations and conclusions are made based on the outcome of the models.

Keywords: credit risk, classification, evaluation, default, models

Introduction

When evaluating the range of data available for analysis, the authors sought data that could provide insight into business decision making as it pertains to non-discrimination. The credit risk dataset enabled the authors to examine the relationship betweem loan approval and other factors such as credit history, gender or education. In general, building effective models of this sort is an important metric of "fairness". Stakeholders want business decisions to be consistent with the available data, not grounded in personal biases. However, consistent decision making may still factor traits like gender into model creation. To evaluate the degree to which factors has influenced the loan approval process within the dataset, we shall construct an optimal logistic model where loan approval is measured against the other predictors provided in the credit risk dataset. The model's performance will then be measured based on the metrics used for classsification models when predicting a potential borrower's outcome for approval.

Literature Review

Generally commercial banks play an important role in economic development of a country as they are crticial to the country's performance. Banks facilitate payments and channel credits to households and businesses. Credit risk is the rise or fall of a company's net asset value which appears when the parties of an agreement are not able to fullfil that obligation. The largest source of credit risk in banking institutions arise from loans.

There are many studies that evaluate that there is an association between the default behaviour of the borrower and certain characteristic of their backgrounds such as education, marital status, income and employment. Most researcher believe that education plays a huge role in predicting who will default or deemed high credit risk. Students who are successful in their studies tend to have lower default rates than those who do not. For instance, there was a study based on students in California where failure to complete the academic program for which they signed up for are one of the strongest predictors of credit risk among all types of students (Woo 2002). Usually it is poor academic performance that encourages a student to withdraw, therefore the cause of loan default (Volkwein and Cabrera 1998).

As for employment,Woo found that the strongest post-school variable associated with default is filing for unemployment insurance. Borrowers who experienced unemployment showed an 83 percent increase in their probability of default over their original probability (Woo 2002). Nationally, borrowers indicate that the most important reasons for default are being unemployed (59 percent said this) and working at low wages (49 percent) (Volkwein et al. 1998).

According to Woo, borrowers with high earnings after they leave school are less likely to default on a loan than those with low earnings. A lot of times this can happen when students or rather borrowers take out larger loans and enter low paying career jobs. However, the income predictor was not as strong as the employment predictor (Woo 2002).

When it comes to the personal lives of borrowers in terms of martial status, being separated, divorced or widowed increases the probability over 7 percent by going into default. In addition, having kids increases the probability by 4.5 percent per child (Volkwein and Szelest 1995). Having dependent children when not married increases the default rates above 40 percent (Volkwein et al. 1998).

As for credit history, this plays a large role in the lender's decision to qualify a customer for a loan. It is the first thing lenders look at when assessing the potential borrower's credit history. One's credit history is like a financial track record that shows how well the individual or party manages credit and payments over time. The results may vary but all lenders like to see good payment history, low amounts of debts, no missed or late payments.

Methodology

Research was gathered using secondary resources. The type of research method used in this study is that of quantitative techniques where the aim is to classify features, count them, and construct statistical models in an attempt to explain what is observed. In order for this research to be possible, the data was collected from an online database. In identifying sources for this research, multiple research papers were used.

Experiments and Results

Description of Dataset

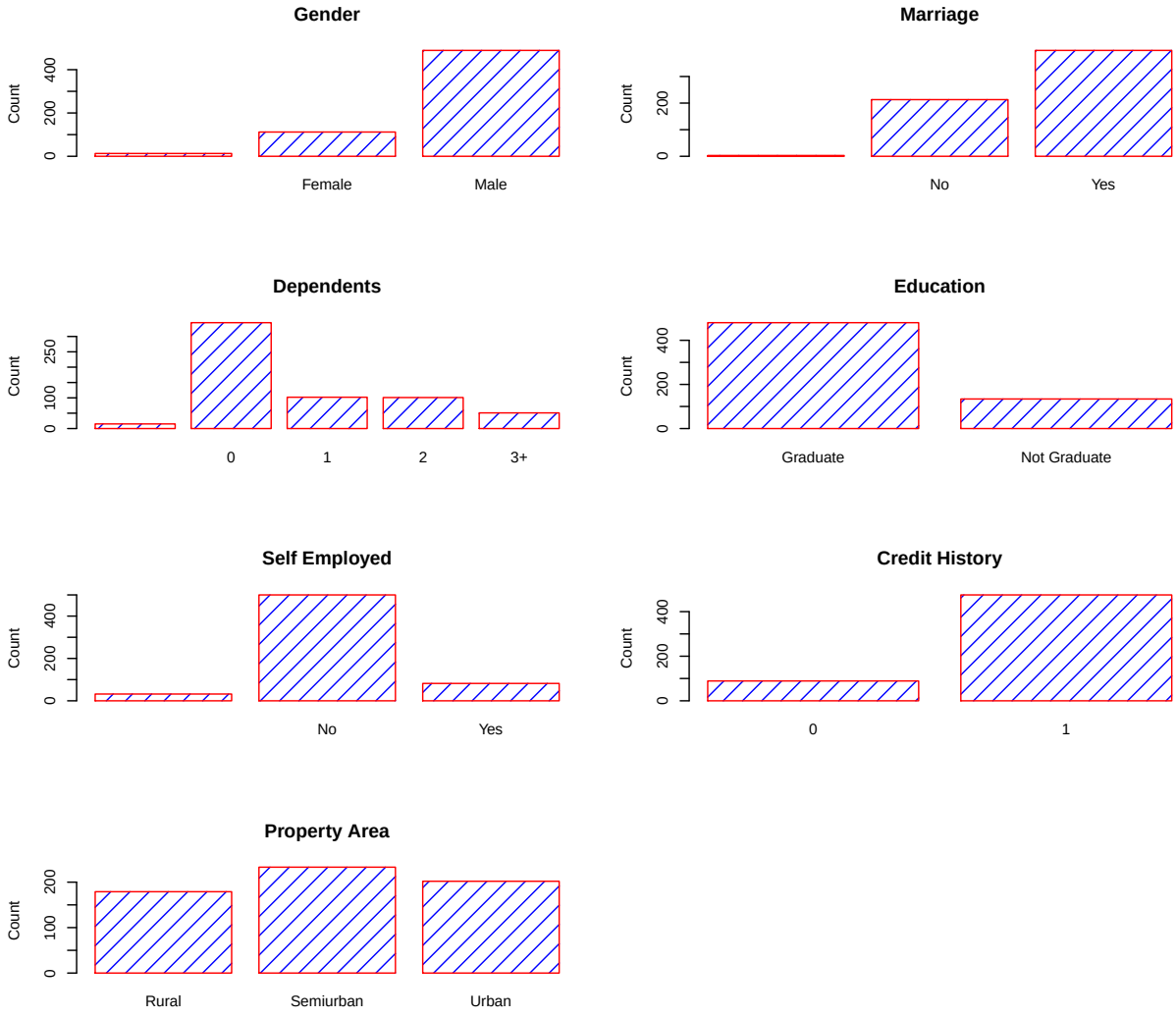
In the dataset, there are 12 variables; 11 predictor variables and the dependent variable. Below is a summary showing the variables. The dependent variable is the `Loan_Status` which contains whether or not the potential borrower may get approved.

Data Summary

```
##      Loan_ID       Gender    Married   Dependents        Education
##  LP001002:  1         : 13      :  3     : 15     Graduate    :480
##  LP001003:  1   Female:112   No :213   0 :345     Not Graduate:134
##  LP001005:  1   Male  :489   Yes:398   1 :102
##  LP001006:  1                           2 :101
##  LP001008:  1                           3+: 51
##  LP001011:  1
##  (Other) :608
##  Self_Employed ApplicantIncome CoapplicantIncome   LoanAmount
##     : 32       Min.   :  150   Min.   :    0     Min.   :  9.0
##  No :500       1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.0
##  Yes: 82       Median : 3812   Median : 1188     Median :128.0
##                Mean   : 5403   Mean   : 1621     Mean   :146.4
##                3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:168.0
##                Max.   :81000   Max.   :41667     Max.   :700.0
##                                                  NA's   :22
##  Loan_Amount_Term Credit_History     Property_Area Loan_Status
##  Min.   : 12      Min.   :0.0000   Rural    :179   N:192
##  1st Qu.:360      1st Qu.:1.0000   Semiurban:233   Y:422
##  Median :360      Median :1.0000   Urban    :202
##  Mean   :342      Mean   :0.8422
##  3rd Qu.:360      3rd Qu.:1.0000
##  Max.   :480      Max.   :1.0000
##  NA's   :14       NA's   :50
```

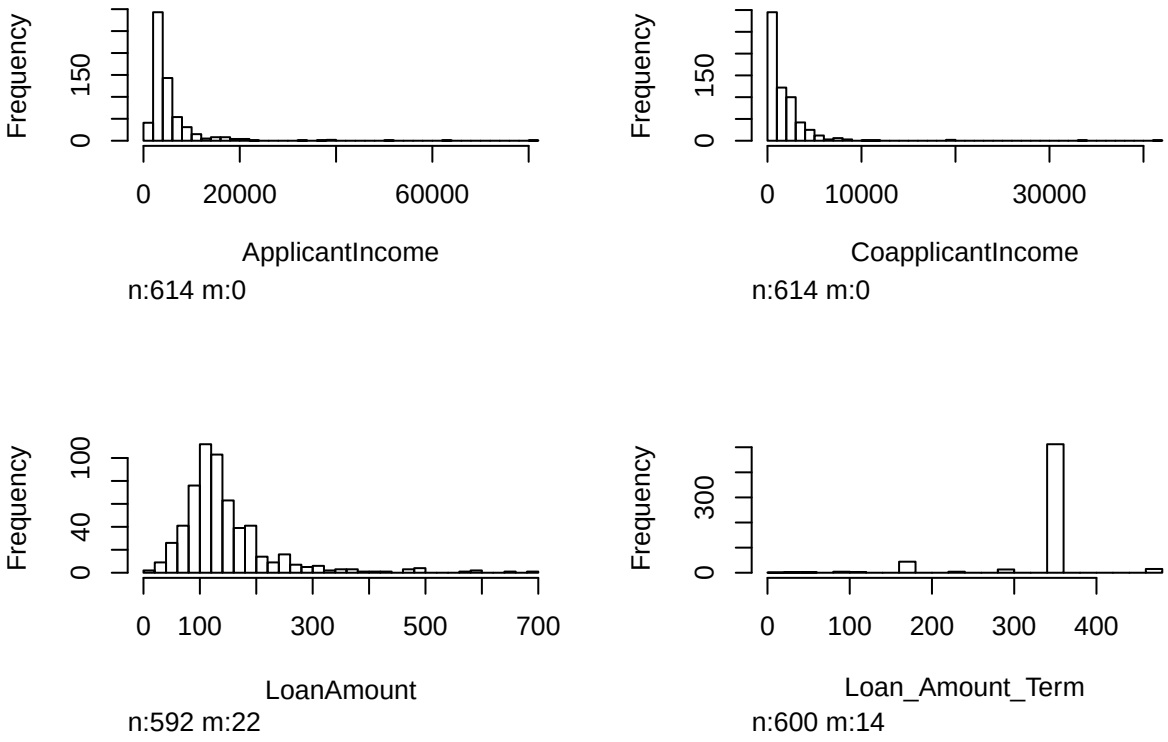
`Credit_History`, `LoanAmount` and `Loan_Amount_Term` had missing values. As we progressed along the project we worked on the missing data.

Frequences for categorical variables



From the plot above we can infer that majority of the persons observed in the dataset are Male, married, no kids, graduated, self employed, have credit history and lives in the semi-urban areas.

Distribution for numeric variables

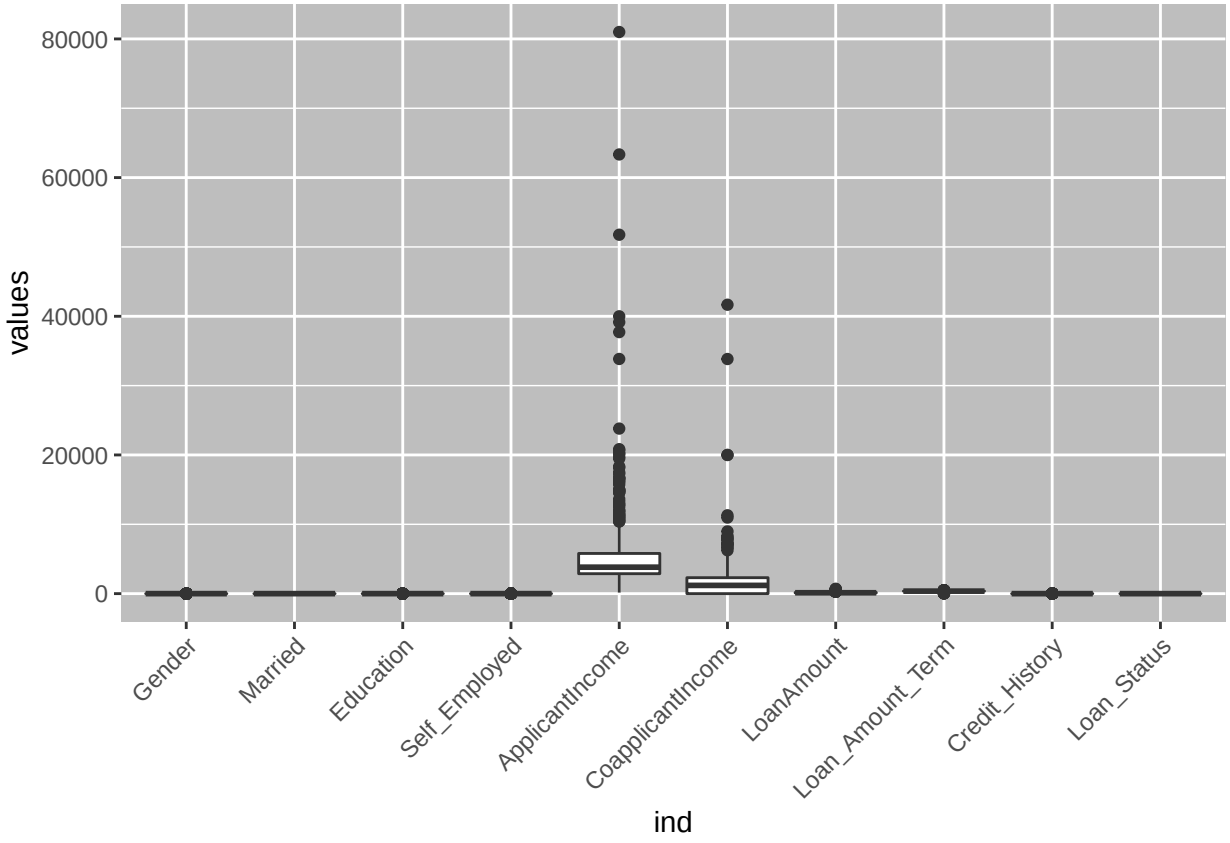


1. The distributions are quite skewed.
2. Majority of the applicant does not make a lot of money and the Coapplicants make less.
3. The loan amount seems to usually be in the 100s. To confirm this, if you look at the summary earlier, you'll see that the median is at `128` (thousand).
4. As far as the Loan Amount term is concerned, the terms averages up to 3 - 3.5 years.

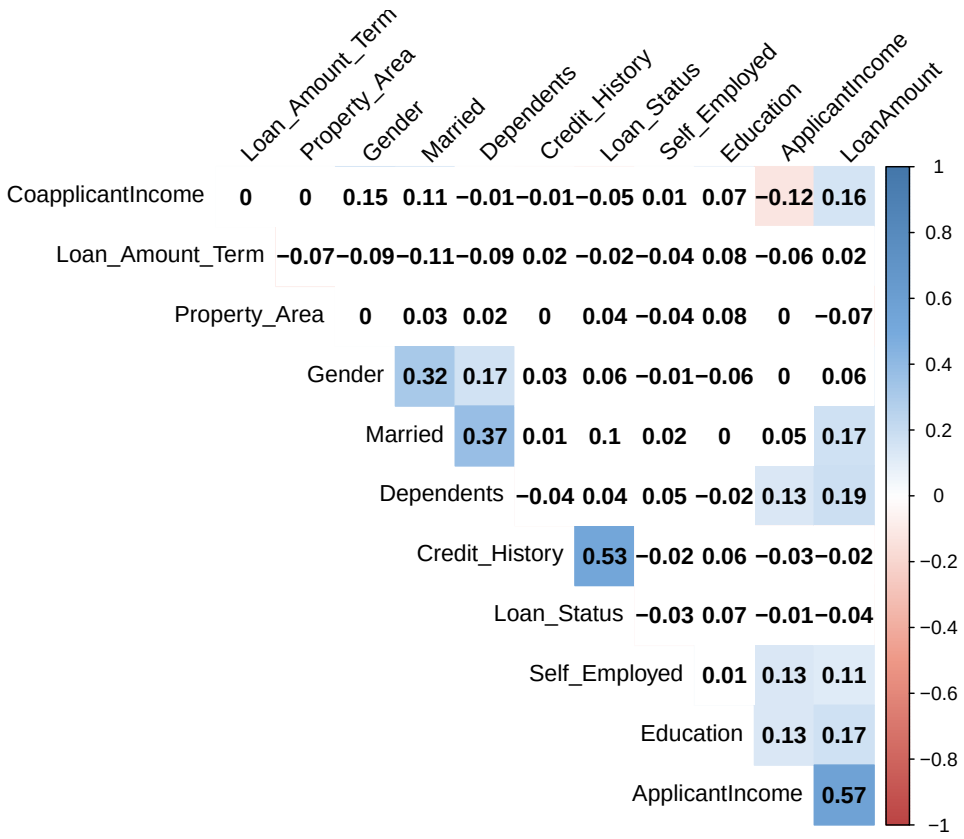
Preprocessing Data First variable to be removed from the dataset is `Loan_ID` because it is unique and has no relevance to the target variable.

All variables that are non-numeric are converted to numeric using multiclass or binary method.

For further exploraration, below is a boxplot of each numeric variable in the dataset clearly outlining any outliers present.



As we move forward, we look at correlations among the various predictors and the target variable.



Only significant correlations are highlighted. For instance we can see some correlation between `LoanAmount` and `ApplicationIncome` at `0.57`. `Credit_History` is the only predictor that has a moderately positive relationship with the target `Loan_Status` at `0.53`. `Married` and `Gender` also somewhat have a relationship `0.32` that is positive.

Earlier we saw the graph showing which variables had missing values. Before applying logistic regression model, we will do further analysis on missing values and outliers by imputation. To take care of this, the predictive mean will be imputed. In order to do this, the mice package is used.

Modeling

Two models were build to determine which would be more optimal in predicting credit risk or chance of loan default of participants. The first model was built using all predictors in the dataset against the dependent variable. On the other hand, the second model was built using the stepwise regression algorithm to determine the best predictors.

When the first model was built, predictor variables such as dependents (2 or more), gender and self employed are not statistically significant. Keeping them in the model may contribute to overfitting. The *AIC* score was `575.33` on `599` degrees of freedom. There was room for the model to improve using statistical techniques, such as stepwise regression to eliminate them. With this technique an optimal model was be generated with only the significant predictors.

Optimal Model

```
##
## Call:
## glm(formula = Loan_Status ~ Married + Education + CoapplicantIncome +
##     LoanAmount + Credit_History + Property_Area, family = binomial,
##     data = cred_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1828  -0.3672   0.5475   0.7094   2.5112
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.328e+00  5.088e-01  -6.540 6.14e-11 ***
## Married            6.155e-01  2.230e-01   2.761  0.00577 **
## Education          4.690e-01  2.580e-01   1.818  0.06911 .
## CoapplicantIncome -5.514e-05  3.409e-05  -1.618  0.10570
## LoanAmount        -1.928e-03  1.198e-03  -1.609  0.10755
## Credit_History     3.975e+00  4.155e-01   9.568  < 2e-16 ***
## Property_Area2     8.428e-01  2.681e-01   3.143  0.00167 **
## Property_Area3     2.290e-01  2.593e-01   0.883  0.37716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 762.89  on 613  degrees of freedom
## Residual deviance: 549.65  on 606  degrees of freedom
## AIC: 565.65
##
## Number of Fisher Scoring iterations: 5
```

The top 7 predictors generated from the second model are what were considered as important variables in the first model. The AIC did decrease therefore indicating that the model did improve a bit.

Multicollinearity

```
##                       GVIF Df GVIF^(1/(2*Df))
## Married           1.041142  1        1.020363
```

```
## Education         1.063355  1        1.031191
## CoapplicantIncome 1.032455  1        1.016098
## LoanAmount        1.101334  1        1.049444
## Credit_History    1.018050  1        1.008985
## Property_Area     1.026903  2        1.006659
```

There does not seem to be an multicollinearity as all the resulting values are below the threshold of 5.

Variable Importance for Optimal Model

```
##                   X   Overall
## 5    Credit_History 9.5678147
## 6    Property_Area2 3.1432539
## 1           Married 2.7605665
## 2         Education 1.8176798
## 3 CoapplicantIncome 1.6178305
## 4        LoanAmount 1.6093203
## 7    Property_Area3 0.8831338
```

Model Evaluation, Selection and Diagnostics

Comparing first and second model,

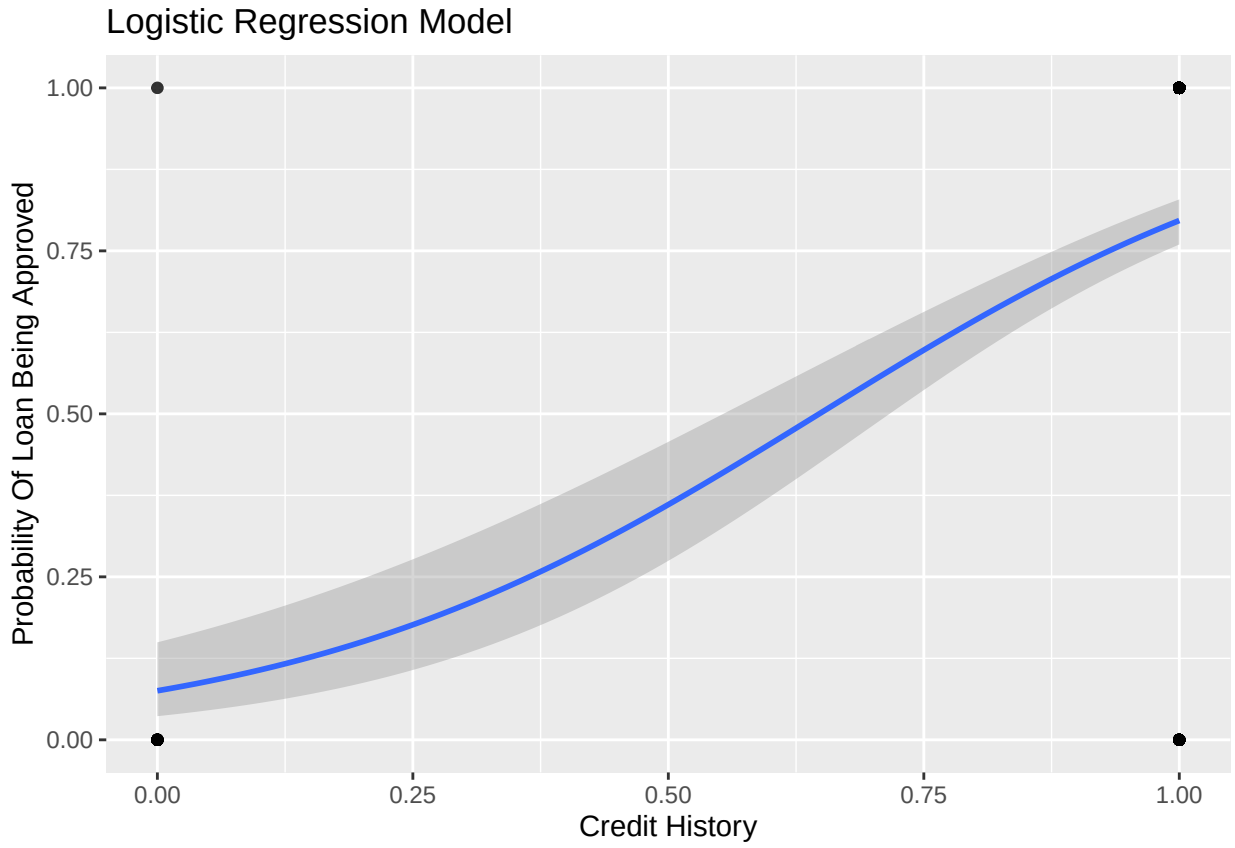
```
## Analysis of Deviance Table
##
## Model 1: Loan_Status ~ Married + Education + CoapplicantIncome + LoanAmount +
##     Credit_History + Property_Area
## Model 2: Loan_Status ~ Gender + Married + Dependents + Education + Self_Employed +
##     ApplicantIncome + CoapplicantIncome + LoanAmount + Loan_Amount_Term +
##     Credit_History + Property_Area
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       606     549.65
## 2       599     545.33  7   4.3204   0.7422
```

the results shows a non-significant result of (`p=0.7422`) against model one. Thus, we reject **model 1** and keep **model 2** as the optimal model going forward.

Interpretation After using the stepwise regression function, we have `7` predictors remaining that are considered strong enough to predict whether or not a person can be approved for a loan. From the looks of the model, `Credit_History` with a positive coefficent of `3.975` has the most impact on the target variable while having the smallest p-value. This means that an increase a good credit history is associated with an increased chance of getting approved for a loan. This makes sense as lot of finance companies tend to look at your credit history to find out how responsible you are. On the other hand, with `LoanAmount` the coefficient is negative with `-0.001928`. This means that an increase in the amount of loan being requested by the borrower will be associated with a decrease in the probability of getting approved for a loan. This predictor definitely helps in the outcome of getting approved for a loan.

To confirm assumptions made about the model, the odds ratio may come in handy. Odds ratio is a statistic that quantifies the strength of the association or relationship between predictor and target variable. For example, based on the regression coefficient of `Credit_History`, a one unit increase will increase the odds of getting loan approved is 53.25 times.

As mention earlier in the literature review, credit histoy plays a major role in determine who gets approved for a loan. It is the first thing the lender look at when assessing the potential buyer's qualification. With `Credit_History` being the most significant predictor, let's view the probability of being approved based on this predictor.



When fitting the line to the points , a slight a S-shape curve is produced.

Metrics

Confustion Matrix The confusion matrix describes the performance of the classifier. It is a table with four different combinations of predicted and actual values. It is extremely useful for measuring Recall, Precision, Specificity, Accuracy and most importantly AUC-ROC Curve. Below is the confusion matrix for the binary classifier.

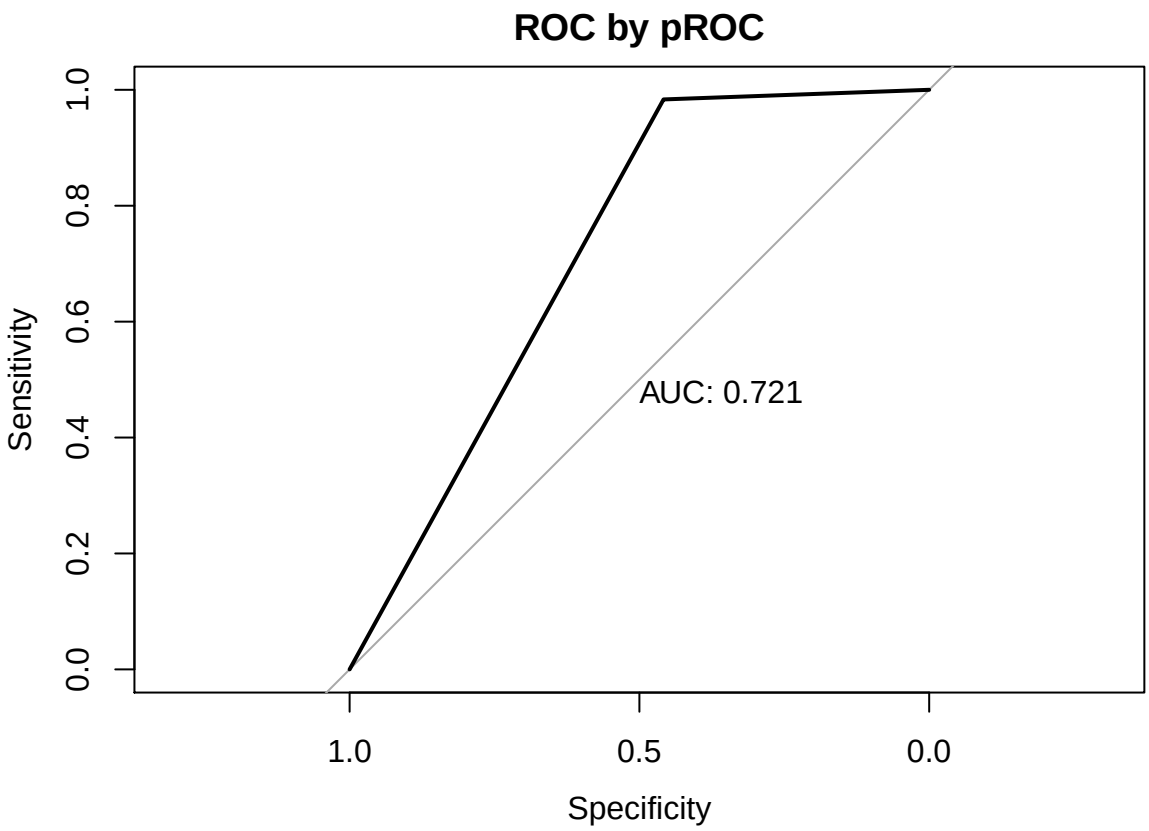
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  88   7
##          1 104 415
##
##                Accuracy : 0.8192
##                  95% CI : (0.7865, 0.8489)
##     No Information Rate : 0.6873
##     P-Value [Acc > NIR] : 9.197e-14
##
##                   Kappa : 0.5123
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9834
##             Specificity : 0.4583
##          Pos Pred Value : 0.7996
##          Neg Pred Value : 0.9263
##              Prevalence : 0.6873
##          Detection Rate : 0.6759
##    Detection Prevalence : 0.8453
##       Balanced Accuracy : 0.7209
##
##        'Positive' Class : 1
##
```

There are two possible predicted classes: "**1**"/ *yes* and "**0**"/ *no*. If we were predicting the presence of a disease, for example, "**1**" means the customer was approved for a loan, and "**0**" means the customer did not get approved for a loan. The classifier made a total of 614 predictions. Out of those 614 cases, the classifier predicted *yes* 422 times, and *no* 192 times. In reality, 519 potential borrowers in the sample were approved, and 95 were not.

Let's dig further:

1. Accuracy - 81.92 % of the time the model is correct.
2. Sensitivity - 98.34% of the time when the model predicts yes, it is actually yes.
3. Specificity - 45.83 % of the time when the model predicts no, it is actually a no.
4. Prevalence - 68.73% of the sample predictions are classified as yes or approved.

ROC CURVE Shows how the true positive rate against the false positive rate at various threshold settings. The AUC (Area Under Curve) tells how much model is capable of distinguishing between classes. Higher the AUC is better, that is, how well the model is at predicting 0s as 0s and 1s as 1s.



A curve pulled close to the upper left corner indicates (an AUC close to 1 and thus) a better performing test. The Area Under the Curve (AUC) is only at `0.721` which is acceptable.

Discussion and Conclusions

The finding of this study provides greater understanding into discovering what factors have signifcant impacts on predicting which borrower will likely to be approved for a loan based on their backround. The model also pointed out that gender is not a major factor for deciding who will get approved. It does make sense however, that education, income, martial status and credit history helps the financial institution in determing who they should loan money to. The model also included the borrower's living arrangement as significant too, that is, those who live in semi-urban neighborhoods. This concludes that folks with a responsible and financially stable background are likely to be approved for loan and are therefore a lower risk to the financial institution.

References

Answers Ltd. (2020, January 15). Literature Review of Risks in Banking. Retrieved from https://ukdiss.com/litreview/literature-review-of-risks-in-banking-finance.php?vref=1

Pandey, Vikas. "Test Preds." Kaggle, 30 May 2017, www.kaggle.com/vikasp/loadpred.

R, L. (2020). Logistic Regression Essentials in R - Articles - STHDA. Sthda.com. Retrieved 7 May 2020, from http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/.

Volkwein, J. F., & Szelest, B. P. (1995). Individual and campus characteristics associated with student loan default. Research in Higher Education, 36(1), 41-72. h

Volkwein, J. F., Szelest, B. P., Cabrera, A. F., & Napierski-Prancl, M. R. (1998). Factors associated with student loan default among different racial and ethnic groups. The Journal of Higher Education, 69(2), 206.

Woo, Jennie H. (2002). Factors Affecting the Probability of Default: Student Loans in California. Journal of Student Financial Aid 32 (2): 5-25.

Appendices

First 10 Predictions On Test Set

Table 1: Table continues below

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
1	1	0	1	0	5720
1	1	1	1	0	3076
1	1	2	1	0	5000
1	1	2	1	0	2340
1	0	0	0	0	3276
1	1	0	0	1	2165
0	0	1	0	0	2226
1	1	2	0	0	3881
1	1	2	1	0	13633
1	0	0	0	0	2400

Table 2: Table continues below

CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	110	360	1
1500	126	360	1
1800	208	360	1
2546	100	360	NA
0	78	360	1
3422	152	360	1
0	59	360	1
0	147	360	0
0	280	240	1
2400	123	360	1

Property_Area	prob	pred.class
3	0.8518	1
3	0.8369	1
3	0.8116	1
3	NA	NA
3	0.6739	1
3	0.733	1
2	0.7984	1
1	0.04762	0
3	0.8055	1
2	0.7541	1

Source Code GITHUB

Technical Report Rpubs