

## Section 2. Mathematical and instrumental methods of economics

<https://doi.org/10.29013/EJEMS-22-2-7-15>

Hantong Li

### IMPLICATIONS OF CREDIT APPROVAL

**Abstract.** Since the advance of the information era, much research has been conducted to help applicants successfully apply for loans. While studies have examined financially related factors, such as income and socioeconomic status, few studies explore aspects outside the financial realm, such as gender and a few other factors. With the loan application data from the Dream Housing Finance Company, we used a logistic regression model to analyze the association between loan approval and other variables. Then, to examine the high correlation impact on the model result, a lasso logistic regression is also fitted, and the out-of-sample performance is compared. We have found that among all the factors, marriage status, the presence of credit history, and geographical locations have significant effects on the application results. In the end, the data is segmented according to male and female groups to investigate the difference in feature importance across gender. The result indicates that female applicant is additionally judged by their income and their co-applicant income, which suggests that women are still, or perceived as, financially unstable in modern society.

**Keywords:** Logistic regression, credit application, gender inequality.

#### Statistical Analysis on Credit Application

##### 1. Introduction

Credit application refers to how a customer obtains loans from a bank. Loaning bank credit is critical in helping people make purchases over their budget. With credit loan data provided by the US Dream Housing Finance Company, we sought to investigate the important factors for credit approval [1]. To begin with, we built a logistic regression model, which allows us to compute the level of significance of multiple factors, including credit history, property area, gender, income distribution, loan amount, number of dependents, etc. To account for the high correlation among independent variables, we cross-validated the model results with a lasso logistic regression model and found that the findings are highly aligned.

Loan approval is an extensively investigated field, while many researchers focus on building models to predict the individual probability of approval. For example, in the paper *Loan Approval Prediction based on Machine Learning Approach*, Arun et al. discuss the application of advanced machine learning models, such as random forest, in credit approval prediction [2]. In terms of the applicant's characteristics, Marcelo et al. analyze the relationship between socioeconomic factors and the probability of loan approval. They find that in addition to the usual financial performance variables, business and social relationships between lenders and prospective borrowers significantly affect the likelihood of loan approval [3].

Compared to previous research on credit approval, our research models the approval probability from a

statistics perspective with careful attention paid to data cleaning and multi-collinearity. In addition, our research has extended the investigation scope to a wider range of variables, such as gender and geographical locations, and analyzed their relationship with credit approval. Additionally, the dataset was divided into male and female groups to investigate the discrepancy in influential factors across different genders. The research is thus helpful in answering the following:

1. Finding the correlation between various factors and an applicant's credit acceptance ratio;
2. Analyzing if gender plays a role among these factors;
3. Analyzing if male and female applicants' applications are reviewed on the same standard.
4. Drawing implications from the differences between how the male group and the female group are assessed.

The paper is then organized as follows: section 2 introduces the dataset; section 3 summarizes the re-

sults of an exploratory data analysis and some background research; section 4 presents and interprets the results of a logistic regression and a lasso logistic regression model; section 5 compares model performance and analyzes the model results with female/male group segmentation; section 6 concludes the paper and discusses some limitations as well as future directions.

## 2. Dataset

This paper has applied data provided by the US Dream Housing Finance company that deals with all home loans. Customers only apply for a home loan after the company validates the customer's eligibility. As Dream Housing Finance company has a presence across all urban, semi-urban, and rural areas, this dataset provides valuable information on all applicants in the United States. We can thus use the dataset to build a predictive model for credit approval. The dataset contains the following information about the applicants:

Table 1. – Variables contained in the dataset

Variable	Type	Description
Credit History	Binary	Whether the applicant's history meets guidelines
Property Area	Categorical	Urban/ Semi Urban/ Rural
Gender	Binary	Male/ Female
Income	Continuous	Applicants' income per month
Loan Amount	Continuous	Loan amount in thousands
Loan Amount Term	Continuous	Term of loan in months
Dependents	Categorical	Number of dependents
Loan Status	Binary	Whether the loan approved
Loan ID	–	Unique Loan ID for each applicant
Education	Categorical	Applicant's education
Coapplicant's Income	Continuous	Coapplicant's income per month
Married	Binary	Whether the applicant married
Self Employed	Binary	Whether the applicant is self-employed

## 3. Preliminary Analysis

This section summarizes some findings from our background research, where we completed an analysis of skewness within data and utilized the data provided to compare the approval ratios across different groups.

### 3.1 Exploratory Data Analysis

Before proceeding with a rigorous statistics model, we first conducted an exploratory data analysis to understand the data better. We visualized the distribution scatter plot using R:

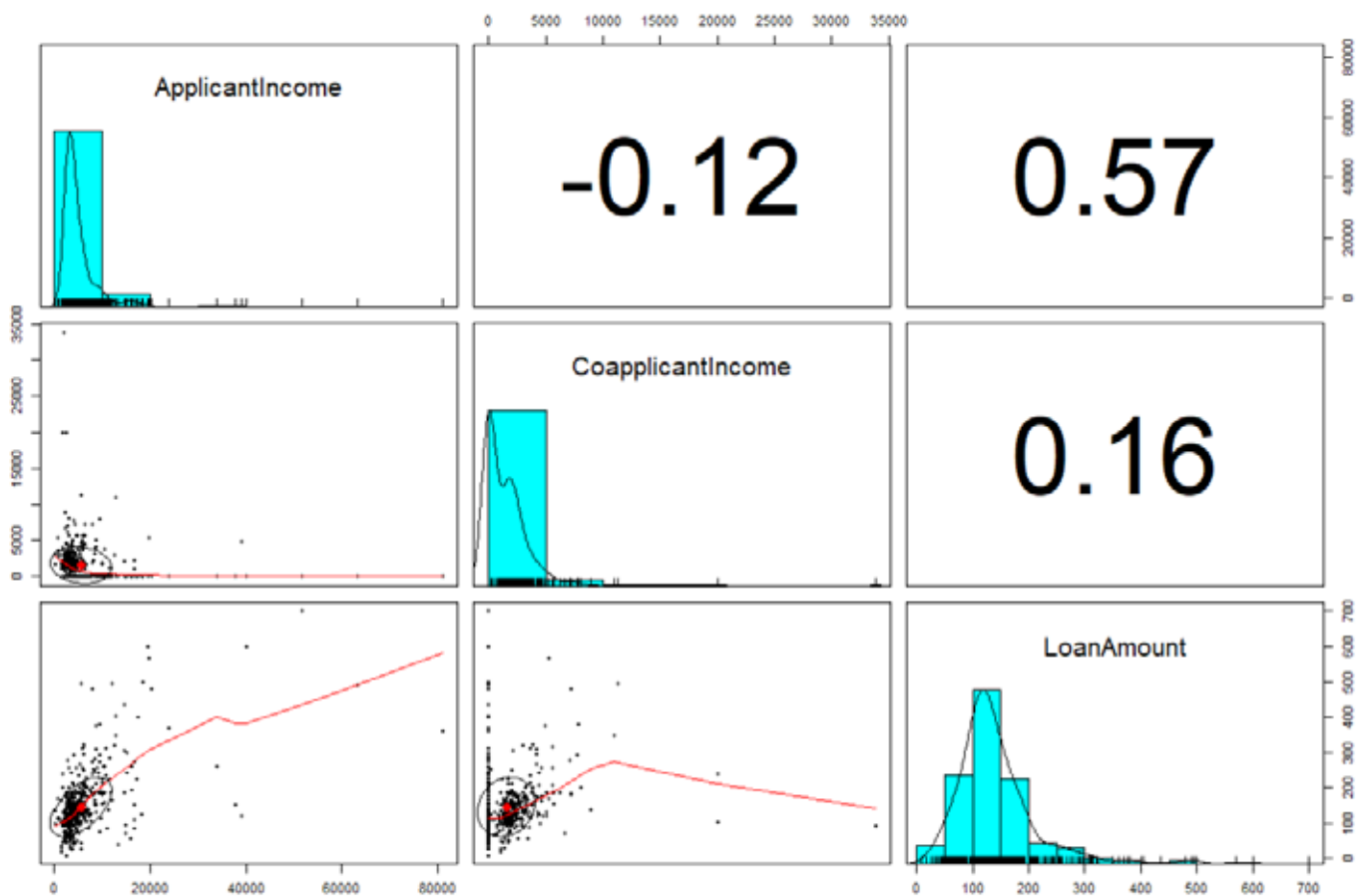


Figure 1. Preliminary analysis results

The plot indicates that the loan amount variable is normally distributed while the applicant's and co-applicant's income distribution demonstrates skewness. This is consistent with reality as the income distribution in modern society is skewed. Another observation from this exploratory data analysis is that the applicant's income is positively correlated with the loan amount. The correlation between them is 0.57, which suggests that the applicant with more income is seeking a higher credit amount. We then decided that a careful examination of the highly correlated independent variable should be addressed from the result. Considering the skewed distribution behind the co-applicant and applicant's income, we dropped the observations with missing income without outlier filtering because the outliers are expected to appear in the applicant's income, while filtering outliers will skew data distribution.

### 3.2 Credit-Debt Ratio

Theoretically speaking, in the credit application process, when assessing applicants' credit applications, lenders especially value their ability to repay loans. This can be illustrated by applicants' debt-to-income ratio. The debt-to-income ratio is the amount of debt that applicants have relative to their income [4]. Thus, we hypothesized that a factor's significance to the credit application is associated with how it reveals about the applicants' financial ability.

To test this hypothesis, we computed the ratio between applicant income and the loan amount and divide the data into five groups according to the calculated quantile statistics. The first group thus represents the applicant with the lowest debt-to-income ratio.

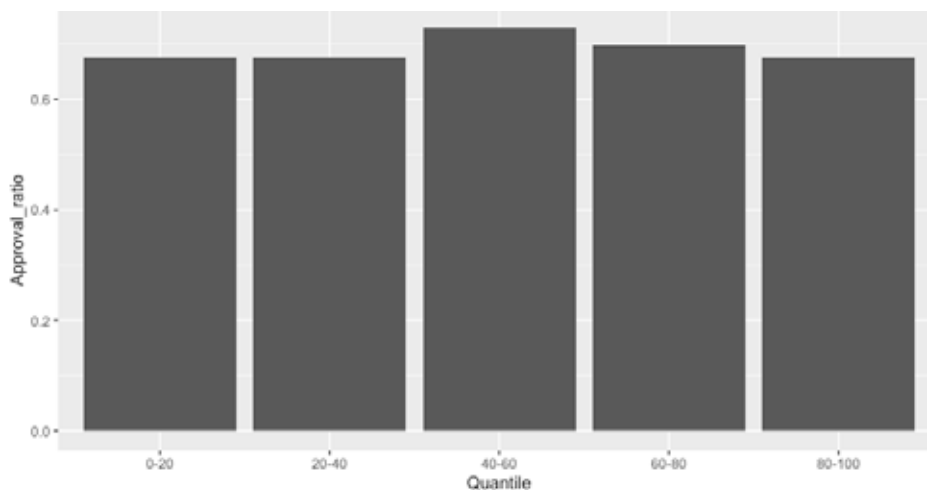


Figure 2. Application approval ratios for different credit-to-debt ratios

However, as the figure indicates, there seems to be no clear correlation between the debt-to-income ratio and the application result.

### 3.3 Credit History

Another potentially relevant factor is whether the applicant maintains a good credit history. Credit history refers to all the information stored in the credit report, such as credit accounts, balances due, bankruptcies, etc. An applicant's credit history, or credit report, then translates into a nu-

merical calculation known as the credit score. This score is used to assess applicants' creditworthiness [5]. In other words, maintaining a good credit history showcases the applicant not only has the "ability to repay debts" but also has "demonstrated responsibility in repaying them," which accounts for why it's crucial when applicants apply for credits [6]. We used credit history as an example to test this hypothesis. The results are shown in the following graph.

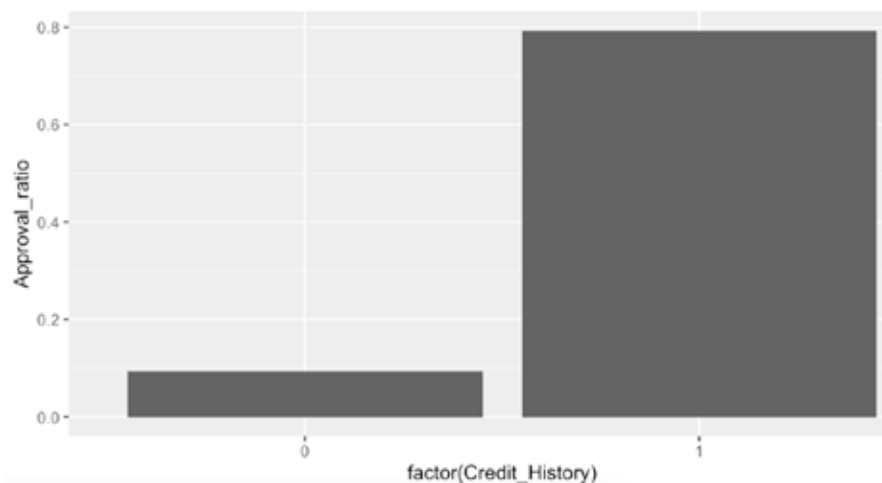


Figure 3. Application approval ratios for different credit history

This graph illustrates that of all applicants whose credit history meets guidelines, roughly 80% of them were approved for the loan. In contrast, only about 10% of those whose credit history is below the guidelines were approved for their applications.

### 4. Logistic Regression Models

To carefully examine the influential factors of credit approval, we adopted a more rigorous model that analyzes all the variables in-depth. The logistic regression model achieves this by connecting the

credit approval probability with all the independent variables. A lasso logistic regression is also studied and compared with the results from vanilla logistic regression model to investigate whether the model performance can be improved by taking the high correlation among variables into consideration.

**4.1 Model Setup**

Researchers have used logistic regression to predict the probability of certain events. In this paper, the model will help determine whether an applicant's characteristics will increase or decrease the likelihood of getting their application approved. The model starts by assuming:

$$\text{Log} \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

where:

$p$  – is the probability for the applicant to have its credit application approved;

$X_1, X_2, \dots, X_p$  are  $p$  variables relevant to the credit application.

Note that one can obtain the probability of approval by taking the

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}$$

Taking derivative with respect to  $\beta_1$ , the model transforms into

$$\frac{dp}{dX_1} = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}{(1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p))^2} \beta_1$$

The model is thus assuming that one unit of change of  $X_1$  will change  $\frac{dp}{dX_1}$  unit of probability.

**4.2 Binomial Probability**

We are given the outcome  $Y$  of individual credit application result. We use  $Y = 0$  to indicate that the application being denied and  $Y = 1$  to indicate the application being approved. With only one data and probability  $p$  to model the instance's chance of getting approved, we could use Bernoulli distribution to describe the data:

$$P(Y) = p^Y (1-p)^{1-Y}$$

Now suppose we are given application result  $(Y_1, Y_2, Y_3, \dots, Y_n)$  with their corresponding application approval probability  $(p_1, p_2, p_3, \dots, p_n)$ , the binomial model could thus be used to describe the data:

$$P(Y_1, Y_2, Y_n, \dots, Y_n) = \pi_{i=1}^n p_i^{Y_i} (1-p_i)^{1-Y_i}$$

Combined with the result of 4.1, we know that the probability to have its application approved is given by

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})}$$

**4.3 Model Solution**

Since from logistic regression model we have the probability of credit approval  $p(x, \beta)$  as a function of data  $x$  and coefficients  $\beta$ , we could thus solve for  $\beta$  by maximizing the binomial probability model. For my research, we used R glm function to solve for  $(\beta_0^*, \beta_1^*, \dots, \beta_p^*)$  that maximize  $\pi_{i=1}^n p_i^{Y_i} (1-p_i)^{1-Y_i}$

with  $p_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})}$

The output indicates:

```

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.429e+00 9.312e-01 -2.609 0.00909 **
GenderMale 3.254e-01 3.309e-01 0.983 0.32548
MarriedYes 5.739e-01 2.924e-01 1.963 0.04970 *
Dependents1 -3.756e-01 3.460e-01 -1.085 0.27771
Dependents2 2.770e-01 3.782e-01 0.733 0.46378
Dependents3+ 1.884e-01 4.874e-01 0.386 0.69915
EducationNot Graduate -4.210e-01 3.033e-01 -1.388 0.16510
Self_EmployedYes -1.492e-01 3.523e-01 -0.423 0.67202
ApplicantIncome 6.945e-06 2.862e-05 0.243 0.80827
CoapplicantIncome -5.143e-05 4.307e-05 -1.194 0.23246
LoanAmount -2.737e-03 1.773e-03 -1.544 0.12270
Loan_Amount_Term -9.253e-04 2.032e-03 -0.455 0.64885
Credit_History 3.650e+00 4.331e-01 8.427 < 2e-16 ***
Property_AreaSemiurban 9.873e-01 3.036e-01 3.253 0.00114 **
Property_AreaUrban 1.511e-01 3.007e-01 0.503 0.61527
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Figure 4. Logistic regression results

Notice that here the Intercept represents the baseline, single female applicants without dependents and are not self-employed.

**4.4 Results Interpretation**

A lot of the predictor variables are indicator variables, which facilitate the presence/absence analysis.

$$\text{Log} \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

The odds ratio (probability of approval vs denied)  $\frac{p}{1-p}$  increases according to the increase of  $X_1$  if  $\beta_1$  is positive, and the odds ratio decrease according to the increase of  $X_1$  if  $\beta_1$  is negative. From the fitted  $\beta$ , we observed that  $\beta_2 = 0.5739$ ,  $\beta_{13} = 3.6$ ,  $\beta_{14} = 0.98$  are all significantly away from 0 (indicated by the Z statistics), those parameters indicate that while all the other conditions equal:

- The presence of  $\beta_2$  (marriage) increases the credit card approval probability
- The presence of  $\beta_{13}$  (credit history) increases the credit approval probability
- The presence of  $\beta_{14}$  (living in a semi-urban area) increases the credit approval probability
- The loan amount  $\beta_{11}$  and loan terms  $\beta_{12}$  negatively affect the application result but the effect is not very significant.
- Male applicants  $\beta_1$  tend to have higher approval probability (indicated by positive slope) but the effect is not statistically significant (as it is indicated by the z value).

The results are generally consistent with the preliminary analysis results in section 3. There are also some findings that deserve more in-depth analysis, such as the effect of gender and credit loan amount in application approval.

#### 4.5 Lasso Logistic Regression

Analysis in section 3 indicates that some continuous variables demonstrate significant correlation. This is problematic for logistic regression. We thus further researched the penalized logistic regression, which seeks the balance between maximizing the probability and minimizing the adequate number of parameters:

$$\text{argmin}_{\beta} -\frac{1}{N} \sum_{i=1}^n y_i \log(p_i) + (1-y_i) \log(1-p_i) + \lambda (|\beta_0| + |\beta_1| + \dots + |\beta_p|)$$

Where the first summation term is the negative of the log-likelihood from binomial probability, the

second term is a penalization term that increases when some of  $(|\beta_0|, |\beta_1|, \dots, |\beta_p|)$  are not 0. The  $\lambda$  measures the penalization strength. As a result, the minimization of the above equation strikes a balance between maximizing the probability and minimizing the effective number of parameters.

We adopt R glmnet package to find optimal  $\lambda = 0.0271032$ .

	s1
(Intercept)	-2.205579e+00
Gender	.
Married	3.762623e-01
Dependents	.
Education	-4.340732e-02
Self_Employed	.
ApplicantIncome	.
CoapplicantIncome	-8.038692e-06
LoanAmount	-8.497286e-04
Loan_Amount_Term	.
Credit_History	3.075231e+00
Property_Area	.
deb_income_group	.

Figure 5. Lasso logistic regression results

The outputted lasso logistic regression model is:

$$\text{Log} \frac{p}{1-p} = \beta_0 + \beta_2 X_2 + \beta_{13} X_{13}$$

Where

- The presence of  $\beta_2$  (marriage) increases the credit approval probability;
- The presence of  $\beta_{13}$  (credit history) increases the credit approval probability;
- All the other coefficients have minor impact on the credit approval probability.

This conclusion differs slightly from the logistic regression model where the presence of  $\beta_{14}$  (living in a semi-urban area) increases the credit approval probability in the logistic regression model. The lasso logistic regression model tends to attribute the “in a semi-urban area” to marriage or the intercept variable.

#### 5. Model Comparison and Further Analysis

After careful analysis of the logistic and lasso logistic regression, we found a similar conclusion regarding the feature importance. However, it is

also important to compare those two models to determine which model will be used for follow-up research. Therefore, we used this section to evaluate and compare the predictive performance of the two models. We found that logistic regression has performed slightly better than the lasso logistic regression. As a result, we decided to apply logistic regression for male/female segmentation research.

### 5.1 Train-Test Dataset Split

The dataset is partitioned into two smaller datasets for training and test purposes: the training dataset for model development and the test dataset for model test and validation. Specifically, we randomly selected 75% of the 480 data sample to fit the model and used the rest 25% of the 480 data to validate the result.

### 5.2 Accuracy Score

Consider a two-class prediction problem, where the outcomes are labeled either as positive or negative. There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and the actual value is also positive, then it is called a true positive (TP); however, if the actual value is negative, then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are negative, and false negative (FN) is when the pre-

diction outcome is negative while the actual value is positive. In this way, the true positive rate (TPR) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN}$$

And the false positive rate (FPR) can be calculated as:

$$FPR = \frac{FP}{TN + FP}$$

Then the accuracy score is commonly used to judge for the model performance, which is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The logistic model achieved an accuracy score 0.800 on the test set, while the lasso logistic model outputted an accuracy score of 0.798. The out of sample result indicates great predictability for both models, and the logistic regression has performance slightly better than the other.

### 5.3 ROC plot

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold varies. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [7].

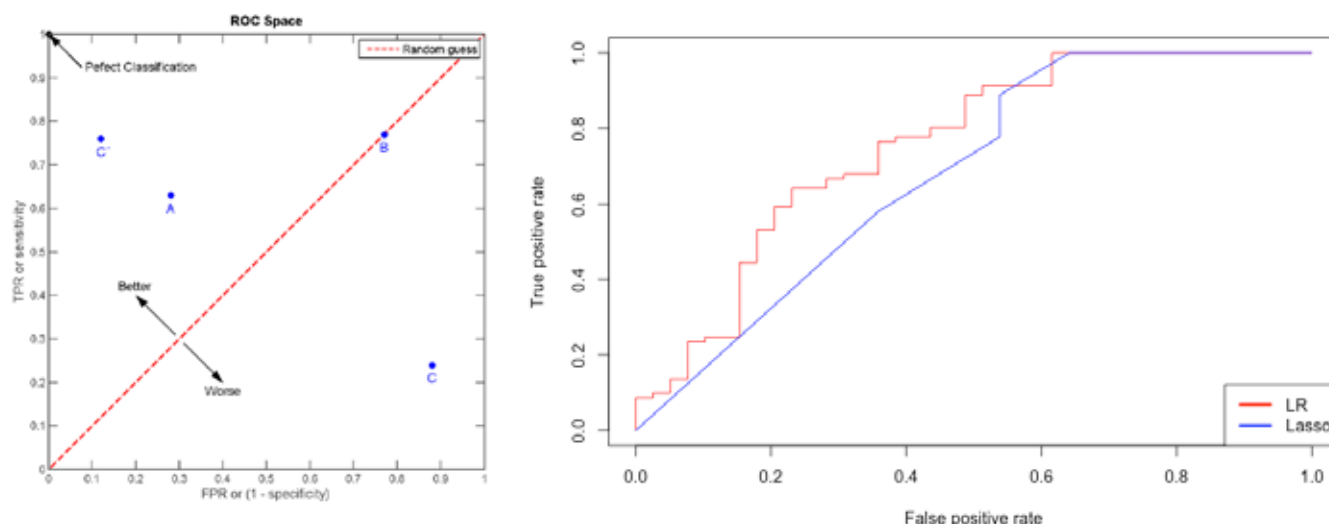


Figure 6. Sample ROC plot (left), ROC plot for logistic and lasso logistic regression (right)

The best possible prediction method would yield a point in the upper left corner of the ROC space. A random guess would give a point along a diagonal line from the left bottom to the top right corners. Points above the diagonal represent better than random classification results, while points below the line represent worse than random results. A sample ROC plot is shown below. In general, ROC analysis is one tool to select possibly optimal models and to discard suboptimal ones independently from the class distribution. Sometimes, it might be hard to identify which algorithm performs better by directly looking at ROC curves. Area Under Curve (AUC) overcomes this drawback by finding the area under the ROC curve, making it easier to find the optimal model.

As shown on the right of figure 6, the logistic regression model has performed better than the lasso logistic regression model at the most discrimination threshold. Combined with section 5.2, we concluded that the logistic regression model has an overall better performance than lasso logistic regression. Therefore, we only applied logistic regression in the following sections for further analysis.

#### 5.4 Logistic Regression on Segmented Data

This section will build on the previous model solution. However, instead of assessing the importance of each factor on the approval ratio for all applicants, this section will break the data according to a specific group (i.e. female versus male) and attempt to investigate whether male and female applicants' are assessed differently.

#### 5.5 Findings and Implications of the Results

From the results above, credit history and the property area of the applicants continue to have a significant impact on credit approval despite splitting the dataset into the male group and the female group. However, male and female applicants are assessed rather differently despite these two factors. Male applicants' approval rates are significantly impacted by their educational background. Specifically, the approval probability for male applicants without a

graduate degree is much lower, suggesting that banks value their academic experience. On the other hand, interestingly, two factors that play a crucial role in assessing female applicants are their income and co-applicant income. While it is true that banks especially value applicants' loan to income ratio, the income factor does not play a role when assessing both male and female applicants, nor was it shown critical in the preliminary analysis in section 3. This suggests that the credit quality of women is still poor, and they are still financially prejudiced by banks. Suppose such discrimination against the financial situation of women persists. In that case, it will lead to a detrimental cycle with women having more difficulty applying for loans and thus contribute negatively to their financial situation.

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.324e+00  2.662e+00  -2.751  0.00594 **
MarriedYes   4.531e-01  7.578e-01   0.598  0.54984
Dependents1  -6.758e-01  8.155e-01  -0.829  0.40727
Dependents2  -2.248e+00  1.402e+00  -1.603  0.10895
Dependents3+ 1.637e+01  1.455e+03   0.011  0.99103
EducationNot Graduate 6.089e-01  8.621e-01   0.706  0.47996
Self_EmployedYes -1.649e+00  1.059e+00  -1.556  0.11960
ApplicantIncome 2.063e-04  1.254e-04   1.645  0.09990 .
CoapplicantIncome 4.586e-04  2.608e-04   1.758  0.07868 .
LoanAmount   -3.370e-03  4.915e-03  -0.686  0.49289
Loan_Amount_Term 4.212e-03  5.073e-03   0.830  0.40642
Credit_History 5.228e+00  1.680e+00   3.112  0.00186 **
Property_AreaSemiurban 1.859e+00  7.669e-01   2.424  0.01536 *
Property_AreaUrban 1.043e+00  8.594e-01   1.214  0.22471
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 7. Logistic regression results for female applicants

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.374e+00  9.973e-01  -1.378  0.1682
MarriedYes   5.539e-01  3.262e-01   1.698  0.0895 .
Dependents1  -2.305e-01  4.008e-01  -0.575  0.5653
Dependents2  4.964e-01  4.080e-01   1.217  0.2237
Dependents3+ 1.829e-01  5.051e-01   0.362  0.7173
EducationNot Graduate -6.359e-01  3.393e-01  -1.874  0.0609 .
Self_EmployedYes -1.247e-01  3.968e-01  -0.314  0.7533
ApplicantIncome -5.732e-06  3.383e-05  -0.169  0.8655
CoapplicantIncome -6.657e-05  4.556e-05  -1.461  0.1440
LoanAmount   -3.377e-03  2.085e-03  -1.620  0.1053
Loan_Amount_Term -2.266e-03  2.333e-03  -0.971  0.3314
Credit_History 3.641e+00  4.764e-01   7.643  2.12e-14 ***
Property_AreaSemiurban 8.791e-01  3.455e-01   2.544  0.0109 *
Property_AreaUrban 4.431e-02  3.329e-01   0.133  0.8941
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 8. Logistic regression results for male applicants



## 6. Conclusion

In this project, we studied the driving forces behind credit applications. With the home loan dataset, we completed preliminary research and conducted an in-depth analysis using logistic regression and penalized logistic regression for credit approval prediction. we found:

1. The geographical location of the applicants, the marriage status of the applicant and the credit history of the applicant all have significant effects on applicants' application result.

2. Application income are surprisingly not the first concerns when financial associates approve for credit application.

3. Among all the factors affecting credit application, credit history is the most important factor when banks review their applicants.

4. Female applicant is additionally assessed by their income and their co-applicant income.

5. Male applicants are additional assessed by their education and marital status.

Though with two models to reach the conclusion, it is important to recognize that this dataset

does not represent the application process for all types of loans and across all countries and under all economic circumstances. Thus, further research using a variety of dataset is needed to confirm the findings of this paper. Another limitation of this study is that data entries with missing values are excluded for analysis. This is a timesaving but defective approach. Depending on the number of data entries with missing values, we may have removed too many sample points, which may weaken the conclusion we draw from the model. Therefore, for future studies, we may use more advanced techniques such as mean value imputation or  $k$ -nearest neighbors (kNN) to impute a value for the missing entries. The mean value imputation method completes missing values with the mean of the entire feature. This is a simple and effective way to make those entries usable by the logistic regression model. Other techniques include the  $k$ -nearest neighbor approach, which replaces missing values with the mean of  $k$  (a value assigned by users) nearest neighbors of that sample [8]. This technique requires more effort but can generally achieve better performance.

## References:

1. Kaggle. "Loan Data Set." kaggle.com. <https://www.kaggle.com/burak3ergun/loan-data-set>.
2. Arun Kumar, Garg Ishan and Kaur Sanmeet. "Loan approval prediction based on machine learning approach." IOSR J. Comput. Eng – 18.3. (2016): 18–21.
3. Siles Marcelo, Steven D. Hanson and Lindon J. Robison. "Socio-economics and the probability of loan approval." Applied Economic Perspectives and Policy – 16.3 (1994): 363–372.
4. Investopedia and Thomas Brock. "Credit History Definition." Investopedia, 2019. URL: <https://www.investopedia.com/terms/c/credit-history.asp>
5. Insurance Information Institute. "Credit Score vs Credit History | III." www.iii.org. Accessed July 17, 2021. URL: <https://www.iii.org/article/what-difference-between-my-credit-score-and-my-credit-history>
6. Tatham Matt. "How Lenders View Your Credit." Experian.com, September 6, 2018. URL: <https://www.experian.com/blogs/ask-experian/how-lenders-view-your-credit>
7. Google. Classification: ROC Curve and AUC | Machine Learning Crash Course. Accessed November 25, 2021. URL: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>
8. Kozma Laszlo. "k Nearest Neighbors algorithm (kNN)." Helsinki University of Technology. (2008).