UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

Report

LOAN APPROVAL PREDICTION

Natalia TSEPELEVA

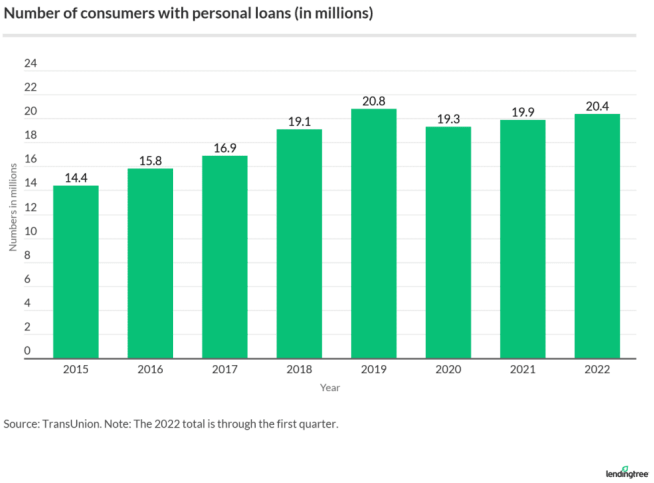
M1 BDEEM

January'23

Index

Introduction

Nowadays the banking industry is growing and more and more people are applying for bank loans. According to the TransUnion more than 20 million Americans have a personal loan as of the first quarter of 2022, up from 19 million in the first quarter of 2021.



Large portions of the bank's assets directly come from the interest earned on the loans they give out. Therefore, loan approval is an essential process for financial institutions.

But since the bank has limited resources they can grant loans to only a limited number of people. At the same time it is very challenging to predict whether a customer will be able to repay a loan or not. So lending carries a high level of risk for the banks. Determining to whom the loan can be granted is a complex process for the banks. They have to take into account and consider a wide range of factors. So there is a large use of analytics and data science methods in the banking industries.

In recent years many researchers have worked on prediction of loan approval systems. But still there is a lot of room for improvement in the way banks manage loan approval processes now.

That is why it seems interesting to work on the data regarding the loan approval process which I have chosen for my project.

Using the data I wanted to describe a profile of the loan applicants and find out what customer details (criteria) the banks use to make their decisions whether the applicant is eligible for a loan or not.

To solve these questions I used statistical analysis and estimated a few econometric models through the Ordinary Least Squares method and logical regression models.

Data Analysis

Data description

The dataset I used is from the website kaggle.com. This dataset consists of 615 rows (observations) and 15 features (variables) to predict whether the loan was approved or not.

The structure of the dataset is presented below in the table.

#	Number of the applicant
Loan_ID	Unique loan application ID
Gender	Gender of the applicant
Married	Marital status of the applicant
Dependents	Applicant's number of dependents
Education	Education level of the applicant (Graduated/Not graduated)
Self_Employed	Whether the applicant is self employed or not
ApplicantIncome	Income of the applicant
CoapplicantIncome	Income of the co applicant
LoanAmount	Amount of the loan the applicant applied for
Loan_Amount_Term	Repayment period of the loan
Credit_History	Whether the applicant has credit history of the repayments of his debts
Property_Area	Where the applicant has a property (Semiurban/Urban/Rural)
Loan_Status	The loan is approved or not
Total_Income	Sum of the applicant and co applicant income

Importing and cleaning the data

First of all, I imported the dataset to RStudio and imported all the necessary packages.

Then I cleaned the data and prepared it for the further analysis: deleted missing values, created some dummy variables, renamed or deleted some variables, changed the datatype of the variables where it was needed.

After this step the number of the observations became equal 390 and the structure of the dataset became the following:

```
loan_data          390 obs. of 15 variables
 $ Gender           : chr  "Male" "Male" "Male" "Male" ...
 $ Married          : num  1 1 1 0 1 1 1 1 1 1 ...
 $ Dependents       : num  1 0 0 0 2 0 3 2 1 2 ...
 $ Education        : chr  "Graduate" "Graduate" "Not Grad...
 $ Self_Employed    : num  0 1 0 0 1 0 0 0 0 0 ...
 $ ApplicantIncome  : int  4583 3000 2583 6000 5417 2333 3...
 $ CoapplicantIncome: num  1508 0 2358 0 4196 ...
 $ LoanAmount       : num  128 66 120 141 267 95 158 168 3...
 $ Loan_Amount_Term : num  360 360 360 360 360 360 360 360...
 $ Credit_History   : num  1 1 1 1 1 1 0 1 1 1 ...
 $ Property_Area    : chr  "Rural" "Urban" "Urban" "Urban"...
 $ Approv_Loan      : num  0 1 1 1 1 1 0 1 0 1 ...
 $ Total_Income     : num  6091 3000 4941 6000 9613 ...
 $ Female           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Graduated        : num  1 1 0 1 1 0 1 1 1 1 ...
```

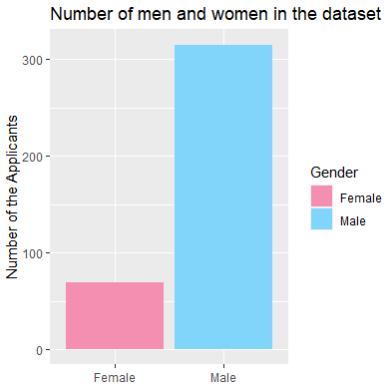
Statistical visualization

Most of the variables in the dataset are dummy variables. To describe the profile of the applicants I visualized all the unique values in the columns using barplots.

This shows which value is dominated in the dataset.

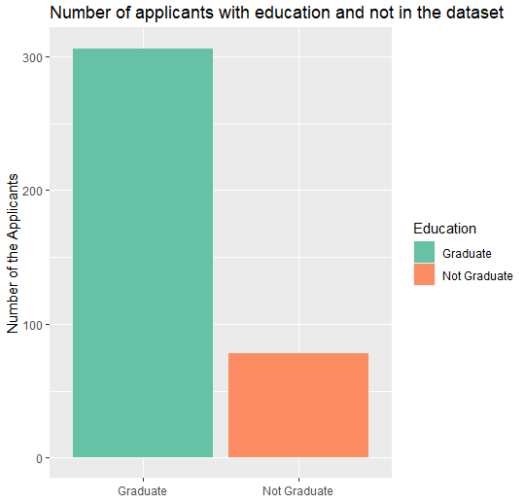
All the graphs are presented below.

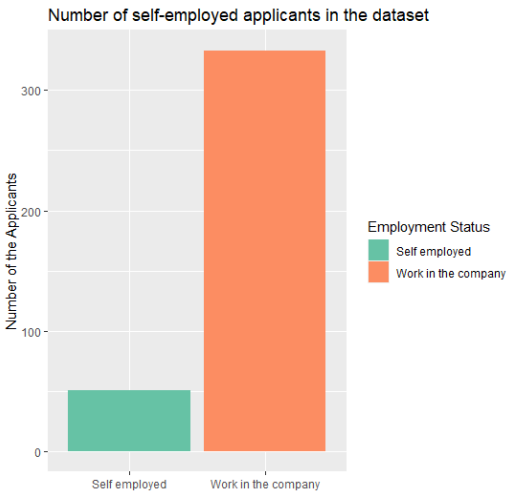
So we can see that the vast majority of the applicants in the dataset are males. There are only 69 women in the dataset, which is only around 18% of all applicants.



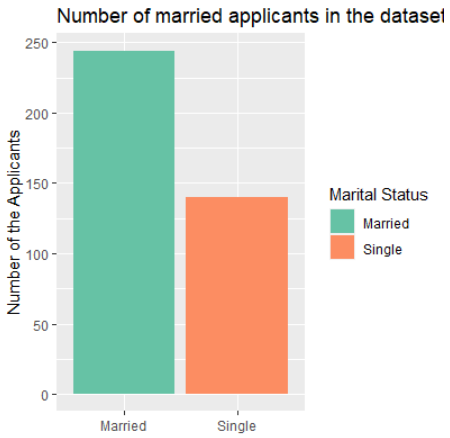

```
Female   Male
    69    315
> mean(loan_data$Female) #proportion of women
[1] 0.1796875
>
```

Also most of the applicants are graduates and work for the company (not self-employed). It is graphically represented on the barplots below.



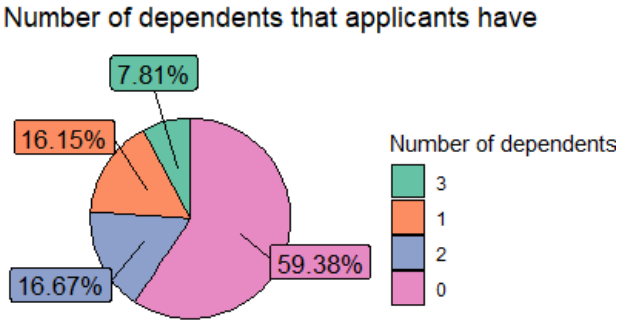


When we have a look at the marital status, we can see that almost twice more of the applicants in the dataset are married.



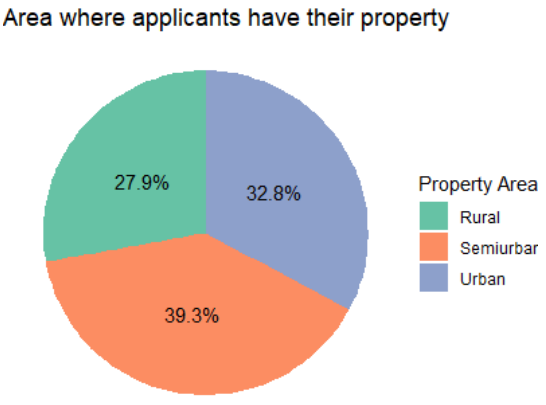
Also it seems interesting to have a look at the statistics of variables “Dependents” and “Property_Area”.

To see what values are dominating for these variables I decided to use pie charts.



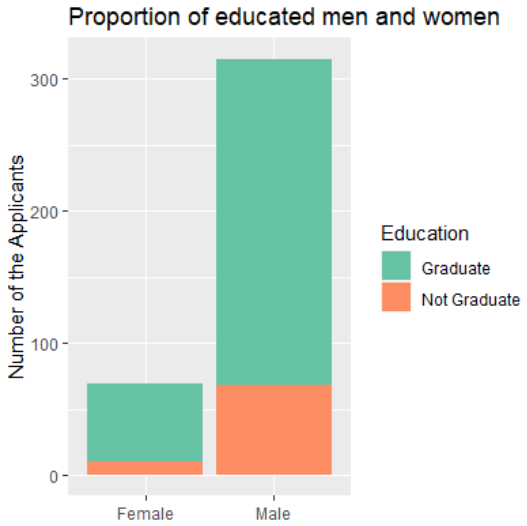
A bit more than half of the applicants (59,38%) do not have any dependents. The equal proportion of the applicants have one and two dependents - 16,15% and 16,67% respectively. And only 7,81% have three and more dependents.

Concerning the area where applicants have their property, we can say that it is equally proportioned: almost a third of the applicants have property in rural areas, third - in semi urban and third - in urban. But still most of the people have their property in semi urban areas (39,3%).



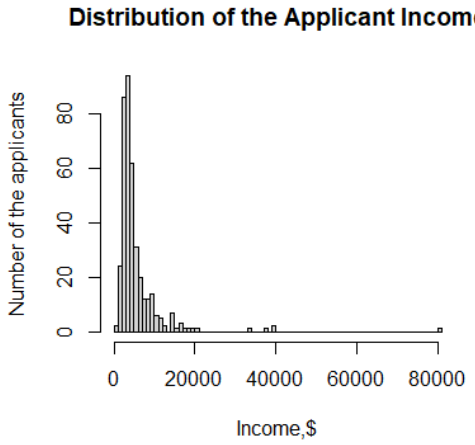
Statistical analysis

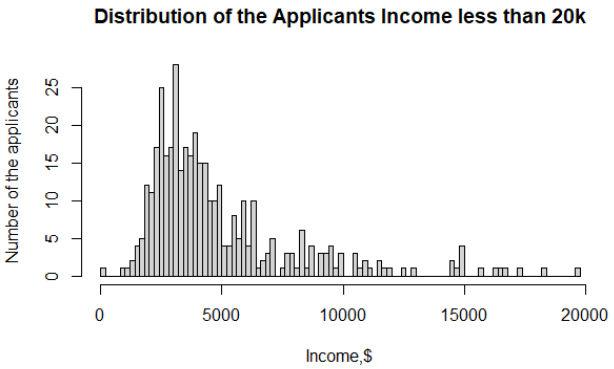
I found it interesting to consider what proportion of men and women graduated. Of course people with an education level prevail both in men and in women. 59 females have graduated and 10 have not. Among men 247 applicants have graduated while 68 applicants do not have an education level.




```
        Graduate Not Graduate
Female        59             10
Male         247             68
> 
```

After that I decided to consider the applicant's income. On the graph of the applicant's income distribution we can see that there are only a few observations with extremely high income. So it will be reasonable to drop these observations from the dataset in order not to influence the overall statistics.





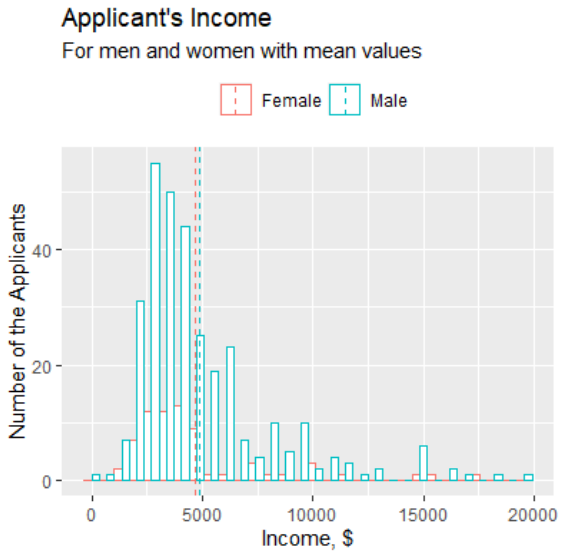
The statistics of the Applicant income became the following:

```
> summary(loan_data$ApplicantIncome)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    150    2865    3863    4849    5712   19730
> |
```

The minimum income is 150 dollars while the maximum is 19730 dollars.

After deleting the extreme values the average of the income has decreased from 5493,64 dollars to 4849,03 dollars.

However, if we consider the mean income for men and women, they have almost the same values. The average income for women in the dataset is equal to 4696 dollars and for men - 4884,5 dollars.




```
  Gender grp.mean
1 Female 4696.159
2   Male 4882.514
> |
```

To confirm that the income for men and women are the same in the dataset, I performed a statistical t-test with two hypothesis:

H_0: income means of men and women are equal

H_1: income means of men and women are not equal

At the 5%-level we fail to reject H_0. We do not have sufficient evidence to say that the mean income between men and women is different.

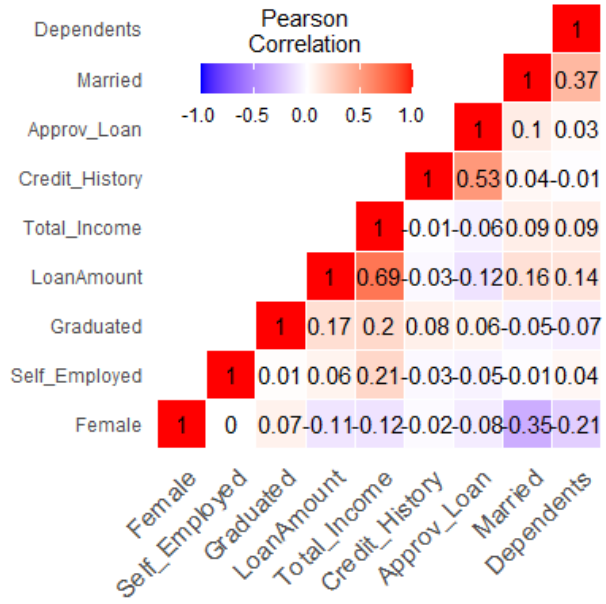
```
        Welch Two Sample t-test

data:  ApplicantIncome by Female
t = 0.43274, df = 98.337, p-value = 0.6662
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
 0
95 percent confidence interval:
 -668.2007 1040.9105
sample estimates:
mean in group 0 mean in group 1
       4882.514        4696.159
```

Econometric regressions

In this part I wanted to answer 2 questions.

1. What factors influence the amount of loan people apply for?
2. What is the probability for the loan to be approved taking into account the applicant's data?

Before doing regressions to answer these econometric questions I built correlation matrix to get some clues to what the data might yield and an explanatory overview of any dependence between variables in the dataset.



The heatmap of the correlation matrix is showing the correlation between loan amount and applicant income.

Also it shows that credit history has a high impact on loan status.

To see if there is any causation between variables I estimated econometric models.

1. What factors influence the amount of loan people apply for?

The average loan amount people applied for is 139 dollars. But what influences this amount?

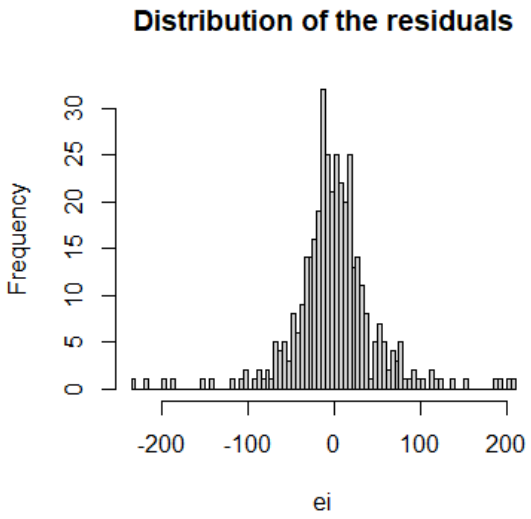
To answer this question I estimated the econometric model through the Ordinary Least Squares methods.

$$LoanAmount = \beta_0 + \beta_1 * Total\_Income + \beta_2 * Married + \beta_3 * Graduated + \beta_4 * Dependents + \beta_5 * Female + \varepsilon$$

```
Call:
lm(formula = LoanAmount ~ Total_Income + Married + Graduated +
    Dependents + Female, data = loan_data)

Residuals:
     Min       1Q   Median       3Q      Max
-232.656  -21.020   -1.622   19.242  209.328

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.199e+01  7.922e+00   5.301 1.96e-07 ***
Total_Income 1.274e-02  7.187e-04  17.720  < 2e-16 ***
Married      1.169e+01  5.980e+00   1.955   0.0513 .
Graduated    7.106e+00  6.512e+00   1.091   0.2759
Dependents   3.571e+00  2.764e+00   1.292   0.1971
Female       1.141e+00  7.166e+00   0.159   0.8736
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.96 on 378 degrees of freedom
Multiple R-squared:  0.4911,    Adjusted R-squared:  0.4844
F-statistic: 72.97 on 5 and 378 DF,  p-value: < 2.2e-16
```

Only total income has a significant influence on the amount of loan. If the total income is increased by 12$ the amount of loan will increase by 1000$ holding other criteria constant.

The distribution of the residuals for this regression model is the following:



It is not obvious from the graph if the residuals are normally distributed or not so I performed the Jarque Bera Test.

```
        Jarque Bera Test

data:  myreg$residuals
X-squared = 437.49, df = 2, p-value < 2.2e-16
```

The p-value is very small so we cannot say that the residuals are normally distributed. This means that the regression model I considered does not explain all trends in the dataset. So the amount of loan can not be totally explained by the applicant’s criteria presented in the dataset.

2. What is the probability for the loan to be approved taking into account the applicant's data?

The second econometric question I considered is the question of loan approval: whether a loan is approved according to the applicant's information.

To predict the probability of loan approval I used logit model:

$$logit(Approv\_Loan) = \beta_0 + \beta_1 * Total\_Income + \beta_2 * Credit\_History + \beta_3 * Married + \beta_4 * Graduated + \beta_5 * Self\text{-}Employed + \beta_6 * Female$$

```
Call:
glm(formula = loan_data$Approv_Loan ~ Total_Income + Credit_History +
    Married + Graduated + Self_Employed + Female, family = "binomial",
    data = loan_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0215  -0.4161   0.5899   0.6869   2.5094

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.324e+00  5.870e-01  -3.959 7.53e-05 ***
Total_Income   -5.813e-05  3.518e-05  -1.652   0.0985 .
Credit_History  3.657e+00  4.919e-01   7.434 1.06e-13 ***
Married         3.898e-01  2.897e-01   1.345   0.1785
Graduated       2.951e-01  3.331e-01   0.886   0.3757
Self_Employed  -1.461e-01  3.853e-01  -0.379   0.7045
Female         -4.299e-01  3.496e-01  -1.229   0.2189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 476.99  on 383  degrees of freedom
Residual deviance: 363.64  on 377  degrees of freedom
AIC: 377.64

Number of Fisher Scoring iterations: 4
```

Credit history is statistically different from zero and has an influence on loan approval. If the applicant already has a credit history his chances of getting a positive decision on the loan increases by 3,66.

Conclusion

Through this project I was interested in considering which criteria a bank takes into account in order to approve a loan or not.

I found out that while loan approval process banks take into account the previous credit history of the applicants. If the applicant has credit history, the chances to get a loan from the bank increases for him.

Also in the project I tried to find out on which criteria the amount of the loan people apply for depends on. Although we can not totally answer this question with the data presented in the dataset, we can say that the more income the applicant and his co applicant get, for the bigger loan amount he applies for.

Moreover during work on the project, I found out some additional trends regarding the education and income between men and women. Thus I got results that average income for men and women are equal.

To sum up, for this project the data analysis has been performed and all the research questions have been answered to this or that extent through statistical analysis and estimation of the econometric models.

References

Data source:
https://www.kaggle.com/datasets/vipin20/loan-application-data

Practical information for data visualization in R:
http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization
https://r-charts.com/part-whole/pie-chart-percentages-ggplot2/
http://www.sthda.com/english/wiki/ggplot2-pie-chart-quick-start-guide-r-software-and-data-visualization#:~:text=start%3D0
https://r-charts.com/part-whole/pie-chart-labels-outside-ggplot2/
http://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

Statistical Analysis:
https://stats.oarc.ucla.edu/r/dae/logit-regression/
https://www.statology.org/two-sample-t-test/

Additional information:
https://www.lendingtree.com/personal/personal-loans-statistics/#americansowebillioninpersonalloandebt
https://www.ijraset.com/research-paper/bank-loan-approval-prediction-using-data-science-technique