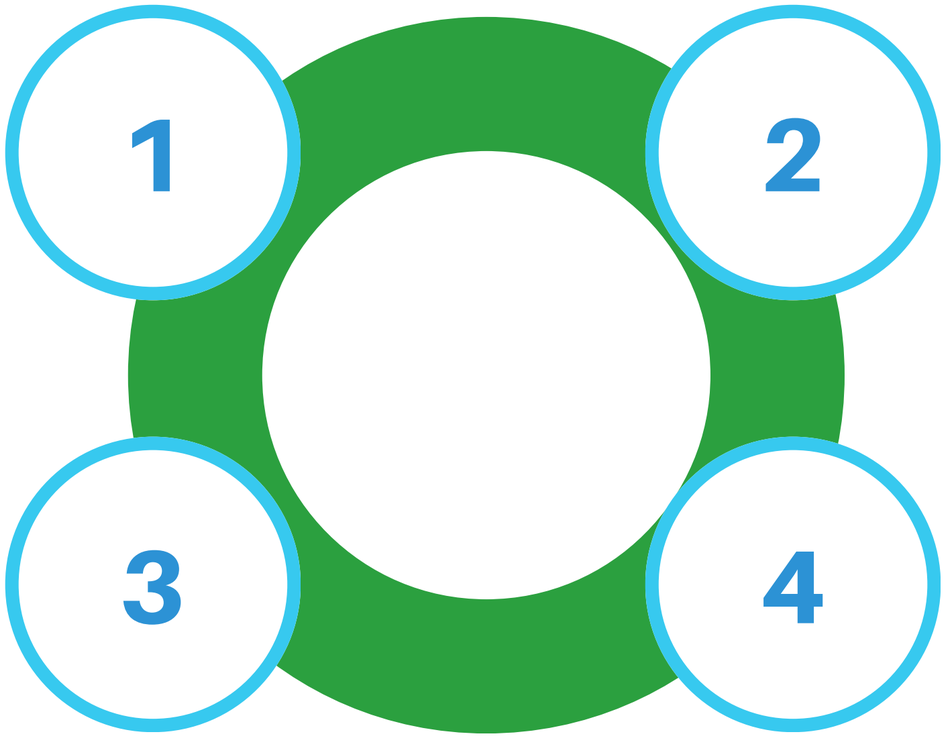
Loan Eligibility

Will you get the loan or not ?!

Ehsan Mokhtari
Omid Ghorbani
Sohrab Seyyedi Parsa

WHAT ARE THE SIGNIFICANT DETERMINANTS OF LOAN APPROVAL



WHATS IS LESS EFFECTIVE IN LOAN APPROVAL

CREDIT HISTORY IMPACT

LOAN APPROVAL AND INCOME RELATIONSHIP

Goals

1

analyze the data and find out which factors has the most effect on the loan eligibility.

2

1 - understand if gender effect the loan status
2 - design a model that can predict loan status base on relevant factors

Dataset

We found the "Loan Eligible Dataset" on Kaggle, which consists of 12 columns: "Loan ID", "Gender", "Married", "Dependents", "Education", "Self Employed", "Applicant Income", "Coapplicant Income", "Loan Amount", "Loan Amount Term", "Credit History", "Property Area", and "Loan Status" with a total of 614 rows.

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	L
Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	
Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	
Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	
Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	
Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	

This dataset was not ready for processing, so we preprocessed it to ensure its suitability for analysis.

PreProcessing

During preprocessing, we addressed missing values by filling NaNs, removed the "Loan_ID" column as it was not needed for analysis, and converted categorical string values (such as "Gender" and "Married") to numerical values to facilitate more effective data processing and analysis.

01. Dropping “Loan ID” column, because it was useless and does not have any effect on the loan eligibility process.

02. Filling all NaN values with “fill with mean” and “forward fill” methods.

03. Converting all string values to numerical values for example all “No” and “Yes” values are replaced with 0 and 1.

Variables

Gender

Credit History

self employed

married

Loan Eligibility

Applicant Income

dependents

Coapplicant Income

education

Property Area

Loan Amount Term

Loan Amount

Structure of Dataset

12 colAumns

614 rows

Gender

This columns shows if the gender of the person is male “1” or it is female “0”

Married

This column shows if the person is married “1” or not married “0”

Dependants

Shows the number of the dependents of the person :
“0” for 0 dependent
“1” for 1 dependent
“2” for 2 dependents
“3” for 3 and more dependents

Education

Shows if the person has university graduation “1” or not “0”

Self Employed

Shows if the person is self emploed “1” or not “0”

Structure of Dataset

Applicant Income

Shows the monthly income of the person

CoApplicant Income

Shows the monthly income of perosn's partner if he or she is married

Loan Amount

The amount of loan which the applicant requested

Loan Amount Term

Shows the number of months you will return the loan

Credit History

Does the applicant has a credit history "1" or not"0"

Property Area

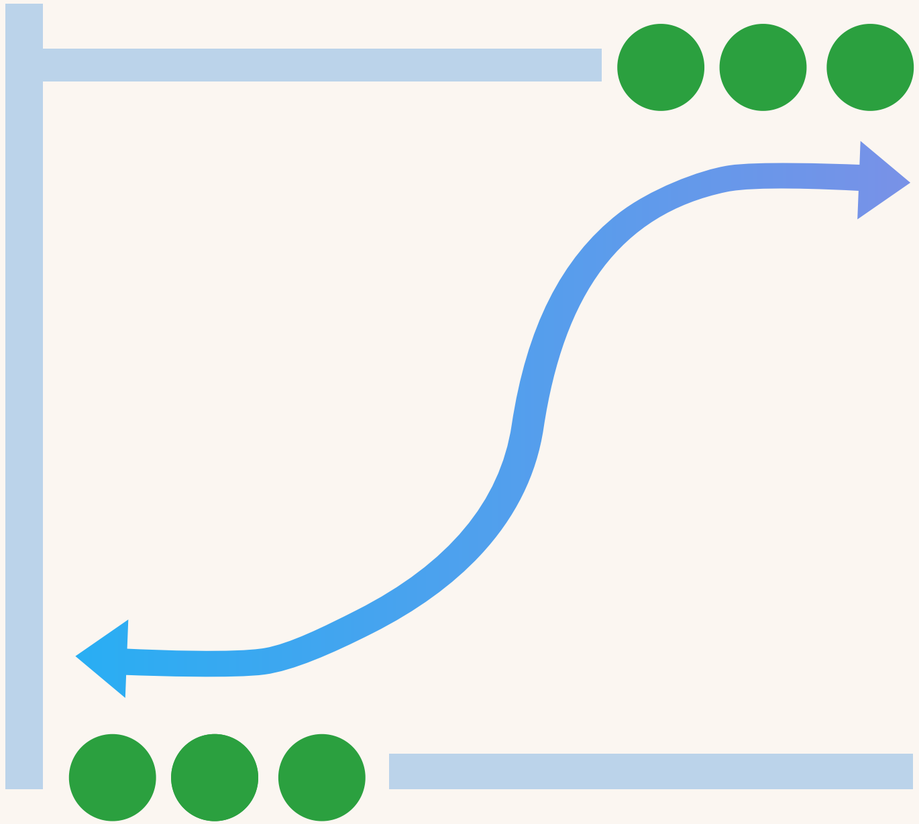
Where the applicant lives :

0 : Urban

1 : Rural

2 : SemiUrban

Model



Y = B1 * (Credit history) + B2 * (married status) + ...

the out put goes into a sigmoid function

$$S(x) = \frac{1}{1 + e^{-x}}$$

generate result between 0 and 1

Result of Fitting

1 - model is valid
2-loss decresse over iteration
3-some feature can be deleted

```
Iteration 0:   log likelihood = -381.44553
Iteration 1:   log likelihood = -290.94707
Iteration 2:   log likelihood = -290.20862
Iteration 3:   log likelihood = -290.20577
Iteration 4:   log likelihood = -290.20577

Logistic regression                                     Number of obs =    614
                                                        LR chi2(11)   = 182.48
                                                        Prob > chi2   = 0.0000
Log likelihood = -290.20577                             Pseudo R2     = 0.2392
```

loan_status	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
property_area	.352856	.1273992	2.77	0.006	.103158	.6025539
credit_history	3.266335	.3175302	10.29	0.000	2.643987	3.888682
loan_amount_term	-.000863	.0017359	-0.50	0.619	-.0042654	.0025393
loanamount	-.0019678	.0016218	-1.21	0.225	-.0051466	.0012109
coapplicantincome	-.0000476	.000035	-1.36	0.173	-.0001161	.0000209
applicantincome	7.91e-06	.0000218	0.36	0.717	-.0000349	.0000507
self_employed	.1133592	.3106051	0.36	0.715	-.4954157	.7221341
education	.4197628	.2540669	1.65	0.098	-.0781993	.9177248
dependents	.0375209	.1149171	0.33	0.744	-.1877125	.2627543
married	.5696293	.245844	2.32	0.021	.087784	1.051475
gender	-.0025668	.2878694	-0.01	0.993	-.5667804	.5616468
_cons	-2.351179	.7521847	-3.13	0.002	-3.825434	-.8769243

Check Correlation

	gender	married	depend~s	educat~n	self_e~d	applic~e	coappl~e	loanam~t	loan_a~m	credit~y	proper~a	loan_s~s
gender	1.0000											
married	0.3715	1.0000										
dependents	0.1645	0.3338	1.0000									
education	-0.0495	-0.0141	-0.0549	1.0000								
self_emplo~d	0.0117	-0.0003	0.0445	0.0087	1.0000							
applicanti~e	0.0462	0.0491	0.1150	0.1408	0.1227	1.0000						
coapplican~e	0.0870	0.0778	0.0267	0.0623	-0.0218	-0.1166	1.0000					
loanamount	0.0987	0.1470	0.1588	0.1670	0.1118	0.5656	0.1878	1.0000				
loan_amoun~m	-0.0754	-0.0953	-0.0847	0.0772	-0.0280	-0.0452	-0.0597	0.0388	1.0000			
credit_his~y	-0.0085	0.0074	-0.0703	0.0846	-0.0108	-0.0202	0.0094	-0.0188	-0.0184	1.0000		
property_a~a	-0.0860	-0.0008	0.0085	0.0036	0.0204	-0.0079	-0.0284	0.0138	0.0895	0.0162	1.0000	
loan_status	0.0122	0.0891	-0.0034	0.0859	0.0090	-0.0047	-0.0592	-0.0364	-0.0210	0.5252	0.1033	1.0000

.

Low Linear correlation - proved what we said logistic regression

checking multi collinearity

mean vif for all variable is low

we don’t need regularization, simple Logistic Regression can be enough

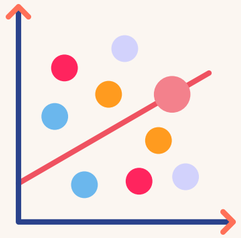
```
. vif

    Variable |      VIF       1/VIF
-------------+----------------------
  loanamount |     1.71    0.586481
applicanti~e |     1.63    0.613659
  dependents |     1.16    0.861350
     married |     1.15    0.868504
coapplican~e |     1.14    0.878360
   education |     1.06    0.946800
loan_amoun~m |     1.05    0.954999
self_emplo~d |     1.02    0.979180
credit_his~y |     1.01    0.985428
property_a~a |     1.01    0.989761
-------------+----------------------
    Mean VIF |     1.19

.
```

Chi sqare test – insight from data

Loan status

&

Credit history



```
. tabulate loan_status credit_history, chi2
```

Loan_Statu s	Credit_History 0	1	Total
0	87	105	192
1	14	408	422
Total	101	513	614

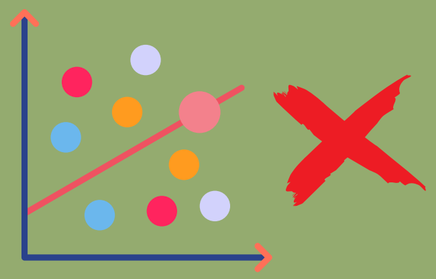
Pearson chi2(1) = 169.3315 Pr = 0.000

Chi sqare test – insight from data

Gender

&

Loan status



```
. tabulate loan_status gender, chi2
```

Loan_Statu s	Gender 0	1	Total
0	37	155	192
1	77	345	422
Total	114	500	614

Pearson chi2(1) = 0.0916 Pr = 0.762

result : No gender discrimination

Linear Restriction Test

```
. test gender coapplicantincome self_employed loanamount loan_amount_term dependents education applicantincome

 ( 1)  gender = 0
 ( 2)  coapplicantincome = 0
 ( 3)  self_employed = 0
 ( 4)  loanamount = 0
 ( 5)  loan_amount_term = 0
 ( 6)  dependents = 0
 ( 7)  education = 0
 ( 8)  applicantincome = 0

       F(  8,    602) =    0.96
            Prob > F =    0.4638
```

we can not reject null hypothesis

$$F = \frac{(\mathrm{SSR}_{\text{restricted}} - \mathrm{SSR}_{\text{unrestricted}})/q}{\mathrm{SSR}_{\text{unrestricted}}/(n-k)}$$

This means there is no huge diffrence between SSR of restrict & unrestricted model

Chi sqare test – insight from data

Although we remove 8 features

R2 is almost the same as previous model

```
. logit loan_status married credit_history property_area

Iteration 0:   log likelihood = -381.44553
Iteration 1:   log likelihood = -294.32994
Iteration 2:   log likelihood = -293.75556
Iteration 3:   log likelihood = -293.75285
Iteration 4:   log likelihood = -293.75285

Logistic regression                                     Number of obs =    614
                                                        LR chi2(3)    =  175.39
                                                        Prob > chi2   =  0.0000
Log likelihood = -293.75285                             Pseudo R2     =  0.2299

------------------------------------------------------------------------------
 loan_status | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     married |   .5294961   .2125203     2.49   0.013     .1129639    .9460282
credit_history |   3.259495   .3138874    10.38   0.000     2.644288    3.874703
 property_area |    .342592   .1242201     2.76   0.006     .0991251    .5860588
       _cons |  -2.567037    .361648    -7.10   0.000    -3.275854   -1.85822
------------------------------------------------------------------------------
```

Heteroscedasticity

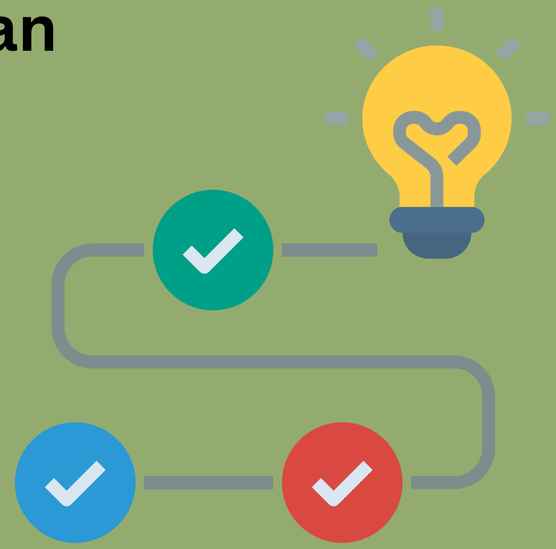
we failed to reject null hypothesis.

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of loan_status

         chi2(1)      =     0.91
         Prob > chi2  =   0.3406
```

we can not say the variance of data change when independent variable increase

Conclusion

1-we find out the both model are statisticaly signifact and valid
2-we can make our model simpler using linear restriction test
3- No gender discrimination have been observed’
4- credit history is the most import factor in giving loan



Thank you