LOAN APPROVAL PREDICTIVE MODEL

DANA 4820 Fall 2002 Project

Predict odds of loan approval using logistic regression models

Andrew Liu 100390239
Aswinee Rath 100389210
Patrick Ipac 100385706

Table of Contents

Project Background

Banks earn significant revenue from lending loans, but it is often associated with risk. The borrowers may default on the loan. To mitigate this issue, the banks use various parameters to decide whether the loan application should be approved. In this project, we used sample data available on Kaggle (https://www.kaggle.com/datasets/yasserh/loan-default-dataset) to evaluate what parameters influence the approval of a loan application.

The dataset consists of historical data of loan applicants of 148670 observations of 34 variables. The data has multiple deterministic factors (e.g., borrower's income, gender, loan purpose, etc.). The dataset is subject to solid multicollinearity & empty values.

Exploratory Data Analysis and Cleanup

The original dataset we used is a CSV file which consists of 148670 rows and 34 columns. The data dictionary provided by the data provides us with the following description for all the columns.

- ID = Customer ID of Applicant
- year = Year of Application
- loan limit = maximum available amount of the loan allowed to be taken
- Gender = sex type
- approv_in_adv = Is the loan pre-approved or not
- loan_type = Type of loan
- loan_purpose = the reason you want to borrow money
- Credit_Worthiness = is how a lender determines that you will default on your debt obligations, or how worthy you are to receive new credit.
- open_credit = is a pre-approved loan between a lender and a borrower. It allows the borrower to make repeated withdrawals up to a certain limit.
- business_or_commercial = Usage type of the loan amount
- loan_amount = The exact loan amount
- rate_of_interest = is the amount a lender charges a borrower and is a percentage of the principal—the amount loaned.
- Interest_rate_spread = the difference between the interest rates a financial institution pays to depositors and the interest rate it receives from loans
- Upfront_charges = Fee paid to a lender by a borrower as consideration for making a new loan
- term = the loan's repayment period
- Neg_ammortization = refers to a situation when a loan borrower makes a payment less than the standard installment set by the bank.
- interest_only = amount of interest only without principles
- lump_sum_payment = is an amount of money that is paid in one single payment rather than in installments.
- property_value = the present worth of future benefits arising from the ownership of the property
- construction_type = Collateral construction type
- occupancy_type = classifications refer to categorizing structures based on their usage
- Secured_by = Type of Collateral
- total_units = number of units
- income = refers to the amount of money, property, and other transfers of value received over a set period of time
- credit_type = type of credit
- co-applicant_credit_type = is an additional person involved in the loan application process. Both applicant and co-applicant apply and sign for the loan
- age = applicant's age
- submission_of_application = Ensure the application is complete or not

- LTV = lifetime value (LTV) is a prognostication of the net profit
- Region = applicant's place
- Security_Type = Type of Collateral
- ***status*** = Loan status (Approved/Declined)
- dtir1 = debt-to-income ratio

These variables capture all the details of a loan application. The status field indicates if the loan was approved or not. So, we will use the status field as our response variable and the rest of the variables as our explanatory variables.

In our preliminary analysis, we understood that these data elements are collected as part of the loan application. Only some of the variables directly influence the loan status approval. The supporting literature, and our investigation of financial services documents, indicate the following factors are used in the loan application approval process.

1. Credit score.
2. Income and employment history.
3. Debt-to-income ratio.
4. Value of your collateral.
5. Size of down payment.
6. Liquid assets.
7. Loan term.

We short-listed the following variables for our model analysis based on our literature study.

Variable Name	Type	Measurement
Status	Categorical	Nominal
Loan_type	Categorical	Nominal
Loan_Amount	Numerical	Continuous
Rate_Of_Interest	Numerical	Continuous
Term	Numerical	Discrete
Property_value	Numerical	Continuous
Income	Numerical	Continuous
Credit_Score	Numerical	Continuous
Age	Categorical	Ordinal
Dtir1	Numerical	Continuous

We used the Status field as our predictor or response variable.

The summary of the variables is as below.

```
     Status          loan_type         loan_amount     rate_of_interest      term       property_value        income       Credit_Score      age                dtir1
 Min.   :0.0000   Length:148670      Min.   :  16500   Min.   :0.00     Min.   : 96.0   Min.   :    8000   Min.   :     0   Min.   :500.0   Length:148670      Min.   : 5.00
 1st Qu.:0.0000   Class :character   1st Qu.: 196500   1st Qu.:3.62     1st Qu.:360.0   1st Qu.:  268000   1st Qu.:  3720   1st Qu.:599.0   Class :character   1st Qu.:31.00
 Median :0.0000   Mode  :character   Median : 296500   Median :3.99     Median :360.0   Median :  418000   Median :  5760   Median :699.0   Mode  :character   Median :39.00
 Mean   :0.2464                      Mean   : 331118   Mean   :4.05     Mean   :335.1   Mean   :  497893   Mean   :  6957   Mean   :699.8                      Mean   :37.73
 3rd Qu.:0.0000                      3rd Qu.: 436500   3rd Qu.:4.38     3rd Qu.:360.0   3rd Qu.:  628000   3rd Qu.:  8520   3rd Qu.:800.0                      3rd Qu.:45.00
 Max.   :1.0000                      Max.   :3576500   Max.   :8.00     Max.   :360.0   Max.   :16508000   Max.   :578580   Max.   :900.0                      Max.   :61.00
                                                       NA's   :36439    NA's   :41      NA's   :15098      NA's   :9150                                          NA's   :24121
```

Figure 1: Summary of data before cleaning

The summary indicates we have null values for rate_of_interest, term, property_value and dtir1. The loan amount and property value have outlier values that may impact our analysis.

	missing <int>
Status	0
loan_type	0
loan_amount	0
rate_of_interest	36439
term	41
property_value	15098
income	9150
Credit_Score	0
age	0
dtir1	24121

Figure 2: Missing Values

We used the available R function to calculate imputed values and used them to fill in the missing data for our columns.

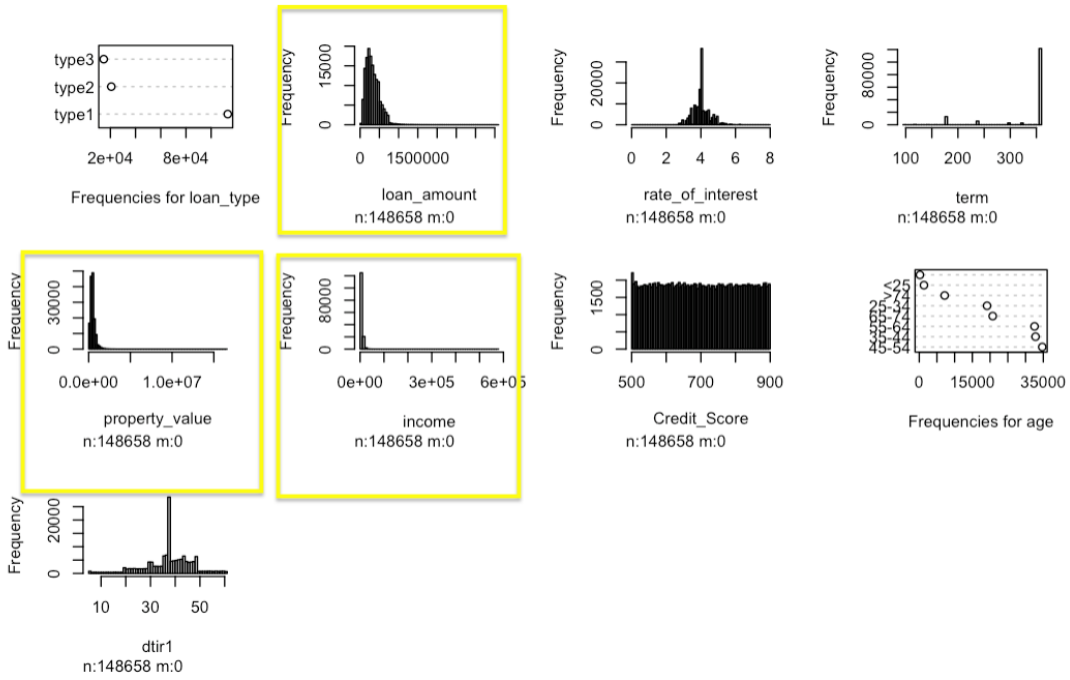


Figure 3: Outliers in our observations

We used R function to identify the outlier (Q3 + (1.5*IQR)) and remove outliers.

The post-cleanup summary of our dataset looks as below.

```
     Status          loan_type         loan_amount     rate_of_interest      term       property_value       income       Credit_Score       age                dtir1       
 Min.   :0.0000   Length:135126      Min.   : 16500   Min.   :0.000   Min.   : 96.0   Min.   :  8000   Min.   :    0   Min.   :500.0   Length:135126      Min.   : 5.00  
 1st Qu.:0.0000   Class :character   1st Qu.:186500   1st Qu.:3.750   1st Qu.:360.0   1st Qu.:278000   1st Qu.: 3660   1st Qu.:599.0   Class :character   1st Qu.:34.00  
 Median :0.0000   Mode  :character   Median :286500   Median :4.045   Median :360.0   Median :428000   Median : 5640   Median :699.0   Mode  :character   Median :37.73  
 Mean   :0.2499                      Mean   :306503   Mean   :4.048   Mean   :335.6   Mean   :435356   Mean   : 5944   Mean   :699.8                      Mean   :38.16  
 3rd Qu.:0.0000                      3rd Qu.:406500   3rd Qu.:4.250   3rd Qu.:360.0   3rd Qu.:548000   3rd Qu.: 7500   3rd Qu.:800.0                      3rd Qu.:44.00  
 Max.   :1.0000                      Max.   :766500   Max.   :8.000   Max.   :360.0   Max.   :998000   Max.   :14940   Max.   :900.0                      Max.   :61.00  
```

Figure 4: Post-Clean data summary

Our data also looked relatively normalized.

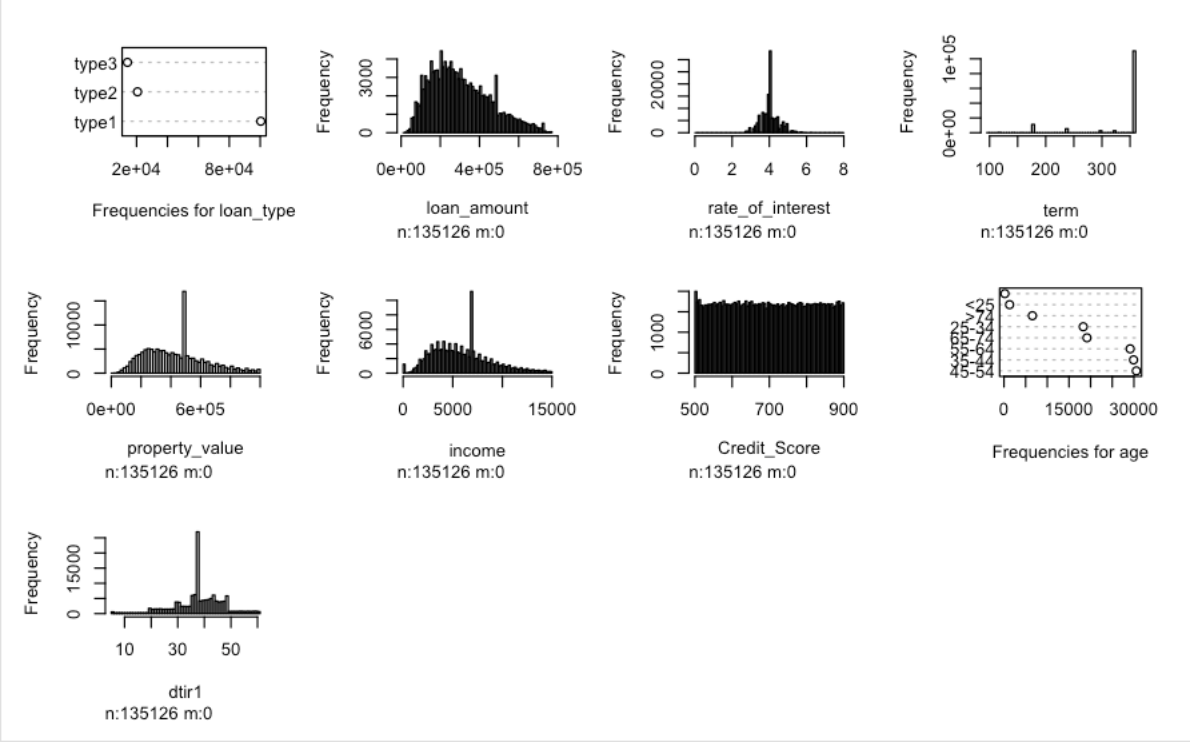


Figure 5: Post-cleanup data distribution

We also observed that in vast of our observations, the loan term is 360 (30 years), a common loan term in the USA.

Variable Selection

After data cleanup, we validated that our short-listed variable has a relationship between our explanatory variable and the response variable.

Categorical variable Analysis

A chi-square test of independence was conducted between the categorical variables and the response variable. The following hypotheses were developed:

Ho: The categorical variable and Status are independent of each other.
Ha: The categorical variable and Status are not independent of each other.

```
  type1 type2 type3
0 77103 13498 10526
1 23010  7125  3542

        Pearson's Chi-squared test

data:  df_imputed$Status and df_imputed$loan_type
X-squared = 1220.7, df = 2, p-value < 2.2e-16
```

```
           <25   >74 25-34 35-44 45-54 55-64 65-74
0     0   918  4565 14210 23139 22983 21392 13920
1   199   374  1997  4058  6689  7473  7679  5208

        Pearson's Chi-squared test

data:  df_imputed$Status and df_imputed$age
X-squared = 977.68, df = 7, p-value < 2.2e-16
```

Figure 6: Chisquare test for Categorical Variables and Status (response var)

Based on the results above, the p-value is <0.05. As such, we reject Ho and we conclude that Loan Type and Age are related to Status, based on 5% level of significance.

Furthermore, we did a two-sample t-test to check for association between numerical variables and the response variable. Since we are unsure if the variances between the numerical variables and the response variable are equal, we conducted an F-Test to determine what type of t-test will be performed. The following hypotheses were developed:

Ho: The population variances are equal.
Ha: The population variances are not equal.

```
        F test to compare two variances

data:  imputed_data$Status and imputed_data$rate_of_interest
F = 0.81703, num df = 134803, denom df = 134803, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8083508 0.8257972
sample estimates:
ratio of variances
         0.8170274
```

```
        Welch Two Sample t-test

data:  imputed_data$Status and imputed_data$rate_of_interest
t = -2160.1, df = 266900, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.801616 -3.794723
sample estimates:
mean of x mean of y
 0.249822  4.047991
```

```
        F test to compare two variances

data:  imputed_data$Status and imputed_data$dtir1
F = 0.0021585, num df = 134803, denom df = 134803, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.002135583 0.002181675
sample estimates:
ratio of variances
        0.002158506
```

```
        Welch Two Sample t-test

data:  imputed_data$Status and imputed_data$dtir1
t = -1492.6, df = 135385, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -37.97111 -37.87152
sample estimates:
mean of x mean of y
 0.249822 38.171140
```

```
        F test to compare two variances

data:  imputed_data$Status and imputed_data$Credit
F = 1.3958e-05, num df = 134803, denom df = 134803, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.380978e-05 1.410784e-05
sample estimates:
ratio of variances
       1.395801e-05
```

```
        Welch Two Sample t-test

data:  imputed_data$Status and imputed_data$Credit
t = -2216.5, df = 134807, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -700.1443 -698.9072
sample estimates:
 mean of x  mean of y
  0.249822 699.775585
```

Figure 7: t-test for Numerical variables

Based on the above tests, since the p-value is less than 0.05 for all the variables of our interest, we reject Ho and conclude that these variables are valid for our model testing, based on 5% level of significance.

We further did normality testing of our datasets to make sure the data was normally distributed for the numerical variables.

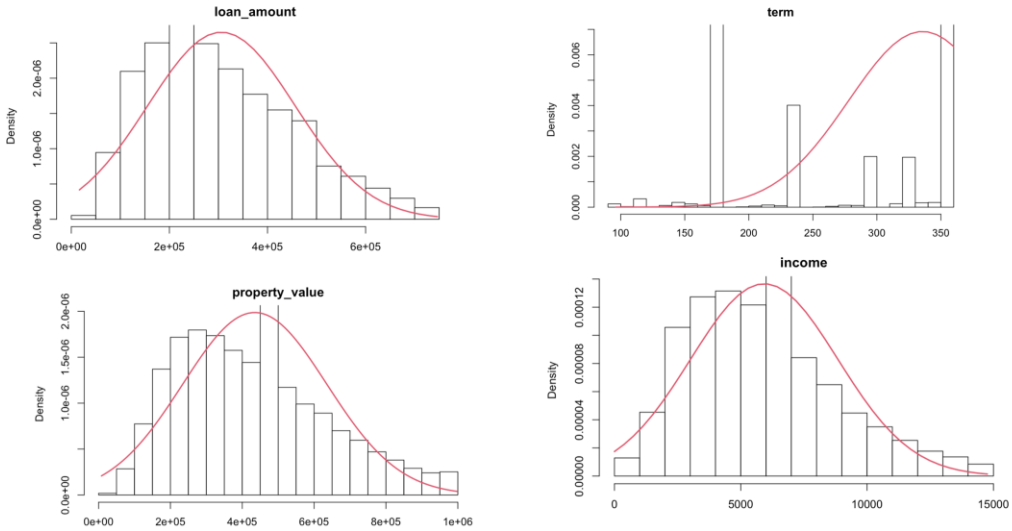


Figure 8: Normality Distribution of numerical data

The above plots indicate that the data is close to normal distribution for our numerical data except Term. However, Term represents the number of months for the loan, and loans are typically given for 30 years; we expect this data to be skewed.

Please see our R-code for other tests and test methods we performed to validate our variable selection process.

We split the data into two datasets, one with 80% of the data for our model training and the other with 20% to validate our model.

Model Evaluation

1 Methodology

The dataset has been split into a training dataset and a test dataset. The training dataset contains 80% of the original dataset. We used the training dataset to generate two binomial GLM models, one containing one interaction term and the other without interaction terms. The test dataset contains the other 20% of the data exclusive to the training dataset. We ran the training dataset-generated models on test data to evaluate the fitness of the models.

2 Non-Interaction and Interaction Model Comparison

2.1 The non-interaction model

We run our model with all the variables in our dataset to identify if all the variables are significant in our analysis.

```
glm(formula = Status ~ loan_type + loan_amount + rate_of_interest +
    term + property_value + income + Credit_Score + age + dtir1,
    family = binomial(link = "logit"), data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.31928  -0.78862  -0.68627   0.00154   2.36731

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.392e+01  4.184e+01   0.333   0.7394
loan_typetype2    5.337e-01  2.030e-02  26.290  < 2e-16 ***
loan_typetype3    1.405e-01  2.539e-02   5.536 3.10e-08 ***
loan_amount      -4.369e-07  9.119e-08  -4.791 1.66e-06 ***
rate_of_interest  1.936e-02  1.644e-02   1.178   0.2388
term             -7.061e-04  1.357e-04  -5.202 1.97e-07 ***
property_value    1.042e-06  5.929e-08  17.569  < 2e-16 ***
income           -1.274e-04  3.486e-06 -36.548  < 2e-16 ***
Credit_Score      5.483e-05  6.164e-05   0.890   0.3737
age<25           -1.458e+01  4.184e+01  -0.349   0.7274
age>74           -1.455e+01  4.184e+01  -0.348   0.7281
age25-34         -1.480e+01  4.184e+01  -0.354   0.7236
age35-44         -1.472e+01  4.184e+01  -0.352   0.7250
age45-54         -1.464e+01  4.184e+01  -0.350   0.7264
age55-64         -1.460e+01  4.184e+01  -0.349   0.7271
age65-74         -1.466e+01  4.184e+01  -0.350   0.7261
dtir1             1.844e-03  8.264e-04   2.231   0.0257 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 121686  on 108091  degrees of freedom
Residual deviance: 117881  on 108075  degrees of freedom
AIC: 117915
```

Figure 9: Initial model (all variables)

The above **model** indicated that rate_of_interest, credit_score and Age were not significant variables (p-value < 0.05)

We re-ran the model only with variables that have significance.

```
 Coefficients:
                      Estimate     Std. Error z value Pr(>|z|)
 (Intercept)      -0.70937761573  0.08236585916  -8.613  < 2e-16 ***
 loan_typetype2    0.56268816317  0.02015887817  27.913  < 2e-16 ***
 loan_typetype3    0.15483693245  0.02514254645   6.158 7.35e-10 ***
 loan_amount      -0.00000069630  0.00000008803  -7.910 2.58e-15 ***
 rate_of_interest  0.04155310026  0.01638267316   2.536   0.0112 *
 term             -0.00090882943  0.00013466971  -6.749 1.49e-11 ***
 property_value    0.00000118995  0.00000005790  20.553  < 2e-16 ***
 income           -0.00012814756  0.00000345978 -37.039  < 2e-16 ***
 dtir1             0.00143572162  0.00082646865   1.737   0.0824 .
```

Figure 10: Model with Significant variables

The non-interaction model equation:

$$\log\left(\frac{\pi_{status}}{1-\pi_{status}}\right) = -0.7093 + 0.5626 * loantype2 + 0.1548 * loantype3 - 0.0000 * LoanAmount + 0.041 * rateOfInterest + 0.0009 * term + 0.0000 * property\ value - 0.000128 * income + 0.0014 * dtir1$$

The non-interaction model interpretations:

From the RStudio model output, we can see that all the p-values corresponding to the factor coefficients are close to 0. This suggests that all the variables in the model are influential to the approval status outcome.

In this baseline logistic model, there are six numerical variables: loan_amount, rate_of_interest, term, property_value, income, dtir1. There is also 1 categorical variable: loan type. The loan type baseline response is type 1.

The coefficient for loan_amount is close to zero but has a negative value, indicating that our probability equation will indicate that when the loan amount increases, the loan approval probability will decrease linearly. Similarly, property value coefficient is close to zero and positive, indicating the approval status will tend to be positive as property value increases, subjected to all other values remaining the same.

The log-linear model is as follows: Y = intercept + βlog(x)

The effect of increasing 1 unit for log(x) is: Log(x) + 1 = log(x) + log(e) = log(ex)

According to the model:

Factors	Multiplicative Effect
Loan Type: Type 2	0.56268
Loan Type: Type 3	0.1548
loan amount	-0.00000
rate_of_interest	0.0415
term	-0.0009
Property value	0.00000
Income (Increase by $1000)	-0.000128
dtir1	0.0014

Table summary of the effect of an increase in each factor when holding other factors fixed

- When the loan type is type2 as opposed to type1, the odds of loan approval increase by $e^{0.555}$ times, or 1.741 times.
- When the loan type is type3 as opposed to type1, the odds of loan approval increase by $e^{0.1685}$ times, or 1.183 times.
- When the income of the borrower increases by 1000 dollars, the odds of loan approval decrease by $e^{1000*(-0.000124)}$ times, or 0.8833 times.

- When dtir1 value increases by the unit of 1, the odds of loan approval increase by $e^{0.0020}$ times, or 1.002 times.
- When the loan amount increases by the unit of 1, the odds of loan approval decrease by $e^{-0.0000}$ times, or one time.
- When the property value increases by the unit of 1, the odds of loan approval increase by $e^{0.0000}$ times, or one time.

2.2 The interaction model

We investigated to see if any two of our explanatory variables have an interaction, i.e., where the interpretation of the effect of one variable depends on the value of another variable and vice versa.

We did ANOVA test between all the explanatory variables to investigate if any of the variables have interaction.

loan_type:loan_amount	2	56.53	108090	117924	5.297e-13 ***
loan_type:rate_of_interest	2	1842.30	108088	116082	< 2.2e-16 ***
loan_type:term	2	59.39	108086	116023	1.272e-13 ***
loan_type:property_value	2	1806.71	108084	114216	< 2.2e-16 ***
loan_type:income	2	208.79	108082	114007	< 2.2e-16 ***
loan_amount:rate_of_interest	1	1076.38	108081	112931	< 2.2e-16 ***
loan_amount:term	1	127.41	108080	112804	< 2.2e-16 ***
loan_amount:property_value	1	1223.09	108079	111580	< 2.2e-16 ***
loan_amount:income	1	601.55	108078	110979	< 2.2e-16 ***
rate_of_interest:term	1	2721.99	108077	108257	< 2.2e-16 ***
rate_of_interest:property_value	1	512.53	108076	107744	< 2.2e-16 ***
rate_of_interest:income	1	65.43	108075	107679	6.034e-16 ***
term:property_value	1	192.67	108074	107486	< 2.2e-16 ***
term:income	1	0.56	108073	107486	0.4526
property_value:income	1	1372.80	108072	106113	< 2.2e-16 ***

Figure 11: ANNOVA test for interaction

Based on the ANOVA test conducted above, most of the interactions can be deemed significant based on their deviances (>10), as well as p-values(<0.05).

For our test we evaluated the interaction between loan type and income.

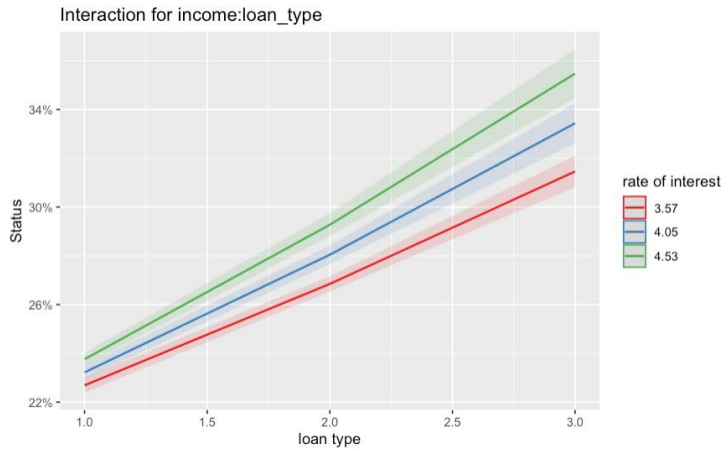


Figure 12: Interaction FIT model test

The above fit model did not show significant interaction between loan type and income.

The interaction model results:

```
Coefficients:
                          Estimate     Std. Error z value        Pr(>|z|)
(Intercept)          -0.69163018429  0.07715116789  -8.965         < 2e-16 ***
loan_typetype2        1.04398042700  0.04124553918  25.311         < 2e-16 ***
loan_typetype3        1.11002513117  0.05804903412  19.122         < 2e-16 ***
loan_amount          -0.00000058138  0.00000008627  -6.739 0.0000000000159 ***
rate_of_interest      0.00530927707  0.01658410188   0.320           0.749
term                 -0.00077371955  0.00013489614  -5.736 0.0000000097129 ***
property_value        0.00000108695  0.00000005810  18.708         < 2e-16 ***
income               -0.00010145180  0.00000353053 -28.736         < 2e-16 ***
loan_typetype2:income -0.00010069035  0.00000777153 -12.956         < 2e-16 ***
loan_typetype3:income -0.00018789547  0.00001059740 -17.730         < 2e-16 ***
```

Figure 13: Interaction model

The interaction model equation:

$$\log\left(\frac{\pi_{status}}{1-\pi_{status}}\right) = -0.6916 + 1.044 * loantype2 + 1.11 * loantype3 \; - \; 0.000 * loan\_amount + \; 0.005 * rate\_of\_interest - 0.0007 * \text{term} + 0.0000 * propertyValue - 0.0001 * loantype2 * income - 0.00018 * loantype3 * income$$

The interaction model interpretation:

- When the loan type is type2 as opposed to type1, the odds of loan approval increase by $e^{1.043}$ times, or 3.736 times.
- When the loan type is type3 as opposed to type1, the odds of a loan approval increase by $e^{1.110}$ times, or 3.706 times.

- When the income of the borrower increases by 1000 dollars, the odds of loan approval decrease by $e^{-0.0001}$ times, or 1 times.
- When the income of the borrower increases by 1000 dollars, and the loan type is fixed at type2, the odds of loan approval decrease by $e^{-0.0001}$ times, or 1 times.
- When the income of the borrower increases by 1000 dollars, and the loan type is fixed at type3, the odds of loan approval decrease by $e^{-0.0001}$ times, or 1 times.
- When the loan amount increases by the unit of 1, the odds of loan approval decrease by $e^{-0.0000}$ times, or 1 times.
- When the property value increases by the unit of 1, the odds of loan approval increases by $e^{0.0000}$ times, or 1 times.

2.3 Likelihood Ratio Test

H_o: The reduced model is appropriate. The coefficient of the interaction term is equal to 0.

H_a: The full model is appropriate. The coefficient of the interaction term is not equal to 0.

```
Likelihood ratio test

Model 1: Status ~ loan_type + loan_amount + rate_of_interest + term +
    property_value + income + dtir1
Model 2: Status ~ loan_type + loan_amount + rate_of_interest + term +
    property_value + income + income:loan_type
  #Df LogLik Df Chisq Pr(>Chisq)
1   9 -58989
2  10 -58764  1 450.8  < 2.2e-16 ***
```

Figure 14: Likelihood ratio test for model comparison

From the test results, we reject the null hypothesis since the p-value <2e-16 is less than the alpha value of 0.05.

At 5% significance, we conclude that sufficient evidence supports that the full model is appropriate.

4 ROC Curve

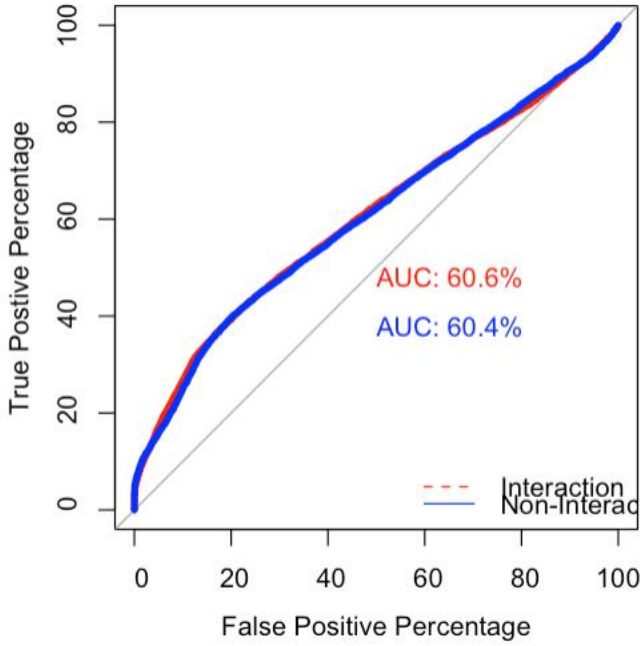


Figure 15L ROC curve

The area under the curve (AUC) for the non-interaction model is 60.4%, while the area under the curve (AUC) for the interaction model is 60.6%. This result agrees with the likelihood ratio test. The ROC curve also suggests that the interaction term contributes to loan approval status prediction. Since there is a very minimal difference between the interaction and non-interaction models, we can use the interaction model to evaluate loan applications.

3 Classification Report

		observed	
	Approval	0	1
predicted	0	20185	6832
	1	1	8

		observed	
	Approval	0	1
predicted	0	20184	6613
	1	2	227

Classification tables of the non-interaction model (left) and the interaction model (right)

Predictive Power Summary				
	Accuracy (%)	Misclassification Rate (%)	Estimated Specificity (%)	Estimated sensitivity (%)

Non-interaction model	74.7	25.3	11	99
Interaction model	75.5	24.5	3	99

Predictive Power Summary table of the non-interaction table and the interaction table

Sensitivity measures the model's power to accurately predict true positives while specificity measures the model's power to accurately predict true negatives.

From the results, the interaction model has an accuracy of 75.5%, which is less than 1% higher than the non-interaction model. The interaction model also has a specificity of 3%, which is about 8% lower than the non-interaction model. Both models have 99% sensitivity. To have a better examination of the predictive power of the two models, ROC curve plots are generated.

5 The Hosmer-Lemshow test

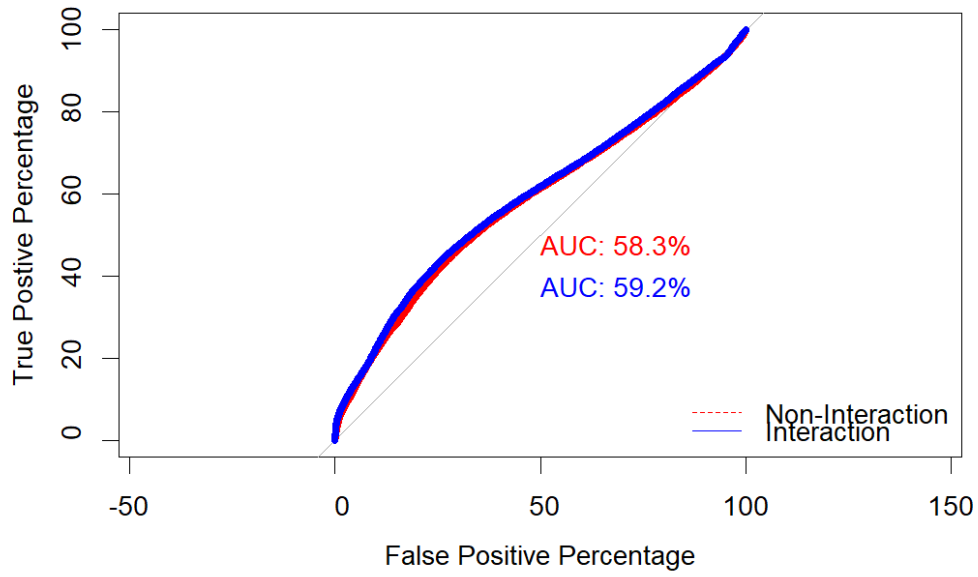
The Hosmer-Lemshow Test by the **generalhoslem** package			
	X-Squared	Degrees of Freedom	p-value
Non-Interaction Model	687.93	8	< 2.2e-16
Interaction Model	447.2	8	< 2.2e-16
The Hosmer-Lemshow Test by the **ResourceSelection** package			
	X-Squared	Degrees of Freedom	p-value
Non-Interaction Model	687.93	8	< 2.2e-16
Interaction Model	447.42	8	< 2.2e-16

Hosmer-Lemshow test results

Both models have p-values less than 0.05, indicating the presence of a lack of fit for both models.

6 Additional Experimentations

There is a presence of lack of fit when testing models built on the train dataset. Could this result be caused by randomness or volatile exploratory variables? What if we run the train model on train data?

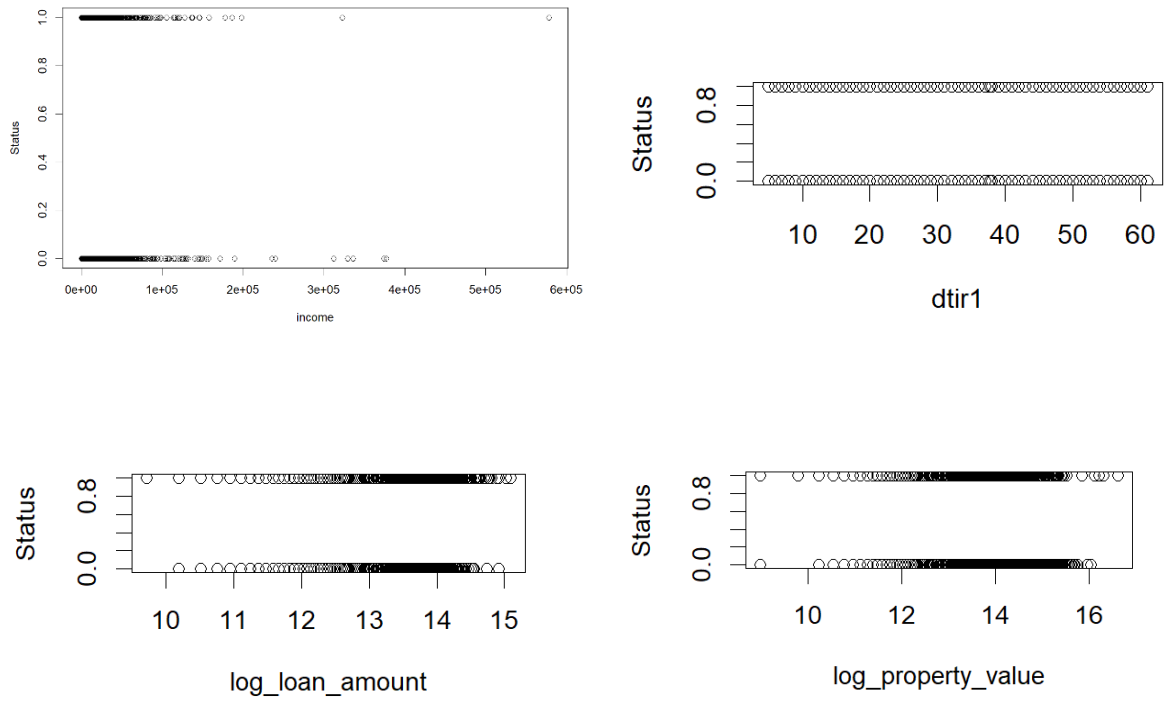


The ROC curve of testing the model on the train dataset

The Hosmer-Lemshow Test by the **generalhoslem** package			
	X-Squared	Degrees of Freedom	p-value
Non-Interaction Model	1637.4	8	< 2.2e-16
Interaction Model	1417.2	8	< 2.2e-16

Hosmer-Lemshow test results for train data

The ROC curve results do not improve when the model is run on the train dataset. There is a lack of fit even when the model results are tested against the train dataset.



Scatterplot of: status and income (top left), status and dtir1 (top right), status and log(loan amount) (bottom left), and status and log(property value) (bottom right)

We re-examined the relationship between Status and the influential variables. As shown in the scatterplots above, the resulting points with Status 1 and Status 0 are rather parallel to each other. Because of this, it is difficult for a logistic model to fit through most of these points.

Conclusion

The aim was to build a model that predicts the loan approval status. A real-life dataset was used for this report. The dataset used for this report contained less than 10% missing values and outlier data. Because of the low numbers of missing values and outliers, a sound data cleaning process could be completed.

For the variable selection method, 9 variables out of 33 variables were selected by domain knowledge. After a stepwise selection process and examining coefficient p-values of the GLM model, 7 influential variables remained. Numerous interactions were observed in this model by conducting the ANOVA test. We selected 2 interactions for examining the effect of interactions. Results show that the interaction terms were significant in predicting loan approval status.

Even though all variables were influential in this model, and the selected variables did pass t-tests and chi-square tests, a lack of fit was present in both the non-interaction model and the interaction model.

Results indicate that interaction terms improve prediction results and reduces the lack of fit, and as such the interaction model is preferred over the non-interaction model in predicting odds of loan approval. We recommend for further analysis a model including all interaction terms should be built and evaluated.