|  |  |
| --- | --- |
| A black umbrella over a piggybank  Loan Approval PREDICTIVE MODEL  DANA 4820 Fall 2002 Project | Abstract  Predict odds of loan approval using logistic regression models  Andrew Liu 100390239​​ Aswinee Rath 100389210​​ Patrick Ipac 100385706 |

|  |  |
| --- | --- |
|  |  |

[Project Background 2](#_Toc121512153)

[Exploratory Data Analysis and Cleanup 2](#_Toc121512154)

[Variable Selection 3](#_Toc121512155)

[Model Evaluation 3](#_Toc121512156)

[1 Methodology 3](#_Toc121512157)

[2 Non-Interaction and Interaction Model Comparison 3](#_Toc121512158)

[2.1 The non-interaction model 3](#_Toc121512159)

[2.2 The interaction model 4](#_Toc121512160)

[2.3 Likelihood Ratio Test 6](#_Toc121512161)

[3 Classification Report 6](#_Toc121512162)

[4 ROC Curve 7](#_Toc121512163)

[5 The Hosmer-Lemshow test 7](#_Toc121512164)

[6 Additional Experimentations 8](#_Toc121512165)

[Conclusion 10](#_Toc121512166)

## 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.

Table

Description automatically generated

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.

Background pattern

Description automatically generated

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.

Diagram, schematic

Description automatically generated

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.

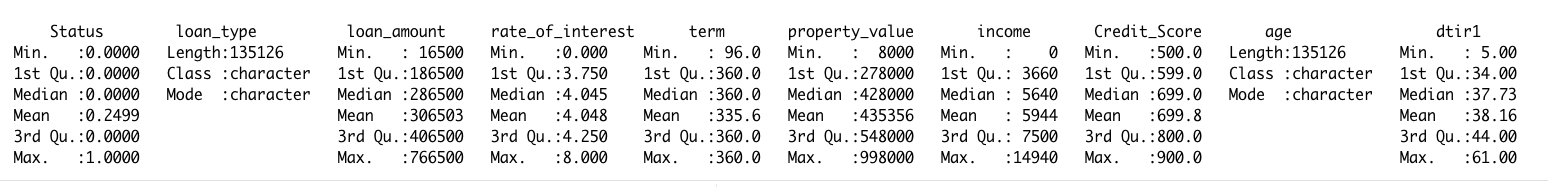


Figure 4: Post-Clean data summary

Our data also looked relatively normalized.

Diagram

Description automatically generated

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.

Text

Description automatically generated

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.

Graphical user interface, text, application

Description automatically generated

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.

Chart, histogram

Description automatically generated

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.

Table

Description automatically generated

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.

Table

Description automatically generated

Figure 10: Model with Significant variables

**The non-interaction model equation:**

**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 e0.555 times, or 1.741 times.
* When the loan type is type3 as opposed to type1, the odds of loan approval increase by e0.1685 times, or 1.183 times.
* When the income of the borrower increases by 1000 dollars, the odds of loan approval decrease by e1000\*(-0.000124) times, or 0.8833 times.
* When dtir1 value increases by the unit of 1, the odds of loan approval increase by e0.0020 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 e0.0000 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.

Table

Description automatically generated

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.

Chart, line chart

Description automatically generated

Figure 12: Interaction FIT model test

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

**The interaction model results:**

Text, table

Description automatically generated

Figure 13: Interaction model

**The interaction model equation:**

**The interaction model interpretation:**

* When the loan type is type2 as opposed to type1, the odds of loan approval increase by e1.043 times, or 3.736 times.
* When the loan type is type3 as opposed to type1, the odds of a loan approval increase by e1.110 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 e0.0000 times, or 1 times.

### 2.3 Likelihood Ratio Test

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

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

Text

Description automatically generated

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

Chart, line chart

Description automatically generated

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?

Chart, line chart

Description automatically generated

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.

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart  Description automatically generated |
| Chart  Description automatically generated with medium confidence | Diagram  Description automatically generated with medium confidence |

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.