STAT 410 PROJECT REPORT

Assessing Loan Status

Model applied

Binary Logistic + Probit + Complementary Log Log Regression Models

Submitted to

Prof. Dr. Olga Korosteleva

Report Prepared

by Chi Nguyen

November 30, 2022

CONTENTS

I. Introduction

Financial decision-making has always intrigued me, particularly when it comes to understanding risks and probabilities. While I am primarily focused on statistical analysis, I find it fascinating to apply these concepts to real-world financial problems. I recently found a dataset that examines loan outcomes, with loan status modeled as a binary response variable. This dataset, with 12 predictors and more than 32,000 observations, presents an excellent opportunity to explore credit risk. By analyzing applicant details and loan characteristics, I aim to uncover the key factors that influence loan decisions and enhance predictive accuracy.

II. Background

Loan status indicates whether borrowers are repaying their loans or have defaulted (not likely to pay back). When borrowers make their payments on time, it strengthens banks and supports the economy by making it easier for others to get loans. However, when borrowers default, it creates problems for lenders and makes it harder and more expensive for future borrowers to get loans. Predicting whether a borrower will repay a loan is challenging because it depends on factors like income, existing debts, credit history, and changes in the economy.

Understanding loan repayment is important because it helps lenders manage risks and make better decisions about who to lend to. Studying loan status data can uncover patterns that show why some loans succeed while others fail. With tools like logistic regression, we can use these patterns to predict which loans are likely to be repaid and which might default, helping lenders create smarter lending strategies.

III. Data description

The dataset, sourced from Kaggle, focuses on analyzing individual borrowers' loan repayment behavior and the risk of default. The main variable of interest is loan status, which indicates whether a borrower has repaid the loan (status = 0) or defaulted (status = 1). This binary response variable is used to model repayment likelihood.

For this analysis, a random sample of 1,000 observations was selected from the original dataset, which contains 32,000 rows. This subset was created using SAS for data cleaning and preprocessing and then exported to R for further analysis. The key variables included in the dataset are:

- ***Loan Status:*** Indicates whether the loan was repaid (0) or defaulted (1).
- ***Loan Amount*** (loan_amnt_scaled): The total loan amount requested, scaled by dividing by 10,000 for easier interpretation.
- ***Income*** (person_income_scaled): The borrower's annual income, scaled by dividing by 10,000.
- ***Loan Interest Rate*** (loan_int_rate): The interest rate applied to the loan, which influences the borrower's ability to repay.
- ***Years of Employment*** (person_emp_length): The number of years the borrower has been employed, representing job stability.
- ***Loan Intent*** (loan_intent): The reason for taking the loan, such as EDUCATION, HOME, or PERSONAL (with EDUCATION set as the reference category).
- ***Home Ownership*** (person_home_ownership): Indicates whether the borrower owns a home, which is linked to financial stability (with OTHER set as the reference category).

The dataset was preprocessed to remove any missing or inconsistent values, making it ready for analysis. This analysis aims to predict loan repayment based on the predictors provided, helping lenders understand the factors influencing loan defaults and repayments

IV. Result

In this analysis, I explored various models to predict the likelihood of loan repayment using a dataset sourced from Kaggle. After comparing multiple models based on criteria such as AIC, AICC, and BIC, the complementary log-log model was found to have the best fit, as it showed the lowest values in all three criteria. This indicated that the complementary log-log model was the most appropriate for predicting loan repayment based on the available data.

A. Significant Predictors

The analysis revealed several key predictors that significantly influence the probability of loan repayment. At the 5% significance level, the following variables were found to be significant:

- Person Income
- Loan Amount
- Loan Interest Rate
- Loan Intent for Debt Consolidation

In the SAS model, Home Ownership specifically with categories 'Mortgage' and 'Owned' is also a significant predictor at the 5% level.

B. The fitted model (using coefficients from R output)

$$
\begin{aligned}
&1 - \widehat{P}(not\ likely\ to\ pay\ back) \\
&= exp(-\ exp((-\ 18.665579)\ +\ (0.002273).Age\ +\ (-\ 0.324959).PersonIncome \\
&+\ (0.92129).LoanAmt\ +\ (0.302909).LoanRate\ +\ (0.001474).EmpLength \\
&+\ (0.893259).LoanIntentConsolidation\ +\ (0.382571).LoanIntentHomeImprove \\
&+\ (0.383549).LoanIntentMedical\ +\ (0.236034).LoanIntentPersonal\ +\ (0.263717).LoanIntentVenture \\
&+\ (13.938772).HomeMortgage\ +\ (12.106410).HomeOwn\ +\ (14.318881).HomeRent))
\end{aligned}
$$

C. Interpretation of the estimated significant regression coefficients

- **Person Income:** The estimated probability of no default in loan status (repay the loan) for each additional 10,000-unit increase in income is the old one raised to the power exp(-0.3250) = 0.7225.
  This indicates that as a borrower's income increases, the probability they are likely to pay back increases (loan_status = 0)

- **Loan Amount:** The estimated probability of no default in loan status (repay the loan) for each additional 10,000-unit increase in loan amount is the old one raised to the power exp(0.92129) = 2.5125.
  This indicates that as a borrower's loan amount increases, the probability of they are likely to pay back decreases (loan_status = 0)

- **Loan Interest Rate:** The estimated probability of no default in loan status (repay the loan) for each additional unit increase in loan interest rate is the old one raised to the power exp(0.302909) = 1.35379.

This indicates that as a borrower's loan interest rate increases, the probability that they are likely to pay back decreases (loan_status = 0)

- **Loan Intent for Debt Consolidation**: (With "EDUCATION" as the reference) The estimated probability of no default in loan status (repay the loan) of loan intent for debt consolidation purposes is that for those loan intent for education purposes raised to the power exp(0.893259) = 2.4431
  This indicates that loans for debt consolidation are likely to repay the loan compared to those for education.

D. Data for Prediction

To assess the model's prediction capabilities, I used a set of sample data for prediction.

- Age: 30 years
- Person Income: 8 (scaled by dividing by 10,000, representing an annual income of $80,000)
- Loan Amount: 5 (scaled by dividing by 10,000, representing a loan amount of $50,000)
- Loan Interest Rate: 11.0529 (the mean interest rate from the dataset)
- Years of Employment Length: 8 years
- Loan Intent: EDUCATION
- Home Ownership: OWN

E. Results and Interpretation

Using the complementary log-log model, the predicted probability of the borrower **n**ot likely to repay the loan (loan_status = 1) was approximately 27.75%. This means that there is a 72.25% chance that the borrower is classified as likely to repay the loan. This probability seems plausible and aligns with the expected behavior of borrowers with these characteristics, indicating the model is functioning as anticipated

F. Sensitivity of the Model

An interesting observation came from testing how changes in home ownership status influenced the model's predictions. When the Home Ownership variable was changed from "OWN" to "RENT," while keeping all other variables the same, the probability that the borrower would not repay the loan increased dramatically to 94.87%. This highlights the model's sensitivity to this variable, showing that home ownership plays a crucial role in determining loan repayment likelihood. Borrowers who rent may face more financial instability, which increases the likelihood of default.

V. Conclusion

Reflecting on the analysis, the process of predicting loan repayment was both interesting and educational. The model performed well, and I was able to predict the probability of a borrower repaying a loan based on realistic characteristics. However, there is room for improvement. The dataset was imbalanced, with more borrowers classified as repaying their loans (loan_status = 0) than defaulting (loan_status = 1). Balancing the data before modeling, perhaps by sampling more default cases, could have made the model more accurate.

It also would have been useful to test the model's predictions on borrowers with a higher risk of default to better evaluate its accuracy. Comparing these results could have provided additional insight into how well the model performs in predicting challenging cases.

Overall, this project was a great opportunity to apply what I've learned to a real-world problem. It gave me valuable experience in working with data and building predictive models, while also showing me areas to focus on improving in the future.

VI. Appendix

A. R code for selecting the fitted model

```r
# Load the data
data <- read.csv("/Users/christinenguyen/Downloads/credit_risk_dataset.csv")

# Remove rows with missing values
data <- na.omit(data)

# Randomly sample 1,000 rows
set.seed(123)
data <- data[sample(nrow(data), 1000), ]

# Convert loan_status to a factor
data$loan_status <- as.factor(ifelse(data$loan_status == 1, 1, 0)) # loan_status: default = not pay back = 1, pay = 0

# Ensure loan_intent and person_home_ownership are factors
data$loan_intent <- as.factor(data$loan_intent)
data$person_home_ownership <- as.factor(data$person_home_ownership)

# Scale large numerical variables for easier interpretation
data$loan_amnt_scaled <- data$loan_amnt / 10000  # Scale loan amount
data$person_income_scaled <- data$person_income / 10000  # Scale income

# Fit the Logistic Regression (Logit) model
logit_model <- glm(loan_status ~ person_age + person_income_scaled + loan_amnt_scaled +
                     loan_int_rate + person_emp_length + loan_intent + person_home_ownership,
                   data = data, family = binomial(link = "logit"))

# Fit the Probit model
probit_model <- glm(loan_status ~ person_age + person_income_scaled + loan_amnt_scaled +
                      loan_int_rate + person_emp_length + loan_intent + person_home_ownership,
                    data = data, family = binomial(link = "probit"))

# Fit the Complementary Log-Log (Cloglog) model
cloglog_model <- glm(loan_status ~ person_age + person_income_scaled + loan_amnt_scaled +
                       loan_int_rate + person_emp_length + loan_intent + person_home_ownership,
                     data = data, family = binomial(link = "cloglog"))
```

```r
# Define calc_aicc function
calc_aicc <- function(model) {
  n <- nrow(model$model)
  k <- length(coef(model))
  aic <- AIC(model)
  aicc <- aic + (2 * k^2 + 2 * k) / (n - k - 1)
  return(aicc)
}

# Collect AIC, AICc, and BIC for all models into a data frame
model_metrics <- data.frame(
  Model = c("Logit", "Probit", "Cloglog"),
  AIC = c(AIC(logit_model), AIC(probit_model), AIC(cloglog_model)),
  AICc = c(calc_aicc(logit_model), calc_aicc(probit_model), calc_aicc(cloglog_model)),
  BIC = c(BIC(logit_model), BIC(probit_model), BIC(cloglog_model))
)

# Print the metrics table
print(model_metrics)
```

Picture 1 & 2: R code for AIC, AICC, and BIC to choose the fitted model

	Model	AIC	AICc	BIC
1	Logit	752.5679	752.9943	821.2765
2	Probit	761.2723	761.6987	829.9809
3	Cloglog	743.9654	744.3918	812.6740

Picture 3: R output, showing that the complementary log-log model has a better fit.

B. SAS code - Fitted model

```
proc import datafile="//vdi-fileshare02/UEMprofiles/028631185/Desktop/credit_risk_dataset.csv"
     out=credit_risk
     dbms=csv
     replace;
     getnames=yes;
run;

/* Remove rows with missing values */
data credit_risk_clean;
     set credit_risk;
     if cmiss(of _all_) then delete;
run;

/* Randomly sample 1,000 rows */
proc surveyselect data=credit_risk_clean out=credit_risk_sample
     method=srs n=1000 seed=123;
run;

data credit_risk_sample;
     set credit_risk_sample;
     loan_amnt_scaled = loan_amnt / 10000; /* Scale loan amount */
     person_income_scaled = person_income / 10000; /* Scale income */
run;

proc genmod data=credit_risk_sample;
     class loan_intent (ref="EDUCATION")  /* Set EDUCATION as reference */
           person_home_ownership (ref="OTHER") / param=ref; /* Set OTHER as reference */
     model loan_status(event='1') = person_age person_income_scaled loan_amnt_scaled
                                    loan_int_rate person_emp_length
                                    loan_intent person_home_ownership
                                    / dist=binomial link=cloglog;
run;
```

Picture 4: SAS code for the fitted model

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-26.5307	0.5296	-27.5687	-25.4928	2509.80	<.0001
person_age		1	0.0023	0.0130	-0.0231	0.0277	0.03	0.8607
person_income_scaled		1	-0.3250	0.0403	-0.4039	-0.2460	65.10	<.0001
loan_amnt_scaled		1	0.9214	0.1296	0.6675	1.1754	50.57	<.0001
loan_int_rate		1	0.3029	0.0273	0.2494	0.3564	123.03	<.0001
person_emp_length		1	0.0015	0.0196	-0.0370	0.0399	0.01	0.9401
loan_intent	DEBTCONSOLIDATION	1	0.8933	0.2508	0.4016	1.3849	12.68	0.0004
loan_intent	HOMEIMPROVEMENT	1	0.3826	0.2729	-0.1524	0.9175	1.96	0.1610
loan_intent	MEDICAL	1	0.3835	0.2440	-0.0947	0.8618	2.47	0.1160
loan_intent	PERSONAL	1	0.2360	0.2520	-0.2579	0.7299	0.88	0.3489
loan_intent	VENTURE	1	0.2637	0.2603	-0.2466	0.7740	1.03	0.3111
person_home_ownershi	MORTGAGE	1	21.8039	0.1691	21.4726	22.1353	16634.2	<.0001
person_home_ownershi	OWN	1	19.9716	0.5112	18.9696	20.9736	1526.08	<.0001
person_home_ownershi	RENT	0	22.1840	0.0000	22.1840	22.1840	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Picture 5: SAS output

C. R code - Fitted model

```r
# Load the data
data <- read.csv("/Users/christinenguyen/Downloads/credit_risk_data_modify.csv")

# Remove rows with missing values
data <- na.omit(data)


# Convert loan_status to a factor
data$loan_status <- as.factor(ifelse(data$loan_status == 1, 1, 0)) # loan_status: default = not pay back = 1, pay = 0

# Ensure loan_intent and person_home_ownership are factors
data$loan_intent <- as.factor(data$loan_intent)
data$person_home_ownership <- as.factor(data$person_home_ownership)

# Set reference levels for factor variables
data$loan_intent <- relevel(data$loan_intent, ref = "EDUCATION")  # Set "EDUCATION" as reference for loan_intent
data$person_home_ownership <- relevel(data$person_home_ownership, ref = "OTHER")  # Set "OWN" as reference for person_home_ownership

# Scale large numerical variables for easier interpretation
data$loan_amnt_scaled <- data$loan_amnt / 10000  # Scale loan amount
data$person_income_scaled <- data$person_income / 10000  # Scale income

# Fit the Complementary Log-Log (Cloglog) model
cloglog_model <- glm(loan_status ~ person_age + person_income_scaled + loan_amnt_scaled +
                       loan_int_rate + person_emp_length + loan_intent + person_home_ownership,
                     data = data, family = binomial(link = "cloglog"))

# Display model summary
summary(cloglog_model)
```

Picture 6: R code for the fitted model

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -18.665579 662.561209  -0.028 0.977525
person_age                    0.002273   0.012420   0.183 0.854792
person_income_scaled         -0.324959   0.038916  -8.350  < 2e-16 ***
loan_amnt_scaled              0.921429   0.127063   7.252 4.11e-13 ***
loan_int_rate                 0.302909   0.026598  11.388  < 2e-16 ***
person_emp_length             0.001474   0.019338   0.076 0.939261
loan_intentDEBTCONSOLIDATION  0.893259   0.247952   3.603 0.000315 ***
loan_intentHOMEIMPROVEMENT    0.382571   0.274380   1.394 0.163224
loan_intentMEDICAL            0.383549   0.240074   1.598 0.110125
loan_intentPERSONAL           0.236034   0.255024   0.926 0.354687
loan_intentVENTURE            0.263717   0.258728   1.019 0.308069
person_home_ownershipMORTGAGE 13.938772 662.561013   0.021 0.983216
person_home_ownershipOWN     12.106410 662.561191   0.018 0.985422
person_home_ownershipRENT    14.318881 662.561007   0.022 0.982758
---
```

Picture 7: R output

D. Computed by hand - Prediction

$$P^0(\textit{not likely to pay back})$$

$$= 1 - exp(- exp((- 18.665579) + (0.002273).(30) + (- 0.324959).(8) + (0.92129).5$$

$$+ (0.302909).(11.0529) + (0.001474).(8) + (12.106410)))$$

$$= 0.277366 \approx 27.7366\%$$

E. SAS code - Prediction

```
/* Create new data for prediction */
data new_data;
    person_age = 30;
    person_income_scaled = 8;  /* Income 80k */
    loan_amnt_scaled = 5;      /* Loan amount 50k */
    loan_int_rate = 11.0529;     /*From R
    person_emp_length = 8;     /* Employment length 8 years */
    loan_intent = "EDUCATION"; /* EDUCATION reference level */
    person_home_ownership = "OWN"; /* OWN for prediction */
    drop _TYPE_ _FREQ_;
run;

/* Combine original and new data for prediction */
data combined_data;
    set credit_risk_sample new_data;
run;

/* Predict probabilities using the fitted model */
proc genmod data=combined_data;
    class loan_intent (ref="EDUCATION")  /* Set EDUCATION as reference */
          person_home_ownership (ref="OTHER") / param=ref; /* Set OTHER as reference */
    model loan_status(event='1') = person_age person_income_scaled loan_amnt_scaled
                                   loan_int_rate person_emp_length
                                   loan_intent person_home_ownership
                                   / dist=binomial link=cloglog;
    output out=predicted_out p=predicted_prob;
run;

proc print data=predicted_out (firstobs = 1001) noobs;
    var predicted_prob;
run;
```

Picture 8: SAS code for prediction

predicted_prob
0.27753

Picture 9: SAS output for prediction

F. R code - Prediction

```
# Calculate the mean of scaled variables
loan_int_rate_mean <- mean(data$loan_int_rate, na.rm = TRUE)
print(loan_int_rate_mean)

# Create new data for prediction
new_data <- data.frame(
  person_age = 30,                  # Age of the loan applicant (30 years old)
  person_income_scaled = 8,         # Scaled income of the applicant (80k)
  loan_amnt_scaled = 5,             # Scaled loan amount requested (50k)
  loan_int_rate = loan_int_rate_mean, # Average interest rate for the loan (calculated from data)
  person_emp_length = 8,            # Length of employment of the applicant (8 years)
  loan_intent = factor("EDUCATION", levels = levels(data$loan_intent)), # loan intent
  person_home_ownership = factor("OWN", levels = levels(data$person_home_ownership)) #home ownership
)

# Predict loan status probability using the Complementary Log-Log model
predicted_risk_cloglog <- predict(cloglog_model,
                                  newdata = new_data,
                                  type = "response")
print(predicted_risk_cloglog)
```

Picture 10: R code for prediction

```
> print(predicted_risk_cloglog)
        1
0.2775282
```

Picture 11: R output for prediction

G. R code testing the Sensitivity of the model

```
# Create new data for prediction
new_data <- data.frame(
  person_age = 30,                      # Age of the loan applicant (30 years old)
  person_income_scaled = 8,             # Scaled income of the applicant (80k)
  loan_amnt_scaled = 5,                 # Scaled loan amount requested (50k)
  loan_int_rate = loan_int_rate_mean,   # Average interest rate for the loan (calculated from data)
  person_emp_length = 8,                # Length of employment of the applicant (8 years)
  loan_intent = factor("EDUCATION", levels = levels(data$loan_intent)), # loan intent
  person_home_ownership = factor("RENT", levels = levels(data$person_home_ownership)) #home ownership
)

# Predict loan status probability using the Complementary Log-Log model
predicted_risk_cloglog <- predict(cloglog_model,
                                  newdata = new_data,
                                  type = "response")
print(predicted_risk_cloglog)
```

Picture 12: R code change Home Ownership variable to 'RENT', keep others variables fixed

```
        1
0.9487295
```

Picture 13: R output with prediction Home Ownership as 'RENT'

VII. Reference

Lao Tse. *Credit Risk Dataset*. Kaggle, https://www.kaggle.com/datasets/laotse/credit-risk-dataset. Accessed 2 Dec. 2024.