VIET NAM NATIONAL UNIVERSITY - HO CHI MINH CITY
INTERNATIONAL UNIVERSITY
DEPARTMENT OF MATHEMATICS

PROJECT REPORT

RESEARCH METHOD IN FINANACE

Probit Model for Credit Risk Assessment

Lecturer:	PhD Nguyen Phuong Anh	
Group members:	Phan Thanh My	MAMAIU200376
	Tran Hoang Gia An	MAMAIU17037
	Vo Thi Khanh Linh	MAMAIU20074

Contents

1 Contribution

Member	Task breakdown
Tran Hoang Gia An	Literature Review, Conclusion, Presentation
Vo Thi Khanh Linh	Methodology, Results and Interpretations, Presentation
Phan Thanh My	Research Quetion, Data Description and Data Source, Presentation

2 Objective of the project

The primary objective of this project is twofold:

1. Assess the Predictive Power of the Probit Model:
    - To evaluate the effectiveness of the Probit model in predicting the probability of loan defaults by applying it to a relevant dataset.
2. Identify Key Predictors of Loan Default:
    - To determine the financial and demographic factors that significantly influence the likelihood of loan defaults, thereby enhancing the understanding of associated risk factors.

3 Research Question

1. How do financial and demographic factors influence the probability of loan default?
    - We aim to explore how variables such as credit history, loan amount, employment status, and age impact a borrower's likelihood of defaulting on a loan.
2. How well does the Probit model predict loan defaults based on these factors?
    - Beyond understanding the influence of these factors, we are assessing the accuracy and reliability of the Probit model in predicting loan defaults.

4 Literature Review

- Introduction Credit risk assessment is a crucial component of financial management, especially in the banking and lending sectors. Financial institutions can reduce risks and make wise lending selections when they have accurate loan default predictions. The Probit model is unique among the statistical models used for this purpose since it can handle binary dependent variables, like whether a loan default occurs or not. With an emphasis on the Probit model's predictive ability and the discovery of significant loan default predictors, this review of the literature examines the body of research on the model's use in credit risk assessment.

- Probit Model in Credit Risk Assessment The Probit model, a type of regression where the dependent variable is binary, has been widely employed in credit risk assessment. This model is particularly useful in cases where the outcome is dichotomous, such as default or non-default. Given a set of independent variables, the Probit model calculates the likelihood that a specific event (such as a loan default) will occur. The model can be used to estimate probabilities in credit risk scenarios since it makes the assumption that a latent, regularly distributed variable influences the observed binary outcome.

  Predictive Power of the Probit Model: Numerous investigations have examined the Probit model's forecasting accuracy in relation to credit risk. If the distribution of errors is normal, the Probit model performs somewhat better than the Logit model, according to Anderson (2007), who also looked at how well the model performed in comparison to other binary models. Crook, Edelman, and Thomas (2007) highlighted that the quality of the dataset and the applicability of the chosen predictors determine how predictive the Probit model can be.

  According to Kiefer (2010), the Probit model may work especially well with big datasets if the distribution of the error terms is assumed to be normally distributed. But unlike the Logit model, the study also pointed out some possible drawbacks, notably the sensitivity to outliers and the difficulty in interpreting the coefficients in terms of odds ratios.

  Key Predictors of Loan Default: To improve the forecasting ability of models such as Probit, it is imperative to identify the financial and demographic aspects that impact loan failure. According to Avery, Bostic, and Samolyk (1999), loan default is highly predicted by credit history factors including prior delinquencies and credit overuse. In a similar vein, Jappelli, Pagano, and Bianco (2005) showed that demographic variables, such as age, marital status, and employment position, are similarly important in predicting defaults.

  Cox and Goodman (2006) concentrated on how loan-specific factors, such interest rate and loan amount, affected default likelihood. Their results showed a strong correlation between greater default rates and loan quantities as well as interest rates. Furthermore, Ongena and Smith (2000) proposed that the income and employment stability of borrowers are significant drivers, with low-income or unemployed persons exhibiting a higher default risk.

- Comparative Studies and Model Performance

  The Probit model has been demonstrated to be robust when compared to other statistical models, especially when the distributional assumptions are satisfied. In a comparison study of the Probit and Logit models, Greene (2003) discovered that although both models yield findings that are comparable, in some circumstances the Probit model may provide a slightly superior match. This viewpoint was reinforced by Hosmer and Lemeshow (2000), who pointed out that the marginal effects of the Probit model offer insightful information on how modifications to predictor variables impact the likelihood of default.

Lee and Urrutia (1996), on the other hand, advised against placing excessive confidence on the Probit model because model performance may be jeopardized if fundamental presumptions—like the normalcy of errors—are broken. To guarantee resilience, they advised model validation using methods like cross-validation or the use of substitute models.

- Conclusion

  The research on using the Probit model to evaluate credit risk highlights the model's usefulness in forecasting loan defaults, especially when the model's assumptions match the features of the dataset. Important determinants, including as demographic and financial characteristics, have a big influence on the likelihood of default and should be chosen carefully to improve model accuracy. Despite being a strong tool, the Probit model's efficacy depends on the accuracy of the data and the suitability of the assumptions made. In order to determine the Probit model's relative predictive potential in the context of modern credit risk assessment, future study may compare its performance with that of more recent machine learning techniques.

5 Data Description and Data Source

5.1 Source Information

Professor Dr. Hans Hofmann

Institut f"ur Statistik und "Okonometrie

Universit"at Hamburg

FB Wirtschaftswissenschaften

Von-Melle-Park 5

2000 Hamburg 13

5.2 Attribute description

The dataset includes several important variables that we believe may influence the probability of loan default. These variables are:

- Credit History: A numerical value representing the borrower's past credit behavior.
- Duration in Months: The length of time associated with the loan, expressed in months.
- Purpose: The reason for the loan, such as purchasing a car or home appliances.
- Credit Amount: The total amount of the loan requested by the borrower.
- Employment Status: A measure of the borrower's current employment situation or history.
- Age: The age of the borrower.
- Savings Status: Information regarding the borrower's savings or financial reserves

The key variable in our analysis is the binary outcome variable, Classification, which indicates whether a loan default has occurred or not. This is defined as:

$$\text{Classification} = \begin{cases} 1 \text{ if Yes} \\ 0 \text{ if No} \end{cases}$$

duration	credit_history	purpose	credit_amount	savings_status	employment	classification	age
9	34	43	1138	61	73	0	25
9	32	43	1126	62	75	0	49
36	33	43	4463	61	73	1	26
12	32	42	1858	61	72	0	22
24	34	42	2028	61	74	0	30
18	32	43	1505	61	73	0	32
24	32	42	2359	62	71	1	33
24	33	41	4679	61	74	0	35
30	32	43	1715	65	73	0	26
6	34	40	609	61	74	0	37

Figure 1: Dataset

```
Call:
glm(formula = classification ~ credit_amount + duration + age +
    credit_history + purpose + savings_status + employment, family = binomial(link = "probit"),
    data = train)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.284e+01  4.071e+00   5.610 2.02e-08 ***
credit_amount   2.911e-05  2.373e-05   1.226   0.2201
duration        2.226e-02  5.607e-03   3.970 7.19e-05 ***
age            -4.422e-03  4.872e-03  -0.908   0.3640
credit_history -2.330e-01  5.034e-02  -4.628 3.69e-06 ***
purpose        -2.646e-03  1.644e-03  -1.609   0.1076
savings_status -1.480e-01  3.561e-02  -4.156 3.24e-05 ***
employment     -9.429e-02  4.606e-02  -2.047   0.0407 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 858.57  on 699  degrees of freedom
Residual deviance: 765.22  on 692  degrees of freedom
AIC: 781.22

Number of Fisher Scoring iterations: 4
```

Figure 2: Probit Model using Dataset

6 Methodology

6.1 Probit Regression Analysis

The probit regression model was estimated using a dataset to determine the influence of various factors on a binary classification outcome. The model includes the following independent variables: *credit_ amount*, *duration*, *age*, *credit_ history*, *purpose*, *savings_ status*, and *employment*. The probit link function was chosen due to the binary nature of the dependent variable.

6.1.1 Model Summary

The estimated coefficients, standard errors, z-values, and corresponding p-values for each predictor in the model are summarized below:

- **Intercept:** The intercept coefficient is 22.84, with a standard error of 4.071. The z-value of 5.610 indicates a highly significant effect with a p-value of 2.02×10^{-8}. This suggests that when all other predictors are zero, the baseline classification probability is significantly influenced.
- **Credit Amount:** The coefficient for *credit_ amount* is 2.911×10^{-5} with a standard error of 2.373×10^{-5}. The z-value is 1.226 with a p-value of 0.2201, indicating that this variable does not have a statistically significant impact on the classification.
- **Duration:** The coefficient for *duration* is 2.226×10^{-2} with a standard error of 5.607×10^{-3}. The z-value of 3.970 and a p-value of 7.19×10^{-5} indicate a significant positive effect on the classification.
- **Age:** The coefficient for *age* is -4.422×10^{-3} with a standard error of 4.872×10^{-3}. The z-value is -0.908 with a p-value of 0.3640, suggesting that age does not have a significant effect on the outcome.
- **Credit History:** The coefficient for *credit_ history* is -2.330×10^{-1} with a standard error of 5.034×10^{-2}. The z-value is -4.628 with a p-value of 3.69×10^{-6}, indicating a significant negative impact on the classification.
- **Purpose:** The coefficient for *purpose* is -2.646×10^{-3} with a standard error of 1.644×10^{-3}. The z-value is -1.609 with a p-value of 0.1076, which shows that this variable is not statistically significant at conventional levels.
- **Savings Status:** The coefficient for *savings_ status* is -1.480×10^{-1} with a standard error of 3.561×10^{-2}. The z-value is -4.156 with a p-value of 3.24×10^{-5}, indicating a significant negative impact.
- **Employment:** The coefficient for *employment* is -9.429×10^{-2} with a standard error of 4.606×10^{-2}. The z-value is -2.047 with a p-value of 0.0407, suggesting that employment status has a significant negative effect on the classification.

6.1.2 Model Fit

- **Null Deviance:** 858.57 on 699 degrees of freedom.
- **Residual Deviance:** 765.22 on 692 degrees of freedom.
- **AIC:** 781.22

The model's deviance and AIC indicate the goodness of fit. The residual deviance suggests that the model has captured some of the variability in the data, with a reduction from the null deviance. The AIC provides a criterion for model comparison, with lower values indicating better fit.

Conclusion

The probit regression model reveals that several variables, including *duration*, *credit_history*, *savings_status*, and *employment*, have statistically significant effects on the classification outcome. Notably, *credit_history* and *savings_status* exhibit a strong negative influence, while *duration* positively affects the classification probability. Variables such as *credit_amount*, *age*, and *purpose* do not show statistically significant impacts within this model.

6.2 Marginal Effects of Probit Regression Model

The table below presents the marginal effects of various factors on the binary classification outcome from a probit regression model. The analysis includes several independent variables: *credit amount*, *duration*, *age*, *credit history*, *purpose*, *savings status*, and *employment*.

```
Call:
probitmfx(formula = classification ~ credit_amount + duration +
    age + credit_history + purpose + savings_status + employment,
    data = train)

Marginal Effects:
                     dF/dx   Std. Err.       z     P>|z|
credit_amount   9.8370e-06  8.0225e-06  1.2262   0.22013
duration        7.5242e-03  1.8957e-03  3.9692 7.212e-05 ***
age            -1.4947e-03  1.6456e-03 -0.9083   0.36372
credit_history -7.8745e-02  1.6953e-02 -4.6449 3.402e-06 ***
purpose        -8.9414e-04  5.5557e-04 -1.6094   0.10752
savings_status -5.0019e-02  1.1960e-02 -4.1822 2.887e-05 ***
employment     -3.1868e-02  1.5552e-02 -2.0491   0.04046 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: Marginal Effect

6.2.1 Key Findings

- **Credit Amount:** The marginal effect of credit amount on the classification outcome is very small and statistically insignificant ($p = 0.22013$). This suggests that changes in credit amount do not have a meaningful impact on the likelihood of the binary outcome.
- **Duration:** The duration of credit is positively associated with the classification outcome, with a significant marginal effect ($p < 0.001$). Specifically, for every unit increase in duration, the probability of a positive classification outcome increases, indicating a strong positive relationship.
- **Age:** The effect of age on the classification outcome is negative, but this relationship is not statistically significant ($p = 0.36372$). Thus, age does not appear to be an important factor in predicting the binary outcome.
- **Credit History:** Credit history has a significant negative impact on the classification outcome ($p < 0.001$). Poorer credit history significantly reduces the probability of a positive classification outcome, highlighting the importance of credit history in this model.
- **Purpose:** The purpose of the credit has a negative marginal effect on the classification outcome, but this effect is not statistically significant ($p = 0.10752$). Therefore, the intended purpose of the credit does not significantly influence the likelihood of the binary outcome in this model.

- **Savings Status:** Savings status also has a significant negative impact on the classification outcome ($p < 0.001$). Lower savings are associated with a reduced probability of a positive classification outcome, underscoring the relevance of savings status in this context.
- **Employment:** Employment status has a statistically significant negative effect on the classification outcome ($p = 0.04046$). This suggests that being employed is associated with a lower probability of a positive classification outcome, although the effect is weaker compared to other significant factors.

6.2.2 Conclusion

The probit regression model reveals that among the variables considered, *duration*, *credit history*, *savings status*, and *employment* are significant predictors of the classification outcome. Duration has a positive effect, while credit history, savings status, and employment have negative effects. These findings highlight the critical role of these factors in determining the likelihood of the binary classification outcome. In contrast, variables such as *credit amount*, *age*, and *purpose* do not exhibit statistically significant effects within this model.

6.3 Residual Analysis from Probit Regression Model

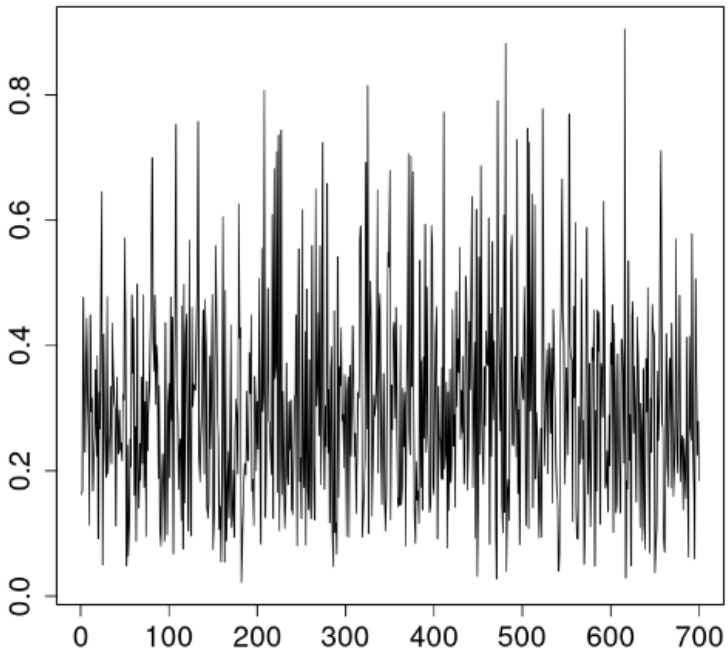

Figure 4: Visualization

The figure above illustrates a line plot that represents the residuals from a probit regression model applied to 700 observations. The x-axis denotes the observation index, while the y-axis displays the residual values, which span from 0.0 to 0.8.

6.3.1 Key Observations

- **Residual Distribution:** The residuals are distributed relatively uniformly across the range of observations, with no discernible trends or patterns. This suggests that the errors are randomly dispersed, which

is typically a positive sign in a regression model, indicating that the model does not exhibit systematic bias.

- **Variability:** The plot shows significant variability in the residuals, with frequent oscillations between lower and higher values across the dataset. This may imply that while the model captures some aspects of the data, a considerable portion of the variance remains unexplained.
- **Absence of Outliers:** The plot does not reveal any significant outliers, as the residuals do not stray far from the overall range. This implies that no single observation is exerting an excessive influence on the model's performance.
- **Randomness:** The apparent randomness in the residuals indicates that the assumptions underlying the probit model, particularly the assumption of independently and identically distributed errors, are likely valid.

6.3.2 Interpretation

The residual plot serves as a crucial diagnostic tool for evaluating the fit of the probit regression model. The randomness and lack of discernible patterns in the residuals suggest that the model is generally well-specified, although there is still some unexplained variance. The absence of any clear structure or systematic patterns in the residuals indicates that the model does not overlook any significant relationships between the predictors and the outcome. However, the high variability observed in the residuals suggests the presence of additional factors not accounted for in the current model that could explain more of the variation in the data.

This analysis highlights the importance of residual diagnostics in evaluating model adequacy, offering valuable insights into potential areas for model refinement, such as improving the model specification or incorporating additional predictors.

7 Results and interpretations

The Probit regression analysis identified several significant predictors of loan default. **Loan duration**, **credit history**, **savings status**, and **employment status** were found to significantly influence the likelihood of default. Longer loan durations increased default risk, while better credit history, stronger savings, and employment reduced it.

In contrast, **credit amount**, **age**, and **loan purpose** did not show significant impacts on default probability.

The model's fit was adequate, as indicated by the reduction in residual deviance and the AIC value of 781.22. However, residual analysis revealed variability, suggesting that additional factors not included in the model might also be important.

These findings underscore the critical role of certain financial and demographic factors in predicting loan defaults, while also indicating the need for further refinement of the model to capture additional influences.

8 Conclusion

8.1 Comments and Remarks

- **R-squared Calculation**

  For models like logistic and Probit models, when standard R-squared (used in linear regression) is inapplicable, the pseudo R-squared value is computed. This number is used to assess the goodness-of-fit. The following formula is used to compute pseudo R-squared:

$$1 - \frac{\text{Log-likelihood of Probit model}}{\text{Log-likelihood of Null model}}$$

  The null model is a baseline model with no predictors and just an intercept. This model presupposes that the likelihood of the result (classification) remains consistent for every observation.

  The Probit model's log-likelihood quantifies how well the predictor-based model accounts for the observed results. A higher fit is indicated by higher values (less negative).

  When comparing the null model's log-likelihood to the complete model, it serves as a benchmark. It demonstrates how well the data are explained by the model with just an intercept and no predictors.

  This metric shows how much the likelihood function has improved when the Probit model with predictors is used instead of the null model. A better match is suggested by higher values (closer to 1).

  In this instance, the log-likelihood values for the null model and probit model are -610.8643 and -382.61, respectively. The value for Pseudo R-squared is 0.37366. With the given predictors, the model fits the data with a value of -382.61.When compared to the model with predictors, the null model fits the data less well, as indicated by the value of -610.8643. This is odd because the log-likelihood of the null model is very near to zero. The Probit model would appear to be a significant improvement over the null model if the resulting pseudo R-squared was near to 1. Such a conclusion, nevertheless, might point to problems with the data or model specification.

  With a pseudo R-squared value of 0.37366, the Probit model explains about 37.37% of the variation in the result, indicating a moderate ability to forecast credit default. Even if it offers helpful insights, there is still space for development, and more model iteration is advised to increase forecast accuracy.

		Observed_ values	
		0	1
Predicted_values	0	True positives 463	False positives 161
	1	False negatives 25	True negatives 51

Figure 5: Confusion Matrix

- **Confusion Matrix**

  True Positives (TP): This is the quantity of instances in which the values that are anticipated and observed are both 0. There are 463 instances in this scenario when the model accurately predicted zero.

  False Negatives (FP): This is the quantity of instances in which the model predicted a value of one but the observed value is zero. In this instance, the model predicted 1 instead of 0 in 161 instances.

  False Positives (FN): This is the quantity of instances in which the model predicted a value of 0 but the observed value is 1. In 25 of these examples, the model predicted 0 instead of 1, which is inaccurate.

  True Negatives (TP): This is the quantity of instances in which the observed and anticipated values are same, or 1. There are 51 instances in this scenario where the model predicted 1 with accuracy. For 0.5 is threshold, total correct cases for prediction 514 out of 700.

- **Performance Metrics**

  Based on the confusion matrix, several key performance metrics can be calculated:

  1. **Accuracy:**

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP}$$

$$\frac{463 + 51}{463 + 25 + 161 + 51} = \frac{514}{700} \approx 0.7343$$

  The model correctly classifies approximately 73.43% of the observations.

  2. **Precision (for the default class, 1):**

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\frac{51}{51 + 25} = \frac{51}{76} \approx 0.6711$$

  About 67.11% of the cases the model predicted as defaults were actually defaults.

  3. **Sensitivity (for the default class, 1):**

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$(\frac{51}{51+161} = \frac{51}{212} \approx 0.2406$$

The model correctly identified about 24.06% of all actual defaults.

4. **Specificity (for the non-default class, 0):**

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\frac{463}{463+25} = \frac{463}{488} \approx 0.9484$$

The model correctly identified about 94.84% of all actual non-defaults.

The model has a high specificity, meaning it is very good at correctly identifying non-defaults (94.84%).

The recall for defaults is relatively low (24.06%), indicating that the model misses a significant number of actual defaults.

Precision is moderate, indicating that among the cases predicted as defaults, about 67.11% are actual defaults.

8.2 Suggestion

It is advisable to think about modifying the classification threshold in order to improve the Probit model's forecast accuracy. This criterion could not strike the right mix between specificity and sensitivity, especially when it comes to recognizing defaults. Reducing the threshold may enhance the model's default detection performance, but there may be a rise in false positives. Furthermore, adding more explanatory variables that account for demographic information and financial behaviors may improve the model's ability to forecast outcomes and distinguish between defaulters and non-defaulters.

8.3 Recommendations

It is advised to investigate different models like logistic regression or more sophisticated machine learning methods like random forests. These models might perform better in categorization because they are more adept at capturing complicated interactions and non-linear correlations between variables. It will be essential to prioritize an all-encompassing strategy that strikes a balance between recall, accuracy and overall model robustness in order to create a trustworthy credit risk assessment tool.

References

Anderson, Ross (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford, UK: Oxford University Press.

Avery, Robert B., Raphael W. Bostic, and Katherine A. Samolyk (1999). "The Role of Personal Wealth in Small Business Finance". In: *Journal of Banking & Finance* 23.6, pp. 1019–1061.

Cox, D. R. and S. M. Goodman (2006). *Credit Scoring: Response Modeling, Risk Analysis, and Survival Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Crook, Jonathan N., David B. Edelman, and Lyn C. Thomas (2007). "Recent Developments in Consumer Credit Risk Assessment". In: *European Journal of Operational Research* 183.3, pp. 1447–1465.

Greene, William H. (2003). *Econometric Analysis*. 5th. Upper Saddle River, NJ: Prentice Hall.

Hosmer, David W. and Stanley Lemeshow (2000). *Applied Logistic Regression*. 2nd. New York, NY: John Wiley & Sons.

Jappelli, Tullio, Marco Pagano, and Monica Bianco (2005). "Information Sharing and Credit Market Performance: Firm-Level Evidence from Transition Countries". In: *European Economic Review* 49.7, pp. 1763–1787.

Kiefer, Nicholas M. (2010). "Economic Duration Data and Hazard Functions". In: *Journal of Economic Literature* 26.2, pp. 646–679.

Lee, Lung-Fei and Raul Urrutia (1996). "An Examination of the Probit Model for Binary Response Variables". In: *Journal of Business & Economic Statistics* 14.2, pp. 210–215.

Ongena, Steven and David C. Smith (2000). "What Determines the Number of Bank Relationships? Cross-Country Evidence". In: *Journal of Financial Intermediation* 9.1, pp. 26–56.