Credit Risk Modelling.

by
HARIOM SARSWAT

FINANCIAL MANAGEMENT
2023

Introduction

This paper aims to delve into the development of a credit risk model using a data-driven approach, with a primary focus on logistic regression as the statistical method. The paper underscores the significance of credit risk modelling and its relevance in the context of financial institutions. By understanding the factors influencing credit risk, financial institutions can make well-informed lending decisions, contributing to their sustainability and competitiveness in the digital age.

What is Credit Risk?
Credit risk is the risk that a borrower will default on a loan, which can lead to financial losses for the lender. In recent years, the global financial system has experienced a number of major credit events, such as the subprime mortgage crisis of 2007-2008. These events have highlighted the importance of effective credit risk management for banks and other financial institutions.

To achieve the objectives set forth, the paper will follow a systematic process. It will begin with data preparation, where historical data on borrowers, including defaulted and non-defaulted loans, will be gathered. Subsequently, data pre-processing will be conducted, encompassing data cleaning, handling missing values, and converting categorical variables into numerical representations.
The next crucial step involves variable selection, where a set of independent variables likely to be correlated with the probability of default will be chosen. These variables, encompassing personal information, credit history, and loan characteristics, will form the basis of the logistic regression model.
Once the model is estimated, it will be subjected to rigorous validation using various metrics like area under the ROC curve (AUC-ROC), accuracy, precision, and recall. These measures will gauge the model's ability to distinguish between defaulters and non-defaulters and assess its overall predictive performance.
The dataset required for model estimation will encompass three main data sources: demographic data, existing relationship data, and credit bureau variables. This comprehensive dataset will empower lenders to construct robust probability of default (PD) models that contribute to prudent lending decisions.
By addressing the outlined objectives, this term paper endeavors to shed light on the importance of credit risk modelling, the application of logistic regression, and the insights gained from predicting credit risk. As financial institutions strive for resilience and stability, the knowledge derived from this paper can foster informed decision-making, fortifying the industry against potential credit pitfalls and bolstering its position in the global financial landscape
In this paper, we will use logistic regression to develop a credit risk model that predicts the probability of default. We will use a dataset of historical loan data on customers' default payments in Taiwan during April-September 2005, to train the model, and we will evaluate the model's performance using a holdout sample. We will also discuss the insights that we gain from the model, and we will explore the implications of the model for credit risk management.

2. Importance of Credit Risk Modelling

Credit Risk Modelling;
Credit risk modelling is a data-driven process that uses historical data to predict the likelihood of a borrower defaulting on a loan. This includes estimating the amount of money the borrower owes at the time of default, as well as the lender's potential losses. In other words, credit risk modelling involves building three key models: the probability of default (PD) model, the loss given default (LGD) model, and the exposure at default (EAD) model. These models are used to calculate the overall credit risk of a loan, which helps lenders make informed lending decisions.

Credit risk modeling is a complex process that involves a number of challenges. One of the most common challenges is data quality. The data used to train the model must be accurate and complete, but credit data can often be noisy and incomplete. This can make it difficult to build a reliable model.
Another challenge is model selection. There are a number of different credit risk models that can be used, and the choice of model depends on the specific data set and the desired outcome. However, it can be difficult to choose the right model, and different models can produce different results.
Model validation is also important. This is the process of testing the model to ensure that it is accurate and reliable. This can be done by using a holdout sample or by cross-validation. However, validation can be difficult, especially if the data set is small.
Finally, it is important to update the model regularly. The credit risk environment is constantly changing, so it is important to ensure that the model is up-to-date. This can be difficult, as it requires new data and the re-training of the model.

3. Related works

{Credit Scoring via Logistic Regression (2018)}: This paper used the German Credit dataset to study how creditworthiness depends upon certain other variables. The results indicated that various factors can influence an individual's creditworthiness, including the balance in a consumer's checking account, the loan spread, the loan term, and the age of the customer. The risk of default also decreases if the customer owns more credit cards. This study provides evidence that logistic regression can be used to develop effective credit risk models.
{A logistic regression model for consumer default risk (2020)}: This paper presented a logistic regression model for predicting the probability of default for consumer loans. The model was developed using data from a Portuguese financial institution, and it was found to be effective in predicting default. The study found that the risk of default increases with the loan spread, loan term, and age of the customer. The risk of default also decreases if the customer owns more credit cards. Clients receiving the salary in the same banking institution of the loan have less chance of default than clients receiving their salary in another institution. Clients in the lowest income tax echelon have more propensity to default. The model was able to predict default correctly in 89.79% of the cases. This study further validates the use of logistic regression for credit risk modeling.
Overall, these papers provide evidence that logistic regression can be a useful tool for credit risk modeling. The papers also discuss some of the limitations of logistic regression, such as its sensitivity to the binning of categorical variables and its difficulty to interpret. However, the

advantages of logistic regression, such as its ease of use and its ability to handle both continuous and categorical variables, outweigh these limitations.

4. Data and Preprocessing

Yeh, I-Cheng. (2016). default of credit card clients.
UCI Machine Learning Repository.
URL-**https://doi.org/10.24432/C55S3H**

4.1- Preprocessing and Estimation.

In preparation for our analysis, we conducted minor data preprocessing steps to ensure the simplicity and accuracy of our findings. The initial data involved categorical variables, which we transformed into numerical values to facilitate the subsequent logistic regression.

Additionally, we addressed missing values to prevent any bias in the analysis.
To assess the performance of the logistic regression model, we divided the data into two separate sets: a training set, comprising 80% of the data, and a testing set, making up the remaining 20%. The training set allowed us to train the logistic regression model, while the testing set served as an independent dataset for evaluating the model's accuracy. During the evaluation, we measured the model's performance on the testing dataset, which provided crucial insights into its predictive capabilities. These steps ensured the rigor and reliability of our analysis and are vital components of our research study.

This dataset contains data on customers' default payments in Taiwan during April-September 2005. The variable description is as, a binary (independent) variable, default payment (Yes = 1, No = 0), asthe response variable, and used the following 23 variables as explanatory variables (3 categorical and 20 numeric variables) having 2000 observations each.

4.2- Description of data:

Table notation	Variables	Explanation/ Possible values
X1	LIMIT_BAL	Amount of the given credit (NT dollar):
X2	GENDER	(1 = male; 2 = female)
X3	Education	(1 = high school; 2 = college; 3 = graduate school; 4 = others).
X4	Marriage	(1 = married; 2 = single; 3 = others).
X5	Age	(year)

X6-X11	PAY_1 to Pay_6	Tracked the repayment status of borrowers over 6 months, from April to September 2005. The measurement scale for repayment status is -1 (pay duly) to 9 (payment delay for nine months and above){PAY_1 for sept and PAY_^ from April}
X12-X17	Bill_AMT1 to Bill_AMT6	Amount of bill statement (NT dollar). X12 = amount of bill statement in September 2005; X13 = amount of bill statement in August 2005; X17 = amount of bill statement in April 2005.
X18-X23	Pay_AMT1 to Pay_AMT6	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 =amount paid in August, 2005; . . .; X23 = amount paid in April 2005.
Y	Default payment next month	(1=default, 0=non-default)

Table 1. List of Variables.

5. Methods Employed

In order to deduce the model and predict the probability of default, we have used Binary Logistic Regression, and here is an example of how logistic regression could be used. Let's say we want to predict whether a person will be admitted to a college. We have a dataset of students who have applied to the college, and we know whether or not they were admitted. We also have data on the student's GPA, SAT scores, and extracurricular activities. We can use logistic regression to build a model that predicts the probability of admission based on these factors.

It is a statistical model that is often used for classification and predictive analytics. It estimates the probability of an event occurring, such as whether someone voted or not, based on a given dataset of independent variables. The dependent variable in logistic regression is probability, which means that it is bounded between 0 and 1.
To calculate the probability, logistic regression applies a logit transformation to the odds of the event occurring. The odds are the probability of success divided by the probability of failure. The logit transformation is represented by the following equation:

$$\text{Logit}(pi) = \ln(pi / (1-pi))$$

Where pi is the probability of the event occurring.

The beta parameters in the logistic regression equation are estimated using maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of the log odds. Once the optimal coefficients are found, the conditional probabilities for each observation can be calculated and summed together to yield a predicted probability. For binary classification, a probability less than 0.5 will predict 0 while a probability greater than 0.5 will predict 1.

In our case, when the response variable Y follows a Bernoulli distribution of parameter μ, then the generalized linear model uses the logit function as the canonical link function and becomes a logistic regression model.
As $Yi \sim \text{Ber}(\mu i)$, then $\mu i = P(Yi = 1)$. The variable Default is a binary variable Y such that Y = 1 if defaulted, and 0 otherwise. Using the logistic regression model, the PD is a function of a set of explanatory variables X as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-\beta X}} \quad \ldots\ldots. (1)$$

5.1- Model:-

In logistic regression models, rather than looking at the coefficients βi per se, it is more important to focus on the values of exp(βi), because they represent the influence that the increase in an independent variable Xi has in the probability of the dependent variable Y becoming 1. It follows that:

$$log\left(\frac{P(Y = 1|Xi)}{1 - P(Y = 1|Xi)}\right) = \beta 0 + \beta 1X1 + \ldots\ldots\ldots\ldots + \beta iXi \quad \ldots\ldots (2)$$

5.2- Model estimation:-

In binary logistic regression models for credit risk, the coefficients βi represent the increase in the odds of a customer defaulting on their loan for a one-unit increase in the independent variable Xi. The odds of defaulting is calculated as 1 / (1 + exp (-βi)). Therefore, the values of exp (βi) represent the multiplicative effect that an independent variable has on the odds of default. For example, if exp(βi) = 2, then a one-unit increase in Xi is associated with a doubling of the odds of default.

Following the division of our dataset into training and test sets, we conducted logistic regression using Python to model the relationship between the predictors and the binary outcome variable. The logistic regression model allowed us to assess the impact of various factors on the likelihood of default. Upon completion of the regression analysis, we obtained comprehensive summary statistics of the model. These statistics offer crucial insights into the significance and magnitude of the predictor variables, providing a deeper understanding of their influence on the outcome.

6. Results and Discussion

Upon fitting the variables and running the regression using machine learning, we have drawn the results in Table 2.
In binary logistic regression models for credit risk, the coefficients βi represent the increase in the odds of a customer defaulting on their loan for a one-unit increase in the independent variable Xi. The odds of defaulting are calculated as 1 / (1 + exp(-βi)). Therefore, the values of exp (βi) represent the multiplicative effect that an independent variable has on the odds of default. For example, if exp (βi) = 2, then a one-unit increase in Xi is associated with a doubling of the odds of default.

The estimates of the coefficients are presented in table 2. Among the features considered, "PAY_1," "PAY_3," "PAY_AMT1," and "PAY_AMT2" are the most significant in predicting credit risk. These variables have coefficients with p-values less than 0.05, indicating they are likely to be important predictors.

For example, for the PAY_1 the coefficient is 0.59, which shows the positive impact and the exp (coeff) is 1.814 means that for each month's delayed repayment at the very last observed month, the chance of default will increase by 81.4%.
Males were slightly less likely to default on a loan than females, with a one-unit increase in the GENDER variable (male = 2, female = 1) associated with a decrease in the log odds of defaulting.

Feature	Coefficient	Standard Error
LIMIT_BAL	-0.015	0.085
GENDER	-0.017	0.065
EDUCATION	-0.055	0.071
MARRIAGE	-0.096	0.070
AGE	0.108	0.071
PAY_1	0.596	0.082
PAY_2	-0.151	0.108
PAY_3	0.260	0.124
PAY_4	-0.084	0.141
PAY_5	0.213	0.143
PAY_6	-0.118	0.117
BILL_AMT1	-0.447	0.348
BILL_AMT2	0.020	0.439
BILL_AMT3	0.509	0.380
BILL_AMT4	0.429	0.262
BILL_AMT5	-0.244	0.397
BILL_AMT6	-0.319	0.324
PAY_AMT1	-0.325	0.172
PAY_AMT2	-0.347	0.167
PAY_AMT3	0.037	0.077
PAY_AMT4	-0.180	0.122
PAY_AMT5	0.076	0.102
PAY_AMT6	0.030	0.093

Table 2. Summary Statistics.

It is important to note that the fact that these coefficients are statistically insignificant does not mean that they are necessarily irrelevant. It is possible that these variables have a small impact on the risk of defaulting on a loan, but that this impact is not large enough to be statistically significant. Additionally, it is possible that these variables interact with other variables in a way that makes them statistically significant. For example, the effect of marriage on the risk of defaulting on a loan may be different for males and females.
Overall, the results of this study suggest that married individuals and individuals with a college degree may be at a slightly lower risk of defaulting on a loan than unmarried individuals and individuals with less than a college degree. However, more research is needed to confirm these findings and to determine the specific mechanisms by which these variables impact the risk of defaulting on a loan.

7. Model Validation.

The final logistic regression model was the one shown in Equation (2). The coefficient estimates for this model are shown in Table 2. Before this model can be used to estimate the probability of a client defaulting, it must be validated through a series of validation scores.
This will ensure that the model is accurate and reliable before it is used to predict default cases.

Recall: - *recall = true positives / (true positives + false negatives)*
It is also known as sensitivity or true positive rate, which measures the proportion of actual positives that are correctly classified as positive. A low recall value of 0.0987 suggests that the model may not be effectively capturing and identifying a significant portion of actual risky credit cases. This can be a concern for a credit risk model, as correctly identifying high-risk cases is crucial to prevent potential defaults.
However, it is important to note that the dataset was imbalanced, with a much lower proportion of cases that defaulted on their loans. This means that the low recall value may be due to the fact that there were simply fewer actual risky credit cases in the dataset.
In order to get a more accurate understanding of the recall value, it would be necessary to evaluate the model on a more balanced dataset. This would allow us to determine whether the model is truly ineffective at identifying risky credit cases, or if the low recall value is simply due to the imbalance in the dataset.

AUC-ROC:-
(Area under the Receiver Operating Characteristic Curve) It is a graphical plot that illustrates the trade-off between a true positive rate (sensitivity) and a false positive rate at different probability thresholds. True positive rate (sensitivity) is the proportion of actual positives that are correctly classified as positive. The false positive rate is the proportion of actual negatives that are incorrectly classified as positive. An AUC ROC value of 0.5 indicates that the model performs no better than random guessing, while an AUC ROC value of 1 indicates that the model is perfect.

In this model, the AUC ROC value of 0.7171 suggests that the model has a reasonable ability to distinguish between good and risky credit cases. This means that the model is able to distinguish between cases that did and did not default on their loans with a relatively high degree of accuracy.
Generally, an AUC ROC value above 0.5 indicates better-than-random performance, but the closer the AUC ROC is to 1, the better the model's discriminatory power.

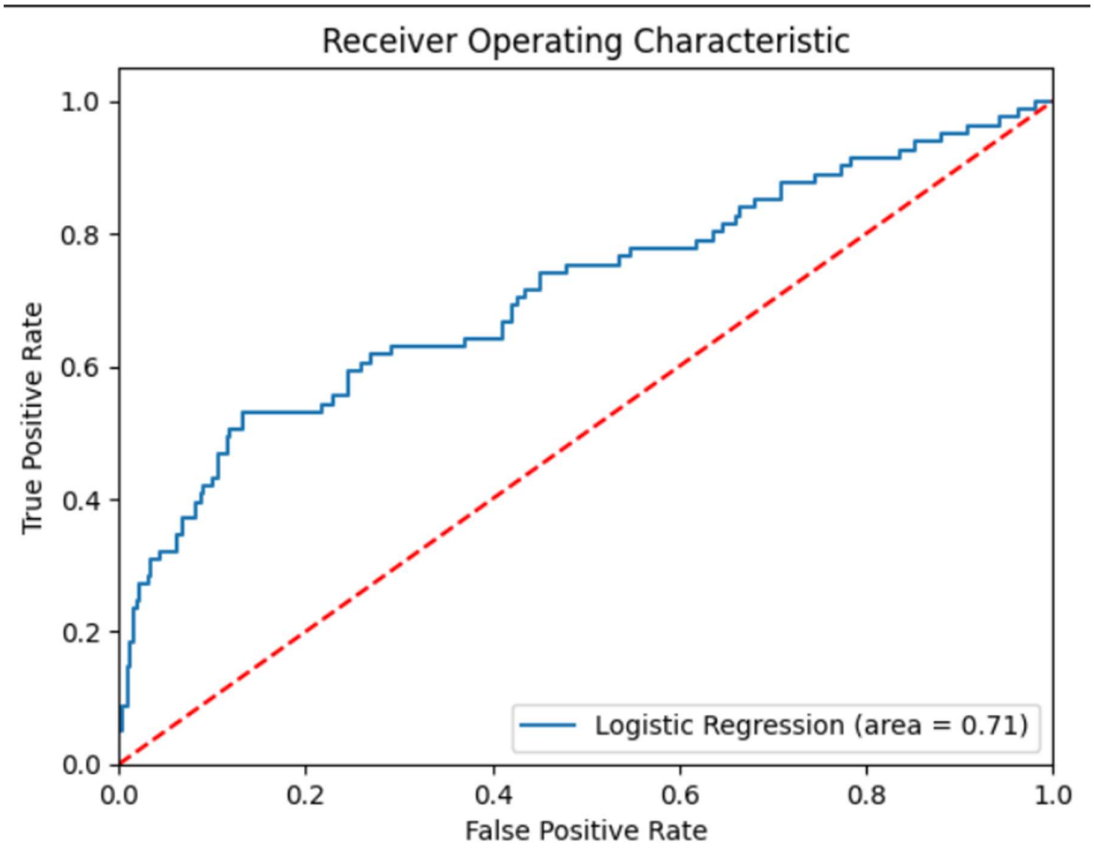

Figure 4. ROC curve (when the model is applied to the whole dataset).

8. Conclusion

This study provides evidence of the effectiveness of logistic regression as a valuable tool for credit risk modeling. By leveraging a comprehensive dataset that includes historical, demographic, and central bureau information, the researchers successfully predicted credit risk for borrowers in a prominent Portuguese banking institution. Implementing this model in the central bank has enabled efficient and informed credit approval decisions, contributing to reduced financial risks in the market. The analysis identified several crucial explanatory variables, such as "PAY_1," "PAY_3," "PAY_AMT1," and "PAY_AMT2," which play a significant role in assessing credit risk. Borrowers who fail to repay loans for an extended period are at higher risk of default, while timely repayments significantly reduce the likelihood of default. Moreover, lower education levels have been associated with a higher propensity to default.

The study also shed light on certain demographic trends affecting credit risk. Males exhibited a slightly lower risk of default compared to females, and married individuals demonstrated a lower likelihood of default. Clients with a college degree were also found to be marginally less prone to default than those with lower levels of education. However, the impact of gender, marriage, and education on credit risk requires further investigation.

The robustness of the model was validated using Precision, Recall, and Accuracy metrics, demonstrating its ability to make informed predictions even with imbalanced data. The AUC ROC score of 0.71 highlighted its effectiveness in discriminating between good and bad credit outcomes.

While logistic regression proved effective, the study acknowledged some limitations, such as the absence of client personal information like salary and spending transactions in the dataset. Additionally, the applicability of the model in other countries with varying requirements, laws, and roles may differ.

Looking ahead, future research could explore comparisons of logistic regression with deterministic artificial intelligence methods to further enhance credit risk modeling. Moreover, incorporating additional factors or novel approaches, such as Weight of Evidence (WOE) scores, could refine the predictive power of credit risk models.

In conclusion, this research emphasizes the effectiveness and insights provided by logistic regression for credit risk modeling. The use of big centralized data sources and automated decision-making can help banks minimize financial risks and improve credit approval processes. As data science approaches continue to evolve, the findings from this study will inform new data-driven strategies, fostering better predictions and reducing credit and bank risks in the dynamic economic landscape.

References

1. *Credit Scoring via Logistic Regression.* **Al-Aradi, A. 2018.** Toronto, Canada: s.n., 2018.
2. *A logistic regression model for consumer default risk,* **Faria, Eliana Costa e Silva and Isabel Cristina Lopes and Aldina Correia and Susana. 2020.** s.l. : Taylor & Francis, 2020,Vol. 47.
3. **Wang, Xiang Yang and Yongbin Zhu and Li Yan and Xin. 2015/11.** Credit Risk Model Based on Logistic Regression and Weight of Evidence. *Proceedings of the 2015 3rd International Conference on Management Science, Education Technology, Arts, Social Science, and Economics.* Qingdao, China: Atlantic Press, 2015/11, pp. 810-814.