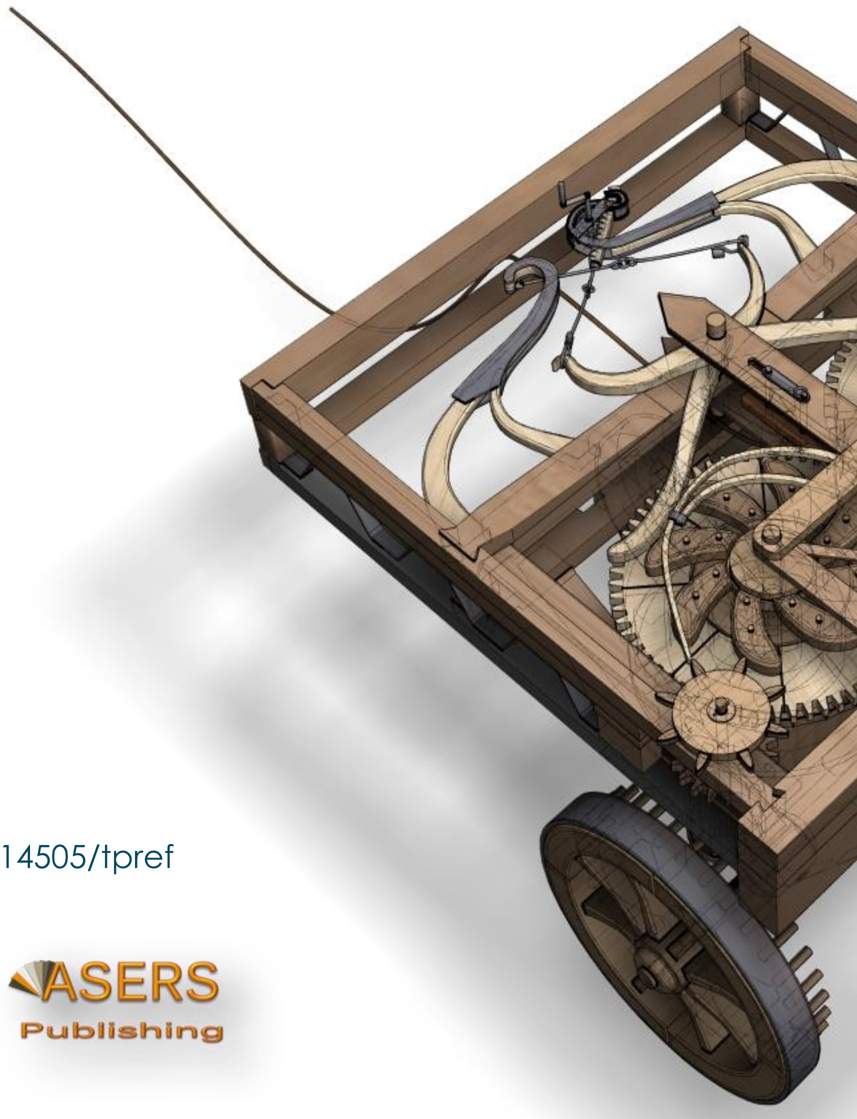
Theoretical and Practical Research in Economic Fields

Special Issue



Quarterly
Volume XV
Issue 2(30)
Summer 2024

ISSN: 2068 – 7710
Journal DOI: https://doi.org/10.14505/tpref

ASERS Publishing

Volume XV
Issue 2(30)
Summer 2024

ASERS Publishing
ISSN 2068 – 7710
Journal's Issue DOI:
https://doi.org/10.14505/tpref.v15.2(30).00

Table of Contents

Volume XV
Issue 2(30)
Summer 2024

ASERS Publishing
http://www.asers.eu/asers-publishing
ISSN 2068 – 7710
Journal's Issue DOI:
https://doi.org/10.14505/tpref.v15.2(30).00

Call for Papers
Fall Issue 2024
Theoretical and Practical Research in Economic Fields

Many economists today are concerned by the proliferation of journals and the concomitant labyrinth of research to be conquered in order to reach the specific information they require. To combat this tendency, **Theoretical and Practical Research in Economic Fields** has been conceived and designed outside the realm of the traditional economics journal. It consists of concise communications that provide a means of rapid and efficient dissemination of new results, models, and methods in all fields of economic research.

Theoretical and Practical Research in Economic Fields publishes original articles in all branches of economics – theoretical and practical, abstract, and applied, providing wide-ranging coverage across the subject area.

Journal promotes research that aim at the unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking. It explores a unique range of topics from the frontier of theoretical developments in many new and important areas, to research on current and applied economic problems, to methodologically innovative, theoretical, and applied studies in economics. The interaction between practical work and economic policy is an important feature of the journal.

Theoretical and Practical Research in Economic Fields is indexed in SCOPUS, RePEC, ProQuest, Cabell Directories and CEEOL databases.

The primary aim of the Journal has been and remains the provision of a forum for the dissemination of a variety of international issues, practical research, and other matters of interest to researchers and practitioners in a diversity of subject areas linked to the broad theme of economic sciences.

At the same time, the journal encourages the interdisciplinary approach within the economic sciences, this being a challenge for all researchers.

The advisory board of the journal includes distinguished scholars who have fruitfully straddled disciplinary boundaries in their academic research.

All the papers will be first considered by the Editors for general relevance, originality, and significance. If accepted for review, papers will then be subject to double blind peer review.

This Special Issue was created at the request of a group of researchers from Ukraine. It is a response to the challenging situation of Ukrainian scholars due to the Russian invasion as well as the growing demand for knowledge on Ukrainian issues.

We would like to express our endless thank to our colleagues, scholars from Ukraine who are working amid the war on topics that are important for all. Also, we thank all our international authors for their valuable contributions to this Issue.

Deadline for submission of proposals:	10th August 2024
Expected publication date:	September 2024
Website:	http://journals.aserspublishing.eu/tpref
E-mail:	tpref@aserspublishing.eu

To prepare your paper for submission, please see full author guidelines in the following file: https://journals.aserspublishing.eu/tpref/Template_for_Authors_TPREF_2024.docx on our site.

ASERS Publishing

DOI: https://doi.org/10.14505/tpref.v15.2(30).25

Evaluating the Impact of Borrower Characteristics, Loan Specific Parameters, and Property Conditions on Mortgage Default Risk

Ali Mahmoud ANTAR
Department of Economics, Faculty of Business Administration
Beirut Arab University, Lebanon
ORCID: 0009-0001-2807-1543
ama935@student.bau.edu.lb

Article info: Received 23 April 2024; Received in revised form 6 May 2024; Accepted for publication 3 June 2024; Published 28 June 2024.

Abstract: This study empirically examines the impact of borrower, loan, and mortgage parameters on default risk in residential mortgage loans. Using 6743 individual housing loan accounts data from Housing Finance Institutions in Lebanon, we develop a comprehensive model using the multivariable binary logistic regression, best subset logistic regression, and stepwise regression analysis procedures to investigate the impact of 21 predictors and 29 sub-predictor parameters on log odds of default risk. In addition, the study conducted a model diagnosis using the Hosmer - Lemeshow Goodness of fit test, Likelihood Ratio Test, Model accuracy- Classification Table, Statistically Significant Test- ROC curve, and Pregibon Delta Beta Statistics. The study aims to assist financial institutions in managing and assessing the default risk more effectively and develop effective strategies to mitigate this risk. The empirical results suggest that the estimated probability of defaulting on a housing loan is approximately 3.8% when all predicted variables are set at their lowest value. In addition, marital status and the existence of dependence have a positive impact on default risk. The higher the number of dependents is, the higher the risk of default. Moreover, a widowed borrower has a higher log odd of default compared to single, married, and divorced borrowers. Furthermore, the results revealed that self-employed borrowers positively impact the risk of default due to the absence of a steady flow of income. In addition, there is an inverse relationship between the market price-to loan ratio and the log odds of default since the borrower's equity will increase when the house price increases. However, log odds of default will increase when the loan value is higher than the mortgage market price. Moreover, the result shows that the nature of the borrower's occupation has a positive relationship with log odds of default where borrowers working in real estate and construction sectors have lower default rates than borrowers working in other industries. In addition, a high interest rate increases the loan's monthly payment and therefore increases the probability of default. Furthermore, the loans granted for purchase and renovation purposes have a lower risk of default than the ones given for construction and under-construction. In addition, the model's overall accuracy was demonstrated by a 91.61 percent visible correct classification rate.

Keywords: binary logistic regression; loan default; credit risk; housing finance; risk management; STATA statistical software.

JEL Classification: C23; C52; C53; C54; C55; C58; G21; G28.

Introduction

Housing loan default has become a significant problem in many countries, with borrowers facing financial distress due to the inability to make timely loan payments. Many of the financial institutions are at risk of incurring financial losses from the high rate of loan defaults. Therefore, it is crucial to identify the factors that contribute to housing loan default to develop effective strategies to mitigate this risk.

Housing loans are a critical source of financing for homeownership, enabling individuals and families to purchase their own homes. Financial institutions grant housing loans to individual borrowers after studying their financial capabilities and ensuring their financial eligibility and ability to pay back the loan principal amount, interest rate, and other lending noninterest expenses. In addition, the lender applies lending policies at the loan origination stage and therefore policymakers need to explore most of the possible parameters that might increase the probability of the risk of default during the loan duration to maturity.

Identifying parameters that contribute to default risk can assist financial institutions in managing and assessing the risk of default more effectively. It helps to develop models that can predict the likelihood of default and enhance decision-making procedures. In addition, lenders can set optimal interest rates and fees that reflect the level of risk associated with housing loans. In addition, exploring borrower, loan, and property parameters that drive default risk helps lenders allocate credit more efficiently. It allows them to grant loans to borrowers who are less likely to default. Furthermore, assessing default risk helps maintain economic, housing market, and financial system stability.

While previous studies have dedicated attention to quantitative factors affecting default risk in residential mortgage loans such as borrower credit score, loan-to-value, and debt ratio, there is a growing recognition of the need to examine socio-economic categorical parameters that might influence default risk besides quantitative variables. This research study aims to fill this gap by empirically investigating the impact of both quantitative and categorical dummy variables related to borrower, loan, and property characteristics on the risk of default in residential mortgage loan

The study examines factors contributing to default to help lenders identify and avoid risky loans, which can reduce their overall losses and improve their profitability. In addition, this can also help to ensure that borrowers are not given loans that they are unlikely not able to repay, which can prevent financial hardship for individuals and families. This can also assist in minimizing the negative impact on borrowers, who may face financial and legal consequences if they default on their loans. Second, understanding the risk factors affecting housing loan default can also inform public policy decisions related to housing finance and consumer protection. By identifying trends and patterns in loan default, policymakers can develop policies and regulations that better protect consumers and promote a stable and sustainable housing finance system, and be able to develop effective strategies to mitigate this risk.

This research aims to identify and analyze the key determinants of default risk in housing loans and to understand how these factors impact the likelihood of default. The study seeks to develop a comprehensive model that can predict default risk based on 41 quantitative and categorical variables related to various borrower characteristics, loan-specific variables, and property parameters. We will use cross-sectional data of 6743 individual housing loan accounts for loans granted by the Housing Financial Institutions in Lebanon during the period extending from 2005 to 2020.

The importance of this research is to understand the interplay between various determinants driving housing loan default since it has a direct negative effect on all parties involved in the lending process, banking system, and the whole economy. As for the lender, the default will decrease the bank's capital since both loan principal and interest payment will fail to be repaid by the borrower, affecting negatively any future funding, and leading to a drop in investment rate. In addition, defaulted loans will shrink the bank's equity since the loss shall be deducted from the equity as a provision. Furthermore, failure to repay the mortgage loan will also have a direct negative impact on the borrowers themselves. Borrowers will lose homes, have lower credit scores, and as a result, will not be able to borrow again in the future.

1. Literature Review

Many studies were conducted to predict the factors that increase the default risk of residential loans. The current literature has identified various factors that affect the likelihood of defaulting on a housing loan. The literature suggests that borrower characteristics, and loan characteristics, play a significant role in predicting the likelihood of defaulting on a housing loan. These factors are interconnected, and understanding their interactions is crucial in managing the risk of housing loan defaults. The findings of this literature review can inform the development of models and policies aimed at mitigating the risk of housing loan defaults.

Studies have consistently shown that borrower characteristics play a key role in housing loan default. Lower credit scores, higher debt-to-income ratios, and unstable employment are all associated with a higher likelihood of default. Moreover, borrowers with a history of missed payments or defaults are more likely to default again in the future. In contrast, as indicated by Li and Yang (2018), borrowers with higher incomes and more stable employment are less likely to default. In addition, the literature also suggests that borrower characteristics, such as credit score, income, employment status, and debt-to-income ratio, are significant factors in predicting the likelihood of defaulting on a housing loan. Another study was conducted by Mayer and Pence (2009) and empirical results suggest that borrowers with low credit scores and high debt-to-income ratios were more likely to default.

Additionally, borrowers with unstable employment and income were also more likely to default. Furthermore, a study by Sandar *et al.* (2010) was conducted to predict the borrower-related determinants that affect the default risk of micro-finance loans. The empirical results revealed that default risk is mainly linked to borrowers who suffer from health problems and therefore have high medication expenses, female borrowers who have lower years of experience, poor educated borrowers, and those who have outside loans with high interest rates. Moreover, another study was conducted by Canepa and Khaled (2018), and the result revealed, regarding borrowers' default risk, that the higher the debt of borrowers, the harder it becomes for them to pay their scheduled obligations.

Further, the results of a study by Bandyopadhyay and Saha (2009), which aimed to examine major factors affecting housing demand, and default risk, revealed that the default risk will be decreased if the borrower submits a collateral additional to the main guarantee. Further, the presence of a greater number of co-borrowers in the loan will significantly decrease the risk of default. This is because the existence of a co-borrower will include many incomes. Moreover, the results show that the borrower's age and the existence of dependence have an impact on default risk. The older the borrower and the higher the number of dependents is, the higher the risk of default is. Also, the study examined the effect of employment on housing default, and the results revealed that employment has a significantly negative impact on default risk while self-employed borrowers have a positive impact on the risk of default due to the absence of a steady flow of income.

In addition, the result of a study conducted by Levy and Kwai-Choi stated that the 'purpose of purchase' plays an important role in residential default risk. Borrowers who apply for a housing loan for personal investment rather than to own an occupied residence prefer to pay a smaller amount of the property's selling price as a down payment and thus decrease the initial equity commitment. Therefore, when the value of the property (collateral) falls due to any economic distress or due to a supply-demand mismatch in the housing market, borrowers will settle the housing loan to limit their loss and therefore this will decrease the risk of defaulting.

Loan characteristics such as loan-to-value (LTV) ratio, loan term, and interest rate also influence the likelihood of default. Higher LTV ratios are associated with higher default rates, as borrowers with less equity in their homes have less to lose by defaulting. Longer loan terms also increase the likelihood of default, as borrowers are more likely to experience income shocks over a longer period. Moreover, adjustable-rate mortgages (ARMs) have been found to have higher default rates than fixed-rate mortgages, particularly during periods of economic stress. In addition, the features of the loan itself, such as the loan-to-value ratio and interest rates, have also been found to affect the likelihood of default. Studies, such as that of Foote *et al.* and Gerardi *et al.* (2008) have shown that a high loan-to-value ratio and adjustable-rate mortgages are associated with a higher risk of default.

Moreover, property characteristics such as location, type, and condition can also affect the likelihood of default. Properties located in areas with higher unemployment rates, crime rates, and declining property values are more likely to experience default. Moreover, properties with structural or maintenance issues are more likely to be subject to default. Additionally, LaCour-Little *et al.* show that borrowers who live further away from their properties are more likely to default, as distance makes it harder to monitor and maintain the property.

Furthermore, Canepa and Khaled also examined the relationship between the housing default risk and the collateral value. The result revealed that there is an inverse relationship between the change in house prices and the housing risk of default. An increase in house prices will lead to an increase in the value of the collateral and vice versa. Therefore, borrowers' probability to default will decrease to avoid losing their housing. In addition, it has been found that housing loans granted to borrowers where the house subject of the loan is located in sub-urban and rural areas are significantly riskier than the loans provided for housing located in urban areas.

Moreover, a study was conducted in the Irish mortgage market revealed that housing loan default risk increases when the housing market price worth less than the repayment value of the mortgage. In addition, the vulnerability of banks depends on the correlation between the falling in housing prices and the borrower's level of equity invested when granted the housing loan. When housing price increase, borrowers will accumulate wealth created by home-price appreciation. However, as mentioned by Rosengren, if price decreases, the probability of default will rise. Additionally, Foote *et al.* found that a decline in housing prices increases the likelihood of default.

2. Research Methodology

This chapter describes the research methodology used in the current study to meet the objective of the study which is to develop a prediction model from a cross-sectional historical dataset collected for existing individual customers from Lebanese financial institutions. The methodology includes the research design, research model, sources of data collecting, and estimating techniques.

2.1. Research Design

This study finds out the determinants that lead to housing loan default in the Lebanese market as well as discovers the impact of each determinant on the risk of default to assist lending policies and strategies adopted by Banks to decrease and /or hedge against such risk.

The research design section describes the overall strategy adopted to accumulate the various components of the research in a comprehensible and analytical way. It is a rational technical plan that ensures the accuracy, reliability, and validity of the survey that shall stay intact. The present research seeks to investigate the determinants of default housing loans in Lebanon. To achieve that aim, the researcher measures the impact of different parameters on house loan default. This is conducted by taking into consideration borrower, loan, and mortgage predictors. These three predictors consist of 20 parameters and 21 sub-predictor parameters.

Borrower Characteristics include the following predictors and sub-parameters: age, gender with sub-categories (male, female, couples as co-borrowers, marital status which categorized into single, married, divorced, widow borrowers, number of children, borrower's job economic sectors including the banking, construction, industrial, service, public, private sectors, borrower's occupation category as employed, self-employed, and freelance, job location whether residential or expatriate, income, the existence of additional guarantees, and debt ratio. In addition, Loan parameters consist of the loan amount, monthly installment, tenor, interest rate, loan type which is categorized as purchase, construction, under-construction, renovation loans, and monthly installment to income ratio. Furthermore, the mortgage parameters include: market price, book value, location (Beirut, Mount Lebanon, North, South), loan-to-value ratio, and loan-to-market price ratio.

The dependent variable is default risk: This variable is binary and represents whether or not the borrower defaulted on the loan. Default rates will be the focus of this paper because we want to analyze how they could be related to other variables.

2.2. Research Model

We will use binary logistic regression analysis to examine the relationship between the predictor variables and the log-odds of the outcome using the logistic function (also known as the sigmoid function). The logistic function ensures that the predicted probabilities range between 0 and 1, which is suitable for binary outcomes. The logistic regression model estimates the coefficients (log-odds ratios) associated with each predictor variable, indicating the direction and strength of their relationship with the outcome. These coefficients are typically estimated using maximum likelihood estimation (Hosmer *et al.* 2013).

The binary logistic regression model function is derived based on the principles of maximum likelihood estimation. It assumes that the log odds of the probability of the outcome dependent binary variable occurring are a linear function of the predictor independent variables.

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k \qquad (1)$$

where: P is the probability of the outcome variable occurring

$x_1 , x_2 , \ldots\ldots , x_k$ are the predictor variables

β_0 is the intercept

$\beta_1, \beta_2 , \ldots\ldots \beta_k$ are the coefficients to be estimated.

The logistic function is then applied to transform the linear combination of predictor variables into a probability between 0 and 1:

$$P = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k)}} \qquad (2)$$

where e is the base of the natural logarithm. The logistic function above ensures that the estimated probabilities fall within the range of 0 and 1, which makes it suitable for modeling binary outcomes. The method used to estimate the coefficient $\beta_0 , \beta_1, \beta_2, \ldots \beta_k$ is the maximum likelihood estimation method. This method pursues to find the set of coefficients that maximizes the likelihood of observing the given data under the assumed logistic regression model to get the best fitting model parameters (Moore *et al.* 2018).

2.3 Data Collection

Data is collected from Lebanese financial Institutions between 2005 and 2020 using 6743 individual housing loan accounts. There are 566 accounts that are defaulted representing 8.4% of the total number of loans. The predictors that might influence the odds of default in this study are 20 independent variables out of which 12 variables are continuous explanatory variables and 8 are discrete categorical independent variables.

2.4. Estimation Technique

2.4.1. Selection Criteria and Model Development

To determine which binary logistic regression model is best, we first conduct univariable regression analysis for each explanatory variable then we select predictors for the multivariable regression analysis. A predictor is a candidate for the multivariable model if its univariable test yields a p-value less than 0.25. The suggestion that a screening criterion for variable selection be employed at a 0.25 level stems from the research conducted by Mickey and Greenland[13] as well as Bendel and Afifi (2017). Then we will conduct a multivariable regression analysis and choose variables with p-value < 0.1 to develop the candidate fitted model. Then model will be built using the best subset model based on the lowest AIC. First, we run a model that includes variables with a p-value less than 0.25 after performing univariable regression analysis, then the new model eliminates variables with a p-value greater than 0.1 after performing multivariable regression. Finally, we will add to the model the eliminated variables and run a combination of variables for those who were eliminated since their p-value is greater than 0.25 and those with a p-value greater than 0.1 to select the best model based on the one that has the lowest AIC.

Next, we will have refined the main effects model and check for interactions among the predictors in the model. We include the interaction variable in the new model and compare it to the previous model using the likelihood ratio test. After obtaining the fitted model, next we perform a diagnosis test for the fitted model, as indicated in Neyman's (2023) study.

2.4.2. Model Diagnosis

After obtaining the candidate-fitted model, next, we perform a diagnosis test. The diagnosis procedures will be applied in this study including the following: Hosmer-Lemeshow test for overall goodness of fit, likelihood ratio test, model adequacy test through Link test, model accuracy and classification table, and Roc curve.

2.4.2.1. Goodness of Fit Tests

Goodness-of-fit tests are used to evaluate how well a statistical model fits the observed data. It provides a measure of the discrepancy between the observed data and the estimated values predicted by the model. We will apply in this study the chi-square goodness-of-fit test for hypothesis testing, and the Hosmer Lemeshow test to assess the adequacy of the fitted model. The model's suitability for fitting the data is the null hypothesis. A small p-value from the Hosmer Lemeshow test indicates a poor fit of the logistic regression model. In this paper, we will perform the Test of Overall Goodness of Fit using Stata Statistical software using the command 'estat gof'.

H0: Data is correctly fitted in the current model

HA: Not

We will reject the null hypothesis if the outcome of the chi-square statistic is high.

2.4.2.2. Likelihood Ratio Test (LRT)

The likelihood ratio test (LRT) is a statistical technique used to compare the fit of two nested models: a full model and a reduced model. The full model includes all predictor variables of interest, while the reduced model is a simplified version of the full model with fewer predictor variables. The likelihood ratio test measures whether the additional variables in the full model significantly improve the fit of the model compared to the reduced model. It compares the likelihood of the observed data under both models and determines whether the improvement in fit is statistically significant. It is obtained as the result of the difference in log-likelihoods between the full model (L_f) and the reduced model (L_r).

$$\text{LR} = -2 \text{ X } (L_f - L_r) \tag{3}$$

Under the null hypothesis that the reduced model is sufficient to explain the data, the value of 'LR' approximately follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the full and reduced models. A small p-value from the likelihood ratio test indicates that the full model provides a significantly better fit to the data than the reduced model, suggesting that the additional predictor variables contribute significantly to explaining the outcome variable. The likelihood ratio test, as mentioned by Hosmer *et al.*, is widely used in logistic regression for model comparison, variable selection, and assessing the overall goodness of fit of the model to the data.

2.4.2.3. Model Accuracy - Classification Table

The Contingency or classification table is a tabular representation of the performance of a classification model. It compares the predicted classifications predicted by the logistic regression model with the actual classifications observed in the data. It has four categories decomposed as below:

True positive represents the number of cases where the logistic regression model correctly predicts a positive outcome when actual outcome is positive; true negative represents the number of cases where the logistic regression model correctly predicts a negative outcome when actual outcome is negative, false positive represents the number of cases where the logistic regression model incorrectly predicts a positive outcome when actual outcome is negative, and false negative represents the number of cases where the logistic regression model incorrectly predicts a negative outcome when actual outcome is positive. Based on the outcomes of classification table, many performance metrics can be obtained including model accuracy, sensitivity, precision, specificity, and others (Hosmer *et al.*).

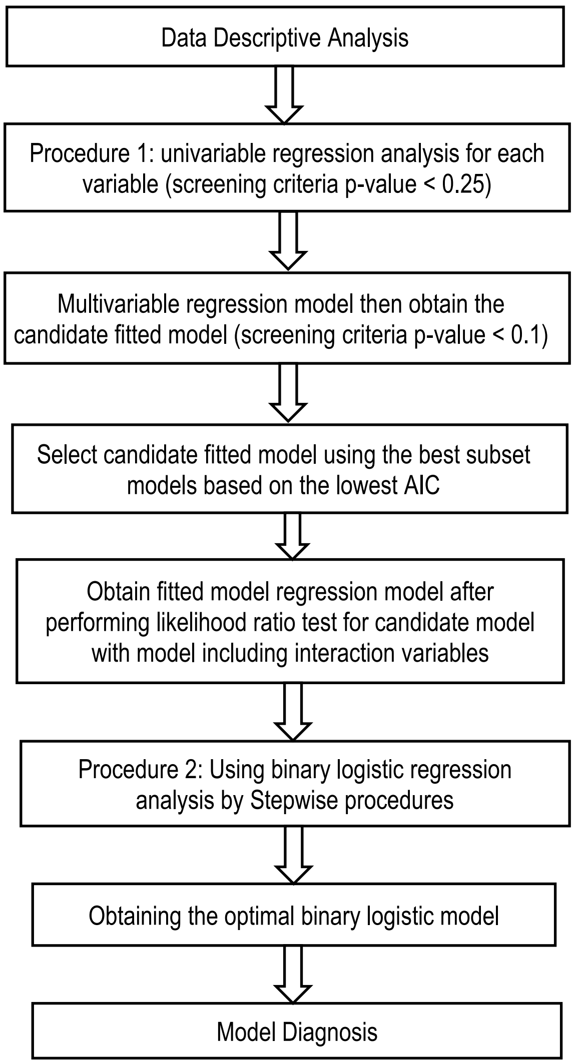
2.4.2.4. Statistically Significant Test – Roc Curve and Pregibon Delta Beta Statistics

A receiver operating characteristic (ROC) curve is a graphical representation of tradeoff between true value rate and false positive rate for different threshold values of the predicted probabilities. The area under ROC curve (AUC) is a commonly used metric to quantify the overall performance of the logistic regression model. A higher AUC value indicates better discrimination ability of the model, with an AUC value of 1 representing perfect discrimination and that of below or equal to 0.5 representing poor discrimination and that better fit is indicated by larger value than 0.5. Therefore, the ROC curve and AUC values assist to evaluate the model discriminating power and select an optimal threshold based on specific requirements of the analysis (Hosmer *et al.*). In addition, Pregibon Delta Beta Statistics helps in identifying any influential points.

2.4.2.5. Methodology Framework

The methodology framework of this study is displayed as below:

Figure 1. Methodology Framework



3. Results and Discussion

3.1. Variables Code Book

In this study, we will try to find the impact of any of 20 independent variables on Loan default. These variables are divided between categorical and continuous variables. Discrete variables amounted eight which are: Gender, Marital Status, Existence of Additional Guarantee, Job category, Economic Sector, Country, Mohafaza, and Loan Type. However, continuous variables amounted 12 which are: Age, Number of Children, Loan amount, monthly payment, income, debt ratio, Book value, LTV, Market price, loan value to market price, Interest Rate, and Loan Tenor.

Table 1. Code Book

Variable	Label	Range /Codes
Default Loan	Default	1=yes, 0 =no
Gender	Female, Male, MF co-borrower	1= Female, 2 =Male, 3= MF
Marital Status	Divorced, Married, Single, Widow	1= Divorced, 2= Married, 3=Single, 4= Widow
Additional Guarantee	Additional Guarantee	1= No additional Guarantee, 2= Yes
Job category	Employee, freelance, Self-employee	1= Employee, 2= Freelance, 3= Self employee
Economical Sector	Banking, Commercial, Construction, Industrial, Public, Service	1 = Banking , 2 = Commercial,3 = Construction, 4 = Industrial, 5 = Public ,6 = Service
Country	Expatriate, Lebanon	1= Expatriate, 2= Lebanon
Mohafaza	Beirut, Bekaa, Mount Lebanon, North, South	1=Beirut, 2= Bekaa, 3= Mount Lebanon ,4= North , 5=South
LoanType	Purchase, Renovation, Under-construction	1 = Purchase, 2= Renovation, 3= Under-construction
Age	Age, years	Continuous
Number of children	Number of children	Continuous
Loan Amount	Loan Amount, millions of Lebanese pounds	Continuous
Monthly Payment	Monthly Payment, thousands of Lebanese pounds	Continuous
Income	Income, thousands of Lebanese pounds	Continuous
debt Ratio	Debt Ratio, Monthly Payment/Income, Ratio	Continuous
Book Value	Book Value, millions of Lebanese pounds	Continuous
LTV	Loan to Value , loan amount/Book Value , Ratio	Continuous
Market Price	Market Price, millions of Lebanese pounds	Continuous
Loan to Market Price	Loan to Market Price , loan/Market Price , Ratio	Continuous
Interest Rate	Interest Rate , cost of money, percentage	Continuous
Loan Tenor	Loan Tenors , years	Continuous

3.2. Descriptive Statistics

3.2.1. Quantitative Data Analysis

Table 2. Summary Table Quantitative Variables

Variable	N	Mean	Std. dev.	Min	Max
Y=0 Non-Default Loan (91.60%)					
Age	6,177	37	7	19	63
Number of Children		2	1	0	8
Loan Amount		200,000,000	130,000,000	18,000,000	800,000,000
Monthly Payment		1,400,000	920,000	120,000	10,000,000

Variable	N	Mean	Std. dev.	Min	Max
Income		6,000,000	4,600,000	687,000	57,000,000
Debt Ratio		0.26	0.07	0.042	0.62
Book Value		340,000,000	240,000,000	24,000,000	3,600,000,000
LTV		0.63	0.16	0.09	1.05
Selling Price		330,000,000	250,000,000	36,000,000	8,000,000,000
Loan to Market value		0.65	0.14	0.06	1.05
interest rate		5.04	0.92	1.628	6.5
Loan Tenor		19.65	5.2	5	30
Y=1, Default Loans (8.40%)					
Age	566	37	7	20	59
Number of Children		1.95	1.26	0	6
Loan Amount		170,000,000	124,000,000	25,000,000	800,000,000
Monthly Payment		1,200,000	920,000	150,000	8,200,000
Income		5,300,000	5,000,000	750,000	51,000,000
Debt Ratio		0.2475483	0.07	0.07	0.5
Book Value		282,000,000	240,000,000	41,000,000	2,600,000,000
LTV		0.64	0.15	0.14	0.844
Selling Price		280,000,000	250,000,000	39,000,000	2,600,000,000
Loan to Market value		0.64	0.15	0.14	1
Interest Rate		5.24	0.9	1.63	6
Loan Tenor		20	4	7	30

3.2.2. Categorical Variables

Categorical variables are: Gender, Marital Status, Existing of guarantee, Existing of children, job type, job industry, country, property location, loan type.

Table 3. Summary Table Categorical Variables

Discrete Independent Variable		Freq.	Percent	Cum.
Gender	FEMALE	377	5.59	5.59
	MALE	669	9.92	15.51
	MF	5,697	84.49	100
Marital Status	Divorced	225	3.34	3.34
	Married	5,697	84.49	87.82
	Single	776	11.51	99.33
	Widow	45	0.67	100
Additional Guarantee	NO	6,255	92.76	92.76
	YES	488	7.24	100
Job category	Employee	5,179	76.81	76.81
	Freelance	846	12.55	89.35
	Self-employed	718	10.65	100
Economic Sector	Banking	456	6.76	6.76
	Commercial	382	5.67	12.43
	Construction	72	1.07	13.5
	Industrial	74	1.1	14.59
	Public	605	8.97	23.57
	Service	5,154	76.43	100
Country	Expatriate	1,349	20.01	20.01
	Lebanon	5,394	79.99	100

Property Location	Beirut	660	9.79	9.79
	Bekaa	289	4.29	14.07
	Mount Lebanon	4,760	70.59	84.67
	North	603	8.94	93.61
	South	431	6.39	100
Loan Type	Construction	430	6.38	6.38
	Purchase	6,183	91.7	98.07
	Renovation	65	0.96	99.04
	Under Construction	65	0.96	100

Comparing defaulting to non-defaulting borrowers, data revealed that 91.60% of the lending portfolio is non-default loan borrowers and 8.40% are default borrowers. In addition, defaulted borrowers are slightly older than non-defaulted ones and have more dependents, and were granted less loan amounts than those of non-defaulting borrowers. In addition, the mean of monthly installment of defaulted borrowers is less than the one associated with performing one. Furthermore, in comparison with that of non-defaulted borrowers, defaulted borrowers have less income.

Debt ratio which reflects the portion of monthly installment out of the monthly income is higher in non-defaulted borrowers. In addition, the book value of the mortgage subject of the loan which is based on the Bank's estimated value is lower in defaulted loans in comparison to non-defaulted ones. Since the book value of the mortgage subject of the loan for defaulted loans is lower than those of non-defaulted, the LTV is higher in non-performing loans.

Housing prices of defaulted borrowers are lower than those of non-defaulted borrowers. In addition, the average mortgage price of performing borrowers is higher with comparison to those of default ones, and since their associated loan to market value is higher, this suggests that defaulted borrowers granted loan amounts are lower than those of performed borrowers.

The Average loan tenor is the same for both defaulted and non-defaulted loans. Married borrowers constitute the highest percentage of default clients with 83.57% followed by single 358 borrowers with 10.25%. In addition, there is 90.64% of defaulted borrowers do not have backed-up additional guarantees other than the main guarantee which is the mortgage subject of the housing loan.

Moreover, most of defaulted borrowers are categorized as employees' workers which constitute to 64.5% of the total number of defaulted borrowers. Most defaulted borrowers work in the service sector (71.38%) followed by those who work in the commercial sector with 10.95% out of the total number of defaulted borrowers. There is 13.25% of defaulted borrowers work outside Lebanon. There is around 70% of defaulted borrower properties are located in Mount Lebanon area and 89.22% of defaulted loans are granted for purchase purposes.

3.3. Model Development

3.3.1. Run single predictor regression models: drop predictors with significance levels > 0.25

Table 4. Single Predictors Regression Models

Predictor	Significance of Wald Z	Remark
Age	0.356	drop
No. of Children	0.002	Consider further – p-value is < .25
Loan Amount	0	Consider further – p-value is < .25
Monthly Payment	0	Consider further – p-value is < .25
Income	0.001	Consider further – p-value is < .25
Debt Ratio	0.003	Consider further – p-value is < .25
Book Value	0	Consider further – p-value is < .25
LTV	0.101	Consider further – p-value is < .25
Market price	0	Consider further – p-value is < .25
LTMP	0.234	Consider further – p-value is < .25
Interest Rate	0	Consider further – p-value is < .25
Loan Tenor	0.898	drop

Predictor	Significance of Wald Z	Remark
Gender		
M	0.426	drop
MF	0.814	drop
Marital Status		
Married	0.087	Consider further – pvalue is < .25
Single	0.054	Consider further – pvalue is < .25
Widow	0.129	Consider further – pvalue is < .25
Additional Guarantee		
Yes	0.042	Consider further – pvalue is < .25
Job Category		
Freelance	0.007	Consider further – pvalue is < .25
Self-employed	0	Consider further – pvalue is < .25
Economical Sector		
Commercial	0	Consider further – p-value is < .25
Construction	0.26	Consider further – p-value is < .25
Industrial	0	Consider further – pvalue is < .25
Public	0.011	Consider further – p-value is < .25
Service	0.033	Consider further – pvalue is < .25
Country		
Lebanon	0	Consider further – pvalue is < .25
Mohafaza		
Bekaa	0.112	Consider further – pvalue is < .25
Mount Lebanon	0.538	drop
North	0.564	drop
South	0.318	drop
Loan Type		
Purchase	0.013	Consider further – p-value is < .25
Renovation	0.061	Consider further – p-value is < .25
Under construction	0.185	Consider further – pvalue is < .25

Following the selection criteria, the following variables are not significant since their p-values are greater than 0.25. These variables are Age, Gender, and loan tenor. Note that for this stage, we will keep the property location since the Bekaa area is significant and has p-vales less than 0.25. After dropping parameters that are associated with p-values > 0.25, we run a logistic regression with predictors of p-values <0.25 and refer to it as the Full Model

Full Model: The likelihood and deviance are calculated below:

(-2) log Likelihood = (-2) (-1841.7963) = 3683.58
Deviance d f = number of observation – number of predictors = 6743 - (26) = 6717

3.3.2. Test for multi-collinearity:

Collinearity occurs when the predictors are themselves interrelated with each other's. If extreme, this is a problem for at least 2 reasons:

1) the model is unstable; and/or
2) it is uninterpretable.

Multi-collinearity problem is suggested if VIF > 10 or Tolerance < .10

Collinearity Diagnostics is performed for this study after examining the 20 variables (Age Number of Children Monthly Payment Income Debt Ratio LTV Selling Price Loan to Market Price Ratio Interest Rate Loan Tenor Gender _code Marital Status _code Additional Guarantee _code Job category _code Economical Sector _code Country _code Mohafaza _code Loan Type _code).

The result above concludes that there are three variables which are Loan amount, Monthly payment, and Book value have VIF greater than 10 and are associated with a Tolerance value < .10. Consequently, we perform more diagnoses to check which of these variables have a higher impact. As per the results, we will drop variables Loan amount and Book value. After performing a Collinearity diagnosis, and dropping variables associated with a p-value greater than 0.25, the dropping variables are: Age, gender, loan tenor, loan amount, and book value.

3.3.3. Run multiple remaining predictors and drop predictors with p-value > 0.1

Table 5. Multiple Regression Analysis

Predictor	Significance of Wald Z	Remark
Number of Children	0.02	Consider further – p-value is < .10
Monthly Payment	0.001	Consider further – p value is < .10
Income	0.336	Drop
Debt Ratio	0.871	Drop
LTV	0	Consider further – p value is < .10
Selling Price	0.583	Drop
Loan to Market Price Ratio	0.003	Consider further – p-value is < .10
Interest Rate	0	Consider further – p-value is < .10
Marital Status		
Married	0.073	Consider further – p-value is < .10
Single	0.448	Drop
Widow	0.179	Drop
Additional Guarantee		
YES	0.186	Drop
Job category		
Freelance	0.007	Consider further – p-value is < .10
Self-employed	0	Consider further – p-value is < .10
Economical Sector		
Commercial Sector	0.009	Consider further – p-value is < .10
Construction Sector	0.994	Drop
Industrial Sector	0.006	Consider further – p-value is < .10
Public Sector	0.026	Consider further – p-value is < .10
Service Sector	0.215	Drop
Country		
LEBANON	0.139	Drop
Mohafaza		
Bekaa	0.921	Drop
Mount Lebanon	0.778	Drop
North Lebanon	0.809	Drop
South Lebanon	0.983	Drop
Loan Type		
Purchase	0.212	Drop
Renovation	0.081	Consider further – p-value is < .10
Under Construction	0.169	Drop

In step three, the following variables are not significant since their P values are greater than 0.10. These variables are: Income, Debt Ratio, Selling Price, Additional Guarantee, Country, and Mohafaza (property location). We will next run logit regression after dropping these six variables.

Reduced Model: Run logit regression to find the relationship between defaulted loans and the following remaining predictors: Number of Children, Monthly Payment, LTV, Loan Market Price Ratio, Interest Rate, Marital Status, Job Category, Economical Sector, and Loan Type. The result is presented below:

Table 6. Regression Full Model Predictors

Logistic regression				Number of obs	=	6,743
				LR chi2(18)	=	199.16
				Prob > chi2	=	0
Log likelihood = -1844.3293				Pseudo R2	=	0.0512
Loan Status	**Coefficient**	**Std. err.**	**z**	**P>z**	**[95% conf.**	**interval]**
Number of Children	0.104183	0.0434807	2.4	0.017	0.0189624	0.1894036
Monthly Payment	-3.67E-07	6.07E-08	-6.04	0	-4.85E-07	-2.48E-07
LTV	2.287339	0.5318519	4.3	0	1.244928	3.329749
LTMP Ratio	-2.008557	0.5997135	-3.35	0.001	-3.183974	-0.8331406
Interest Rate	0.210041	0.0549681	3.82	0	0.1023055	0.3177765
Marital Status						
Married	-0.4217655	0.2209292	-1.91	0.056	-0.854787	0.0112477
Single	-0.2213838	0.2634957	-0.84	0.401	-0.737258	0.2950582
Widow	0.600315	0.4391309	1.37	0.172	-0.260658	1.460996
Job category						
Freelance	0.419258	0.1355891	3.09	0.002	0.1535083	0.6850078
Self-employed	0.9437229	0.1277156	7.39	0	0.693405	1.194041
Economical Sector						
Commercial Sector	0.7148428	0.268354	2.66	0.008	0.1888787	1.240807
Construction Sector	-0.0039821	0.4887424	-0.01	0.993	-0.961995	0.9539353
Industrial Sector	1.006652	0.3687148	2.73	0.006	0.283984	1.729319
Public Sector	0.5975713	0.2627259	2.27	0.023	0.0826381	1.112505
Service Sector	0.2542959	0.2244384	1.13	0.257	-0.185952	0.6941871
Loan Type						
Purchase	-0.2495123	0.178292	-1.4	0.162	-0.589582	0.0999336
Renovation	-1.250974	0.7425358	-1.68	0.092	-2.76317	0.2043694
Under Construction	0.5290428	0.3773	1.4	0.161	-0.214516	1.268537

Reduced Model Likelihood and Deviance Value:(-2) ln L = (-2) (-1844.3293) = 3688.64

Deviance d f = 6743-(18) = 6725

Likelihood ratio test comparing the above two regression models manual calculation:

LR Test = [(-2) ln (L) REDUCED] - [(-2) ln (L) FULL] = = 3688.64 - 3683.58 = 5.06
LR Test d f = change Deviance d f = change in numbers predictors in model = 6725 - 6717= 8
p-value = Pr {Chi square with 8 degree of freedom > 5.06} = 0.7511

Results: This is not significant. Possibly, we can drop 9 variables Age, Gender, Loan Tenor, Income, Debt Ratio, Selling Price, Additional Guarantee, Country, and Mohafaza.

Likelihood ratio test comparing REDUCED and Full Model Using STATA Software:

To perform the Likelihood test ratio using Stata software, first, we find a REDUCED model using the command quietly: to suppress output. The reduced model will be produced using the variables: Number of Children, Monthly Payment, LTV, Loan to Market Price Ratio, Interest Rate, Marital Status, Job Category, Economical Sector, and Loan Type. Second, we find the FULL model using command quietly using the following variables: Number of Children, Monthly Payment, Income, Debt Ratio, LTV, Selling Price, Loan to Market Price Ratio, Interest Rate, Marital Status, Additional Guarantee, Job Category, Economical Sector, Country, Mohafaza, and Loan Type. Obtain the LR test using Stata command

Likelihood-ratio LR chi2(8) = 5.07 (Assumption: reduced nested in full)	
Prob > chi2 = 0.7505 match	

Therefore, the candidate significant parameters that affect log odds of default are included in the reduced regression model which are the quantitative predictors: Number of children, Monthly Payment, Loan to Value, Loan to Market Price, Interest rate, and the categorical predictors which are: Marital Status, Job type, Job economic sector, and the housing loan type.

3.3.4. Investigate Confounding

A 'good' final model is the nine predictors model mentioned above. However, we need to explore possible confounding of the nine predictor models by the omitted variables Age, Gender, Loan Tenor, Income, Debt Ratio, Selling Price, Additional Guarantee, Country, and Mohafaza. We will assess these variables as a potential confounder using 2 criteria:

1. Likelihood Ratio test < .10
2. Relative Change in estimated betas > 15% using the following formula:

Where change in estimated betas: $\Delta\hat{\beta} = \frac{|\hat{\beta}\text{without confounders} - \hat{\beta}\text{with confounders}|}{\hat{\beta}\text{with confounders}}$ x 100

Run regression of candidate final model:

Table 7. Candidate Final Model

Loan Status	Coefficient	Std. err.	Z	P>z	[95% conf.	interval]
Number of Children	0.104183	0.0434807	2.4	0.017	0.0189624	0.1894036
Monthly Payment	-3.67E-07	6.07E-08	-6.04	0	-4.85E-07	-2.48E-07
LTV	2.287339	0.5318519	4.3	0	1.244928	3.329749
Loan to Market price Ratio	-2.008557	0.5997135	-3.35	0.001	-3.183974	-0.8331406
Interest Rate	0.210041	0.0549681	3.82	0	0.1023055	0.3177765
Marital Status						
Married	-0.4217655	0.2209292	-1.91	0.056	-0.8547787	0.0112477
Single	-0.2213838	0.2634957	-0.84	0.401	-0.7378258	0.2950582
Widow	0.600315	0.4391309	1.37	0.172	-0.2603658	1.460996
Job category						
Freelance	0.419258	0.1355891	3.09	0.002	0.1535083	0.6850078
Self-employed	0.9437229	0.1277156	7.39	0	0.693405	1.194041
Economical Sector						
Commercial Sector	0.7148428	0.268354	2.66	0.008	0.1888787	1.240807
Construction Sector	-0.0039821	0.4887424	-0.01	0.993	-0.9618995	0.9539353
Industrial Sector	1.006652	0.3687148	2.73	0.006	0.283984	1.729319
Public Sector	0.5975713	0.2627259	2.27	0.023	0.0826381	1.112505
Service Sector	0.2542959	0.2244384	1.13	0.257	-0.1855952	0.6941871
Loan Type						
Purchase	-0.2495123	0.178292	-1.4	0.162	-0.5989582	0.0999336
Renovation	-1.250974	0.7425358	-1.68	0.092	-2.706317	0.2043694
Under Construction	0.5290428	0.3773	1.4	0.161	-0.2104516	1.268537

Table 8. Run Regression full model which includes predictors of candidate fitted model in addition to omitted predictors:

Loan Status	Coefficient	Std. err.	Z	P>z	[95% conf.	interval]
Number of Children	0.1018338	0.0436941	2.33	0.02	0.0161949	0.1874728
Monthly Payment	-5.00E-07	1.57E-07	-3.19	0.001	-8.08E-07	-1.93E-07
Income	2.62E-08	2.72E-08	0.96	0.336	-2.71E-08	7.96E-08
Debt Ratio	0.1758393	1.086102	0.16	0.871	-1.952882	2.304561
LTV	2.234713	0.5389621	4.15	0	1.178367	3.291059
Selling Price	1.55E-10	2.83E-10	0.55	0.583	-3.99E-10	7.10E-10
LTMPR	-1.966337	0.6505957	-3.02	0.003	-3.241481	-0.6911934
Interest Rate	0.2067166	0.056264	3.67	0	0.0964412	0.316992
Marital Status						
Married	-0.3979486	0.2216966	-1.8	0.073	-0.832466	0.0365688
Single	-0.2004013	0.2643156	-0.76	0.448	-0.7184503	0.3176478
Widow	0.5903815	0.4394251	1.34	0.179	-0.2708759	1.451639
Additional Guarantee						
YES	0.2100846	0.1588032	1.32	0.186	-0.1011639	0.5213331
Job category						
Freelance	0.3748148	0.1394592	2.69	0.007	0.1014798	0.6481497
Self-employed	0.8757572	0.1338527	6.54	0	0.6134108	1.138104
Economical Sector_						
Commercial Sector	0.7050568	0.2698374	2.61	0.009	0.1761851	1.233928
Construction Sector	0.0038314	0.4904956	0.01	0.994	-0.9575222	0.9651851
Industrial Sector	1.010965	0.3699979	2.73	0.006	0.2857828	1.736148
Public Sector	0.5858204	0.263109	2.23	0.026	0.0701361	1.101505
Service Sector	0.2797729	0.2258699	1.24	0.215	-0.1629239	0.7224697
Country						
LEBANON	0.2168795	0.1464982	1.48	0.139	-0.0702517	0.5040107
Mohafaza						
Bekaa	0.0251559	0.2535913	0.1	0.921	-0.471874	0.5221857
Mount Lebanon	-0.0458258	0.1629166	-0.28	0.778	-0.3651364	0.2734849
North Lebanon	-0.0527893	0.2182368	-0.24	0.809	-0.4805257	0.374947
South Lebanon	-0.0050185	0.2346429	-0.02	0.983	-0.4649102	0.4548731
Loan Type						
Purchase	-0.2294727	0.1840196	-1.25	0.212	-0.5901445	0.1311991
Renovation	-1.298583	0.7437918	-1.75	0.081	-2.756388	0.1592224
Under Construction	0.5245479	0.381181	1.38	0.169	-0.2225531	1.271649
_cons	-3.495022	0.615263	-5.68	0	-4.700915	-2.289128

Next, we need to check for a greater than 15% Change in Betas for Predictors in the Model. Potential confounding of predictors: Number of Children, Monthly Payment, LTV, Loan to Market Price Ratio, Interest Rate, Marital Status, Job Category, Economical Sector, Loan Type **By:** Age, Gender, Loan Tenor, Income, Debt Ratio, Selling Price, Additional Guarantee, Country, Mohafaza.

Table 9. Confounding Predictors Coefficient Change in Beta

Predictors	Coefficient without confounding	Coefficient with confounding	Change in Beta's	Change
number of children	0.104183	0.1018338	0.023069	2.306896
Monthly Payment	-3.67E-07	-5.00E-07	-0.266	-26.6
LTV	2.287339	2.234713	0.023549	2.354933

Predictors	Coefficient without confounding	Coefficient with confounding	Change in Beta's	Change
LTMP ratio	-2.008557	-1.966337	0.021471	2.147139
Interest Rate	0.210041	0.2067166	0.016081	1.608192
MaritalStatus_code				
Married	-0.4217655	-0.3979486	0.059849	5.984918
Single	-0.2213838	-0.2004013	0.104702	10.47024
Widow	0.600315	0.5903815	0.016825	1.682556
Jobcategory_code				
Freelance	0.419258	0.3748148	0.118573	11.85737
Self-employed	0.9437229	0.8757572	0.077679	7.760726
EconomicalSector_code				
Commercial Sector	0.7148428	0.7050568	0.01397	1.79733
Construction Sector	-0.0039821	0.0038314	-2.039329	-203.9333
Industrial Sector	1.006652	1.010965	-0.002662	-0.426622
Public Sector	0.5975713	0.5858204	0.020589	2.0058878
Service Sector	0.2542959	0.2797729	-0.0910631	-9.106314
LoanType_code				
Purchase	-0.2495123	-0.2294727	0.03289	8.7328907
Renovation	-1.250974	-1.298583	-0.036623	-3.666227
Under Construction	0.5290428	0.5245479	0.0085691	0.8569093

Results: The relative change in the beta's in the good model are less than 15.6%. Therefore, greater chances are there for candidate model to be the best fitted model. Next, we need next to investigate about interaction between predictors of the candidate model.

3.3.5. Investigate Effect Modification

Are individuals who are both unemployed and with low income more likely to be defaulted? For this illustration, we will create a new variable called 'low' to capture borrowers whose monthly payment is less than or equal to 900.000 LBP. Then we will create an interaction of lowpay and Job categories. Then we run a regression model that includes the main effects of both of the variables contributing to the interaction. Thus, this model includes the main effects of low-pay and in addition to the interaction low-pay _Job category. Next, we generate a new variable labeled low-pay where it is a Monthly installment associated with an amount less than or equal to 900.000 LBP

Table 10. Summary Statistics for Variable lowpay

lowpay new variable		**Monthly Payment**	**Freq.**	**Percent**	**Valid**	**Cum.**
0	Other	> 900.000 LBP	4550	67.47	67.48	67.48
1	Lowpay	<=900.000 LBP	2193	32.52	32.52	100
		Total	6743	99.99	100	

We are interested to generate an interaction variable for borrowers those who work as employees and their loan monthly payment less than or equal to 900.000 LBP.We generate new variable low pay _ Job category so that borrowers are assigned zero if their monthly payment greater than 900,000 LBP and those who monthly payment less than or equal to 900,000 to assigned 1 for employees, 2 for freelancer, and 3 for self-employed.

Table 11. Interaction Variable

	lowpay_Jobcategory	Freq.	Percent	Valid	Cum.
0	Other	4550	67.47	67.48	67.48
1	employee and lowpay	1853	27.48	27.48	94.96
2	Freelancer	184	2.73	2.73	97.69
3	Self-employed	156	2.31	2.31	100
	Total	6743	99.99	100	

Table 12. Descriptive Statistics for Interactive Variable

Job category	other	Monthly pay <=900.000	Total
Employee	3,326	1,853	5,179
Freelance	662	184	846
Self-employed	562	156	718
Total	4,550	2193	6,743

We can conclude from the above figure that 1853 borrowers are both employees and have loan monthly payments less than or equal to 900.000 LBP. Next, we need to perform a likelihood ratio test to check the significance of the model after adding a new variable with the new variable interaction inserted into the candidate's final model. Therefore, we will add to the final model an additional variable (low-pay) and the interaction variable (lowpay_jobcategory) and check its significance. After running the logistic regression including the variables: Number of Children, Monthly Payment, LTV, Loan to Market Price Ratio, Interest Rate, Marital Status, Job Category, Economical Sector, Loan Type, lowpay, lowpay _ Jobcategory, the following results are presented below:

Table 13. Regression Model Including Interaction Variable

Iteration 0: log-likelihood	=	-1943.9103				
Iteration 1: log-likelihood	=	-1864.8602				
Iteration 2: log likelihood	=	-1835.6704				
Iteration 3: log-likelihood	=	-1835.5781				
Iteration 4: log-likelihood	=	-1835.578				
Logistic regression			Number of obs	=	6,743	
			LR chi2(20)	=	216.66	
			Prob > chi2	=	0	
Log likelihood = -1835.578			Pseudo R2	=	0.0557	
Loan Status	**Coefficient**	**Std. err.**	**z**	**P>z**	**[95% conf.**	**interval]**
Number of Children	0.1078049	0.0435149	2.48	0.013	0.02251	0.1930
Monthly Payment	-1.54E-07	7.39E-08	-2.09	0.037	-2.99E-07	-9.40E-09
LTV	2.335896	0.5324232	4.39	0	1.292365	3.379426
Loan to Market price Ratio	-1.94	0.599954	-3.23	0.001	-3.115888	-0.7641
Interest Rate	0.1909533	0.0546323	3.5	0	0.083876	0.2980305
Marital Status _code						
Married	-0.4329017	0.2214756	-1.95	0.051	-0.866986	0.0011826
Single	-0.2225849	0.2640039	-0.84	0.399	-0.7400	0.2948533
Widow	0.5780773	0.4380881	1.32	0.187	-0.2805	1.436714
Job category _code						
Freelance	0.3788648	0.1452397	2.61	0.009	0.0942003	0.6635293
Self-employed	0.8520233	0.1580499	5.39	0	0.5422511	1.161795
Economical Sector _code						
Commercial Sector	0.7158771	0.2689094	2.66	0.008	0.1888244	1.24293
Construction Sector	-0.0677168	0.4903858	-0.14	0.89	-1.028855	0.8934217
Industrial Sector	0.9659601	0.3700067	2.61	0.009	0.2407602	1.69116
Public Sector	0.5429135	0.2641617	2.06	0.04	0.0251661	1.060661
Service Sector	0.2574801	0.2246614	1.15	0.252	-0.18284	0.6978083
Loan Type _code						
Purchase	-0.2337222	0.1785744	-1.31	0.191	-0.5837	0.1162771
Renovation	-1.226884	0.7454903	-1.65	0.1	-2.688	0.2342504
Under Construction	0.4800383	0.3775808	1.27	0.204	-0.2600	1.220083
Lowpay	0.3361596	0.2117022	1.59	0.112	-0.0787	0.7510884
Lowpay _Jobcategory	0.1269765	0.1210548	1.05	0.294	-0.1102	0.3642395
_cons	-3.692678	0.5161525	-7.15	0	-4.7043	-2.681038

Performing likelihood ratio test of interaction variable: Adding the low pay variable and the interaction variable lowpay_jobcategory to the candidate final model and assume it as **Full model**. This model will include the variables: Number of Children, Monthly Payment, LTV, Loan to Market Price Ratio, Interest Rate, Marital Status, Job category, Economical Sector, Loan Type, low pay, lowpay _ Jobcategory.

Adding the low pay variable to the candidate final model and assume it as **reduced model**. This model will include the variables: Number of Children, Monthly Payment, LTV, Loan to Market Price Ratio, Interest Rate, Marital Status, Job category, Economical Sector, Loan Type, low pay, run quietly regression to both full and reduced model and conduct likelihood ratio test for both models (lrtest reduced full).

Likelihood-Ratio Test: Assumption: reduced nested within full model LR chi2 (1) = 1.09; Prob > chi2 = 0.2955

Results: Not significant so we drop the induced variable. Therefore, as a conclusion: A reasonable multiple predictor model of default in this study contains the following predictors: Number of Children, Monthly Payment, LTV, Loan to Market Price Ratio, Interest Rate, Marital Status, Job category, Economical Sector, Loan Type.

Let's fit the final model one more time, in two ways: (1) using the command logit to obtain the prediction equation and (2) using the command logistic to obtain odds ratios instead of betas.

Table 14. Best Fitted Model

Loan Status	Coefficient	Std. err.	z	P>z	[95% conf.	interval]
Number of Children	0.104183	0.0434807	2.4	0.017	0.01896	0.1894036
Monthly Payment	-3.67E-07	6.07E-08	-6.04	0	-4.85E-0	-2.48E-07
LTV	2.287339	0.5318519	4.3	0	1.24492	3.329749
LTMP Ratio	-2.008557	0.5997135	-3.35	0.001	-3.183974	-0.8331406
Interest Rate	0.210041	0.0549681	3.82	0	0.1023055	0.3177765
Marital Status						
Married	-0.4217655	0.2209292	-1.91	0.056	-0.8547787	0.0112477
Single	-0.2213838	0.2634957	-0.84	0.401	-0.7378258	0.2950582
Widow	0.600315	0.4391309	1.37	0.172	-0.2603658	1.460996
Job category						
Freelance	0.419258	0.1355891	3.09	0.002	0.1535083	0.6850078
Self-employed	0.9437229	0.1277156	7.39	0	0.693405	1.194041
Economical Sector						
Commercial Sector	0.7148428	0.268354	2.66	0.008	0.1888787	1.240807
Construction Sector	-0.0039821	0.4887424	-0.01	0.993	-0.9618995	0.9539353
Industrial Sector	1.006652	0.3687148	2.73	0.006	0.283984	1.729319
Public Sector	0.5975713	0.2627259	2.27	0.023	0.0826381	1.112505
Service Sector	0.2542959	0.2244384	1.13	0.257	-0.1855952	0.6941871
Loan Type						
Purchase	-0.2495123	0.178292	-1.4	0.162	-0.5989582	0.0999336
Renovation	-1.250974	0.7425358	-1.68	0.092	-2.706317	0.2043694
Under Construction	0.5290428	0.3773	1.4	0.161	-0.2104516	1.268537
_cons	-3.256975	0.505895	-6.44	0	-4.248511	-2.265439

3.3.6. Model Equation:

Logit {Pr[default=1]} = - 3.25 + 0.10 X Number of Children – 0.000000367 X Monthly Payment + 2.28 X LTV - 2 X Loan to Market Price Ratio + 0.21X Interest Rate - 0.42 X (married borrower) + 0.42 X (freelance borrower) +0.94 x (self-employed borrower) + 0.7148428 X (Borrower works in Commercial Sector) + 1 X (Borrower works in industrial Sector) + 0.25 X (Borrower works in Public Sector) -1.25 X (renovation loans)

3.4. Model Diagnosis

3.4.1. The Hosmer-Lemeshow Test of Goodness-of-Fit

The Hosmer-Lemeshow Goodness of Fit Test compares observed versus predicted counts of outcome events in each of several 'meaningful' subgroups of the data, like the Chi-Square Goodness of Fit. If the fit is good (the null hypothesis is true), the observed and (model-based) expected counts will be close and their differences will be small. The actual test statistic is a sum of (observed – expected)/expected2 and is distributed chi square under the null hypothesis.
Hosmer-Lemeshow Goodness of Fit Test
HO: The current model is a 'good' fit for the data.
HA: not.
Rejection occurs for large values of the chi-square statistic with associated small p-value

3.4.2. Goodness-of-fit Test after Logistic Model

Test Results: The sample number of observation is 6743 associated with eight number of groups with Hosmer-Lemeshow chi2(6) is 5.61, The Hosmer_Lemeshow test (p=0.4686) suggests no statistically significant departure from a good fit. The null hypothesis of 'good fit' is NOT rejected.

3.4.3. The Link Test (check if there is another model that best fits the data)

The Link Test is a simple check of the fitted model. It assesses whether or not the fitted model is an adequate fit (null hypothesis) to the data or, if not if there is still some additional modeling that needs to be done (alternative hypothesis).
HO: The current model is an adequate fit for the data.
HA: Alternative modeling is needed.
A Likelihood Ratio (LR) Test is performed and compares a 'null hypothesis' adequate model (reduced) with an 'alternative hypothesis enhanced (full) model:

Reduced: logit[π] = β (0) + β (1) [$\hat{\pi}$ model]
Full: logit[π] = β (0) + β (1) [$\hat{\pi}$ model] + β (2) [$\widehat{\pi 2}$ model]
H0: β2 = 0
H 1 = not
$\hat{\pi}$: This is the predicted probability from our model; we hope this is significant.
π ˆ2: If the null is true (the model is adequate), this should be non-significant
Thus, Rejection of the null occurs for large values of the LR Test and associated small p-values.

Table 16. Table for Link test

Link test						
Log likelihood = -1844.049			Pseudo R2	=	0.0514	
Loan Status	Coefficient	Std. err.	Z	P>z	[95% conf.	interval]
_hat	1.267682	0.3613615	3.51	0	0.5594264	1.975937
_hatsq	0.0619295	0.0818108	0.76	0.449	-0.0984166	0.2222757
_cons	0.2640519	0.3867142	0.68	0.495	-0.4938941	1.021998

_hat = $\hat{\pi}$ model : This is marginally significant (p=.0)
_hatsq = πˆ2 : This is non-significant (p=0.449)
Test Results: The Link Test (p=0.449) suggests no statistically significant departure from model adequacy. The null hypothesis of 'model adequacy' is NOT rejected.

3.4.4. The Classification Table

The classification table describes the predicted number of successes compared with the observed number of successes. Likewise, it compares the predicted number of failures with the actual number of failures observed. Stata software by default chooses a threshold probability for an event as 0.5. This probability can be amended when needed. In this study, we use the default cutoff which is 0.4

Table 17. Classification Table

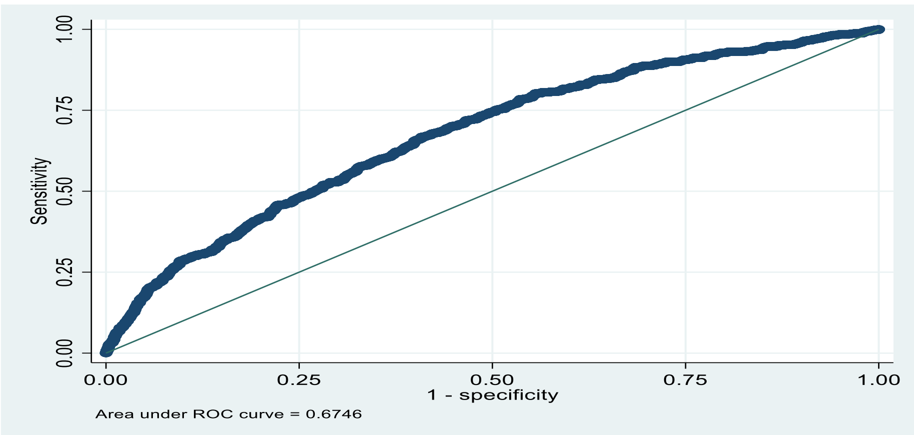
		TRUE		
Classified	D	~D	Total	
+	2	2	4	
-	564	6175	6739	
Total	566	6177	6743	
Classified + if predicted Pr(D) >=	0.4			
True D defined as Loan Status !=	0			
Sensitivity		Pr(+ D)	0.35%	
Specificity		Pr(-~D)	99.97%	
Positive predictive value		Pr(D +)	50.00%	
Negative predictive value		Pr(~D -)	91.63%	
False + rate for true ~D		Pr(+~D)	0.03%	
False - rate for true D		Pr(- D)	99.65%	
False + rate for classified	+	Pr(~D +)	50.00%	
False - rate for classified	-	Pr(D -)	8.37%	
Correctly classified				**91.61%**

Manual Calculation implies: (True Negative + True positive) / Total Sample Size = (2+ 6175) / 6743 = 91.61 % **match**

3.4.5. The ROC Curve and The Pregibon Delta Beta Statistic

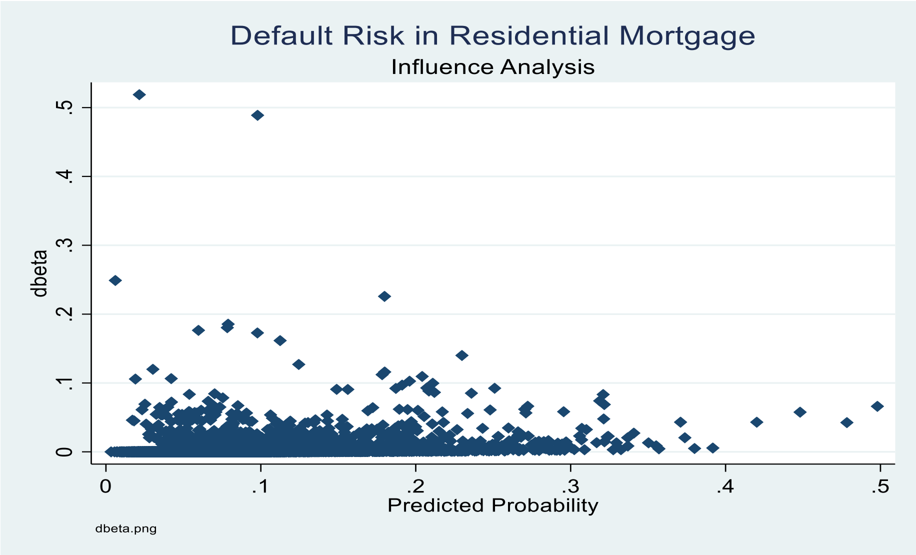
A ROC curve, short for Receiver-Operating Characteristic, serves as a graphical representation showcasing the overall effectiveness of a logistic model and the corresponding equation for predicted probabilities. This visualization accounts for two types of prediction errors: (1) incorrectly predicting a true event as a non-event (false negative), and (2) incorrectly predicting a true non-event as an event (false positive, equivalent to 1 - specificity). For different selected ‘cut-off’ values (where a predicted probability is designated as a predicted event), the ROC curve illustrates the relationship between false positive (X-axis) and true positive (Y-axis) values across various ‘cut-off’ choices. The ROC Curve c-statistic is equal to the overall percentage correctly classified which is reflected by the area under the curve.

Figure 2. ROC Curve



The straight line with slope =1 is a reference line; it corresponds to the ROC curve where chance alone is operating (coin toss with probability heads = .50). The ROC c-statistic = 0.6746 says that the overall percentage who are correctly classified is 67.46% close to 70%.

Figure 3. The Pregibon Delta Beta Statistic



Results: The dbeta values are all less than 0.25, suggesting the absence of influential points.

Conclusion and Further Research

This study empirically examines the impact of borrower, loan, and mortgage parameters on default risk in residential loans. Using 6743 individual housing loan accounts data for the selected period from 2005 to 2020 from Housing Finance Institutions in Lebanon, we use the multivariable logistic regression model, best subset logistic regression model, and stepwise regression analysis procedures to investigate the impact of 21 predictors on log odds of default risk. The empirical results suggest the following: the estimated probability of defaulting on a housing loan is approximately 3.8% when all predicted variables are set at their lowest value. In addition, gender, the existence of additional guarantees, geographical location, property location, age, loan amount, income, debt ratio, book value, selling price, and loan tenor parameters have no impact on the risk of default risk. However, the log odds of defaulting on loans increased by 10.5% for every increase in the number of dependents, decreased by approximately 0.0000367% for every one dollar increase in the monthly payment, increased by 876.1 % for one percent increase in loan to book value ratio, decrease by approximately 86.5 % for one percent increase in loan to market price ratio, and increase by approximately 23.3% for every increase of one percent of interest rate. Furthermore, for the marital status predictor, having the divorced category as a reference variable, the odds of default on loans for married and single borrowers are lower than the divorced borrower by about 34.4% and 19.9% percent respectively. However, the odds of default on loans of widow borrowers have approximately 82.2% higher log odds of default than divorced borrowers. Second, for the job category predictor, having the employed borrower category as the reference variable, the odds of default for freelance and self-employed borrowers are higher than for employee borrowers by 52.1% and 156.1 % respectively. Furthermore, for borrower's job industry, having the banking industry as the reference variable, the log odds of default on loans for borrowers working in commercial, industrial, public, and service sectors have higher log odds of default by 105.3%, 171.8%, 80.6%, 28.4% respectively. However, borrowers working in the construction industry have lower log odds of default compared to those working in the banking sector by approximately 0.395. Moreover, for the housing loan type predictor, having construction loans as a reference variable, the log odds of default for purchase and renovation loan borrowers are about 22.12% and 71.35% respectively lower than the construction loans. However, the log odds of default for under-construction loan borrowers are about 68.2 % higher than the construction loan borrowers. Furthermore, the model's overall accuracy was demonstrated by a 91.61 % visible correct classification rate.

The limitation of the study is reflected by the lack of information about borrower's credit scoring and the lending criteria as they were mentioned in many studies to have a high impact on the log odds risk of default. In addition, it is highly recommended for further work to include macroeconomic factors to examine its impact on default risk in mortgage loans. For instance, high inflation will decrease borrower income, any raise in interest rate will be associated with an increase in the monthly payment, and the high unemployment rate will affect the borrower's income and therefore the ability to pay back the loan.

Acknowledgment

I would like to thank the management team of housing bank in Lebanon who helped me in providing the data used in this research paper.

Declaration of Competing Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Use of Generative AI and AI-assisted Technologies

The author declares that he has not used generative AI and AI-assisted technologies during the preparation of this work.

References

[1] Bandyopadhyay and A. Saha (2009). *Factors driving demand and default risk in residential housing loans: Indian evidence.* MPRA Paper 14352, University Library of Munich, Germany.

[2] Bendel R. B. and Afifi A. A. (1977). Comparison of stopping rules in forward 'Stepwise' regression. *Journal of the American Statistical Association, 72*(357): 46–53. DOI: https://doi.org/10.2307/2286904

[3] Canepa and F. Khaled (2018). Housing, housing finance and credit risk. *International Journal of Financial Studies, 6*(2): 50. DOI: https://doi.org/10.3390/ijfs6020050

[4] Foote L. *et al.* (2008). Just the facts: An initial analysis of subprime's role in the housing crisis. *Journal of Housing Economics, 17*(4): 291-305. DOI: https://doi.org/10.1016/j.jhe.2008.09.001

[5] Gerardi, K. *et al.* (2008). Making sense of the subprime crisis. Brooking Papers on Economic Activity, Fall.

[6] Hosmer Jr, D. W. *et al.* (2013). *Applied logistic regression.* John Wiley and Sons.

[7] Lacour-Little, M. *et al.* (2016). What role did piggyback lending play in the housing bubble and mortgage collapse? *Journal of Housing Economics, 34:* 94-117. DOI: https://doi.org/10.1016/j.jhe.2016.08.003

[8] Levy D.S. and Lee C. Kwai-Choi (2004). The influence of family members on housing purchase decisions. *Journal of Property Investment and Finance, 22*(4): 320-338. DOI:https://doi.org/10.1108/14635780410550885

[9] Li Y.and J. Yang (2018). Default risk prediction and feature analysis: A comprehensive study. *Journal of Finance, 72*(3): 456-473.

[10] Mayer, Christopher *et al.* (2009). The rise in mortgage defaults. *Journal of Economic Perspectives, 23* (1): 27-50. DOI: https://doi.org/10.1257/jep.23.1.27

[11] Mickey, R. M. and S. Greenland (1989). The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology, 129*(1): 125-137. DOI: https://doi.org/10.1093/oxfordjournals.aje.a115101

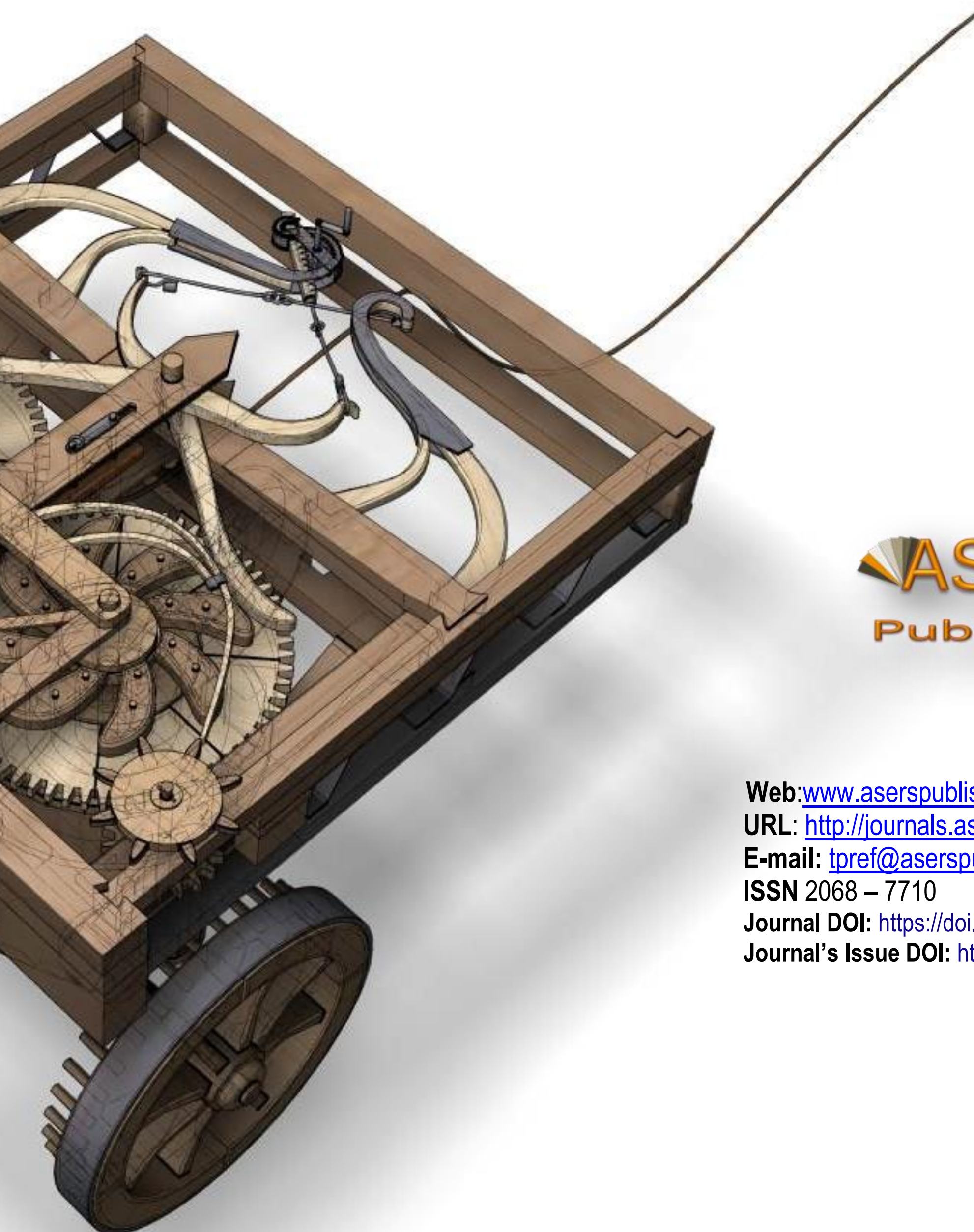
[12] Moore *et al.* (2018). *Introduction to the practice of statistics.* Publisher: W. H. Freeman Publication.

[13] Neyman, J. (2023). *Unit 7: Logistic regression.* Available at: https://people.umass.edu/biep640w/pdf/7.%20%20Logistic%20Regression%202023.pdf

[14] Rosengren E. S. (2008). The mortgage meltdown - Implications for credit availability. Discussion presented at the U.S. Monetary Policy Forum, New York, NY. Federal Reserve Bank of Boston. Available at: https://fraser.stlouisfed.org/title/statements-speeches-eric-rosengren-9015/mortgage-meltdown-implications-credit-availability-667430

[15] Sandar, T.*et al.* (2010). Factor affecting the borrowers' repayment performance of microfinance in sub-urban areas of Yangon, Myanmar. *Journal of Economics, 14*(2): 125.

ASERS



ASERS Publishing

Web:www.aserspublishing.eu
URL: http://journals.aserspublishing.eu/tpref
E-mail: tpref@aserspublishing.eu
ISSN 2068 – 7710
Journal DOI: https://doi.org/10.14505/tpref
Journal's Issue DOI: https://doi.org/10.14505/tpref.v15.2(30).00