Application of Logistic Regression in Enhancing Digital Credit Risk Management in Commercial Banks

January 11, 2025

Abstract: For a long time, credit risk has been and so remains, the main cause of concern among lenders, whether digital or conventional. The greatest among bank officers' key performance indicators is effective management of the quality of loan portfolio. The officers are therefore required at all times, to be vigilant and get to know which borrower will fail to honor the contract by failing to repay the borrowed amounts as and when it falls due. Failure by the borrower to repay the loan results in a credit risk where the bank writes down the value of its assets through impairment. At the same time, the bank's revenue diminishes as a result of reduced income from the expected interests. The bank is therefore interested in designing a model that will robustly predict the probabilities of default accurately. The study aims to provide an assessment of credit risk on digital loans in commercial banks using logistic regression model.The study employed logistic regression, a widely used statistical method in credit risk modeling, due to its effectiveness in predicting binary outcomes such as loan default versus repayment, and its ability to provide easily interpretable results for key stakeholders.The model was fitted with simulated data obtained from commercial banks' digital loan portfolio. The dataset consisted of twelve independent variables representing income to loan ratio, debt to income ratio, risk taking behavior, past loan behavior, gender, employment status, credit score, multiple loan applications, application timing and the dependent variable, being the digital loans probability of default. Analysis was done using R software and the parameters estimated using Maximum Likelihood method. The results from the study showed that income to loan ratio and credit score were critical variables in predicting digital loan default. However, variables such as debt to income ratio and past loan behavior did not show significance in determining success or failure of the digital loan outcome in the model.The performance of the model was measured with two mathematical approaches. First, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was scored at 0.9579, which implies the model has been very good in predicting the defaults and the robust model in separating the defaulting borrowers from the non-defaulting ones. Second, the Hosmer-Lemeshow GOF test yielded the p value equals to 0.8475 which means that there are no significant differences between observed and predicted probability of the data and so the model is appropriate to the data.

Keywords: Digital Credit Provider, Logistic Regression, Maximum Likelihood, Loan-to-Income Ratio.

1 Introduction

Digital credit is a phenomenon that has gained popularity in Kenya and in the process revolutionized the financial system in the world. The advent of this new technology has greatly enhanced access to credit and those previously locked out of the formal financial system can now borrow easily and conveniently (Johnston, 2024, p. 12). However, with this advancement, there has come a rise in bad loans, a factor that continues to make credit risk in the commercial banks as well as other digital credit providers a critical concern (Barbiero & Dimou,2024). In this study, we employed simulated dataset consisting of 6,000 digital loan facilities to

estimate the credit risk using the logistic regression model. Consequently, the study seeks to contribute to the development of effective risk control measures, thereby increasing the quality of digital loan portfolios and the confidence of stakeholders.

For a long time, commercial banks act as intermediaries between fund suppliers and borrowers and earn revenue from loan interest (KPMG, 2022). Through the advancement of mobile banking technology, commercial banks have experienced growth in digital loan portfolio, which has been caused by development in digital loans (Chen & Faz,2015). Nevertheless, the digital lending has caused defaults especially for the subprime borrowers, which poses risks to the balance sheets of banks (Nation. Africa, 2024). Credit assessment is therefore essential in managing these risks (Fundingo, 2024) since poor credit decisions harm loan quality, cash flows and the reputation of the lending organization (Alshebmi et al., 2020).

To address these challenges, the study employed logistic regression model to assess credit risk. Logistic regression is more preferred in the credit risk evaluation because it is easy to implement and appropriate in situations where the outcome is binary such as loan defaults and repayments as seen in Das (2024) as well as Eberhardt and Breiwick (2012).

Using a dataset of 6,000 simulated records, key predictors of loan default were identified. These included the income to loan ratio and credit score, which were found to be critical determinants of a borrower's likelihood to default. Borrowers with low income relative to their loan amounts or poor credit histories were significantly more likely to default, underscoring the importance of these variables in risk evaluation. Interestingly, some of the credit risk factors as measured by the level of debt to income and past loan behavior did not produce the expected results in this case. This implies that digital lending is a different prospect all together and may not necessarily operate within credit assessment parameters. Although a clear picture emerged from the logistic regression model for these key predictors. It also highlighted some weaknesses. Even though it assumes linear relationship between variables, it does not address the timing issue of defaults, which is an important consideration in credit risk management.

Due to these shortcomings, there is a compelling need to go for more complex techniques that can capture the timing of loan default events. Survival analysis presents a good solution. Compared to logistic regression, survival analysis looks at the time to the occurrence of an event, which in this context, is loan default, and as such enables lenders not only to determine if a default is going to occur, but when specifically, the default is most likely to happen. This time to event path approach affords a better understanding of the social and temporal characteristics of credit risk and enables lenders to develop preventing and intervening strategies.

2 Literature Review

2.1 Introduction

Commercial banks rely on credit as a major form of income by means of interests, loan application and credit evaluation fees, penalties charged on late repayments as well as other charges (KPMG, 2022). Nonetheless, loan defaults present these institutions with certain inherent credit risks such as deterioration of assets, loss of interest income, high recovery costs and dilution of reputation (Alshebmi et al., 2020). To manage these risks, it is necessary for the lending institutions to put credit risk assessment models into place to scan high risk loans and loan seekers.

2.2 Logistic Regression model

Logistic regression models are employed in assessment of credit risks and are, particularly appropriate for use in case of categorical predictors. These models exclude the normality assumptions, required for other methods such as discriminant analysis. According to Anderson, R. (2007) Logistic regression model is a statistical method applied in credit risk management to predict the likelihood of a borrower defaulting on a loan. Developed by David Cox (1958) with the aim of modelling dichotomous outcomes. The method has been advanced over time and has become popular due to its simplicity in interpretation and communication of the results to the stakeholders and regulators. Recent development and application of logistic regression modelling in credit risk assessment remains work in progress with the contemporary study focusing on application of this model in assessment of credit risk in digital loans in commercial banks. Nelder, J. A., &

Wedderburn, R. W. M. (1972) extended this model through generalized linear models allowing for broader application way beyond the binary outcome.

3 MATERIAL AND METHOD

3.1 Data

The study made use of simulated dataset consisting of 6,000 digital loans, with variables representing income to loan ratio, debt to income ratio, risk taking behavior, past loan behavior, gender, employment status, credit score, multiple loan applications and application timing.

3.2 Model Development

This study utilized the multivariate logistic regression analysis that is a statistical technique used for analyzing the dependent variable with more than one independent variable. This model is particularly useful because the outcome variable (loan default, in this case) is dichotomous. Logistic regression provides a more detailed understanding of how various factors contribute to the likelihood of an event occurring and allows for adjustment of intervening variables, enabling better predictive models. The model was developed to predict the likelihood of digital loan default, subject to the following explanatory variables: The variables are X_1: Income to Loan Ratio, X_2: Debt to Income Ratio, X_3: Past Loan Behavior, X_4: Gender, X_5: Risk Taking Behavior, X_6: Credit Score, X_7: Multiple Loan Applications, X_8: Application Timing, X_9: App Usage, X_{10}: Digital Transaction History, X_{11}: Communication Patterns, X_{12}: Employment Status, and X_{13}: Location.

Default means failure by a borrower in meeting his obligations in terms of payments of the debt borrowed or failure to pay the amount owing by the agreed upon due date. In such cases, the loan must be written down on the balance sheet and a loan loss provision, an expense to the income statement, is created. This provision quantifies possible losses which arise due to defaults and thereby serves as a buffer for the lender. Provisions for impairment of loans are core business in the context of risk management and guarantee the conformity to IFRS or central banking guidelines. The practice of reserves and provisions depend on the severity of the default, internal policies, and legislation enabling lenders to retain adequate capital subordinate and secure. Let the loan default outcome be denoted as $\mathcal{Y}$, where:

$$(Y) = \begin{cases} 1 & \text{if the borrower defaults on the loan,} \\ 0 & \text{if the borrower did not default on the loan.} \end{cases}$$

$$P(Y) = \frac{1}{1 + e^{-\beta_0 - \sum_{i=1}^{n} \beta_i X_i}}, \tag{1}$$

where $P(Y)$ is the probability of default, β_0 is the intercept, and β_i are coefficients for predictor variables X_i.

3.2.1 Parameter Estimation

The parameters used in model 3.2 were estimated using Maximum Likelihood Method. The estimation was as follows: The logistic regression model is defined as follows:

$$p(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_{12} X_{12})}} \tag{2}$$

This can be simplified as:

$$p(Y = 1|X) = \frac{1}{1 + e^{-X\beta}} \tag{3}$$

where X is the vector of independent variables, and $\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_{12})^T$ is the vector of parameters to be estimated.

The likelihood function, representing the probability of observing the data given the parameter estimates, is given by:

$$L(\beta) = \prod_{i=1}^{n} \left[p(Y_i = 1|X_i)^{y_i} \left(1 - p(Y_i = 1|X_i)\right)^{1-y_i} \right] \tag{4}$$

Substituting the logistic model for $p(Y_i = 1|X_i)$, the likelihood becomes:

$$L(\beta) = \prod_{i=1}^{n} \left[\left(\frac{1}{1 + e^{-X\beta}} \right)^{y_i} \left(1 - \frac{1}{1 + e^{-X\beta}} \right)^{1-y_i} \right] \tag{5}$$

Taking the natural logarithm of the likelihood function:

$$\log L(\beta) = \sum_{i=1}^{n} \left[y_i \log p(Y_i = 1|X_i) + (1 - y_i) \log \left(1 - p(Y_i = 1|X_i)\right) \right] \tag{6}$$

Substituting $p(Y_i = 1|X_i) = \frac{1}{1+e^{-X_i\beta}}$, we get:

$$\log L(\beta) = \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1 + e^{-X_i\beta}} \right) + (1 - y_i) \log \left(\frac{e^{-X_i\beta}}{1 + e^{-X_i\beta}} \right) \right] \tag{7}$$

Simplifying further:

$$\log L(\beta) = \sum_{i=1}^{n} \left[y_i (X_i\beta) - \log(1 + e^{X_i\beta}) \right] \tag{8}$$

This gives the log-likelihood function to maximize when estimating the parameters β.

The values of β that maximize the log-likelihood function are obtained by taking the partial derivatives of the log-likelihood with respect to each parameter and equating them to zero:

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = 0, \quad \forall j = 1, 2, 3, \ldots, n \tag{9}$$

The derivative of the log-likelihood with respect to β_j is:

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\left(y_i - \frac{1}{1 + e^{-X\beta}} \right) X_{ij} \right] \tag{10}$$

where X_{ij} is the j-th independent variable for the i-th observation.

3.3 Model Fitting and Evaluation

The logistic regression model was fitted with glm function and binomial family of R Studio. The goodness of fit tests applied in this research were the residual deviance and the Akaike Information Criterion (AIC). Coefficient estimates, p-values and standard errors were calculated in order to test the significance of the independent variables

4 RESULTS AND DISCUSSION

4.1 Introduction

This section presents the results of the study aiming at identifying factors causing digital loan defaults in the commercial banks in Bungoma County by employing logistic regression techniques. The paper aims at analyzing relationships between various predictor variables and the default risk with particular emphasis on the impacts of the predictor variables towards enhancing the quality of the digital loan portfolio. Use of the model coefficients and evaluating for statistical significance gives the expected and the unexpected results. This is done with regards to the borrowers' financial and economic environment and in comparison with findings from prior studies and related literature. To answer the central research questions and provide actionable insights, the study holds the following implications:

4.2 Logit Model Fitting

4.2.1 Model Identification Process

We conducted preliminary tests to ensure that the model was appropriately identified, variables were correctly selected, and conditions such as the exclusion of multicollinearity and sufficient variation in the dependent variable were met. These tests helped in model specification, variable selection, and adherence to critical identification conditions. The weak correlations observed in most variables suggest that multicollinearity was not a significant issue in our data, which is advantageous for conducting a logistic regression analysis. The absence of strong correlations among predictors indicates a relatively clean model specification, making it less likely that multicollinearity would distort the regression results. The correlation matrix of the variables is shown in Figure 1 below:

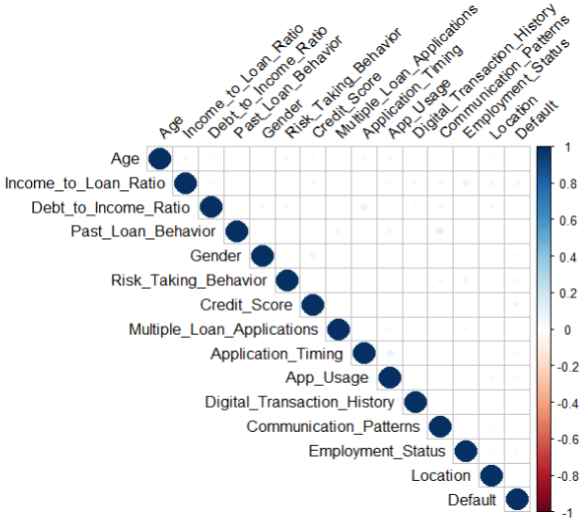

Figure 1: Correlation plot showing the relationship between variables.

4.2.2 Exploratory Data Analysis

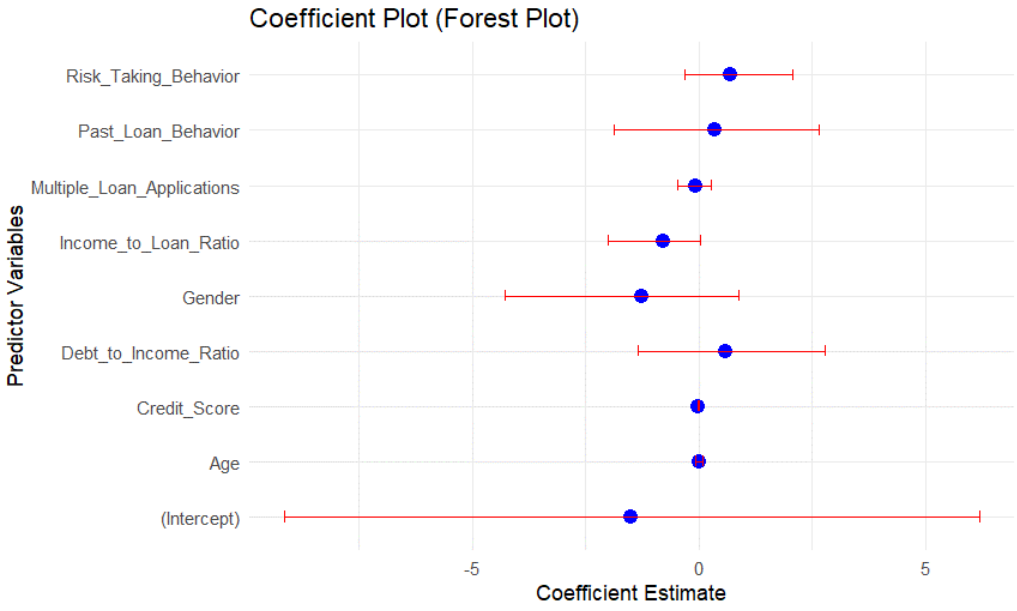

Figure 2: Forest plot showing the results of the analysis.

From figure 2 above,we can clearly see that significant predictors were Variables like Risk taking behavior,Past loan behavior, Income to Loan ratio, and Multiple loan applications while non significant predictors were variables like gender, age and app Usage.The significance of the predictor variables in the model is explained by its strong correlation with the default event. Risk taking behavior is significant because borrowers who are more risk tolerant are likely to default. Past loan behavior is especially important because borrowers who

successfully repaid the loan or defaulted on it is predictable in his or her behavior. The debt to income ratio is another important metric that describes the ratio of the income of the borrower to the loan amount; the bigger the value the lower the risk to face problem with payments and default. Multiple loan applications are also important, which indicate that the customer may have a huge burden and is likely to default hence high risk.

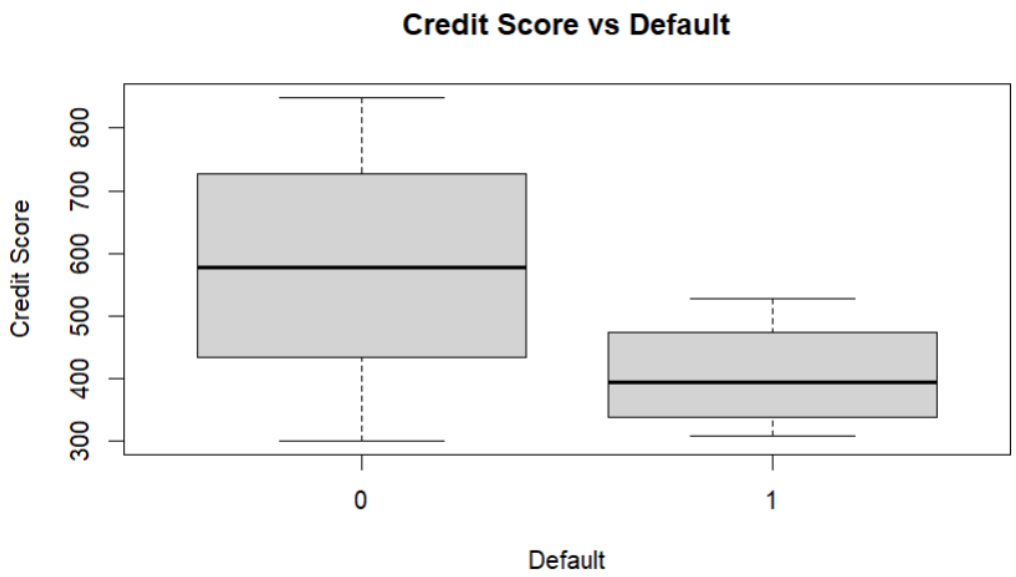

Figure 3: Box plot showing the distribution of data.

The relationship between credit score and default was shown by Figure 3.The box plot for no default (0) depicts a wider range with a median credit score higher than that for defaulting borrowers. This means that customers who do not default on loans do have a higher credit score. On the other hand, the box plot for borrowers who recorded default event is a narrower implying that borrowers who default have a lower credit score.

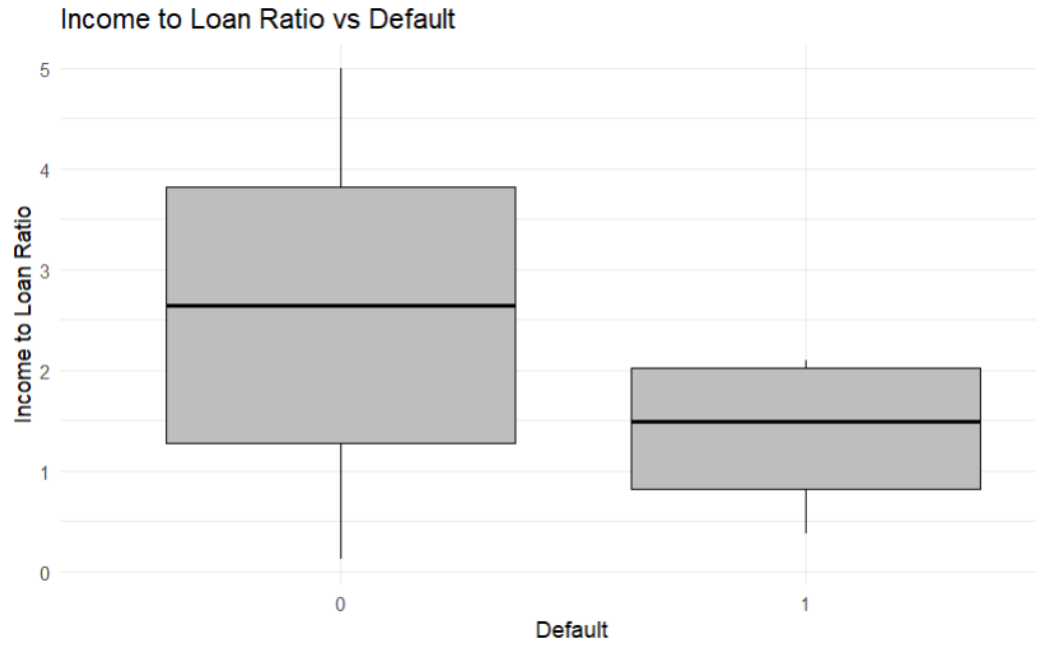

Figure 4: Box plot illustrating the data distribution.

From Figure 4 above , we can see that the box for no default is wider with a median of approximately 3. This means that borrowers who don't default have a more varied income to loan ratio. The box plot for borrowers who defaulted however is narrower suggesting that defaulting borrowers generally have a lower income to loan ratio. The results observed above strongly supports the idea that borrowers with a fair financial situation tend to have lower credit risk as compared to ones in an unstable financial state.

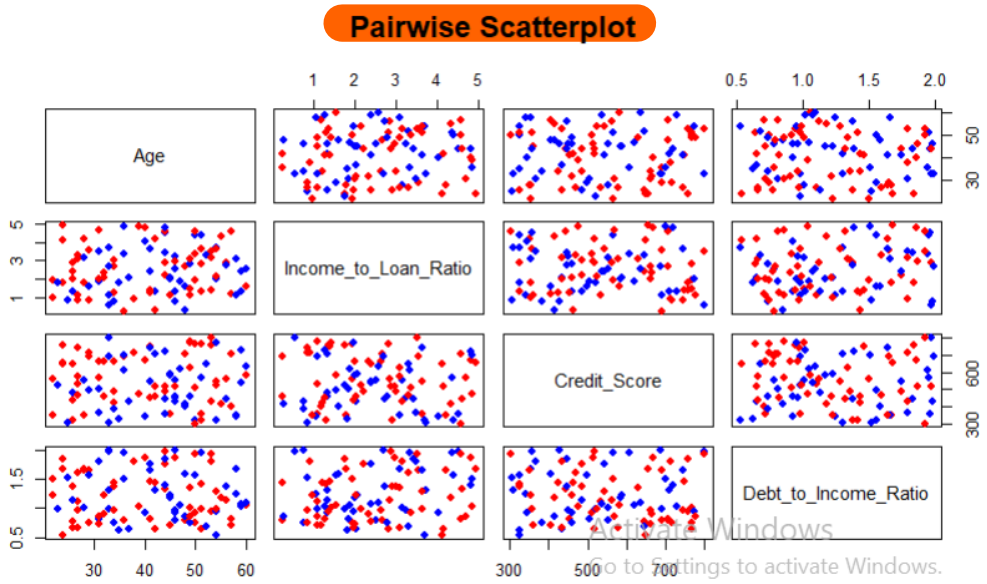

Figure 5: Scatter plot showing the relationship between variables.

Figure 5 above shows a scatter plot matrix displaying pairwise relationships between age, income to loan ratio, credit score and debt to income ratio. From the figure, we can clearly see that there is no clear relationship between age and Income to loan ratio, age and credit score, age and debt to income ratio. This is evidenced by the fact that the points on the scatter plots are actually scattered across the plot suggesting that age is not related to either of the above variables. The relationship between age and credit score exists but is weak. There is however a slight trend between credit score and debt to income ratio where higher credit score seems to correlate with a lower debt to income ratio. The correlation is however not strong.

4.2.3 Model Building Process

Since weak correlations were observed in most variables in our dataset, a logistic regression analysis was found to be suitable since multicollinearity would not distort the regression results. The logit model was fitted to our data set and the results were as shown in Table.1 below;

Variable	**Estimate**	**Std. Error**	**z value**	$\Pr(>\|z\|)$
(Intercept)	-4.712	1.531	-3.078	0.002085
Age	0.05529	0.01397	3.958	0.0000755
Income_to_Loan_Ratio	-1.145	0.1963	-5.831	5.5E-09
Debt_to_Income_Ratio	0.7543	0.3287	2.295	0.021753
Past_Loan_Behavior	-1.19	0.4208	-2.829	0.004676
Gender	0.3868	0.3515	1.1	0.271025
Risk_Taking_Behavior	0.8887	0.1981	4.486	0.000007
Credit_Score	-0.01087	0.001861	-5.839	5.25E-09
Multiple_Loan_Applications	0.2262	0.06678	3.388	0.000705
Application_Timing	0.04598	0.02593	1.774	0.076146
App_Usage	-0.01018	0.01188	-0.857	0.391367
Digital_Transaction_History	-0.00015	0.00006234	-2.359	0.018322
Communication_Patterns	0.03446	0.02925	1.178	0.238748
Employment_Status	-0.5883	0.3614	-1.628	0.103569
Location	-0.02918	0.2107	-0.138	0.889859

Table 1: Regression Results

4.2.4 Logistic Regression Model Fitting

The Logit Model was expressed as;

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_9+\beta_{10}X_{10}+\beta_{11}X_{11}+\beta_{12}X_{12}. \tag{11}$$

Based on the results in Table 1 above, the fitted model was found to be;

$$\begin{aligned}\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = & -4.712 + 0.05529 \cdot \text{Age} - 1.145 \cdot \text{Income to Loan Ratio} \\ & + 0.7543 \cdot \text{Debt to Income Ratio} \\ & - 1.190 \cdot \text{Past Loan Behavior} + 0.3868 \cdot \text{Gender} + 0.8887 \cdot \text{Risk Taking Behavior} \\ & - 0.01087 \cdot \text{Credit Score} + 0.2262 \cdot \text{Multiple Loan Applications} \\ & + 0.04598 \cdot \text{Application Timing} \\ & - 0.01018 \cdot \text{App Usage} - 0.0001471 \cdot \text{Digital Transaction History} \\ & + 0.03446 \cdot \text{Communication Patterns} \\ & - 0.5883 \cdot \text{Employment Status} - 0.02918 \cdot \text{Location}\end{aligned}$$

From the fitted model above, it was observed that;

- **Intercept:** The estimated intercept was found to be -4.712.The depression of the intercept suggests that loan default is unlikely in general when all the predictors are at their average levels. p-value 0.0021: Significant. While predisposition to default increases at every level of control, even with no other factors accounted for, the baseline tendency is quite low.

- **Age:**.An estimated value of 0.0553. A positive coefficient means that as borrowers age,the default probability rises slightly with a p-value of $7.55e-05$,Very high, which shows that age is a good predictor of default.This supports the idea that as borrowers attain old age,their income streams tend to narrow hence higher default probability.

- **Income to Loan Ratio:**.An estimated value of : -1.145.The negative sign indicates the fact that the higher income to loan ratio is associated with a lower probability to default by the borrowers.The reduced risk is attributed to financial stability. p-value $5.50e-09$: Very significant, thus it is one of the most important factors that greatly leads to positive outcomes.

- **Debt to Income Ratio :**.An estimated value of :0.754.If the number of borrowers who defaulted was larger,for the borrowers with higher debt to income ratio, then coefficient will be positive. p-value 0.0218: Indeed, major, proving it has a predictive value on defaults.

- **Past Loan Behavior:**.An estimated value of: -1.190.This means that if the signs in front of period t measures are positive, then this has a negative impact on the likelihood of default in t period adding that positive past loan behavior (for example, timely re-payment) decreases the likelihood of default. p-value 0.0047: which is statistically significant.

- **Risk taking Behavior:**.An estimated value of : 0.889.Credit risk is positively related to borrowers' risk-taking, indicating that borrowers with propensity to take higher risk are more likely to default. p-value $7.24e-06$:implies a strong relationship, which is highly significant.

- **Credit Score**.An estimated value of -0.0109.The negative coefficient implies that higher values of credit score, means lower probability of defaulting. p-value $5.25e-09$: Very significant as this is also another predictor variable.

- **Multiple Loan Applications** .An estimated value of 0.226.A positive coefficient means that borrowers who apply for multiple credit products are more likely to default, which may be explained by desperation. p-value 0.0007: Highly significant.

- **Digital Transaction History**.The results show an estimated value of −0.000147.This is a small negative coefficient indicating that borrowers with detailed transaction histories (which is suggestive of stability) are less likely to default. p-value 0.0183: Statistically meaningful, but trivial, to use the slang, in terms of estimated measure of impact.

- **Application Timing**.From the analysis of the results,it was estimated at 0.04598.A positive value of the coefficient.This therefore implies that borrowers who apply for loans at certain times (for instance during financial difficulties) are slightly more likely to repay their loans. p-value 0.0761: Some of them are just marginally significant and thus merit further research. The following variables were found to be non-significant predictors;Gender

- **Gender**.The coefficient (positive) is not significantly different from zero Therefore, gender as a covariate does not predict default in this model.

- **App Usage**The negative coefficient means that the app usage has a tiny negative impact on the defaults where the higher the usage the lower the default rates. This may mean that borrowers are slightly less likely to be a defaulter if they are more active in use of the loan app, in terms of balance checks, repayments and checking Money Management tools. The coefficient is so small that it can be inferred that, app usage has little effect on the default behavior. With increments of app usage—depending on what it is defined (usage logins or time spent, for instance)—the difference in the change in the log-odds of default is very close to zero. However, the given p value 0.391367, which is well above the conventional value of 0.05 signifies that app usage is not significantly related to loan default. It is also possible that there exists no genuine link between the study factors and the outcome, and the metrics derived from the observed data reflect only random fluctuations.

- **Communication Patterns**.A positive estimate of 0.03446: The positive sign of the coefficient implies that the higher the communication frequencies or a greater communicating intensity, default rates increase just slightly. This could mean that borrowers who frequent the system, lenders or the loan system via inquiries, messages or notifications may be in a position to default the loans. Thus, it may indicate such aspects as financial pressure, in which a higher rate of communication means such problems or hesitation of repayment. Using the conventional test, the p-value 0.238748 greater than 0.05 and therefore, show that the pattern of communication and default is not significantly related. That is, the results of the current study indicate that communication patterns are not viable to predict loan default in the intended sense. It may have been caused by chance factors in the results that make them appear otherwise related when they are not.

- **Employment Status**Negative but near non significant; unemployment has a marginal negative effect on default risk, perhaps slightly offset by a positive effect.

- **Location**.Very little impact and statistically significantly close to being worthless.The location of the borrower therefore does not impact in any way the probability of default on digital loans.

4.3 Model Adequacy

- **Hosmer and Lemeshow goodness of fit (GOF) test**
  Table 2 below shows summary statistic results for Hosmer and Lemeshow goodness of fit;

Statistic	Value
χ^2	4.1054
df	8
p-value	0.8475

Table 2: Summary Statistics

The Hosmer-Lemeshow test above shows that the logistic regression model is a good fit of the data since p-value = 0 .8475, is greater than the significance level of 0.05 which mean that there is no significant difference between the observed and the predicted default probability.

- **Receiver Operating Characteristic (ROC) curve**

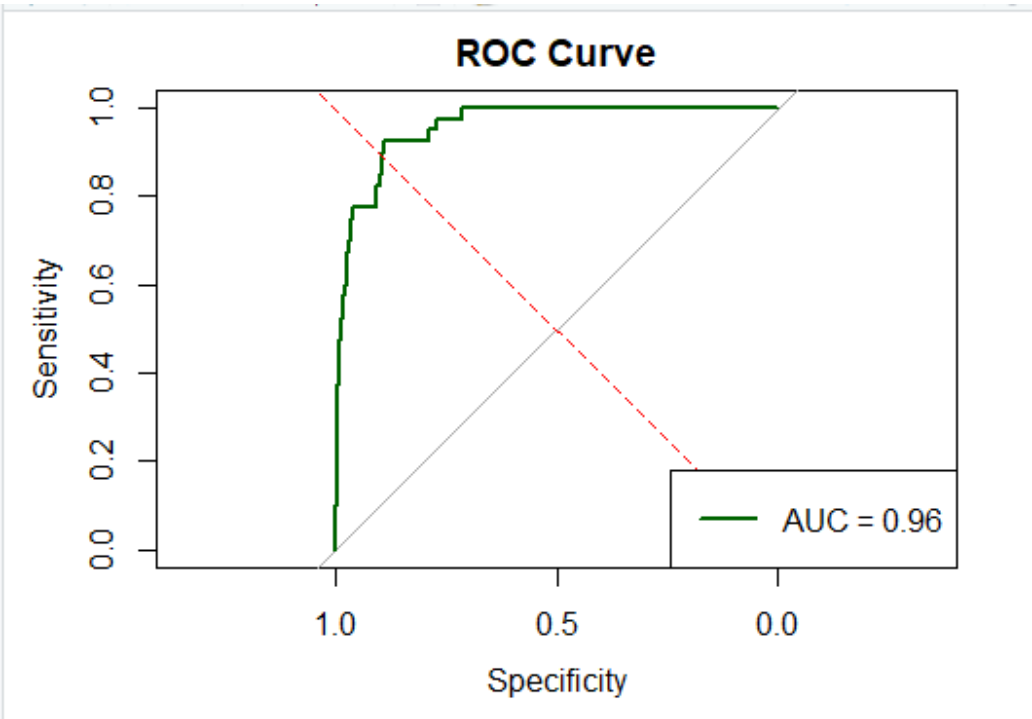

Figure 6: Receiver Operating Characteristic (ROC) Curve

We can deduce the following from figure.6 above;

- **High Predictive Power**:
  According to the Area under the Cover (AUC) score, with the absolute score of 0.9579 the model shows perfect ability in classifying between the default and non-default borrowers. For any randomly chosen pair of borrowers: one in default and the second not in default, the model always assigns a higher default probability to the defaulting borrower in 95.79% of cases.
- **Model Performance:** This high AUC value points to the fact that the proposed model provides an excellent ranking of borrowers according to their risk of defaulting on their loans. They show that the model has managed to capture most of the underlying patterns in the data.
- **Robustness:** As a rule, such a high AUC suggests that the model can generalize well to and perform well on new data excluding overfitting or data problems. It also has good accuracy in sorting out positive and negative cases, therefore, the logistic regression model with an AUC of 0.9579 can be relied upon to predict loan default.This assessment, coupled with the favourable Hosmer-Lemeshow test values, also indicate good calibration and practical usefulness in decision making.

- **Residual Analysis: Deviance Residuals**
  Deviance residuals are normally shaped with no apparent trends and located at 0.Figure 7 below indicates that the fitted model is indeed adequate. This tally with the previous observation such the good AUC and the Hosmer-Lemeshow.

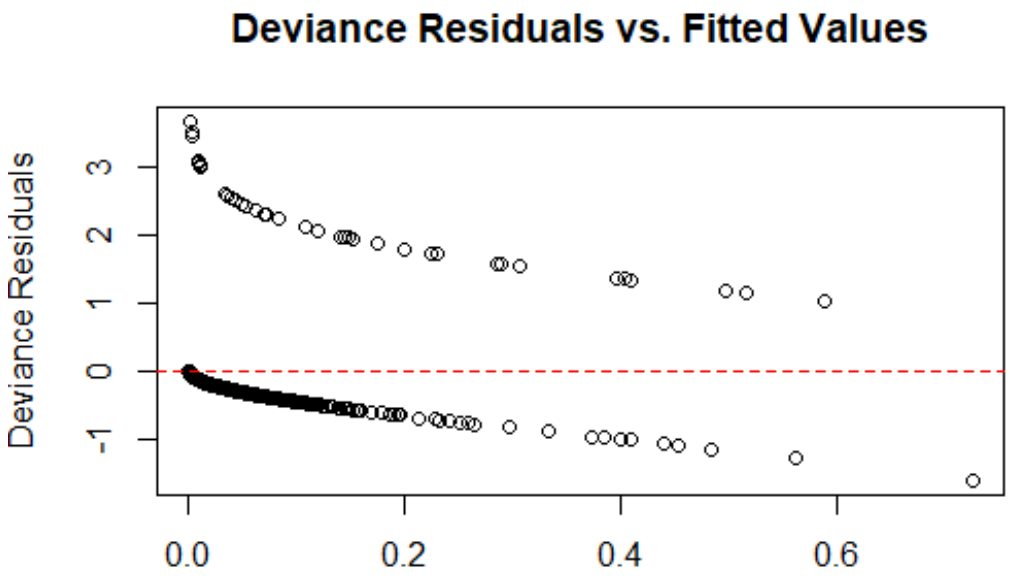

Figure 7: Deviance Plot

4.3.1 Adding Interaction Terms in Logistic Regression

Table 3 below shows the interaction effects between age and income to loan ratio.

Variable	**Estimate**	**Std. Error**	**z value**	Pr(> |z|)
(Intercept)	-3.263	-1.902	0.057	
Age	0.0256	1.217	0.224	
Income_to_Loan_Ratio	-2.671	-2.798	0.005	
Debt_to_Income_Ratio	0.734	2.243	0.025	
Past_Loan_Behavior	-1.173	-2.796	0.005	
Gender	0.3997	1.139	0.255	
Risk_Taking_Behavior	0.9	4.532	0.000	
Credit_Score	-0.01091	-5.841	5.18E-09	
Multiple_Loan_Applications	0.2288	3.443	0.0006	
Application_Timing	0.04751	1.828	0.068	
App_Usage	-0.01064	-0.898	0.369	
Digital_Transaction_History	-0.000152	-2.432	0.015	
Communication_Patterns	0.03713	1.267	0.205	
Employment_Status	-0.5774	-1.6	0.110	
Location	-0.02564	-0.122	0.903	
Age:Income_to_Loan_Ratio	0.02924	1.706	0.088	

Table 3: Regression Results

From table 3,we can clearly see that the variable income to loan ratio has a marginally significant interaction with the variable age, which means that the effect of the income-to-loan ratio variable on default may depend on the age of the borrower.$p = 0.02924$, $\beta = 0.0880$. For the interaction model the AIC was determined to be 319.5686 while for the no-interaction model it was 320.6856. AIC penalizes complexity thus the fact that it reduces shows that its addition helps in boosting model fit while at the same time explaining the need for the extra complexity.

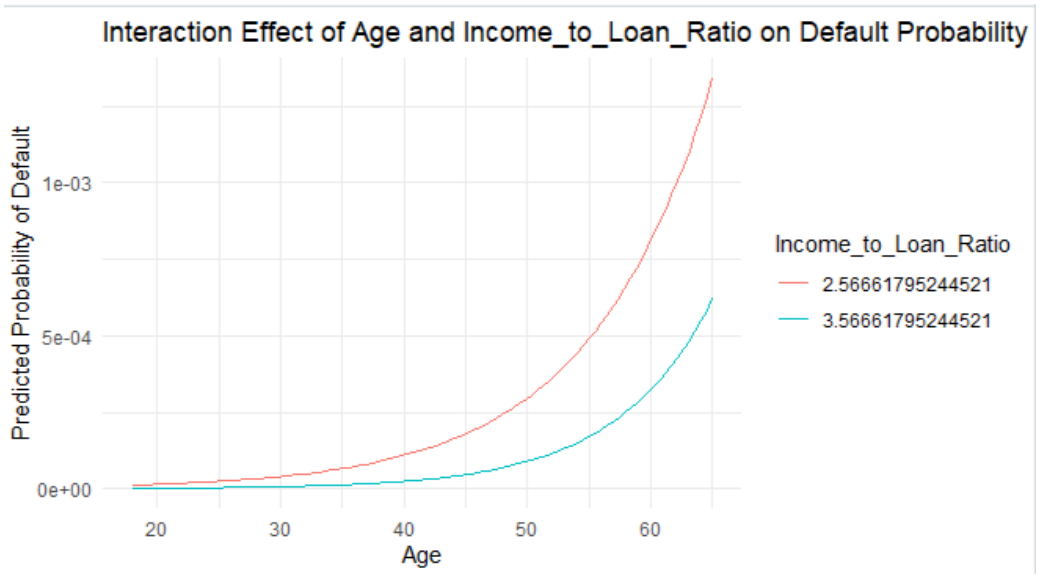

Figure 8: Interaction Curve.

4.3.2 Perform likelihood ratio test

The value for p is 0.07748.07748,slightly higher than the conventional 0.05 levels, meaning that the additional of an interaction term enhances the fitness of the model only marginally. However, by including the interaction term into the equation the model does improve, but the improvement is not to a great extent that makes the statistical significance at 0.05 level. Nonetheless, the p-value = 0.075 < 0.1, which implies that at best there is weak evidence that we should incorporate the interaction variable.

Resid. Df	**Resid. Dev**	**Df**	**Deviance**	$\Pr(>\|Chi\|)$
5985	290.69			
5984	287.57	1	3.117	0.07748

Table 4: Model comparison table showing residual degrees of freedom, residual deviance, degrees of freedom, deviance, p-value, and significance.

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

Logistic regression analysis was employed to model the relationship between digital loan predictor variables and loan default as the dependent variable. The logistic regression analysis identified key factors influencing loan outcomes. The income to loan ratio and credit score were significant predictors of loan default, with higher income ratios and credit scores reducing default risk. On the other hand, variables like debt to income ratio, past loan behavior and gender did not show statistical significance, suggesting they may not be reliable predictors in this context.

5.2 Recommendations

The contemporary study used logistic regression analysis to model the relationship between digital loan predictor variables and loan default as the dependent variable. Therefore, future studies can improve the predictability of the model by:

1. Excluding non-significant variables like past loan behavior and gender, or exploring alternative measures to enhance predictive power and simplify the model.

2. Reviewing the role of the debt to income criteria since it does not seem to play any critical role, while, possibly, decreasing its importance or adding to this criterion other financial performance coefficients.

3. Targeting predictors like income-to-loan ratio and credit score which is the most preferred and confirmed predictor of loan outcomes.

More so, the study focused on digital loans in three major commercial banks in Bungoma County. Therefore, future studies can expand the literature by considering quasi-digital loan products such as Buy-Now-Pay-Later (BNPL) (Watu Simu & Delights Solar products) and ecosystem loans.

References

[1] Alshebmi, A. S., Adam, M. H. M., Mustafa, A. M., & Abdelmaksoud, M. T. D. O. E. (2020). Assessing the non-performing loans and their effect on banks' profitability: Empirical evidence from the Saudi Arabia banking sector. *International Journal of Innovation, Creativity and Change, 11*(8), 69-93.

[2] Anderson, D. R. (2007). *Model based inference in the life sciences: A primer on evidence*. Springer Science & Business Media.

[3] Barbiero, F., & Dimou, M. (2024). Credit risk and bank lending conditions. *ECB Economic Bulletin, Issue 4/2024*.

[4] Chakroun, F., Abid, L., Elarbi, D., & Masmoudi, A. (2024). Gamma–lindley regression cure model for corporate credit default prediction. *Expert Systems with Applications, 257*, 125004.

[5] Challoumis, C., & Eriotis, N. (2024). A historical analysis of the banking system and its impact on the Greek economy. *Edelweiss Applied Science and Technology, 8*(6), 1598-1617.

[6] Chen, G., & Faz, X. (2015). The potential of digital credit to expand financial inclusion in emerging markets. *Consultative Group to Assist the Poor (CGAP)*.

[7] Das, A. (2024). Logistic regression. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 3985-3986). Cham: Springer International Publishing.

[8] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179-188.

[9] Fundingo. (2024). *Mastering credit risk assessment: Evaluating creditworthiness*.

[10] Johnston, Mary. Financial Inclusion through Digital Innovation. New York: Economic Press, 2024.

[11] KPMG. (2022). *Banking industry outlook: Revenue streams and profitability drivers*.

[12] McLachlan, G. J. (2005). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.

[13] Nation, Africa. (2024). Digital lenders should use AI to reduce default rates.

[14] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society, 135*(3), 370-384.

[15] Njenga, M., & Kavindah, P. (2021). Digital lending in Kenya: A review of credit access and challenges.

[16] Pocriciuc, A. (2024). Financial reporting in the era of AI: The response of companies in the Netherlands to the challenges posed by machine readership (Master's thesis, University of Twente).

[17] Quost, B., Denoeux, T., & Li, S. Parametric classification with soft labels using the Evidential EM algorithm.

[18] Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2019). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control, 52*, 456-462.

[19] Tam, K. Y., & Kiang, M. (1990). Predicting bank failures: A neural network approach. *Applied Artificial Intelligence an International Journal, 4*(4), 265-282.

[20] Uniify. (n.d.). Best practices for credit assessments. Uniify.

[21] McLachlan, G. J. (2005). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.