Financial risk assessment for loan application

using logistic regression and factor analysis

1. Abstract :

Loan officer at any bank would be interested in detecting the factors which can identify people who are likely to default on loans, consequently good and bad credit risks. Moreover, he will also be interested in designing model which can predict chances of default with reasonable accuracy. A crucial step in the loan application process is financial risk assessment, which helps financial institutions minimize credit risk and allocate resources as efficiently as possible. This paper creates a predictive framework for estimating the probability of loan defaults by combining factor analysis and logistic regression. While Factor Analysis lowers dimensionality and finds latent characteristics influencing financial risk, like credit history and debt-to-income ratio. Logistic Regression is a strong classification technique is used to estimate default probabilities by balancing sensitivity and specificity, the suggested model showed excellent predictive accuracy. It offers a data-driven and interpretable way to expedite loan approval decisions, guaranteeing a more fair and effective procedure.

2. Introduction :

The overall purpose of this project is to analyse and build a model on loan application data that can be useful for decision-making purposes. Traditional methods rely on credit scoring models that consider a range of financial and demographic variables. However, these models often face challenges due to multicollinearity, high dimensionality, and the lack of interpretability of complex relationships among variables. To address these limitations, advanced techniques of multivariate analysis are applied here: Factor Analysis & Logistic regression. These will help identify patterns, relationships, and predict from the dataset, thus supporting managerial decision making in risk assessment and loan approval processes. This term paper simulates the real-world scenario wherein a financial institution aims to optimize its loan approval system. The analysis would identify the key factors that influence loan decisions, segment customers into meaningful groups, and build predictive models for loan approval within the restrictions of risk management standards.

**Objective:**

1. To identify the key underlying factors that influence loan approval decisions

based on borrowers' financial profiles, credit history, and demographic characteristics.

**2.** To predict loan approval likelihood based on financial and demographic

factors, identifying significant predictors such as credit score, debt-to-income ratio,

homeownership, and loan term to inform loan decision-making.

3. Literature Review:

Statistical and machine learning techniques for evaluating loans have been thoroughly examined in previous studies. Because of its ease of use and interpretability, logistic regression has been a mainstay. Conversely, factor analysis has been used in many fields to extract latent variables from correlated data. In order to improve performance, recent research emphasize the advantages of combining dimensionality reduction approaches with predictive models. By preprocessing loan data using factor analysis before utilizing logistic regression for prediction, this study expands on previous discoveries.

3.1. Credit Assessment :

The process of assessing a person's or an organization's creditworthiness in order to ascertain their capacity to repay debt is known as credit evaluation. To ascertain the degree of risk involved in making a loan to a borrower, it entails examining their income, assets, liabilities, and financial history because it enables them to reduce risk and make well-informed lending decisions, credit evaluation is essential for financial institutions. (Nsengiyera 2020) Financial firm risk losing money if they lend money to people or organizations that might not be able to repay the loan if their credit is not properly evaluated.

Credit Assessment Types —

There are several methods for evaluating credit, and each has pros and cons of its own. These are a few of the most prevalent kinds:

‍

**Traditional Credit Assessment:** This kind of evaluation is predicated on a borrower's financial records, payment history, and credit history. The evaluation entails examining the borrower's credit utilization ratio, existing debts, payment history, and credit score. Banks and other financial organizations frequently conduct traditional credit evaluations.

**Behavioural Credit Assessment:** This method determines a borrower's creditworthiness by analysing information about their habits and behaviour. In order to learn more about the borrower's reliability and character, this method looks at non-financial data including social media activity and internet purchases.

**Alternative Credit Assessment:** This method of determining a borrower's creditworthiness uses non-traditional data sources, like utility bills, rent payment history, and other information. This strategy is especially helpful for borrowers who do not have a lot of credit history or who have little access to conventional credit information.

* 1. Logistic Regression:

Credit risk Analysis can be carried out through numerous methods. The logistic regression and neural network models both are alike & worthy. Though the prior one is marginally better. Moreover, the genetic algorithm model is somewhat inferior but efficient. (Gouvea 2007). Logistic regression and multicriteria decision making have been combined to create a proficient strategy to achieve high performance. Logistics regression is used to find probability of default whereas multicriteria decision is used to classify firms for credit scoring based on predefined criteria. (Jerić 2009). Credit Risk Analysis is modelled using an artificial neural network. In commercial banks, logistic regression modelling is more effective than logistic regression modelling at identifying problematic clients. Higher credit risk results from higher interest rates and longer repayment terms. Customers in the industry sector are more affected. With a longer customer relationship history with the bank and a longer loan repayment period, it decreases. (Karimi 2014).

* 1. Factor Analysis :

Factor analysis, which includes methods like principal component analysis and common factor analysis, is a statistical technique for identifying interrelations among a large number of variables. It explains these variables in terms of underlying dimensions or factors. The primary goal is to reduce a large dataset into fewer variables (factors) while retaining most of the original information. By providing a quantitative estimation of the variables' structure, factor analysis offers an objective method to create summarized scales. For instance, in a study of customer ratings for a fast-food restaurant, six variables—food taste, food temperature, freshness, waiting time, cleanliness, and employee friendliness—could be analysed. (Joseph F. Hair 2019) The analysis might reveal that food taste, temperature, and freshness cluster into a single factor, termed "food quality." Similarly, waiting time, cleanliness, and employee friendliness might form another factor, referred to as "service quality." This approach helps researchers better understand the underlying relationships among variables.

4. Methodology:

Dataset Definition and Source :

The dataset used for this analysis includes detailed information on loan applications from borrowers and various types of variables- financial and credit profile, demographic characteristics, and loan-specific details.

Source:

The dataset has been curated and shared for academic purposes to simulate loan approval data in a financial institution.

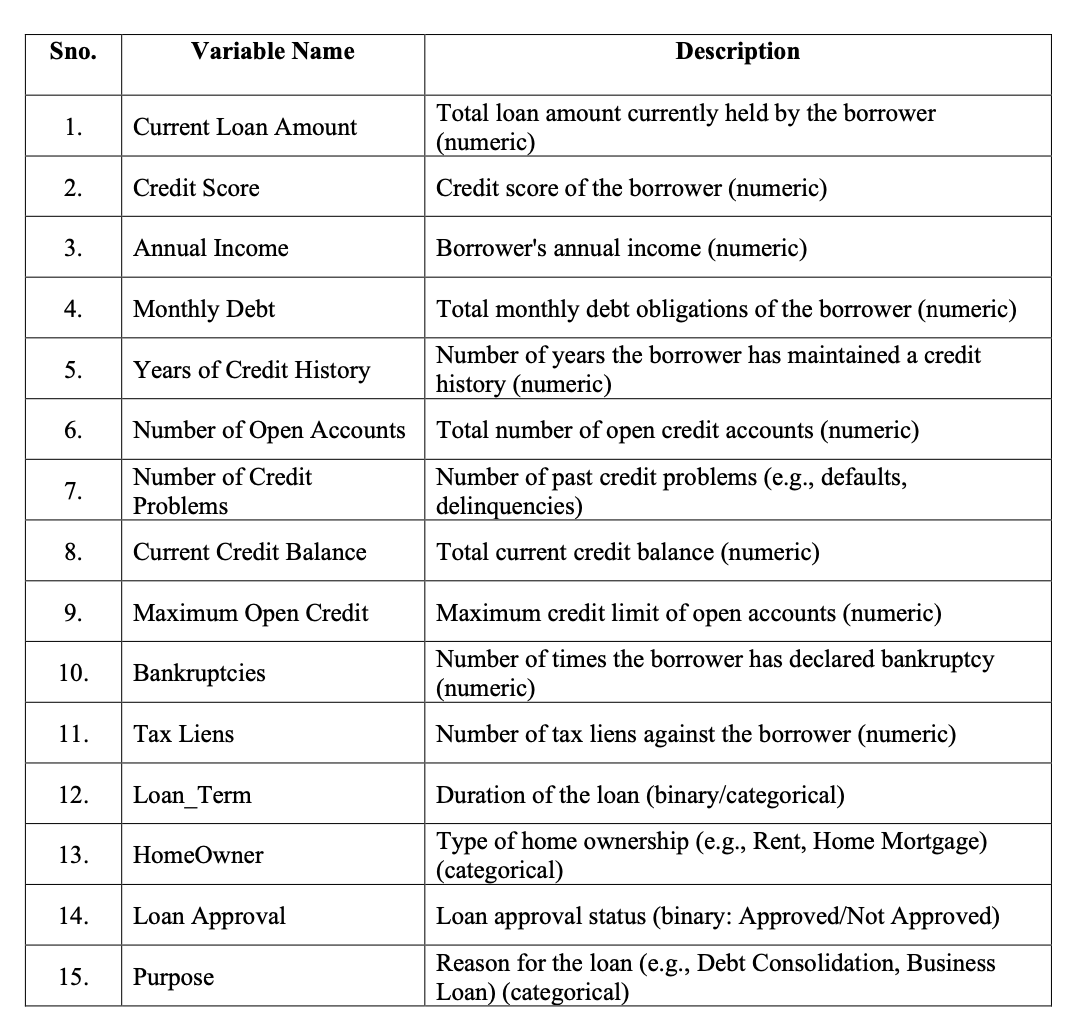
Dataset Objective:

This dataset will be analysed to identify key factors influencing loan approval decisions,

predict loan outcomes, and segment borrowers into actionable groups based on financial and credit profiles.

Dataset Description :

The dataset consists of 1,099 records, with the following key variables:



4.1 Factor Analysis :

**Objective:** To identify the key underlying factors that influence loan approval decisions based on borrowers' financial profiles, credit history, and demographic characteristics.

**SPSS Output:**

1. **KaiserMeyerOlkin (KMO) :** Measures the sampling adequacy to ensure that the data is suitable for factor analysis. A value close to 1 indicates that factor analysis is appropriate.

• **KMO Value**: **0.553**

• The KMO value above **0.5**, indicates that the dataset is **adequate for factor analysis**.

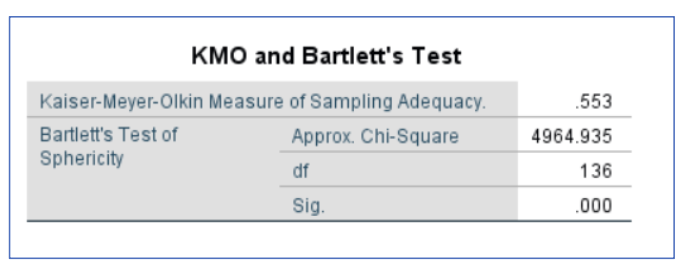
**b) Bartlett's Test of Sphericity :** Tests the null hypothesis that the correlation matrix is an identity matrix (no correlation between variables). A significant result (p-value < 0.05) suggests that the data is suitable for factor analysis.

• Chi-square (Approx.): 4964.935

• Degrees of Freedom (df): 136

• Significance (Sig.): 0.000

• The test is **highly significant** (**p < 0.05**), indicating that the correlation matrix is not an identity matrix. This means there are significant relationships among the variables, justifying the use of factor analysis.



**c) Principle Component Analysis:**

• **Eigenvalues and Factor Retention:**

• Components with eigenvalues above 1 are typically considered significant and considered for further analysis in Factor Analysis. In this case, seven components have an eigenvalue above 1, so they were taken as the significant factors.

• This states, complexity in your dataset can be reduced to about seven core factors without information loss being too consequential

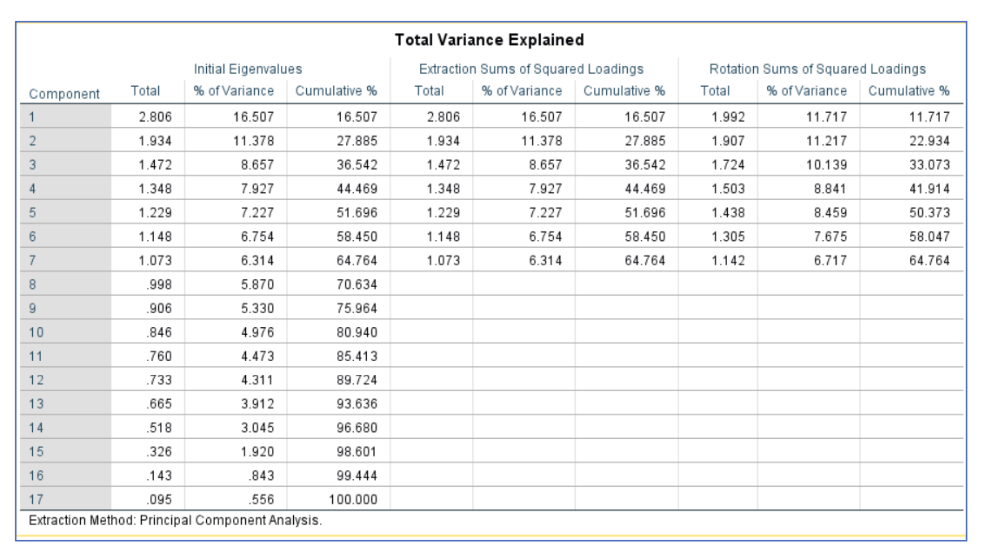
**Scree Test:** A visual inspection of the eigenvalues in a scree plot. The point where the curve flattens helps to determine the number of factors to retain.

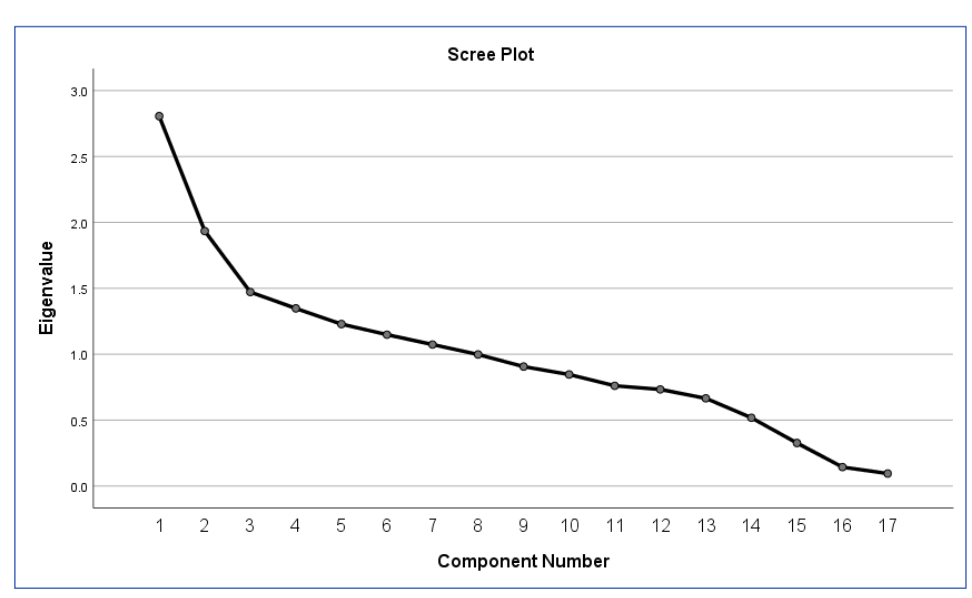
**Variance Explained by Each Factor:**

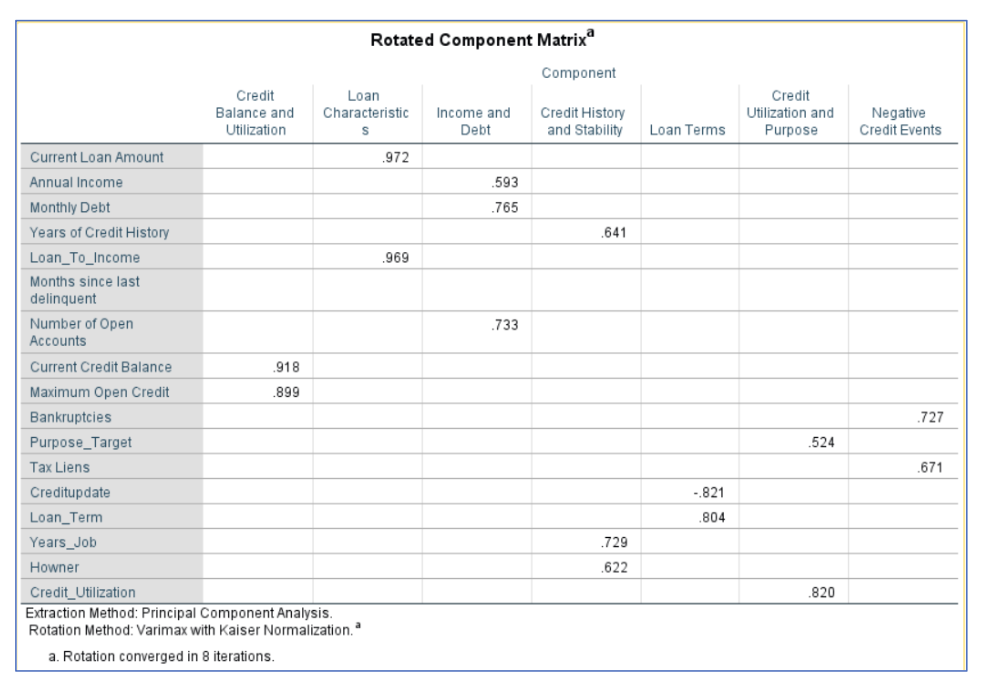
• The initial eigenvalues column shows that the first component explains 16.507% of the

variance, while the cumulative variance explained by the seven factors is 64.764%. This

means that these seven components capture around 65% of the total variance in the dataset.





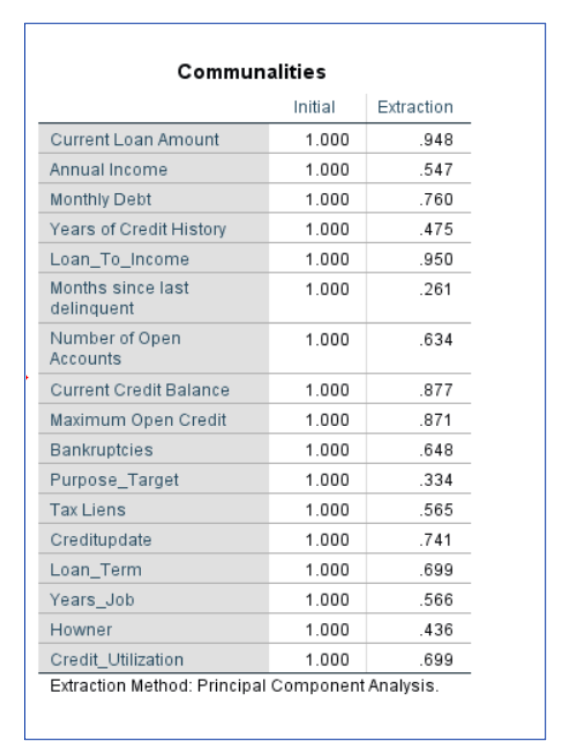


**Communities Analysis**

• **Variables with High Communalities:** These variables are strongly captured by the factors, indicating they share a lot of variances with the latent constructs.

• **Variables with Moderate Communalities:** These are reasonably well explained by the factors but could be influenced by other unmeasured variables.

• **Variables with Low Communalities:** These are poorly explained by the factors, suggesting that either these variables are not highly related to the latent constructs or there are additional underlying factors not captured in this analysis.



**Interpretation :**

**a) Factor 1:** Credit Balance and Utilization

This factor reflects the borrower's credit capacity and current usage of available credit

limits. High values in this factor might indicate individuals with high credit balances

and greater credit limits, which in turn may suggest a strong credit profile.

b) **Factor 2:** Loan Characteristics

This factor describes the size and burden of the loan relative to the borrower's income.

It captures how well borrowers are able to handle their loan obligations relative to

what they earn.

c) **Factor 3:** Income and Debt

This factor evaluates the creditworthiness of the borrower. It emphasizes income and

current obligations. The number of accounts open represents the borrower's activity in

the credit system.

d) **Factor 4:** Credit History and Stability

This would reflect the ability to commit to savings and dependability. A longer credit

history, stable employment, and homeownership suggest a borrower is likely more

financially dependable.

e) **Factor 5:** Loan Terms

This reflects the terms of the loan and recent updates in the credit profile of the borrower. The negative loading on "Creditupdate" suggests the relationship between recent updates to credit files and length of terms of loans might be inversely related.

f) **Factor 6:** Credit Utilization and Purpose

This factor focuses on how credit is being used for specific goals and the proportion

of credit being utilized. It ties the borrower's credit behavior to their purpose for borrowing.

g) **Factor 7:** Negative Credit Events

This factor measures the presence of adverse financial events, including bankruptcies

and tax liens, which directly impact creditworthiness and the borrower's ability to secure loans.

4.2 Logistic Regression :

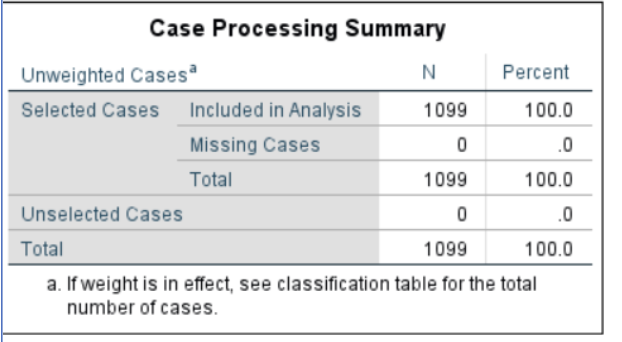
**Objective:** To predict loan approval likelihood based on financial and demographic

factors, identifying significant predictors such as credit score, debt-to-income ratio,

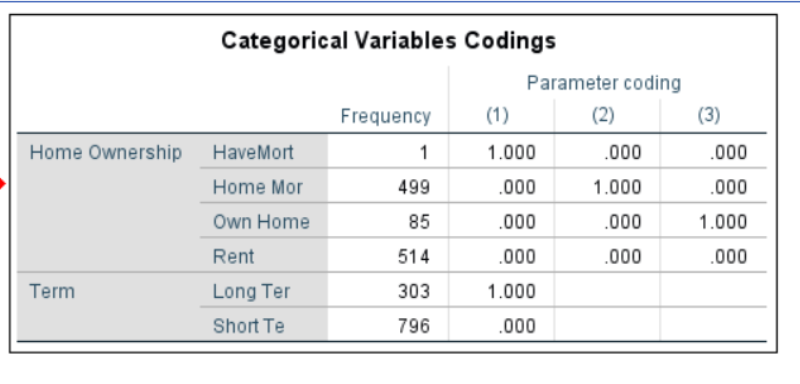
homeownership, and loan term to inform loan decision-making.

**SPSS Output:**

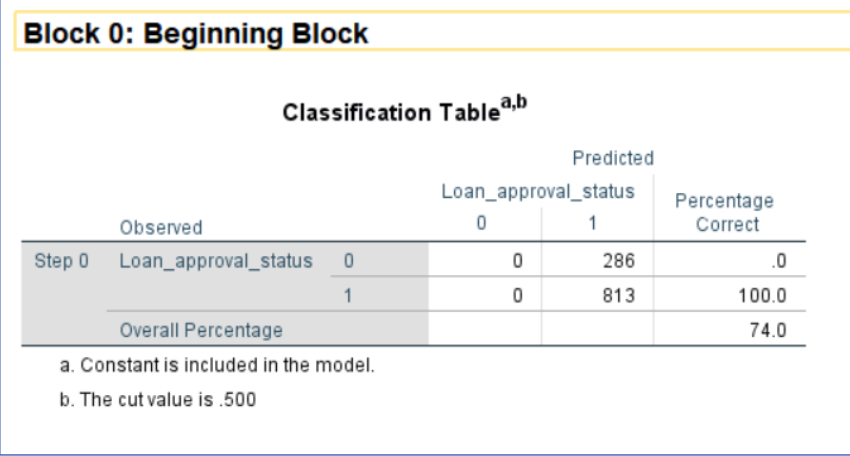
**a)**



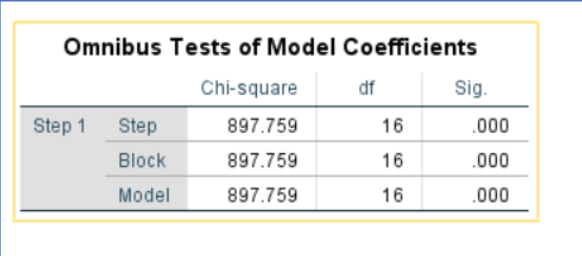
b)



**c) Baseline Comparison:** The baseline accuracy of 74.0% serves as a benchmark for evaluating the logistic regression model.

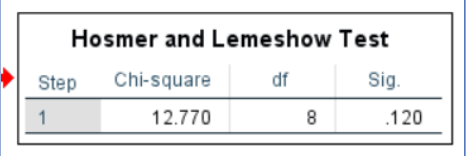


**d) Omnibus Test of Significance :**



The p-value = 0.000 (less than 0.05), which indicates that the model with predictors significantly improves the prediction compared to the null model.

**e) Hosmer and Lemeshow Test :**

****

**Interpretation :**

**1. ChiSquare Value (12.770):**

• This measures the difference between observed and predicted probabilities of the dependent

variable (loan approval). A smaller value indicates a better fit.

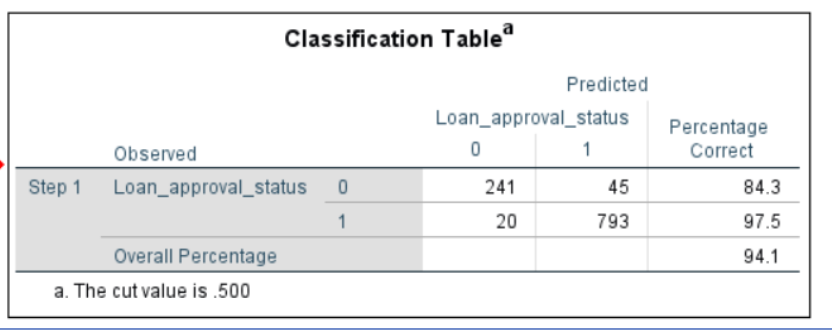
2. **Significance (p = 0.120)**:

• Since p=0.120p = 0.120p=0.120 is **greater than 0.05**, there is **no statistically significant**

**evidence** to suggest a lack of fit.

• This means the model's predictions are consistent with the observed outcomes.

3.

****

The Classification Table provides the model's prediction accuracy for the dependent variable

(in this case, Loan Approval Status) based on the logistic regression model with predictors.

Here’s how to interpret the results:

**1. Loan Approval Status = 0** (Not Approved):

• Observed = 286, of which:

▪ Correctly predicted = 241.

▪ Incorrectly predicted = 45 (predicted as approved when not approved).

• Accuracy for Not Approved cases = 84.3%.

**2.** **Loan Approval Status = 1** (Approved):

• Observed = 813, of which:

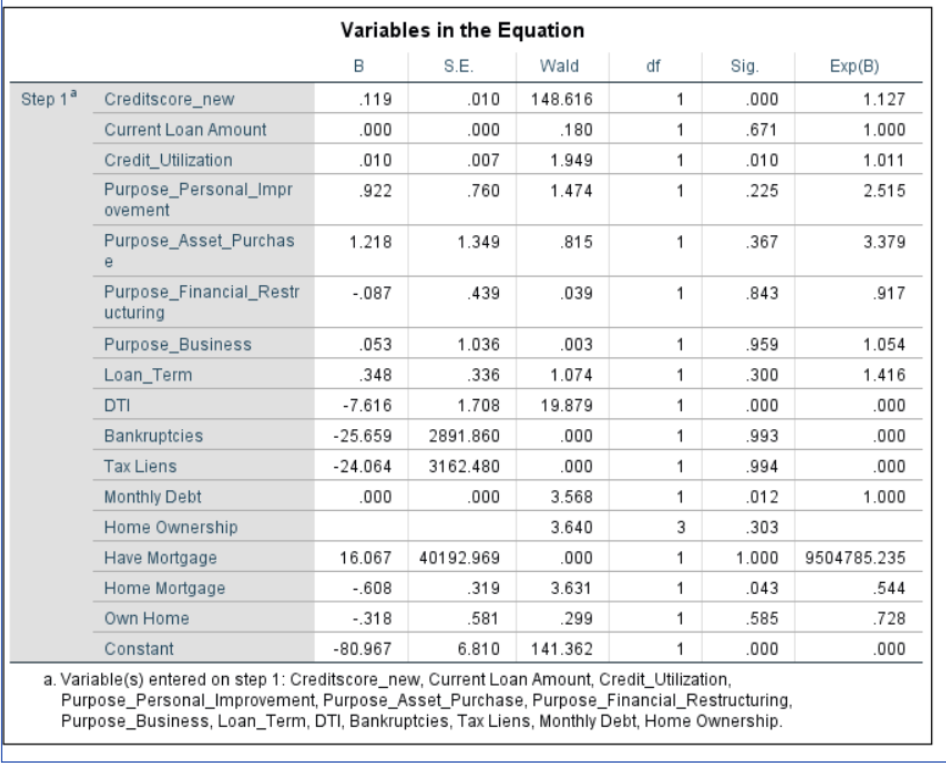
• Correctly predicted = 793.

• Incorrectly predicted = 20 (predicted as not approved when actually approved).

• Accuracy for Approved cases = 97.5%.

**3.** **Overall Percentage:**

• 94.1% of all cases (both approved and not approved) were correctly classified by the model.



• **Significant Predictors (p < 0.05)**

1. **Credit Score:**

a. **B = 0.119**, **p < 0.001**, **Exp(B) = 1.127**

b. A one-unit increase in the credit score increases the odds of loan approval by 12.7%.

c. **Insight**: High credit scores are positively associated with loan approvals.

2. **Debt-to-Income Ratio (DTI) :**

a. **B = -7.616**, **p < 0.001**, **Exp(B) = 0.000**

b. A higher DTI decreases the odds of loan approval to almost zero.

c. **Insight**: Applicants with manageable DTI are significantly more likely to get loans.

3. **Monthly Debt:**

a. **B = 0.000**, **p = 0.012**, **Exp(B) = 1.000**

b. Monthly debt, though appearing small, is significant in predicting approvals.

c. **Insight**: Applicants with consistent monthly debt management may positively influence

approval decisions.

4. **Credit Utilization:**

a. **B = 0.010**, **p = 0.010**, **Exp(B) = 1.011**

b. A slight increase in credit utilization increases the odds of loan approval by 1.1%.

c. **Insight**: Moderate credit utilization indicates responsible financial behaviour.

5. **Home Mortgage:**

a. **B = -0.608**, **p = 0.043**, **Exp(B) = 0.544**

b. Applicants with mortgages have 45.6% lower odds of approval compared to renters.

c. **Insight**: Mortgage holders may be seen as riskier due to existing liabilities.

• **Non-Significant Predictors (p ≥ 0.05)**

**1. Loan Purpose (e.g., Personal Improvement, Asset Purchase, Financial Restructuring):**

a) These purposes show no significant impact on loan approval decisions individually.

b) Action: Reassess purpose-specific lending policies or use alternate modeling techniques.

**2. Loan Term:**

a) B = 0.348, p = 0.300, Exp(B) = 1.416

b) Longer-term loans slightly increase the odds of approval but are not statistically significant.

**3. Bankruptcies and Tax Liens:**

a) Both predictors were not significant, with p > 0.99.

b) Action: While these variables may not directly predict approvals, they could influence

other loan terms (e.g., interest rates).

**4. Current Loan Amount:**

a) p = 0.671, indicating no significant impact on approvals.

**5. Home Ownership Categories (Other than Mortgage) :**

a) Categories such as "Own Home" did not significantly affect loan approvals.

• **Final Logistic Regression Equation**

**Log(P/1-P) = -80.967+(0.119×Credit Score) −(7.616×DTI) +(0.000×Monthly Debt) +(0.010×Credit Utilization) −(0.608×Home Mortgage).**

5. Result :

A) Factor Analysis:

* This factor is critical for assessing creditworthiness and risk of default, as it shows whether the borrower is nearing their credit limit or managing balances responsibly.
* A balanced mix of income and debt obligations may be healthy, indicating a stable financial situation. A high debt-to-income ratio might indicate potential difficulties in discharging the debt and consequently, in servicing the debt balance.
* These factors strongly predict trustworthiness and consistency of credit behavior. Borrowers scoring high here could be considered lower risk, since they have confirmed stability over time.
* A high score here may mean that there are significant loan obligations relative to income, potentially signifying financial strain. This factor is vital for assessing affordability of loans and the debt-to-income ratio, which is an essential metric for lenders.
* A borrower who has several new credits may have secured some new financial obligations or changes that have altered his credit profile. This factor brings in the implication of the current credit events on the longterm obligation.

B) Logistic Regression :

**a. Creditworthiness Assessment -**

* **Credit Score** is the most significant predictor.

**Action:** Develop exclusive products or expedited processes for applicants with high credit

scores.

* **Debt-to-Income Ratio** shows a strong negative relationship.

**Action:** Use DTI as a primary eligibility filter in automated approval systems.

**b. Loan Purpose Analysis -**

* **Asset Purchase** loans show a trend toward approval, although not statistically significant.

**Action**: Consider creating specialized loan offerings or discounts for asset purchases.

**c**. **Home Ownership Insights -**

* **Mortgage Holders** are less likely to receive approvals.

**Action**: Introduce tailored products like mortgage-backed loans or restructured payment plans for this segment.

* **Renters** are viewed more favourably.

**Action**: Explore why renters are approved more frequently and refine criteria for homeowners.

**d. Operational Enhancements -**

* Automate loan approvals based on significant predictors (e.g., high credit score, low DTI, moderate credit utilization).

**Action**: Deploy a predictive model to flag high-probability approvals and reduce

processing time.

**e. Risk Management -**

* **High DTI** and excessive credit utilization are key risk factors.

**Action**: Offer credit counseling and risk-based pricing for high-risk customers.

Monitor customers with significant monthly debts to identify those suitable for cross-selling other financial products.

**f. Policy Refinement -**

* Reassess the role of non-significant variables (e.g., bankruptcies, tax liens) in risk scoring.

**Action**: Explore alternative datasets or variable transformations to improve predictive power.

6. Conclusion:

The aim of the project was to combine factor analysis and logistic regression to create a reliable framework for evaluating financial risk in loan applications. The probability of loan defaults was successfully determined using logistic regression, which is known for its resilience in binary classification situations. Factor analysis made it possible to reduce dimensionality and identify important latent factors—like debt-to-income ratio, income stability, and credit history—that affect financial risk.

The findings showed that combining these techniques can greatly improve loan default prediction interpretability and predictive accuracy. The model provided financial organizations with a useful tool for data-driven decision-making by striking a balance between specificity and sensitivity. The framework ensures better financial resource allocation, streamlines loan approval procedures, and reduces credit risk.

7. Recommendation for future research :

* **Integration of advanced machine learning models:** Adding sophisticated algorithms like ensemble techniques or Gradient Boosting Machines (like XGBoost and LightGBM) could improve predicted accuracy and adaptability to complicated datasets.
* **Dynamic factor models:** Using dynamic factor analysis can help models perform better in changing financial markets by taking time varying elements into consideration.
* **Real-time risk assessment:** By integrating streaming data into real-time scoring systems, it may be possible to obtain immediate insights while processing loan applications.
* **External data incorporation:** The model could be enhanced and non-traditional creditworthiness signals could be captured by incorporating alternative data sources such social media activity, transaction patterns, and utility payment histories.
* **Explainability and AI trust:** By providing clear insights into the variables affecting risk projections, explainable AI (XAI) solutions can increase stakeholders' trust.
* **Fairness and Regulatory Compliance:** Future research should concentrate on making sure the model conforms with ethical norms and financial rules, especially when it comes to biases in loan acceptance decisions.‍
* **Geographic and Demographic Expansion:** By evaluating the model in other geographical locations and demographic groupings, its generalizability can be confirmed and market-specific modifications can be found.
* **Integration with Blockchain Technology:** Investigating blockchain's potential for safe data exchange and improved loan processing transparency can increase public confidence in financial institutions.

# Bibliography

Gouvea, M. A. 2007. *Credit Risk Analysis Applying Logistic Regression.* Neural Networks and Genetic Algorithms Models. POMS 18th Annual Conference.

Jerić, Silvija & vSarlija, Natavsa & vSori'c, Kristina & Rosenzweig, Vivsnja. 2009. *Logistic Regression and Multicriteria Decision Making in Credit Scoring,.* Proceedings of the 10th International Symposium on Operational Research SOR '09.

Karimi, A. 2014. *Credit Risk Modeling for Commercial Banks.* International Journal of Academic Research in Accounting Finance and Management Sciences, .

Joseph F. Hair, William C. Black, Barry J. Babin, Rolph E Anderson. 2019. *Multivariate Data Analysis .* Cengage Learning.

Begiev, Zaur. 2017. *Bank Loan Status Dataset.* Edited by kaggle.com/datasets/zaurbegiev. Kaggle.

Nsengiyera, Daniel. 2020. *THE ASSESSMENT OF LOAN MANAGEMENT ON FINANCIAL PERFORMANCE OF BANKING INSTITUTIONS IN RWANDA CASE STUDY: BANK OF KIGALI HEADQUARTER.* Daniel Nsengiyera University of Science and Technology of China .