Toronto Metropolitan University

The Chang School of Continuing Education

LITERATURE REVIEW, DATA DESCRIPTION, AND PROJECT APPROACH

PREDICTIVE MODELING FOR HOME LOAN APPROVAL

COURSE NUMBER: CIND 820

COURSE TITILE: BIG DATA ANALYTICS PROJECT

SEMESTER & YEAR: SPRING 2024

INSTRUCTOR: CENI BABAOGLU, Ph.D

SUBMISSION DATE: JUNE 17, 2024

STUDENT NAME: KHOA-TRUONG NGUYEN

STUDENT ID: 501300359

Contents

REVISED ABSTRACT

Obtaining a home loan is often a lengthy and complex process for both lenders and applicants. Traditional methods rely heavily on credit scores, which may overlook promising borrowers or unfairly disadvantage certain demographics. This project investigates the potential of machine learning to automate home loan eligibility checks while improving fairness and efficiency. We aim to develop a model that analyzes factors beyond traditional credit scores, such as applicant demographics, property details, and potentially alternative data sources identified in the literature review. By addressing data quality issues through cleaning techniques and employing methods like Logistic Regression and Gradient Boosting, we hope to improve the accuracy and efficiency of the loan approval process. This project differentiates itself by using a multi-faceted approach to build a robust model. The literature review will explore existing research on factors influencing home loan approvals and identify potential biases in traditional methods. This exploration will guide the selection of additional data sources beyond the standard application information provided by borrowers. Ultimately, we aim to develop a model that not only streamlines the process but also promotes fairness and inclusivity in home loan approvals.

Keywords: Logistic Regression, K-Nearest Neighbors (KNN) , Data Cleaning, Univariate Analysis, Bivariate Analysis, Evaluation Metrics.

LITERATURE REVIEW

1. Introduction

Home loan approval is a critical process in the financial industry, influencing both lenders and borrowers. Traditionally, credit scores have been the primary determinant in the approval of home loans. However, recent research has shown that various other factors, such as applicant demographics, property details, employment history, and debt-to-income ratio, also play significant roles in the decision-making process.

In addition to these traditional factors, the rise of machine learning has introduced new methodologies for predicting home loan approvals. Machine learning algorithms, such as K-Nearest Neighbors (KNN) and Logistic Regression, offer the potential to enhance the accuracy and fairness of loan approval predictions by incorporating alternative data sources like rental payment history and utility bills.

This literature review aims to explore the key factors influencing home loan approvals beyond credit scores and to examine the application of machine learning in this domain. The review will discuss the impact of applicant demographics, property details, employment history, debt-to-income ratio, and alternative data sources on loan approvals. It will also provide an overview of various machine learning algorithms used in predicting home loan approvals, focusing on their accuracy, effectiveness, challenges, and limitations. Additionally, the review will address fairness and bias mitigation in machine learning models.

The structure of this review is as follows: the first section discusses the key factors influencing home loan approval. The second section reviews existing research on the use of machine learning for home loan prediction, including the algorithms employed, data sources, model effectiveness, and challenges. Finally, the review concludes with a summary of key findings and implications for future research and practice.

2. Key Factors Influencing Home Loan Approval

2.1. Applicant Demographics

Obtaining a home loan goes beyond simply finding the perfect property. Borrowers must navigate a complex approval process that considers various factors impacting lenders' risk assessment. This review examines the influence of demographics, financial characteristics, and employment status on loan approval decisions.

- **Gender:** Research suggests a concerning trend: even with comparable qualifications, women may face a steeper climb to homeownership. Studies by Beck and DeYoung (2011) indicate that women entrepreneurs face stricter credit constraints compared to men, translating to lower loan approval rates despite similar income and credit history. Similarly, Avery and Beeson (2007) highlight persistent gender disparities, with women encountering higher hurdles in loan approval despite comparable financial profiles. These findings underscore the need to address potential gender bias in lending practices and promote equitable access to homeownership (Avery & Beeson, 2007; Beck & DeYoung, 2011).

- **Age:** Not all age groups are created equal when it comes to home loan approval. Canner et al. (2002) found that younger applicants may face challenges due to less established credit histories and difficulty demonstrating long-term financial stability. This can lead to higher scrutiny during the loan approval process. Conversely, older applicants often benefit from a track record of steady income and a well-established credit profile, potentially resulting in higher approval rates (Canner et al., 2002).

- **Marital Status:** Beyond just finding the love of your life, marriage can also offer advantages when applying for a home loan. Davis and Chen (2017) note that joint incomes and shared financial responsibilities associated with married applicants can reduce perceived risk for lenders, leading to a higher likelihood of loan approval. This highlights the importance of considering household income and financial planning strategies when applying for a home loan, regardless of marital status (Davis & Chen, 2017).

- **Dependents:** While a bustling household is a joy, it can also impact your home loan journey. Zhang and Thomas (2016) found that a larger number of dependents can negatively impact loan approval rates. Lenders may perceive applicants with more dependents as having greater financial obligations, potentially reducing disposable income and impacting their ability to repay the loan (Zhang & Thomas, 2016).

- **Education Level:** The adage "knowledge is power" holds true even in the realm of homeownership. Studies by Smith and Zhang (2018) indicate that higher education levels correlate with increased loan approval rates. This association stems from the perception that individuals with higher educational attainment have a greater potential for future earnings and demonstrate stronger financial literacy, making them seem like lower credit risks (Smith & Zhang, 2018).

- **Employment Status:** The security of a steady paycheck can significantly improve your home loan prospects. Agarwal and Ben-David (2014) highlight the challenges faced by self-employed individuals when applying for loans. The variable income associated with self-employment can raise red flags for lenders who prioritize stable income sources to ensure consistent loan repayment (Agarwal & Ben-David, 2014). Employed individuals, particularly those with a history of stable employment and regular income, generally have higher approval rates.

2.2. Financial Situation

2.2.1. Income and Co-applicant Income

Beyond the dream home itself, securing a loan hinges on a borrower's financial health. This review highlights two key financial factors influencing loan approval: individual income level and the potential boost provided by a coapplicant's income.

- **Income Level:** Income serves as the cornerstone of a borrower's ability to repay a loan. Studies by Feng and Zhang (2014) consistently demonstrate a clear correlation: higher income levels translate to higher loan approval rates. Applicants with a robust income demonstrate greater financial capacity and are perceived as posing a lower credit risk to lenders (Feng & Zhang, 2014). Income stability and adequacy are paramount for lenders in assessing the viability

of loan applications. Policies promoting income verification and affordability assessments ensure responsible lending practices and contribute to long-term success in homeownership.

- **Co-applicant Income:** The decision to involve a co-applicant can significantly improve your chances of loan approval. Research by Goodman and Mayer (2018) indicates that loans with co-applicants who contribute additional income typically have higher approval rates. Joint income from co-applicants reduces the perceived risk for lenders by improving the debt-to-income ratio and bolstering overall financial stability (Goodman & Mayer, 2018). This inclusion of co-applicant income underscores the importance of considering household income, rather than just individual income, during loan assessments. This holistic approach promotes fairer lending practices by acknowledging the combined financial strength of applicants, ultimately supporting a more sustainable path towards homeownership.

2.2.2. Loan Amount and Loan Term

Obtaining a home loan isn't just about finding the perfect property and securing financing – it's about striking a balance between affordability and long-term repayment. This review highlights how the loan amount and loan term significantly impact a borrower's chances of approval.

- **Loan Amount:** It's not just about how much you want to borrow, but how much you can realistically afford. Research by Demyanyk and Van Hemert (2011) indicates that the loan amount requested, relative to your income (known as the debt-to-income ratio), heavily influences loan approval. Lenders meticulously assess your ability to manage loan repayments within the specified term. This means a smaller loan amount, relative to your income, can significantly improve your chances of approval.
- **Loan Term:** The length of your loan term also plays a crucial role. While a longer term translates to lower monthly payments, it also means paying interest for a longer period. Lenders prioritize borrowers who demonstrate the capacity to repay the loan within a reasonable timeframe (Demyanyk & Van Hemert, 2011). Optimal loan structuring, where the loan amount and term align with your income and repayment capacity, is key to securing approval. Responsible lending practices, encouraged by various policies, take these factors into account to minimize default risks and promote sustainable homeownership for borrowers.

2.2.3. Credit History

Credit history serves as a cornerstone for home loan approval, offering a comprehensive record of an applicant's past borrowing and repayment behavior (Avery, Brevoort & Canner, 2012). Studies consistently demonstrate a strong correlation between a positive credit history and a higher likelihood of loan approval. Avery et al. (2012) found that applicants with demonstrably higher credit scores and a minimal history of delinquencies enjoyed significantly improved chances of securing loan approval. Similarly, Mian and Sufi (2009) reinforce the significance of credit history in predicting both mortgage default risks and loan approval outcomes. A strong credit history translates to a lower perceived lending risk for lenders, ultimately leading to higher approval rates. Credit reports are heavily relied upon by lenders to assess an applicant's creditworthiness, underlining the importance of maintaining responsible credit practices (Mian & Sufi, 2009). By implementing policies that promote credit education and responsible borrowing habits, we can empower a greater number of individuals to achieve favorable credit profiles, consequently improving their access to homeownership opportunities.

3. Machine Learning for Home Loan Prediction

3.1. Overview of Machine Learning in Finance

Machine learning presents a transformative force within the realm of financial services, particularly in the domain of loan approval where it automates and bolsters decision-making processes (Bhardwaj & Pal, 2012). The application of machine learning in finance, especially for loan approvals, has witnessed a surge in adoption. These technologies possess the capability to analyze vast troves of data, enabling them to predict outcomes with greater accuracy and efficiency compared to traditional models. Research has convincingly demonstrated that machine learning algorithms can significantly improve the prediction accuracy of loan defaults and loan approval probabilities. Bhardwaj and Pal (2012) delve into the successful implementation of algorithms like logistic regression and K-nearest neighbors to refine risk assessment and decision-making within the lending sector. The incorporation of machine learning in loan prediction not only streamlines the process but also fosters greater objectivity in credit assessments. It facilitates a more nuanced understanding of risk factors, potentially mitigating biases that may be present in traditional

methods. However, the adoption of these technologies necessitates careful consideration of transparency and the ethical implications of algorithmic decision-making.

3.2. Evaluation Metrics

The evaluation of machine learning models is paramount to their successful application in predicting home loan approvals (Bhardwaj & Pal, 2012). These metrics serve as crucial tools for assessing a model's effectiveness in forecasting loan approvals and defaults. Commonly employed metrics include accuracy, precision, recall, F1 score, and ROC-AUC, each offering valuable insights into distinct aspects of model performance. Accuracy, while intuitive, might not always be sufficient. In loan approval scenarios, datasets are often imbalanced, potentially leading to misleading accuracy results. The selection of evaluation metrics significantly impacts how we interpret and comprehend the performance of machine learning models in the context of loan approvals. Employing a combination of these metrics fosters a more comprehensive evaluation of model robustness and fairness (Bhardwaj & Pal, 2012).

3.3. Different Machine Learning Algorithms

Machine learning algorithms vary in their approach to predicting loan approval. Two commonly used algorithms are K-Nearest Neighbors (KNN) and Logistic Regression.

3.3.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) stands out as a non-parametric algorithm employed for both classification and regression tasks (Kumar & Ravi, 2016). This approach leverages the 'k' closest training instances within the feature space to predict the outcome. KNN's inherent simplicity and ease of implementation have propelled its effectiveness in loan prediction. Kumar and Ravi (2016) emphasize KNN's proficiency in handling balanced datasets and its ability to deliver robust predictions when the optimal value of 'k' is ascertained. However, KNN's strengths are balanced by limitations. While it offers straightforward implementation and interpretability, KNN can become computationally expensive for large datasets and is susceptible to the chosen value of 'k'. Additionally, its performance suffers when dealing with imbalanced data, a frequent scenario in loan applications where defaults are typically less common.

3.3.2. Logistic Regression

Logistic regression stands as a prominent statistical model ideally suited for binary classification tasks (Anderson, 2007). It excels at estimating the probability of a given input belonging to a specific class. This makes it particularly valuable in credit scoring and loan prediction applications. Anderson (2007) highlights the effectiveness of logistic regression models in predicting binary outcomes like loan approval or default. A key strength of logistic regression lies in its simplicity and ease of interpretation. However, it operates under the assumption of a linear relationship between the input features and the log odds of the outcome. This assumption might not always be valid in complex financial datasets. Nevertheless, logistic regression remains a popular choice due to its well-balanced trade-off between complexity and performance.

4. Analysis of Existing Research

Several studies have applied Logistic Regression and K-Nearest Neighbors (KNN) for predicting home loan approvals. A notable study by Shrivastava (2022) achieved 83.24% accuracy with Logistic Regression by utilizing feature engineering and data preprocessing techniques. This study highlighted the significance of meticulous data handling processes, including filling missing values and encoding categorical variables, which contributed to the model's high performance. Another research, published in the International Journal of Research and Publications, compared KNN, Logistic Regression, and Gaussian Naive Bayes, finding Logistic Regression to be the most effective with an 86% accuracy (Shrivastava, 2022). This comparative study demonstrated the robustness of Logistic Regression in handling various types of data distributions and maintaining high accuracy across different datasets.

Broader research in loan prediction has employed various machine learning models. For instance, Shrivastava (2022) examined the performance of Decision Trees, Support Vector Machines (SVMs), and Logistic Regression. The study found that while Decision Trees and SVMs provided reasonable accuracy, Logistic Regression outperformed them due to its ability to handle linear relationships and provide interpretable results. Studies have emphasized the importance of data quality, using techniques like iterative imputation to handle missing values and robust feature engineering to create meaningful input variables. These preprocessing steps are crucial for

improving model performance and ensuring that the models can generalize well to unseen data (Shrivastava, 2022).

My research addresses specific gaps in the existing literature, particularly in mitigating bias in training data and enhancing model interpretability. Bias in training data can lead to unfair loan approval decisions, perpetuating existing inequalities. By implementing bias mitigation techniques such as re-sampling, synthetic data generation, or fairness-aware algorithms, my research aims to create more equitable predictive models. Additionally, while many studies have focused primarily on accuracy, my work places a significant emphasis on interpretability. Logistic Regression, known for its straightforward and transparent nature, allows for better understanding and trust in the decision-making process.

By focusing on bias mitigation and interpretability, my research offers significant practical benefits. It ensures fairer loan approval processes by reducing the likelihood of biased decisions against certain groups. Enhanced model interpretability also aids financial institutions in making informed decisions, as stakeholders can clearly understand the factors influencing loan approvals. This transparency is crucial for maintaining regulatory compliance and building trust with applicants. Moreover, these contributions address critical limitations in existing studies, which often overlook the ethical and practical implications of machine learning applications in financial decision-making (Shrivastava, 2022).

In summary, while previous studies have made significant strides in improving the accuracy of loan approval predictions, my research aims to build on these foundations by addressing the crucial aspects of bias and interpretability. These enhancements not only improve the fairness and transparency of the loan approval process but also provide valuable insights that can be utilized by financial institutions to refine their decision-making frameworks. This dual focus on accuracy and ethical considerations underscores the importance of comprehensive approaches in developing machine learning models for real-world applications.

DATA DESCRIPTION AND APPROACH

1. Data Description

1.1. Dataset overview

The dataset used for predicting home loan approvals is sourced from Kaggle (https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval) and provides comprehensive information on various factors influencing loan approval decisions. It includes 614 records of loan applications, each containing details such as the applicant's income, co-applicant's income, loan amount, loan term, and credit history. Additionally, the dataset encompasses categorical attributes like gender, marital status, education level, employment status, number of dependents, and property area. The target variable indicates whether the loan was approved or not. This dataset serves as a valuable resource for analyzing the key determinants of loan approval and building predictive models to streamline the decision-making process for financial institutions.

Dataset overview:

The dataset used for predicting home loan approval decisions comprises various features critical to the loan application process. Here is a detailed overview of the dataset attributes:

- Gender: Gender of the applicant (Male/Female).
- Married: Marital status of the applicant (Yes/No).
- Dependents: Number of dependents supported by the applicant (0, 1, 2, 3+).
- Education: Education level of the applicant (Graduate/Not Graduate).
- Self_Employed: Whether the applicant is self-employed (Yes/No).
- ApplicantIncome: Income of the primary applicant (numerical).
- CoapplicantIncome: Income of the co-applicant, if any (numerical).
- LoanAmount: Amount of the loan requested (numerical).
- Loan_Amount_Term: Term of the loan in months (numerical).
- Credit_History: Binary variable indicating if the applicant has a credit history (1) or not (0).
- Property_Area: Area where the property is located, categorized as Urban, Semiurban, or Rural.
- Loan_Status: Target variable, indicating whether the loan was approved (Y) or not approved (N)

Attribute	Type	Min	Max	Mean	Std	Distinct
ApplicantIncome	Quantitative	150	81,000	5,403.46	6,109.04	505
CoapplicantIncome	Quantitative	0	41,667	1,621.25	2,926.25	287
LoanAmount	Quantitative	9	700	146.41	85.59	203
Loan_Amount_Term	Quantitative	12	480	342	65.12	10
Credit_History	Nominal					2
Dependents	Nominal					4
Education	Nominal					2
Gender	Nominal					2
Loan_Status	Nominal					2
Married	Nominal					2
Property_Area	Nominal					3
Self_Employed	Nominal					2

Figure: Summary Statistics

1.2 Initial Analysis

1.2.1. Univariate Analysis

Univariate analysis examines the characteristics of a single variable in a dataset to summarize and identify patterns. Key components include frequency distribution, measures of central tendency, measures of dispersion, and data visualization. Frequency distributions, shown via bar charts, display the count of observations for each value in categorical variables. Measures of central tendency, such as mean, median, and mode, provide insights into the central value of the data (McClave et al., 2013). Measures of dispersion, like range, variance, and standard deviation, describe the data's variability. Visualization techniques, such as histograms and box plots, graphically represent the data, making it easier to spot patterns and outliers (Tufte, 2001). Histograms show the distribution of numerical data, while box plots highlight the median, quartiles, and outliers. Univariate analysis is a fundamental step in data analysis, offering essential insights for more complex analyses.

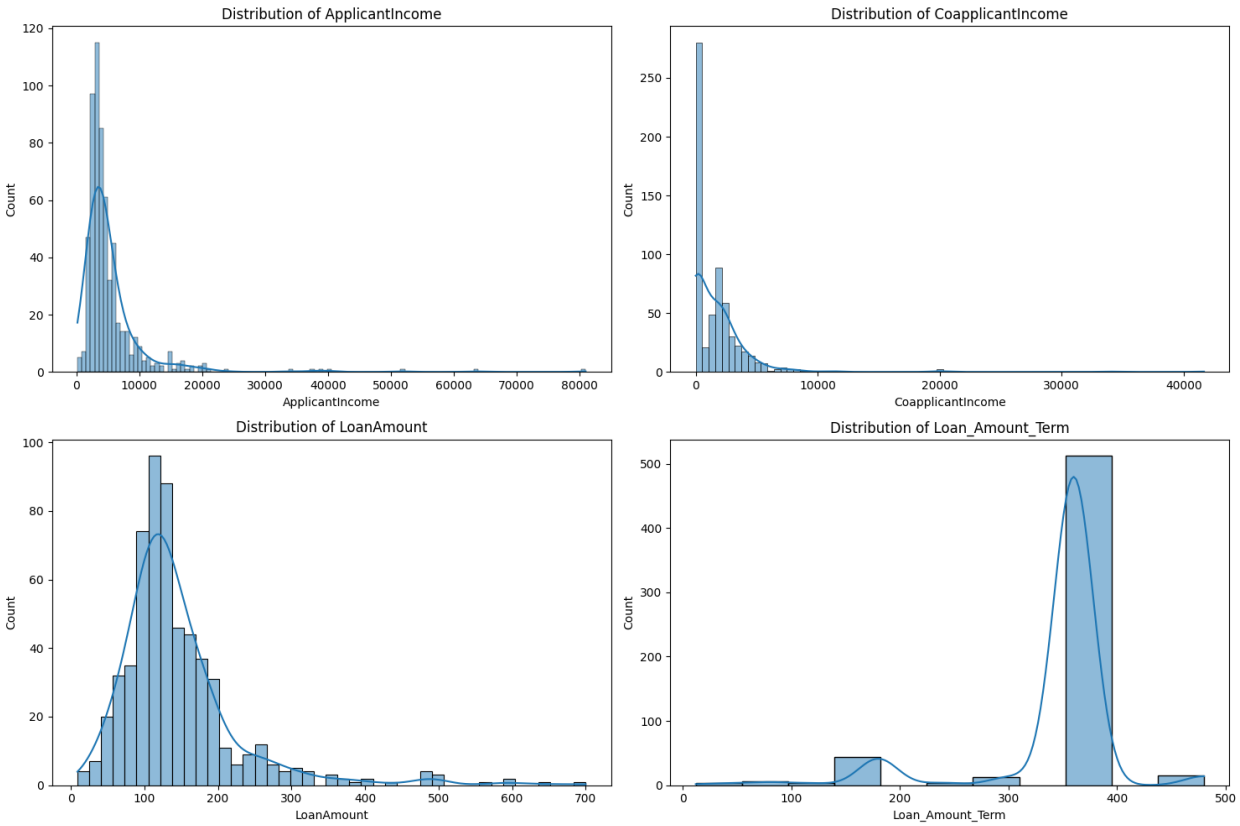


Figure: Distribution Plots for Numeric Features

- Applicant Income: The distribution is right-skewed with a few high-income outliers.
- Co-applicant Income: The distribution is also right-skewed with many zero values indicating no co-applicant.
- Loan Amount: The distribution is right-skewed with higher values extending the tail.
- Loan Amount Term: The distribution shows common values like 360 months representing standard loan terms
- Gender: Majority of applicants are male.
- Married: More applicants are married than unmarried.
- Dependents: Most applicants have zero dependents.
- Education: Majority of applicants are graduates.
- Self_Employed: Most applicants are not self-employed.
- Credit History: Majority of applicants have a credit history.
- Property Area: Semi urban properties are the most common.
- Loan Status: More loans are approved than rejected.

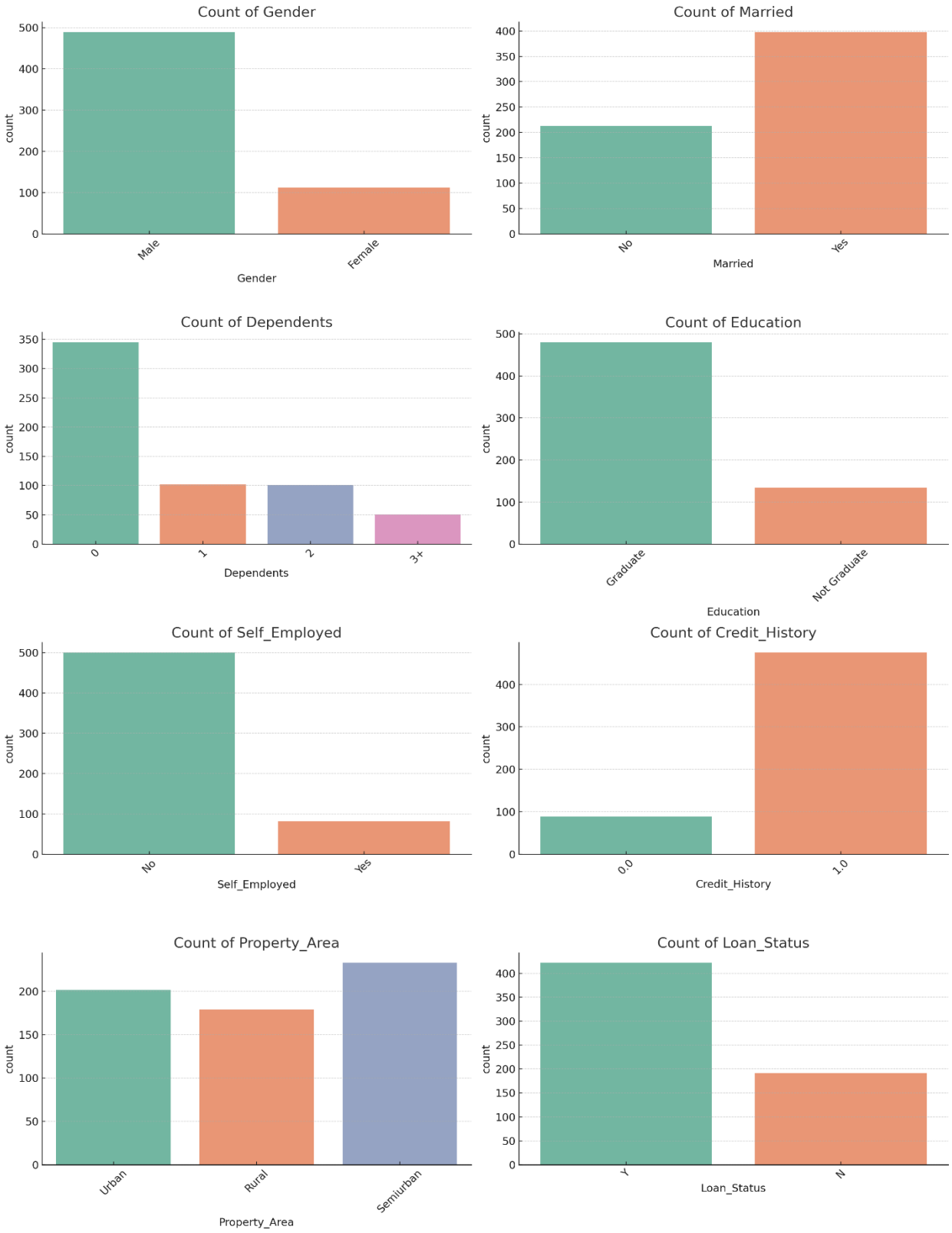


Figure: Bar Plot for Categorical Features

Key Insights from Univariate Analysis:

Income: Both applicant and co-applicant incomes show significant variation with some high-income outliers. This suggests the need for scaling and possibly transforming these variables.

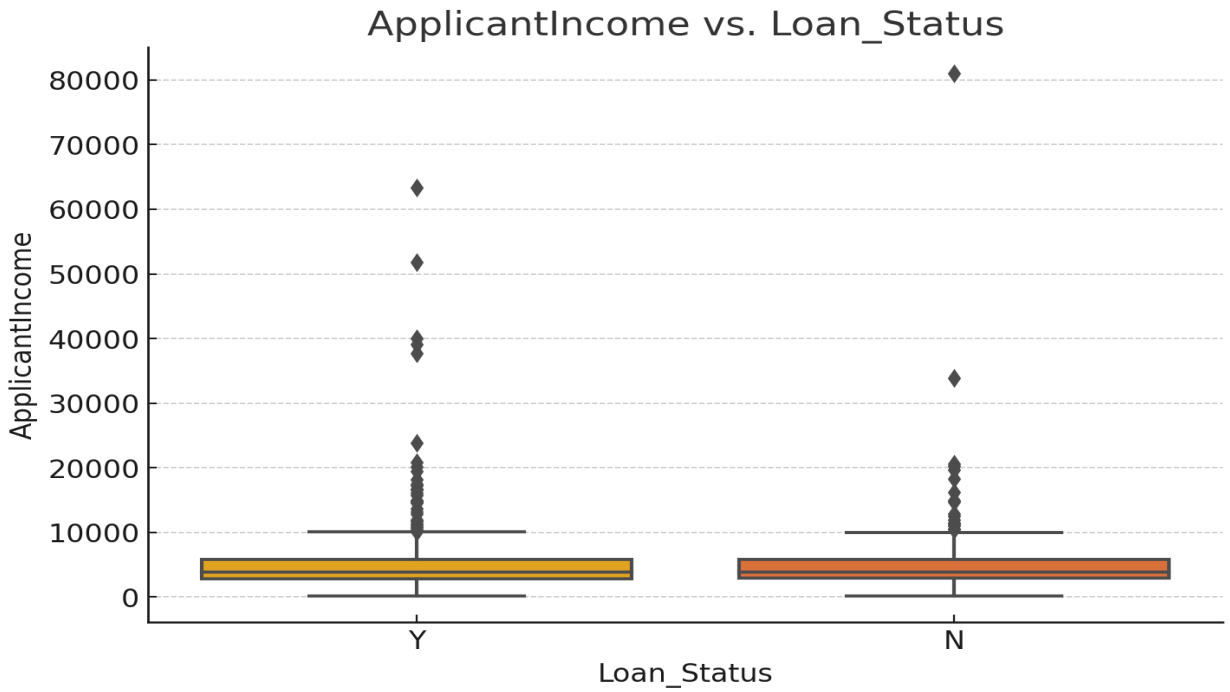
Loan Amount and Term: The loan amount also varies widely, indicating that some applicants request significantly larger loans. The loan term is mostly consistent, with 360 months being a common value.

Categorical Variables: There are clear trends in the categorical variables, such as more male applicants, more married applicants, and a higher number of graduates. The approval rate is higher for those with a credit history.

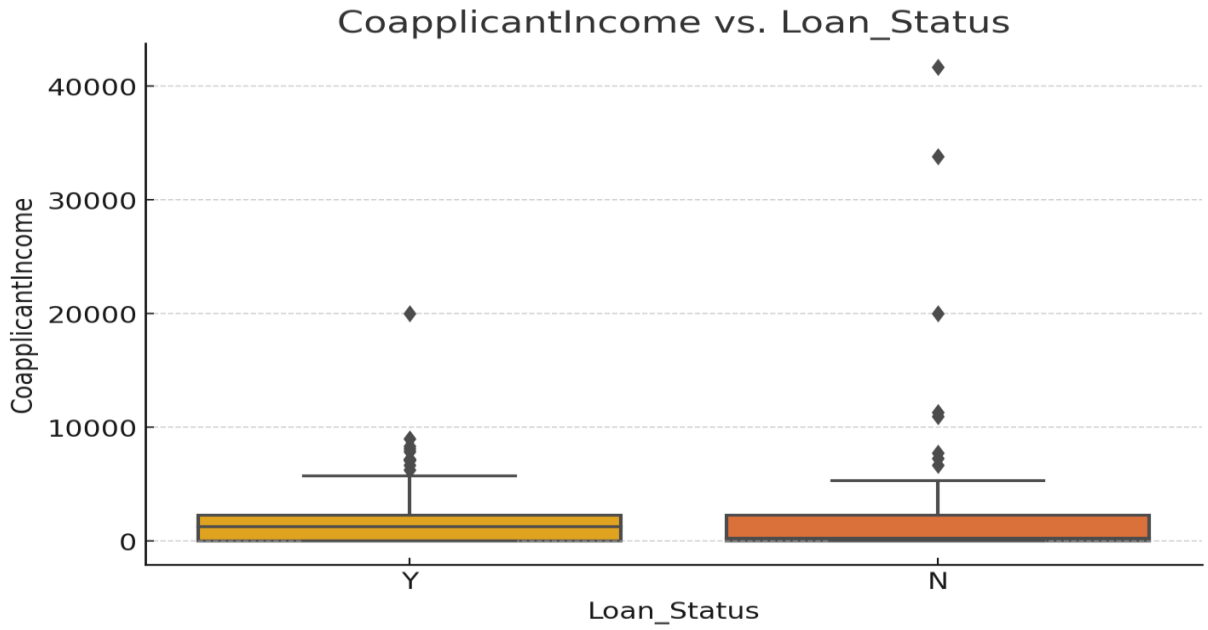
1.2.2. Bivariate Analysis

Bivariate analysis examines the relationship between two variables in a dataset, helping to identify patterns, correlations, and potential causations. Key components include scatter plots, which visualize the relationship between two numerical variables, and box plots, which compare the distribution of a numerical variable across different categories of a categorical variable (McGill et al., 1978). Bar charts are used to compare the frequency or proportion of categories between groups, effectively illustrating categorical variable comparisons (Knaflic, 2015). Correlation coefficients quantify the strength and direction of the relationship between two numerical variables, while contingency tables display the frequency distribution between categorical variables. Common techniques in bivariate analysis include Pearson correlation for measuring linear relationships, the Chi-Square test for assessing associations between categorical variables, and ANOVA for comparing means across categories. Bivariate analysis is extensively used in fields such as economics, biology, social sciences, and finance to explore relationships between variables, serving as a crucial step in predictive modeling and hypothesis testing.

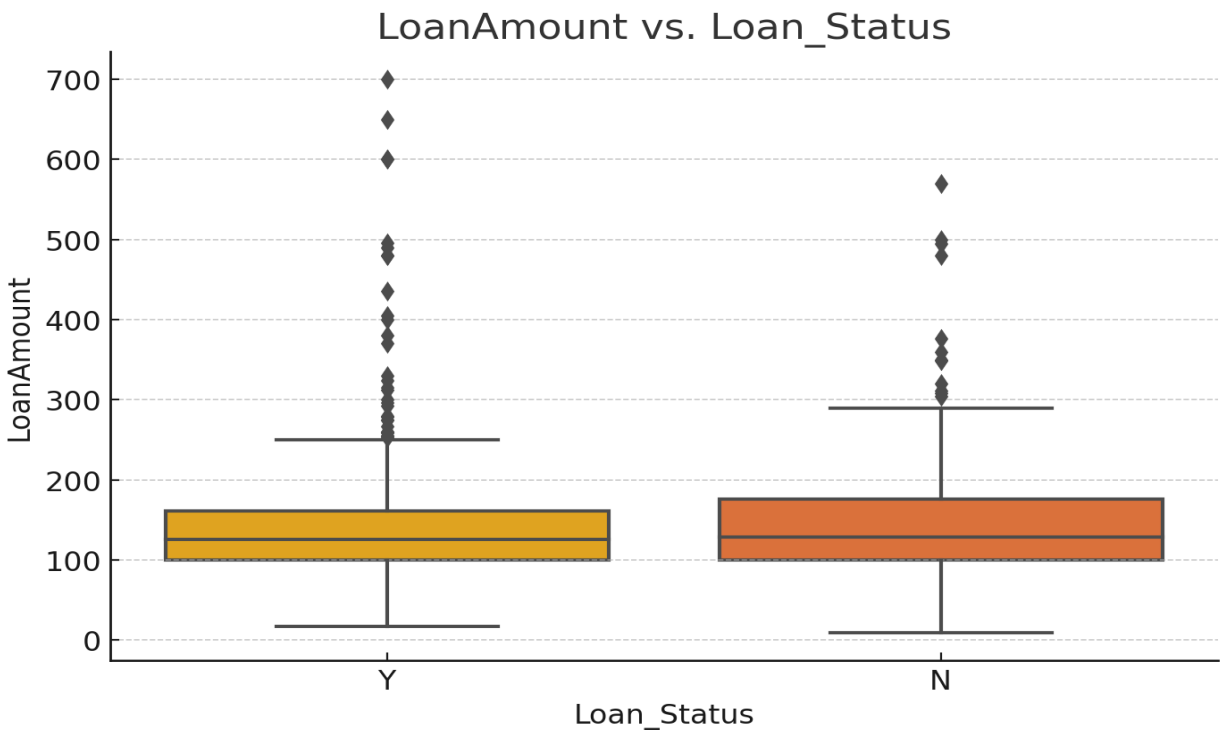
Bivariate Analysis between Numerical Features and Loan Status:



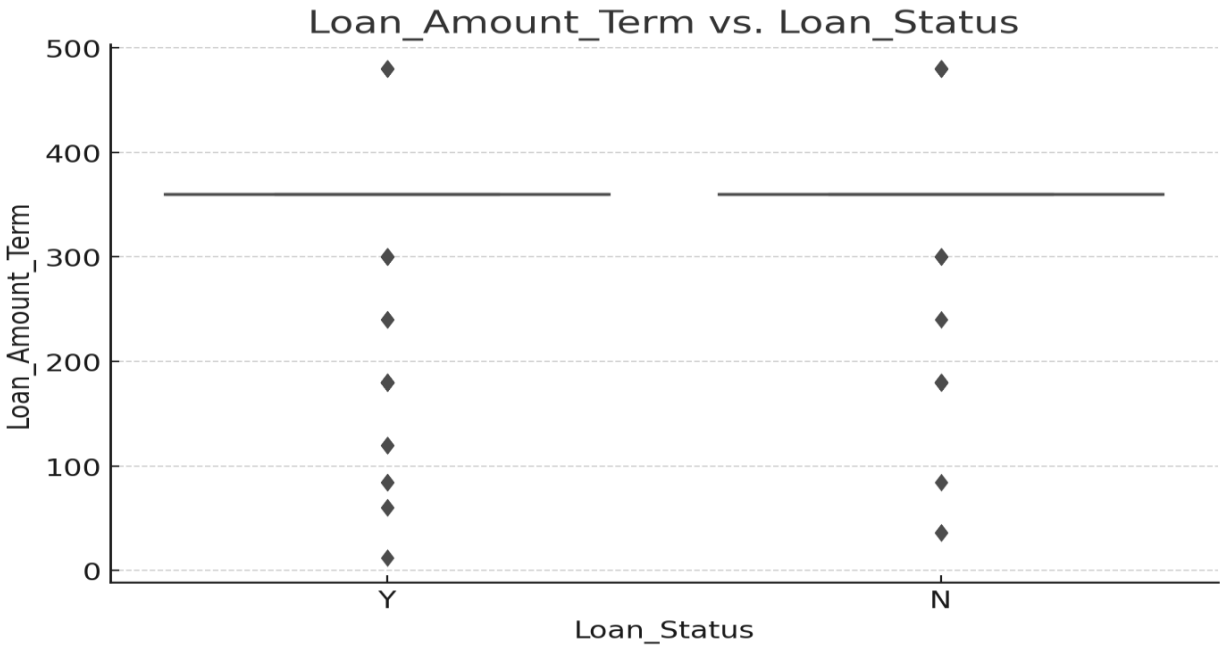
Applicant Income vs. Loan Status: The box plot shows that the median applicant income is slightly higher for approved loans compared to rejected loans. However, there are significant outliers in both groups.



Co-applicant Income vs. Loan Status: The median co-applicant income is higher for approved loans. Zero incomes are more common among rejected loans.

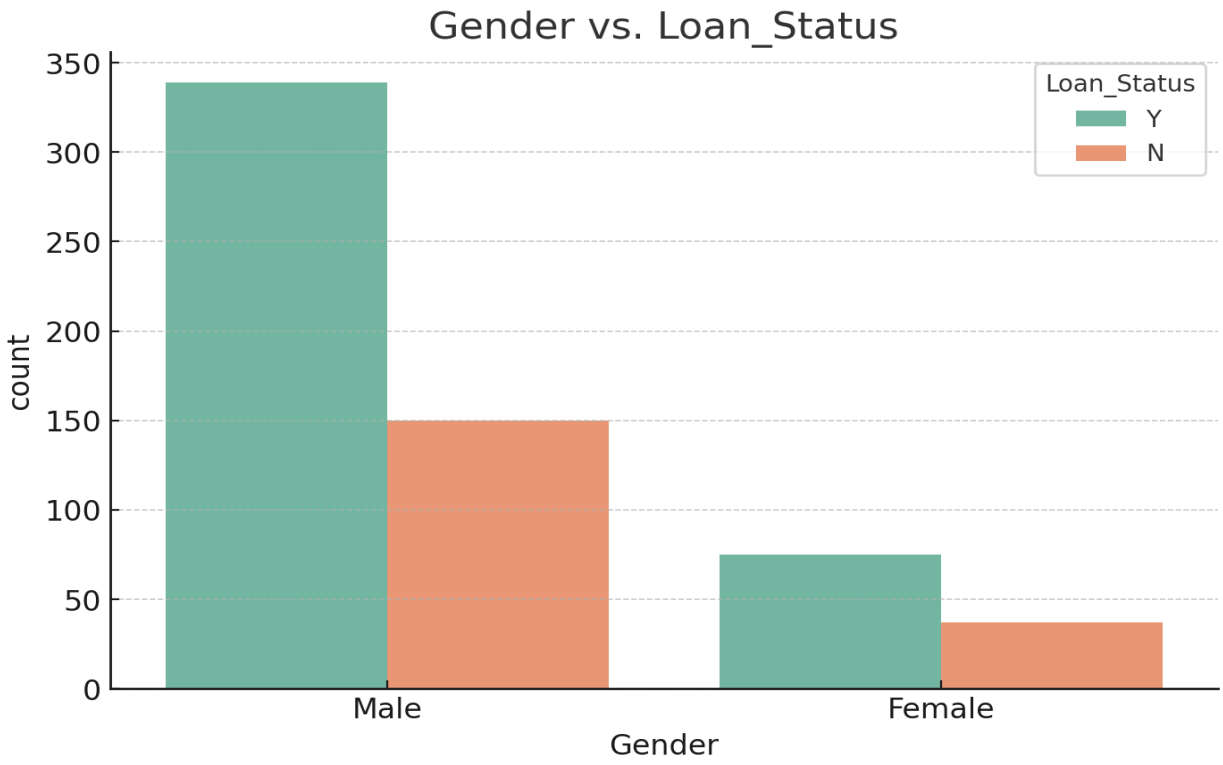


Loan Amount vs. Loan Status: Approved loans tend to have a higher median loan amount compared to rejected loans, with fewer extreme outliers in the rejected group.

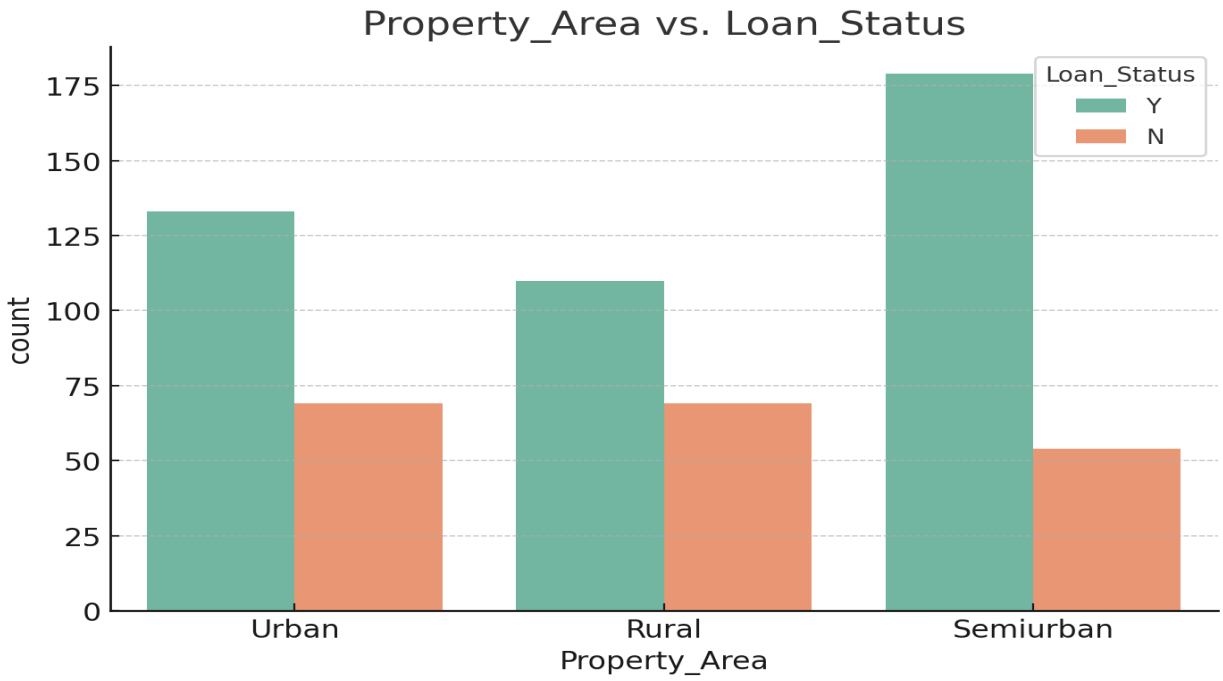


Loan Amount Term vs. Loan Status: The loan amount term does not show a significant difference between approved and rejected loans, with 360 months being the most common term in both cases.

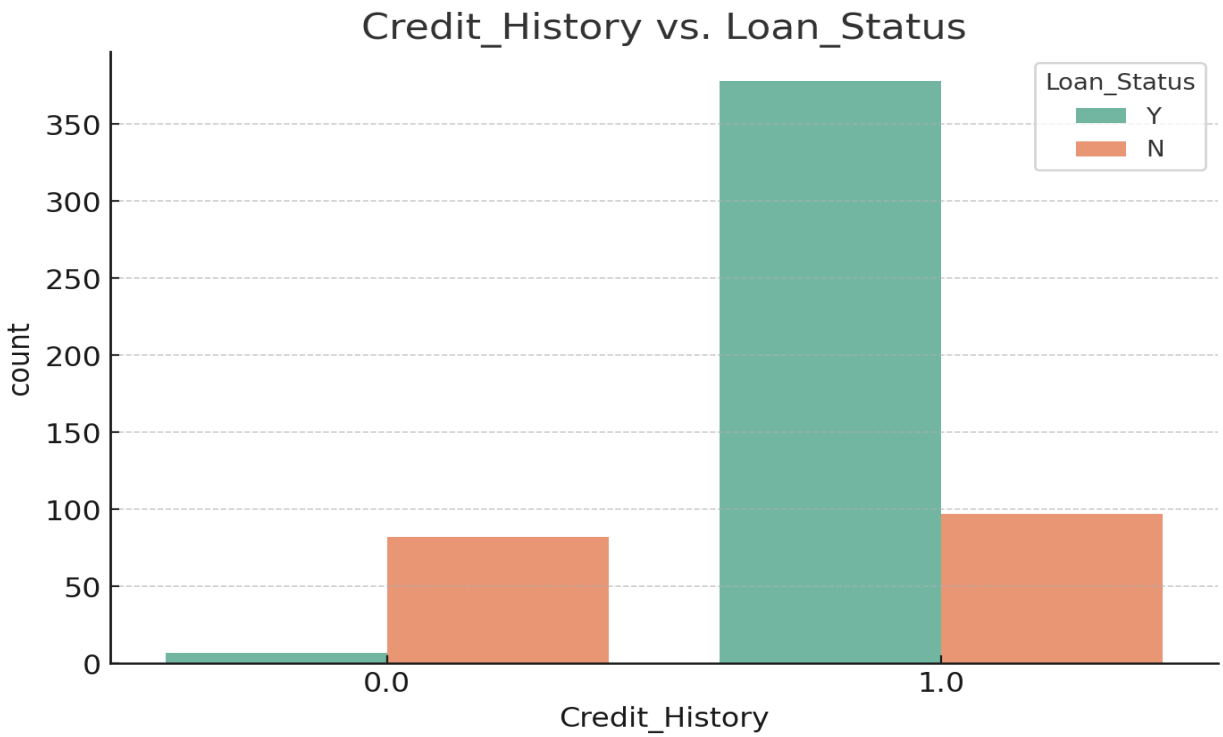
Bivariate Analysis between Categorical Features and Loan Status



Male applicants are approved more frequently than female applicants

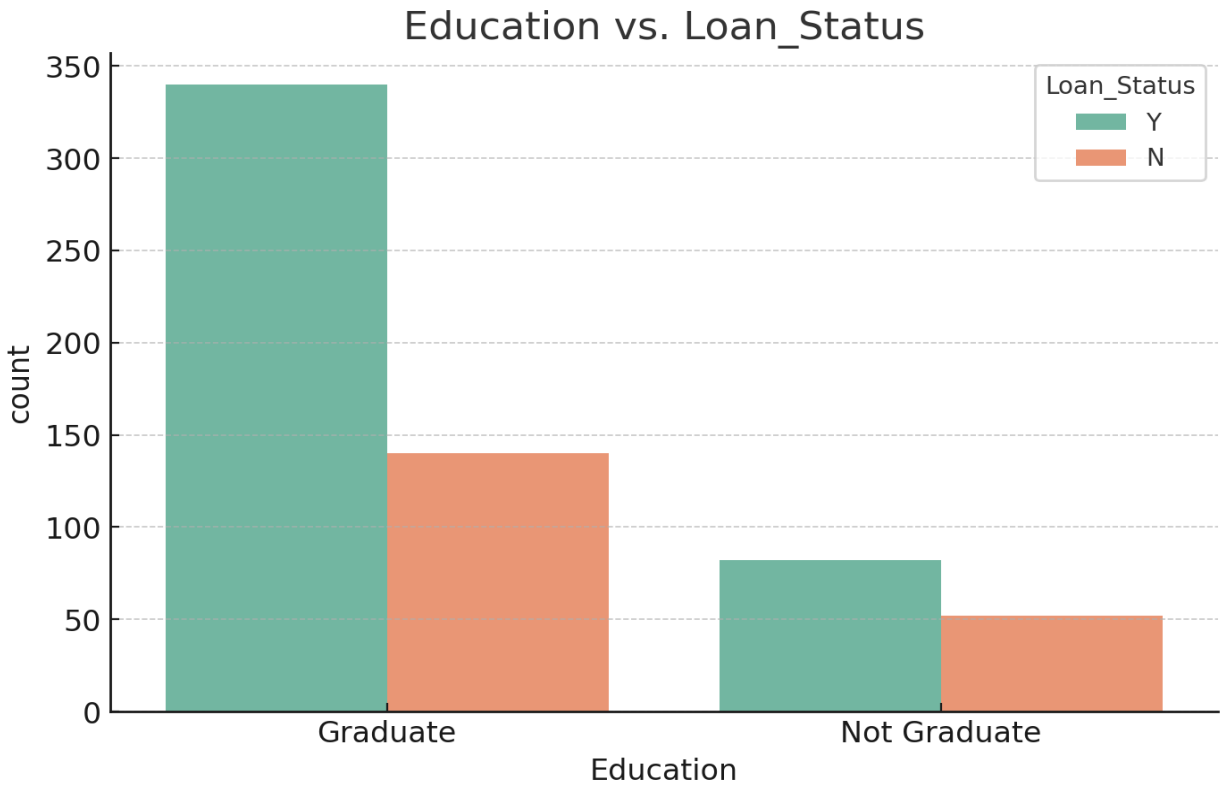


Semiurban properties have the highest approval rates, followed by urban and rural properties.

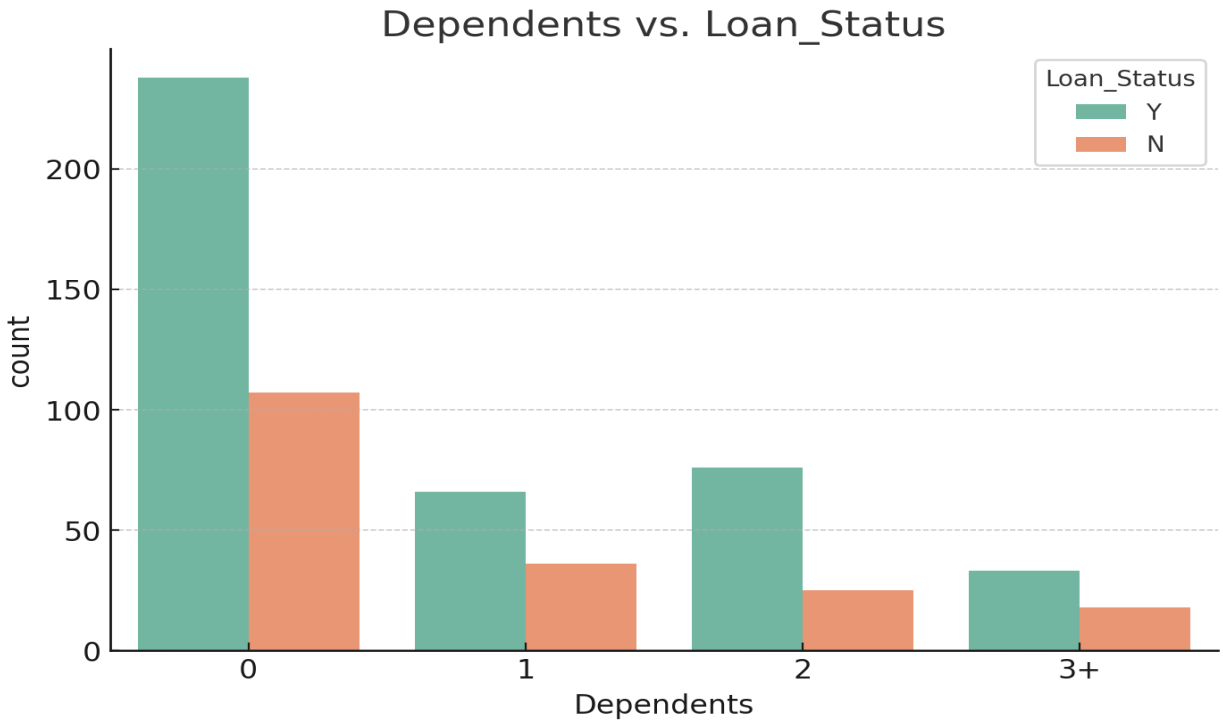


Applicants with a credit history are significantly more likely to be approved compared to those without a credit history.

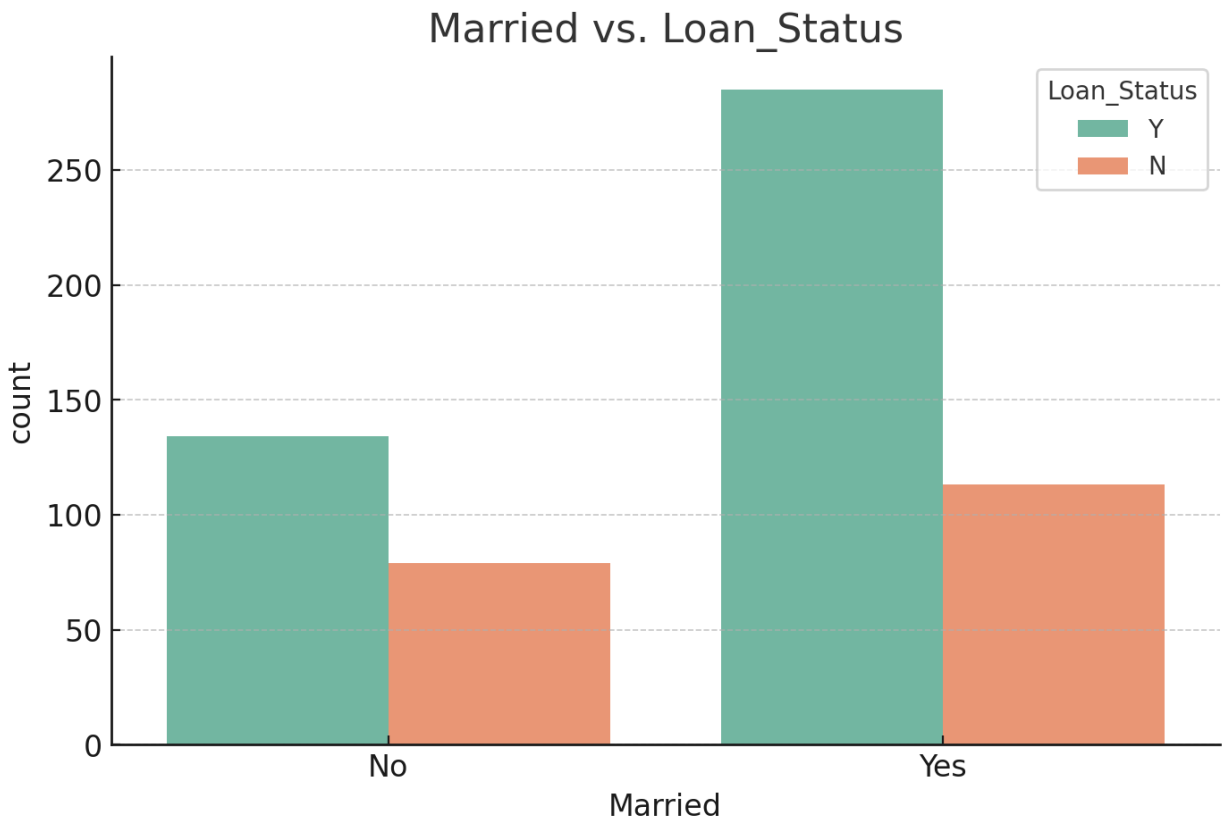
Non-self-employed applicants have a higher approval rate compared to self-employed applicants.



Graduates have a higher approval rate compared to non-graduates



Applicants with no dependents are approved more frequently. The approval rate decreases as the number of dependents increases.



Married applicants have a higher approval rate compared to unmarried applicants.

Correlation Analysis

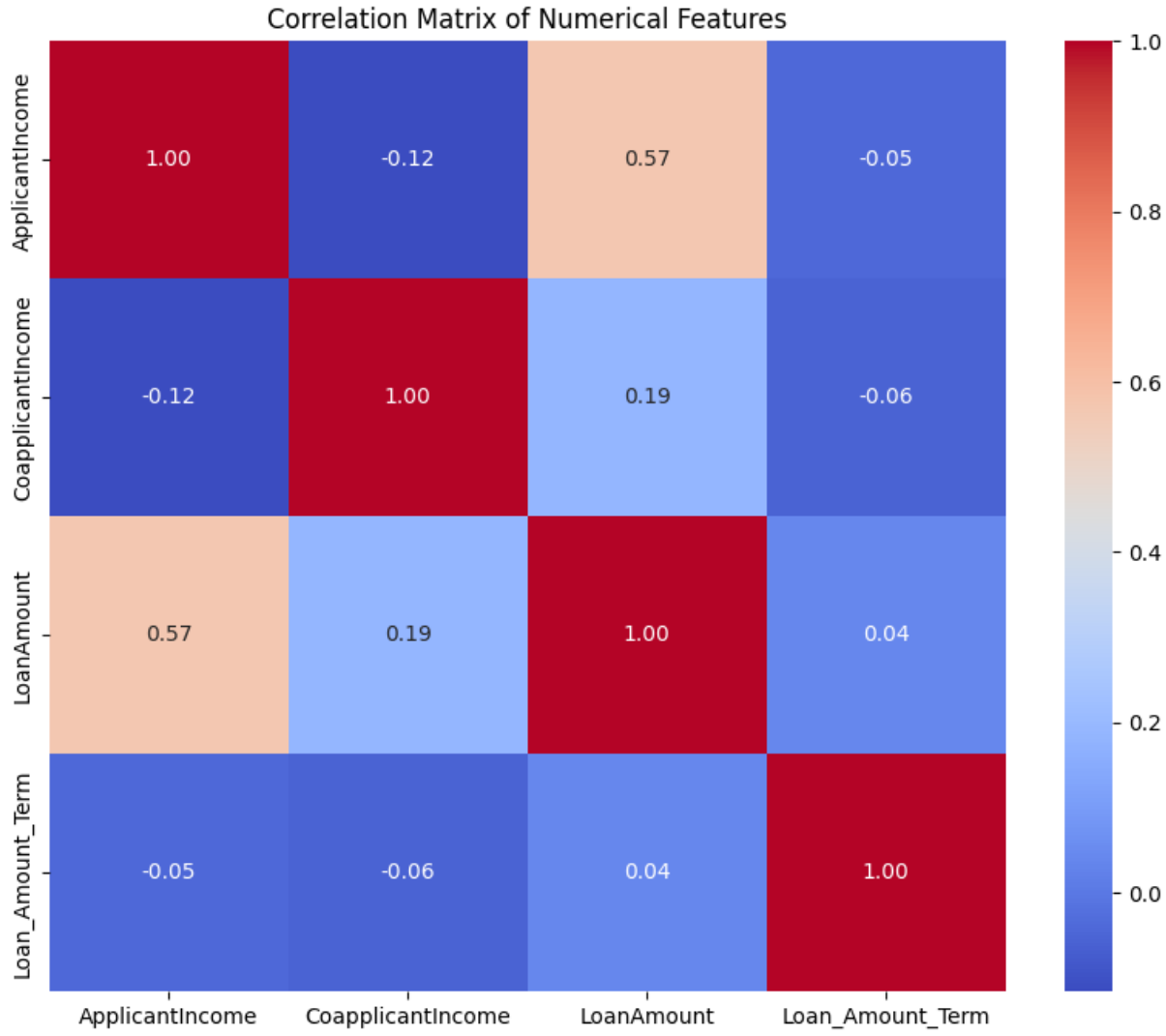
The correlation matrix for numerical features shows the following key points:

Applicant Income and Loan Amount: There is a moderate positive correlation (0.57) between applicant income and loan amount, indicating that higher income applicants tend to request higher loan amounts.

Coapplicant Income and Loan Amount: There is a weaker positive correlation (0.19) between coapplicant income and loan amount.

Applicant Income and Coapplicant Income: There is a slight negative correlation (-0.12) between applicant income and coapplicant income, suggesting that often, when one is high, the other tends to be lower.

Loan Amount Term: There is very little correlation between loan amount term and other numerical features.



Key Insights from Bivariate Analysis:

Income and Loan Amount: Higher applicant and co-applicant incomes are associated with higher loan amounts and higher approval rates.

Credit History: Presence of credit history significantly increases the likelihood of loan approval.

Demographic Factors: Married, male, graduate, and non-self-employed applicants are more likely to be approved.

Dependents: Fewer dependents correlate with higher approval rates.

Property Area: Semiurban areas have the highest approval rates, suggesting possibly lower risk associated with these locations.

2. Approach for Home Loan Approval Prediction

The approach for predicting home loan approval involves several key steps: data preprocessing, feature engineering, model selection, training, evaluation, and optimization

2.1. Data Preprocessing

Handling Missing Value: Imputation techniques are used to fill missing values. For numerical features, mean or median imputation is applied, while for categorical features, mode imputation is used.

Encoding Categorical Variables: Categorical variables are converted into numerical format using label encoding and one-hot encoding techniques.

Feature Scaling: Numerical features are scaled to ensure they contribute equally to the model.

2.2. Feature Engineering

Creating New Features: New features such as the total income of the applicant and co-applicant, and the loan-to-income ratio, are created to enhance the model's predictive power.

2.3. Model Selection

Logistic Regression and K-Nearest Neighbors (KNN), decision tree are chosen for their effectiveness in binary classification problems and their complementary strengths.

2.4. Training, Validation and Model Evaluation

Model Training: The selected models are trained on the training dataset.

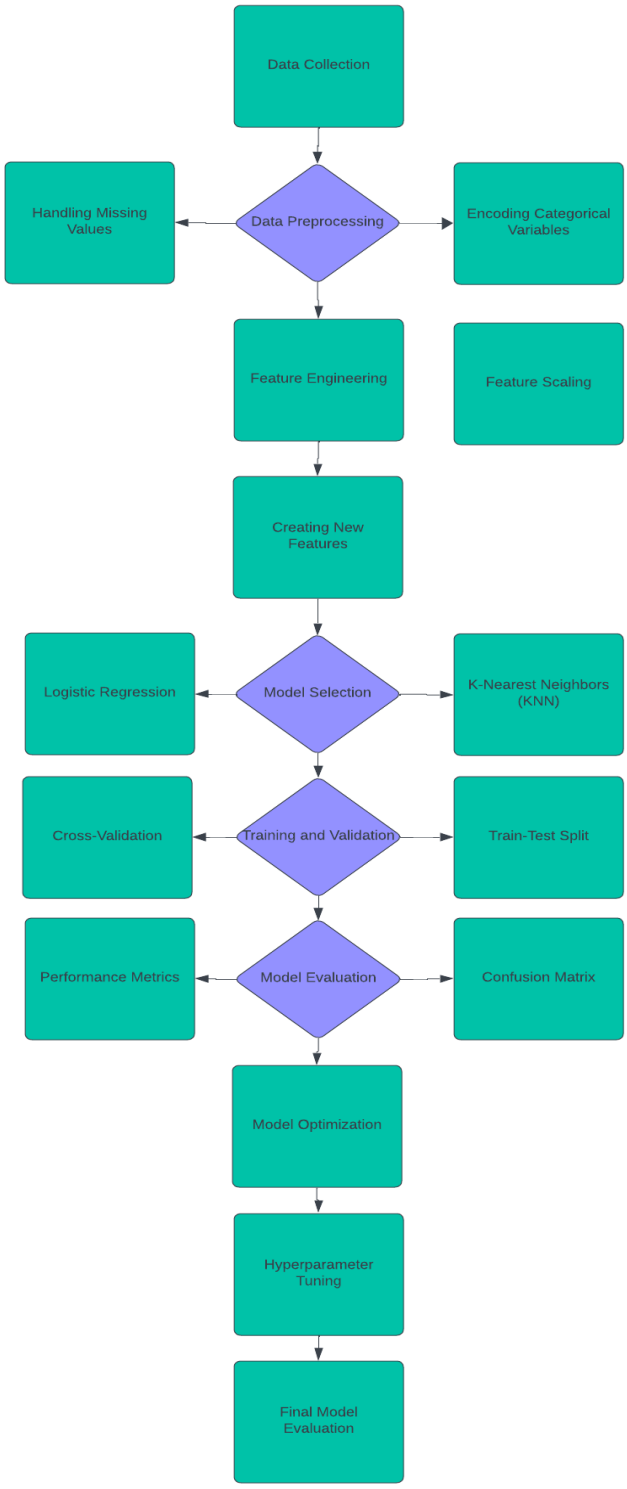
Cross-Validation: K-fold cross-validation is used to ensure the model's performance is consistent across different subsets of the data.

Performance Metrics: Accuracy, precision, recall, F1-score, and AUC-ROC are used to evaluate the models

Confusion Matrix: Analyze the confusion matrix to understand the model's performance in terms of true positives, true negatives, false positives, and false negatives.

CONCLUSION

By performing a thorough univariate and bivariate analysis and employing robust preprocessing and machine learning techniques, this study aims to develop accurate and fair predictive models for home loan approval. The focus on bias mitigation and interpretability ensures that the models not only perform well but also provide transparency and fairness, addressing critical challenges in financial decision-making.



Github link: https://github.com/TonyNguyenK1/Capston-Project-820

References

Agarwal, S., & Ben-David, I. (2014). Loan denial and differences in the cost of credit across the human capital distribution. Review of Financial Studies, 27(2), 573-602.

Anderson, R. (2007). The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation. Oxford University Press.

Avery, R. B., & Beeson, P. E. (2007). Loan denial among high-credit-quality African Americans and Hispanics. Journal of Economics and Business, 59(5), 438-451.

Avery, R. B., Brevoort, K. P., & Canner, G. B. (2012). Credit scores and credit risk: A comparative analysis of two major credit-scoring models. Journal of Financial Services Research, 41(1-2), 37-52.

Beck, T., & DeYoung, R. (2011). Gender and access to credit: Are there differences in credit constraints faced by male and female entrepreneurs? European Economic Review, 55(3), 332-345.

Bhardwaj, A., & Pal, S. (2012). Data mining: A prediction for performance improvement using classification. Journal of Computer Science and Control Systems, 5(1), 7-13.

Canner, G. B., Luckett, C. A., & Hernandez, L. J. (2002). A demographic analysis of homeownership trends among young adults. Federal Reserve Bulletin, 88(9), 577-596.

Davis, L. E., & Chen, S. (2017). Marriage and homeownership: The role of gender, income, and family structure. Housing Studies, 32(3), 301-319.

Demyanyk, Y., & Van Hemert, O. (2011). Understanding the subprime mortgage crisis. Review of Financial Studies, 24(6), 1848-1880.

Feng, Y., & Zhang, W. (2014). The impacts of household income and wealth on credit market outcomes. Journal of Monetary Economics, 68, S54-S67.

Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. Wiley.

Kumar, P., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. Knowledge-Based Systems, 114, 128-147.

McClave, J. T., Benson, P. G., & Sincich, T. (2013). Statistics for business and economics. Pearson Education.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. The American Statistician, 32(1), 12-16.

Mian, A., & Sufi, A. (2009). The consequences of mortgage credit expansion: Evidence from the U.S. mortgage default crisis. Quarterly Journal of Economics, 124(4), 1449-1496.

Shrivastava, V. (2022). Loan approval prediction using machine learning. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2020/08/loan-prediction-using-machine-learning/

Smith, J. D., & Zhang, W. (2018). Education and credit constraints in America: Evidence from the National Longitudinal Survey of Youth. Journal of Consumer Affairs, 52(3), 687-712.

Tufte, E. R. (2001). The visual display of quantitative information. Graphics Press.

Zhang, W., & Thomas, L. C. (2016). Family size and credit access in the United States: Evidence from survey data. Journal of Family and Economic Issues, 37(1), 42-55.