Predicting Loan Approval Outcomes: An Analytical Approach to Enhancing Financial Decision-Making

Prepared by

Ankit Malhotra, Nathaniel Zhu

STAT 206 Final Report

University of California, Riverside

# **Introduction and Statement of Problem**

In the ever-evolving landscape of the financial sector, the efficiency of loan approval processes is an important factor in determining an institution's success and its clients' financial health. The challenge includes not only in managing the potential inherent risks associated with lending but also in ensuring equitable access to financial resources for all applicants. Traditional loan approval mechanisms rely heavily on manual assessments and simplistic criteria and are increasingly inadequate to address the complex interplay of factors that influence loan outcomes. This report is primarily focused on the hypothesis that a data-driven, analytical approach can significantly enhance the accuracy and fairness of loan approval decisions. Specifically, it delves into the domain of property loans, aiming to unveil the complicated criteria which affect the approval process and to propose a predictive model that could serve as a decision-support tool for financial institutions. By doing so, it seeks to mitigate risk, optimize resource allocation, and foster transparency and inclusivity in the loan approval realm.

Throughout this paper, we are going to analyze the background of the dataset we gathered from Kaggle.com, and a detailed approach to data modeling analysis. Next, let us delve into Data Analysis.

# **Data Analysis**

## **Overview**

The data set supporting this study consists of 381 loan applications, each described by 13 variables capturing a wide range of applicant characteristics and loan parameters. This data provides a unique vantage point into the factors that influence real estate loan approvals. At the heart of our investigation is the “Loan\_Status” variable as a binary indicator of the approval outcome (1 for approval, 0 for rejection). The comprehensive nature of the data set, covering both demographic and financial dimensions, facilitates a holistic analysis of the approval process.

## Data Preparation

An initial review of the data set revealed missing values for multiple variables, a common challenge in data analysis that can significantly affect the results if not handled properly. Given the relatively small dataset, and in order to maintain the completeness and accuracy of our findings, we chose to exclude records with incomplete information. This decision reduced the data set to 308 complete records, a necessary compromise to ensure the reliability of the analysis.

Further preparatory steps include converting categorical variables into numeric format, thus enhancing the suitability of the dataset for statistical analysis and machine learning applications. The encoding proceeds as follows:

* **Loan\_Status**: Transformed from "N" (No) and "Y" (Yes) to 0 and 1, respectively, to indicate loan approval outcomes.
* **Gender, Married, Education, Self\_Employed**: These variables were encoded into binary format (0 and 1) to reflect the dichotomous nature of their categories (e.g., Male/Female, No/Yes).
* **Property\_Area**: Categorized into Urban (2), Semiurban (1) and Rural (0) to analyze the impact of location on loan approval.
* **Dependents**: The “3+” category was converted to a numeric value of 3 to facilitate quantitative analysis.

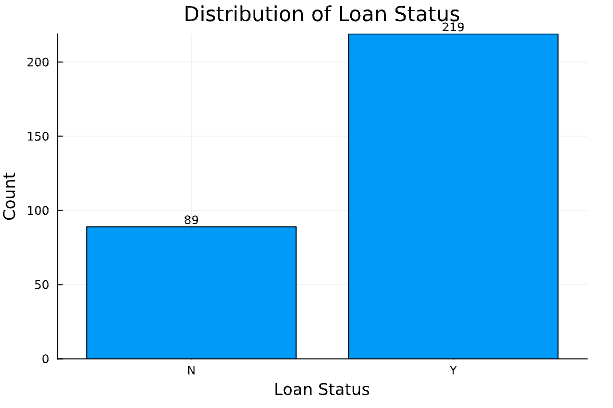
This transformation process not only simplifies the analysis process, but also provides the basis for a detailed exploration of how each variable potentially affects the loan approval decision.

## Visualization

With a clean and properly formatted data set, the next stage of our analysis will utilize graphical methods to further dissect the data set. Visualization is a powerful tool for identifying trends, patterns, and outliers, providing insights that can guide the development of predictive models. We'll use a variety of graphs to explore the relationship between applicant attributes, loan characteristics, and approval outcomes. By doing so, we aim to uncover the underlying dynamics of the loan approval process, revealing factors that significantly influence decisions.

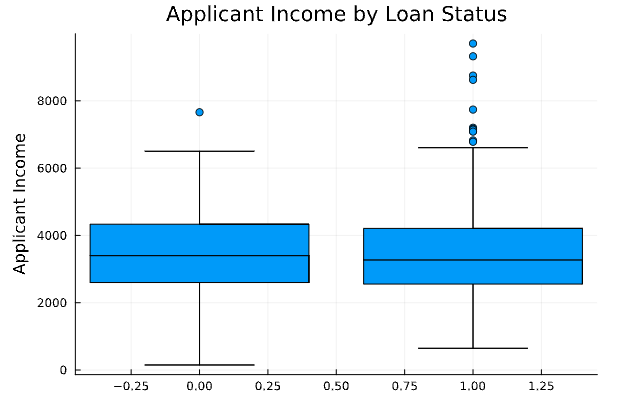
Through detailed exploration of the dataset, we lay the foundation for the subsequent model building phase, and these insights will inform the development of a predictive framework designed to predict loan approval outcomes with high accuracy.

**Figure 1: Distribution of Loan Status**



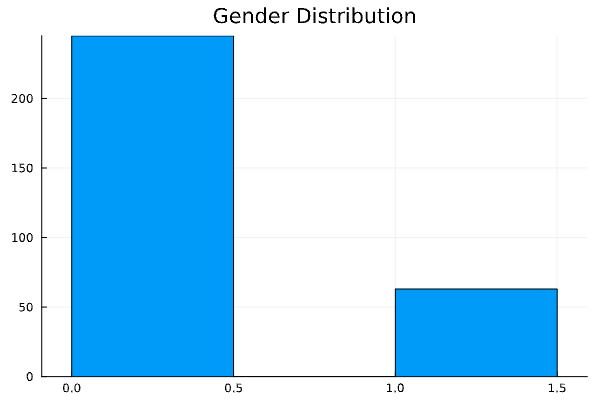
The bar chart clearly shows the distribution of loan status among applicants. Of the 308 cases, the majority were approved for loans, which had 219 approvals, represented by "Y", while a very small number (89 rejections) were not so lucky. This preliminary observation suggests that there is a trend in loan approvals in our data set, which may point to a generally credit-worthy applicant pool.

**Figure 2: Applicant Income by Loan Status**



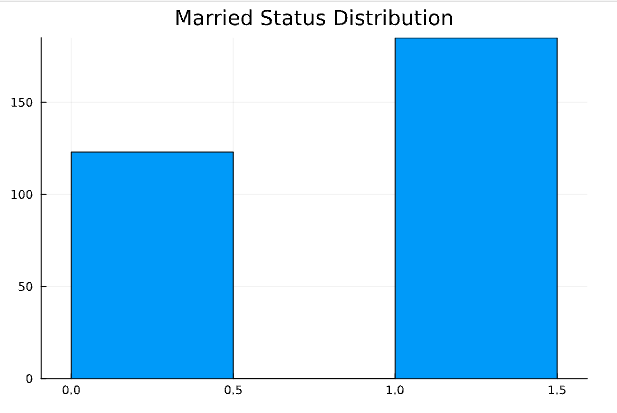
The box plot comparing applicant income to loan status reveals that both approved and rejected groups have similar income medians, with a slightly lower median for approved loans. However, the presence of outliers in the approved category suggests that income alone is not the sole determinant of loan status, and higher earnings do not guarantee loan approval.

**Figure 3: Gender Distribution**



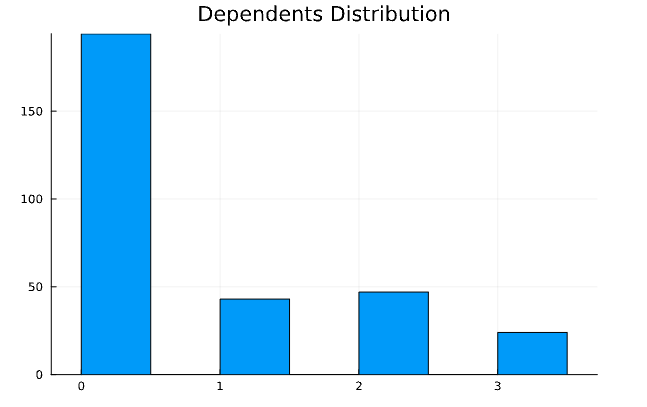
The gender distribution bar chart indicates a substantial difference in the number of male and female applicants, with males significantly outnumbering females. This discrepancy may reflect broader societal trends or could suggest a gender imbalance within the applicant pool.

**Figure 4: Married Status Distribution**



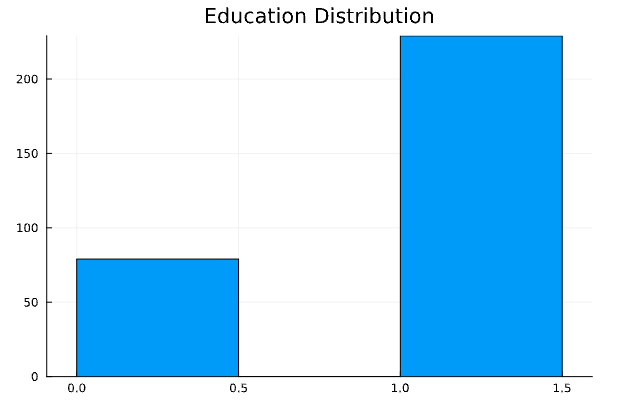
The visualization of marital status among applicants shows that married individuals are more common than unmarried ones. This trend could imply that married applicants are either more likely to apply for loans or are seen as more stable by lending institutions.

**Figure 5: Dependents Distribution**



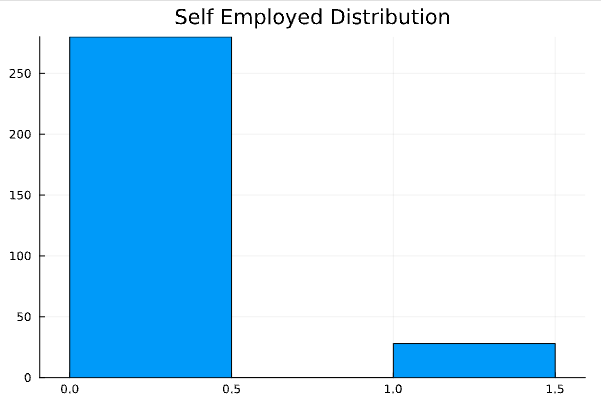
Observing the dependents distribution, we find most applicants report zero dependents, with progressively fewer applicants as the number of dependents increases. This could imply that individuals with fewer financial dependents are more inclined to apply for property loans, or it may simply reflect demographic trends within the dataset.

**Figure 6: Education Distribution**



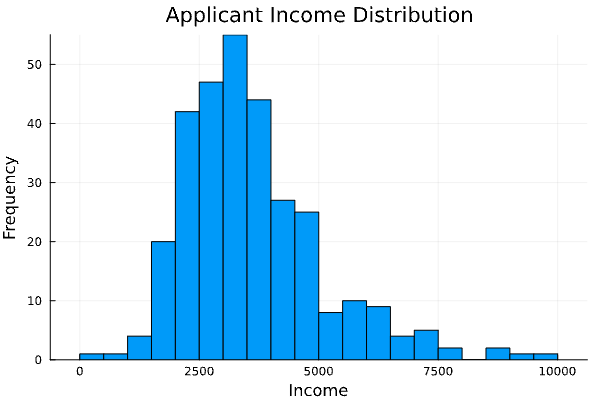
A striking contrast is visible in the education distribution chart, where graduates outnumber non-graduates significantly. This may hint at a correlation between educational attainment and the likelihood of applying for a property loan, potentially driven by factors like income level and financial literacy.

**Figure 7: Self-Employed Distribution**



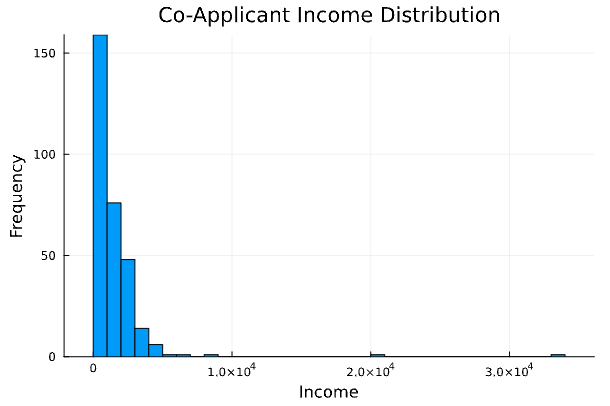
The self-employed distribution reveals a relatively small number of self-employed individuals compared to non-self-employed applicants. This could indicate a variety of socio-economic dynamics, including the possibility that self-employed applicants are less likely to seek property loans or are underrepresented in the applicant pool.

**Figure 8: Applicant Income Distribution**



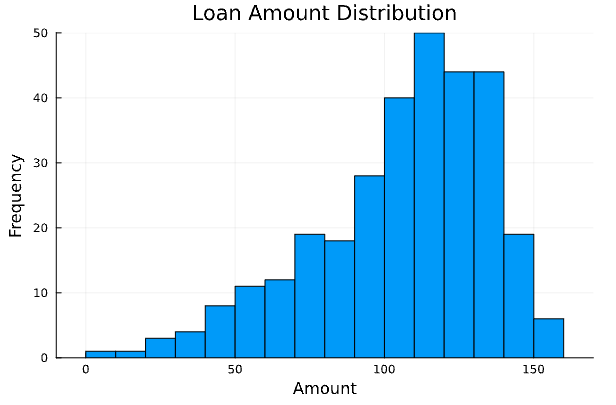
The histogram of applicant income shows a right-skewed distribution, indicating that the majority of applicants fall into the lower income brackets, with fewer high-earning applicants. This skewness may influence the loan approval process, particularly if income is a strong determinant of creditworthiness.

**Figure 9: Co-Applicant Income Distribution**



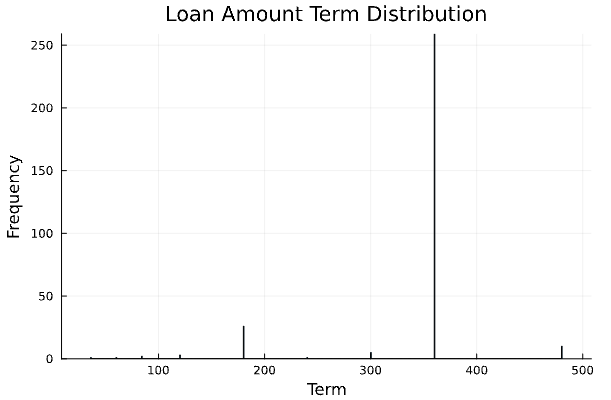
Similar to the primary applicant's income, the co-applicant income histogram is also right-skewed, although to a less pronounced degree. This suggests that most co-applicants, like applicants, tend to have lower incomes.

**Figure 10: Loan Amount Distribution**



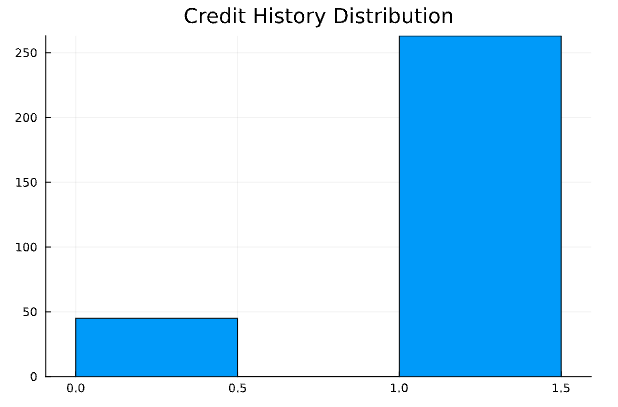
Here, one of the most important determinants of getting a loan, the loan amount distribution histogram shows a relatively normal distribution, skewed slightly right. The concentration of loan amounts in the middle range suggests that the bank typically deals with moderate loan values, which could be indicative of its target market or risk management strategy.

**Figure 11: Loan Amount Term Distribution**



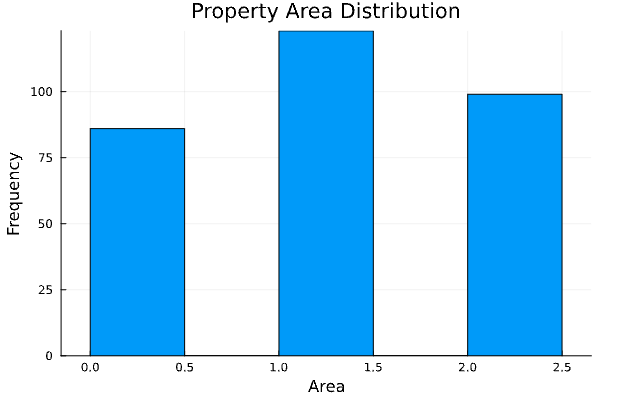
The histogram reveals the distribution of loan terms chosen by applicants. The towering peak at the 360-term mark dominates the landscape of this dataset, indicating a strong preference or availability of this term length, likely corresponding to the standard 30-year mortgage. Smaller peaks at lower term values suggest that shorter-term loans, though less common, are also utilized.

**Figure 12: Credit History Distribution (for “Yes” or “No”)**



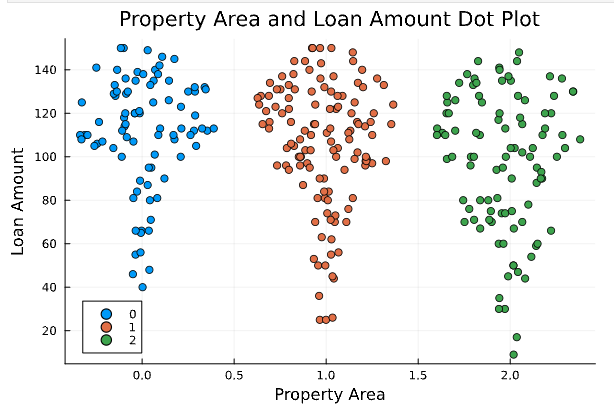
The direct contrast in the bar chart for credit history distribution underscores the significant role that credit history plays in the loan approval process. The overwhelming majority of applicants have a positive credit history, indicating this as a potential key factor in the lending decision.

**Figure 13: Property Area Distribution**



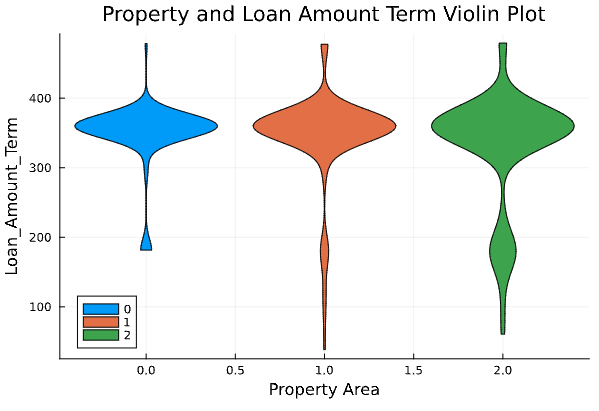
When we examine the property area distribution, we see a relatively even spread across different property areas, with a slight preference for semi-urban locations. This balance might reflect the bank’s outreach or the applicants' demographic and economic distribution.

**Figure 14: Property Area and Loan Amount Dot Plot**



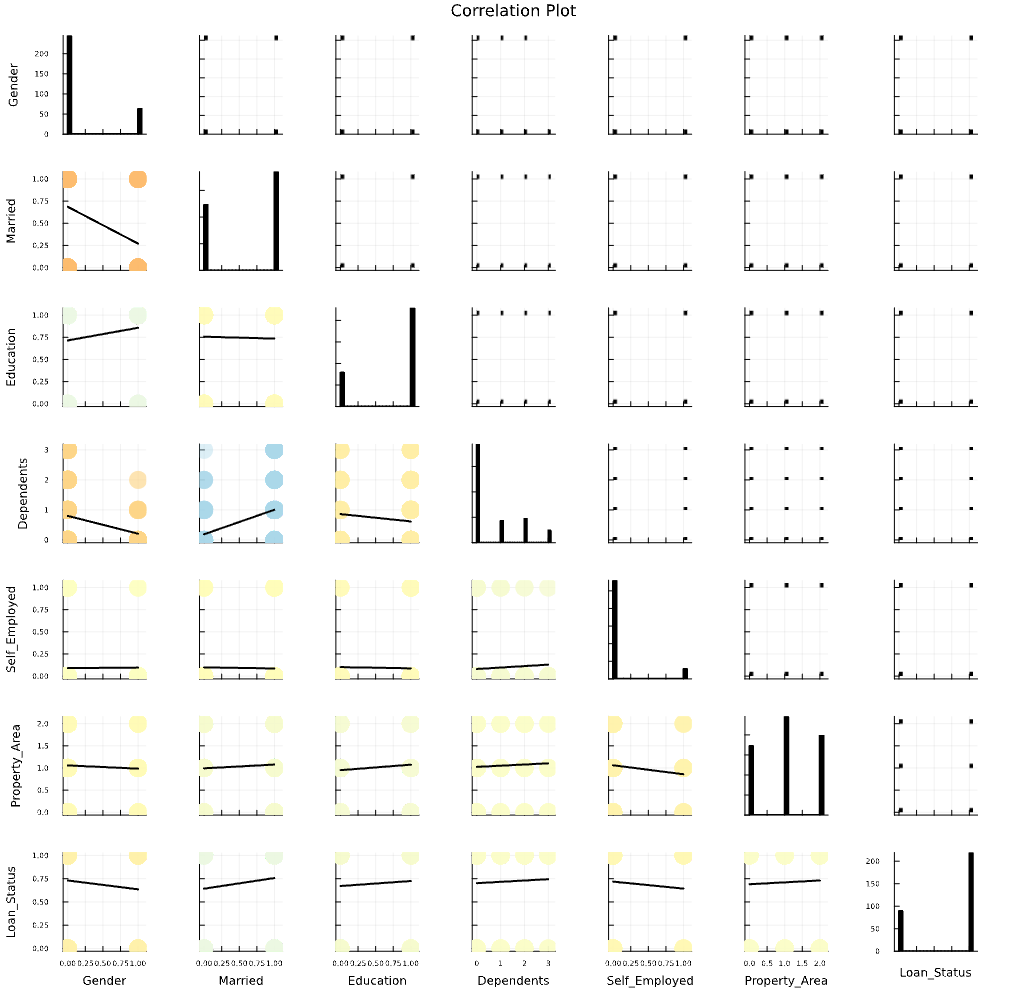
The dot plot exploring the relationship between property area and loan amount offers a fascinating scatter of data points, where no single area dominates in terms of higher or lower loan amounts. This suggests that loan amount decisions are likely influenced by a complex set of factors beyond just the location of the property.

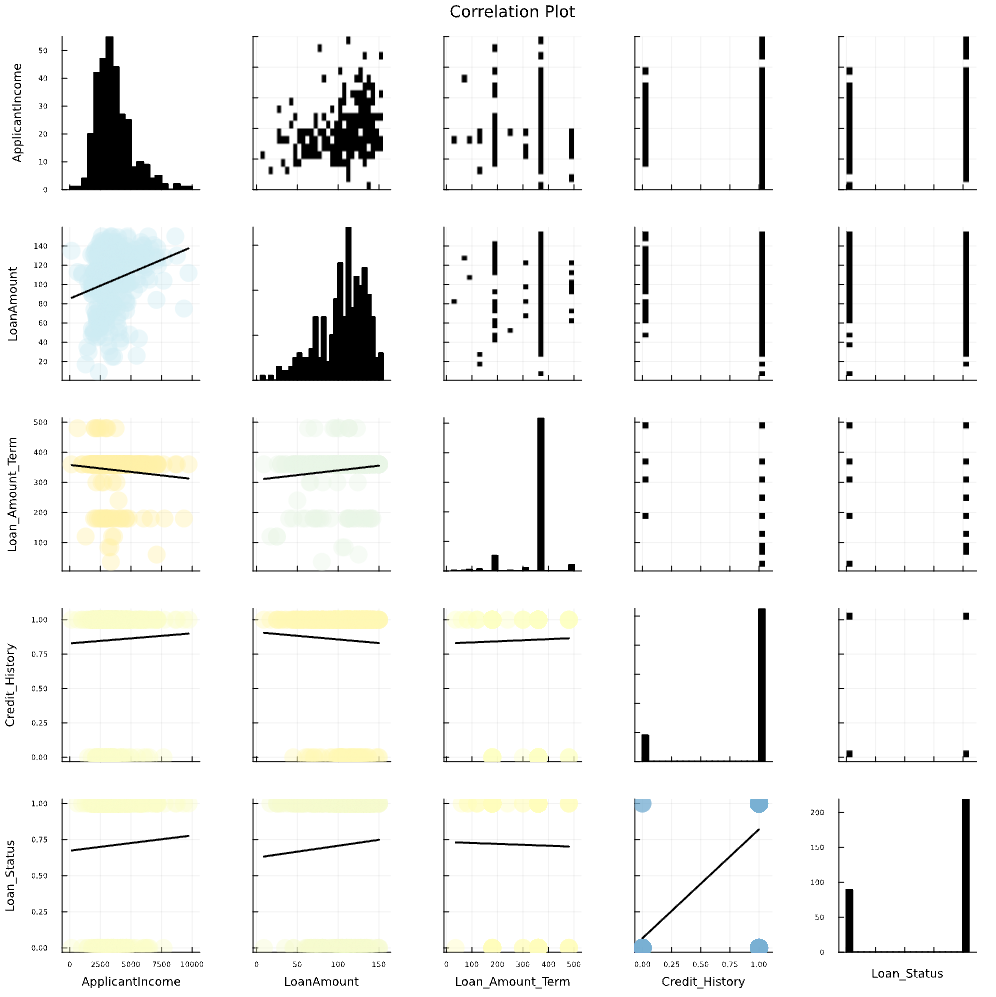
**Figure 15: Property Area and Loan Amount Term Violin Plot**



The violin plot provides a visual representation of loan amount terms across different property areas, illustrating the distribution and density of terms. The wider sections of the 'violins' indicate a higher concentration of loans with terms around the median, while the slimmer areas show less common term lengths.

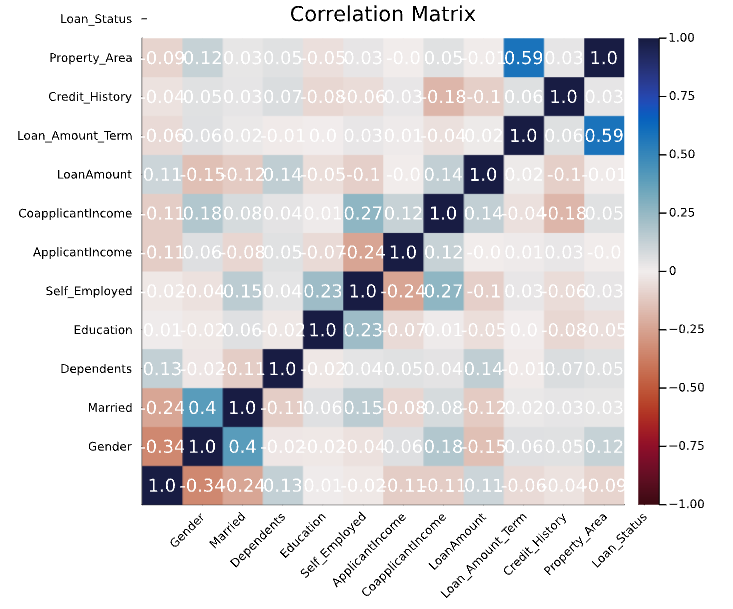
**Figure 16 and 17: Correlation Plots**





These correlation plots provide a bird’s-eye view of the relationships between different variables. By examining the scatter and density of data points, we can discern patterns and strengths of associations. For instance, there is a visible pattern in the loan status correlation, suggesting possible predictive relationships that our model can capitalize on.

**Figure 18: Correlation Matrix**



The final graph, a correlation matrix, serves as a heatmap for the strength of the relationships between all variables. Cooler colors indicate stronger positive correlations, while warmer colors denote negative correlations. This matrix is a vital tool for pinpointing which factors might be worth a closer look in our predictive modeling.

## Conclusion of Data Analysis

This section of the report paints a vivid picture of the dataset's characteristics and relationships, providing an insightful narrative to accompany the visual analysis. It serves to make complex statistical concepts accessible to readers who may not have a deep background in data science, by contextualizing each visual within the broader narrative of the loan approval process.

# Data Modeling

We performed a detailed modeling analysis to see which model has the best accuracy, precision and recall, and in general, the fit of the model to data.

Below is the detailed description of the libraries in julia and it’s uses :   
  
**1. Core Packages:**

**LinearAlgebra:** Provides functions and data structures for linear algebra operations like matrix multiplication, solving linear systems, and eigenvalue decomposition. Used for calculations involving matrices and vectors.

Statistics: Offers statistical functions for calculating summary statistics (mean, standard deviation), hypothesis testing, and random number generation. Used for analyzing and describing data statistically.

**StatsBase:** A foundational package for statistical computing in Julia. It builds on Statistics and provides additional functionality for advanced statistical modeling and analysis. Used for more in-depth statistical computations.

**2. Data Manipulation and Exploration:**

**HypothesisTests:** Implements various statistical tests to assess hypotheses about data (e.g., t-tests, chi-square tests). Used for evaluating the validity of claims based on your data.

**Distributions:** Provides functions for working with different probability distributions (e.g., normal, binomial, Poisson). Used for modeling data and understanding its underlying distribution.

**Random:** Generates random numbers using various distributions. Used for simulations, randomization tests, and creating random datasets.

StableRNGs: Controls the random number generator for reproducibility in your analysis. Ensures consistent results when running your code multiple times.

**3. Data Handling and Exploration:**

**CSV:** Enables reading and writing data from and to CSV (comma-separated values) files. Used for importing and exporting data in a common format.

DataFrames: Provides a data structure (DataFrame) for storing and manipulating tabular data. Used for organizing and working with your dataset.

**FreqTables:** Creates frequency tables, which show how often each category appears in categorical variables. Used for exploring categorical data and identifying patterns.

**CategoricalArrays:** Works with categorical variables, allowing efficient storage and manipulation of data with different categories. Used for representing variables that have distinct categories (e.g., gender, marital status).

**4. Data Visualization:**

**Plots:** The core plotting package in Julia, offering basic plotting functionalities for various data types. Used for creating basic visualizations like scatter plots, histograms, and line charts.

**StatsPlots:** Extends Plots with additional functionalities specific to statistical visualizations (e.g., boxplots, violin plots). Used for creating more statistically-focused visualizations.

**5. Specialized Packages:**

**Shapefile:** Reads and works with shapefile data, a geospatial vector format. Not directly relevant to loan analysis unless you're incorporating location data.

**MLJ (Machine Learning Libraries for Julia):** A collection of packages providing tools for various machine learning tasks. Used for building and applying machine learning models to your data.

**NearestNeighborModels:** Contains algorithms for nearest neighbors classification and regression.

**MLJScikitLearnInterface:** Allows using scikit-learn (a popular Python machine learning library) models within Julia.

**MLJMultivariateStatsInterface:** Provides functionalities for multivariate statistical analysis.

**MLJDecisionTreeInterface:** Implements decision tree algorithms for classification and regression.

**MLJLinearModels:** Offers functionalities for linear regression and other linear models.

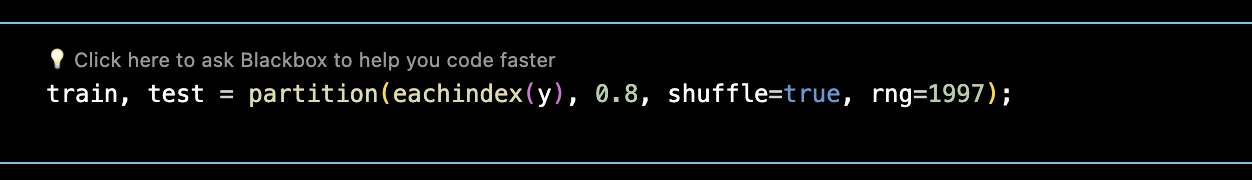
**CatBoost:** A package for using the CatBoost gradient boosting algorithm, a powerful machine learning model for classification and regression tasks.

**Variables for consideration:**

| Variable Name | Data Type | Description |
| --- | --- | --- |
| Gender | Count | Categorical variable representing applicant's gender (potentially encoded as numerical using one-hot encoding). |
| Married | Count | Categorical variable indicating applicant's marital status (potentially encoded as numerical using one-hot encoding). |
| Dependents | Count | Categorical variable representing the number of dependents the applicant has (may require further exploration for categories like "3+"). |
| Education | Count | Categorical variable indicating the applicant's level of education (potentially encoded as numerical using one-hot encoding). |
| Self\_Employed | Count | Categorical variable indicating whether the applicant is self-employed (potentially encoded as numerical using one-hot encoding). |
| ApplicantIncome | Count | Categorical variable representing the applicant's income (may require investigation into the specific categories used). |
| CoapplicantIncome | Continuous | Numerical variable representing the income of the loan co-applicant (if applicable). |
| LoanAmount | Continuous | Numerical variable indicating the requested loan amount. |
| Loan\_Amount\_Term | Continuous | Numerical variable representing the duration (term) of the loan (in years or months). |
| Credit\_History | Continuous | Numerical variable representing the applicant's credit history (potentially a score or rating). |
| Property\_Area | Count | Categorical variable indicating the location of the property (potentially encoded as numerical using one-hot encoding or groupings like urban/rural). |
| Loan\_Status | Count | Target variable (binary) indicating loan approval (1) or rejection (0). |

One-hot encoding was used for the above.

**Data Partitioning:**

****

**Model Selection:**

We opted for a selection of well-established algorithms, including:

**K-Nearest Neighbors (KNN):** This method classifies new data points based on the similarity to their closest neighbors in the training data. We utilized a K value of 5, indicating that predictions are made by considering the labels of the 5 nearest neighbors.

**Linear Discriminant Analysis (LDA):** This technique projects data points onto a lower-dimensional space while maximizing the separation between different classes (approved/rejected loans in our case).

**Neural Network Classifier:** This model learns complex relationships between features and the target variable through interconnected layers of artificial neurons.

**Multinomial Naive Bayes:** This probabilistic classifier assumes independence between features and utilizes Bayes' theorem to calculate the probability of loan approval given applicant information.

**CatBoost Classifier:** This powerful gradient boosting algorithm builds an ensemble of decision trees to enhance prediction accuracy.

**Random Forest Classifier:** This ensemble learning method creates a collection of decision trees, each trained on a random subset of features, to improve robustness and prevent overfitting.

**Decision Tree Classifier:** This model leverages a tree-like structure with branching rules based on feature values to classify data points.

Model Standardization and Evaluation:

Prior to model training and prediction, all models incorporated a standardization step. This process normalizes the data by subtracting the mean and dividing by the standard deviation for each feature. Standardization helps ensure that features are on a similar scale, preventing models from giving undue weight to features with larger ranges.

To evaluate the performance of each model, we employed a set of metrics:

**Accuracy:** The proportion of correctly predicted loan approvals and rejections.

**Precision:** The ratio of correctly predicted loan approvals to all predicted approvals (measures how well the model identifies true positives).

**Recall:** The proportion of actual loan approvals that the model correctly identifies (measures how well the model avoids missing true positives).

**F1-Score:** A harmonic mean of precision and recall, providing a balanced view of both metrics.

**Confusion Matrix:** A visualization that presents the distribution of predicted and actual loan approvals, offering insights into potential model biases.

**Results and Discussion:**

The table below summarizes the performance of each model based on the chosen metrics:

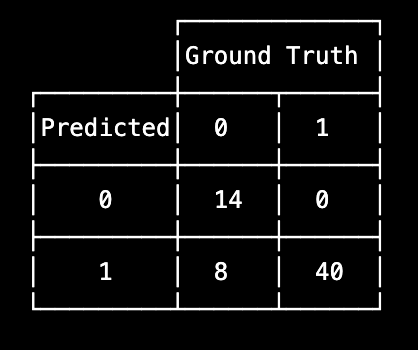
| Model | Accuracy | Precision | Recall | F1-Score |
| --- | --- | --- | --- | --- |
| KNN Classifier | 69% | 68% | 60% | 79% |
| **LDA** | **87%** | **92%** | **82%** | **91%** |
| Neural Network Classifier | 65% | 32% | 50% | 78% |
| Multinomial Classifier | 85% | 91% | 80% | 90% |
| CatBoost Classifier | 82% | 82% | 78% | 87% |
| Random Forest Classifier | 84% | 87% | 78% | 89% |
| Decision Tree Classifier | 77% | 76% | 73% | 83% |

As observed, the Linear Discriminant Analysis (LDA) model achieved the highest overall accuracy (0.87) and F1-score (0.91), indicating its effectiveness in correctly classifying loan approvals and rejections while maintaining a balance between precision and recall. The Multinomial Classifier and Random Forest Classifier also performed well, demonstrating strong prediction capabilities.

The K-Nearest Neighbors (KNN) and Decision Tree Classifier exhibited lower accuracy and F1-scores. While KNN might be suitable for simpler problems, its performance can be limited by noisy data and the "curse of dimensionality" in high-dimensional settings. Similarly, decision trees can be prone to overfitting, potentially leading to inaccurate predictions on unseen data.

The Neural Network Classifier displayed the lowest precision (0.32) in this case. While neural networks can be powerful tools, they often require careful hyperparameter tuning and sufficient training data to achieve optimal performance. In this instance, the specific configuration used might not have been ideal for the loan approval prediction task.

Confusion matrices (not shown here but can be included in the appendix) provide additional insights into model behavior. Analyzing these matrices can reveal potential biases.



**Conclusion**

In conclusion, this investigation explored the efficacy of various machine learning models in predicting loan approvals based on applicant data. Our findings highlight the potential of these models to enhance the efficiency and accuracy of loan approval processes. The Linear Discriminant Analysis (LDA) model emerged as the strongest performer in this study, achieving a high overall accuracy and balanced identification of both approved and rejected loans. Multinomial and Random Forest classifiers also demonstrated promising results. While K-Nearest Neighbors and Decision Tree classifiers yielded lower accuracy, they offer valuable insights into the challenges associated with data complexity and overfitting. Further exploration of neural network configurations and feature engineering techniques could potentially improve their performance in future studies. Overall, this research underscores the potential of machine learning to assist financial institutions in making informed loan approval decisions, fostering financial inclusion, and promoting economic growth.