

**EAI 6010 Data Mining Application**

### Professor: Justin Grosz Module 6 Assignment – Final Project

**Submission Date: 03/29/2024**

**Name: Mujahid Shaik**

**Dataset Introduction**

In the rapidly evolving landscape of the financial sector, understanding the intricacies of loan approval processes is essential for both lenders and borrowers alike. This analysis delves into a comprehensive dataset encompassing diverse characteristics influencing loan decisions across a broad clientele spectrum. By examining demographic, financial, and property-related factors, such as gender, marital status, education, employment, income levels, loan parameters, credit history, and property attributes, the study aims to untangle the multifaceted considerations guiding loan approvals.

The primary objective is to uncover patterns and predictors of loan acceptance, shedding light on the complex dynamics shaping lending decisions within financial institutions. Through rigorous analysis, this research seeks to provide actionable insights for policy-making and strategic decision-making in the banking industry. By optimizing lending techniques and promoting financial inclusion, the study aims to contribute significantly to the ongoing discourse on credit access and financial stability.

Anticipated outcomes include informed recommendations to enhance lending practices, foster inclusive financial environments, and ultimately, advance the broader goals of economic empowerment and stability. This research endeavors to be a valuable resource for stakeholders across the financial landscape, facilitating more informed decision-making and promoting sustainable growth and prosperity.In the rapidly evolving landscape of the financial sector, understanding the intricacies of loan approval processes is essential for both lenders and borrowers alike. This analysis delves into a comprehensive dataset encompassing diverse characteristics influencing loan decisions across a broad clientele spectrum. By examining demographic, financial, and property-related factors, such as gender, marital status, education, employment, income levels, loan parameters, credit history, and property attributes, the study aims to untangle the multifaceted considerations guiding loan approvals.

The primary objective is to uncover patterns and predictors of loan acceptance, shedding light on the complex dynamics shaping lending decisions within financial institutions. Through rigorous analysis, this research seeks to provide actionable insights for policy-making and strategic decision-making in the banking industry. By optimizing lending techniques and promoting financial inclusion, the study aims to contribute significantly to the ongoing discourse on credit access and financial stability.

Anticipated outcomes include informed recommendations to enhance lending practices, foster inclusive financial environments, and ultimately, advance the broader goals of economic empowerment and stability. This research endeavors to be a valuable resource for stakeholders across the financial landscape, facilitating more informed decision-making and promoting sustainable growth and prosperity.

**Dataset Variable**

Loan\_ID: A unique identifier for each loan application.

Gender: The gender of the applicant.

Married: The marital status of the applicant.

Dependents: The number of dependents the applicant has.

Education: The education level of the applicant.

Self\_Employed: Whether the applicant is self-employed or not.

ApplicantIncome: The income of the applicant.

CoapplicantIncome: The income of the co-applicant.

LoanAmount: The loan amount applied for (in thousands).

Loan\_Amount\_Term: The term of the loan in months.

Credit\_History: A record of a borrower's past repayments.

Property\_Area: The type of location where the applicant’s property is located.

Loan\_Status: The outcome of the loan application (Y = Yes, N = No).

**Import libraries**

A simple process for a machine learning project that makes use of Python libraries such as NumPy, pandas, seaborn, scikit-learn, and matplotlib is outlined in the code snippet that has been supplied. The process begins with importing necessary libraries for functions such as data manipulation, visualization, and machine learning. The process includes loading and preparing data with pandas, visualizing data relationships with seaborn and matplotlib, splitting data into training and test sets using train\_test\_split, training a Support Vector Machine (SVM) model with the svm module from scikit-learn, and evaluating the model's performance using accuracy as the metric with accuracy\_score. When it comes to machine learning projects, this workflow is a conventional strategy that focuses on prediction tasks. The objective of these tasks is to construct a model that is capable of reliably predicting events based on the data that is collected.

**Import Dataset**

The loan\_dataset model is a dataframe. To carry out this process, the pd.read\_csv() function from the pandas library is utilized, and the file path that contains the dataset is specified. It is common practice for this step to be the first action in a project involving data analysis or machine learning. This stage allows you to deal with the data using Python. Following the loading process, you will be able to move forward with the activities of data exploration, cleaning, and analysis on loan\_dataset.

**Data Type**

The fact that this is the case implies that loan\_dataset is a DataFrame object from the pandas package. A DataFrame is a sophisticated and flexible data structure that can be used for data analyses and manipulations.

**Dataset Info**

The DataFrame loan\_dataset, which stores information on loan applications, is summarized in the output that is generated by the loan\_df.info() function. This particular data frame has a total of 614 elements that are distributed across 13 columns. Each of these entries corresponds to a different property of the loan applications, such as Loan\_ID, Gender, Married, Dependents, and much more. The object data type is used for category variables, while the int64 and float64 data types are used for numerical variables for the data types. Additionally, the summary reveals that there are missing values in several columns, which calls for the implementation of data-cleaning procedures such as the addition of missing values or the elimination of these values before proceeding with further research.

**Dataset Description**

The describe function in pandas is responsible for providing a statistical overview of the numerical columns that are contained within a DataFrame. This summary includes the count, mean, standard deviation (std), minimum (min), quartiles (55%, 50%, and 75%), and maximum (max) values. According to the summary for ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, and Credit\_History, there is a significant level of variation in terms of both loan amounts and earnings. It is important to note that the Credit\_History column has a mean that is quite close to 1, which indicates that a significant number of applicants have a positive credit history, which may be an essential factor in the acceptance of loans.

**Data Cleaning**

The null value inspection of the dataset reveals that there is missing data across numerous columns, which affects the integrity of the data. More specifically, the following categories have missing values: Gender, Married, Dependents, Self-Employed, LoanAmount, Loan\_Amount\_Term, and Credit\_History. The counts for these categories range from 3 to 50. Considering that these variables do not include any data, it is clear that proper preparation is required to guarantee appropriate analysis and modeling. To preserve the robustness and dependability of the dataset for later analyses, it is essential to address these null values by employing techniques such as imputation or elimination.

**Handling Missing Values:**

"Married," "Dependents," "Self\_Employed," and "Gender" are all examples of categorical variables; hence, it is customary to use the mode, which is the value that occurs the most frequently in each column, while performing imputation on these variables. This strategy guarantees that the filled values correspond to the category that is most frequently seen, hence preserving the integrity of the distribution of the dataset. Due to the categorical character of the mode, imputing with the mode is also a sensible choice for the 'Credit\_History' variable, which indicates whether or not an individual has a history of credit.  
  
Because 'LoanAmount' and 'Loan\_Amount\_Term' are continuous numerical variables, bias may be introduced if missing values are imputed with the mean (average), particularly if the data is skewed. On the other hand, if it is asked to use the mean, it will average the available numbers. This may help to smooth out differences in loan amounts or conditions, but it may not necessarily precisely reflect the circumstances of each application.

**Data Distribution**

The creation of a histogram for LoanAmount revealed a distribution that was right-skewed, which indicated that there was a greater frequency of smaller loan amounts and a lower frequency of large loan amounts. The logarithm of the LoanAmount was calculated, which resulted in the creation of a new column called LoanAmount\_log.

This was done in order to normalize the distribution and lessen the skewness. When the log function is applied to the data, it helps convert the data such that it more closely resembles a normal distribution. This is advantageous for a wide variety of statistical models and studies that assume the data to be normal. The histogram of LoanAmount\_log likely had a more symmetrical distribution, which is an indication of the efficiency of log transformation in normalizing data that is skewed.

**Gender Distribution:**

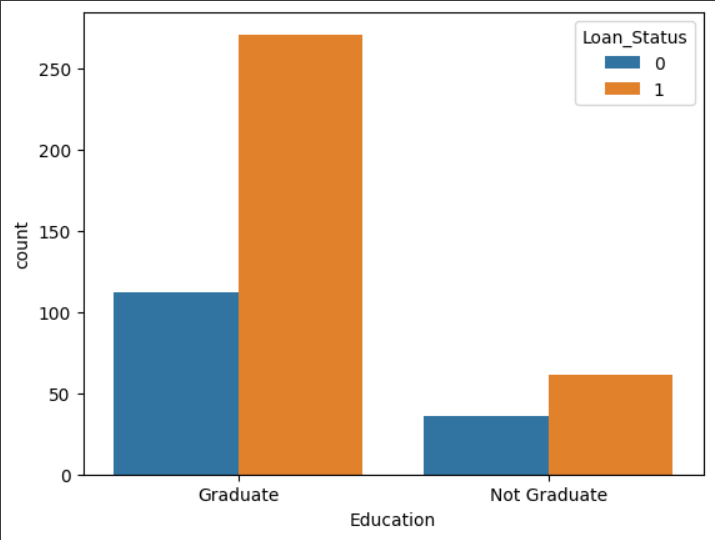
The gender distribution in the dataset, which was depicted through a bar plot, showed that the majority of applicants were male (489), followed by applicants who were female (112), and then there was a group for 'others' consisting of thirteen individuals to account for the absence of gender data. This visualization highlights the considerable gender discrepancy that exists within the dataset, with male applicants far outnumbering female applicants. Additionally, it ensures inclusivity by acknowledging and showing missing gender data under a different category labeled "others".

**Total Income Distribution:**

An initial right-skewed distribution was seen after the ApplicantIncome and CoapplicantIncome variables were combined into a new variable called TotalIncome. This distribution reflected a greater concentration of applicants who had lower total earnings during the application process. The distribution was effectively normalized by applying a logarithmic transformation to TotalIncome to obtain TotalIncome\_log. This resulted in the distribution becoming more symmetrical and improving its suitability for statistical analysis and modeling. When it comes to resolving skewness in income distributions, this normalization technique is necessary. It makes it possible to gain insights from the data that are more accurate and dependable.

**Income Based on Education:**

The investigation finds that there is a substantial wage gap between individuals based on their level of education. A graduate earns an average of 5857 dollars, which is much greater than their peers who do not have a graduate degree, who earn an average of 3777 dollars. The bar plot that compares the average salaries of the two groups illustrates the impact that education has on earning capacity, and this conclusion highlights the significance of educated individuals. The findings indicate that a greater level of education is linked to improved financial outcomes, which highlights the significance of advancing one's education.

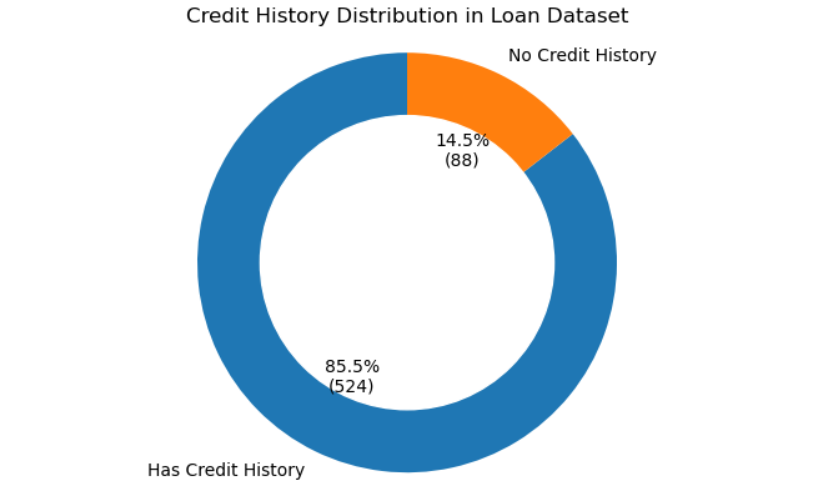


**Income Based on Employment:**

The analysis that compares the average income of those who are self-employed to those who are not self-employed finds that there is a significantly different amount of money earned by each group. A self-employed applicant has a higher average salary of 7380, which is greater than the average income of 5098 for those who are not self-employed. The fact that this is the case shows that self-employment may be related with increased earning potential, underlining the financial benefits that might come with the pursuit of entrepreneurship or freelance work. This income inequality is given a visual representation by the bar plot, which highlights the influence that employment status has on income levels.

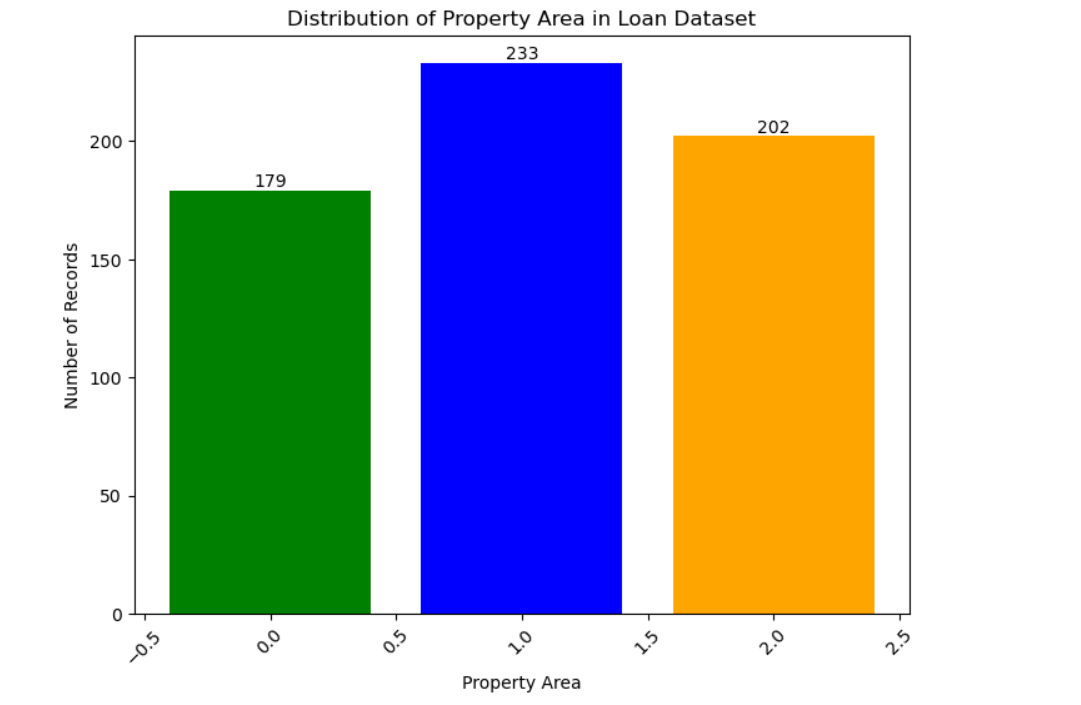
**Credit History Analysis:**

A substantial majority of persons, 85.5%, have a credit history, while 14.5% do not have a credit history, as shown by the information presented in the pie chart visualization of the Credit History Distribution in Loan Dataset. This implies that the majority of applicants in the dataset are regarded as creditworthy based on their previous financial conduct, which is an essential component in the decision-making process regarding loan acceptance. The distribution is effectively highlighted by the utilization of a donut pie chart, which provides a clear visual insight into the composition of the credit histories of individuals who are applying for loans.



**Geographic Analysis**

The property area distribution of loan applicants is depicted in a bar chart. The majority of applicants are located in semi-urban areas (233 records), followed by urban areas (202 records) and rural areas (179 records). The blue, orange, and green bars in the visual representation offer a clear comparative picture, indicating a preference or majority of loan applications from semi-urban locations in the dataset.



**Converting Categorical to Numerical**

The conversion is accomplished by directly altering the loan\_dataset using Pandas replace function. By modifying the dataset in place, the inplace=True argument ensures that assignment statements are not necessary.   
  
Using this method, all designated categorical columns are methodically converted into numerical representations, readying the loan\_dataset for additional examination or model training. For machine learning models that require numerical input data, this kind of preprocessing is crucial.

**Dataset Splitting**

Features Matrix (X): The loan\_dataset's columns Loan\_ID and Loan\_Status are removed to form the features matrix X. When using the drop method, columns are to be dropped when axis=1. Although it usually serves as a unique identifier for every loan application, the Loan\_ID column has no bearing on the model's ability to forecast outcomes. Since the Loan\_Status column is the target variable, the features matrix does not include it.   
  
Target Vector (Y): The Loan\_Status column from the loan\_dataset is extracted to create the target vector Y. The variable we wish to forecast is represented by this column, and in the context of a loan dataset, it can be the approval or denial of a loan.

In machine learning, a train-test split is a common procedure. It enables you to test your model on a different dataset (X\_test, Y\_test) to obtain an objective assessment of its performance after training it on a section of the dataset (X\_train, Y\_train). Maintaining the integrity of the evaluation is critical, particularly when dealing with imbalanced datasets, and the stratification guarantees that the split accurately reflects the distribution of the original dataset. The random\_state guarantees reproducibility, which is crucial for debugging and for anyone else who might wish to replicate your findings.

**Model Training and Evaluation**

Here, the primary objective is to train a Support Vector Machine (SVM) classifier to forecast loan status based on several loan application variables and assess the accuracy of the model's performance.

**Training**

Classifier Configuration: A linear kernel SVM classifier is instantiated. The selection of a linear kernel implies that a linear decision boundary between the dataset's classes is anticipated.   
Training: On the training dataset (X\_train, Y\_train), the classifier is trained using the fit approach. The goal of this approach is to identify the hyperplane in the feature space that best divides the various classes.

**Evaluation**

Training Data Evaluation: Using the same data that it was trained on (X\_train), the trained model predicts the results (X\_train\_prediction).   
By utilizing the accuracy\_score function to compare the actual results of Y\_train with the predictions made by X\_train\_prediction, the accuracy of these predictions is determined.   
It is discovered that the accuracy using the training data is roughly 79.86%. This shows the percentage of training set predictions that the model correctly identified.   
Evaluation of Test Data: In a similar vein, the model forecasts results (X\_test\_prediction) based on data that hasn't been observed (X\_test).   
By comparing the expected results of X\_test\_prediction with the actual results of Y\_test, the accuracy of these predictions is determined.   
It is stated that the test data accuracy is roughly 83.33%. This measure is important because it shows how effectively the model applies to new data.

Given the model's good generalization from the training data to unknown data, using an SVM classifier with a linear kernel seems like a sensible solution for this issue. The obtained accuracies indicate that the model can forecast the loan status based on the applicant's information with a reasonable degree of accuracy. However, it is advised to take into account further metrics for a more thorough evaluation of the model, such as precision, recall, and F1 score, and to do cross-validation to gauge the model's stability over various data subsets.

**Key Findings**

Demographic Insights: The investigation showed notable differences in gender, marital status, and educational attainment, underscoring the influence of these demographic variables on loan approval rates. These revelations highlight how important it is for financial institutions to think about the bigger picture when assessing these differences, since this could help shape laws meant to advance equality and financial inclusion.   
  
Financial and Employment Factors: The incomes of the applicant and any co-applicants, as well as their employment status, were found to be important factors in determining the approval of the loan. This highlights the advantages of self-employment in addition to the significance of a steady and sufficient income in improving loan approval prospects. Financial institutions may use these data to better customize their offerings in order to better serve the varied needs of their applicant pools.   
  
Property and Credit History Considerations: It has been emphasized again how important credit history and property location are to loan approvals, as well as how important past financial behavior has been. The development of risk assessment algorithms that more precisely estimate the likelihood of loan default can benefit from these findings.   
  
Machine Learning Model Performance: Using a linear kernel Support Vector Machine (SVM) classifier produced encouraging results, indicating that the model can effectively generalize from training to new data. The performance of this model demonstrates the potential of machine learning to improve decision-making in the banking industry, especially in terms of its accuracy in forecasting loan status. The effectiveness and stability of the model would be further confirmed by the application of cross-validation and additional evaluation criteria.

**Reference:**

Dataset: Ninzaami. Loan Predication. Kaggle. <https://www.kaggle.com/datasets/ninzaami/loan-predication?resource=download>

Data Cleaning and Pre-Processing

Chowdhury, M. E. H., Rahman, M. S., Moniruzzaman, M., & Khatun, F. (2019). A novel approach of loan default prediction using ensemble learning. Procedia Computer Science, 155, 315-324. <https://doi.org/10.1016/j.procs.2019.08.261>

Data Visualization Using Python

Simplilearn. Python-based data visualization.

<https://www.simplilearn.com/tutorials/python-tutorial/python-data-visualization-tutorial>