LOAN ELIGIBILITY PREDICTIONS

# Project Proposal: Loan Eligibility Predictions

# Dataset:

This dataset offers a special chance to investigate how factors like gender, income, education, credit history, and the location of the property affect loan acceptance decisions. Are those with good credit records more likely to get loans? Do gender-based disadvantages still exist in the lending sector? How do different property types affect the results of borrowing?

# Question:

Can we predict whether a loan application will be approved or denied based on various applicants and loan-related factors?

# Hypothesis:

There is a relationship between the various factors in this dataset (e.g., Gender, Education, Credit History, etc.) and the likelihood of loan approval. Specifically, certain characteristics such as credit history, income, and education level may significantly influence the approval or rejection of loan applications.

The number of elements, including income, education, and credit history, may have a big impact on the loan acceptance process.

# Variables in Dataset:

Loan\_ID: A special number assigned to every loan application.

Gender: A categorical variable indicating the applicant's gender. There are two groups: men and women.

Married: A categorical variable expressing the applicant's marital status. There are two groups: Yes and No.

Dependents: A categorical variable that indicates the applicant's total number of dependents. These are the categories 0–1, 2– and 3+.

Education: A categorical variable that reflects the applicant's degree of education. Graduate and Non-Graduate are the two classifications.

Self\_Employed: This is a categorical variable that indicates if the candidate works for themselves or not. There are two groups: Yes and No.

ApplicantIncome: A numerical variable that indicates the applicant's income.

CoapplicantIncome: A numeric variable that represents the coapplicant's income

LoanAmount: A numerical variable that represents the desired loan amount.

Loan\_Amount\_Term: A numerical variable that shows how long the loan will last in months.

Credit\_History: It is a binary variable that returns 1 if the applicant has credit history and 0 otherwise.

Property\_Area: The property area type is represented by the categorical variables which has three divisions: Urban, Semiurban, and Rural

Loan\_Status: The label or target variable indicating the approval status of the loan. Yes (Y) and No (N).

**Variables relevant to the question:**

***Credit History***:

When approving a loan, credit history is a major consideration. Those with a good credit history typically have their loans approved more quickly by lenders since it shows their previous financial responsibility and likelihood of making timely repayments.

***Applicant's Income:***

A larger salary indicates a stronger ability to repay the debt. As a result, applicants with greater earnings might stand a better chance of getting a loan.

***Income of Co-Applicant***:

The applicant's and co-applicant's combined income may influence the financial stability of the household as a whole and, consequently, the loan repayment capacity.

***Loan Amount:***

Higher loan amounts may be subject to higher requirements for approval.

***Loan Length:***

The applicant’s capacity for repayment and stability may impact the length of the loan. While lengthier loan periods may result in lower initial interest rates, but greater total interest paid over time; shorter loan durations may indicate higher EMI payments.

**Identifying and handling NULL values:**

Missing values in this dataset cannot be removed as they affect the entire dataset. Here are the methods to handle it:

*Gender,* *Married,* *Dependents, Self Employed,* *'Loan Amount Term’, and* ‘*Credit History columns* are handled by replacing with the most frequent value in that particular field. The missing values are replaced with the median value for the' Loan Amount' column.

**Identifying and handling outliers:**

|  |  |
| --- | --- |
| **Variables** | **Outliers identified** |
| ApplicantIncome | 50 |
| CoapplicantIncome | 18 |
| LoanAmount | 41 |
| Loan\_Amount\_Term | 88 |
| Credit\_History | 89 |

For handling the outliers:

* Select only the columns with numeric data types.
* For every numerical column, find Q1, Q3, and IQR.
* The dataset identifies the data points that are either above or below the lower or upper bound as possible outliers.
* Assign the nearest non-outlier value to the outliers.

**Visualization Summary:**

*Histograms:*

These histograms shed light on the distributions of important variables found in the dataset. To comprehend the properties, it can be helpful to look at the distributions and the concentration of values inside particular ranges.

A graph of a diagram

Description automatically generated with medium confidence

Applicant Income Histogram: It displays the number and frequency of different income levels. The height of each bar, representing a range of salaries, shows how many applications fall into that range. It is left-skewed which deviates from normality.

Loan Amount Histogram: This is the visual representation of the frequency of various loan amount ranges, which is similar to the first histogram. Every bar symbolizes a range of loan amounts, and the height denotes the number or frequency of loans falling inside that range. Left skewed so no normalization.

A red rectangular object with black text

Description automatically generated

Credit History histogram: The distribution of credit history values is visualized. The histogram shows that the count is high with applicants who have a credit history. More than 500 applicants have a credit history. This is not normalized because applicants who have a credit history are represented by 1 otherwise 0, as the distribution is different.

Loan Length Distribution: It shows the distribution of loan terms, or lengths in days. Most of the applicant's Loan amount term is for 11 months, approximately 360 days. Like the above graph it doesn’t not have normalizations.

*Correlation between variables:*

A screenshot of a computer screen

Description automatically generated

The correlation matrix displays the coefficients of correlation between the numerical features. The values are -1 to 1, with 1 denoting a high positive correlation, -1 a strong negative correlation, and 0 denoting no association. The correlation between loan amount and application income is the strongest compared to other variables.

*Identifying Trends:*

A green circle with blue and white text

Description automatically generated

Impact of Credit History: A considerable impact is seen by the donut chart that shows approval and denial rates according to credit history. An increased approval rate of 85.5% is correlated with a clean credit history.

A graph of a loan approval

Description automatically generated

Property Area Influence: Investigating this may help identify patterns about how the location of a property affects loan acceptance. From the graph, we can see that all areas have higher loan acceptance than denial.

*Outliers:*

A graph of a person and person

Description automatically generated with medium confidence

The distribution of applicant income for loans that are authorized ('Y') and denied ('N') is graphically compared in the boxplot. The median line is located inside the box, which depicts the interquartile range (IQR). The "whiskers" cover the lowest and greatest values that fall into a specific range, which is typically 1.5 times the IQR. All the points that are larger than the whiskers are outliers.

*Handling Outliers:* In this boxplot, values above the upper bound can be substituted with the upper bound value, and values below the lower bound can be substituted with the lower bound value for each column. By doing this, the extreme values outside of the specified bounds are effectively clipped or truncated. This method is frequently used to reduce the influence of outliers on statistical analysis.

**Model Development:**

KNN is a suitable choice for capturing the complexity of the decision-making process since it can successfully model non-linear interactions, as there may not be a link between the criteria impacting loan acceptance.

The process of building a KNN model involves a few important steps:

Data Preparation: Relevant features are chosen and the dataset is preprocessed. Binary numeric values (1 and 0) are transferred to categorical target values (Y and N).

Data splitting: To train and assess the performance of the model, the dataset is split into train and test (70% training, 30% testing).

Standardization: A crucial step for kNN models, feature standardization is used to guarantee that the magnitude of each input feature is the same.

Model Training: Using a selected value of k (7 in this case), the kNN model is trained on the standardized training data.

Evaluation and Prediction: The model forecasts the test set's loan approval status and computes accuracy, precision, and F1 score. An understanding of the model's accuracy can be gained from the confusion matrix.

Model Optimization: The ideal value of K is identified by the F1 scores for various values of k. In this case, the optimal k value is 9.

A graph with a line

Description automatically generated