Loan Application Status Prediction

# Introduction

# Lending out loans is crucial for banks to generate profits, but it's essential to do so strategically to avoid turning them into non-profitable assets. Before approving loan applications, banks need to carefully assess various factors to determine the suitability of the applicant.

# By leveraging machine learning algorithms, banks can automate the verification process based on a range of attributes such as applicant income, loan amount, credit history, gender, coapplicant income, and more. This automated approach streamlines the decision-making process and improves the efficiency of loan approval procedures.

# Project Overview

In this project, we utilize the Loan\_Dataset from Kaggle to predict loan approvals. The project steps include:

1. [**Data Exploration**](#_Data_Exploration)
2. [**Univariate, Bivariate, and Multivariate Analysis**](#_Bivariate_and_Multivariate)
3. [**Handling Missing Values**](#_Handling_Missing_Values)
4. [**Data Preparation**](#_Data_Preparation)
5. [**Model Building**](#_Model_Building)
6. [**Model Evaluation with Confusion Matrix**](#_Model_Evaluation)
7. [**Feature Engineering**](#_Upsampling_the_Imbalanced)
8. [**Upsampling the Imbalanced Dataset**](#_Upsampling_the_Imbalanced)

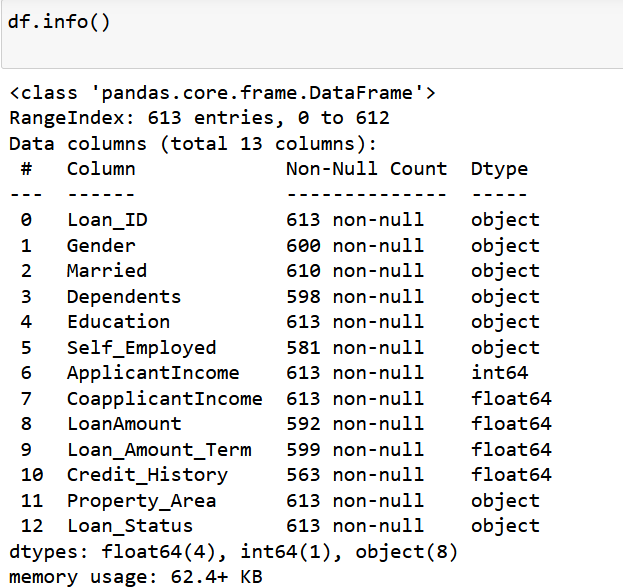
# Starting the Project

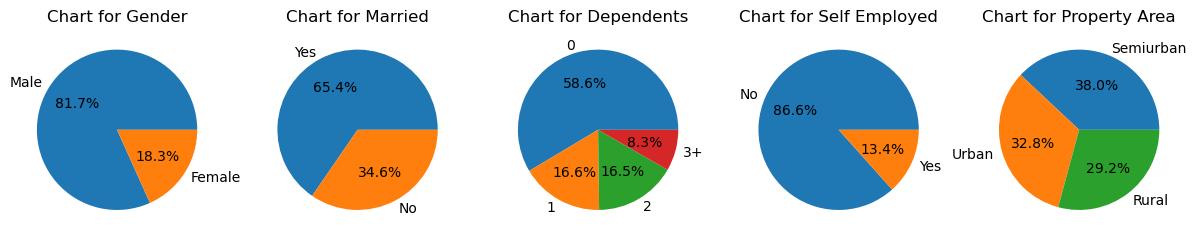
Begin by importing necessary libraries: Pandas, Seaborn, Matplotlib, and Numpy.

# Data Exploration

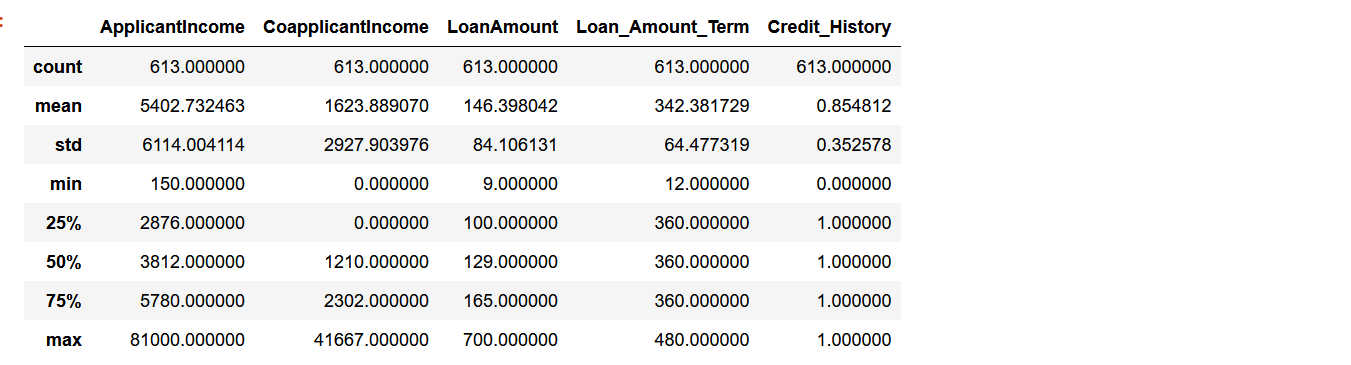
During our initial data exploration:

* We discovered that the dataset contains 614 observations and 12 columns.
* Some columns have missing values, with the 'Credit\_History' column having the highest percentage of missing data (around 8%).
* The dataset is imbalanced regarding loan approval status, with approximately 69% of applications approved and 31% rejected.
* A significant majority of loan applicants are male, comprising 81% of the total, while females make up only 18%.
* The most common characteristics among applicants include:
  + Having no dependents (56%)
  + Being married (64%)
  + Living in semiurban areas (38%)
  + Being non-self-employed (86.6%)





These initial findings provide valuable insights into the dataset's composition and characteristics, setting the stage for further analysis and model development.



Upon analyzing the summary statistics of the dataset using the describe() function, we can make the following important observations:

ApplicantIncome: The average applicant income is approximately 5403, with a wide range of incomes from 150 to 81000. The data is characterized by relatively high variability, as indicated by the standard deviation of 6109.

CoapplicantIncome: The average coapplicant income is approximately 1621, with values ranging from 0 to 41667. Similar to applicant income, this column also exhibits significant variability (standard deviation of 2926).

LoanAmount: The average loan amount applied for is approximately 146, with values ranging from 9 to 700. The standard deviation is approximately 84, indicating variability in loan amounts.

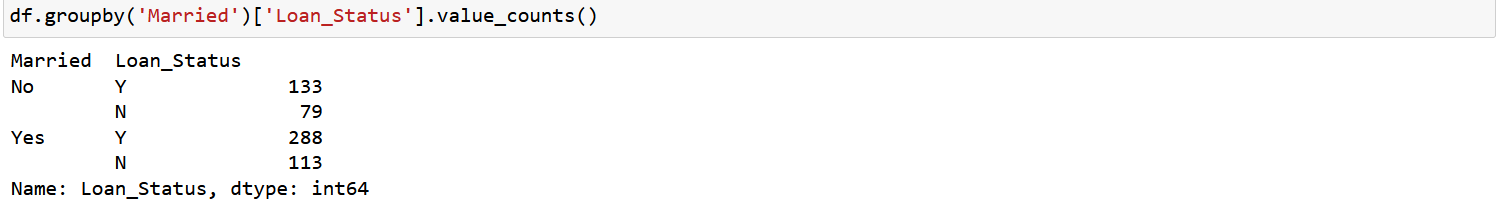
Loan\_Amount\_Term: The majority of loan applicants (75%) have a loan amount term of 360 months (30 years), with some variability around this value. The range of terms is from 12 months to 480 months.

Credit\_History: The majority of applicants (approximately 86%) have a credit history of 1 (good credit), while around 14% have a credit history of 0 (poor credit or no credit history).

These summary statistics provide insights into the central tendency, spread, and distribution of numerical variables in the dataset. The wide ranges and variabilities in applicant income, coapplicant income, and loan amount suggest diversity in the financial profiles of loan applicants. Additionally, the high percentage of applicants with good credit history is noteworthy and may play a significant role in loan approval decisions.

# Bivariate and Multivariate Analysis

## Key insights include:

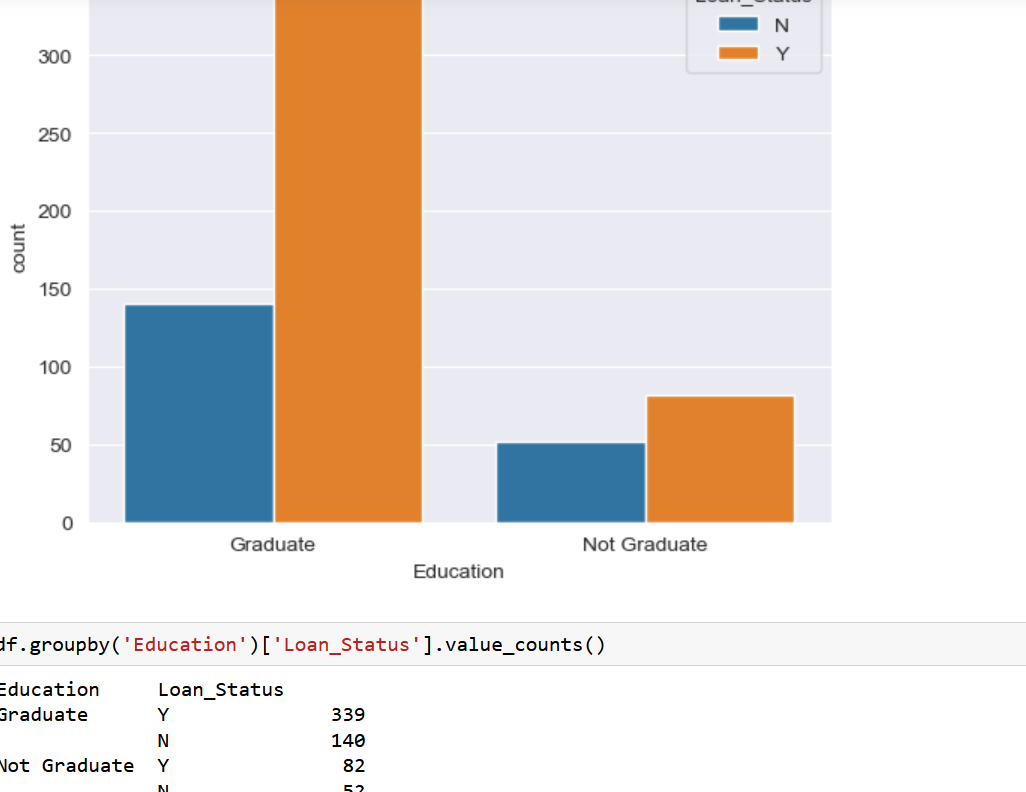


Analysing the distribution of loan status based on marital status, we make the following important observation:

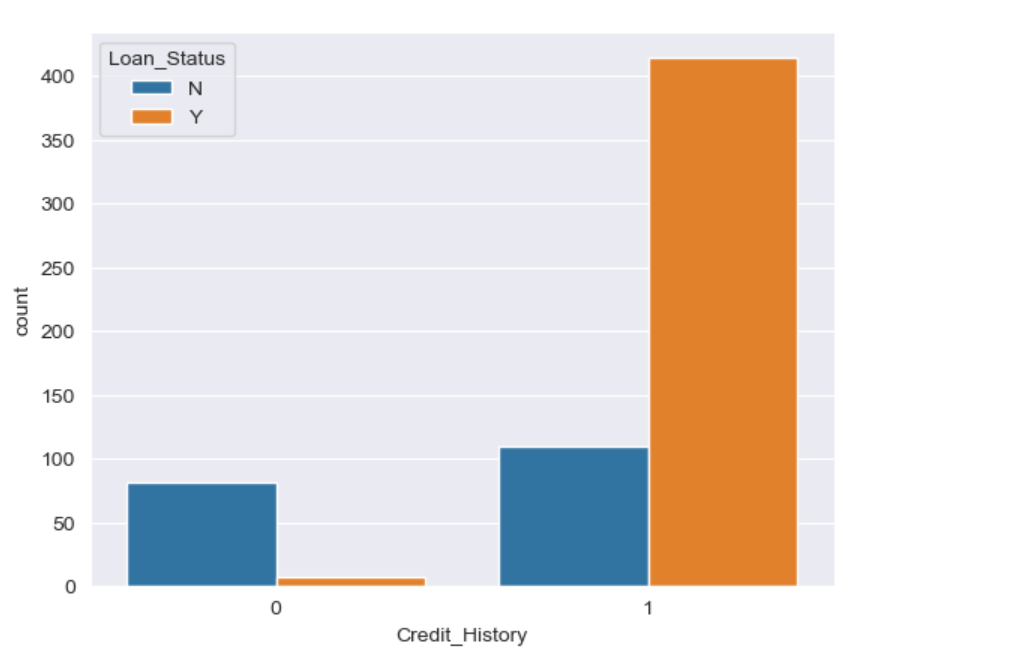
Among applicants who are not married, 134 loans have been approved ('Y') and 79 loans have been rejected ('N'). On the other hand, among married applicants, 288 loans have been approved and 113 loans have been rejected. This observation suggests that a higher proportion of married individuals have been approved for loans compared to those who are not married.

This marital status-based analysis provides insight into the potential influence of marital status on loan approval rates within the dataset. Further investigation could explore the underlying factors contributing to this difference in loan approval between married and unmarried individuals.

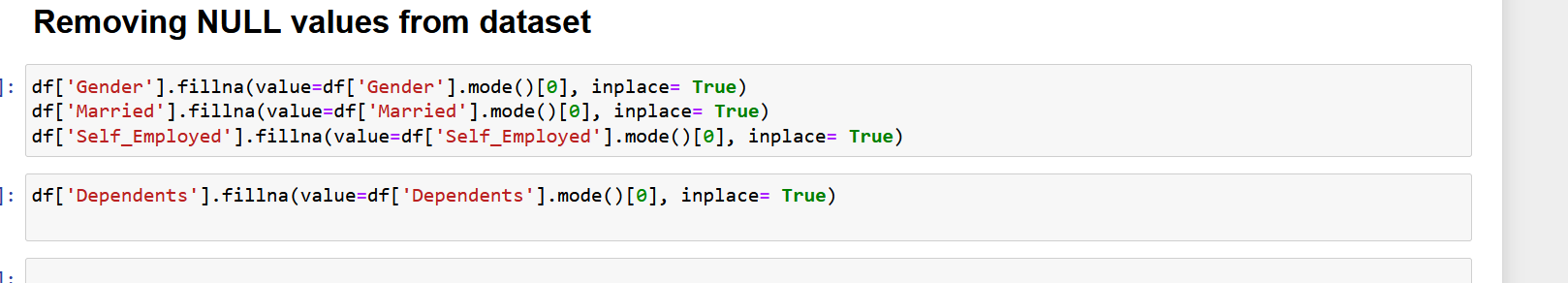
* Graduates generally have higher approval rates.



* High credit\_history is associated with higher approval chances.



# Handling Missing Values



# Data Preparation

# Categorical Data has been converted to ordinal data.

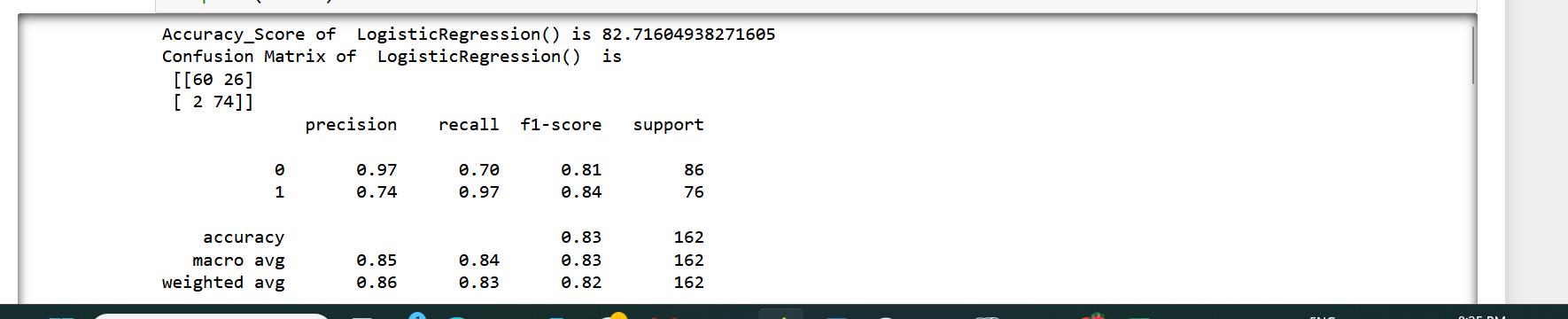
# Outliers are present in the data specifically for the numerical columns like applicant\_income, loan\_amount and coapplicant\_income. Convert these variables into logarithmic data, this will help in removing the skewness that we had observed in data itself. Logarithmic transformation won’t have impact on the smaller values but would reduce the larger values in loan\_dataset.

# Before splitting the data into train and test would create dummy values of the ordinal data.

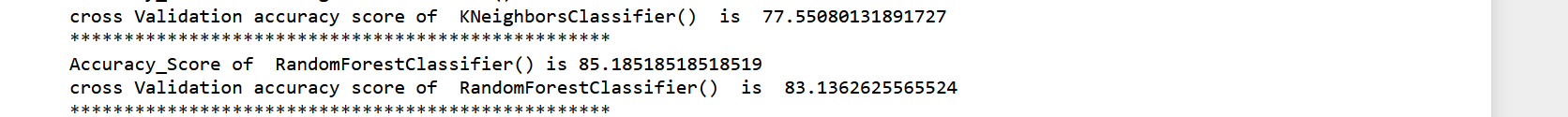
# Model Building

We developed two models:

1. **Logistic Regression**: Achieved approximately 83% accuracy.



1. **RandomForestClassifier**: Achieved approximately 85% accuracy.



* The base model that we’ll be using is Logistic Regression. Train and test split will be of 70:30 ratio. With Logistic Regression we’ll be getting the accuracy of approx. 83%.
* Second model that we have implemented is RandomForestClassifier. This provides us the accuracy of 85%.

# Model Evaluation

Although Logistic Regression showed higher accuracy, the RandomForestClassifier had fewer false positives and higher recall.

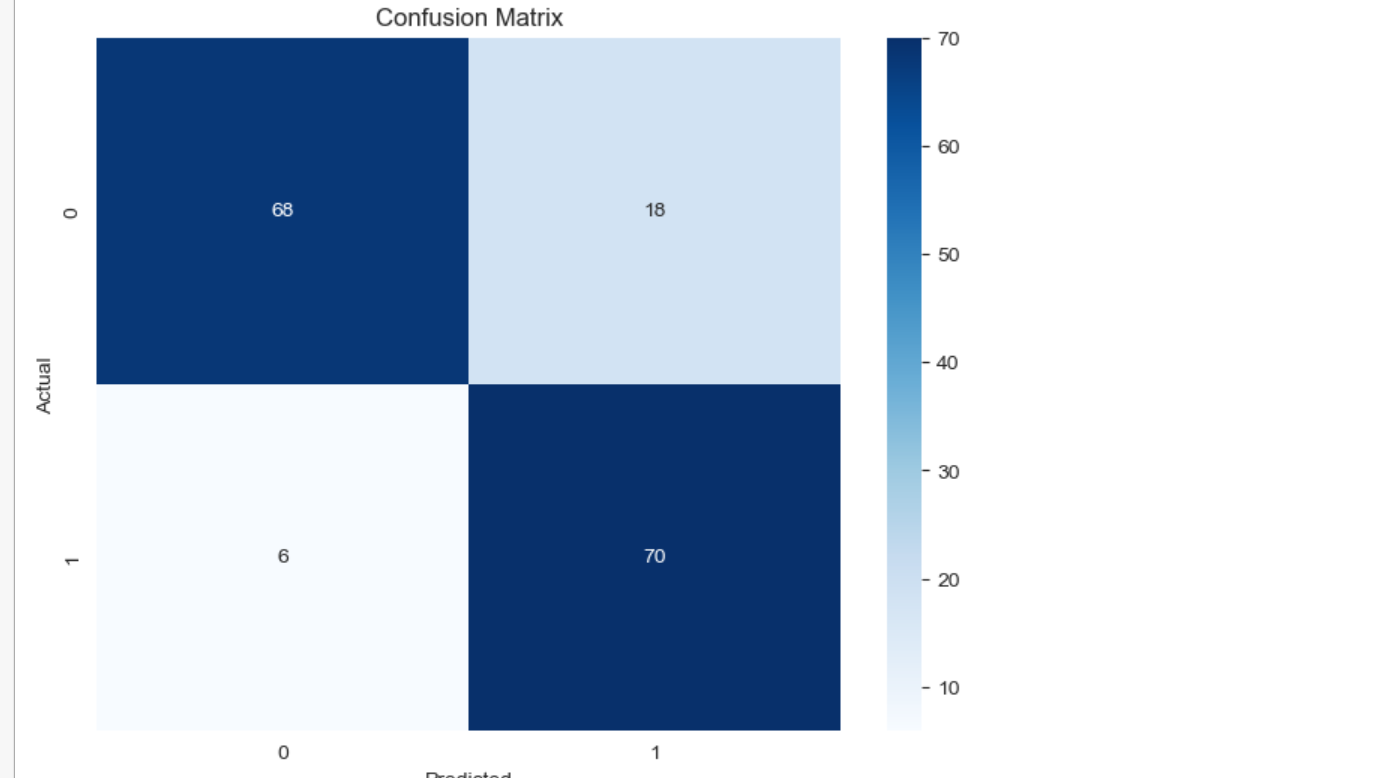
# Upsampling Imbalanced Dataset

Experimenting is the key for building a robust model to do classification or regression. The same technique has been applied over here as well to increase the recall as per our model needs. For that python library ADASYN has been used from imblearn.over\_sampling.

It will use the technique where over\_sampling from the minority class is done, this leads to upsampling of the monority class where very less of loan\_applicants have been rejected.

Results of using the upsampling technique are: -

Recall values has increased and model is able to correctly classify the applicants to whom loan should not be provided.



# Conclusion

* We explored various techniques to handle missing values and outliers.
* Conducted univariate, bivariate, and multivariate analyses.
* Built and evaluated classification models with a focus on recall and false positives.
* Improved model performance through feature engineering and upsampling techniques.

# Key Takeaways

* Precision and recall are crucial in classification models, especially in applications like loan approval.
* Feature engineering and addressing imbalanced datasets can significantly improve model performance.
* Machine learning can enhance and automate decision-making processes in the financial sector.