By: Igord Diamandi

Loan Approval Prediction  
Using Machine Learning

Contents

[Introduction 2](#_Toc175311685)

[Data Overview 3](#_Toc175311686)

[General information 3](#_Toc175311687)

[Categorical Fields 4](#_Toc175311688)

[Continuous Fields 7](#_Toc175311689)

[Target Variable Description 8](#_Toc175311690)

[Design 9](#_Toc175311691)

[General Approach 9](#_Toc175311692)

[Research Subjects 9](#_Toc175311693)

[Data Handling Strategy 9](#_Toc175311694)

[Algorithm Validation 9](#_Toc175311695)

[Data Preparation 10](#_Toc175311696)

[EDA 10](#_Toc175311697)

[Data Cleansing 10](#_Toc175311698)

[Data Imputation 10](#_Toc175311699)

[Data Transformation and Engineering 10](#_Toc175311700)

[Feature Selection 10](#_Toc175311701)

[Algorithms 11](#_Toc175311702)

[Method Selection 11](#_Toc175311703)

[Model Validation 11](#_Toc175311704)

[Hyperparameters Tuning 11](#_Toc175311705)

[Implementation 12](#_Toc175311706)

# Introduction

This project leverages historical loan approval data to develop a machine learning model capable of predicting the likelihood of future loan approvals. This model aims to provide accurate predictions that can support decision-making processes in financial institutions, by analysing patterns and trends embedded within past loan applications.

The dataset used as the foundation for this project contains records of previous loan applications, including details about applicant demographics, financial histories, loan parameters and the outcomes of their loan requests. Through careful examination and preparation of this data, key features have been identified and engineered to serve as inputs for the predictive model.

In the following sections, this document will provide a thorough overview of the project, beginning with a detailed description of the data and the steps taken to prepare it for analysis. We will then define the target variable, outline the design of the machine learning model, and discuss the methodologies employed in its implementation. The document concludes with a comparison of model performances and the selection of the final model for deployment.

By systematically addressing each of these components, this document aims to provide a clear and comprehensive understanding of the project’s scope, objectives, and outcomes.

# Data Overview

## General information

**Dataset Overview**

The loan approval dataset consists of 13 fields, each providing critical information about the loan application and the applicant. Below is a detailed description of each field:

1. **Loan\_ID (object)**  
   This is a unique identifier for each loan application. Since it serves solely as an identification field with no predictive value, it will be excluded from the training dataset.
2. **Gender (object)**  
   This textual field indicates the gender of the applicant. It captures whether the applicant is male or female. Will be converted to binary categorical during the preprocessing.
3. **Marital Status (Married) (object)**  
   Another textual categorical field, this one reflects the marital status of the applicant, indicating whether the applicant is married or not. Will be converted to binary categorical during the preprocessing.
4. **Dependents (object)**  
   This textual categorical field specifies the number of dependents the applicant has. The values range from 0, 1, 2, to '3+' dependents.
5. **Education (object)**  
   A textual field that indicates whether the applicant is a graduate or not. This field will be transformed into a binary categorical variable during the data preparation stage.
6. **Self\_Employed (object)**  
   This textual field denotes whether the applicant is self-employed. Like the Education field, it will be converted into a binary categorical variable during preprocessing.
7. **ApplicantIncome (int64)**  
   A continuous numerical field (int64) representing the applicant’s income. This field will be used to assess the financial capacity of the applicant.
8. **CoapplicantIncome (float64)**  
   This is another continuous numerical field (float) that captures the income of the co-applicant, if any. It complements the applicant's income in determining the total financial standing.
9. **LoanAmount (float64)**  
   A continuous numerical field that indicates the amount of loan requested by the applicant. This field plays a crucial role in determining the loan approval outcome.
10. **Loan\_Term\_Amount (float64)**  
    This categorical numerical field represents the term of the loan in months. The values will be grouped during data preparation to facilitate analysis.
11. **Credit\_History (float64)**  
    A numerical field that records the applicant's credit history. It contains binary values (1.0 and 0.0) that will be converted into a categorical field in later stages of the project.
12. **Property\_Area (object)**  
    A textual categorical field that describes the area type (Urban, Semi-Urban, Rural) of the property for which the loan is being requested.
13. **Loan\_Status (object)**  
    This is the target variable for the model, represented as a textual categorical field. It has two possible values: "Y" for approved loans and "N" for rejected loans. Will be converted to binary categorical during the preprocessing.

There are total of 614 rows in the data set.

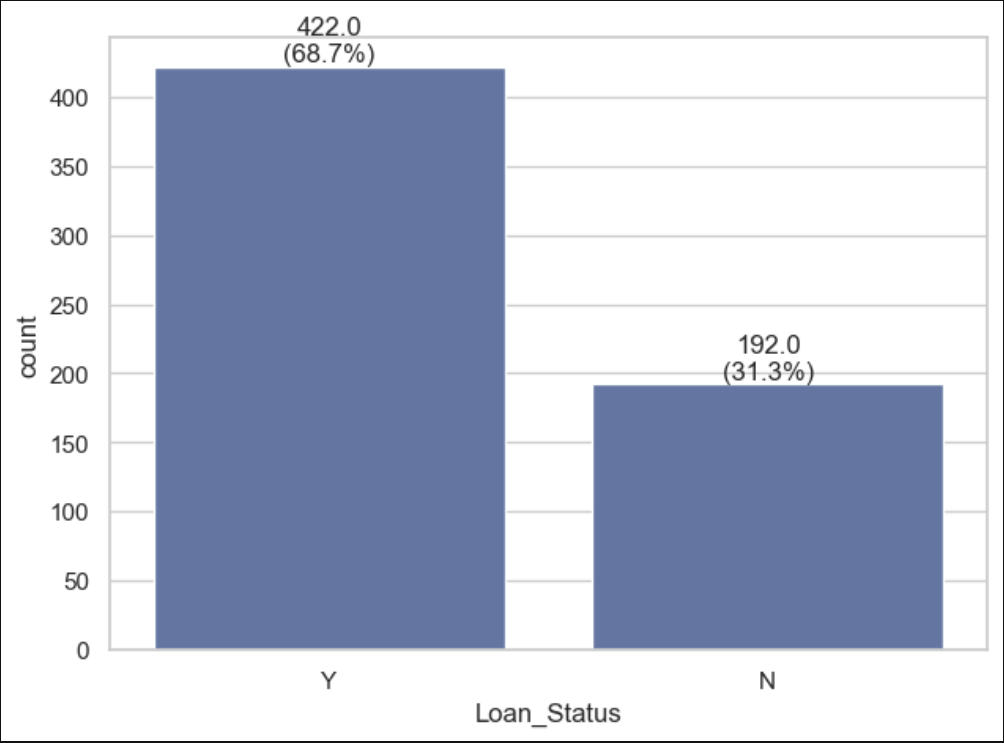
## Target Variable Description

The primary objective of this project is to develop a predictive model that can accurately determine the approval status of loan applications. The target variable for this model is the LoanStatus field, which captures whether a loan was approved or not. This field is a binary categorical variable, indicating either an approval or a rejection, and is crucial for training the machine learning model.

The LoanStatus field is encoded with two possible values:

* **Approved (Y)**: Indicates that the loan application was successful.
* **Rejected (N)**: Indicates that the loan application was not successful.

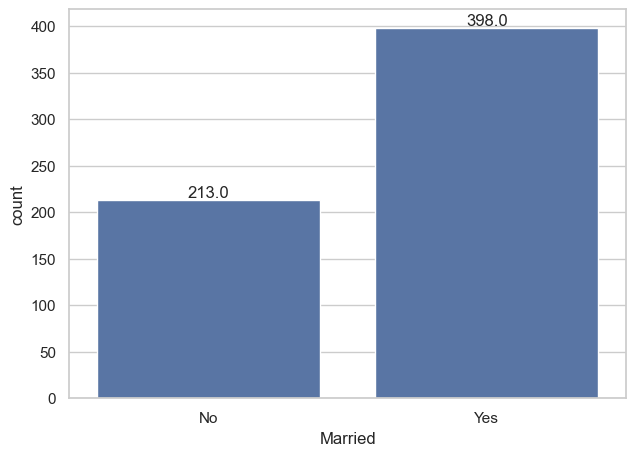
The field is well-prepared for modelling, with no missing values, ensuring that all records in the dataset contribute to the learning process. Additionally, the distribution of the LoanStatus field is balanced, which is advantageous as it prevents the model from being biased towards one outcome. Below is the distribution of the values within this field:

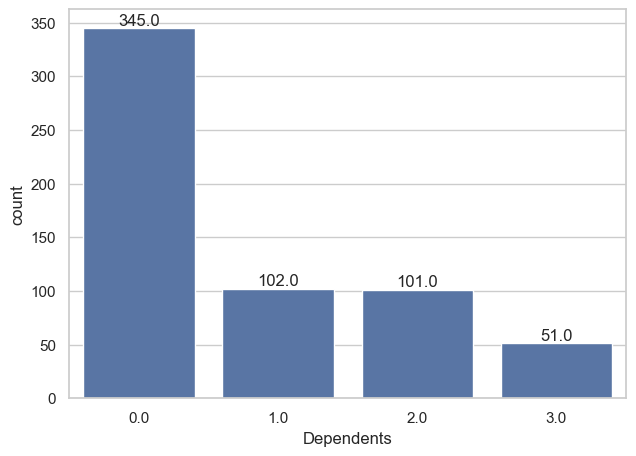


## Categorical Fields

**Gender**  
Unique non-empty values – 2: ‘Male’ and ‘Female’  
Number of empty values – 3 out of 614 (2.117%)  
Value distribution:  
A graph of a number of people

Description automatically generated

**Married**  
Unique non-empty values – 2: ‘Yes’ and ‘No  
Number of empty values – 0 out of 614  
Value distribution:  


**Dependents**  
Unique non-empty values – 4: 0, 1, 2 and 3+   
Number of empty values – 15 out of 614 (2.443%)  
Value distribution:  


**Education**  
Unique non-empty values – 2: ‘Graduate and ‘Not Graduate’  
Number of empty values – 0 out of 614  
Value distribution:  
A graph of a graph with a bar and a number of text

Description automatically generated with medium confidence

**Self\_Employed**  
Unique non-empty values – 2: ‘Yes’ and ‘No  
Number of empty values – 32 out of 614 (5.212%)  
Value distribution:  
A graph of a number of people

Description automatically generated

**Credit\_History**  
Unique non-empty values – 2: 0,0 and 1,0  
Number of empty values – 50 out of 614 (8.143%)  
Value distribution:  
A graph of a credit history

Description automatically generated

**Property\_Area**  
Unique non-empty values – 3: ‘Urban’, ‘Rural’ and ‘Semiurban’  
Number of empty values – 0 out of 614  
Value distribution:  
A graph of a number of blue rectangular objects

Description automatically generated with medium confidence

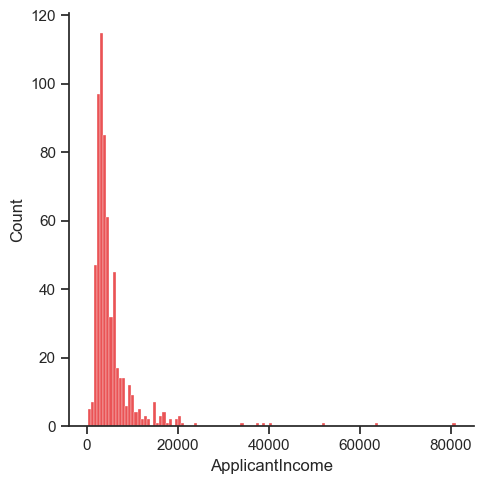
**Loan\_Amount\_Term**  
Unique non-empty values – 10: 12, 36, 60, 84, 120, 180, 240, 300, 360 and 480.  
Number of empty values – 14 out of 614 (2.28%)  
Value distribution:  
A graph of a person and person

Description automatically generated with medium confidence

## Continuous Fields

**ApplicantIncome**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | mean | std | min | 0.25 | 0.5 | 0.75 | max |
| 614 | 5403.46 | 6109.04 | 150 | 2877.5 | 3812.5 | 5795 | 81000 |

Number of empty values – 32 out of 614 (3.583%)   
Distribution charts:  


**CoapplicantIncome**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | mean | std | min | 0.25 | 0.5 | 0.75 | max |
| 614 | 1621.25 | 2926.25 | 0 | 0 | 1188.5 | 2297.25 | 41667 |

Distribution charts ():

A graph of a number of numbers

Description automatically generated

LoanAmount

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | mean | std | min | 0.25 | 0.5 | 0.75 | max |
| 592 | 146.41 | 85.59 | 9 | 100 | 128 | 168 | 700 |

Number of empty values – 32 out of 614 (3.583%)  
Distribution charts:

A graph of a number of red lines

Description automatically generated   
\*\* C*harts are not taking empty values into account for the boundary’s calculation here, so numbers are not the same as what we’ll see after the empty values imputation*  
\*\*\* *Negative boundaries values are meaningless here and will be limited by 0 in the further stages*

# Design

## General Approach

## Research Subjects

## Data Handling Strategy

## Algorithm Validation

# Data Preparation

## EDA

## Data Cleansing

## Data Imputation

## Data Transformation and Engineering

## Feature Selection

# Algorithms

## Method Selection

## Model Validation

## Hyperparameters Tuning

# Implementation