COMP 4948 Assignment 2

Predicting Loan Defaults

Table of Contents

[Problem Definition 1](#_Toc131768339)

[Exploratory Data Analysis 1](#_Toc131768340)

[Data Treatment 1](#_Toc131768341)

[Model Development, Tuning, and Comparison 1](#_Toc131768342)

# Problem Definition

For this assignment, I used a loan default dataset to build several predictive models. The goal was to help predict if loan applicants were likely to default on their loans. The application of this knowledge could help assist banks and those involved in the loan/mortgage process to identify those who are more likely to struggle with the loan they are applying for.

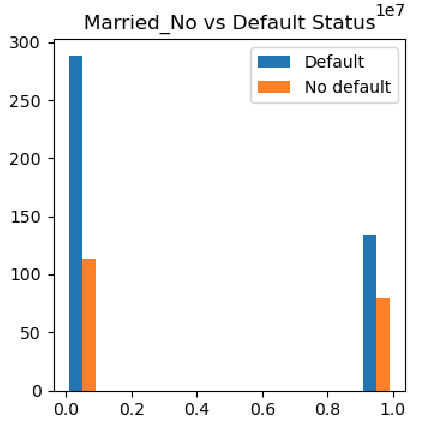
# Exploratory Data Analysis

The loan default dataset features 11 available predictors:

|  |  |
| --- | --- |
| Predictor | Possible values |
| Gender | M or F |
| Married | Yes or No |
| Dependents | 0, 1, 2, or 3+ |
| Education | Graduate or not graduate |
| Self employed | Yes or No |
| Applicant income | Integer ranging from 15,000 to 8,100,000 |
| Coapplicant income | Float ranging from 0 to 4,166,700 |
| Loan amount | Integer ranging from 0 to 70,000,000 |
| Term | Float ranging from 12 to 480 months |
| Credit history | Float ranging from 0 to 1 (binary) |
| Area | Rural, Semiurban, Urban |

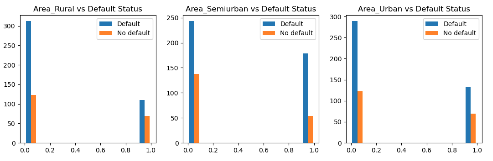
We can examine the relationships between each of these predictors and our target, ‘Status’ (referring to default/non-default status of each loan). Some of the strongest relationships identified are between loan defaults and marriage status, loan amount, credit history, and area (rural, semiurban, or urban).

This distribution shows that those who are married are more likely to default on a loan:



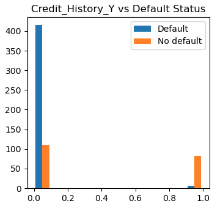
While the proportions look similar between married and not married, there is a slight skew towards non-married loan holders being more likely to default. This may be due to many married households having more than one income, so events like loss of a job are less likely to be as detrimental to loan payments.

Here we can see that those in semiurban areas are more likely to default on their loans.



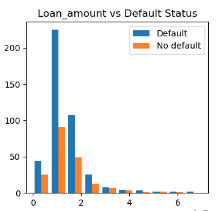
The proportion of defaults vs non defaults is much higher for semiurban areas. Possible explanations for this may include that loan holders may be taking out mortgages or large loans for larger homes relative to their income when compared to urban or rural areas.

Loan holders who did not have previous credit history are very strongly correlated with more loan defaults.



This correlation may speak to an unfamiliarity with loans and financial management; those who don’t have a previous understanding of loans may be more likely to make errors when it comes to managing payments.

The final predictor we will examine here is loan amount; loans of approximately $150,000-$200,000 seem to be the most likely to default.



This might relate to the idea that loan holders who buy homes in this price range might have low-mid range incomes; they are more likely to be impacted by financial events such as job loss and need to resort to defaulting on loans while they sort things out.

# Data Treatment

For this assignment I used several methods to treat the data. One of these methods was dummy variables, replacing categorical (non-numeric) column values with new columns that represent whether or not that data row belongs in which category (for example, if a loan holder is ‘Male’, they would have a 1 under the new ‘Gender\_Male’ column and a 0 under the ‘Gender\_Female’ column). Another method was used on the target column; rather than using these placeholder columns to represent ‘Y’ and ‘N’, I directly replaced those values in the ‘Status’ column with their equivalents (1 for ‘Y’ and 0 for ‘N’).

For missing values, I used KNNImputer from the scikit-learn library. This looks at other rows that have similar values, and uses them to infer what the missing value might be.

\*\*\* talk about scaling if/when used

# Model Development, Tuning, and Comparison