Predicting Loan Eligibility

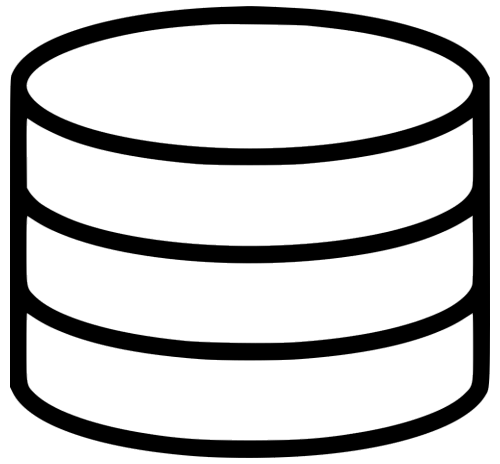
Lighthouse Labs: Mini-Project IV

Michael Halim - July 9, 2021

Introduction

For most people, purchasing property requires a loan from financial companies. It is beneficial for companies and applicants to know if a loan will be approved. The *objective* of this project is to develop and deploy a machine learning model capable of predicting an applicant's loan eligibility based on demographic and financial information.

Outline of Workflow

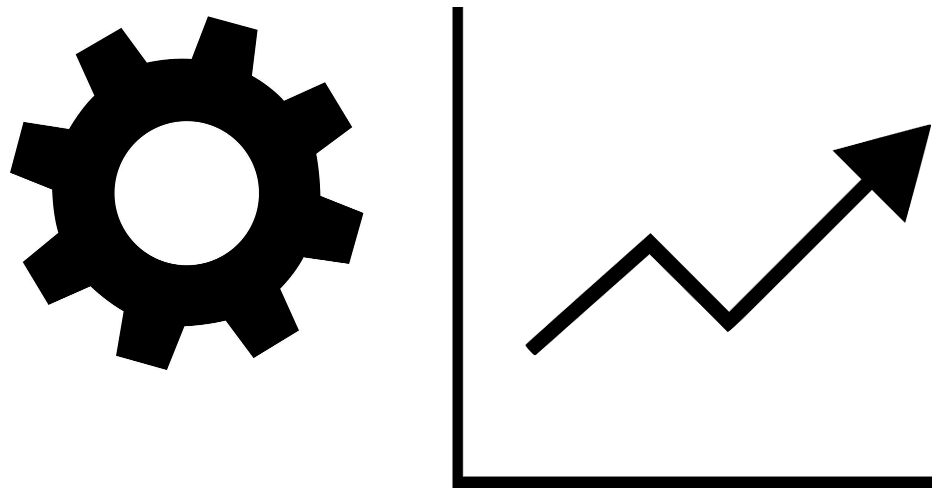


Data Collection

CSV file containing historical applicants' information, such as income and credit history, and loan status

Exploratory Data Analysis

Analyzed trends within the data to determine factors important for modelling



Data Cleaning and Modelling Pipeline

Implemented data cleaning and modelling approaches using a machine learning pipeline

Deployment

Implemented model in a flask web app, which is then deployed to the cloud using AWS.

Hypothesis Generation

1 Applicant Income

Higher income applicants have more money to repay loans

2 Credit History

Applicants with good credit history are more trustworthy with loans

3 Property Area

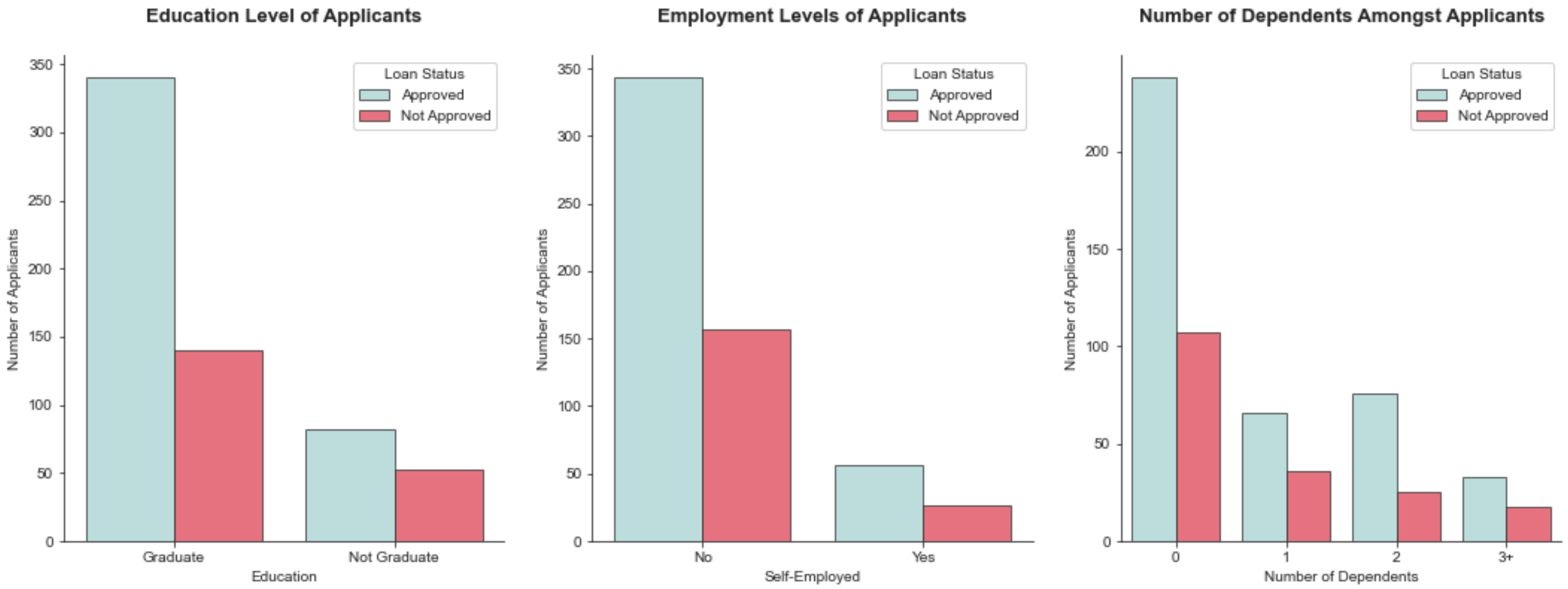
Properties at desired areas are more appealing to financial companies

4 Demographic Properties

Applicants with stable jobs, high education levels, and less dependents may be considered more reliable

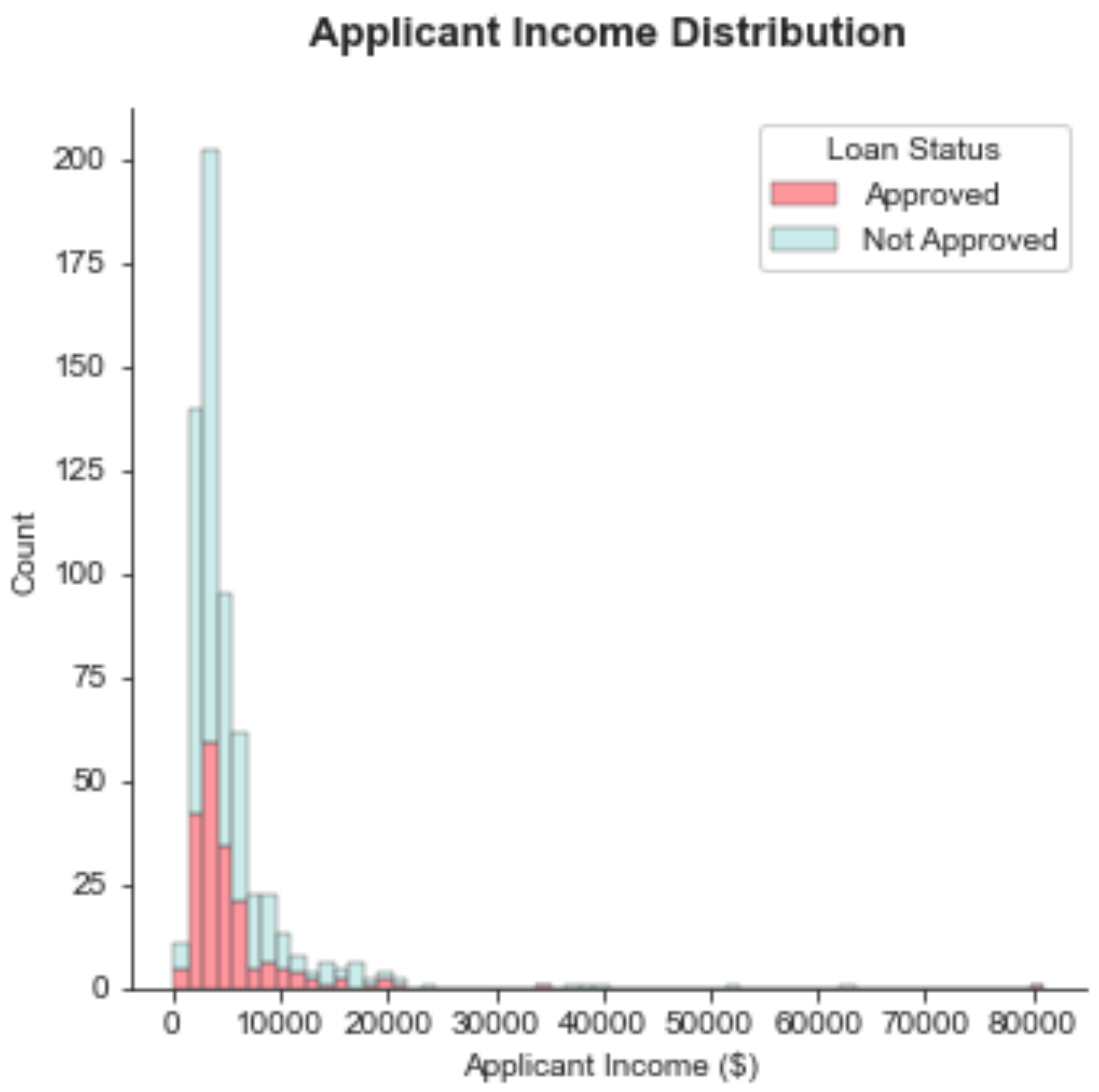
Exploratory Data Analysis

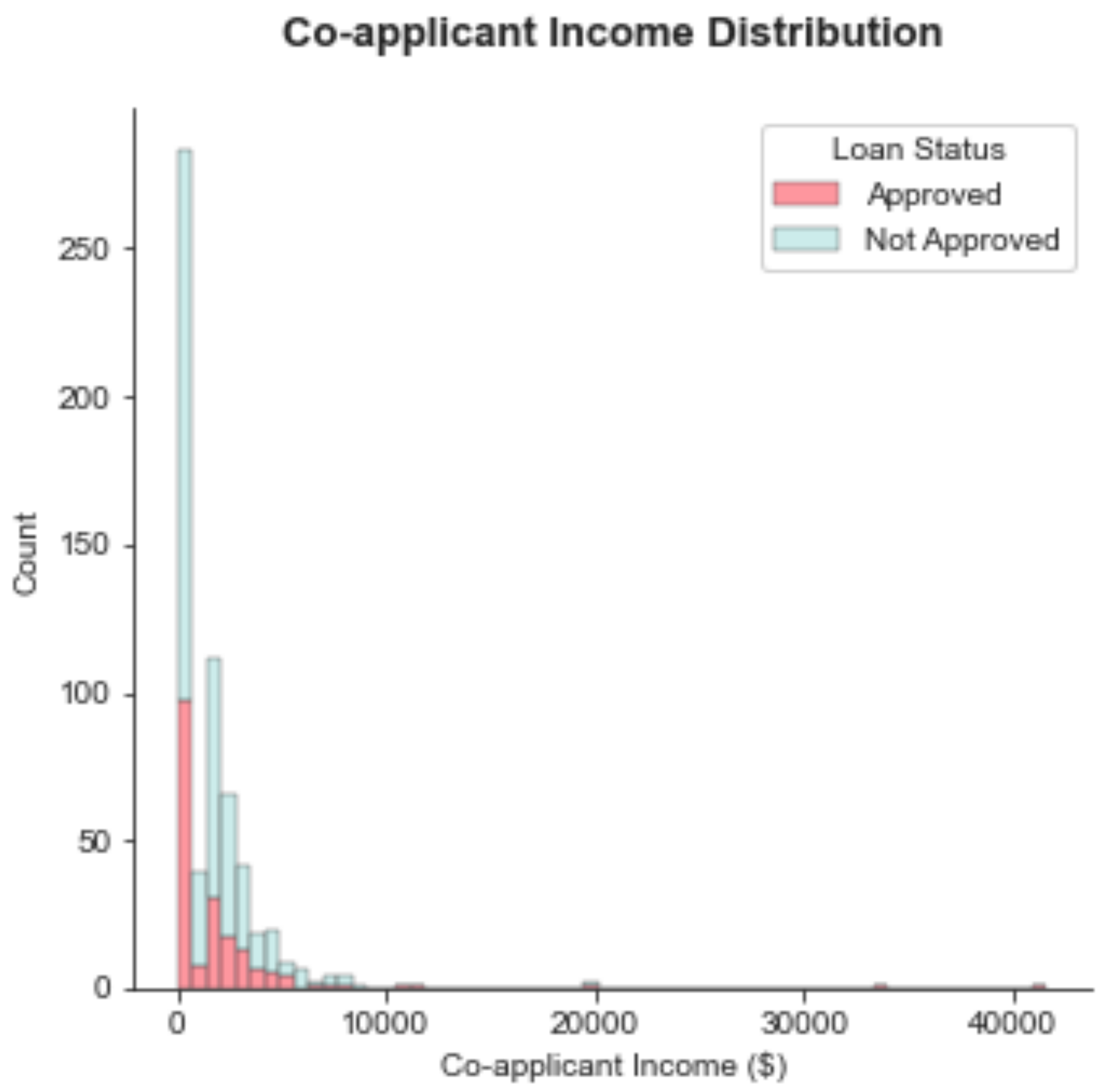
Demographic Patterns

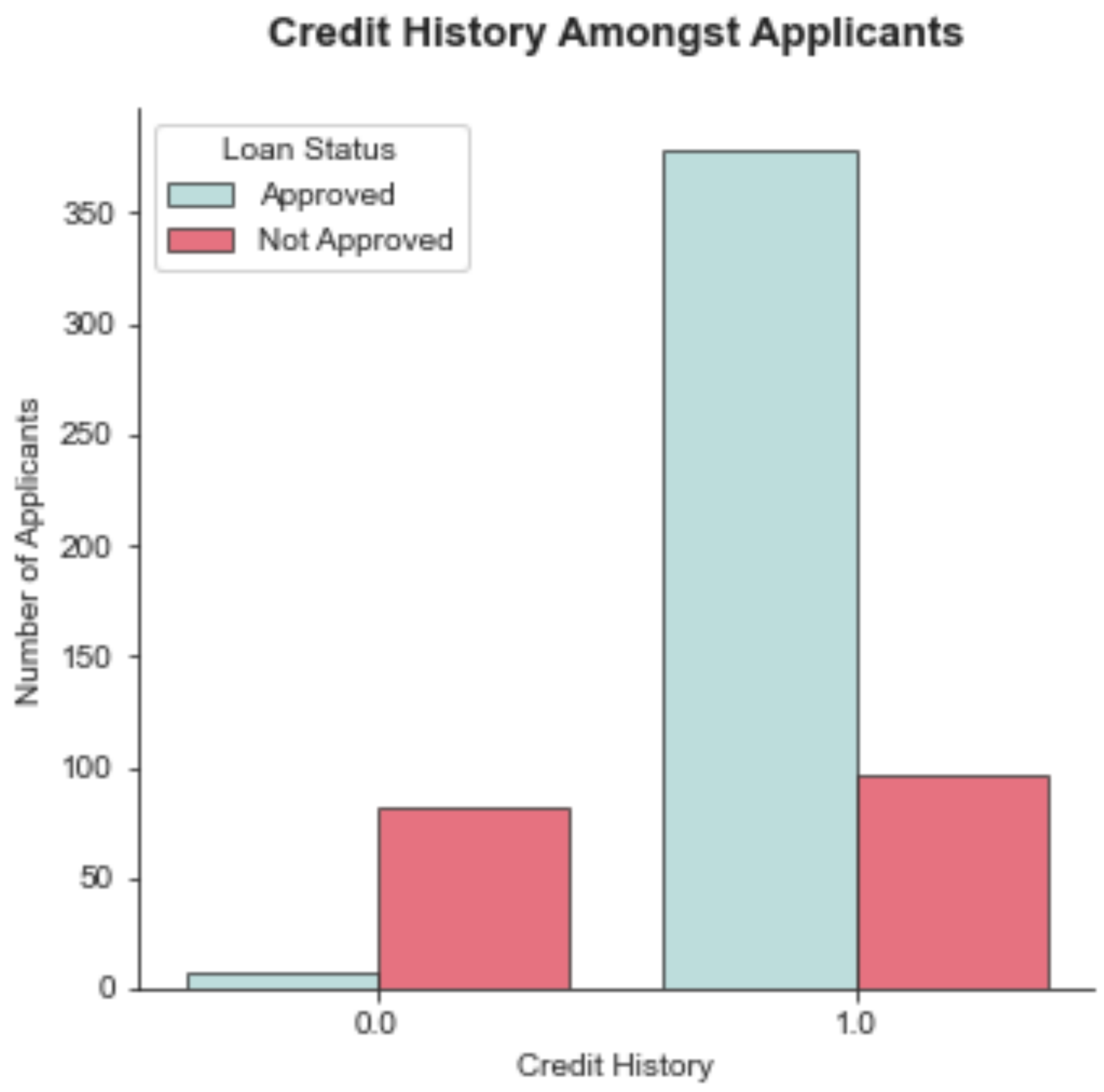


Categorical data is generally unbalanced. Amongst the different groups within a category, the proportions of loan status varies.

Financial Patterns

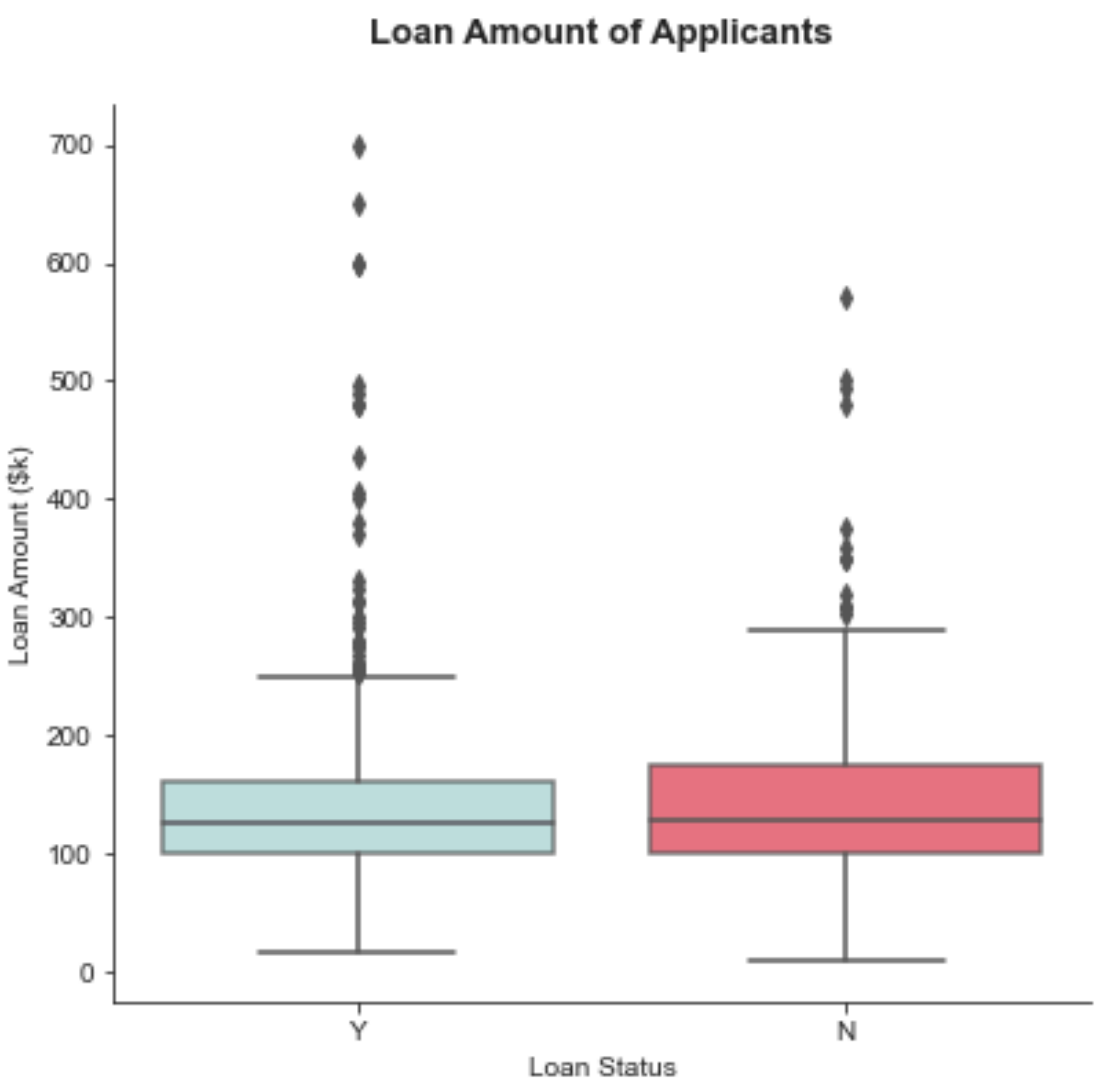






Incomes are heavily skewed; the bulk of applicants make between $0 to $10,000.
Applicants with poor credit history are more likely to get their loans refused.

Loan Patterns



Loans for semi-urban properties are more likely to get approved.

Pipeline Development

Data Transformation Approaches

MISSING DATA

Categorical Data

Impute missing values with most common observation

Numerical Data

Impute missing values with median

TRANSFORMATION

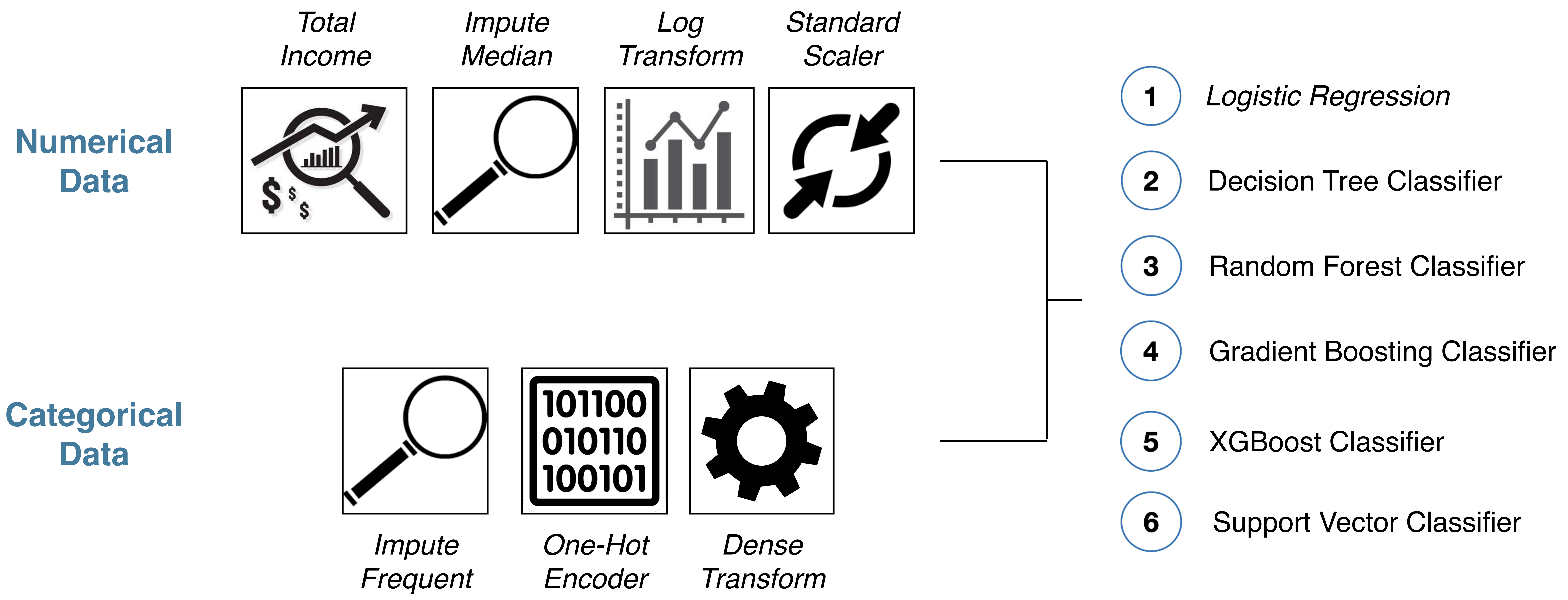
Total Income

Create a total income feature by adding applicant and co-applicant income

Extreme Values

Perform a log-transformation on skewed numerical data

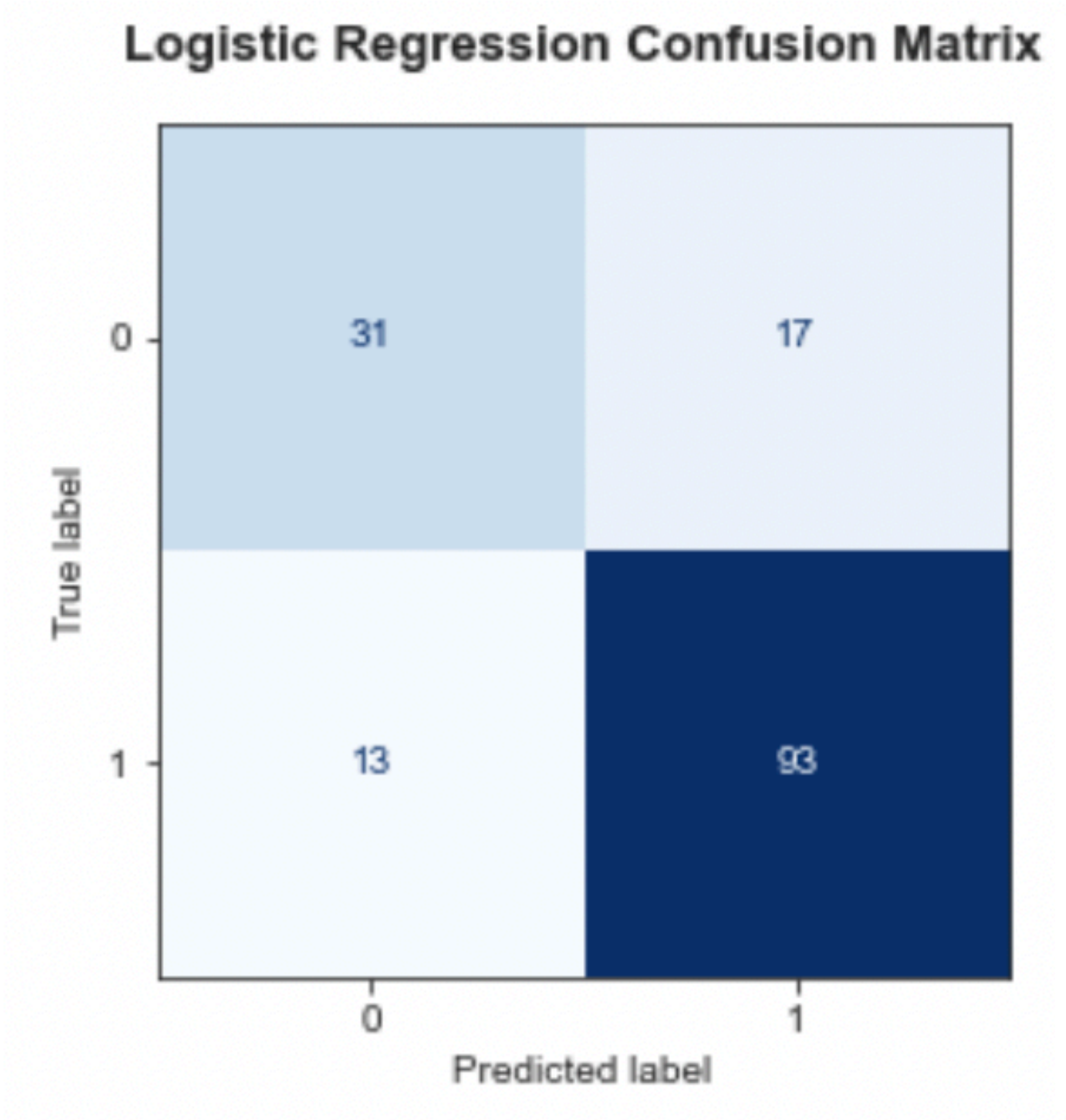
Pipeline

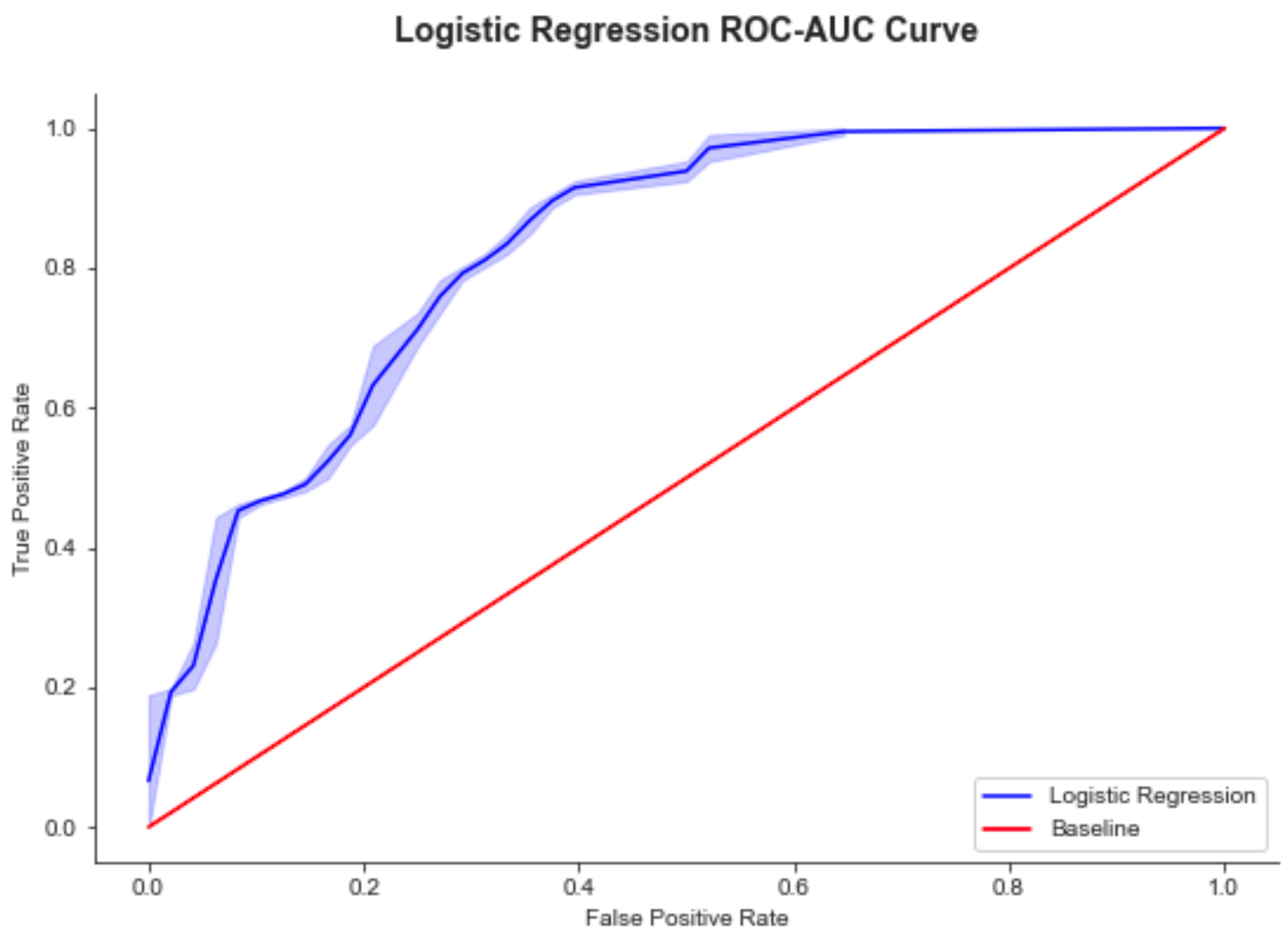


Total Income
Impute Median
Log Transform
Standard Scaler
Numerical Data
Categorical Data
101100
010110
100101
Impute Frequent
One-Hot Encoder
Dense Transform
1 Logistic Regression
2 Decision Tree Classifier
3 Random Forest Classifier
4 Gradient Boosting Classifier
5 XGBoost Classifier
6 Support Vector Classifier

Modelling and Evaluation

Logistic Regression Model





1. *Accuracy Score:* 0.81
2. *Precision Score:* 0.85
3. *Recall Score:* 0.88
4. *F1-Score:* 0.86
5. *ROC-AUC Score:* 0.76

Implemented balanced weighting in logistic regression model (data is slightly unbalanced - 70%-30%).
In doing so, model was able to reduce number of false positives.

Deployment

Quick Demo!

Loan Eligibility Predictor

Basic Information

What is your gender? Male

What is your education level? Graduate

Are you married? Yes

Are you self-employed? Yes

Financial Status

Does your credit history meet guidelines? Yes 4583

Dependent Information

How many dependents do you have? 0 1508

Loan Information

Where is your property located? Semiurban 123 90

Predict Loan Eligibility

Future Improvements

- Model would benefit significantly with larger and more balanced dataset
- Improve model to better predict negative outcomes - reduce false positives - through boosting or oversampling techniques (SMOTE
- Improve web application - features and design