Understanding Loan Approval: a Data Analysis

COGS 108 Fall 2023

Group 119:

- Joseph Lee
- Kenneth Song
- Georgio Feghali
- Lorenzo Ramos
- Sujay Talanki

Research Question:

Is there a relationship between an applicant's credit history (whether or not an applicant′s credit history meets the bank′s guidelines), income, loan amount, employment status, property area, dependents and their loan approval status in America?

Hypothesis:

- We hypothesize that employment status, loan amount, education, income, credit history, and property area collectively influence the likelihood of loan approval

- We expect that applicants with a strong credit history, higher income, and potentially certain demographic attributes, such as being a graduate or residing in specific property areas, are more likely to have their loan applications approved

Brief Background:

- Banks give out loans so that borrowers can repay the money with the interest

- We want to take a look at if we could make a machine learning model to accurately predict whether or not an applicant will receive a loan

Home Loan Approval Dataset Overview

Dataset Source: Kaggle

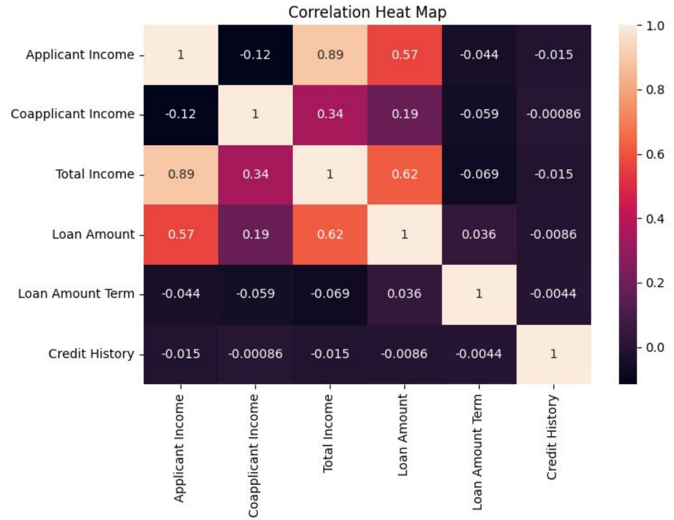
Size: 981 observations (614 train, 367 test)

Variables:

- **Age** (int): Represents the age of the loan applicant in years.
- **Gender** (str): Indicates the gender of the applicant.
- **Marital Status** (bool): Represents whether the applicant is married or not.
- **Education** (str): Indicates the educational status of the applicant (Graduate/Not Graduate).
- **Income** (float): Represents the income of the applicant in USD.
- **Credit History** (bool): Binary indicator of whether the credit history meets the bank's guidelines.
- **Property Area** (string): Represents the area where the property is located.
- **Loan Approval Status** (bool): Binary indicator of whether the loan was approved or not.
- **CoapplicantIncome** (float): Numerical metric representing the income of the co-applicant.
- **LoanAmount** (float): Numerical metric representing the requested loan amount.
- **LoanAmountTerm** (int): Numerical metric representing the term of the loan in months.

EDA

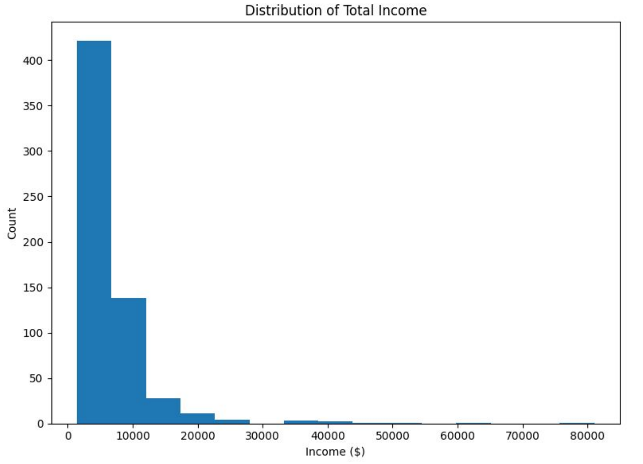
Part 1



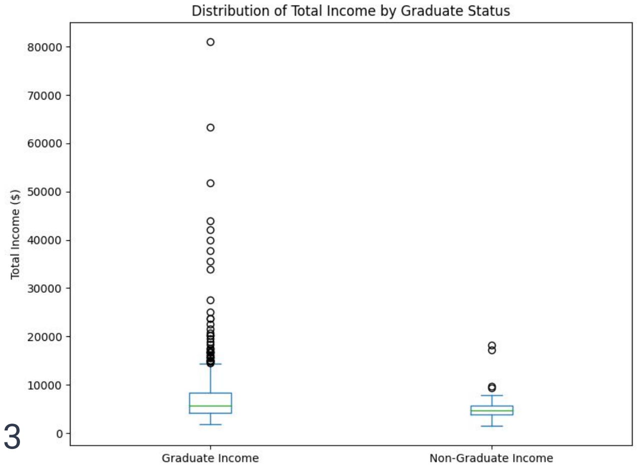
Correlation Heat Map

	Applicant Income	Coapplicant Income	Total Income	Loan Amount	Loan Amount Term	Credit History
Applicant Income	1	-0.12	0.89	0.57	-0.044	-0.015
Coapplicant Income	-0.12	1	0.34	0.19	-0.059	-0.00086
Total Income	0.89	0.34	1	0.62	-0.069	-0.015
Loan Amount	0.57	0.19	0.62	1	0.036	-0.0086
Loan Amount Term	-0.044	-0.059	-0.069	0.036	1	-0.0044
Credit History	-0.015	-0.00086	-0.015	-0.0086	-0.0044	1

Part 2



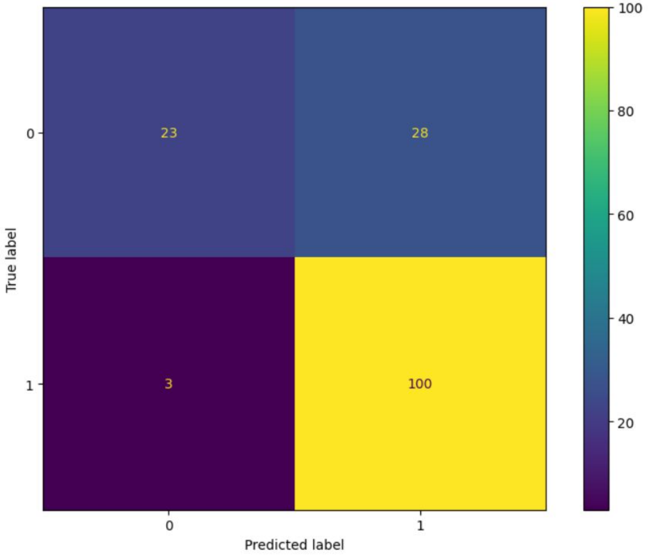
Part 3



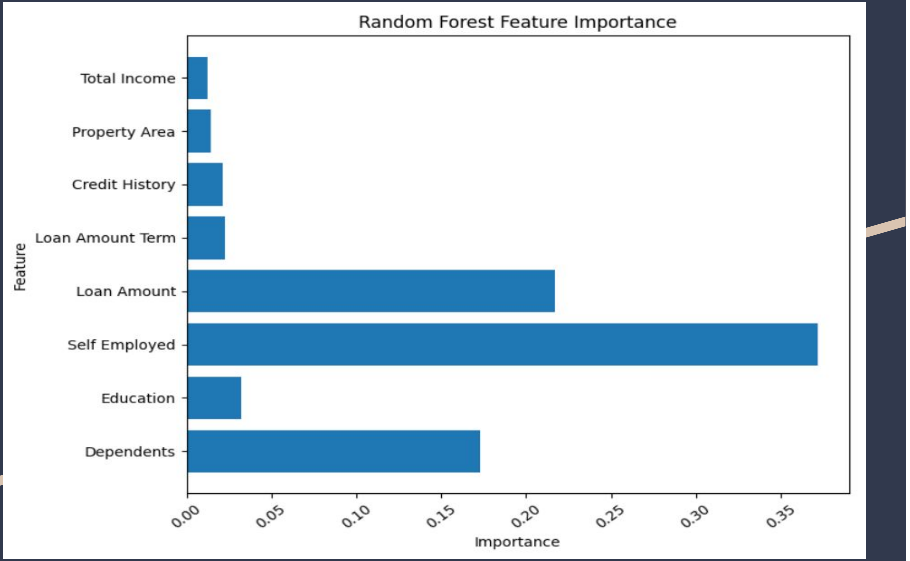
Model Building

- This is a classification task: we need to determine whether a loan will be approved or denied based on the features we are interested in
- We tried two ensemble learning algorithms- (1) LightGBM and (2) RandomForestClassifier. They both combine the results of many decision trees to make an informed decision
- We engineered sklearn pipelines that preprocessed the data (it handled missingness, employed one hot encoding, chose the variables of interest) and compiled the two models.
- We utilized sklearn's GridSearchCV to tune the hyperparameters for each model, and used a 5-fold Cross Validation to determine each model's accuracy
- RandomForestClassifier performed better (84% accuracy vs. 74% accuracy for the LightGBM)

Confusion Matrix



Conclusion



- Throughout our project, we utilized various data visualizations and classification models for our analysis.
- While our analyses suggested that certain predictors–such as Self-Employment, Loan Amount Request, and Number of Dependents–have more impact on loan approval status, due to certain limitations, such as a small dataset and our lack of knowledge of how our data was collected, we cannot be certain whether our findings supports or rejects our hypothesis.
- Moving forward, we believe our project have a lot of potential for expansion. In particular, we believe that having a larger and more reliable dataset would allow us to arrive at a more solid conclusion, as well as implementing advanced modeling techniques.