Project memo

2023-01-16

```
setwd("~/Desktop/Third year/PSTAT 131/project/project data")
train <- read.csv("train.csv"); train
test <- read.csv("test.csv"); test
```

Overview of dataset

- I will be using an open source dataset found on Kaggle, and traced back to a Analytics Vidhya contest.
  - See Kaggle data source here.
  - See original problem here
- The data file is split into a training dataset and a test dataset.
  - The training dataset contains 614 observations and 13 variables.
  - The test datset contains 367 observations and 12 variables (predictors only).
- I will be working with a mix of quantitative and categorical variables.
- There is some missing data. I plan to identify and remove the observations for which there is a missing record.

Research questions

- For my project, I want to build a machine learning model to predict whether a customer will be approved for a loan based on information in their application profile.
- The predictors include Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others.
- The response variable will be Loan_Status (Y/N), and represents loan eligibility.

- I'm also interested in descriptive questions such as:
  - How many applicants are married based on historical data? How many have dependents?
  - How do the incomes of applicants who live in different areas compare?
  - How do the incomes of employed applicants compare to those who are self-employed?
  - Is there a correlation between an applicant's income and the loan amount they apply for?
  - Is there bias? Does eduation level improve loan eligibility? Gender?
  - How do the credit histories of male and female applicants compare?
- Because my response is qualitative (classification) and I have no response for the test set (unsupervised), I will need to use a more complex method such as K-modes clustering.
- However, I want to split my training set into a training set and a test set instead of using the one given in the data file. This way I can use supervised methods such as decision tree, naive bayes, or logistic regression.
- For this project, I will focus on prediction and model selection. But if a functional form for f is avaliable (such as in logistic), some inference would be helpful to my analysis.

Project timeline

- I have already found my dataset. I plan to tidy my data and perform EDA on weeks 2 and 3. For the remainder of the quarter, I will focus on building my model and editing my final draft.
- I expect to be done with my model around week 5-6 and my write up by week 7.