Loan Default Prediction

Milestone 1 - Adithi Mohan

Problem Definition

- **Context**
  - Mortgage loans make up a major chunk of bank profits. However, loan defaults can plummet bank profits, in turn risking the economy.
  - In order to reduce loan defaults, it is imperative to fine-tune the approval process for loan applicants.
  - Pre-IT heuristics were extensive and severely biased; Studies show that customers who belonged to certain minority groups and/or belong to certain income brackets, fell prey to higher mortgages of up to approximately 40% (Bartlett, Morse, Stanton, Wallace, 2019)[1].
  - Data science and machine learning can ensure that the decision making process is unbiased and simple.

- **Objectives**
  - The goal is to create a classification model to predict if a client would default on the loan based on credit history and personal information.
  - BAD (target: # of defaulters) is binary variable, where 1 = client defaulted and 0 = loan repaid
  - The model is intended to help the banks make decisions on approving loans
  - This also ensures profit maximization and, in turn overall reduction of economic risk.

[1] Bartlett, Morse, Stanton, Wallace, 2019; Consumer-Lending Discrimination in the FinTech Era - http://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf

Problem Definition

- **Key Questions/ Insights**
  - What are the values that show strong correlations to highlight loan default status?
  - What are the parameters that make applicants viable candidates for loan approval?
  - What could be possible reasons why applicants may not be viable candidates for loan approval?

- **The Problem Formulation**
  - The focus is to predict clients who are likely to default on their loan. This will allow us to extrapolate the "bad" apples from the bunch in order to create a favorable prediction template.
  - Through the classification model, we can classify whether the loan will be approved or not.
  - Since loan approval has only two outcomes, Binary Classification is the best method of approach.
  - We can set our target as "BAD" for bad credit history
  - The two outcomes from the target are 0 for loan has been repaid and 1 for defaulter.

Data Exploration

- **Data Description**
  - The Home Equity dataset (HMEQ) contains baseline and loan performance information for 5,960 recent home equity loans.
  - The target (BAD) is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. 1 = Client defaulted on loan, 0 = loan repaid
  - 12 input variables were registered for each applicant including loan amount, existing mortgage, property value, credit information, loan purpose (debt consolidation or home improvement), job information, Debt-to-income ratio, etc.

- **Observations & Insights**
  - Under REASON for the loan request approximately 70% of applicants apply for loans for debt consolidation. We can infer that these applicants might pose a problem since they already have existing debt issues.
  - Under JOB approximately 42% of applicants belong in "other" job category. This might need more exploration, however given the dataset, we may only be able to so far.
  - There is a positive correlation between defaulter status and LOAN amount. We can infer that this could be probably due to inability to pay back the high amounts.

Proposed Approach

- **Potential Techniques**
    - Logistic Regression for the probability of falling into either one of the defaulter status
    - SVM would find the boundaries between the observations.
    - Random Forest would perform bagging to find correlating relationships for loan predictions

- **Overall Solution Design**
    - We utilized Confusion Matrix in the design
    - We had to come up with a threshold to create the solution design.
    - In the next steps, we could try playing with threshold value to view possible changes

- **Measures of Success**
    - We could potentially try the above mentioned techniques to see if the results produce less busier and clearer models.
    - The heatmap from the project was very busy and could use some fine tuning if we get rid of some more unwanted variables