Calculating credit worthiness for rural India

Context

In Banking industry, loan applications are generally approved after a thorough background check of the customer's repayment capabilities. Credit Score plays a significant role in identifying customer's financial behavior (specifically default). However, people belonging to rural India don't have credit score and it is difficult to do a direct assessment.

The accompanying file **trainingData.csv** contains some of the information that is collected for loan applications of rural customers. We need to understand the maximum repayment capability of customers which can be used to grant them the desired amount.

Description of variables:

- **Id:** Primary Key
- **Personal Details:** city, age, sex, social_class
- **Financial Details:** primary_business, secondary_business, annual_income, monthly_expenses, old_dependents, young_dependents
- **House Details:** home_ownership, type_of_house, occupants_count, house_area, sanitary_availability, water_availability
- **Loan Details:** loan_purpose, loan_tenure, loan_installments, loan_amount (these contain loan details of loans that have been previously given, and which have been repaid)

Problems:

- Do a descriptive analysis of all the variables.
- There is a new customer who needs a loan. Which models will be best suited to predict the loan_amount that can be granted to the customer?
- Build a model to predict the maximum loan_amount that can be granted to the customer. Which all variables are good predictors?
- Build atleast one model from scratch that fits this data, without using any third party packages like sklearn, glm, lm, rpart, etc. You are free to use linear algebra packages like scipy, numpy or any blas derivative. We would be more interested in the convergence of the algorithm rather than the prediction accuracy.
- Is loan_purpose a significant predictor? The business has insisted on using loan_purpose as a predictor. If it is not already a significant contributor, can we still modify the model to include it?
- How will you measure the fitness of the model? Which metrics (accuracy, recall, etc.) are most relevant?

Expectation:

- You will attempt to solve all the problems. This will include doing some statistical analysis and model building (using R or Python). The code should be documented.

- You will share the R/Python code with us, along with a small document or presentation describing your approach.
- You can reach us regarding any question you might have regarding the exercise.