Calculating credit worthiness for rural India

Context

In the 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 a 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 all of which have been repaid)

Problems:

- Do a descriptive analysis of all the variables.

Expectation: Share the output in either notebook or a presentation format.

- 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?

Expectation: Support the answers based on data analysis done in the previous step.(nothing to be explicitly shared but more of answering the question) **Q&A format**

- Build a model from scratch to predict the maximum loan_amount that can be granted to the customer. Which all variables are good predictors?

Expectation: Don't use 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.
Share the output in notebook/Codebase.

- 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?

Expectation: Support the answers based on data analysis or data pattern(nothing to be explicitly shared but more of answering the question) **Q&A format**

- How will you measure the fitness of the model? Which metrics (accuracy, recall, etc.) are most relevant?

Expectation: Support the answers based on data analysis and business relevance (nothing to be explicitly shared but more of answering the question) **Q&A format**