Loan Interest Rate Prediction

Problem Statement:

Have you ever wondered how lenders use various factors such as credit score, annual income, the loan amount approved, tenure, debt-to-income ratio etc. and select your interest rates?

The process, defined as 'risk-based pricing', uses a sophisticated algorithm that leverages different determining factors of a loan applicant. Selection of significant factors will help develop a prediction algorithm which can estimate loan interest rates based on clients' information. On one hand, knowing the factors will help consumers and borrowers to increase their credit worthiness and place themselves in a better position to negotiate for getting a lower interest rate. On the other hand, this will help lending companies to get an immediate fixed interest rate estimation based on clients information. Here, your goal is to use a training dataset to predict the loan rate category (1 / 2 / 3) that will be assigned to each loan in our test set.

You can create feature engineering by using any combination of the features in the dataset to make your loan rate category predictions accuracy higher. Some features will be easier to use than others.

Variable	**Definition**
Loan_ID	A unique id for the loan.
Loan_Amount_Requested	The listed amount of the loan applied for by the borrower.
Length_Employed	Employment length in years
Home_Owner	The home ownership status provided by the borrower during registration. Values are: Rent, Own, Mortgage, Other.
Annual_Income	The annual income provided by the borrower during registration.
Income_Verified	Indicates if income was verified, not verified, or if the income source was verified
Purpose_Of_Loan	A category provided by the borrower for the loan request.
Debt_To_Income	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income.
Inquiries_Last_6Mo	The number of inquiries by creditors during the past 6 months.
Months_Since_Deliquency	The number of months since the borrower's last delinquency.
Number_Open_Accounts	The number of open credit lines in the borrower's credit file.
Total_Accounts	The total number of credit lines currently in the borrower's credit file
Gender	Gender
Interest_Rate	Target Variable: Interest Rate category (1/2/3) of the loan application

You're required to:

- Build a machine learning model in a language of your choice, preferably R or Python, with using the data provided
- Model Accuracy level minimum 70%-80%

File for Submissions:

1) **Final dataset** with selected features for training & testing the model
2) **Documentation** describing the flow, packages used, functions created, etc. (in Word Document), with contents as follows:

A) Data Exploration

- What did you discover from the datasets given?
- Include any charts that you created as well

B) Data pre-processing

- What data had to be changed / replaced
- What are the new variables as result of feature engineering
- Which datasets did you merge
- How did you derive your new variables (if any)
- How did you prepare your final training dataset

C) Modeling

- Which models were used
- What was the model performance criteria & what were the results
- How did you improve the model performance

3) Presentations (PPT slides)

- Brief slide on business presentation
- Your slide should include
  - ✓ Model development process i.e. missing data assumption, new variables creation, etc. (1 slide)
  - ✓ Model selection process i.e. model assessment, error rate, etc. (1-2 slide)
  - ✓ Final model, key features, and model performance (1-2 slides)
  - ✓ Business problem & solution
- Maximum: 7 slides (excluding cover page/ dividers)