1. Read both the datasets
2. Create Interest Rate Variable in Test set
3. Create a New column representing whether the row coming from train set or test set
4. Check the column names matching or not
5. Concatenate the data frames
6. Check the data types
7. Drop the Id and Remove amount funded by investors because that variable is a future variable.
8. Remove % symbol from Interest rate, Debt to income ratio
9. Convert Amount.Requested', 'Interest.Rate','Debt.To.Income.Ratio', 'Open.CREDIT.Lines', 'Revolving.CREDIT.Balance to Numeric.
10. Create two columns with Fico Range and calculate the mean, create that as new fico column.
11. Clean the employment Length Varaible.
12. Select all the categorical columns in to separate object
13. Check that Object, as it has data column in it, drop the data column.
14. Create dummy variables for Loan Length,Loan purpose, Home ownership and state columns, make sure that there are not too many dummies created, check the frequency and control the dummies based on frequency.
15. Check the missing values and impute them with mean
16. Separate the train and test datasets
17. Remove data column from test and data as well as Interest rate from test
18. Split the train dataset into train1 and train2 datasets,
19. build the model on train1 and check the intercept and coefficients
20. Import the MAE from sklearn and findout the MAE.
21. validate the model on train2.
22. Submit the predicted values
23. Perform regularization techniques like Ridge with different alphas ( penalties)
24. Find out the best estimators
25. Fit the regression model with best estimators, find the coefficients
26. Try with Lasso Regression and find out which variables can be dropped
27. Find out the MAE and select the best estimators, fix the model.