Group Loans: Predicting Risk for a group using random forest classifier on group dataset.

Data cleaning/Feature engineering:

1. Trimming down columns in data-set from 180+ to around 50, based on relevance towards prediction of risk and whether or not it could be turned into a meaningful numeric value which might enhance the model's predictive power.

The following changes were done:

- Selected Variables: ['groupName', 'customereducationalqualification', 'spouseeducationalqualification', 'familymemberscount', 'age', 'everDefault', 'everNPA', 'groupSize', 'maxRelCatProp', 'casteSimilarityProp', 'occupationHomogeneity', 'ownAsset', 'maritalStatus', 'ownHouse', 'farmerCategory', 'totalLand', 'min_age', 'max_age', 'var_age', 'dif_age', 'avg_expenditure', 'min_familymemberscount', 'max_familymemberscount', 'avg_familymemberscount', 'var_familymemberscount', 'dif_familymemberscount', 'log_minTotalAssets', 'logMaxAssets', 'logAvgAssets', 'avg_totalLand', 'var_totalLand', 'ownAssetProp', 'marriedProp', 'ownHouseProp', 'landFarmerProp', 'branchNameLC1', 'categoryLC1', 'districtnameLC1', 'villagetypeLC1', 'villagetypeLC2', '('category', 'GEN')', '('category', 'OBC')', '('category', 'SC')', '('category', 'ST')', 'logIncome', 'logExpend', 'casteHomogeneity', 'religion']

- Dummy-Variables: Caste--SC,ST,GEN,OBC,OTHERS. Since OTHERS were very few combined them with GEN i.e. Dummy for GEN=1 also when dummy for OTHERS =1 & GEN = 1.

- Padded Variables: districtname,villagetype

- Conversions:
    1. customereducationalqualification :: Converted into numerical values by assigning score for the customer's qualification.
       ILLITERATE : 0
       PRIMARY : 1
       SSC : 2
       HSC/HIGH SCHOOL : 3
       GRADUATION : 4
       POST GRADUATION : 5

       Took into account the spouse's education to fill the NaN values for this column. If a specific customer had no record of their qualification, then took the spouse's qualification instead.
    2. Income/Expenditure : Used log(.)

3. Religion : Since the count for other religions were very less than HINDU and MUSLIM hence seperated religion into 3 categories only
   HINDU : 0
   MUSLIM : 1
   OTHERS : 2
4. farmercategory:
   Landless : 0
   Marginal : 1
   Small : 2

Forming Group level data::

1. Grouped the entire data set by groupName column and took mean and median for the rest of the columns.

2. Median is taken for logIncome,logExpenditure,totalAssets,religion

3. Mean is taken for the rest of the variables

Machine Learning Model::

The model used is the Xgboost classifier, we trained the data using ‘default’ as our dependent variable. Default =1 if proportion of default in the group >0 and default = 0 otherwise.
The following columns were dropped on the grounds of redundancy:

var_age,everDefault,logAvgAssets,var_familymemberscount,ownAsset,avg_expenditure,maritalStatus,everNPA,spouseeducationalqualification,casteSimilarityProp,var_totalLand,maxRelCatPr-op

Results of the model

ROC_AUC_score: 72.6 to 74.9%
Fscore: 57.4 to 60.8%
Precision: 49.9 to 52.5%
Recall: 55.4 to 57.6%
Gini: 42.6 to 45.7%

1. Tell about all tables briefly
2. tell about groups - footnote
3. Add a table on occupational categories
4. Table 2 - recurring expenses; Only one NPA people,
5. Talk to Sir-> However, it is also argued that proximity could be a double-edged sword, and that too much proximity might be detrimental (See for example \cite{la2003related}, \cite{haselmann2018rent})

Prevalence of informal risk sharing within networks, especially in developing economies, is well known and could lead to a possible negative `domino effect'\footnote{The domino effect occurs when a member of a credit network subsequently defaults because of defaults by other members} at times of distress.

Castes thus formed tight-knit social groups with informal, mutual arrangement to offer support like loans and jobs to people within them. This enabled villages to be largely self-sufficient economic entities that mostly operated statically for subsistence. And even today, caste-based networks help in finding jobs, accessing credit, starting businesses, providing insurances against income shocks and major contingencies into old age (\cite{munshi2017caste}). Castes constitute exogenously formed networks of informal risk sharing (\cite{mobarak2013informal}).