LOAN DELINQUENCY PREDICTION

-Manpreet Kaur

UNDERSTANDING BANKING TERMS

- - **Debt to income ratio**:

> measures the amount of income a person or organization generates in order to service a debt. A low DTI ratio indicates sufficient income relative to debt servicing, and makes a borrower more attractive.

- **- Loan to Value ratio :**

> A loan-to-value (LTV) ratio compares the amount of a loan you're hoping to borrow against the appraised value of the property you want to buy. Lenders use LTVs to determine how risky a loan is and whether they'll approve or deny it. A higher LTV ratio suggests more risk because there's a higher chance of default. The more money a lender gives you, the higher your LTV ratio and the more risk they're taking. If you're considered a higher risk for the lender, this usually means that:

1. It's harder to get approved for loans.
2. You might have to pay a higher interest rate.
3. You might have to pay additional costs, such as mortgage insurance.

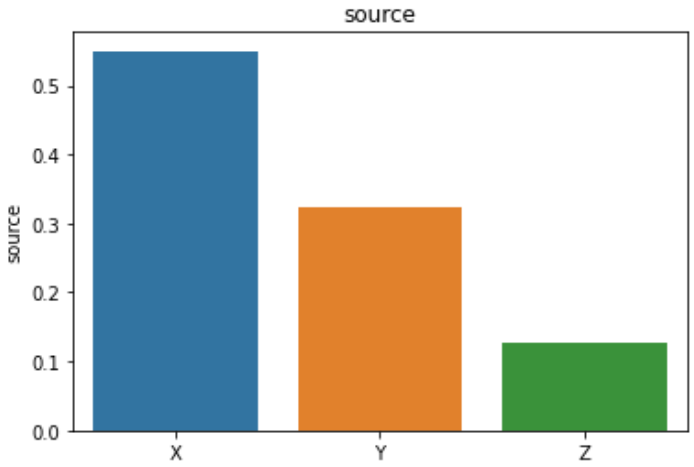
- **- Insurance percent**

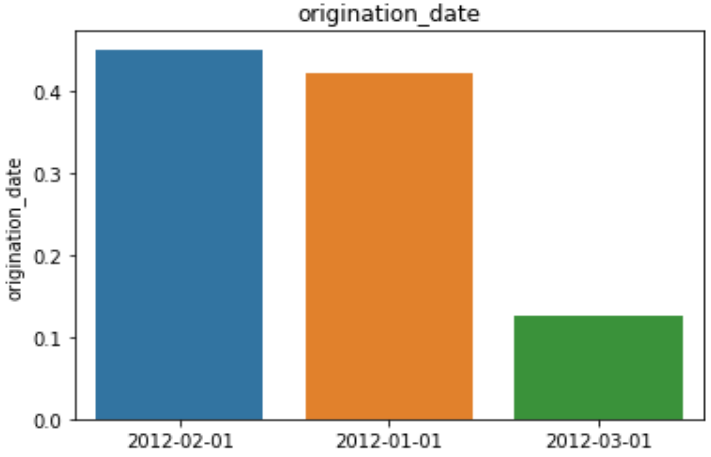
>a provision in a health insurance contract stipulating that the insurer and insured will share covered losses in agreed proportions. For example, the insurer may be required to pay 80 percent of the insured's hospital costs with the insured responsible for the remainder.

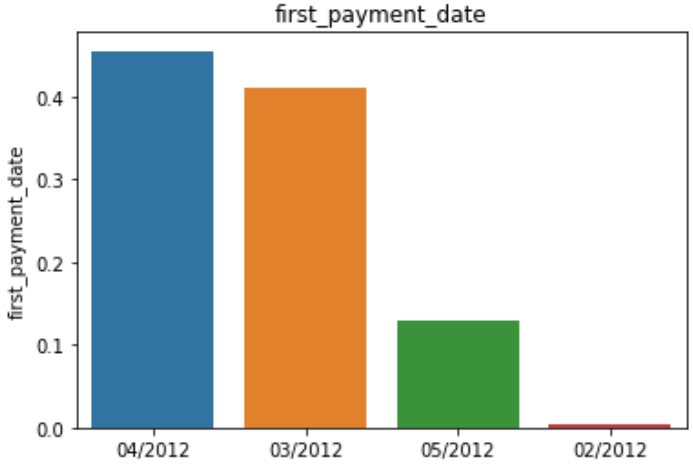
BASIC DATA STUDY

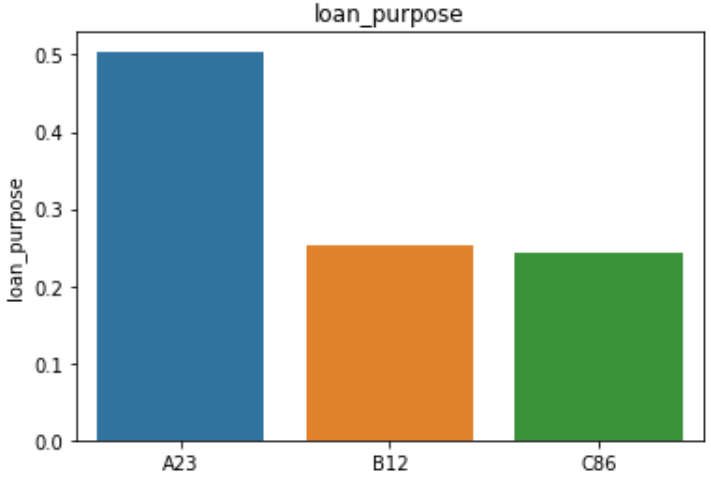
- There are 92846 rows in the train dataset and 23212 rows in the test dataset.
- There are 24 numerical columns and 5 categorical columns in the train dataset.
- Loan_ID does not contribute to the outcome
- interest rate seems to be fairly distributed
- num of borrowers:2 > 1
- DTI Ratio seems to be normally distributed
- borrower_credit_score has outliers on the lower side
- co-borrowers credit score is lesser than the borrowers
- **Data is highly imbalanced**

UNIVARIATE ANALYSIS:
Categorical columns

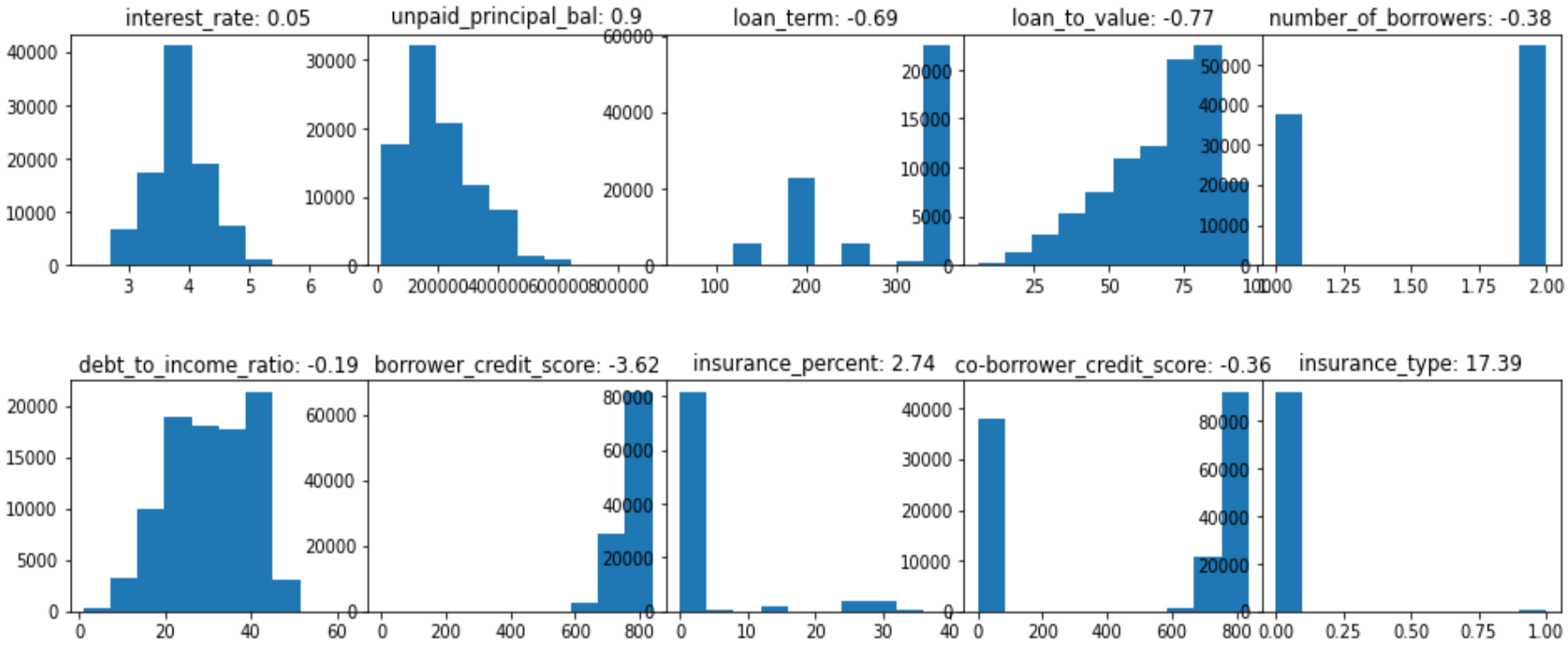






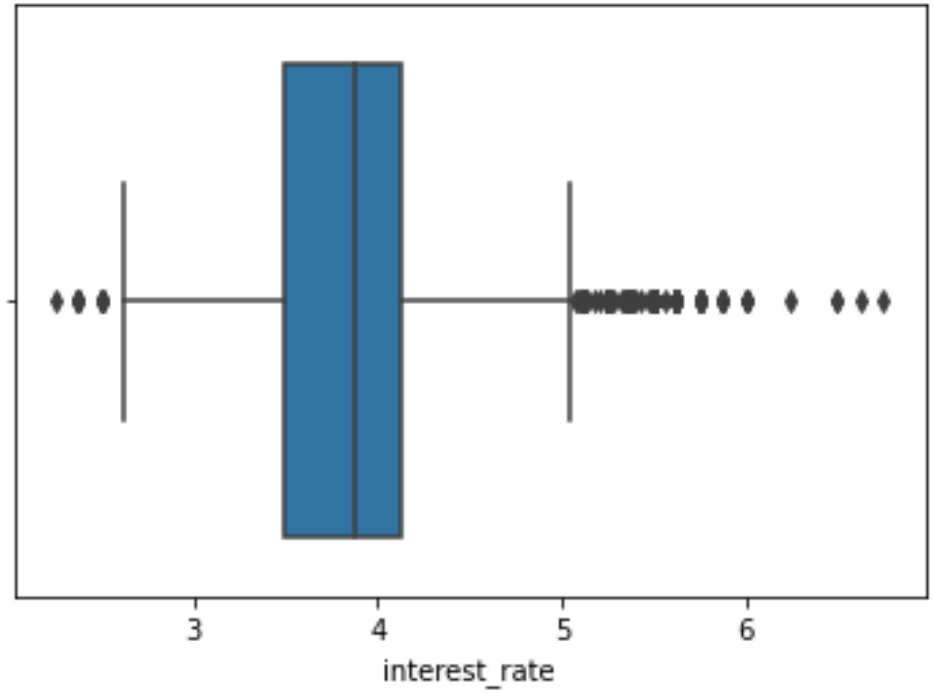


UNIVARIATE ANALYSIS:
Numerical columns



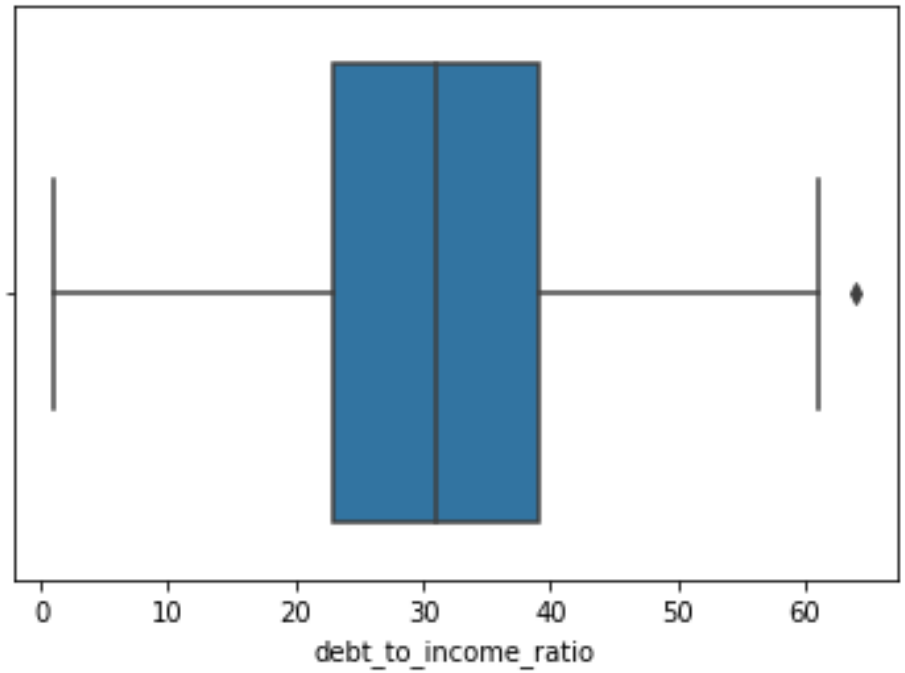
If skewness is less than −1 or greater than +1, the distribution is highly skewed.
If skewness is between −1 and −½ or between +½ and +1, the distribution is moderately skewed. If skewness is between −½ and +½, the distribution is approximately symmetric

Interest_Rate



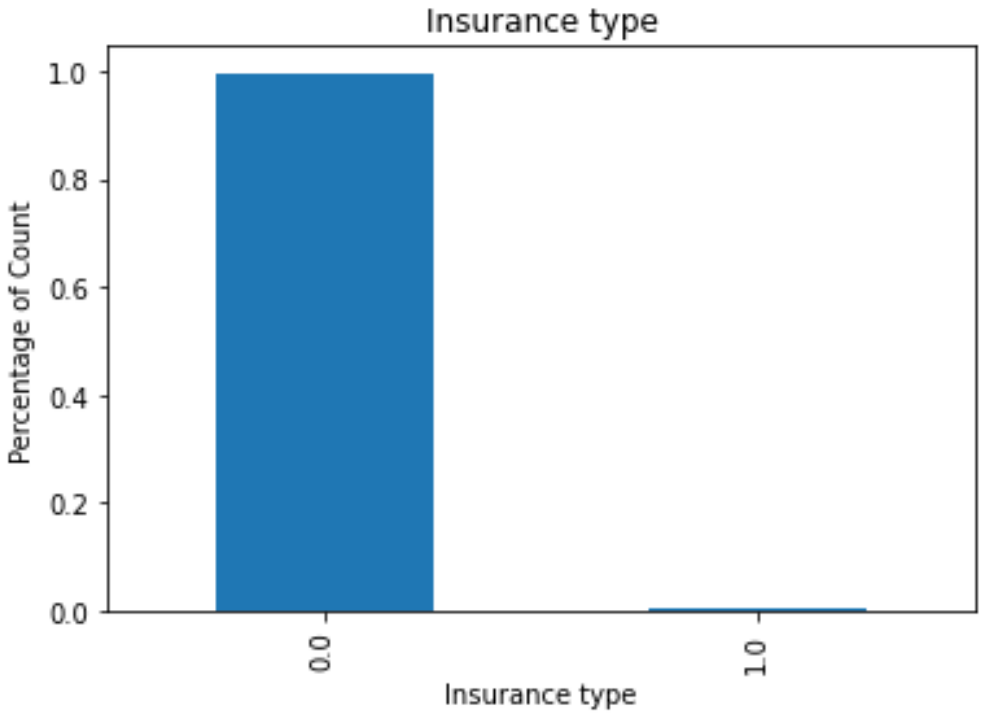
The datapoints outside interest 6 in the train dataset are clearly outliers

Debt_to_income ratio



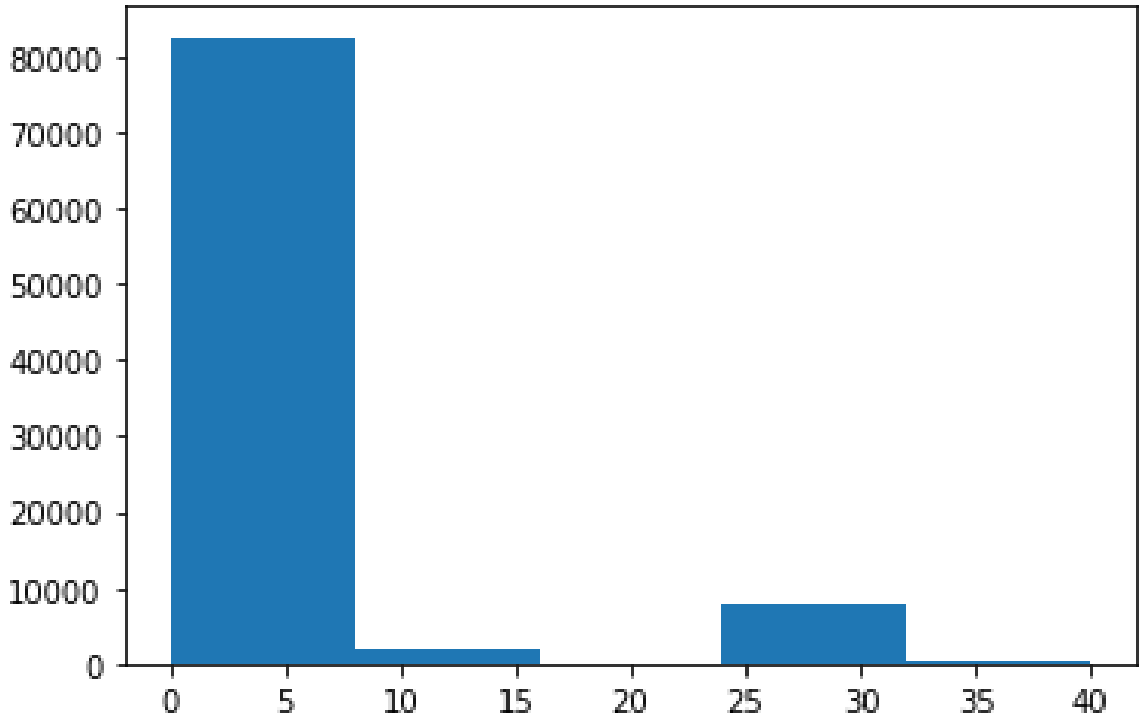
Outlier in the dataset as compared to the test data

Insurance_type



The insurance_type column is highly redundant.
Hence, should be removed from the dataset

Insurance_percent



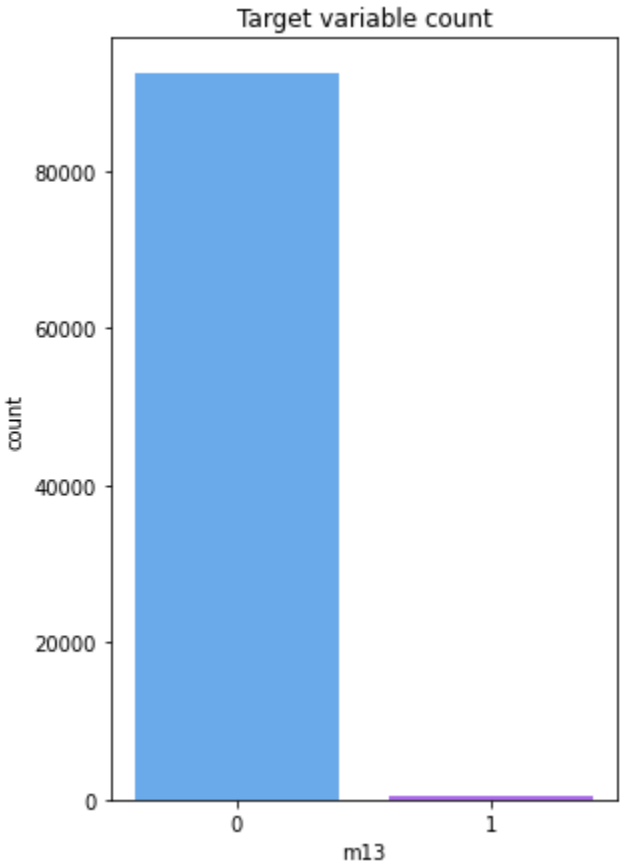
Redundant column too. Hence, dropped during model building.

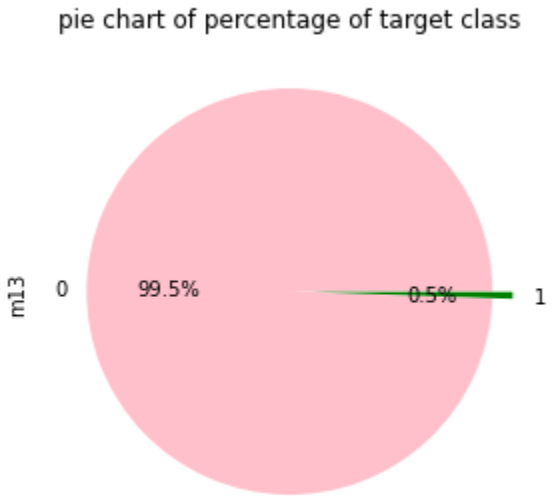
Delinquency for 12 months

	#Unique Values
m1	4
m2	4
m3	5
m4	5
m5	7
m6	7
m7	8
m8	9
m9	10
m10	11
m11	12
m12	12

These are the different delinquency values for each month.
Indicates the different periods of delinquency throughout the dataset

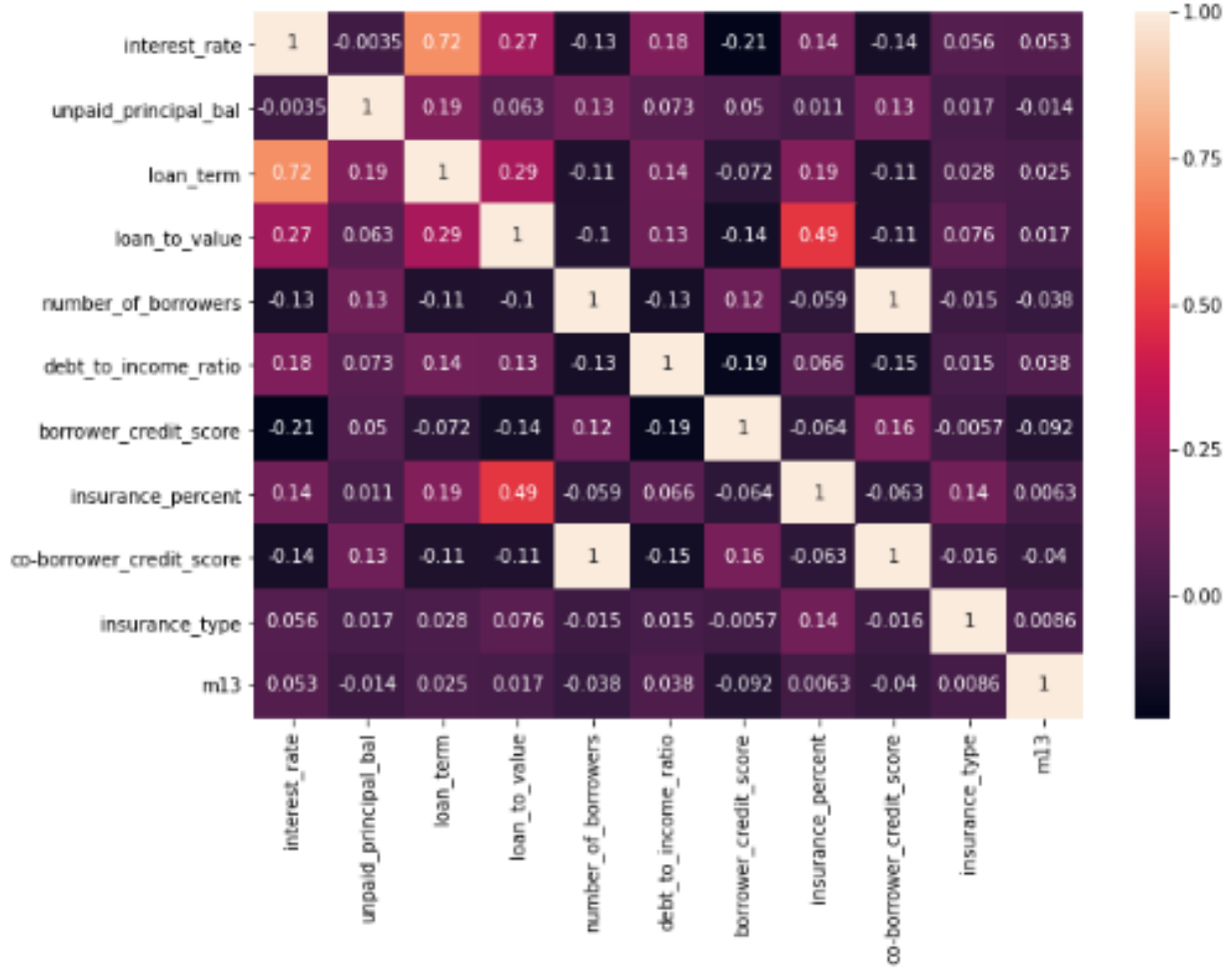
Target Variable – m13





The dataset is highly imbalanced and has majority of rows for which delinquency is 0. Therefore, this dataset will be balanced by oversampling using ADASYN method from imblearn

Correlation between numerical columns

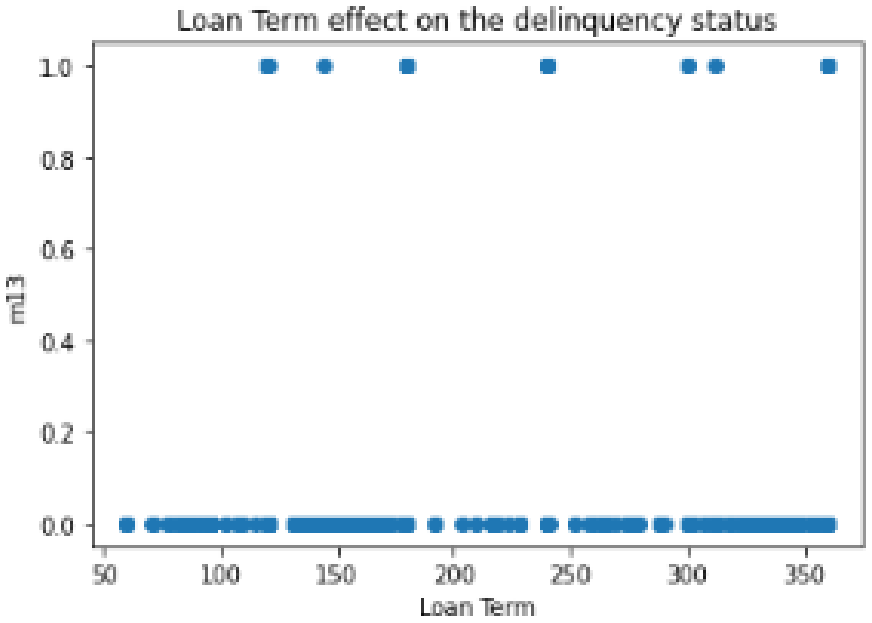


1. loan_term and interest_rate are highly correlated - 0.72

2. LTV ratio and insurance_percent are also sufficiently correlated - 0.49

3. Number of borrowers and co-borrower credit score is 100% correlated - 1

loan_term



The loan_term column is biased towards more number of days and there is no visible correlation between loan_term and m13. Therefore, we can drop this column.

Data Processing

- We drop the outliers observed in the dataset. (observed in previous slides)
- We LabelEncode the financial institutions and give each institute a unique number.
- Source and loan_purpose both have 3 unique values only. These columns are one hot encoded.
- Drop the redundant columns : first_payment_date, origination_date, insurance_percent, number_of_borrowers, insurance_type, loan_term
- The number of borrowers is highly correlated to the co-borrower credit score. Also, the 2 scores of the borrowers can be averaged and combined into one column
- StandardScaler used for unpaid_principal_bal and avg_credit_score

SAMPLING

- The dataset is highly imbalanced and hence a balancing technique has to be applied.
- Oversampling technique is applied to balance this dataset
- ADASYN method from imblearn has been applied to balance the dataset

Splitting the dataset

- The dataset has been split into test and validation set using StratifiedKFold method into 5 folds.
- These 5 sets have been used to test the various classification models used.

DIFFERENT MODELS USED

- 5 different models have been used to predict the values of the validation set:
  - 1. Logistic Regression
  - 2. RandomForest
  - 3. DecisionTree Classifier
  - 4. XGBoost classifier
  - 5. AdaBoost
- RandomForest gave the best model.
- Hyperparameter tuning was done on the model using GridSearchCV and the best estimators were chosen.
  - max_depth : 19
  - n_estimators : 141
- This model was then used to predict values on the test dataset given.

THANK YOU