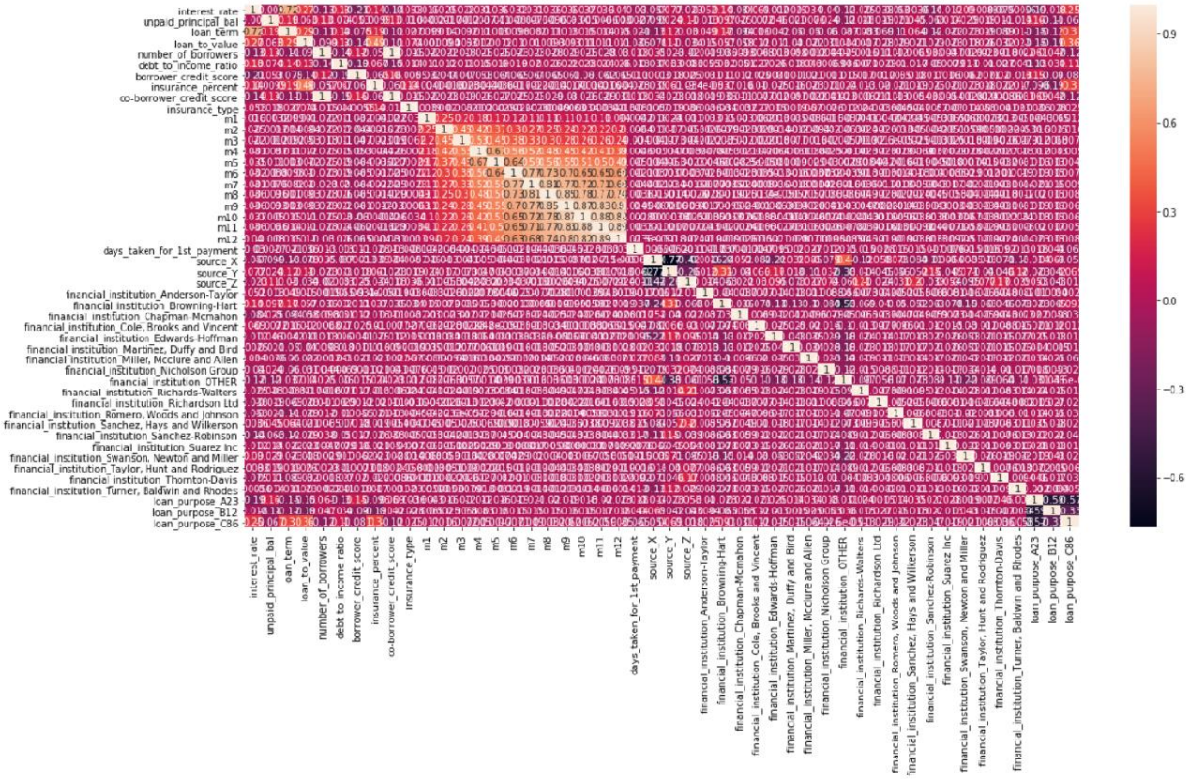
Data Understanding:

- By inspecting the data using head () and info, observed that data is having both numerical and categorical values.
- It is also obvious that there are no null values in the given datasets
- Inspected the unique values in categorical columns and observed the statistical details of numerical columns
- Calculated the default rate and found that the dataset is a unbalanced date and default rate is 0.55%
- Calculated the correlation between the features and found that some of the feature are highly correlated.



Data Preparation:

- Derived a new column day_taken_for_1st_payment which the difference between first_payment_date and origination_date. It signifies the number days taken by the customer for 1st payment.
- Encoded the categorical attributes “source”, “financial_institution” and “loan_purpose” using One-Hot Encoding method.
- Dropped the columns “loan_id”, “origination_date”, “first_payment_date”, “source”, “financial_institution” and “loan_purpose”. Because they are not need for model.

Train-Test Split:

- I have used to test-train split from sklearn to split the dataset into test and train.

Standardization:

- Used standard scalar from sklearn to standardize the features

Seshagiri Rao Suravaram

Model Building:

- Used Logistic Regression for model building
- As there is high correlation multi collinearity effecting the model.

Feature Selection:

- Used Recursive Feature Elimination for reducing the number variables.
- Using RFE with feature = 15, built the model again

Variance Inflation Factor:

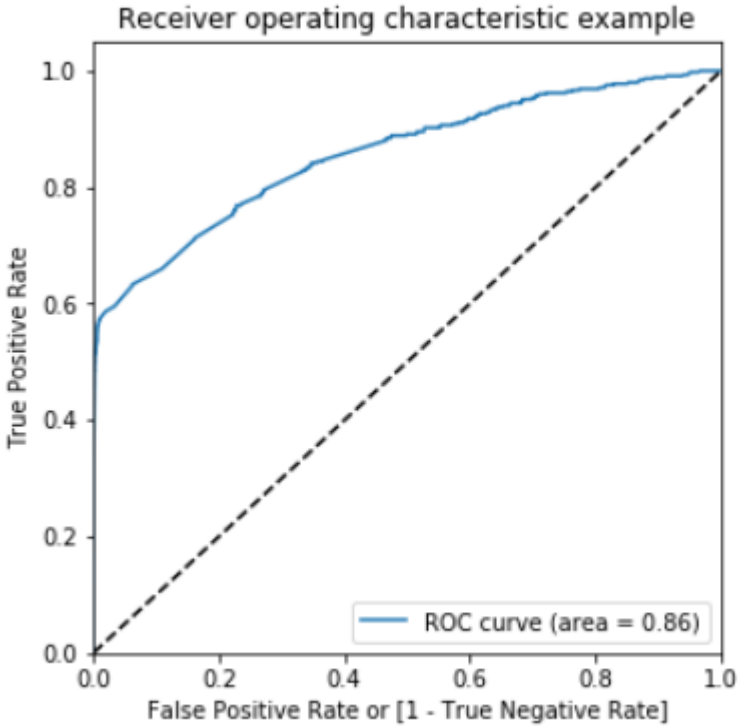
- Calculated VIF values for the above 15 feature and found that some of the feature are having very high VIF values
- Dropped the features one by to get low VIF values (VIF < 5)
- Left with 13 feature varaibles

	Features	VIF
5	m10	4.96
4	m9	4.51
6	m12	3.45
10	loan_purpose_A23	1.44
3	m4	1.29
12	loan_purpose_C86	1.26
11	loan_purpose_B12	1.16
7	source_Y	1.08
8	source_Z	1.08
2	m1	1.04
0	number_of_borrowers	1.02
1	insurance_type	1.01
9	financial_institution_Taylor, Hunt and Rodriguez	1.01

Model Building and Evaluation:

- Built the model with the above selected feature and calculated confusion matrix

ROC Curve:



Optimal Cutoff:

- As the dataset is unbalanced, we cannot rely on accuracy in deciding the optimal threshold. We need to achieve right balance between accuracy, sensitivity and specificity.

	prob	accuracy	sensi	speci
0.0	0.0	0.005539	1.000000	0.000000
0.1	0.1	0.995187	0.488889	0.998007
0.2	0.2	0.995581	0.460000	0.998564
0.3	0.3	0.995815	0.420000	0.999022
0.4	0.4	0.995901	0.333333	0.999592
0.5	0.5	0.995938	0.324444	0.999678
0.6	0.6	0.996012	0.315556	0.999802
0.7	0.7	0.995938	0.286667	0.999889
0.8	0.8	0.995901	0.277778	0.999901
0.9	0.9	0.995778	0.255556	0.999901

- So, from the above the optimal threshold is 0.1

F1- Score:

- Using the above model with optimum threshold I have achieved F1-Score as 0.53.

Predictions:

- Predictions on test dataset are done using the above model and the resulting DF has been exported to a CSV file