XYZCorp_LendingData

By:- Saurav Dash

Points to Cover

- Problem Statement
- Pre-Processing (Cleaning)
- Model Building (Basic Model)
- EDA - Exploratory Data Analysis
- Treatment (Skewness and correlation)
- Model Building After Treatment
- Conclusion

Problem Statement

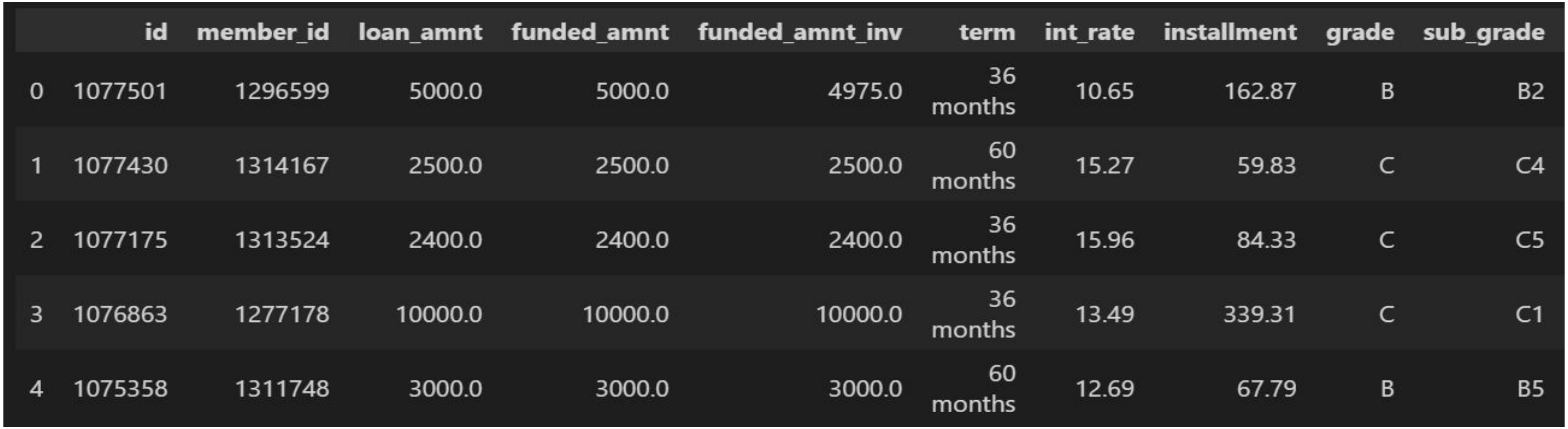
In this project we have to manage credit risk by using the past data and deciding whom to give loan in the future.

Objective : We have to build a data model to predict the probability of default. Alternatively we can also use a modelling technique which gives binary output.

About Dataset

Loan Issuer Dataset broadly available in Kaggle, also known as XYZCorp_LendingData which has **73** columns and **855969** of rows of data.

Problem is to predict the defaulters or non-defaulters present in the loan withdrawing committee. Dataset contains many columns out of which 39 columns are used for predictions using ML in python.



	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade
0	1077501	1296599	5000.0	5000.0	4975.0	36 months	10.65	162.87	B	B2
1	1077430	1314167	2500.0	2500.0	2500.0	60 months	15.27	59.83	C	C4
2	1077175	1313524	2400.0	2400.0	2400.0	36 months	15.96	84.33	C	C5
3	1076863	1277178	10000.0	10000.0	10000.0	36 months	13.49	339.31	C	C1
4	1075358	1311748	3000.0	3000.0	3000.0	60 months	12.69	67.79	B	B5

Cleaning of Data

- Null value Treatment in columns less than 20% of Null Values

  Replaced Null Values with **Mean** and **Median** of Data Samples.

  Total Rows of Data Present = 855969

- Removal of few Rows of Data as those were having Null values.

  Total Rows of Data Present for Computation = 855467

Dropping of Attributes.

- Dropping attributes or columns which are having more than half Null values present in Dataset - used a user defined code to drop the columns.

  Before Dropping - Total No. of Columns = 73

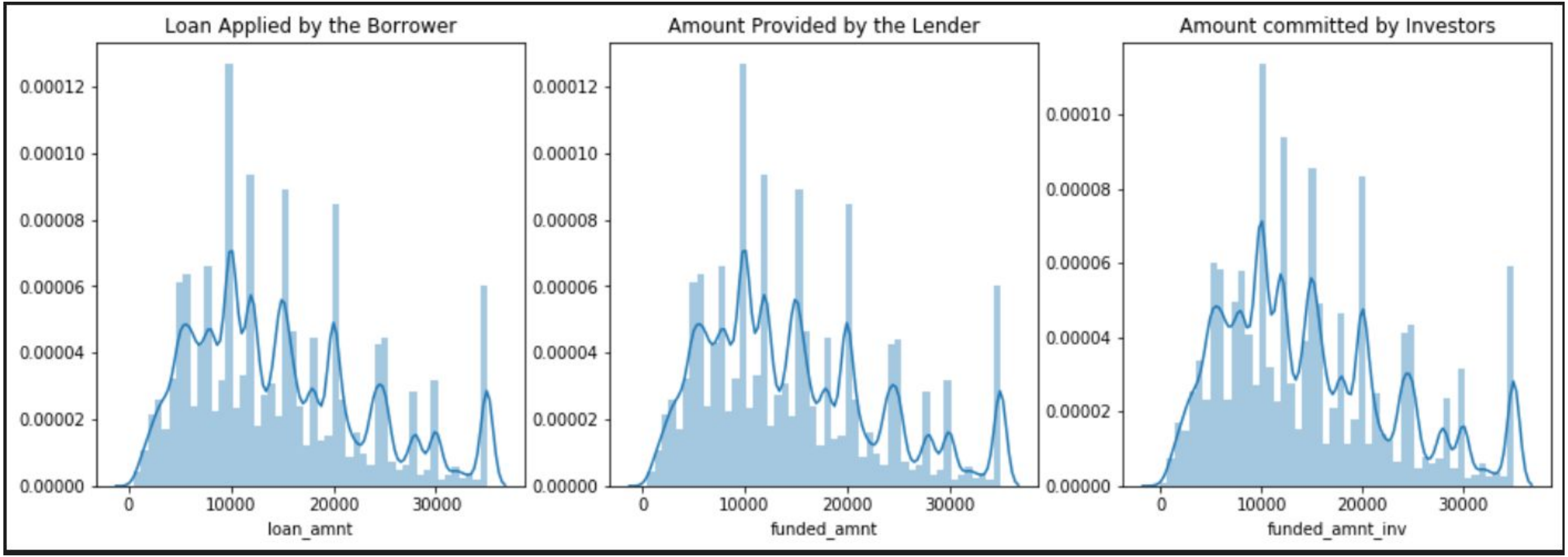
  After Dropping - Total No. of Columns = 52

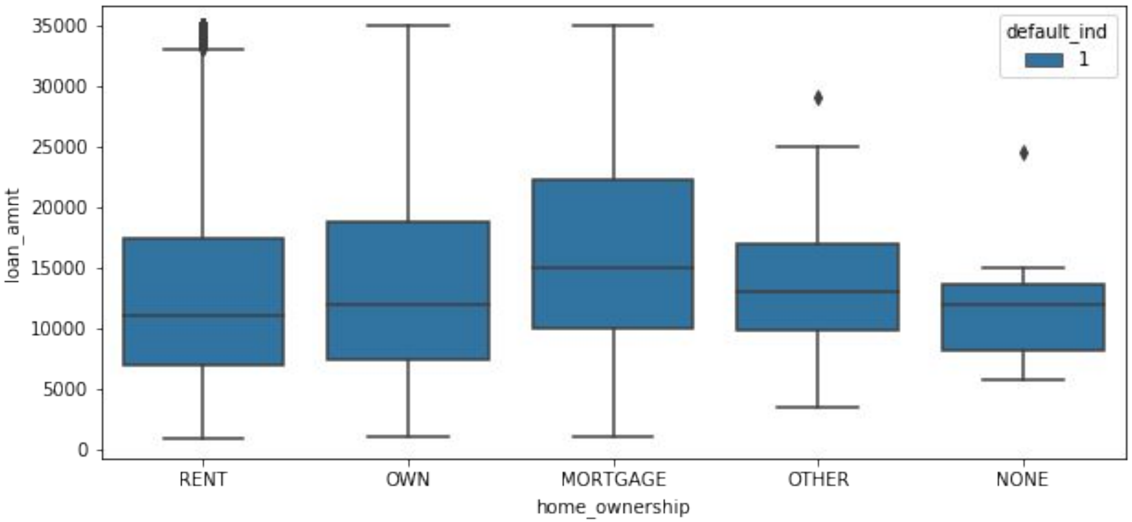
- Dropping few more attributes, which are not required by our model like id, member_id, postal_code,etc

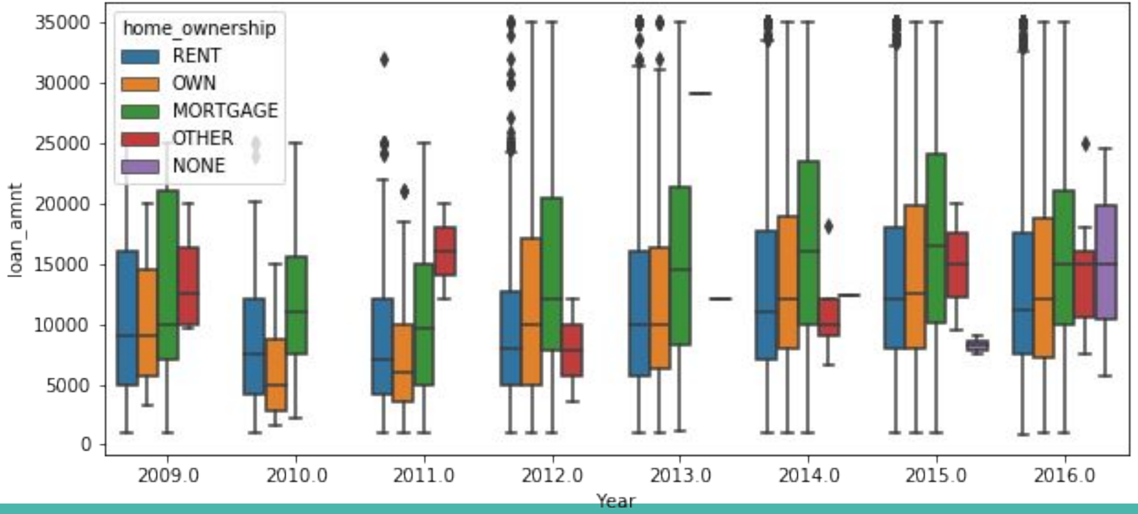
  Final Shape of Columns = 40

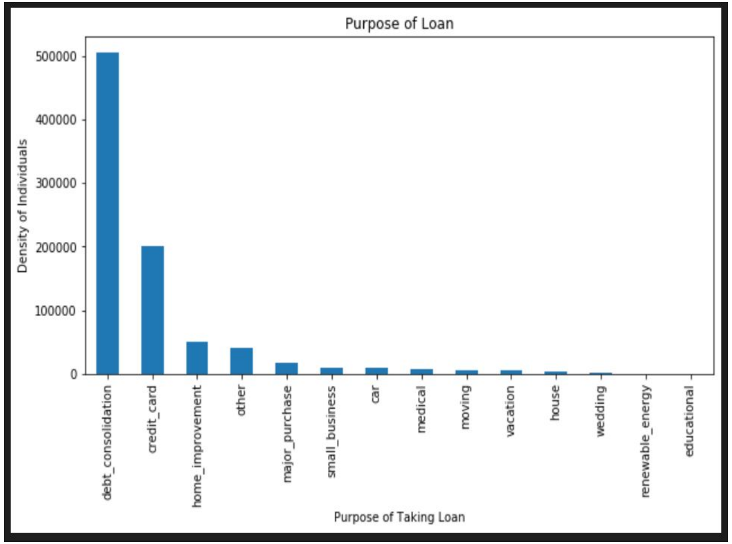
Exploratory Data Analysis

XYZCorp_LendingData



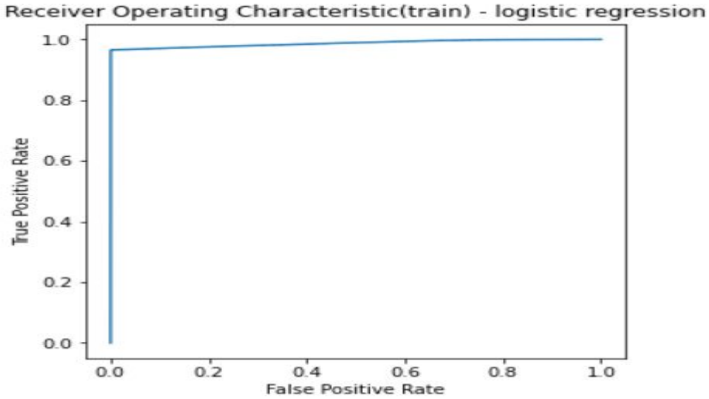

default_ind
1
loan_amnt
35000
30000
25000
20000
15000
10000
5000
0
RENT
OWN
MORTGAGE
OTHER
NONE
home_ownership


home_ownership
RENT
OWN
MORTGAGE
OTHER
NONE
loan_amnt
35000
30000
25000
20000
15000
10000
5000
0
2009.0
2010.0
2011.0
2012.0
2013.0
2014.0
2015.0
2016.0
Year

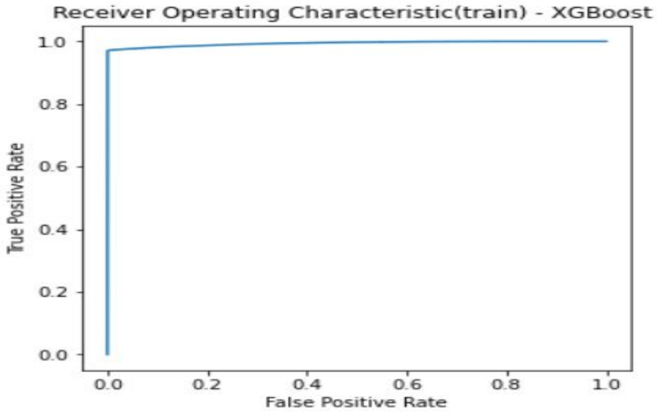

Purpose of Loan
Density of Individuals
500000
400000
300000
200000
100000
0
debt_consolidation
credit_card
home_improvement
other
major_purchase
small_business
car
medical
moving
vacation
house
wedding
renewable_energy
educational
Purpose of Taking Loan

Model Building

ML Model	Accuracy
Logistic Regression	99.75
Random Forest Classifier	99.96
Decision Tree Classifier	100
XGBoost Classifier	99.72
Extra Tree Classifier	100



Receiver Operating Characteristic(train) - logistic regression



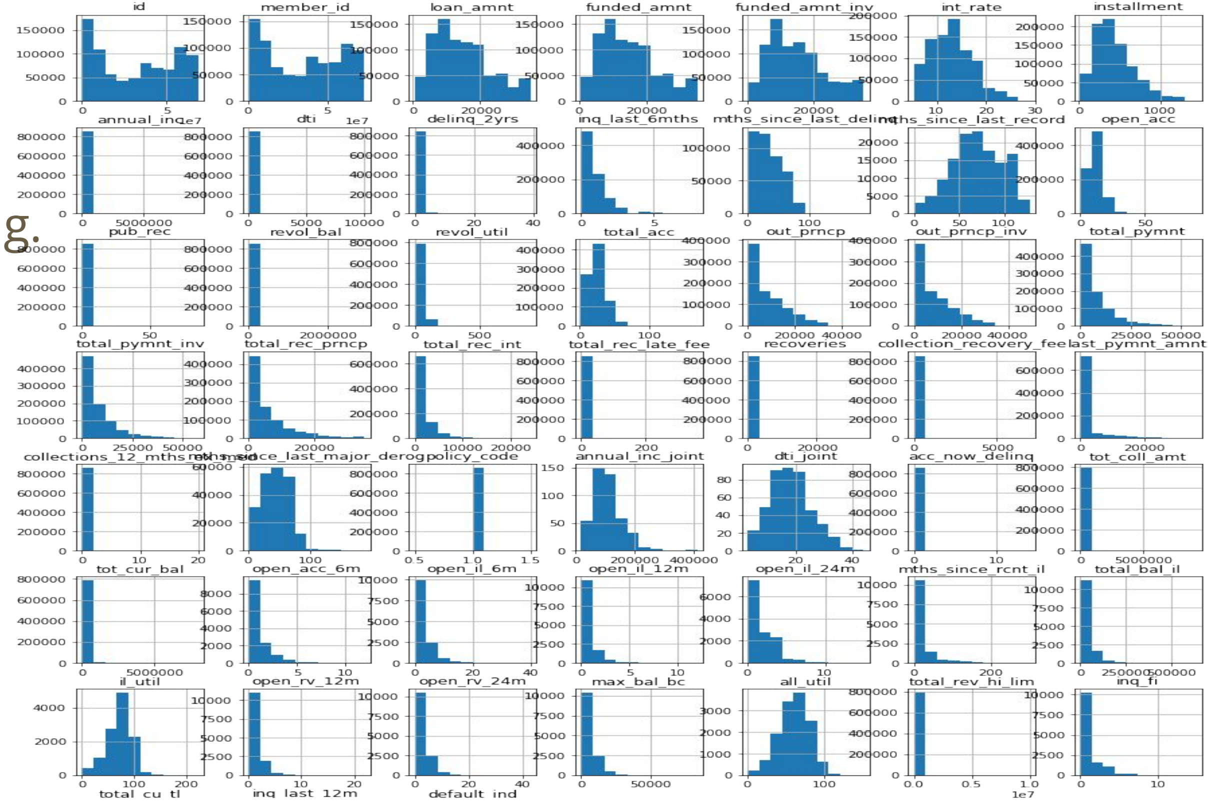
Receiver Operating Characteristic(train) - XGBoost

Treatment (Skewness and Correlation)

Treated Skewness of the attributes present and removed very low correlated values from dataset.

After Treatment

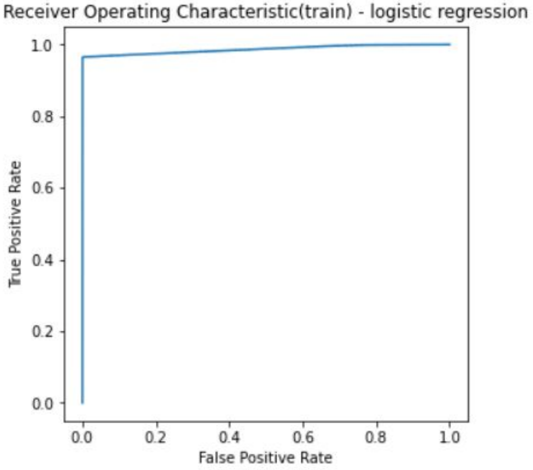
29 columns remaining.



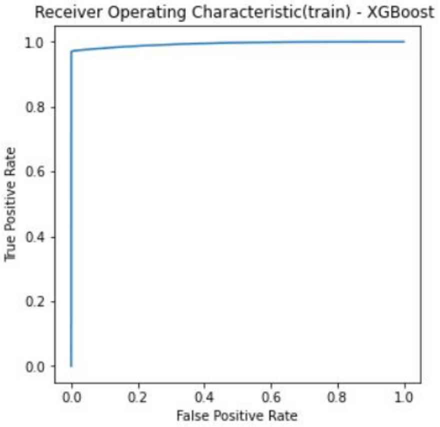
Model Building After Treatment

ML Model	Accuracy
Logistic Regression	98.30
Random Forest Classifier	99.99
XGBoost Classifier	99.94

```
auc_score for Logistic Regression(train):  0.9865922027858597
```



```
auc_score for Xgboost: (train):  0.993612326557153
```



Conclusion

From all above analysis we can conclude that after Treatment of dataset and applying **logistic regression** model gives best result.

Hence logistic model can be used for further predicting.

THANK YOU