GREAT LAKES

INSTITUTE OF MANAGEMENT

Capstone Project–Report

Predicting and Classifying a Loan Borrower as Defaulter or Non-defaulter

Domain – Banking and Finance

Mentored by:

Ms. Akhila Gurre

Submitted by:

Group – 1

Anjali Mahato

Aradhana Coelho

Chitturi Vinay Kumar

Kritika Khanna

Monika Gattu

Neralla Phalgunanand Sharma

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to **MS AKHILA NAGA SAI GURRE**, Data Science Consultant, SYNITI, Bangalore whose insightful leadership and knowledge benefitted to steer out our project successfully. In every phase of this project, her supervision and guidance shaped this report to be completed perfectly. We also thank her for the continuous encouragement and the immense support shown towards us to complete our project work.

We are extremely grateful to all the teaching and non-teaching staff members of **GREAT LEARNING**, who showed keen interest and inquired about our development.
We would also like to thank all teaching and non-teaching staff members of the **Great Lakes Institute of Management**, for providing their support and guidance for our project.

We greatly admire and acknowledge the constant support we received from our friends and team members for all the effort and hard work that they have put into completing this project.

CONTENTS

ABSTRACT

Banks and corporate financial firms are the entities that provide loans to individuals and businesses. The former entities are called lenders and the latter are the borrowers. Borrowers take loans from lenders to satisfy their short-term and long-term financial goals. For individuals, financial goals can be buying a house, higher education, etc. and for businesses, it could be for investing in infrastructure, expanding operations, or increasing working capital. Banks/financial firms make a profit from the interest that is levied upon these loans. In general, these entities carry out analysis to determine if loans are made on feasible terms and if potential borrowers can and are willing to pay back the loan. For this, they use some metrics like current financial standing, previous credit history, and some other variables as input and obtain a measure of safety on the loan that determines the probability that the borrowers will pay back the loan (principal and interest). The main objective of this project is to make predictions on whether an individual/business is eligible for a Loan or not, using machine learning techniques. The motivation for this study is to help bankers make better use of these state of art machine learning algorithms to prevent them from being loan defaulters.

PROBLEM STATEMENT

What would you achieve by this project?

Financial institutions tend to face huge losses whenever the borrowers fail to pay back the loan that has been sanctioned to them. If many such instances pile up, it could lead these lending institutions into bankruptcy and result in the closure of these companies. So, if we would be able to identify defaulters in advance from the larger population of those who are applying for the loan, these institutions can prevent themselves from succumbing to the pitfall and make a profit at the same time.

Here, we use a dataset that contains about 8 lakh records of customers applying for a loan. Each record represents the details of the customer. The dataset comprehends loan data for all loans issued from June 2007 to December 2015.

We aim to build a classification model using the dataset XYZCorp_LendingData and will choose the best model based on different metrics that would help the financial institutions to identify and target genuine borrowers more efficiently as compared to the traditional method

How would this help the business or clients?

The method that we will use in classifying the borrowers as defaulters or non-defaulters will help XYZ corp. to reduce their NPAs and prevent them from making unsecured lendings. It will also enhance the bank's reputation and will lead to high revenue generation.

Through the incorporation of qualitative and quantitative factors with ML, we aim at determining the creditworthiness of clients for our subject, XYZ Corp. While these analyses are not a guarantee against default, our motivation for this project is to sufficiently lower the chances of our subject experiencing high monetary losses.

1. INTRODUCTION

The primary source of income for any bank is on its credit line. Mainly they earn from the interests paid by the loan borrowers. The ultimate task for any bank is to provide their wealth in safer hands. Their profit and loss are necessarily dependent on the borrowers' payback. They approve the loans after verifying and validating the documents provided by the borrowers. Yet, the provided documents don't guarantee the credibility of the borrowers. If the borrowers are defaulting, it results in a direct loss to the bank. Foretelling a borrower's state (defaulter/non-defaulter) in the future is a cumbersome task for any financial institution. Categorizing the borrowers as defaulters or non-defaulters can help the banks to reduce their losses as Non-performing assets.

The objective is to build a machine learning model for classifying and predicting the new applicants of XYZ corp. as defaulters or non-defaulters.

The data is taken from Kaggle named “XYZ Corp. loan lending prediction” to analyze and study the behaviour of the borrowers using different machine learning algorithms. The various machine learning models will be compared based on their performance metrics. The selected model should be able to assess the important attributes of a borrower and help XYZ corp. to target the right customers and minimize the possible losses.

1.1 NEED FOR THE PROJECT

At present, Banks use the manual way of determining the suitability of the borrower for lending money. Before evident economic growth, the manual procedures were quite effective and conducive. Though, with the increase in loan applications, this process has become redundant and is insufficient to allocate loans to the customers. To reduce the waiting time and to make prompt decisions, "loan prediction machine learning models" can be used to assess the customer's status. These models should help to extract the inferences as output, which will accelerate the overall process.

1.2 OBJECTIVE OF THE PROJECT

The main objective of this project is to find the features/variables that affect the borrowers' future status(defaulter/non-defaulter). The goal is to build a machine learning model for classifying and predicting the new applicants of XYZ corp. as defaulters or non-defaulters, thus optimizing the revenue management of the institute.

1.3 SCOPE OF THE PROJECT

From a business point of view, it will be helpful if the lending institutions can identify defaulters in advance from the larger population of those who are applying for the loan. These institutions can then target genuine borrowers. Thus, this project is vital from a revenue optimization standpoint.

1.4 TOOLS USED

- Programming Language: Python
- Visualisation: Python and Tableau

2. DATA DESCRIPTION

2.1 DESCRIPTION

The Dataset consists of 73 attributes and 855969 records, The file contains loan data for all loans issued by XYZ Corporation from June 2007 to December 2015.

The dataset consists of **52 numerical** and **16 categorical** and **5 datetime** attributes.

ATTRIBUTE NAME	DATA TYPE	ATTRIBUTE DESCRIPTION
id	Int	Unique ids for each observation
member_id	Int	Unique ids for each member
loan_amnt	Int	Total amount requested by the borrower
funded_amnt	Int	Total amount committed to that loan at that point of time
funded_amnt_inv	Float	Total amount committed by investors
int_rate	Float	Rate of interest issued on the loan
Installment	Float	The monthly payment owed by the borrower if the loan originates
annual_inc	Float	Self-reported annual income of the borrower
Dti	Float	Ratio of borrower's total monthly debt payments to the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's monthly income
delinq_2yrs	int	Number of 30+ days overdue incidences of delinquency in the borrower's credit file in past 2 years
open_acc	Int	The number of open credit lines in the borrower's credit file
pub_rec	Int	Number of derogatory public records
revol_bal	Int	Total credit revolving balance
revol_util	Float	Revolving line utilization, ie, the amount of credit the borrower is using relative to all available revolving credit
total_acc	Int	Total number of credit lines currently in borrower's credit file
out_prncp	Float	Remaining outstanding principal for total amount funded
out_prncp_inv	Float	Remaining outstanding principal for portion of total

		amount funded by investors
total_pymnt	Float	Payments received to date for total amount funded
total_pymnt_inv	Float	Payments received to date for portion of total amount funded by investors
total_rec_prncp	Float	Principle received to date
total_rec_int	Float	Interest received to date
total_rec_late_fee	Float	Late fees received to date
recoveries	Float	Post charge off gross recovery
collection_recovery_fee	Float	Post charge off collection fee
last_pymnt_amnt	Float	Last total payment amount received
policy_code	Int	Publicly available policy_code=1 New products not publicly available policy_code=2
tot_coll_amt	Float	Total collection amounts ever owned
tot_cur_bal	Float	Total current balance of all accounts
total_rev_hi_lim	Float	Total revolving high credit/credit limit
annual_inc_joint	Float	The combined self-reported annual income provided by the co-borrowers during registration
mths_since_last_delinq	Float	The number of months since the borrower's last delinquency.
mths_since_last_record	Float	The number of months since the last public record.
mths_since_last_major_derog	Float	Months since the most recent 90-day or worse rating
dti_joint	Float	A ratio calculated using the co-borrower’s total monthly payments on the total debt obligations, excluding mortgages and the requested loan, divided by the co-borrower’s combined self-reported monthly income
open_acc_6m	Float	Number of open trades in last 6 months
open_il_6m	Float	Number of currently active installment trades
open_il_12m	Float	Number of installment accounts opened in past 12 months
open_il_24m	Float	Number of installment accounts opened in past 24 months
mths_since_rcnt_il	Float	Months since most recent installment accounts opened
total_bal_il	Float	Total current balance of all installment accounts
il_util	Float	Ratio of total current balance to high credit/credit limit on

		all install acct
open_rv_12m	Float	Number of revolving trades opened in past 12 months
open_rv_24m	Float	Number of revolving trades opened in past 24 months
max_bal_bc	Float	Maximum current balance owed on all revolving accounts
all_util	Float	Retrieving data. Wait a few seconds and try to cut or copy again.
inq_fi	Float	Number of personal finance inquiries
total_cu_tl	Float	Number of finance trades
inq_last_12m	Float	Number of credit inquiries in past 12 months
inq_last_6mths	int	Number of credit inquiries in the past 6 months
collections_12_mths_ex_med	Float	Number of collections in 12 months excluding medical collections
acc_now_delinq	int	Number of accounts on which the borrower is now a delinquent
default_ind (Target)	int	Indicates whether a borrower is a defaulter or non-defaulter. Values are 1 and 0.

Categorical Features in the Dataset:

ATTRIBUTE NAME	DATA TYPE	ATTRIBUTE DESCRIPTION
term	Object	Number of payments of the loan. Values are in months and can be either 36 or 60
grade	Object	Grade assigned to the loan by XYZ_Corp
sub_grade	Object	Sub-grade assigned to the loan by XYZ_Corp
emp_title	Object	Job title of the borrower
emp_length	Object	Work experience of the employee in years. Value ranges from 0 to 10
home_ownership	Object	Housing status of the borrower. Possible values are Rent, Own, Mortgage, Other.
verification_status	Object	Indicates if income source was verified by XYZ_Corp. or not
pymnt_plan	Object	Indicated if a payment plan has been put in place for the loan
purpose	Object	Reason provided by borrower for loan request
title	Object	Loan title mentioned by the borrower

zip_code	Object	First three number of zip code provided by borrower
addr_state	Object	State that the Borrower belongs
initial_list_status	Object	The initial listing status of the loan. Possible values are – W, F
application_type	Object	Indicates whether the loan is an individual application or a joint application with two co-borrowers
desc	Object	Loan description provided by the borrower
verification_status_joint	Object	Indicates if the co-borrower's joint income was verified by XYZ Corp., not verified, or if the income source was verified

<u>Date-time Features in the Dataset</u>:

ATTRIBUTE NAME	DATA TYPE	ATTRIBUTE DESCRIPTION
issue_d	Datetime	The month which the loan was funded
earliest_cr_line	Datetime	Month the borrower's earliest reported credit line was opened
last_pymnt_d	Datetime	Last month payment was received
next_pymnt_d	Datetime	Next scheduled payment date
last_credit_pull_d	Datetime	The most recent month XYZ corp. pulled credit for this loan

Variable Categorization:

No. of Categorical Variables	52
No. of Numerical Variables	16
No. of DateTime Variables	5

2.2 SOURCE

The Dataset is taken from KAGGLE, the file downloaded was in CSV format.

Dataset name: XYZ Corp. Lending Data Prediction

3. LITERATURE SURVEY

All financing institutes know and accept that the lending business is full of risks; thus, it is crucial to carry out excessive credit analysis before sanctioning any loan. For years, machine learning has been used to create credit risk management models, but in recent years its usage has been growing rapidly. These days, most financial institutions use data-driven analysis with 5C discrimination factors to analyze credit risk which can be basically called - "lender's role, capital, collateral, capacity, and environment". Random Tree forests, classified regression, discriminant analysis are mostly used in the data analysis. However, currently, regression (logistic) is a highly preferred technique to prioritize the best features for building better models.

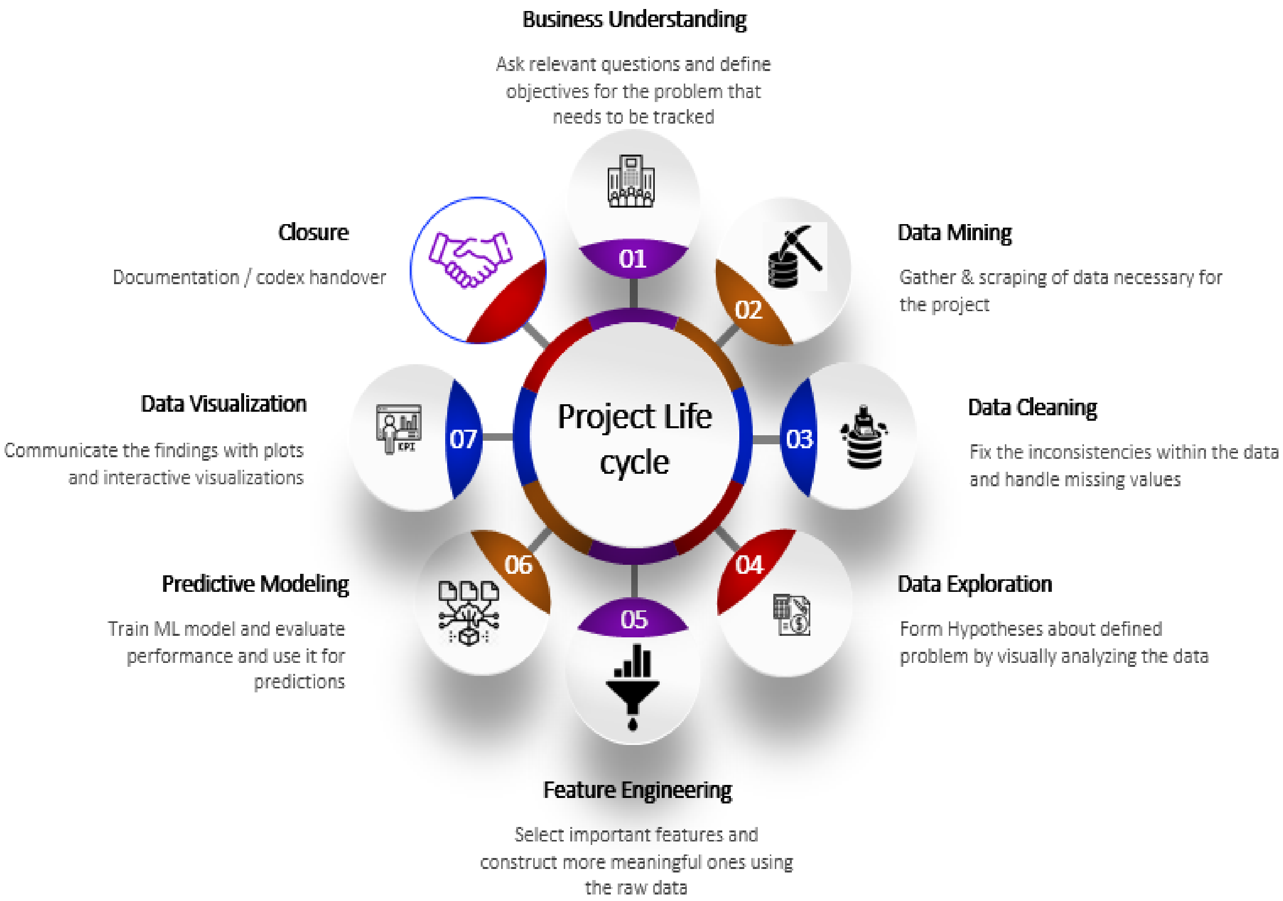
The credit analysis should also include the external factors on which the company does not have control, but which can have a strong impact on the regular repayment of the loan. In fact, this is about analyzing the business environment in which the borrower works.

The following should be assessed:

- The trends in the industry in which the borrower works,
- The technological trends in the industry,
- The position of the borrower on the market,
- The stability of his relations with suppliers and buyers,
- The business cycle phase, the future movement of interest rates.

Through the incorporation of qualitative and quantitative factors with ML, we aim at determining the creditworthiness of clients for our subject, XYZ Corp. While these analyses are not a guarantee against default, our motivation for this project is to sufficiently lower the chances of our subject experiencing high monetary losses.

4. PROJECT METHODOLOGY


Business Understanding
Ask relevant questions and define objectives for the problem that needs to be tracked
01
Data Mining
Gather & scraping of data necessary for the project
02
Data Cleaning
Fix the inconsistencies within the data and handle missing values
03
Data Exploration
Form Hypotheses about defined problem by visually analyzing the data
04
Feature Engineering
Select important features and construct more meaningful ones using the raw data
05
Predictive Modeling
Train ML model and evaluate performance and use it for predictions
06
Data Visualization
Communicate the findings with plots and interactive visualizations
07
Closure
Documentation / codex handover
Project Life cycle

5. DATA PRE-PROCESSING

Data cleaning is a major part of this project. The RAW data cannot be directly given to the Machine learning algorithms. We need to pre-process it and design data in a better way so that the Machine learning algorithm can better understand the data. The dataset contains 73 attributes and 8855969 records.

5.1. REMOVING NULL COLUMNS

Columns that have null values greater than 50% of the total count are dropped, after which, attributes are reduced to 52 from 73. The remaining variables that are left are shown below.

Numerical:

['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', 'installment', 'annual_inc', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'mths_since_last_delinq', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_amnt', 'collections_12_mths_ex_med', 'policy_code', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim', 'default_ind']

Categorical:

['term', 'grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'verification_status', 'pymnt_plan', 'purpose', 'title', 'zip_code', 'addr_state', 'initial_list_status', 'application_type']

DateTime:

['issue_d', 'earliest_cr_line', 'last_pymnt_d', 'next_pymnt_d', 'last_credit_pull_d']

5.2. Removing Insignificant Variables

Columns such as **id,member_id** contain the id of the loan and borrower respectively, which is redundant for further analysis. Thus, we drop these columns from the dataset.

a) Checking unique values present in the columns of the dataset, we find that the column **policy_code** has only one unique value, so we will remove it from the dataset.

b) Checking value counts present in the categorical columns of the dataset, it is observed that the frequency distribution of **pymnt_plan** as well as **application_type columns** is mostly concentrated around one category, so therefore we remove these columns from the dataset.

c) Removing columns based on Domain Understanding.-
We remove categorical columns named **emp_title, emp_length, purpose, title, zip_code, addr_state** and date columns such as **issue_d, earliest_cr_line, last_pymnt_d, next_pymnt_d, last_credit_pull_d** from the dataset, as they are not going to help in our model building.

d) Now let us remove numerical columns based on high correlation.

Let us remove columns that have a correlation greater than 0.9. Columns such as **loan_amnt,funded_amnt_inv,installment** are having high correlation with **funded_amnt**,so we drop these columns and retain only funded_amnt columns. Similarly columns such as **total_pymnt,total_pymnt_inv** are highly correlated with **total_rec_prncp**,so we drop these columns and retain only total_rec_prncp columns. And **out_prncp_inv** is highly correlated with **out_prncp**, therefore we drop this column and retain only the out_prncp column.

e) Checking columns and number of columns after removing all insignificant variables based on different conditions that are discussed above.

Columns left after removing insignificant variables are :

['funded_amnt', 'term', 'int_rate', 'grade', 'sub_grade', 'home_ownership', 'annual_inc', 'verification_status', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'initial_list_status', 'out_prncp', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_amnt', 'collections_12_mths_ex_med', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim', 'default_ind']

The number of columns left after removing insignificant variables is: 30

5.3. Outlier Analysis and Treatment

We need to check the distribution of important columns based on domain understanding, and then remove outliers from each of these columns sequentially without affecting the percentage of the count of categories present in the **default_ind** column.

Columns which we are considering are **funded_amnt,annual_inc,revol_util,dti, int_rate, total_rec_prncp,total_rec_int,open_acc, total_acc.**

Distributions of the above columns before removing outliers are shown below:

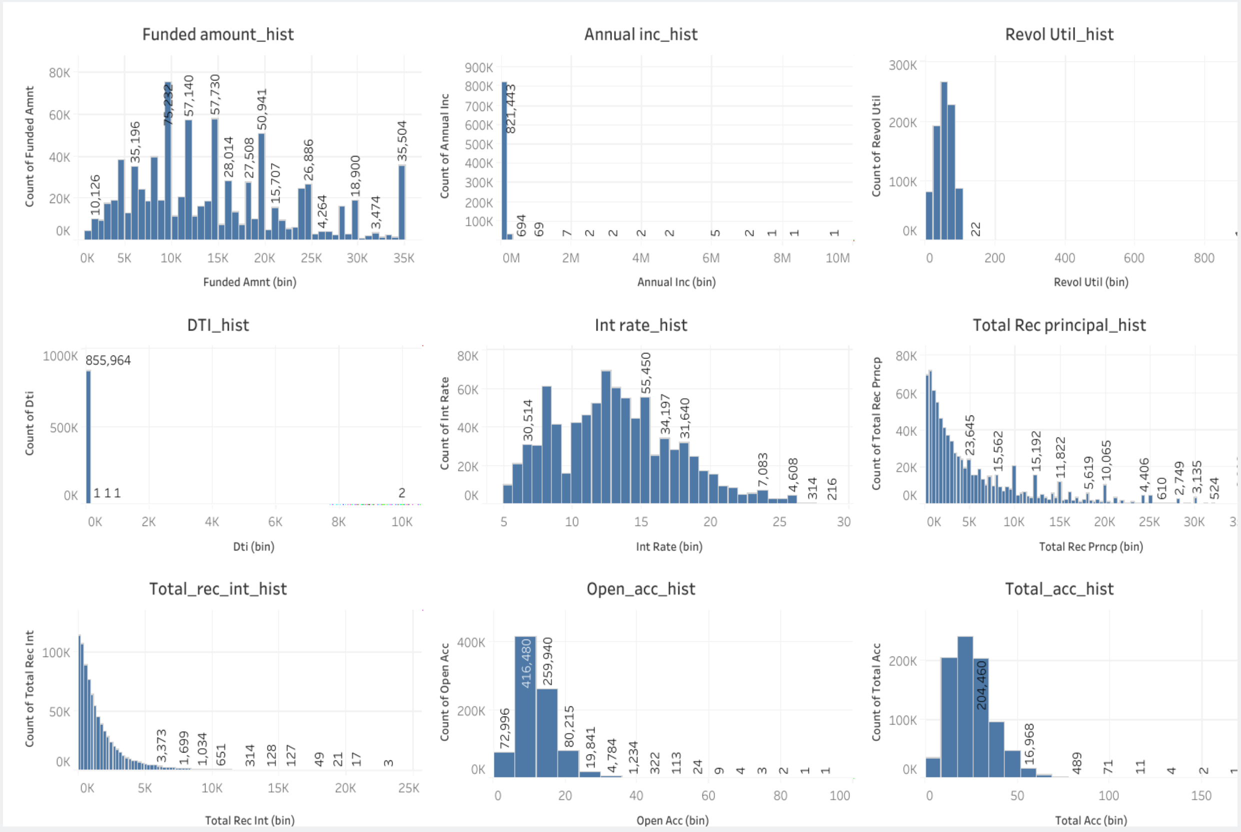
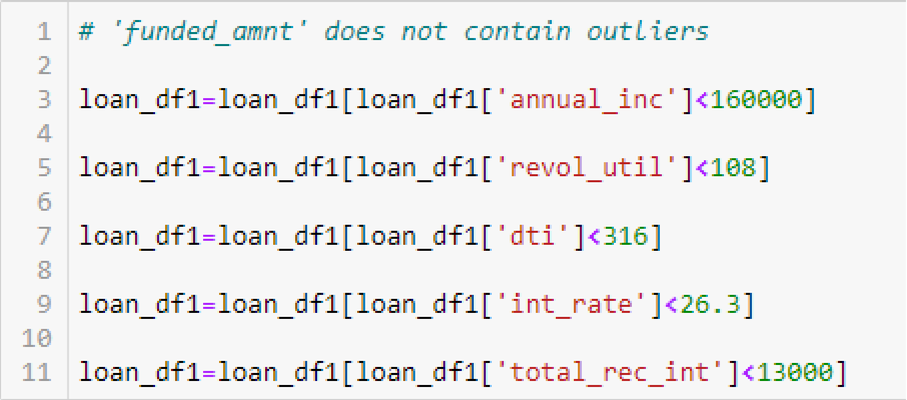


Figure 1. Histogram plots of numerical attributes

Based on the distribution of columns shown above and understanding the range of values columns should have, based on this we remove outliers sequentially as shown below.

Code snippet for removing outliers:



```
# 'funded_amnt' does not contain outliers

loan_df1=loan_df1[loan_df1['annual_inc']<160000]

loan_df1=loan_df1[loan_df1['revol_util']<108]

loan_df1=loan_df1[loan_df1['dti']<316]

loan_df1=loan_df1[loan_df1['int_rate']<26.3]

loan_df1=loan_df1[loan_df1['total_rec_int']<13000]
```

Instead of adopting an ad-hoc method to remove outliers from the data frame, we used understanding of the domain to remove outliers, at the same time, the percentage of the count of categories present in the target column i.e default_ind did not change.

5.4. Missing Value Analysis and Treatment

Missing data in the training data set can reduce the power/fit of a model or can lead to a biased model because we have not analyzed the behaviour and relationship with other variables correctly. It can lead to wrong predictions or classification. Thus, it is important to either remove the records containing missing values or impute them with estimated ones.

Only four columns are seen to have null values. Those are **collections_12_mths_ex_med, tot_coll_amt,tot_cur_bal, total_rev_hi_lim**.

a) We will drop rows for the variables having less than 5% missing values. So, we drop rows with null values in collections_12_mths_ex_med.

b) For the variables having greater than 5% null values, we impute the mean value if the distribution is normal or the median if the distribution is not normal.

Checking the skewness of the columns to determine normality, we saw that the distributions are not normal and hence we will impute the missing values in each column with median values.

By carrying out the above steps, we have ensured that our dataset is free from null values.

5.5. Duplicate Values

Checking for any duplicate values in the dataset.
We see that the dataset has no duplicate records in it.

5.6. Encoding of Categorical Variables

Most ML algorithms work best with numerical inputs. In order to convert categorical columns into numeric forms, we make use of encoding techniques.

a) First, we segregate all categorical columns:

b) Carrying out (n-1) Dummy Encoding on 'term', 'home_ownership', 'verification_status' and 'initial_list_status' columns:

c) Carrying out Ordinal Encoding on 'grade' and 'sub_grade' columns:

5.7. Multicollinearity Testing

Multicollinearity makes it hard to interpret the coefficients, and it reduces the power of the model to identify independent variables that are statistically significant.

Thus, it's important to identify which variables are affected by multicollinearity and the strength of the correlation. This is carried out using a very simple test to assess multicollinearity in the regression model. The variance inflation factor (VIF) identifies a correlation between independent variables and the strength of that correlation.

a) To calculate the VIF value, we concatenate numeric columns from the dataset to the encoded categorical columns.

b) We then calculate VIF values for each numerical column.

The output shows that the variable 'funded_amnt' has the highest VIF. Now, we use the for-loop to find VIF and remove the variables with VIF greater than 10. We set the threshold to 10, as we wish to remove the variable for which the remaining variables explain more than 90% of the variation. One can choose a threshold other than 10. (It depends on the business requirements).

Now, we have all the variables with a VIF of less than 10. So, we can conclude that there is no multicollinearity in the data.

6. EXPLORATORY DATA ANALYSIS

As this part is the initial inference on the data, a higher no. of columns can be observed inclusive of their visualizations.

Target Variable: default_ind

In this section, univariate and bivariate analysis between default_ind, which is a target variable, and all the independent variables in the xyzcorp_lending dataset is performed.

There are 14 categorical columns, 33 numerical columns as shown below.

Categorical columns:

['term','grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'verification_status', 'pymnt_plan', 'purpose', 'title', 'zip_code', 'addr_state', 'initial_list_status', 'application_type']

Numerical columns:

['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', 'installment', 'annual_inc', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'mths_since_last_delinq', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_amnt', 'collections_12_mths_ex_med', 'policy_code', 'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim']

6.1 Univariate Analysis of Target variable

Default_ind: Initially, the classes along with their count were checked in the target variable as follows.

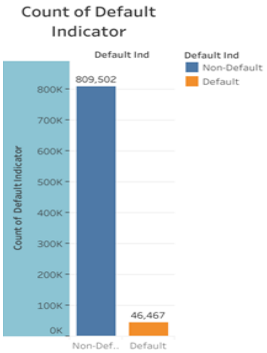


Figure 2. Univariate Analysis of Target Variable

As per the graph, it can be inferred that the variable is highly imbalanced with one class being 5.75% of the other (i.e. Higher observations in one class than the other)

<u>Note</u>:

1. If this is not handled properly, the disparity might cause the algorithm to categorize our predictions/analysis more into one specific class while showing the model in a false tone as being highly accurate i.e. a high bias in the output where the algorithm doesn't learn what makes the other class and its pattern unique for distinguishing. Hence an overfitting issue will arise for the dataset which hampers the main objective.

2. Metrics like Sensitivity and Specificity, F1-score, AUC, Kappa can further help in handling the dataset better.

6.2 Bivariate Analysis

1. <u>Default Indicator vs Employment Length</u>

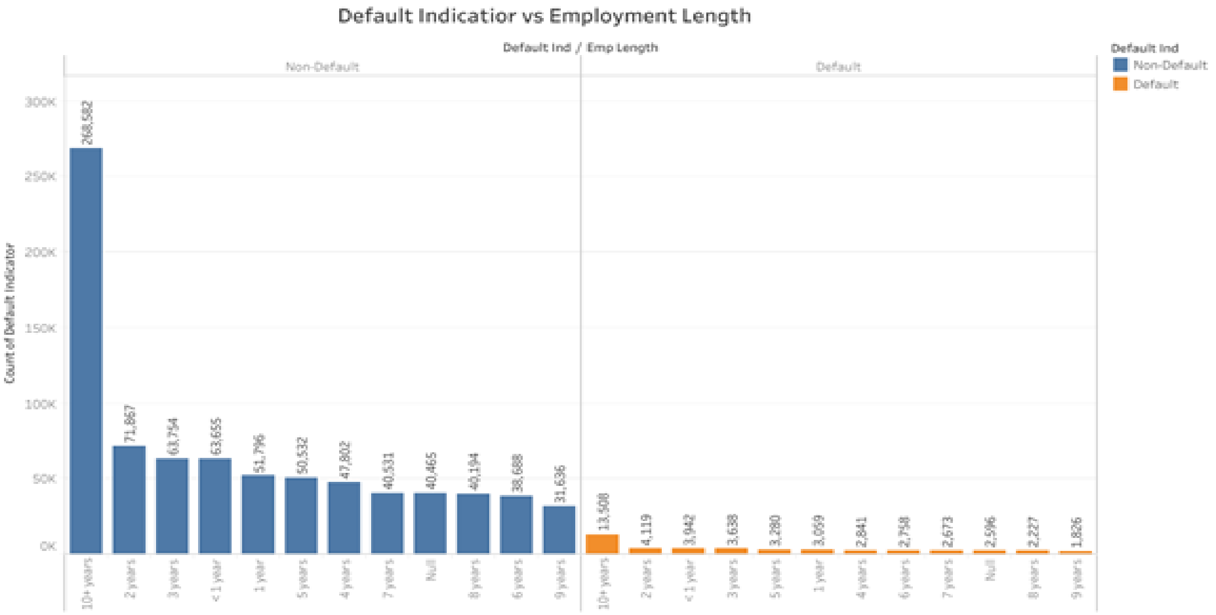


Figure 3. Bivariate analysis between Default Indicator and Employee Length

a. From the employment_length column, we can see that almost 62% of the applicants are employed for less than 10 years, while around 33% have been employed for more than 10 years. The count is fairly distributed among the time period for the first 10 years, but we don't have any idea of how it is distributed beyond 10 years. About 5% of the info is missing.

b. For applications with work experience of more than 10 years, data collection should be stringent and should be properly inherited into the engine to avoid misclassification.

c. For the missing data - Assigning An Unique Category can handle as it can give low variance after one hot encoding — since it is categorical and minimizes the loss of data by adding a unique category.

2. Default Indicator vs Home Ownership

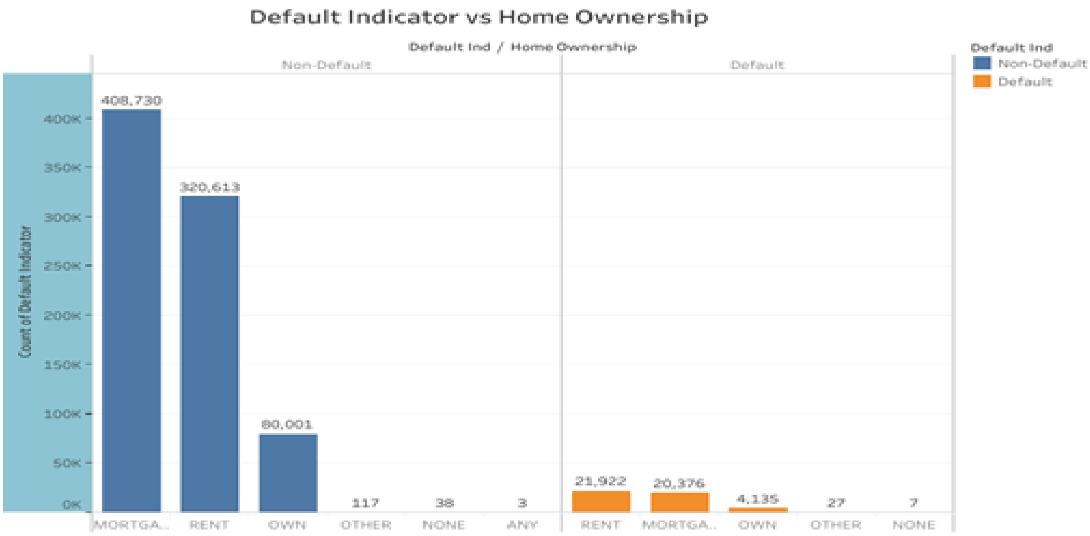


Figure 4. Bivariate analysis between Default Indicator and Home Ownership

a. From the ownership column, we can see that the majority of the applicants (approximately 50%and 40% respectively) have either taken out a mortgage or lived in rented houses. Only around 10% have their own homes and less than 0.1% are miscellaneous.

b. From the above count plot, we can see that a greater number of defaulters and non-defaulters are having homeownership status as ‘Mortgage’ and ‘Rent’.

3. Default indicator vs Loan Grade

Before analyzing loan grade against default indicator let us check how loan grade is associated with an interest rate

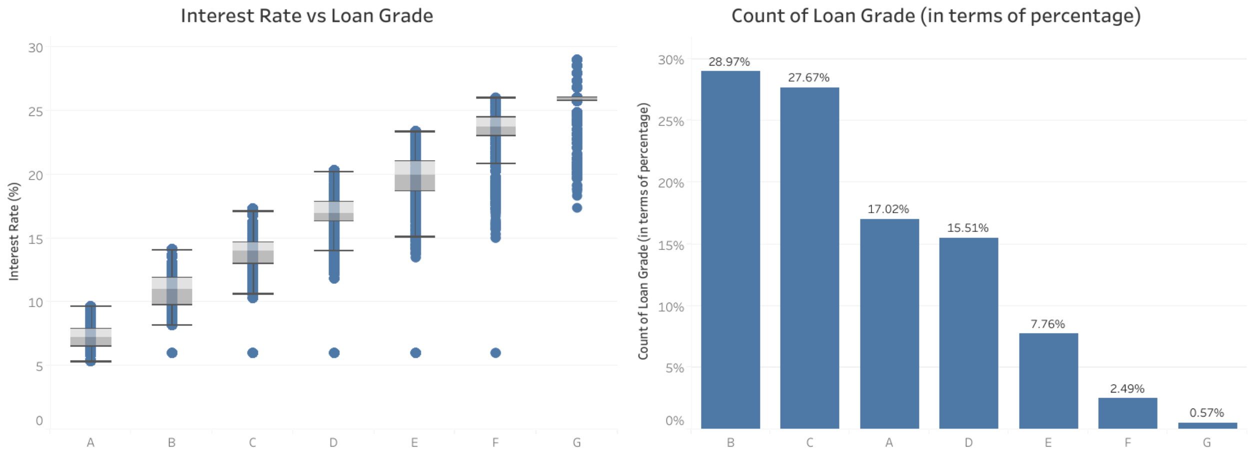


Figure 5. Association of Loan grade with Interest Rate

a. From the box plot, we can see that, as we move from loan grade A to G, the interest rate of the loan increases. This tells us that the lender will have to face high risk and at the same time high returns when he provides a loan in grade G compared to F and so on.

b. Count plot which is expressed in terms of percentage depicts that almost 74 per cent of loans that are provided to borrowers are safer loans as far as the lender is considered, these are considered as safer loans because they are having low-interest rates.

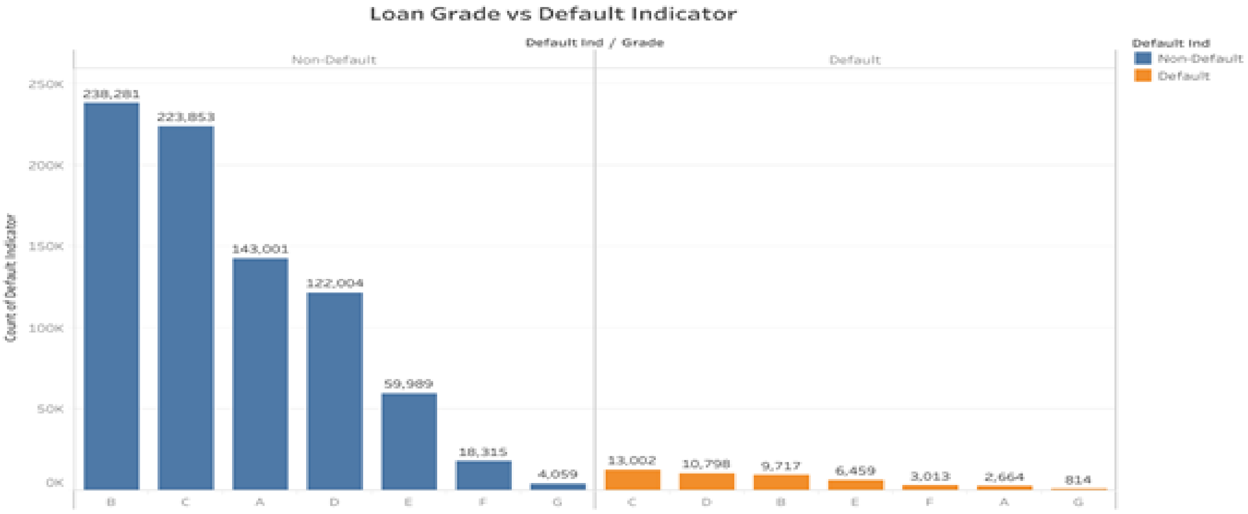


Figure 6. Bivariate Analysis between Loan Grade and Default Indicator

c. As we can see from the above plot, a greater number of loans that are provided to defaulters are having higher interest rates compared to non-defaulters, which signifies loans with higher interest rates are riskier loans for lenders.

4. Default Indicator vs Term

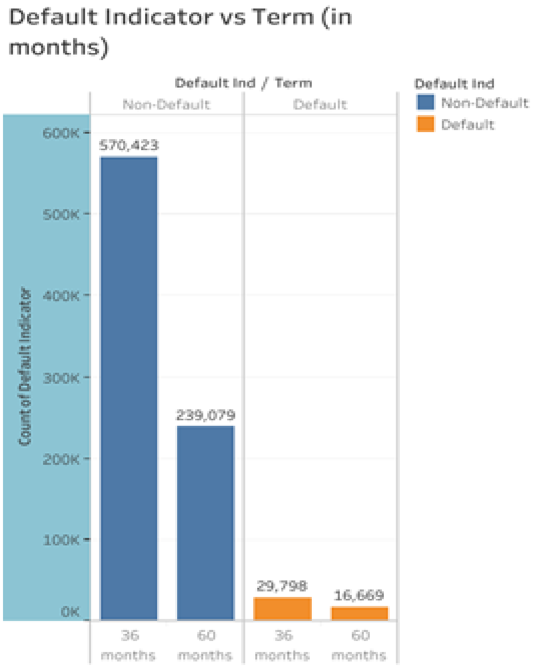


Figure 7. Bivariate Analysis between Term and Default Indicator

a. From the Term column, we can see that about 70% of the applicants have applied for a 3-year term while the remaining 30% have applied for a 5-year term.

b. For a pitch, it can be portrayed as higher returns in lesser time for the investors which in turn grabs their attention as few might be new to the market and can hesitate to incline/invest

c. Secondly, extra benefits need to be designed & marketed as the borrowers generally prefer to pay small amounts over a period of time for borrowers which reduces the risk of default when compared to paying more in a lesser time period.

d. However, a higher default/non-default ratio is observed in the 5-year term compared to the 3-year term.

5. <u>Inquiries (Last 6 months (excluding auto and mortgage inquiries))</u>

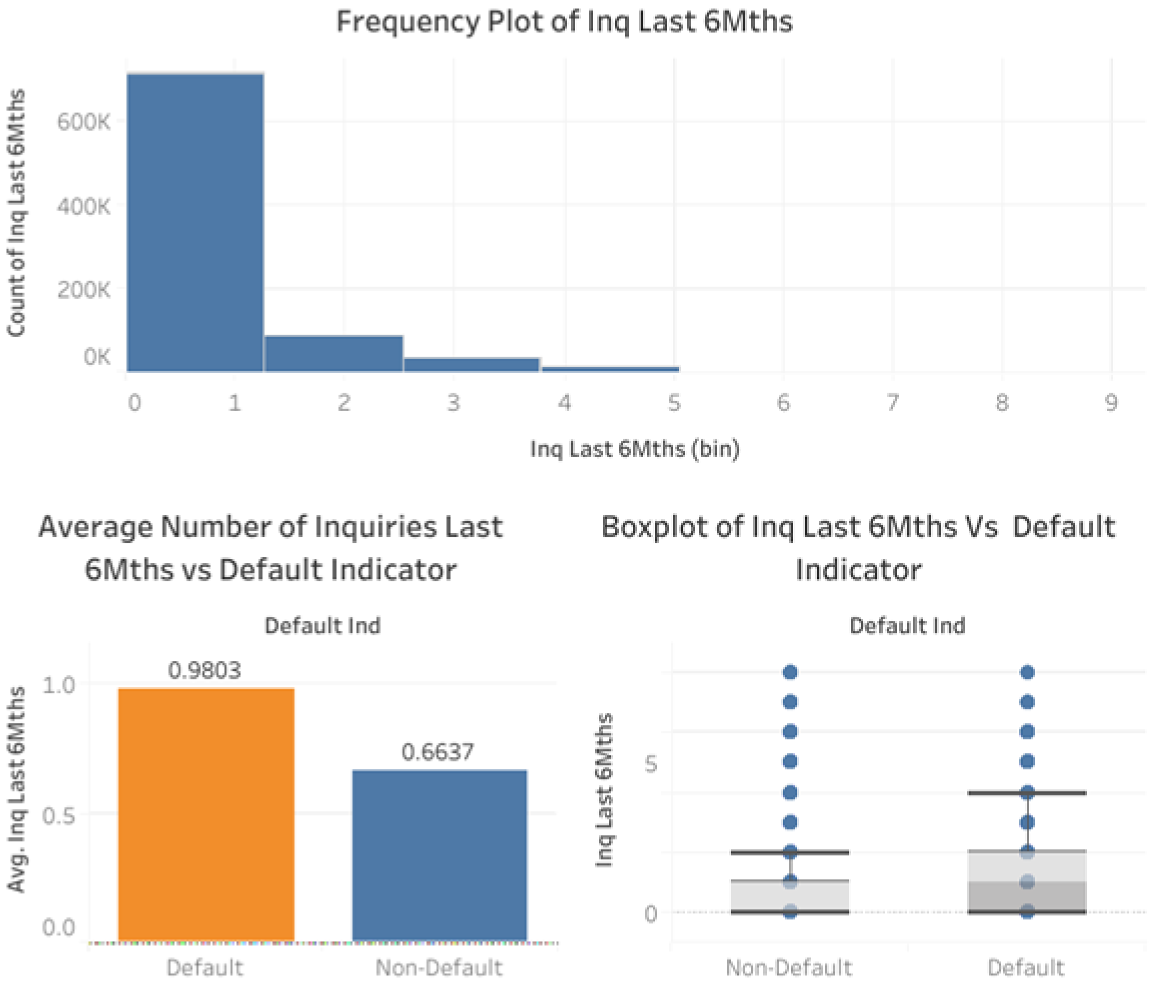


Figure 8. Multiple plots of Inquiries made from the past 6 months

a. Whenever one applies for a credit card, a car loan, a home loan or almost any other kind of loan, one’s credit report gets marked with an inquiry. Any loan application where they do credit report checks is called a hard pull of your credit, it will stay on your credit report for six months.

b. From the frequency plot, we can see that distribution is not normal and a greater number of individuals are having less than 4 inquiries in the past 6 months.

c. Boxplot depicts that the range of values of inquiries for defaulters is more than non-defaulters, and there are a greater number of outliers in the case of non-defaulters compared to defaulters.

d. Bar plot indicates that the average no of inquiries in the past 6 months for defaulters is more than non-defaulters, which makes sense because if someone is out shopping for credit in several places and then they come to Lending looking for a loan then they are a higher risk borrower. They may have some serious financial problems if they are shopping for a lot of credit. But if someone is looking for a loan and comes to p2p lending first then they are a better credit risk

6. Delinquency (Past 2 years)

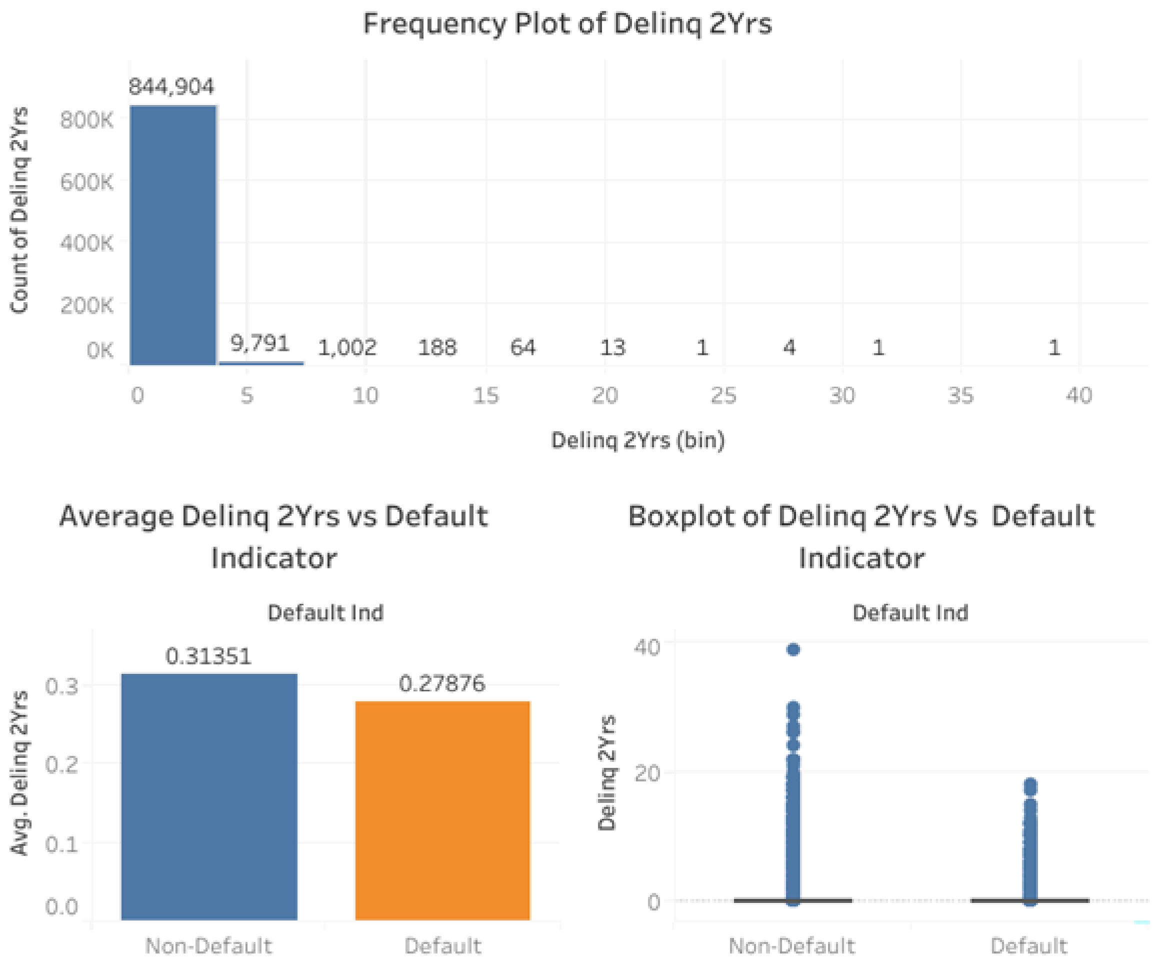


Figure 9. Multiple plots of Delinquency from past 2 years

A delinquent account is the one with pending outstanding payment post due date. “In the context of credit cards, delinquent accounts are **those that have not made at least a minimum payment for 30 days or more.”**

a. What this column depicts is the number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years, delinquency occurs when a borrower falls behind on making required monthly payments. While being 30 days late is generally considered delinquent.

b. From the frequency plot we can see that distribution is not normal and a greater number of individuals are having delinquency less than 5 in the past 2 years.

c. Boxplot depicts that the range of values of delinquency for defaulters and non-defaulters remains almost the same and there are a greater number of outliers in case of non-defaulters compared to defaulters.

d. Bar plot indicates that average delinquency for non-defaulters and defaulters remains almost the same.

7. Public Records (Number of derogatory public records)

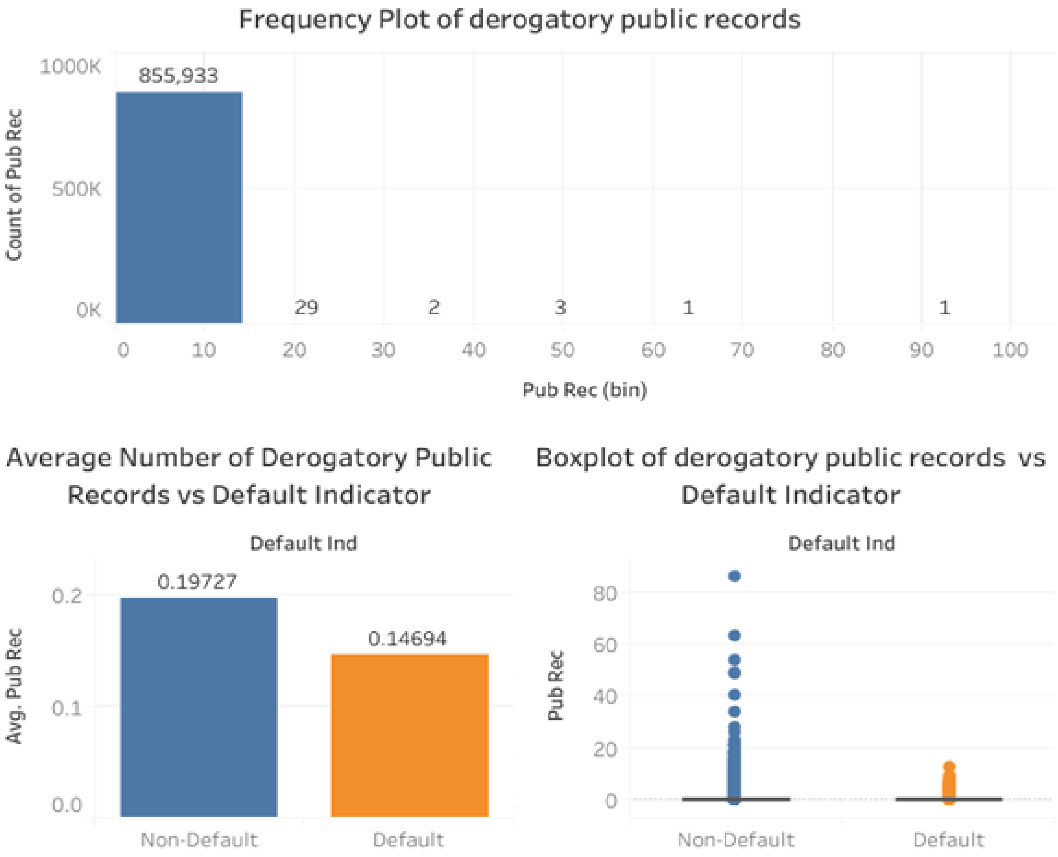


Figure 10. Multiple Plots of Derogatory Public Records

a. From the frequency plot we can see that distribution is mostly concentrated, where individuals are having less than 15 derogatory public records.

b. Boxplot depicts that the range of values of public records for defaulters and non-defaulters remains almost the same and there are a greater number of outliers in case of non-defaulters compared to defaulters.

c. The bar plot indicates that the average number of derogatory public records for non-defaulters is slightly more than defaulters, which doesn’t make sense, but since this dataset is highly imbalanced, there is a chance that this can happen.

Note:
It is also a reg flag for lenders as these are the records where the amount was not paid as promised. Most of them stay on your credit reports for about seven years. and few lenders may view a paid derogatory more favourably than an unpaid one.

Overall it can be by -**“Late payments, Loan and credit defaults, Debts sent to collections, Foreclosures or repossessions, Bankruptcies (past 7 -10 years declaration)”** –

8. <u>Revolving Balance (Total credit revolving balance)</u>

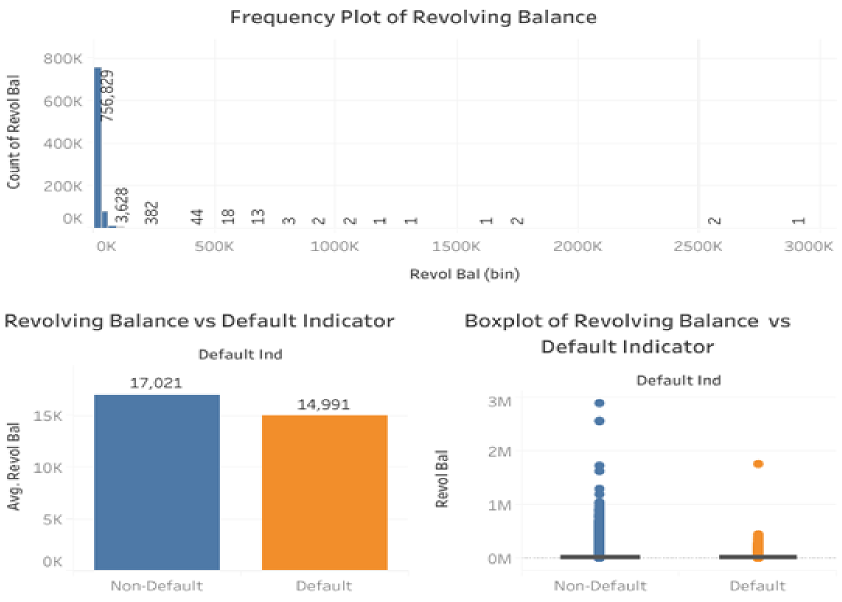


Figure 11. Multiple Plots of Revolving Balance

a. Revolving credit allows a consumer to make purchases up to a certain spending limit and pay down the debt each month. As long as the spending cap has not been reached, the consumer can make purchases using the line of credit. He does not have to pay off the total amount borrowed every month, but any balance that carries over month to month is the revolving balance.

b. From the frequency plot, we can see that distribution is mostly concentrated, where individuals are having a revolving balance of less than 100k

c. Boxplot depicts that the range of values of revolving balance for defaulters and non-defaulters remains almost the same and there are a greater number of outliers in the case of non-defaulters compared to defaulters.

d. The bar plot indicates that the average number of revolving balances for non-defaulters is more than defaulters.

Note:

In general, revolving is usually for credit cards or home equity lines while non-revolving are for car loans or mortgages.

"When a borrower makes a purchase, it increases their outstanding balance and decreases their available balance. When a borrower makes a payment, it decreases their outstanding balance and increases their available balance. Thus, a borrower's balance and available credit will vary each month."

Remark:

Here the majority of the loans are for Mortgages which is a non-revolving line

9. Revolving Utilization Rate

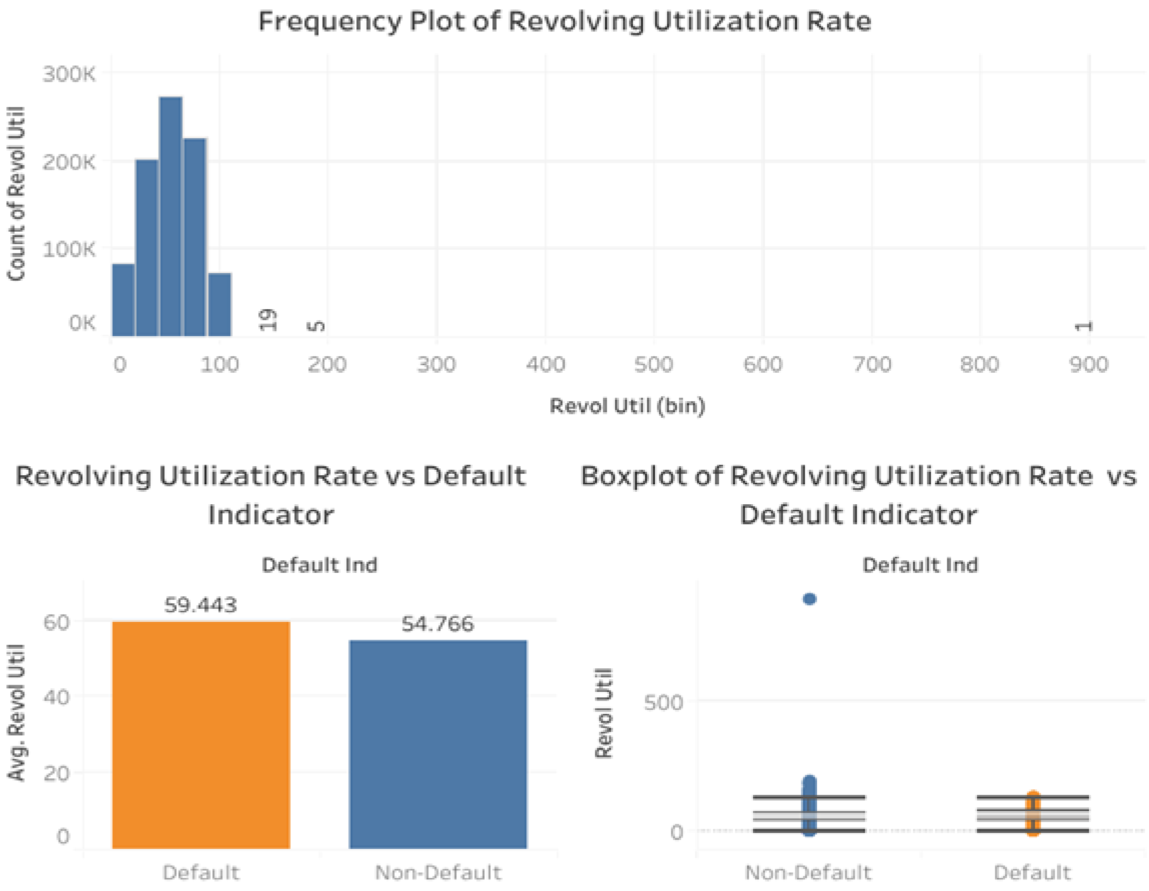


Figure 12. Multiple Plots of Revolving Utilization Rate

a. The credit utilization ratio is the percentage of a borrower's total available credit that is currently being utilized. Lesser the percentage of utilization rate will improve individual cibil score, financial experts recommend having utilization rate below 30%.

b. From the frequency plot, we can see that about 90% of that data has less than 70% utilization rate and a greater number of individuals are having a utilization rate of about 40%.

c. Boxplot depicts that the range of values of revolving utilization rate for defaulters and non-defaulters remains almost the same and there are a smaller number of outliers in the case of non-defaulters and no outliers in the case of defaulters.

d. The bar plot indicates that the average revolving utilization rate for non-defaulters is less than defaulters, which makes sense because individuals who are using a higher percentage of revolving credit amount available to them, are more likely to default.

The utilization ratio is how much one currently owes divided by one's credit limit. In general FICO score says a good total credit utilization rate is below 30%.

A higher rate can flag to lenders that the borrower is having trouble managing your finances.20

6.3 Multivariate Analysis

Here from the below plot, we can see that the median - payment received to date for a portion of the total amount funded by investors, where the return is higher in 5-year term loan for genuine borrowers in the "None" homeownership category. Also, this can be pitched for new creditors as higher returns loan vertical and second highest is in "Other" homeownership category.

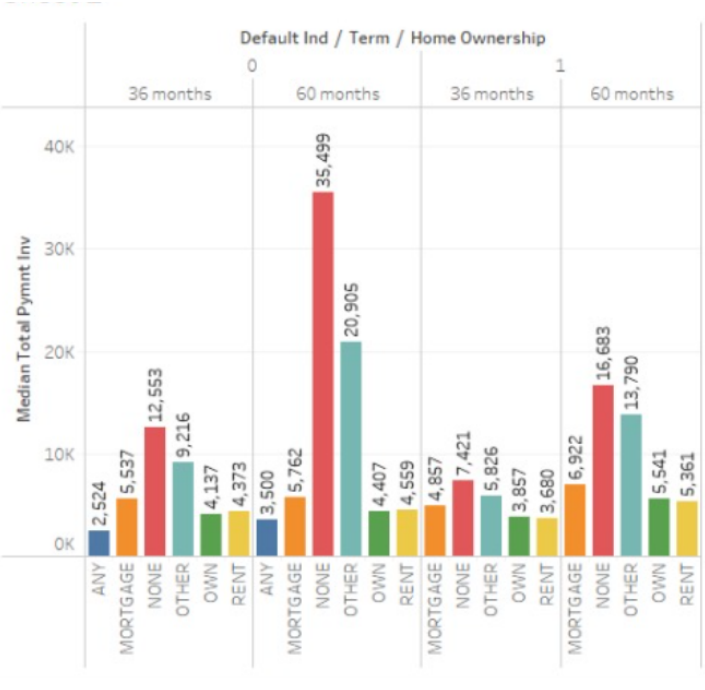
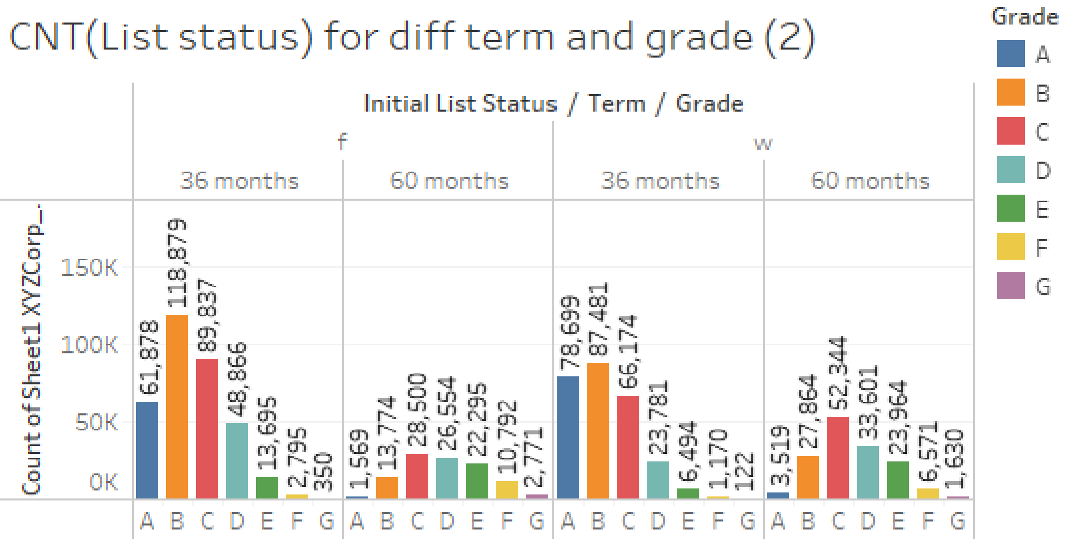
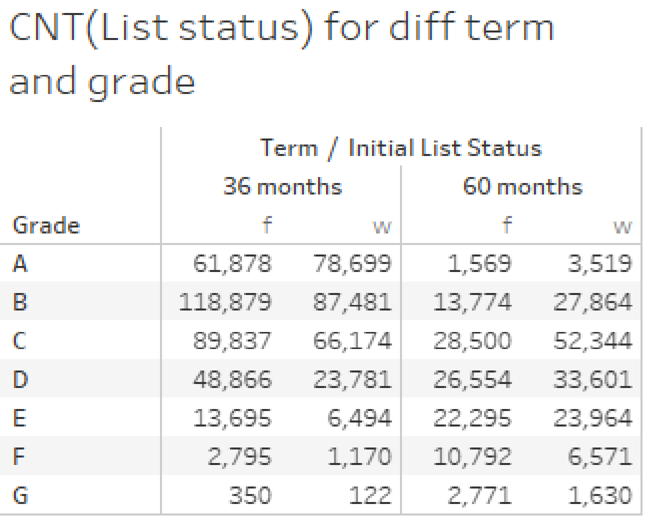


Fig 13. Multivariate analysis of Total Payment,Term, Home Ownership & Default Indicator

Hence other loan categories need to be focused for faster repayment of principal as median payment in the defaulter category is equal to that of genuine borrowers where a unique set of characteristics need to be defined for the machine to distinguish properly to identify better loan categories to be marketed for creditors and borrowers based on the risk level of default in each class.



CNT(List status) for diff term and grade (2)



CNT(List status) for diff term and grade

	Term / Initial List Status			
	36 months		60 months	
Grade	f	w	f	w
A	61,878	78,699	1,569	3,519
B	118,879	87,481	13,774	27,864
C	89,837	66,174	28,500	52,344
D	48,866	23,781	26,554	33,601
E	13,695	6,494	22,295	23,964
F	2,795	1,170	10,792	6,571
G	350	122	2,771	1,630

Figure 14. Multivariate analysis of List status, Term & Grade

From the above plot, we can see that Grade B are the highest loan categories people are willing for in either initial_list_status (whole & fractional) for 3-year term loans. Even though there is a benefit from getting ‘instant funding’ almost 36% more people are biased for the fractional in the respective grade.

Grade C & D are almost equal and highest in 5-year term loans so better benefits should be provided in other grades for an exclusive cash flow.

Also for an investor, it is an opportunity to invest and break even at a good pace with good interest rates for these grade loans as the median returns are also good making them profitable.

7. DATA TRANSFORMATION

The process of modifying data while retaining the information is known as feature transformation. It is nothing more than a function that converts features from one representation to another. When our original data does not follow a normal distribution, we can use various transformations to make it as normal as possible, allowing the normality assumption in various machine learning algorithms to be valid. Transformation reduces the skewness in the distribution of the original data and makes the data more interpretable.

Our dataset has features with high skewness. Thus, to make our data more Gaussian-like, we will carry out a transformation. Some of the transformation techniques that can be used are as follows:

1. **Square root transformation:** Values of a variable are replaced with its square root. It can be applied even when the variable takes a zero value.
2. **Box cox transformation:** Generalized form of logarithmic transformation. The Box-Cox transformation can only be used on positive variables.
3. **Yeo–Johnson transformation:** The Yeo–Johnson transformation allows also for zero and negative values

Out of all these, we initially went for square root transformation as it could be applied even when the variable takes a zero value, which it does in our case. The skewness of a few variables showed a reduction, but some variables were still skewed. Also, it was noticed that square root transformation carried the skewness of few features away from zero.

To curb this issue, we, later on, applied the Yeo-Johnson method using Power Transform which showed improvement in our results. In the Yeo-Johnson method, the optimal parameter for stabilizing variance and minimizing skewness is estimated through maximum likelihood. By default, zero-mean, unit-variance normalization is applied to the transformed data; hence, no scaling is required.

8. MACHINE LEARNING

8.1 TRAIN-TEST SPLIT

The Dataset is stratified into train and test in the ratio of 70:30

8.2 BASE MODEL

Logistic Regression

Logistic Regression is a predictive algorithm that uses independent variables to predict the dependent variable. It is used for classification problems and is based on the concept of probability. Logistic regression is a statistical model that uses the Logistic function to model conditional probability.

Despite being a linear regression model, Logistic Regression uses a complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’ instead of a linear function. The hypothesis of logistic regression tends to limit the cost function (hθ) between 0 and 1.

Representation of Logistic Regression

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

where,

- hθ(x) is the cost function
- b0 is the bias or intercept term
- b1 is the coefficient for the single input value (x)

Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

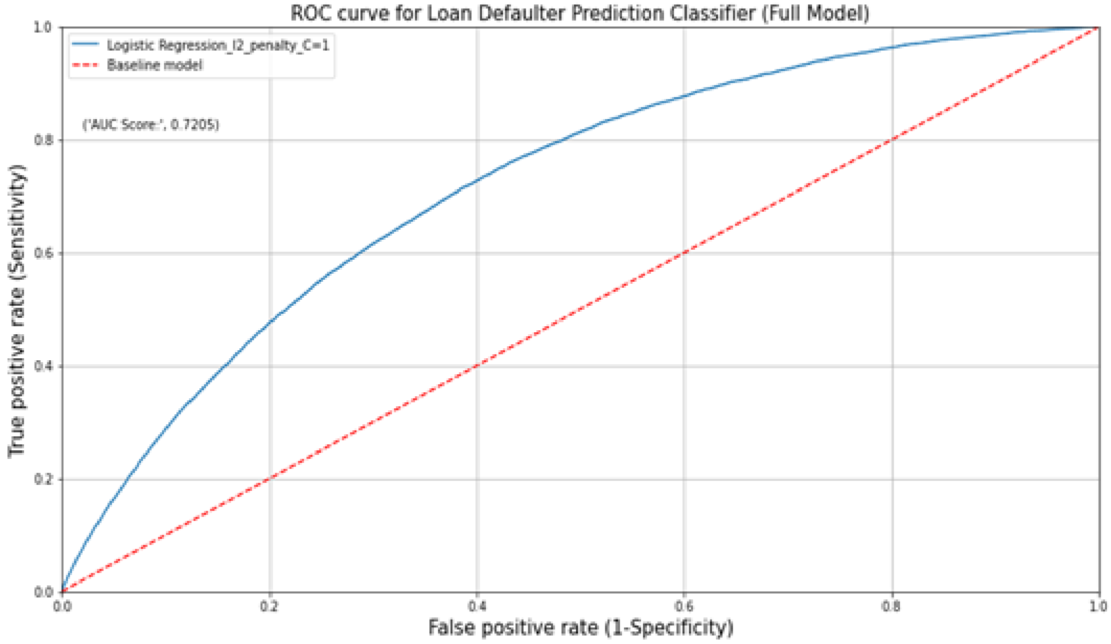
Performance Metrics

For the base model, we fit the training dataset with all features into the Logistic Regression Classifier. Since this is an imbalanced classification problem where class 1 represents just 5% of the dataset, we set the parameter Class_weights='balanced' in the logistic regression model.

The fitted model had the following scores:

- Accuracy = 65%
- Precision = 10%
- Recall = 67%
- F1 score = 18%

ROC-AUC Curve



The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is often used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. Our base model has an AUC Score of 72.05%, which indicates our model can correctly classify both classes correctly almost 72% of the time.

Tuning the HyperParameters:

A hyperparameter is a parameter that is set before the learning process begins. It is given by the user. Here, we use the GridSearchCV technique to select the best combination of hyperparameters in order to obtain the best results from our models.

By giving a range of different hyperparameter values, the GridSearchCV algorithm selects the optimum hyperparameters by trying all the combinations of the values passed in the dictionary and evaluating the model for each combination using the Cross-Validation method. For Logistic regression, penalty=’l2’ and C=0.7 are chosen. For the Decision Tree classifier, max_depth=8 and criterion=’gini’ are chosen. The same is fed into our Random Forest classifier and n_estimators=28 is chosen. For Light GBM, num_leaves= 30 and max_depth=12 are selected. We then pass these hyperparameters to our base model to tune the model before performing any further optimization techniques.

Following are the scores of five tuned base models:

Model	Recall	Precision	F1 Score	ROC-AUC
Logistic Regression_tuned	0.69	0.11	0.2	0.69
NB_tuned	0.46	0.12	0.19	0.6318
Decision Tree_tuned	0.77	0.12	0.21	0.797
Random Forest_tuned	0.72	0.13	0.21	0.7834
Light GBM	0.80	0.13	0.22	0.8179

Even after tuning the hyperparameters of our models, we see that the precision and f1 scores remain low because of the presence of a high imbalance in the dataset. In order to improve our results, we have to carry out some imbalance treatment.

8.3 SMOTE

Standard classifier algorithms like Decision Tree and Logistic Regression have a bias towards classes that have a number of instances. The presence of an imbalance in the data increases the tendencies of the model only predicting the majority class data and ignoring the minority class. If left untreated, there is a high probability of the misclassification of the minority class as compared to the majority class.

In order to treat the imbalance present in our dataset, we carry out SMOTE, which is an oversampling technique where synthetic samples are generated for the minority class. This technique is better than random oversampling as it prevents overfitting. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

Best results were observed when SMOTE was combined with Random Undersampling. In our case, we used a sampling strategy of 0.28 while carrying out SMOTE and clubbed it with Random Undersampling where a sampling strategy of 0.34 was used. These values were obtained by trial and error method, during which we learnt that as the sampling strategy of SMOTE was enlarged, accuracy and precision improved and recall and F1 scores of the minority class reduced; the opposite was observed while increasing the sampling strategy of Random Undersampling method. Appropriate sampling strategies were chosen based on the acceptable tradeoff of precision and recall. Tuning of the ’k_neighbors’ parameter for SMOTE was done by cross-validating the resampled datasets for various values of ‘k’ and k=2 was found to give the highest scores.

Following are the scores after carrying out the imbalance treatment:

Model	Recall	Precision	F1 Score	ROC-AUC
Logistic Regression_tuned	73	10	18	75
NB_tuned	62	10	17	70
Decision Tree_tuned	75	12	21	78
Random Forest_tuned	78	11	19	77
Linear Support Vector_tuned	81	09	17	-

8.4 FEATURE SELECTION

Statistical Testing

1. Categorical Independent Variables

Chi-Square test for independence:

Null hypothesis - The two variables are independent
Alternate hypothesis - The two variables are dependent

[‘'term_ 60 months', 'home_ownership_MORTGAGE', 'home_ownership_NONE', 'home_ownership_OTHER', 'home_ownership_OWN', 'home_ownership_RENT', 'verification_status_Source Verified', 'verification_status_Verified', 'initial_list_status_w', 'grade', 'sub_grade', 'emp_length']

Using the Chi-Square test, it has been found that the above attributes have a significant influence on the target column (p-values are less than 5%).

2. Continuous Independent Variables

Mann-Whitney U Test:

Null hypothesis - The sample distributions are equal.
Alternate hypothesis - The sample distributions are not equal.

['loan_amnt', 'int_rate', 'annual_inc', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim']

Using the Mann-Whitney U test, it has been found that the above continuous attributes have a significant influence on the target column (p-values are less than 5%).

Recursive Feature Elimination (RFE):

Since SMOTE analysis is not successful in improving metric scores, we will use our original dataset and perform feature selection using the Recursive Feature Elimination technique and build classification models on selected features and hyper tune those models.

Total Features:

'loan_amnt', 'int_rate', 'annual_inc', 'dti', 'delinq_2yrs','inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util','total_acc', 'tot_coll_amt', 'tot_cur_bal', 'total_rev_hi_lim','term_ 60 months','home_ownership_MORTGAGE','home_ownership_NONE','home_ownership_OTHER' ,'home_ownership_OWN', 'home_ownership_RENT','verification_status_Source Verified', 'verification_status_Verified','initial_list_status_w', 'grade', 'sub_grade', 'emp_length'

Decision Tree Features:

'loan_amnt', 'int_rate', 'annual_inc', 'dti', 'revol_bal','tot_cur_bal', 'total_rev_hi_lim', 'initial_list_status_w', 'grade','sub_grade'

RF Features:

'loan_amnt', 'int_rate', 'annual_inc', 'dti', 'inq_last_6mths','tot_cur_bal', 'total_rev_hi_lim', 'initial_list_status_w', 'grade','sub_grade'

Logistic Regression Features:

'int_rate','delinq_2yrs','inq_last_6mths', 'open_acc', 'pub_rec''home_ownership_MORTGAGE', 'home_ownership_OWN', 'home_ownership_RENT','verification_status_Source Verified', 'verification_status_Verified','initial_list_status_w', 'grade', 'sub_grade', 'emp_length'

Model	Recall	Precision	F1 Score	ROC-AUC
Logistic Regression_tuned	0.74	0.11	0.18	0.7576
NB_tuned	0.73	0.09	0.16	0.7135
Decision Tree_tuned	0.79	0.12	0.21	0.7978
Random Forest_tuned	0.75	0.12	0.21	0.7878
Light GBM	0.81	0.13	0.22	0.817
Ada Boost	0.78	0.12	0.21	0.7983
XG Boost	0.8	0.13	0.22	0.814
Stacking (DT+XG Boost+light GBM)	0.82	0.12	0.22	0.8175

8.5 FINAL MODEL:

Out of all the above models, it is important to choose one that addresses our problem, in this case, one that correctly predicts Defaulters. Looking at the 'recall' scores, we see that the Stacking model provides the highest value of recall. Thus, our final model is a Stacked model consisting of three base estimators, namely, Decision Tree, XGBoost and Light Gradient Boost; Light GBM is selected to be the final classifier. This model has the advantage of reduced bias. The model is fitted with a train data of best features that are selected using Recursive Feature Elimination with LGBM as the underlying classifier.

From the obtained results, we see that our model is able to predict 82% of all positive cases correctly, which is an important metric for the lending institution to prevent it from incurring losses.

9. BUSINESS SUGGESTIONS

A business insight combines data and analysis to find meaning in and increase understanding of a situation, resulting in some competitive advantage for a business. Simply performing exploratory data analysis, building models and deriving insights won't be of any help if we are not able to leverage these insights into business solutions.

Here, we discuss some of the business solutions that the company i.e. XYZ Corp can implement and improvise upon, based on the insights gathered from our dataset.

Insight: The common best features for predicting a defaulter.
Business Solution: In order to ensure minimum risk of default, the grade & sub-grade of loan needs to be checked while issuing/handling loan of a borrower along with initial listing status (if whole) as it is from a single lender higher the chance of defaulting - higher the loss. Hence, this needs to be focused upon in a dual manner - to attract investors for obtaining instant funding that helps borrowers and also to ensure in minimizing risk as a higher concentration of loans is in Fractional category 'F' which is 51% that annotates more no. of investors involvement.

Insight: The recommended model for the company for future data handling based on the current dataset.

Business Solution: A stacking model can be used (DT+XG Boost+light GBM) based on the current dataset as it is predicting the target with the highest score of around 82%.

Insight: Higher the grade, i.e.; A to G, more the risk of default.

Business Solution: People who requested/opened a line for a loan that is into a lower grade, are not more likely to default. But for higher loan grades, chances are high. To tackle this, the company can make loans issued only after proper collateral so that principal can be recovered and make interest rate feasible with mix and match as per industry standards and run a proper background check, stringent terms & conditions for issuing a loan.

Insight: Helpful metrics found in the analyses.

Business Solution: Since revolving utilization mentions how good a borrower is utilizing the issued loan and how much likely is he/she to default, also since around 62% of the applicants are employed for less than 10 years and most are defaulters as they might be having higher financial dependence and 48% cases are 'W' category. So, the company can focus on all credit line metrics of the borrower to check how many open accounts are present and how much one had utilized and overdue in one's past/current financial history and collateral available (if applicable) like homeownership, bonds/deposits etc.

10. PROJECT OUTCOME

There are different metrics that help in quantifying the performance of the model such as accuracy, recall, precision, F1 score. For our model, the recall score is the most accurate metric to judge performance.

Before explaining the reason why we will talk about the FP(False Positive) and False Negative(FN) predictions made by the model. With respect to our dataset, FP implies that the model has incorrectly labelled a borrower as a defaulter. FN implies that the borrower is a defaulter but identified as a non-defaulter. FP condition may lead the bank to reject the loan application of a genuine borrower, whereas FN may lead to the approval of unsecured lendings.

Both of these are unfortunate and have a high cost attached to them. While precision gives weightage to FP, recall gives weightage to FN. F1 score is the harmonic mean of precision and recall, thus giving weightage to both the conditions. Meanwhile, accuracy doesn't give any weightage to False Positive and False Negative.

For our model, recall is the best score as our main objective is to recognize the unsecured lendings. We have obtained a recall of 0.82 for both the train and test data, which indicates the model is neither underfitting nor overfitting. The ROC-AUC Score for our model is 0.8175 which is closer to 1 and indicates that our model is able to distinguish between the two classes quite correctly.

There is definitely room for improvement here. Since there is such a high ratio of imbalance, such scores are expected. There are chances of improving the model further by carrying out advanced feature engineering, which will be part of our future work.

11. CONCLUSION

By using XYZ Corp. loan lending prediction dataset, we found the main objectives of the research along with the application of data science skills such as data exploratory data analysis, data preprocessing, Feature Engineering & machine learning. We carried out a full data science lifecycle on this dataset.

Various Classification models like logistic regression, naive Bayes, decision tree, random forest, ADA boost, XG boost, light GBM, stacking classifier are being evaluated on the dataset and we found that stacked model of Decision Tree, XG Boost and Light GBM, is a good technique to build a loan defaulter prediction model. The model development revealed that the features had different weights and importance, so we tried to build a model which could satisfy all the criteria.

The main problem with the dataset is the imbalanced nature of the default class variable, to deal with this issue SMOTE analysis is being carried out and classification models are being evaluated, which did not yield satisfactory results.

To improve metric scores feature engineering is performed using an automated feature creation tool, Feature Tools, but did not quite result in satisfactory outcomes. This area will be explored in great depth in future.

The stacked model enables financial institutions to mitigate revenue loss derived from loan defaulters. Prediction of loan defaulters allows the banks to implement less rigid loan policies, without increasing uncertainty. This has the potential to render more sales.

Concurrently, the development of these models should contribute to improving revenue management for the banks and minimize losses caused due to NPAs or bad loans.

REFERENCES

Akash Kesrwani. (2021). XYZCorp_LendingDataPrediction. https://www.kaggle.com/akashkesrwani/xyzcorp-lendingdataprediction

CFI. (n.d.). *Credit Analysis*. Retrieved from Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/knowledge/credit/credit-analysis/

CFI. (n.d.). *Loan Analysis*. Retrieved from Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/knowledge/credit/loan-analysis/

CFI. (n.d.). *Major Risks for Banks*. Retrieved from Corporate Finance Institute: https://corporatefinanceinstitute.com/resources/knowledge/finance/major-risks-for-banks/ 5. Ereiz, Z. (2019, November). Predicting Default Loans Using Machine Learning (OptiML). In *2019 27th Telecommunications Forum (TELFOR)* (pp. 1-4). IEEE.

Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.

Cohen, M. C., Guetta, C. D., Jiao, K., & Provost, F. (2018). Data-driven investment strategies for peer-to-peer lending: A case study for teaching data science. *Big data*, *6*(3), 191-213.

Fan, S., Shen, Y., & Peng, S. (2020). Improved ML-Based Technique for Credit Card Scoring in Internet Financial Risk Control. *Complexity*, *2020*.

HDFC Bank. (n.d.). *7 Factors That Determine Whether Your Loan Gets Sanctioned*. Retrieved from HDFC Bank: https://www.hdfcbank.com/personal/resources/learning-centre/borrow/7-factors-that-determine-wheth er-your-loan-gets-sanctioned?

Legal Match. (2018, June 07). *Long Term vs. Short Term Loans*. Retrieved from Legal Match: https://www.legalmatch.com/law-library/article/long-term-vs-short-term-loans.html

Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012042). IOP Publishing.

Mampatta, S., & Lele, A. (2020, July 14). India saw nearly Rs 25,000-cr wilful defaulter cases even before lockdown. Business Standard.

Mathew,G., Narayan, K. (2021, May 11). Amid Covid effect, bank steps, wilful defaults rise Rs 38,976 crore. *The Indian Express.* https://indianexpress.com/article/business/amid-covid-effect-bank-steps-wilful-defaults-rise-rs-38976- crore-7309969/

Ratings, S. P. G. (2016). S&P global ratings definitions. *URL: https://-www. standardandpoors. com/en_US/web/guest/article/-/view/sourceId/504352, Accessed on*, *15*, 13.

Sastry, D. V. (2020). *Business Analytics and Business Intelligence: Machine Learning Model to Predict Bank Loan Defaults.* Blue Diamond Publishing.

Vikash. V, M. A. (2018). *Loan Default Prediction using Machine Learning Techniques.* 14.
Yash Diwate, P. R. (2021). Loan Approval Prediction Using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)*, 5.

Wah, Y. B., Rahman, H. A. A., He, H., & Bulgiba, A. (2016, June). Handling imbalanced dataset using SVM and k-NN approach. In *AIP Conference Proceedings* (Vol. 1750, No. 1, p. 020023). AIP Publishing LLC.

(2021, June 01). Getting personal loan sanctioned can be a daunting task amid Covid 19. Mint.https://www.livemint.com/money/ask-mint-money/getting-personal-loan-sanctioned-can-be-a-da unting-task-amid-covid-19-11622545327909.html

https://www.business-standard.com/article/finance/number-of-willful-defaulters-rose-before-coronavir us-lockdown-analysis-120070900273_1.html

DECLARATION

This is to declare that the dataset that we are using for our capstone project is publicly available and can be used to showcase the work we do on it as a presentation in Great Learning.

Original owner of the data	Kaggle
Data set information	XYZCorp_LendingData Credit Risk Analysis
Any past relevant articles using the dataset	Vatsal Gala (November 19, 2019).*XYZ Corp Lending [Prezi Slideshow]* https://prezi.com/p/qubjcxvb1kai/xyz-corp-lending/
Link to webpage	https://www.kaggle.com/sonujha090/xyzcorp-lendingdata

****