EDA CASE STUDY

Arijit Mukherjee (arijit.mukherjee142228@gmail.com) (Group Facilitator)

Nilratan Chattopadhyay (Nilratan.Chattopadhyay@yahoo.com)

Shaswata Tripathy (tripathyshaswata@gmail.com)

ML-AI-Cohort 3

29th July 2018

Preface

Problem statement :

Consumer Finance companies run the risk of business loss by granting loans to consumers who are unlikely to repay the loans (defaulters) and also financial loss by denying loans to a good consumer who is willing to repay the loan on time. Can data analytics help solve this problem?

The objective of this case study is to analyze consumer loan data of a NA based consumer finance company to isolate consumer and loan attributes that will help in identifying potential loan defaulters and thus reducing the risk of business loss due to bad loans.

EDA techniques are applied to list out the preliminary set of loan default drivers that can be the inputs to subsequent advanced analysis (not in scope of this study) for decisive action on whether to grant loan and what should be the funded amount if any consumer loan applicant profile analyzed to be unlikely to repay the loan.

Inputs Used:

- Instruction, definitions and process flow of loan application processing
- Loan dataset on CSV format
- Data Dictionary
- Additional study materials referred on consumer finance domain

Constraints and assumptions:

Loan application reject data not available in supplied data set

Did not undertake any text analysis as not part of EDA

Analysis focus in of total number of defaulters and not the total amount (as the problem statement is whether to approve loan or not to a consumer)

Conclusion:

Identified potential attributes like Loan Verification Status, Address States, Home Ownership type, Loan Approval grades as to qualify for loan defaulter. Also additional analysis has been done beyond EDA on what could be the reasons for ineffective loan Verification process. Details in subsequent slides.

The process

We have applied 7 step process

Step 1: Data Understanding and Cleaning

Step 2: Univariate Analysis and Summary

Step 3: Bivariate Analysis of Categorical data

Step 4: Bivariate Analysis of Continuous data

Step 5: Bivariate Analysis Summary

Step 6: Additional Analysis

Step 7: conclusion

Step 1: Data Understanding and Cleaning

It is important to understand the available data, clean the data and create a subset which relevant for analysis and optimal use of resources.

Loan database as provided : 39717 records, 111 variables, data types include Float, Integer and object.

Dropping the records and valiances with no data

- Drop the columns having no data.
- Drop columns that has more than 30% null values (with data understanding that null value has no business meaning)
- Drop the records where loan amount is zero. Assuming that loan reject records are not provided in the supplied data.

Cleaning and formatting the data:

- Replace Null values with zero (eg. pub_rec_bankruptcies)
- Remove special characters (eg. revol_util)
- Split strings to extract numeric values (eg. emp_length)

Step 2: Univariate Analysis and Summary

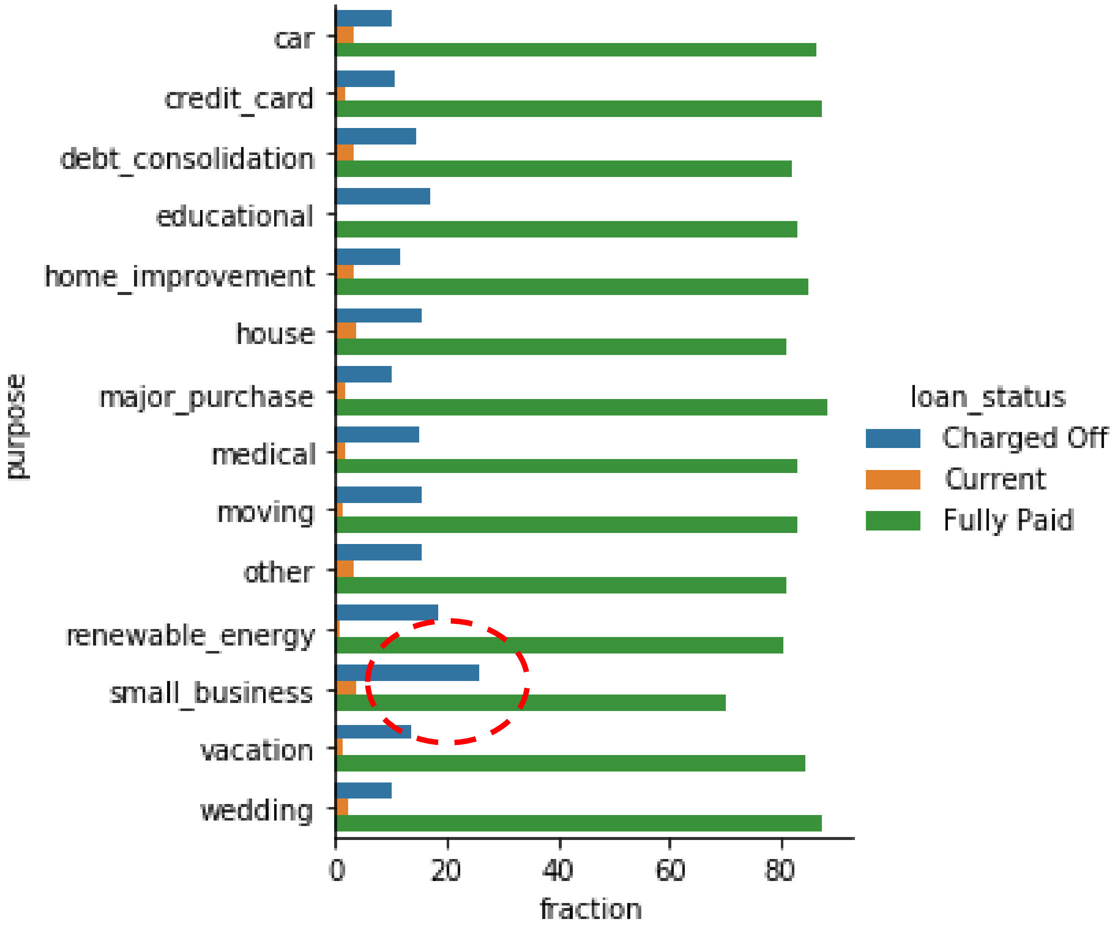
Undertook Univariate EDA to identify driver variables. Enlisting below important few..

Variable name	Description	Observation
funded_amnt_inv	Continuous variable	Max USD 35000: Min 0 USD: Median USD: 8975
Interest Rate	Continuous variable	Max 24.59% : Min 5.42% : Median : 11.86%
Grade	There are 5 grades and interest rates applied is dependent as grade	Grade can be used as category variable for interest rate slabs. G Followed by F has the highest default rate.
Home ownership	There are 5 types of home ownership	Maximum member of loans have home ownership type “RENT”
Verification status	There are 3 types of verification	Maximum member of loans have verification status “Not Verified” followed by ‘Verified’
Loan Status	Loan status are of three types “Charged Off”, “Fully Paid”, and “current”	Data provided has maxim umber of Fully paid records followed by charged off 32950 followed by 5627
Purpose	There are 13 types of purpose	Maximum member of loans taken for Debt_Consolidation purpose 18641
Emp Length	Categorical variable ranges from less than one year to more than 10 years	Default is for 10 years followed by 1 year.

Used above parameters for segmentation. Most important one is LOAN STATUS hence used it throughout all the bivariate analysis in subsequent analysis.

Step 3: Bivariate Analysis of Categorical Data

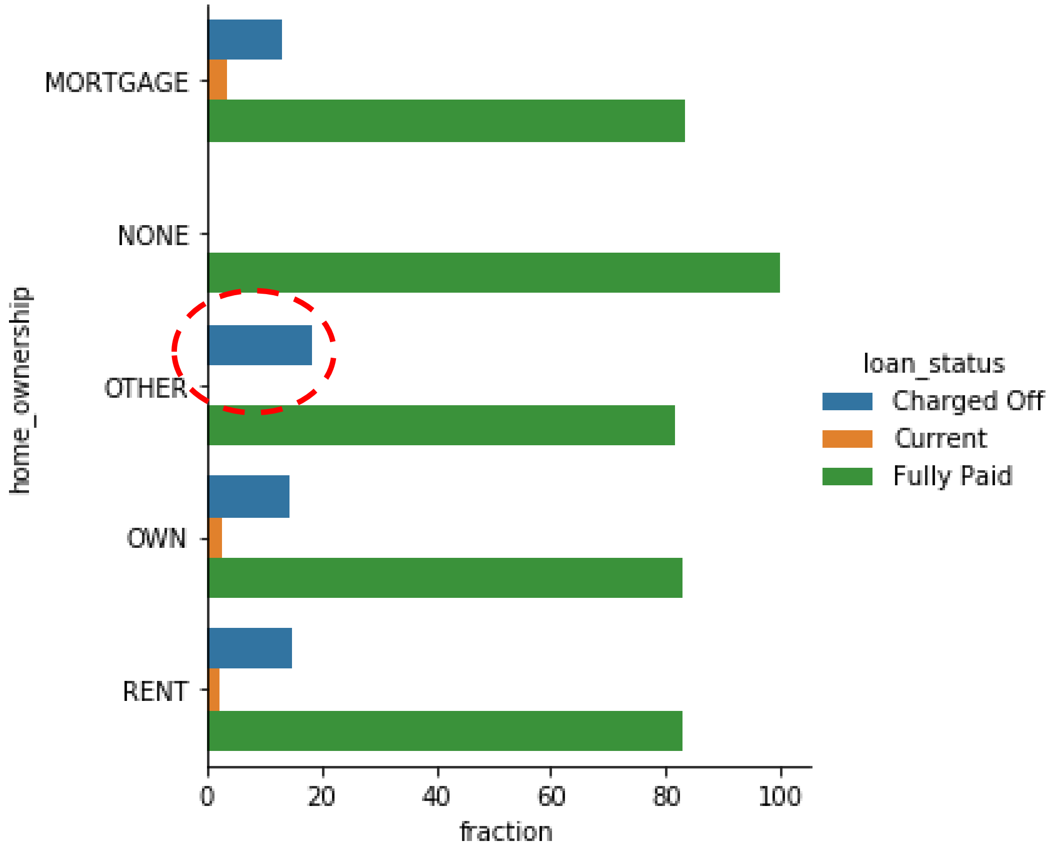
Analysis on purpose for which loan has been taken for different types of loan status..



Consumers taken small business loans have defaulted the most. Educational and Renewable energy types follows then……..

Step 3: Bivariate Analysis of Categorical Data

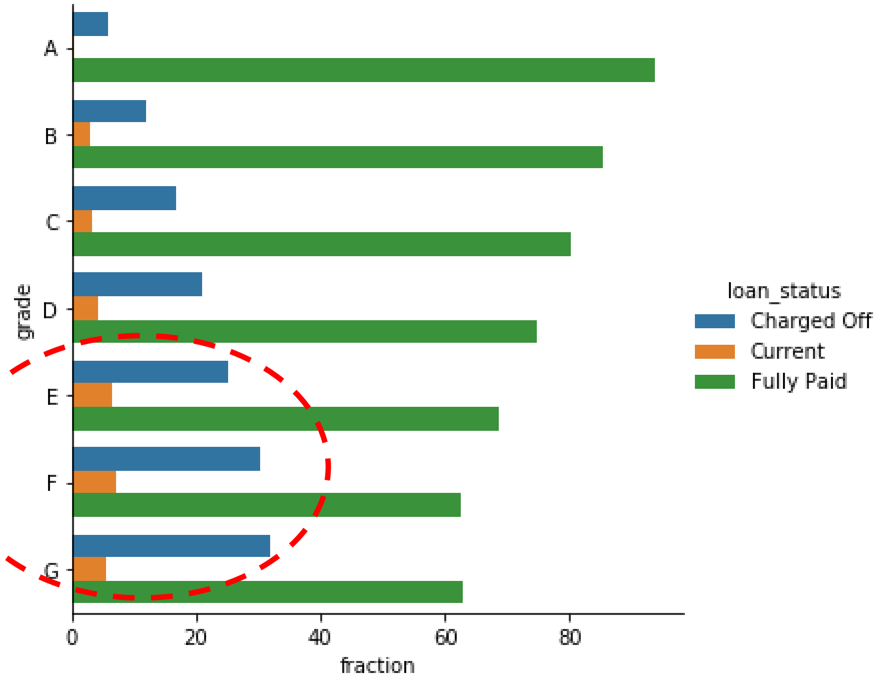
Analysis on home ownership type of consumers for different types of loan status..



Consumers having OTHER type home ownership have defaulted most. Interestingly where home ownership is not mentioned (NONE) there is no defaulter?

Step 3: Bivariate Analysis of Categorical Data

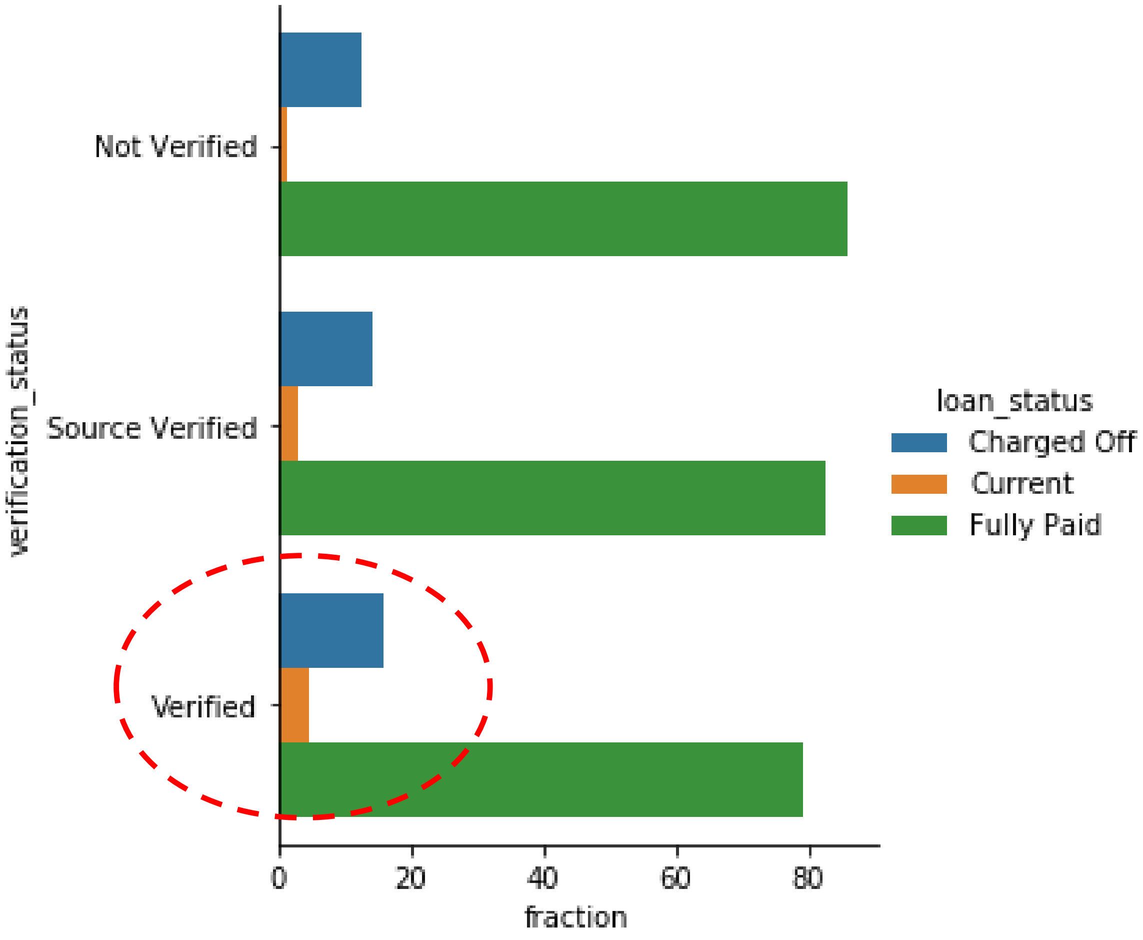
Analysis on loan approval Grade of consumers for different types of loan status..



Consumers being graded as F and E types for loan approval have defaulted the most

Step 3: Bivariate Analysis of Categorical Data

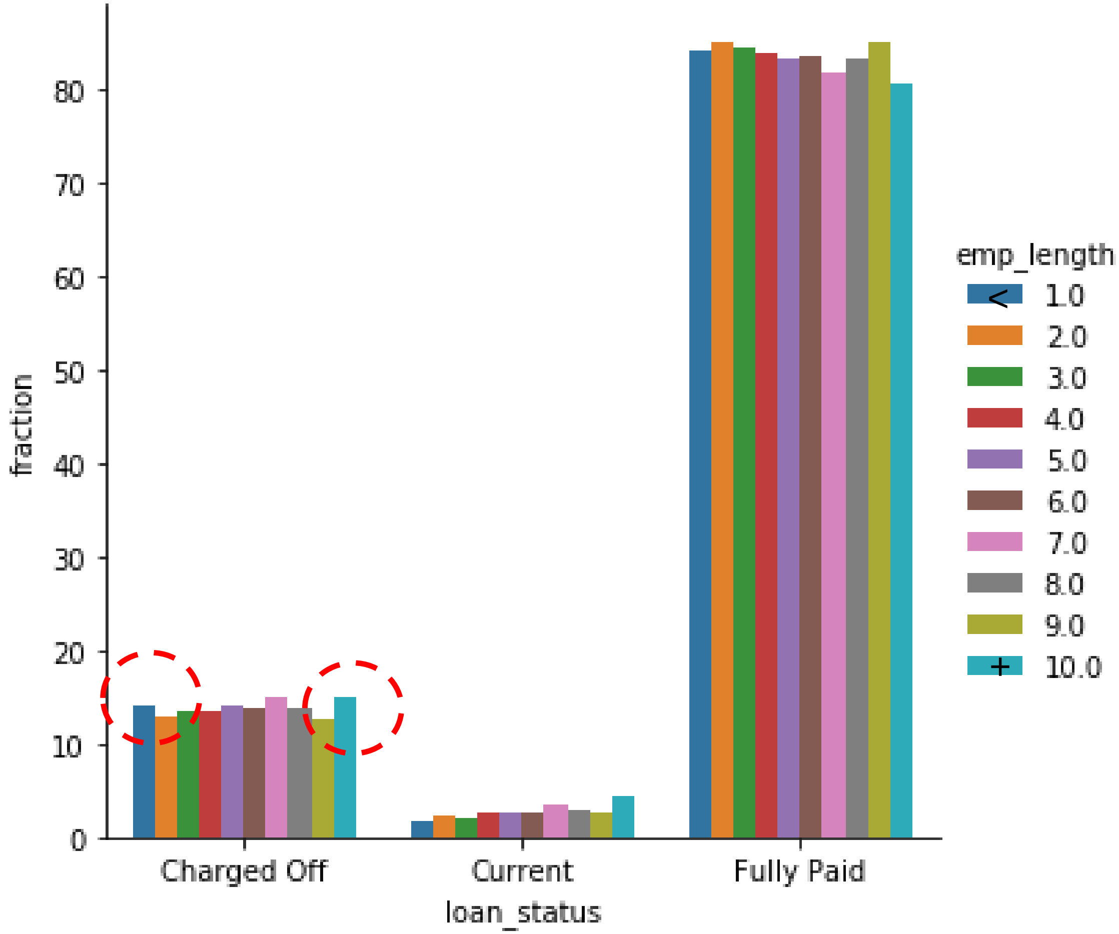
Analysis on loan verification status of consumers for different types of loan status..



Interestingly loans where Consumer Submitted Documents/Data verifies have maximum % of Defaulters

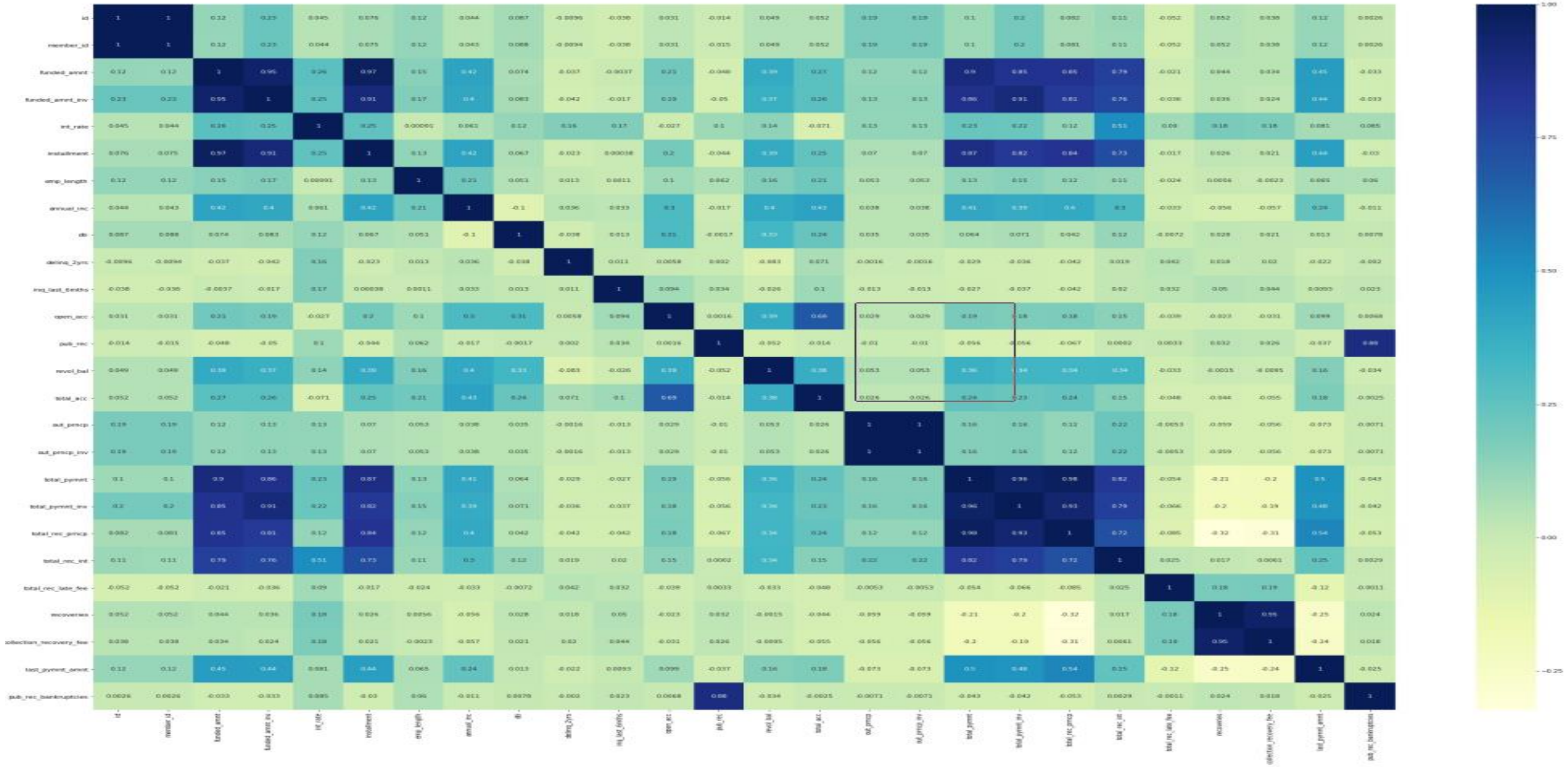
Step 3: Bivariate Analysis Categorical Data

Analysis on loan verification status of consumers for different types of loan status..



More than 10 years and less than one years of employment length consumers are defaulting more

Step 4: Bivariate Analysis of Continuous Data



Could not find interesting strong positive or negative correlation amongst continuous variables (other than the obvious like 0.98 is correlation coefficient between funded amount and loan amount) to recommend for further analysis

Step 5: Bivariate Analysis Summary

Enlisting below the summary of bivariate analysis done to identify potential lead attributes for consumer loan defaulter

Variables used	Outcome	Early indications that needs detail analysis
Verification status vs Loan status	Consumers where loan application documents/data have been completely VERIFIED have defaulted the most	In detail analysis required to identify in which sates verification process is ineffective. Is data getting altered during verification
Loan approval grade vs Loan Status	Consumers having GRADE as F and E types for loan approval have defaulted the most	Interestingly the worst graded loans (G) have relatively less defaults
Home ownership type of consumer and loan status	Consumers having MORTGAGE TYPE home ownership have defaulted most	Interestingly where home ownership is not mentioned (NONE) there is no defaulter?
Loan purpose and loan status	Consumers taken SMALL BUSINESS LOANS have defaulted the most. Educational and Renewable energy types follows there after.	Further analysis required to identify which attributes will help find potential loan defaulter under Shall Business category
Employment length of consumer	Consumers having less than one year and more than 10 years of employment length have defaulted the most	Further analysis required to identify which attributes will help find potential loan defaulter under less than one year employment length category
Loan Term	Consumers taken 60 months loan tend to default more	Consumers taken 60 months loan tend to default more

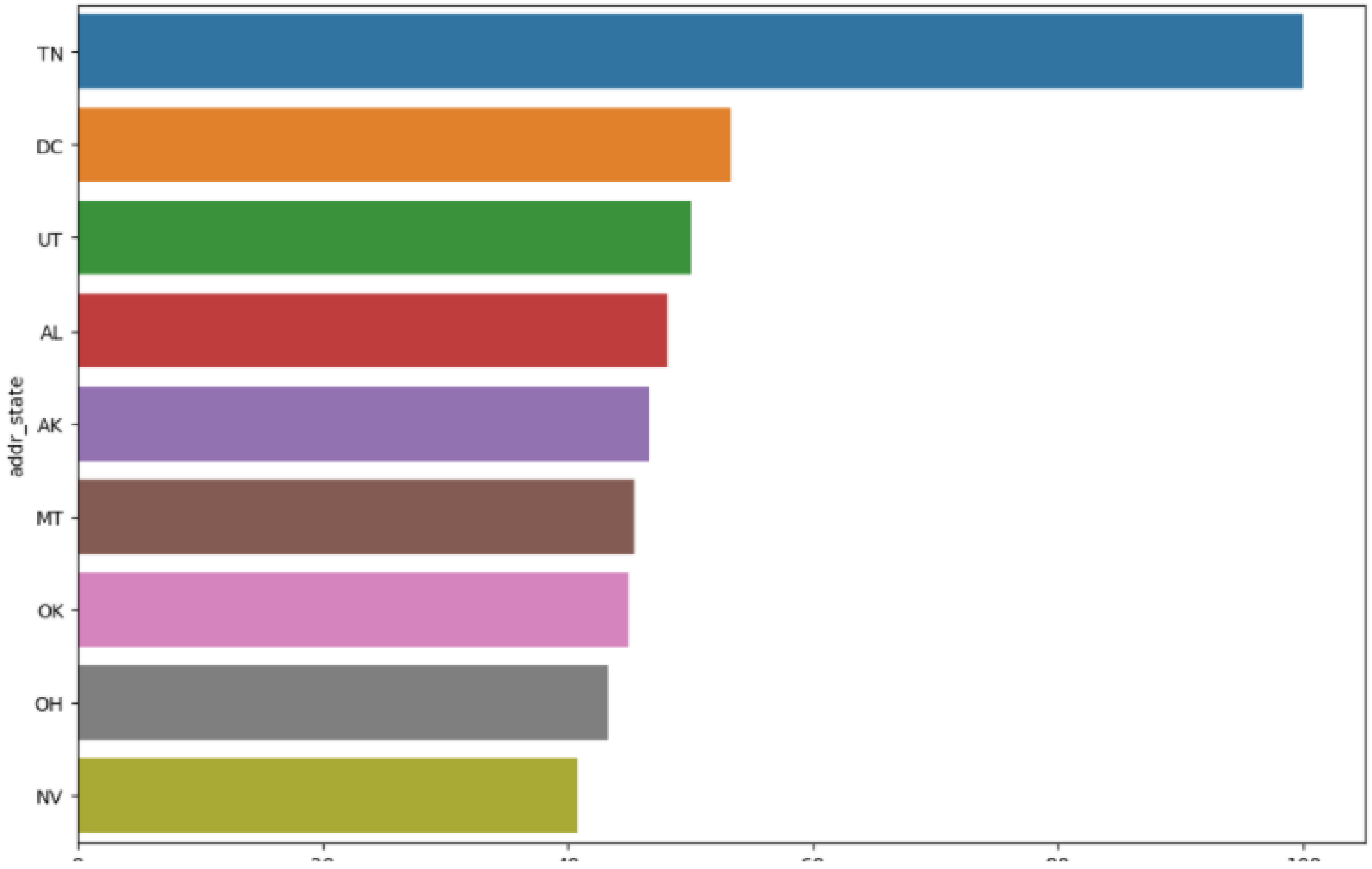
Outcome of all the above bivariate analysis are early indicators to raise questions related to who can be the potential consumer loan applicants for default if approved.

Microsoft Excel Worksheet

The indicator on “ verification Process” ineffectiveness is quite interesting. Hence we have taken up additional analysis for deeper understanding

Step 6: Additional analysis of “Verification Status” Data

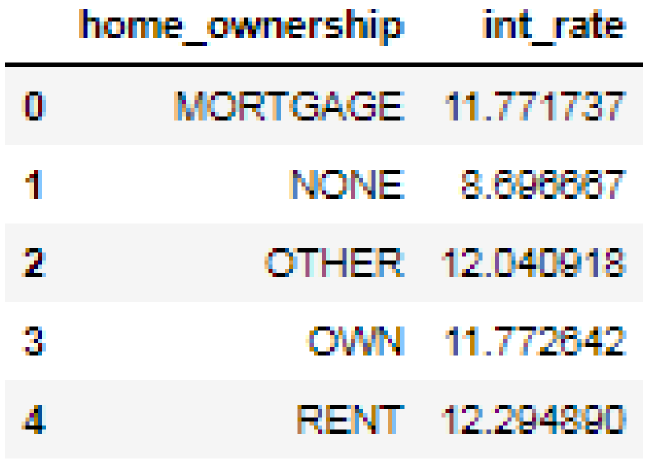
Analysis of Charged off consumers with verification status “Verified” for different consumer address states



The consumer loan approval verification process are poor in top 5 states are TN (Tennessee), DC (District of Columbia), UT(Utah), AL(Alabama), AK (Alaska) ,leading to high default even after complete verification

Step 6: Additional analysis of “verification Status” Data

Analysis of Charged off consumers with verification status “verified” , “home ownership” and “Interest rate”



	home_ownership	int_rate
0	MORTGAGE	11.771737
1	NONE	8.696667
2	OTHER	12.040918
3	OWN	11.772642
4	RENT	12.294890

Observed that majority of Charged Off consumers where data is verified, the home ownership mentioned is MORTGAGE. On the other hand the average interest rate applicable is low for Mortgage type home ownership.

Is the home ownership type getting changed to MORTGAGE during verification process to enable lower interest loan ? This needs more analysis before conclusion

Conclusion

Incase a consumer finance loan applicant data have the following attributes, detail analysis required to qualify for loan sanction to avoid potential default (charged off) leading to business loss.

Attributes	Indicators
Loan Verification status	Percentage of defaulters high with verification status as VERIFIED
Consumer Address State	Loan Verification process may be ineffective in the states like TN, DC, UT, AL, AK
Home ownership type	Home ownership type MORTGAGE if verified may need re-verification
Length of employment of consumers	Consumers having less than ONE YEAR and more than 10 YEARS employment experience has high potential to default
Loan taken for Small Business	Defaulter percentage is high for loan taken to fund SMALL BUSINESS
Loan approval grade	If the loan approval GRADE falls under E and F, high probability to default
Loan Term	Potentials defaulters opt for long tenure (60 months) loans

The EDA case study is well throughout and sequenced in the MLAI program. Enabled us to get hands-on experience in working on data analysis with the real word data leading to deep data understanding, data cleaning, logical thinking and finally applying PYTHON for EDA analytics. Truly appreciate team IIIT-B and Upgrad who have designed this case study.