Lending Club Case Study

Prepared by: Ramesh Krishnan

1. Introduction

- **Problem statement**
  - You work for a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
    - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
    - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

- **Objective**
  - The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Methodology

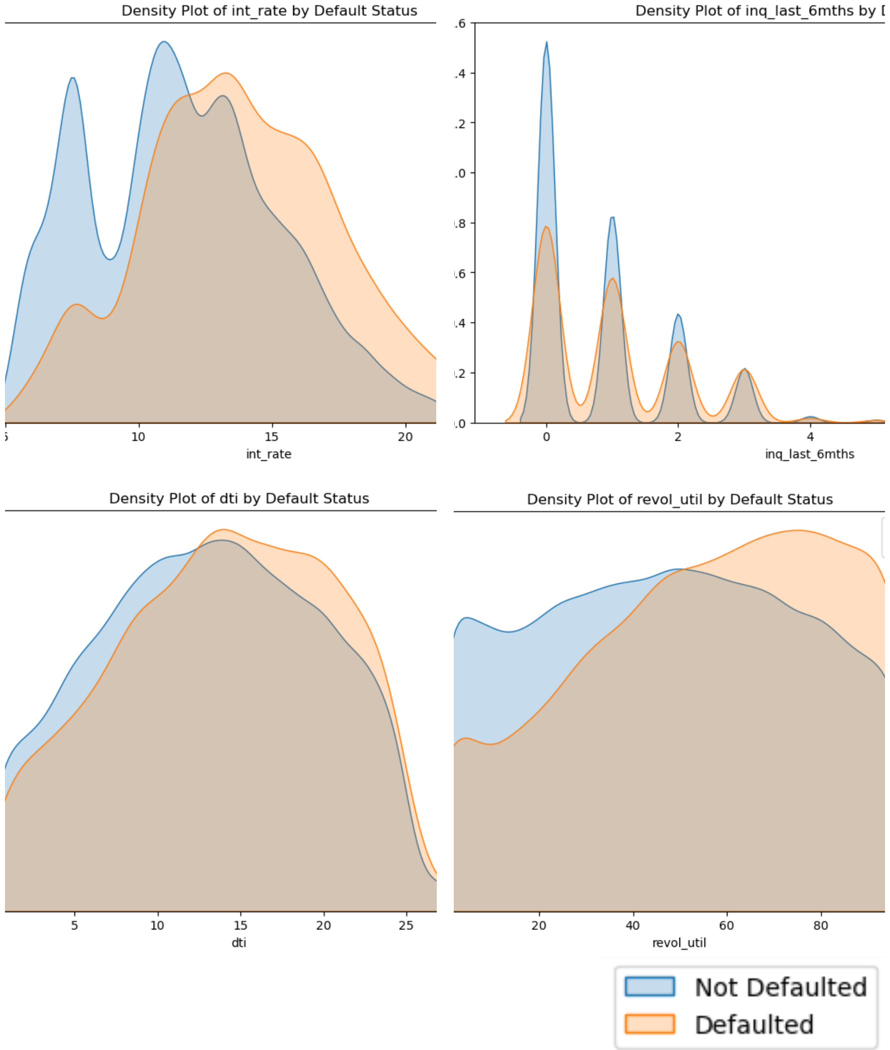
- Data Overview
  - Given data consist of 39717 customer data with 111 columns relating to the customer and loan data. Through data cleaning 34 columns were shortlisted for further analysis. The data formats were fixed and few missing data (very minimal%) were imputed to create a perfect record for analysis.
- Analysis Approach
  - Univariate Analysis: Analysis was done with various Numerical (Distribution and Boxplot) & Categorical columns (Count plot).
  - Bivariate analysis: Analysis was done with various Numerical & Categorical columns against a derived column called 'charged-off' (created from column '***loan_status'*** with value 1 = default and 0 – otherwise. Density plot analysis for numerical column gave new insights that were missing from the other visualizations.
  - Multivariate analysis – Correlation matrix heat map was used to find linear correlationships.

Keyfindings

The Lending Club case study found many relations and dependencies that may lead to a loan default. Following are the key metrics affecting the loan default.

Strong Indicators:

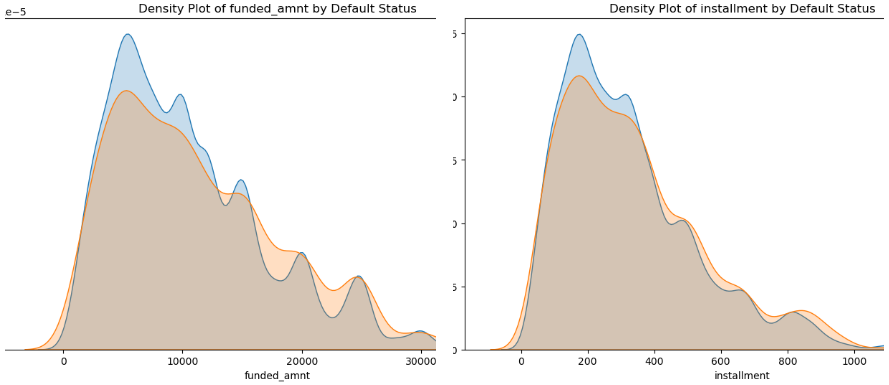
- Debt-to-Income Ratio (DTI): A higher DTI ratio significantly correlates with default risk. Borrowers with a higher percentage of debt relative to their income are more likely to default according to the data.
- Revolving Utilization: High credit utilization rates and balances are strongly associated with default. Borrowers who utilize a large portion of their available credit tend to have higher default rates.
- Inquiries in Last 6 Months: Credit seeking behavior with frequent recent inquiries is an early indicator of financial distress, potentially leading to defaults.
- Interest rates: higher interest rates are more prone to default

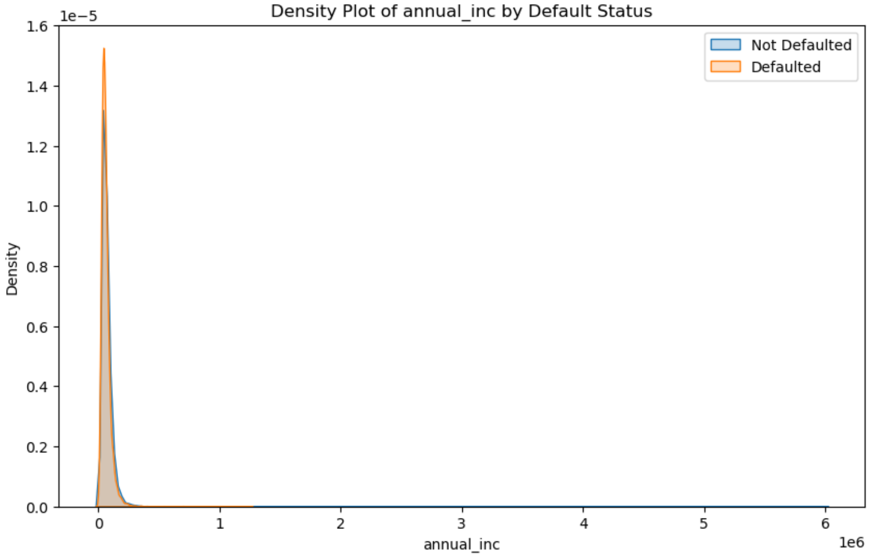


Keyfindings

Moderate Indicators:

- Installment Size: Though broadly similar across defaulters and non-defaulters, variations in installment sizes at extreme values could have a moderating effect on default likelihood, particularly at very high or low ranges.
- Higher funded amount and loan amounts (more than ~17K) are likely to default compared to lower amounts
- Annual Income: While the income levels are similar between defaulters and non-defaulters, few data indicates that higher income individual tend to close the loan on time.





Data cleaning Assumptions

- As a part of data cleaning, columns with all null values (54 nos) were removed.
- As a part of data cleaning, columns with null values were either removed or taken assumptions
- Few irrelevant columns were removed for the analysis

Sl No.	Column	Null%	Action taken	Remarks
1	next_pymnt_d	97.12969	Columns Deleted	Majority null values and irrelevant data to the study
2	mths_since_last_record	92.98537	Columns Deleted	Majority null values and irrelevant data to the study
3	mths_since_last_delinq	64.66249	Columns Deleted	Majority null values. Although the data seems to be relevant to the study, due to majority null values, and assumptions to fill the data may be prone to error.
4	desc	32.58051	Columns Deleted	Majority null values. Redundant column ‘purpose’ already exist
5	emp_title	6.191303	Columns Deleted	Irrelevant data, since the Employer further details are not mentioned like the Employer credibility status
6	emp_length	2.70665	Imputation	Important data, missing values can be imputed with the mode value
7	pub_rec_bankruptcies	1.754916	Imputation	Important data, missing values can be imputed with the median value
8	last_pymnt_d	0.178765	No Action	Irrelevant for Analysis
9	chargeoff_within_12_mths	0.140998	Imputation	Missing values can be imputed with the median value
10	collections_12_mths_ex_med	0.140998	Imputation	Missing values can be imputed with the median value
11	revol_util	0.125891	Imputation	Missing values can be imputed with the median value
12	tax_liens	0.098195	Imputation	Missing values can be imputed with the median value
13	title	0.027696	No Action	Irrelevant for Analysis
14	last_credit_pull_d	0.005036	No Action	Irrelevant for Analysis

Univariate Analysis

Numerical columns

- interest rates range from 5.5% to 24.5% with most common interest offered is 10-11%.
- open_acc - While most borrowers manage few lines of credit, there are a few with an unusually high number of open accounts (going until 40) indicating a potential default risk.
- pub_rec - Few borrowers have public records which may pose a higher default risk, which may need detailed analysis.
- total_acc – most borrowers have 20 and 40 credit lines, but few accounts with higher number indicating a potential credit overextension which might increase in default risk.
- annual_inc – the data contains a significant number of high-income outliers, but most borrowers are clustered at lower end of the scale. There could be a possibility that these data might skew the average income calculation, but it is decided to retain the outliers because of the understanding that a loan defaulting is mainly due to poor money management and this is equally applicable to both the low income and high income earners. However a new derived column with a high income and low income data might give some new insights.
- dti – most borrowers have a particular level of debt relative to their income, which might be typical for a consumer lending market. Higher dti values reaching to 30% indicates increased financial stress indicating a potential default risk. In boxplot analysis, the data is symmetrically distributed around the median which typically reflects a balanced situation in the borrower pool
- revol_bal- while most borrowers manage smaller amounts of revolving debt, a few have very high balances, signalling higher financial risk or greater reliance on credit.
- revol_util – shows a normal distribution with utilization rate peaking around 50%-60%, meaning many borrowers are using about half of their available revolving credit, with a few utilizing upto 100% which can be a risk for default. Majority of the data points are falling between 20% to 80% utilization.
- During the analysis of boxplot, most of the boxplots showed many outliers and it is decided to keep these data rather than removing. The reason behind is that, during the analysis of loan status, only around 3000 accounts were charged off, which accounts to 10% of the available data. Since our objective is to analyse the metric of loan default, it is decided to keep the data as these data might provide valuable insights into the behaviour of borrowers who are more likely to default.

Univariate Analysis

Categorical columns

- term – shows a clear preference for 36 month loan over 60 months among the borrowers
- grade - Grade A & Grade B loans are the most common with B3, B4, B5, A4, & A5 being the highest demand. Lowest demand is F&G graded loans
- emp_length – employees with 10 + years of employment contributes to most data, which is an indicator of lower credit risk due to employment stability. This in turn indicates a good health to the borrower pool due to lower credit risk. There are few borrowers that are less than 1 year of employment leading to a higher risk for loan default.
- home_ownership – Borrowers that have rented their house or taken a mortgage are the highest numbers. The ones that have rented the house may increase the credit risk to the pool.
- verification_status – There are significant number of borrowers that are unverified leading to a higher risk to the lenders.
- loan_status – data shows that the majority share has cleared their loan, while few percent of the borrowers has a charged off status, indicating the importance of risk assessment before granting loan. Analysing the characteristics of charged-off loans can provide insights into the risk factors leading to defaulting a loan.
- purpose – the most common purpose for loan is debt consolidation.
- addr_state – states like CA, NY, TX have the highest counts probably because of the population size of these states.
- issue_d – the number of loans issued over time is peaking in 2011 indicating a steady trend and demand from the past years. The recent demand also highlights the importance of such study to identify risk profile.

Bivariate Analysis

Customer profile

- Dti – Borrowers with Debt-to-income ratio more than 12 are more prone to default.
- Revol_util – higher percentages of revolving line utilization rate are likely to create defaulting of loan
- Annual_inc – non-defaulted has more outliers and high income categories indicating high income owners are not likely to default.
- Delinq_2yrs – the distribution doesnot distinguish between the charged off and non default situation, indicating that it is not a strong factor or may be it needs to combine with other factors for analysis
- Inq_last_6mths – higher median and more variability in defaulted loans suggest that , frequent credit seeking mentality is strongly associated with defaulting or may be it is indicating the stress level of the borrower prior to default
- Pub_rec_bankruptcies – higher counts of public record bankruptcies are more common in defaulted loan
- Employment length – borrowers having shorter tenure (especially less than 1 year) with the company seem to show higher rates of default.
- Purpose - Loans taken for small business and educational purposes show higher default rates. These types of loans might carry higher risks probably due to the less stable financial conditions
- Grade – lower grades (C to G) that are riskier naturally exhibit higher default rates
- Total_acc – boorwers that have significantly higher (60+nos) are more likely to default

Bivariate Analysis

Loan metric

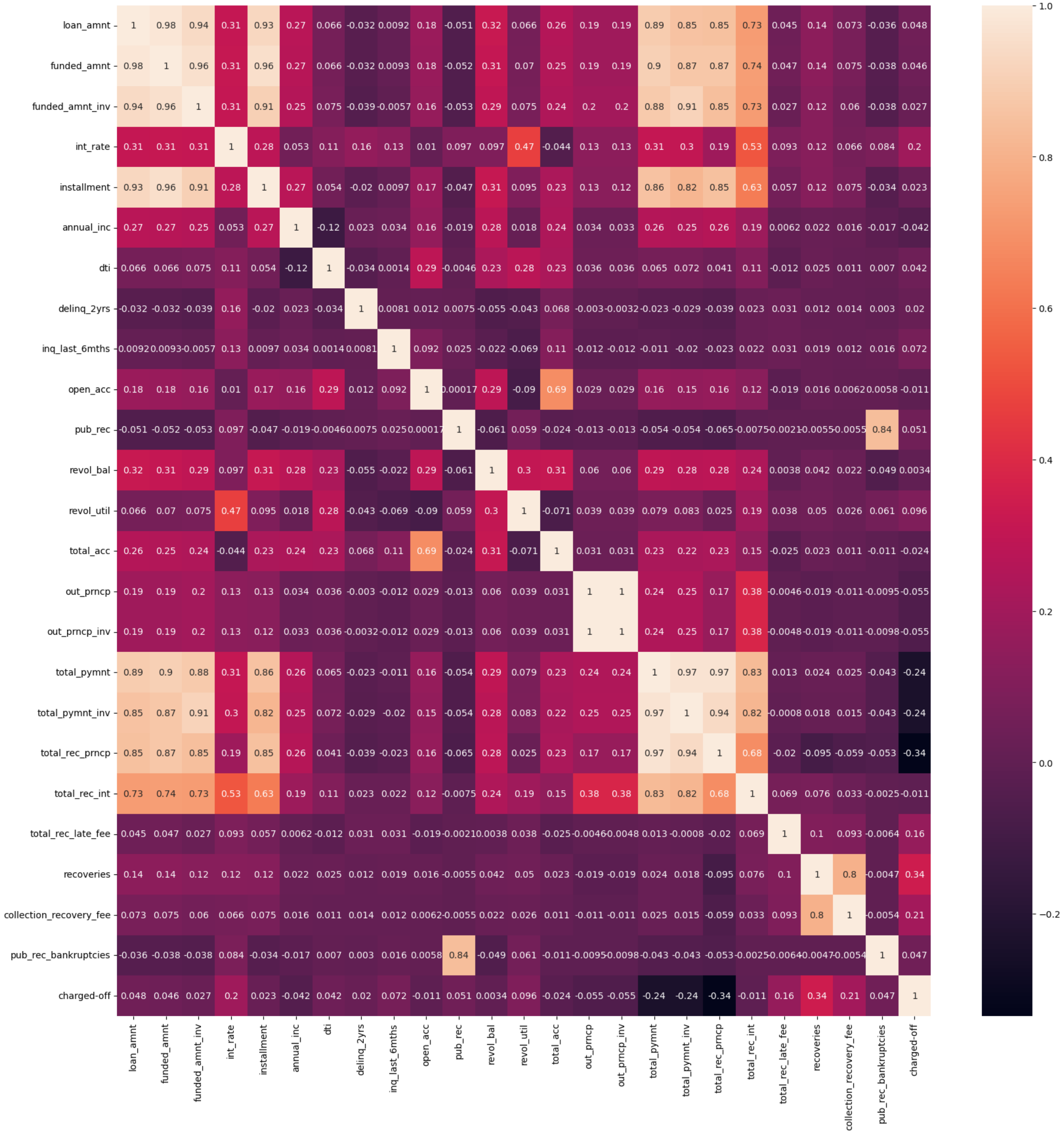
- Int_rate – loans that have defaulted have higher interest rates indicating a strong correlation between the defaulting and the interest rate. Higher interest rate on a high risk profile may lead to defaulting of loan.
- Funded amount & loan amount – higher funded amount and loan amounts (more than ~17K) are likely to default compared to lower amounts
- Installment – higher installment amounts in defaulted loans is a critical indicator of default.
- Total payment – significantly lower payments in defaulted loan suggest a strong link between payment amounts and default status.
- Term – longer durations like 60 months tend to default than shorter 36 month duration

General Observation:

Surprisingly, several factors which were considered as key metric for a loan default didn’t gave strong correlation for the default. Example are Delinquencies in Past 2 Years, Public Records & Bankruptcies , Open Accounts , Total Accounts.

Multivariate Analysis

- For the multivariate analysis, correlation matrix was used. However, the observed linear relationships were obvious like loan default (charged-off) is negatively correlated to total payment. A strong positive correlation (0.93) for loan amounts and installment payments, which is normal. There is very little linear relationship (0.066) between the debt-to-income ratio of borrowers and the loan amount they receive etc. Interest rates show a moderate negative correlation with funded amounts, suggesting that higher loan amounts typically attract lower interest rates.
- In summary, the loan default has much complex relation to other variables, that a direct linear correlation is difficult to establish.



	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	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	pub_rec_bankruptcies	charged-off
loan_amnt	1	0.98	0.94	0.31	0.93	0.27	0.066	-0.032	0.0092	0.18	-0.051	0.32	0.066	0.26	0.19	0.19	0.89	0.85	0.85	0.73	0.045	0.14	0.073	-0.036	0.048
funded_amnt	0.98	1	0.96	0.31	0.96	0.27	0.066	-0.032	0.0093	0.18	-0.052	0.31	0.07	0.25	0.19	0.19	0.9	0.87	0.87	0.74	0.047	0.14	0.075	-0.038	0.046
funded_amnt_inv	0.94	0.96	1	0.31	0.91	0.25	0.075	-0.039	-0.0057	0.16	-0.053	0.29	0.075	0.24	0.2	0.2	0.88	0.91	0.85	0.73	0.027	0.12	0.06	-0.038	0.027
int_rate	0.31	0.31	0.31	1	0.28	0.053	0.11	0.16	0.13	0.01	0.097	0.097	0.47	-0.044	0.13	0.13	0.31	0.3	0.19	0.53	0.093	0.12	0.066	0.084	0.2
installment	0.93	0.96	0.91	0.28	1	0.27	0.054	-0.02	0.0097	0.17	-0.047	0.31	0.095	0.23	0.13	0.12	0.86	0.82	0.85	0.63	0.057	0.12	0.075	-0.034	0.023
annual_inc	0.27	0.27	0.25	0.053	0.27	1	-0.12	0.023	0.034	0.16	-0.019	0.28	0.018	0.24	0.034	0.033	0.26	0.25	0.26	0.19	0.0062	0.022	0.016	-0.017	-0.042
dti	0.066	0.066	0.075	0.11	0.054	-0.12	1	-0.034	0.0014	0.29	-0.0046	0.23	0.28	0.23	0.036	0.036	0.065	0.072	0.041	0.11	-0.012	0.025	0.011	0.007	0.042
delinq_2yrs	-0.032	-0.032	-0.039	0.16	-0.02	0.023	-0.034	1	0.0081	0.012	0.0075	-0.055	-0.043	0.068	-0.003	-0.0032	-0.023	-0.029	-0.039	0.023	0.031	0.012	0.014	0.003	0.02
inq_last_6mths	0.0092	0.0093	-0.0057	0.13	0.0097	0.034	0.0014	0.0081	1	0.092	0.025	-0.022	-0.069	0.11	-0.012	-0.012	-0.011	-0.02	-0.023	0.022	0.031	0.019	0.012	0.016	0.072
open_acc	0.18	0.18	0.16	0.01	0.17	0.16	0.29	0.012	0.092	1	0.00017	0.29	-0.09	0.69	0.029	0.029	0.16	0.15	0.16	0.12	-0.019	0.016	0.0062	0.0058	-0.011
pub_rec	-0.051	-0.052	-0.053	0.097	-0.047	-0.019	-0.0046	0.0075	0.025	0.00017	1	-0.061	0.059	-0.024	-0.013	-0.013	-0.054	-0.054	-0.065	-0.0075	-0.0021	-0.0055	-0.0055	0.84	0.051
revol_bal	0.32	0.31	0.29	0.097	0.31	0.28	0.23	-0.055	-0.022	0.29	-0.061	1	0.3	0.31	0.06	0.06	0.29	0.28	0.28	0.24	0.0038	0.042	0.022	-0.049	0.0034
revol_util	0.066	0.07	0.075	0.47	0.095	0.018	0.28	-0.043	-0.069	-0.09	0.059	0.3	1	-0.071	0.039	0.039	0.079	0.083	0.025	0.19	0.038	0.05	0.026	0.061	0.096
total_acc	0.26	0.25	0.24	-0.044	0.23	0.24	0.23	0.068	0.11	0.69	-0.024	0.31	-0.071	1	0.031	0.031	0.23	0.22	0.23	0.15	-0.025	0.023	0.011	-0.011	-0.024
out_prncp	0.19	0.19	0.2	0.13	0.13	0.034	0.036	-0.003	-0.012	0.029	-0.013	0.06	0.039	0.031	1	1	0.24	0.25	0.17	0.38	-0.0046	-0.019	-0.011	-0.0095	-0.055
out_prncp_inv	0.19	0.19	0.2	0.13	0.12	0.033	0.036	-0.0032	-0.012	0.029	-0.013	0.06	0.039	0.031	1	1	0.24	0.25	0.17	0.38	-0.0048	-0.019	-0.011	-0.0098	-0.055
total_pymnt	0.89	0.9	0.88	0.31	0.86	0.26	0.065	-0.023	-0.011	0.16	-0.054	0.29	0.079	0.23	0.24	0.24	1	0.97	0.97	0.83	0.013	0.024	0.025	-0.043	-0.24
total_pymnt_inv	0.85	0.87	0.91	0.3	0.82	0.25	0.072	-0.029	-0.02	0.15	-0.054	0.28	0.083	0.22	0.25	0.25	0.97	1	0.94	0.82	-0.0008	0.018	0.015	-0.043	-0.24
total_rec_prncp	0.85	0.87	0.85	0.19	0.85	0.26	0.041	-0.039	-0.023	0.16	-0.065	0.28	0.025	0.23	0.17	0.17	0.97	0.94	1	0.68	-0.02	-0.095	-0.059	-0.053	-0.34
total_rec_int	0.73	0.74	0.73	0.53	0.63	0.19	0.11	0.023	0.022	0.12	-0.0075	0.24	0.19	0.15	0.38	0.38	0.83	0.82	0.68	1	0.069	0.076	0.033	-0.0025	-0.011
total_rec_late_fee	0.045	0.047	0.027	0.093	0.057	0.0062	-0.012	0.031	0.031	-0.019	-0.0021	0.0038	0.038	-0.025	-0.0046	-0.0048	0.013	-0.0008	-0.02	0.069	1	0.1	0.093	-0.0064	0.16
recoveries	0.14	0.14	0.12	0.12	0.12	0.022	0.025	0.012	0.019	0.016	-0.0055	0.042	0.05	0.023	-0.019	-0.019	0.024	0.018	-0.095	0.076	0.1	1	0.8	-0.0047	0.34
collection_recovery_fee	0.073	0.075	0.06	0.066	0.075	0.016	0.011	0.014	0.012	0.0062	-0.0055	0.022	0.026	0.011	-0.011	-0.011	0.025	0.015	-0.059	0.033	0.093	0.8	1	-0.0054	0.21
pub_rec_bankruptcies	-0.036	-0.038	-0.038	0.084	-0.034	-0.017	0.007	0.003	0.016	0.0058	0.84	-0.049	0.061	-0.011	-0.0095	-0.0098	-0.043	-0.043	-0.053	-0.0025	-0.0064	-0.0047	-0.0054	1	0.047
charged-off	0.048	0.046	0.027	0.2	0.023	-0.042	0.042	0.02	0.072	-0.011	0.051	0.0034	0.096	-0.024	-0.055	-0.055	-0.24	-0.24	-0.34	-0.011	0.16	0.34	0.21	0.047	1

Recommendations

Following are the recommendations to the Lender/s

1. When considering the loan application:
   - Focus on Credit behavior – Monitor borrowers who exhibit high credit utilization, frequent inquiries, and high debt-to-income ratio
   - Interest – Implementation of higher interest rates for high-risk applicants may not always be a good tactic as the high interest rates are prone to default.
2. Implement Loan Performance Tracking
   - Continuously monitor and assess loan performance indicators like Total payment, Total Principal Received, Total Late Fees etc especially for loans with high interest rates since they tend to default more likely.
3. Better Risk Assessment Model
   - From the study, it is understood that variables affecting Loan default is non linear and some relations might come in to play only under certain conditions. So a better risk assessment model that consider many factors only can reveal the true nature.
4. Credit Counselling
   - Offering financial counseling to high-risk but potential profiles is one way to reduce the loan default

Implications for Risk Analysis

- Variables like total_pymnt, total_rec_prncp, and installment are crucial for understanding the financial performance of loans.
- The high correlations between loan amounts and payment-related variables suggest these are critical in predicting loan performance.
- The relatively low correlation between charged-off and all other variables (generally around 0.04-0.05 with loan amounts) suggests that predicting defaults directly from these individual financial variables may be challenging without more nuanced analysis or additional data such as borrower credit behavior or external economic factors.

THANK YOU