upGrad

#LifeKoKaroLift

Lending Club Case Study: Presentation

Gautam Joshi
Kush Lulla

08-05-2022

upGrad

Course : Machine Learning

Lecture On : Lending Club Case Study

upGrad

Agenda

- Introduction
- Problem Statement
- Approach
- Data Understanding
- Data Cleaning
- Univariate and Bivariate Analysis
- Percentage based bivariate analysis
- Further Analysis
- Observations and Recommendations
- Conclusion

Introduction

upGrad

What is Lending Club?

Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

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

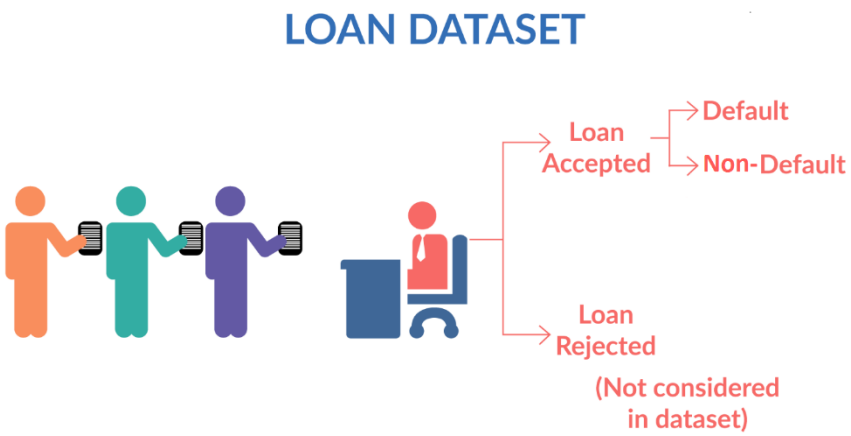
Problem Statement

upGrad

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Problem Statement



Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

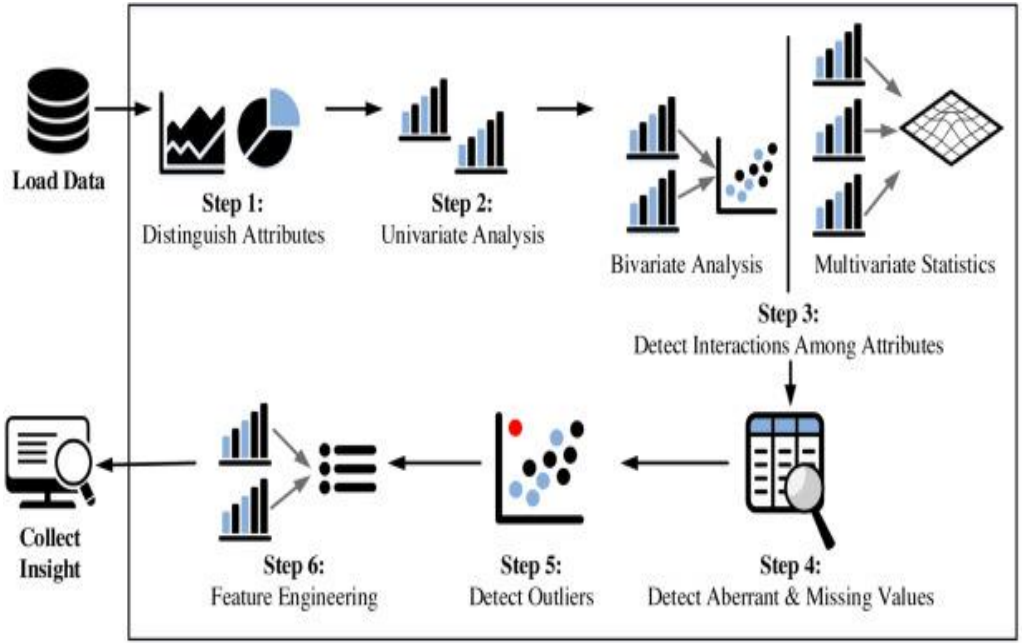
Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

Approach

upGrad

There are four major parts that are needed to be done for this case study:

1. Data understanding
2. Data cleaning (cleaning missing values, removing redundant columns etc.)
3. Data Analysis
4. Recommendations



Lending Club: Approach

upGrad

Our Plan of action

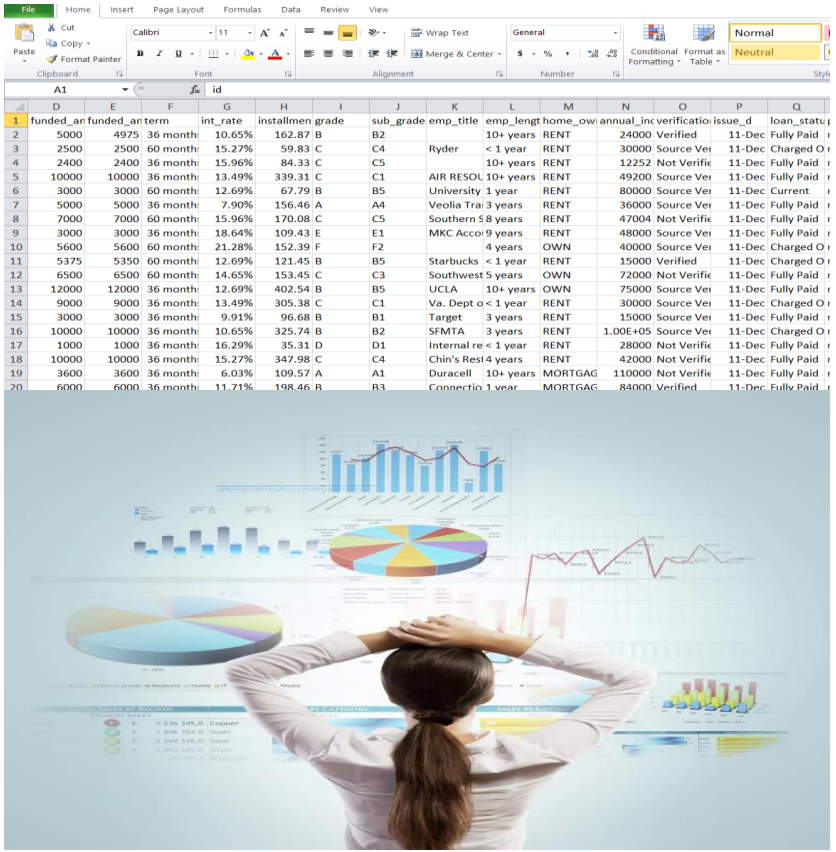
Before we start visualizing and drawing consultations we must:

- Understand the dataset, this involves ,
  - seeing the raw data.
  - finding the columns that might be of interest.
  - seeing the values held by a column.
- Cleaning the data: This would involve ,
  - Dropping columns that dont have any values.
  - Dropping columns that dont add any value to the data set.
  - Standardization of the values in the data set.
  - Replace the null values if that can be done depending on the column.
- Create new columns : We will now create columns that might give us a deeper insight into the data and help create more segmenting attributes.

upGrad

Data Understanding

- Understand the Business problem and expectation from the analysis
- Dataset has a total of 39717 records and 111 columns to begin with.
- We went through the Data dictionary to understand the metadata about the dataset
  - Check the description of each of the variable in Data dictionary, if it makes sense and move ahead
  - If not, then searched the internet to find out more details about that variable
  - Checked the website of the company involved Lending Club
- Visually glanced through the data set to understand the nature of data, null values etc.



Loan Attributes – Important variables	Customer behavior variables (Important, Demography, age, income etc.)	Customer behavior variables (Not important)	Unimportant variables (mostly raw text, indexes, majority null values etc.)
loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, verification_status, issue_d, purpose, loan_status	open_acc, revol_util, total_acc, pub_rec_bankruptcies, emp_length, home_ownership, annual_inc, addr_state, dti, earliest_cr_line, inq_last_6mths, delinq_2yrs, pub_rec	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_d, last_pymnt_amnt, next_pymnt_d, last_credit_pull_d, collections_12_mths_ex_med, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt, tax_liens	Id, member_id, sub_grade, emp_title, url, desc, title, policy_code, application_type pymnt_plan – 1 value zip_code – cant get much out of this mths_since_last_delinq - >60% null mths_since_last_record - >90% null initial_list_status

upGrad

Data Cleaning

- Multiple columns have only null values, we simply deleted those features, after that we were left with 57 columns
- Once we removed the 'all null' columns, we divided the remaining columns into three categories
    - Important variables – Loan amount, term, interest rates , annual income etc.
    - Customer behavior variables (typically generated after the loan is issued)
    - Unimportant variables – URL, title, employee title etc. which has mostly random text data or just 1 unique value
- We can delete the Customer behavior and other unimportant variables
- We calculated the % of null values in the remianing columns and deleted the columns that had more than 60% null values
- We also removed the records for which the loan status was '**Current**' i.e. the loans which were still running, as they couldn't be classified as defaulted loans. So we don't want to consider those loans for our analysis.

* Complete list of features considered for EDA is given in the next page

upGrad

Data Cleaning

- After removing the unwanted rows and columns from the dataset, we moved ahead with cleaning & standardizing the remaining data, it included below steps:
  - Setting up the data types for different columns, float, int, date time etc.
  - Standardize the columns by removing suffixes like '%', 'months'
  - Change the state code to state names for better readability of data
  - Data imputation – for the columns that had relatively low null values we imputed the null values with the mean, median or mode values depending on the type of variable
    - Employee Length with the median employee length
    - Credit utilization with mean credit utilization
    - **Pub record bankruptcies with the mode i.e. maximum occurring value**

New columns

upGrad

Addition of new features for EDA

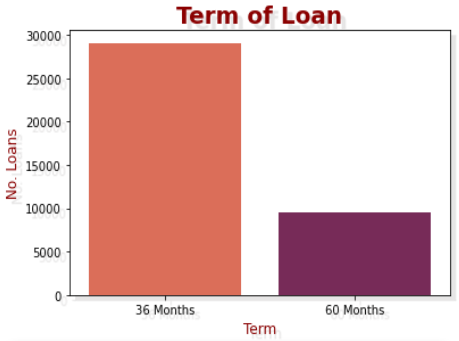
- We also added some new columns to do the analysis
- Issue Date -> Issue month, Issue year
- Loan Income ratio -> Loan amount/Annual income
- Monthly income -> Annual income /12
- Installment to Monthly income
- Credit history – Loan issue date – earliest reported credit line date

We also divided some of the important variables into categories for insights

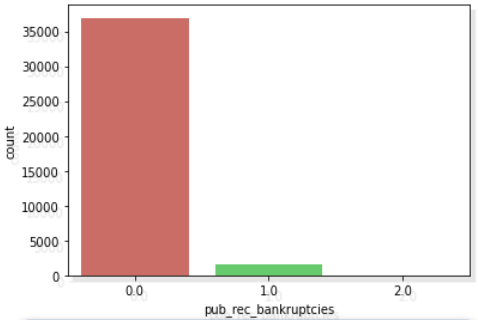
- Annual income category - 3k-30k, 30-60k,60-90k,90-120k,120-150k,150-200k,>200k
- Interest rate category – 5-10%,10-15%,15-20%,20-25%
- Revol_utilisation category – [0-10%,10-20%,….90-100%]
- Loan Amount category – [0-5k,5-10k,….30-35k]

Univariate and Bivariate Analysis

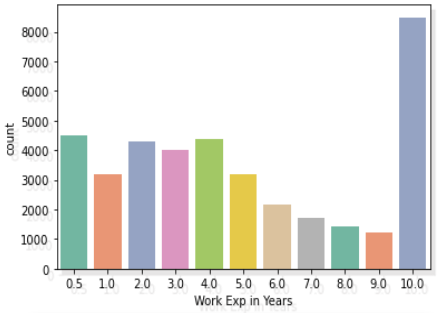
upGrad



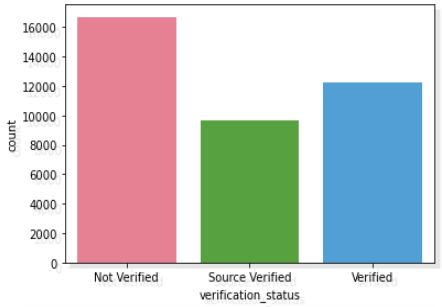
Term of the loan, most Loans are of 3 Years duration



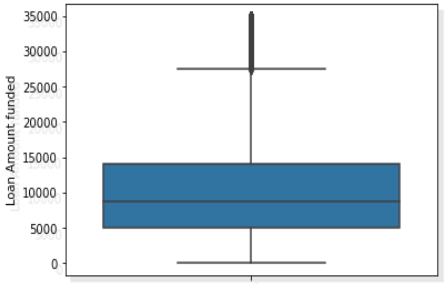
Most of the loan applicants don't have bankruptcy record



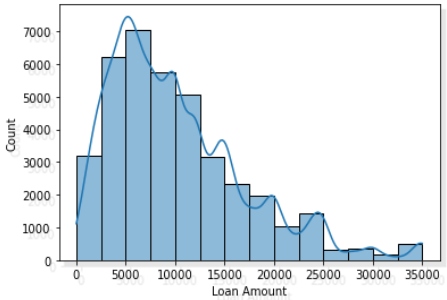
Most of the loan applicants have a 10 year or higher work experience



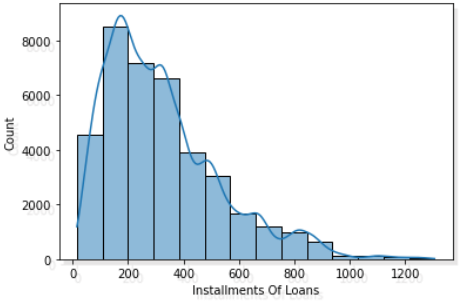
Majority of the loans are unverified



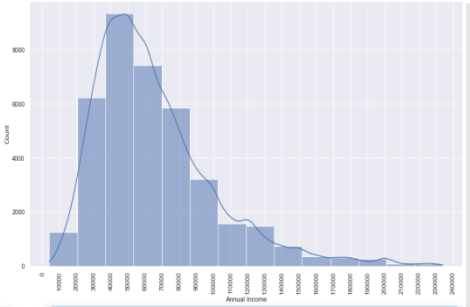
Boxplot of the funded loan amount



Majority of the loans are below $15k



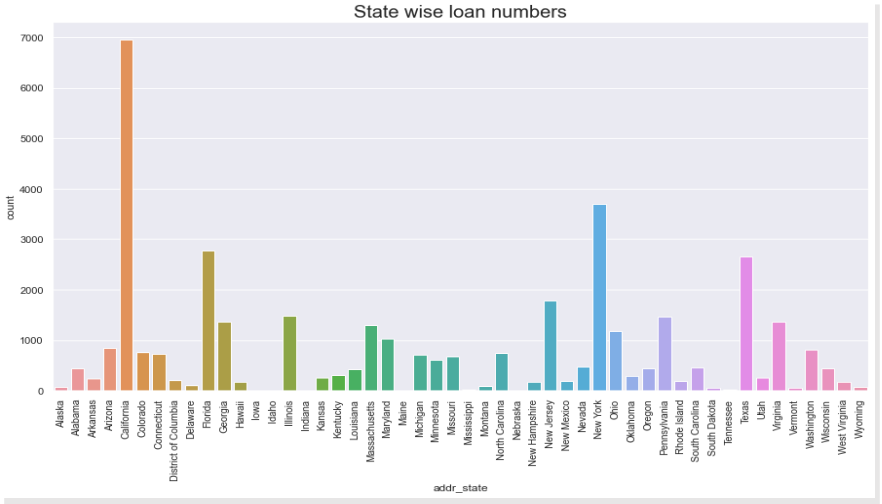
Most of the installment amounts are below $600 per month



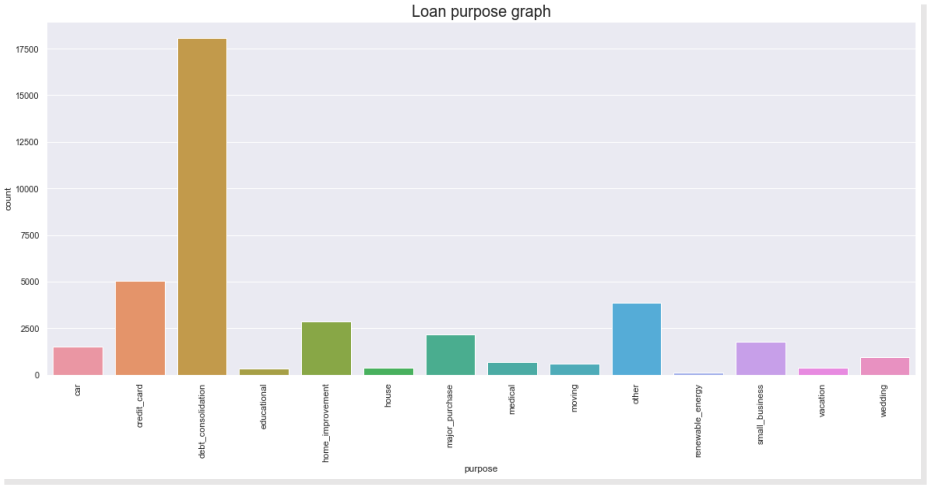
Annual income of the majority of the customers is below 100,000 $

Univariate and Bivariate Analysis

upGrad



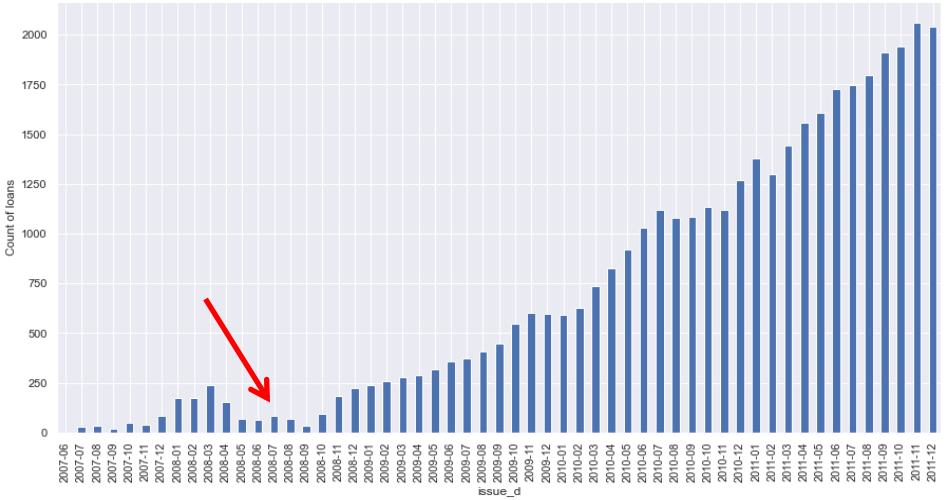
Maximum number of loans are from California (18%) , followed by New York (9.5%) and Texas(6.89%)

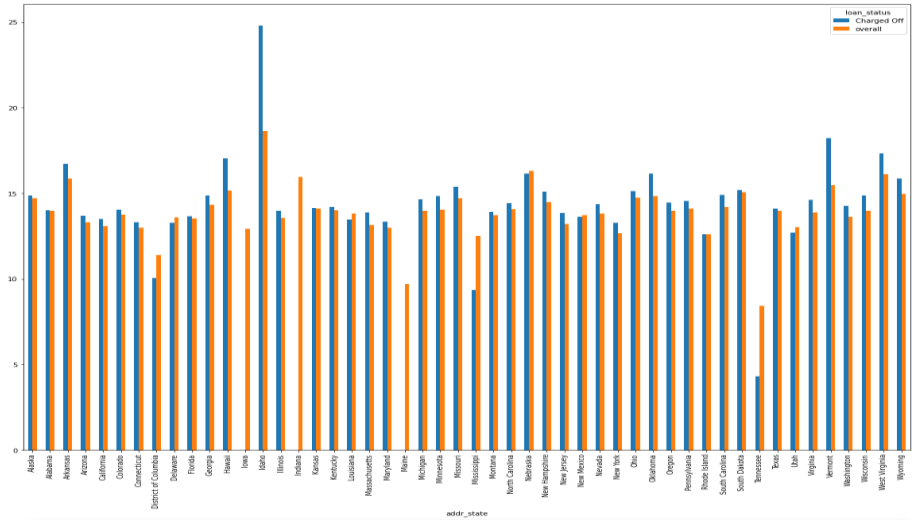


Most of the loans are applied for Deb consolidation (46.8%), followed by Credit card (13%) repayment

Univariate and Bivariate Analysis

upGrad





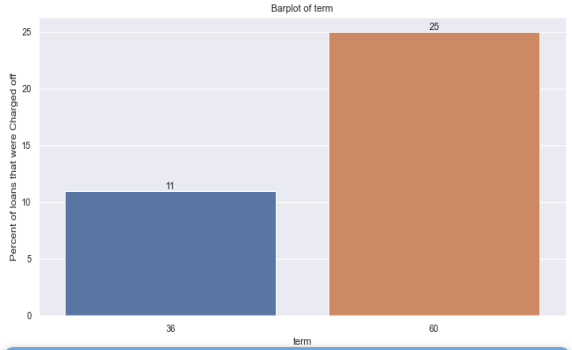
Growth of loans since the start of data, since the beginning till around 2008 March the loans were growing, but then suddenly they started falling sharply and were lowest in Sep 2008. After that it started growing very fast.

What explains the dip around Sep 2008 - > 2008 Financial Crisis

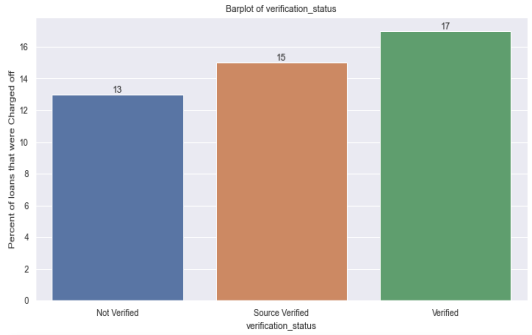
DTI of charged off loans vs Fully paid loans, a general trend is that DTI is higher for charged off loans vs fully paid loans

Percentage based bivariate analysis

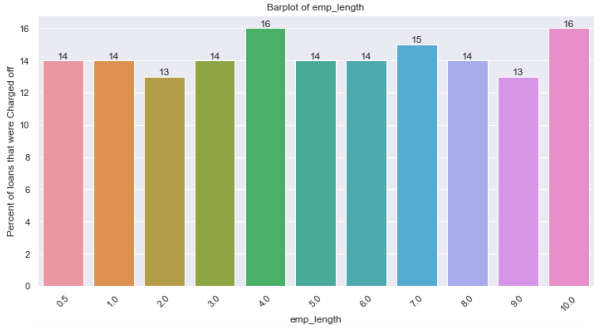
upGrad



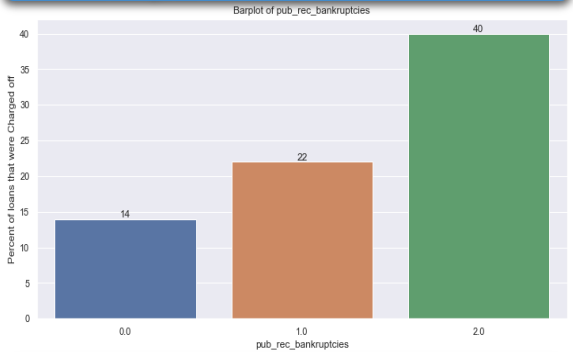
60 month loans charge off rate is 25% while 36 month loans are charge off rate is only 11%



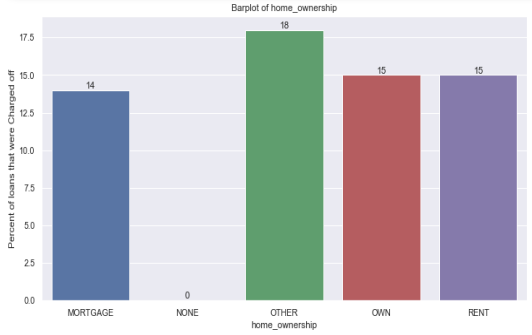
Another worrying trend that the default rates for verified loans is greater than unverified loans



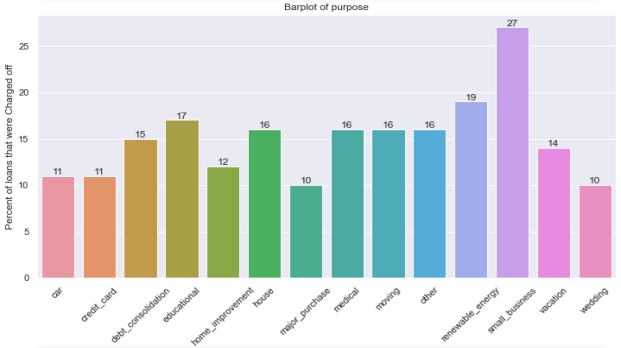
Maximum charge off percentage is for the customers having a work experience of 10+ years



Though the number of customers involved in a bankruptcy is very low, but still for those who have a bankruptcy the Charge off rate is very high.



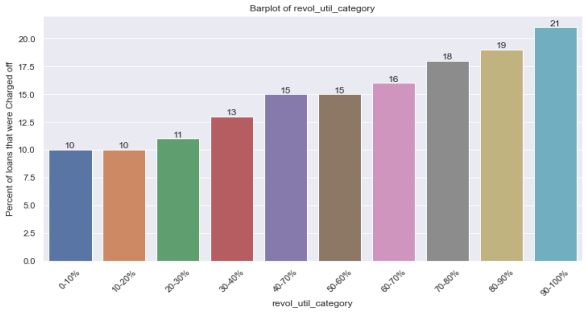
Default rate of an applicant against their home ownership



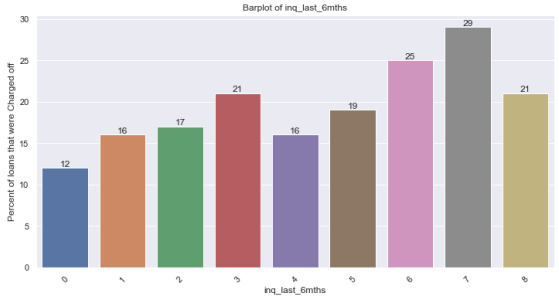
Maximum charge off percentage is for loans taken for small businesses (27%) and renewable energy (19%), while lowest is for wedding and major purchases (at 10% each)

Percentage based bivariate analysis

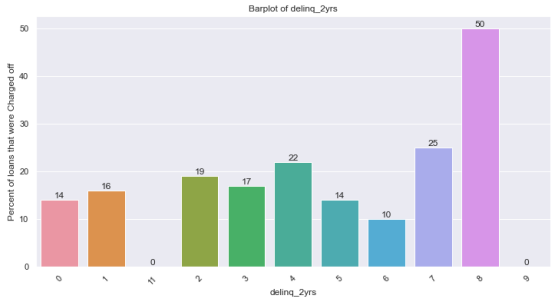
upGrad



Default percentage increases with increase in credit utilization increase



More the number of inquiries in the last 6 months, more the rate of default which is self explanatory i.e. the customer is trying hard to get a loan but isn't getting one

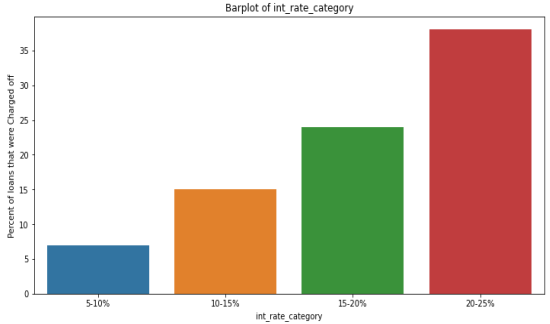


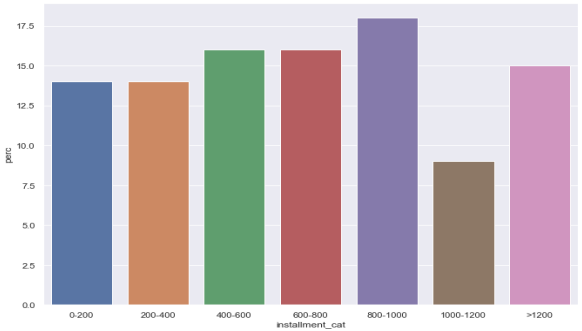
General trend is that if the number of delinquency incidents are higher in the last 2 years, there is a high chance of default.

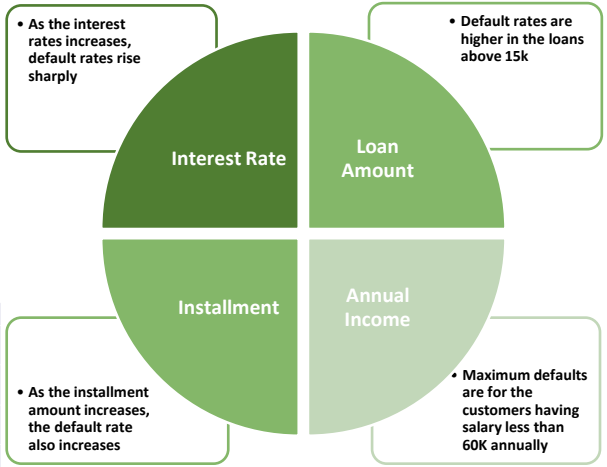
Further Analysis

upGrad

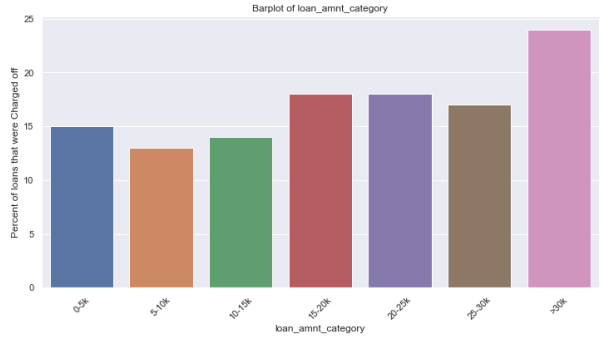
Binning of important variables and analysis of loan status against those bins

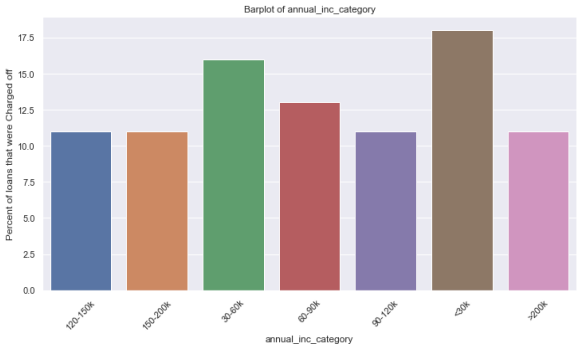






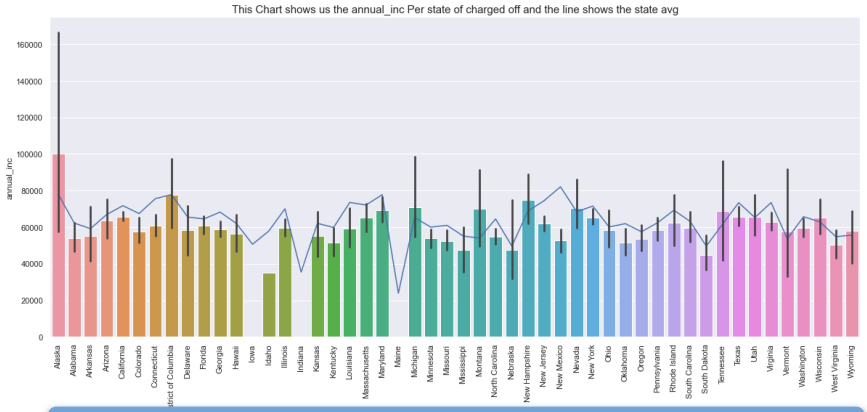
- As the interest rates increases, default rates rise sharply
- Default rates are higher in the loans above 15k
- As the installment amount increases, the default rate also increases
- Maximum defaults are for the customers having salary less than 60K annually



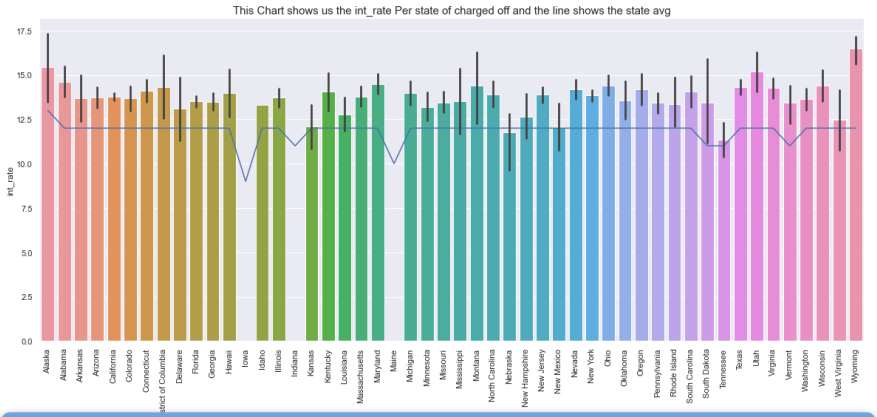


Further Analysis

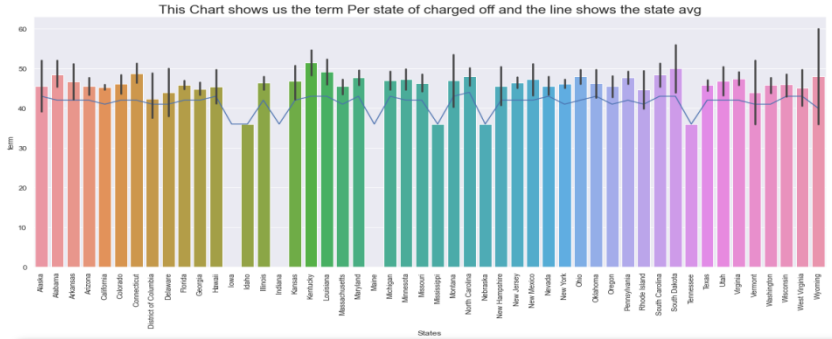
upGrad



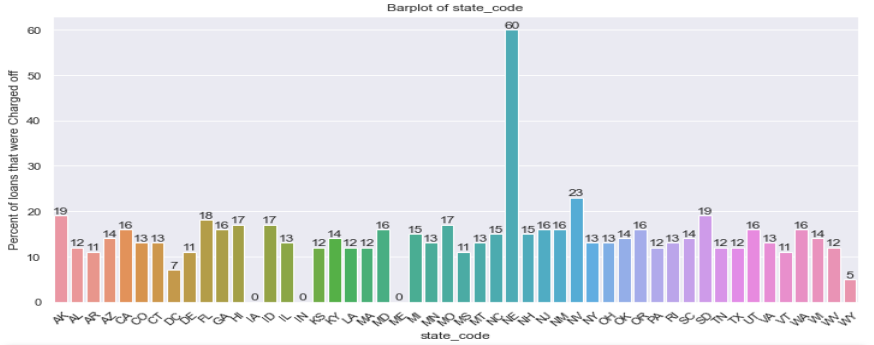
State wise annual income of the state vs charged off loans



State wise interest rates of the state vs charged off loans



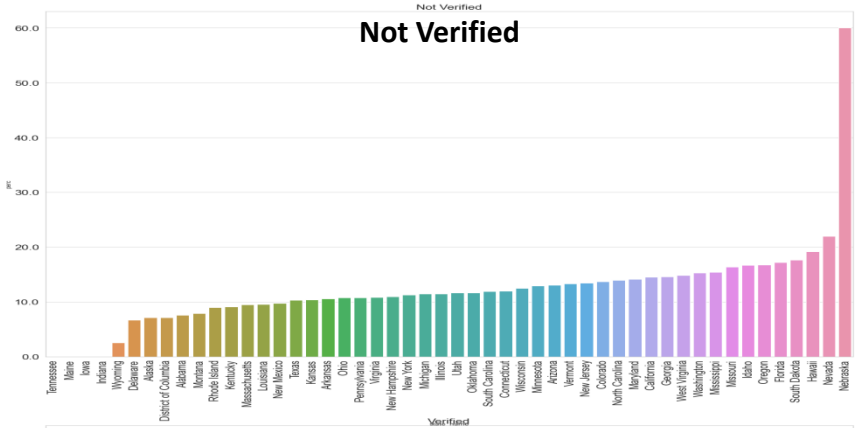
State wise term of loans for the charged off loans vs the state average

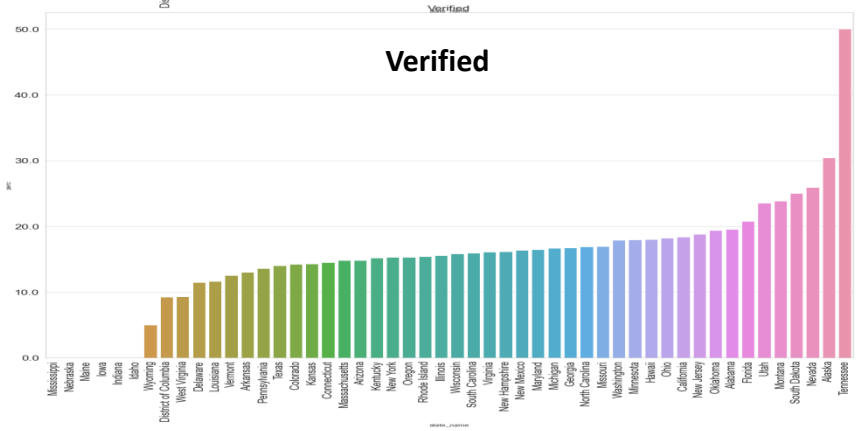


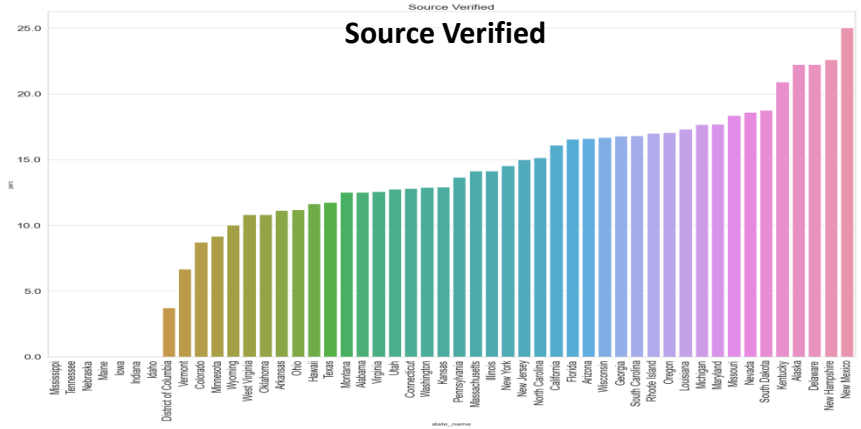
State wise loan default percentages

Further Analysis

upGrad





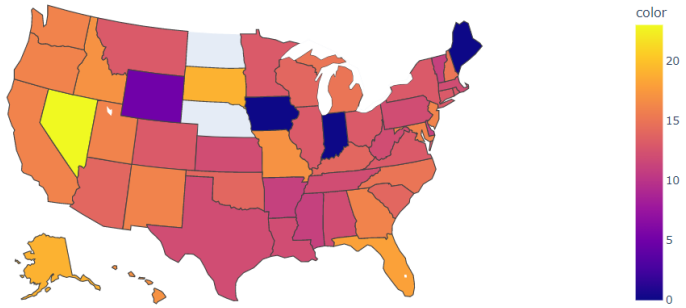


- Unverified loans have a default rate of 12.8%
- Source Verified loans have a default rate of 14.8%
- Verified loans have a default rate of 16.8%
- There is a worrying trend that the default rates of the loans which are verified is higher compared to the non-verified ones. Clearly there are some gaps in the verification process being used in the loan eligibility calculation. Both the 'verified' and 'source verified' categories of the vertification have higher default rates i.e.16% and 14% respectively.

Further Analysis

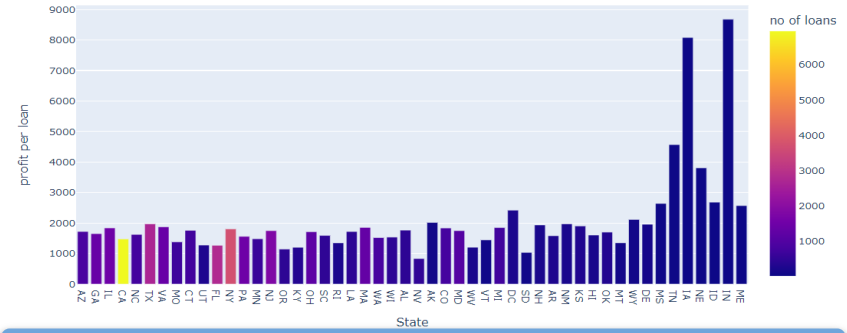
upGrad

State wise default percentage



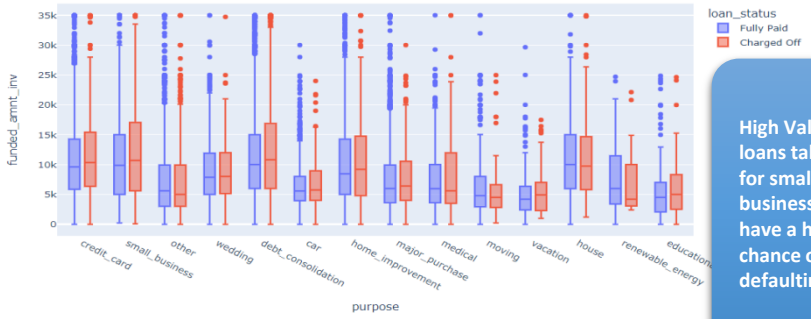
West coast has a high percent of defaults but the difference is marginal

Statewise Profit per loan



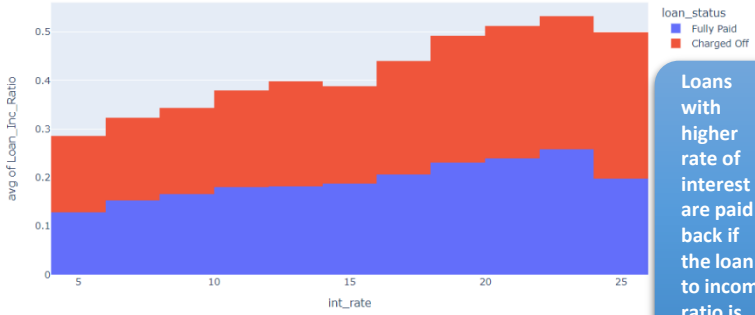
The central part of the country is the most profitable and the west coast is least profitable

Purpose vs loan amount chart



High Value loans taken for small businesses have a high chance of defaulting

Loan income ratio vs interest rates



Loans with higher rate of interest are paid back if the loan to income ratio is low.

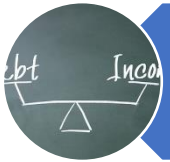
Key Observations

upGrad

We can see that the loan default rates for the Small business, renewable energy and education are very high. These are the riskiest loans as per the data, while categories like wedding, car, credit card and major purchase are least risky loans for the company

There is a rise in charge off percentage for loans above 15K, higher>15k loan amounts are more likely to default

Clearly there is a rise in number of defaults as the revol_util increases, we have assumed that revol_util indicates the utilisation of the current credit line for the currently open loans of the loan applicant

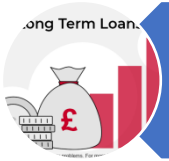


Loans with higher rate of interest are only paid back if the loan to income ratio is low. Meaning income is high and loan value is low . Applicants with low income , large loans and high rate of interest have an extremely high chance of defaulting.

15-20% and 20-25% interest rate category has maximum default rate

Looking at the annual income alone isn't helpful but when we see the same data state wise we see a trend where the annual income below the average of that state tend to default more .

People who file for bankruptcy even once have a higher chance of defaulting on their loans



Loans that were taken for a longer term have a higher percent of default

Recommendations

upGrad

Try to reduce the 60 month loan tenure to 4 years instead of 5 years

Create specific groups to review loans applied for– Small business, renewable energy and education

Need to improve the Source verification and verification process in all the states, as the default rates for verified loans is on the higher side.

Focus on micro loans i.e. <15K loans as they tend to default less.

Customer behavior variables i.e. high dti, high revol_util, frequent credit inquiries in last 6 months, delinquency etc. may be build a model to further strengthen the current verification of these variables to bring down the default rates.

Loans that fall in the higher interest rate category (above 15%), should have a lower loan funded amount.

When approving loans, annual income of the applicant should be compared to the state average.

Conclusion

upGrad

As per our Analysis, we conclude that the main driving factors behind the loan charge offs are as follows:

1. Term of the loan
2. Interest rate applicable on the loan
3. Annual income of the applicant compared to state average
4. Purpose of the loan
5. Loan Amount
6. Credit hungriness of the applicant
7. Bankruptcy and Derogatory records of the applicants
8. Customer behavior aspects like high credit utilization and delinquency tendency.

upGrad

#LifeKoKaroLift

Thank You!