Lending Club Case Study

Presented By:

Pavan Kale
Retheesh Kumar

Data Understanding, Cleaning and Manipulation

- There are 54 columns having all NaN values, dropped those
- Dropped single value column: policy_code, application_type, acc_now_delinq, delinq_amnt, initial_list_status, pymnt_plan
- Dropped other column which has no meaning in analysis or have 0 and NA values only: member_id, url, desc, collections_12_mths_ex_med, chargeoff_within_12_mths, tax_liens, next_pymnt_d
- Drop rows having loan_status=Current as we are interested in loan_status=Charged Off/Fully Paid only
- No duplicate rows in data
- No rows with all NaN values
- Delete rows have emp_length=NA values as this variable impact customer income and cannot fill with average or 0 value
- Change emp_length column from object type to int
- Change int_rate column from object type to int
- Extract month and year from all date columns: issue_d, earliest_cr_line, last_pymnt_d, last_credit_pull_d
- Change revol_util column from object to int
- annual_inc column have outliers, drop rows having value more that 95% percentile
- revol_bal column has large number of outlier, drop rows having value more that 95% percentile
- Add new column ChargedOff, 1 if loan_status=Charged Off else 0

Unordered Categorical Variables - Univariate Analysis

Note: Following analysis is after cleaning the data

- home_ownership
  - Around 50% people are leaving in rented house, and they are buying the first home.
  - Less than 5% of people already have their own home and going for second (or more) buy
- verification_status
  - From all the charged-off borrowers, 45% are not “Not Verified”. Out of those 12.5% are charged off, so verification of borrower is the crucial factor to avoid charged off.
- purpose
  - Around 27.5% of the customer who have taken loan for “small_business” are charged off. So, purpose of loan is the crucial factor to avoid charged off. Out of these around 33% are not verified
  - Around 20% of the customer who have taken loan for “renewable_energy” are charged off, out of those 47% are not verified.
  - Around 50% of the customers have taken loan for “debt_consolidation”
  - So are should be taken while providing loan for small_business and renewable_energy
- addr_state
  - Around 19.6% of the total customers are from state “CA”
  - 20.7% of the total customers from state “NV” are charged of, so special care must be taken while providing loan to this state customers

Ordered Categorical Variables - Univariate Analysis

Note: Following analysis is after cleaning the data

- term
  - Around 75% of the customer take loan for 36 months
  - Around 26% of the customer who have taken load for 60 months are charged off, so longer the duration of loan higher chance of being charged off
- grade
  - Around 30% of the customers are of grade B (B > A > C > D > E > F > G).
  - grade vs frequency of grade (not log of frequency of grade) graph follow 'Power Law Distribution'
  - If we assume A as the highest grade of the customer and G as lowest grade of the customer, lower (E, F, G) the grade of the customer, there is high chance of charged off.
  - Around 37% of the grade G customer are charged off, 32.5% of the grade F customer are charged off and 27% of the grade E customer are charged off
- sub_grade:
  - sub_group have characteristics like grade
- emp_length
  - 21% of the total customers are having 10+ years of employment length
  - 15.5% of 10+ years emp_length are charged off and 4% are still paying loan so some of those may charged off. So, care should be taken while providing loan to high emp_length (age) customers
- delinq_2yrs
  - 10% of customer have failed at least once incidences of delinquency.
  - delinq_2yrs plot follow 'Power law distribution'

Continue…

Ordered Categorical Variables - Univariate Analysis

Note: Following analysis is after cleaning the data

- inq_last_6mths:
  - More than 50% of the customer have at least one inquiry in last 6 months
  - inq_last_6mths plot follow 'Power law distribution'
  - As the number of inquiries increases, number of customers getting charged of increases
    - inq_last_6mths = 7 0r 8 (around 28.5-30% are charged off)
    - inq_last_6mths = 6 (around 21% are charged off)
- pub_rec:
  - Around 5.5% of the customer have derogatory public records
  - pub_rec plot follow 'Power law distribution'
  - If customer has public derogatory record, then there is high chance (around 22%) of charged off (ignore pub_rec=3 or 4 as number of records are very less)
- pub_rec_bankruptcies:
  - 4.5% of the customer have public record of bankruptcies
  - pub_rec_bankruptcies plot follow 'Power law distribution'
  - As the number of public record of bankruptcies increases, chance of charged off increases
    - pub_rec_bankruptcies = 2 (40%)
    - pub_rec_bankruptcies = 1 (22%)

Quantitative/Numeric Variables - Univariate Analysis

Note: Following analysis is after cleaning the data

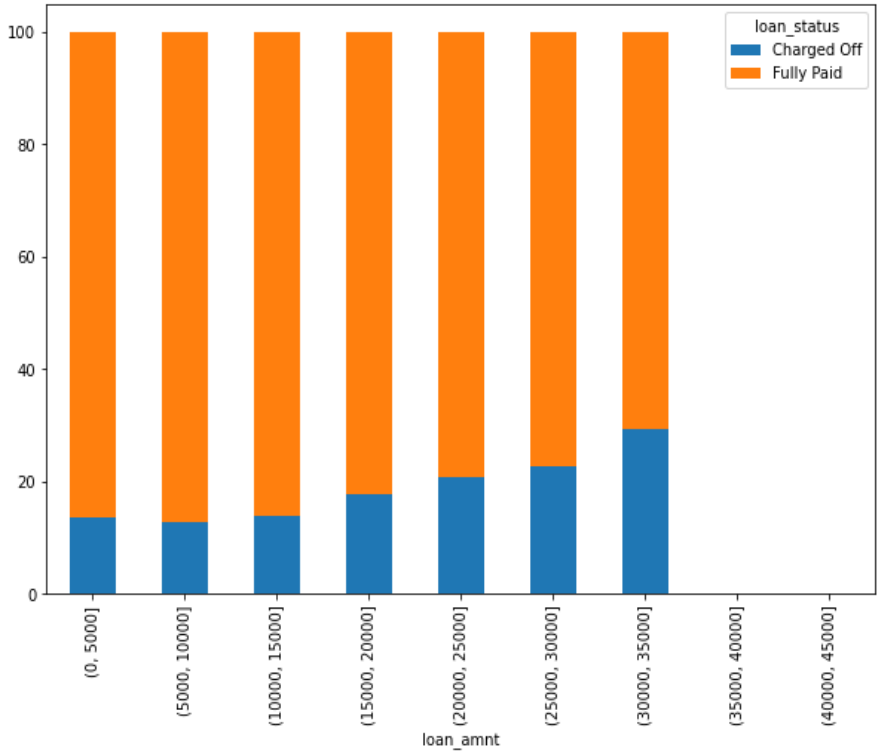
- loan_amnt
  - If we divide the loan amount into bins of 5000 and check loan_status, as loan amount increases, chance of getting charged off increases
  - Around 33% of the customer took loan in between 5000-10000
- int_rate:
  - If we divide the loan amount into bins of 5 and check the loan_status, as interest rate increases, chance of customer getting charged off increases
- annual_inc:
  - Avg income varies from 40000 to 60000
  - Higher the annual income of customer, less chance of getting charged off
- revol_bal and revol_util:
  - Higher the value of revol_bal or revol_util, higher chance of customer getting charged off

Recommendations

- verification_statue: Borrower should be verified before providing loan.
- purpose: If purpose of the loan = small_business or renewable_energy, then special care should be taken while providing loan as there is high chance of getting charged off
- addr_state: If borrower addr_state is NV or FL then there are chances of charged off
- grade: If we consider A being highest and G being lowest grade; lower the grade, high chance of customer getting charged off
- revol_bal and revol_util: higher the value of revol_bal and revol_util for customer, high chance of getting them charged off
- annual_inc: Higher the annual income, less chance of getting charged off
- loan_amnt: Higher the loan amount, high chance of getting charged off
- int_rate: Higher the interest rate, high chance of getting charged off

Loan amount vs Loan status

- As Loan amount ↑, charged off ↑



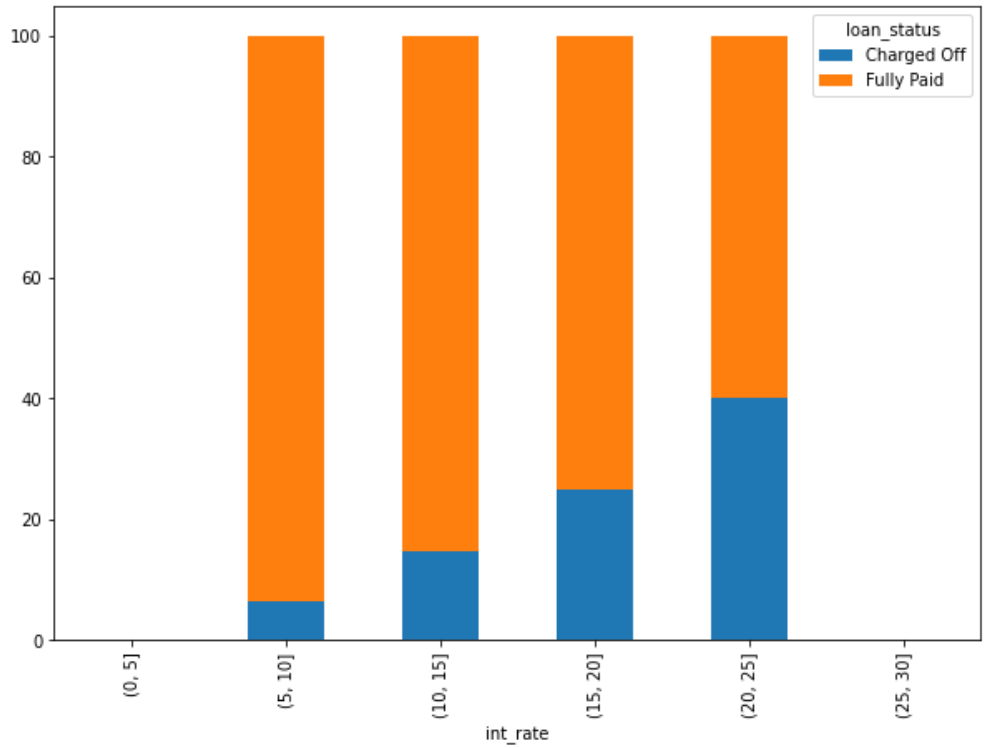
Term vs Loan Status

- As Term increase, chance of charged off increases

term	36	60	All
loan_status			
Charged Off	10.965834	25.733496	14.531554
Fully Paid	89.034166	74.266504	85.468446

Interest rate vs loan status

- As interest rate ↑, charged off ↑



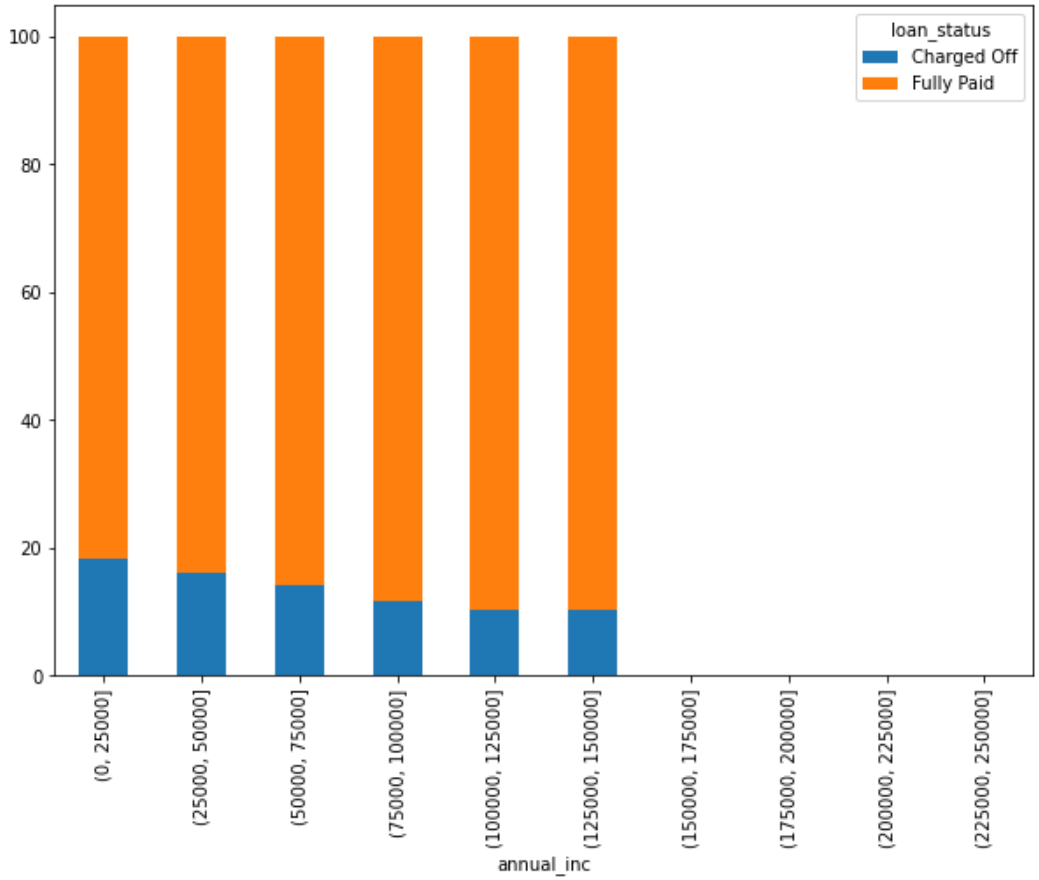
Grade vs loan status

- As Grade ↓, Charged off ↑

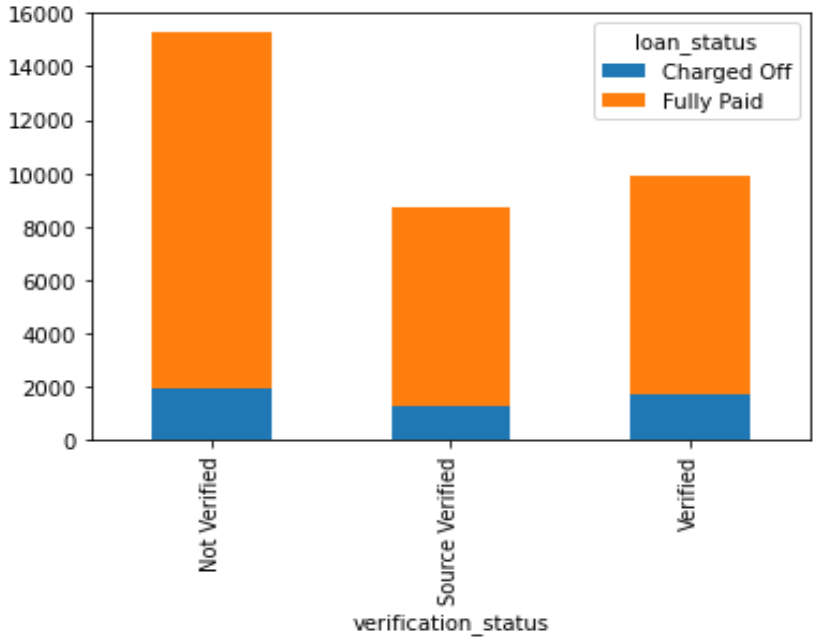
grade	A	B	C	D	E	F	G	All
loan_status								
Charged Off	5.859066	12.129774	17.180996	22.167817	27.120141	32.589839	37.190083	14.531554
Fully Paid	94.140934	87.870226	82.819004	77.832183	72.879859	67.410161	62.809917	85.468446

Annual Income vs Loan status

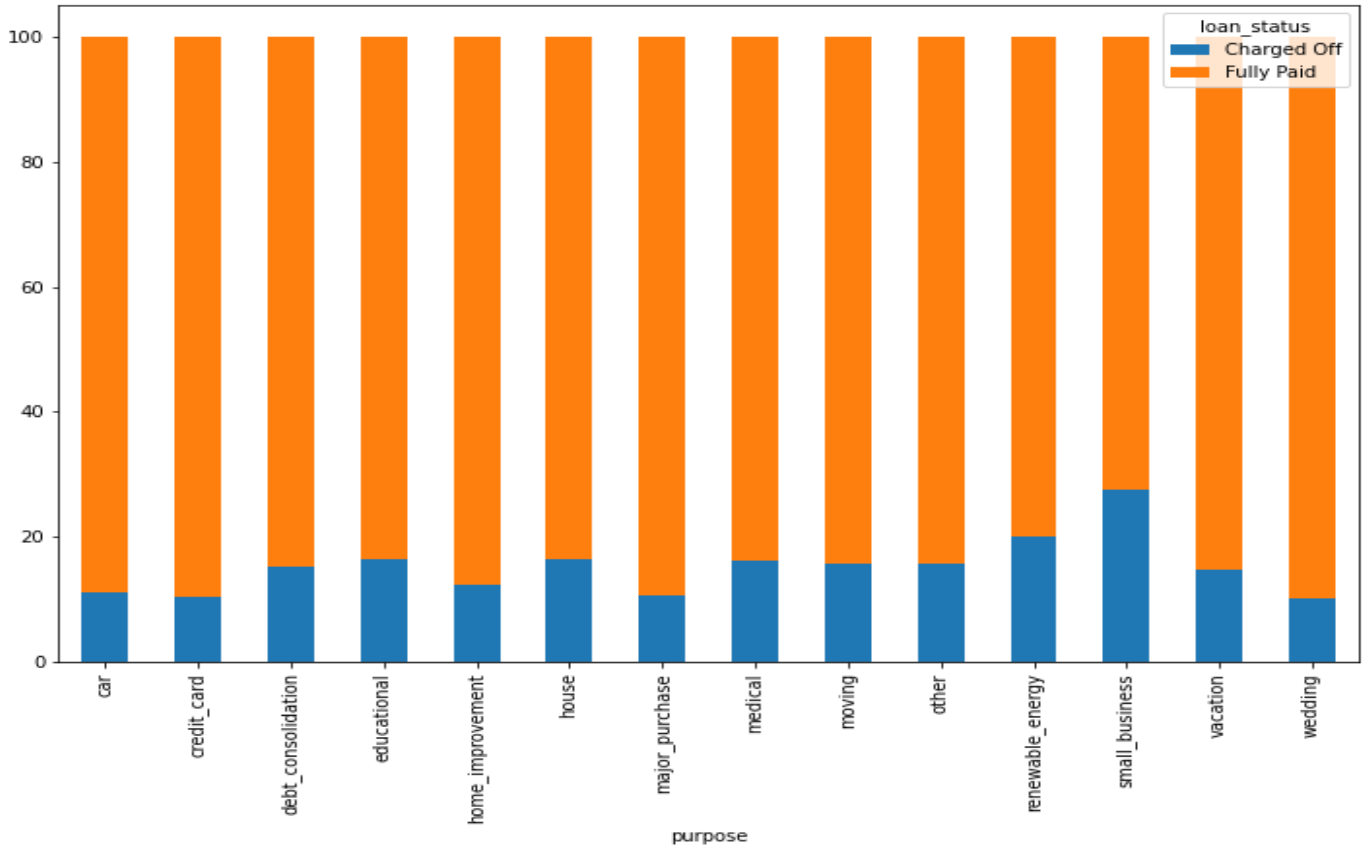
- As annual income ↑, charged off ↓



Verification status vs Loan status



Loan purpose vs Loan status



Thank You