Lending Club – Case study

Saunak MALLIK

IIIT-B/ UPGRAD-ML-C64

23rd June 2024

Objective

- This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

- 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 utilise this knowledge for its portfolio and risk assessment.

Data Understanding

- Load the data into a PANDAS DataFrame
- Perform data exploration
- Perform basic due diligence like shape, info, describe, overall null_checks etc.
- Decide data imputation and data cleansing strategy

Data Cleaning and Manipulation

- Use pandas library to identify NULL values, redundancies, clean data e.g., removal of %, texts etc. in numerical data columns and finally removal of outliers.
- 📋 🛑**Missing value checks**
- Assumptions while taking care of missing values:
- **Any column** having **high % (generally >40%) of missing values** ideally ***should not be considered the part of analysis***, even if it is very relevent column. That's because, the insights coming out from that column might not be reliable.
- 🛑 NULL Imputation strategy; Drop the columns with > 40% missing values from the dataframe in one go

Data Cleaning and Manipulation

- Perform attribute-specific cleansing strategy. E.g., to name a few (refer ipynb file for details)
- 📋 🛑 **column: emp_length (Imputation strategy):**
  - Remove +, < and years
  - Replace rows having null values of emp_length with the MEDIAN(emp_length)
- 📋 🛑 **column: revol_util (Imputation strategy):**
  - Remove %
  - Replace rows having null values of emp_length with the MEDIAN(revol_util)
- 📋 🛑 **column: last_pymnt_d, last_credit_pull_d, issue_d_month (Imputation strategy):**
  - Convert last_pymnt_d to date format
- 📋 🛑 **column: term (Imputation strategy):**
  - Remove 'months' and convert to int
- 📋 🛑 **Create 2 Derived metrics earliest_cr_line_month & earliest_cr_line_year from earliest_cr_line**
- 📋 🛑 **Create a Derived metrics delinquent from loan_status=‘charge off’**

Data analysis - Univariate, Bivariate & Multivariate

📋 Univariate analysis

🪧 🛑**Insights from the distribution plots**

- 1. Loan issuance (issue_d_year): was at the rock bottom during 2007-8 indicating the sub-prime crisis. After 2009, Loan issuance started picking up indicating the resurgence of loan books of banks after the sub-prime crisis was over.
- 2. Bulk of the loans are issued in Q4 (last 3 months of the year). (issue_d_month)
- 3. Smaller loan amounts (<10,000 USD) are more likely to be repaid. Loan Amount (loan_amnt): Larger loan amounts may increase the risk of default if not correlated with the borrower's income and creditworthiness.
- 4. Smaller loan instalments (<500 USD) are more likely to be repaid. (installment)
- 5. Loans are more likely to be repaid if the companies/ individuals haven't declared bankruptcy. (pub_rec_bankruptcies)
- 6. Interest Rate (int_rate): Higher interest rates are often associated with a higher risk of default.
- 7. Annual Income (annual_inc): Higher income levels generally correlate with a lower risk of default.
- 8. Debt-to-Income Ratio (dti): A higher DTI ratio is a significant risk factor for loan default.
- 9. Employment Length (emp_length_cleaned): Longer employment lengths are typically associated with more stable incomes and lower default rates.

Data analysis - Univariate, Bivariate & Multivariate

📋 Univariate analysis ………… continued

🪧 🛑 Insights from the bar plots of categorical variables

- 1. A and B grade loans are more likely to be repaid. As grade diminishes, likelihood of repayment reduces. (grade and sub-grade).
- 2. Those who live on rental are more likely to apply for loan. (home_ownership)
- 3. High proportion of customers' income is not verified. (verification_status)
- 4. High proportion of customers are repaying their loan. (loan_status)

Data analysis - Univariate, Bivariate & Multivariate

📋 Bivariate & Multivariate analysis

🪧 🛑 **Which columns influence Delinquency the most (had been defaulted)**

- **Approach:**
- From above Heatmap and
- create a correlation matrix of df['delinquent’] ** with the other numerical columns of the DataFrame

🪧 🛑 **Insights: Delinquency (had been defaulted) criteria:** based on the **heatmap** and **correlation co-efficient of df['delinquency'],** we find that **Delinquency has very high co-relation with -**

- **recoveries** (corr = 0.340297) - Higher the post charge-off recovery, higher the chance of deliquency.
- **total_rec_prncp** (corr = -0.335019) - lower the loan principal recovery, higher the chances of delinquency
- **total_pymnt** (corr = -0.238844) - lower the loan payments received, higher the chances of delinquency
- **total_pymnt_inv** (corr = -0.236232) - lower the loan funding amount from investors, higher the chances of delinquency
- **last_pymnt_amnt** (corr = -0.214949) - lower the last payment amount received, higher the chances of delinquency

*** df['delinquent’] is a column derived from df[‘status’] = ‘charged off’*

Data analysis - Univariate, Bivariate & Multivariate

📋 Bivariate & Multivariate analysis

🪧 🛑 loan_status Insights -

- 83% of loans are fully paid meaning there's a high proportion of customers repaying their loan, but 14% customers are not repaying their loan.
- 90% of the charged off loans belongs to the grade B - G.

🪧 🛑 Loan term Insights -

- 73% of loans are with with tenor = 36 months i.e., short-term loans.
- 57% of the charged off loans are short-term loans meaning customers are slightly less likely to repay when the loan tenor is smaller.

*** df['delinquent'] is a column derived from df['status'] = 'charged off'*

Conclusions & Recommendations

🪧 🛑 **Loan Amounts:**

- Smaller loans tend to get paid in full more often while higher loan amounts are associated with higher risks.
- Recommendation: Exercise additional due-diligence before approval of larger loan amounts.

🪧 🛑 **Track principal repayments very closely:**

- Likelihood of repayments stopping if one repayment of principal is not done by the customer.
- **Recommendation: Automated system alerts for customers defaulting principal repayment for even 1 cycle.**

🪧 🛑 **Lending cycle:**

- More loans are disbursed in the last Quarter of the year.
- **Recommendation: Run promos in the last quarter of the years to augment the lending book.**

🪧 🛑 **Creditworthiness & Due diligence assessment:**

- Higher grades (A and B) tend to be associated with lower interest rates and lower default rates.
- Customers with a history of bankruptcy tend to default more.
- Customer with high DTI (already having high debt) are likely to default more.
- Recommendation: Lend more to high-grade borrowers. Exercise additional due-diligence before approval of loans to low grade borrowers.
- Avoid customers with history of bankruptcy and high DTI ratio.

🪧 🛑 **Length of Employment:**

- Borrowers with longer employment tend to repay loans better
- Recommendation: Prioritize borrowers with employment lengths of 5+ years.

Conclusions & Recommendations

🪧 🛑 **Existing Home Ownership:**

- Existing Homeowners (with mortgage or owned outright) have lower default rates compared to renters.
- Recommendation: Promo offers of lower interest rates to existing homeowners.

🪧 🛑 **Loan Purpose:**

- Loans purpose is a critical driver for the loan. Each sector has it’s own nuances and cyclicality which needs proper due diligence to be conducted to avoid default.
- Recommendation: Perform strict sectoral due-diligence for each loan purpose.

Thank you