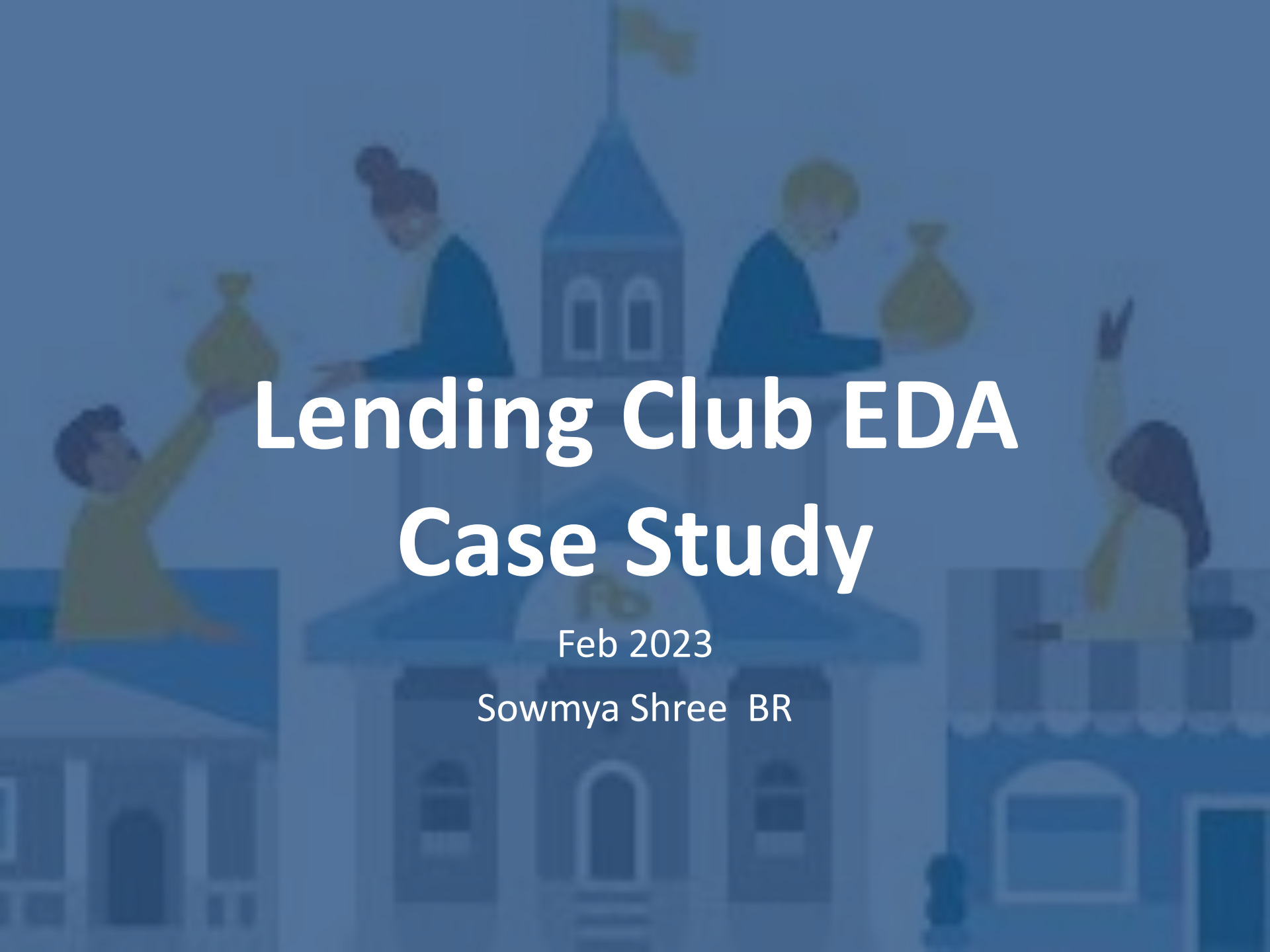
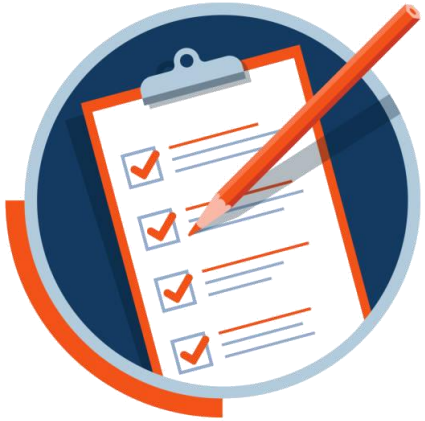


Lending Club EDA Case Study

Feb 2023

Sowmya Shree BR

Agenda



1. Introduction
2. Problem statement & objective
3. Primary goals
4. Approach exploratory data analysis (EDA)
5. Conclusion - recommendations

1 Introduction

Lending Club EDA Case Study

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

It specialises in lending various types of loans to urban customers.

Business understanding

Lending Club company which specialises 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.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
  - **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
- **Loan rejected:** The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

2 Problem statement

- Lending Club company which specialises 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

Business Objectives

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial 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.

3 Primary goal

Reduce the Credit Loss by using EDA techniques to find driving factors for the defaulting of loan.

Associated Risks

- Loss of Business:
  - If the applicant is likely to repay the loan, then rejecting their loan is a loss of business.
  - If the applicant is not likely to repay the loan, then approving loan may lead to financial loss.

4 Approach

Lending Club Company has provided a data set with complete loan data for all loans issued through the time period 2007 t0 2011. Also, provided data dictionary providing business context for the data elements.

Data Understanding

Data Handling

Data visualization & Analysis

Conclusion

Data Understanding

Observations:

- Total # of rows : 39717
- Total # of Columns : 111
- Columns having All null or All 0 values: 64 out of 111
- Loan Terms : 36 and 60 months
- Interest rate: 5.42% - 24.59% based on loan category
- Loan Grades: A,B,C,D,E,F,G

Applicant Income Verification Status

- **Verified:** Verified By LC
- **Source verified:** Verified By Lending Bank
- **Not verified:** Not verified

Loan Status

- **Current:** The applicant who have availed loan and are repaying on-time
- **Charged off:** Indicates the defaulters**.**
- **Fully Paid:** Are the ones who have closed the loan.

Data handling

	Description	Remaining Columns	Remaining Rows
1	Dropped Columns with only (all) NULL Values	57	39,717
2	Dropped Columns with more than 30% of NULL Data	53	39,717
3	Dropped Columns with single constant value like 0,INDIVIDUAL	44	39,717
4	Dropped irrelevant columns which are calculated after loan is approved thus have no relevance to the analysis	39	39,717
5	Excluding rows with loan status = 'Current' as goal is to analyse on default for either fully paid or charged off loans	39	38,577
6	Replaced Missing values in the columns	39	38,577
7	Added Derived Columns (category datatypes) from Numeric, Date and Object Data types.	49	38,577
8	Updated the datatypes for the columns, based on the data content (e.g. Object --> Category, Object --> Numeric, Object --> Date)	49	38,577
9	Checked for Outliers and removed the data.	49	38,577

Data visualization

Visualization Approach Used	Visualization Techniques Used
• Univariate • Segmented Univariate Analysis • Bivariate Analysis	• Box Charts • Bar Charts • Histogram Charts • Pie Charts • Heat Map

Univariate analysis

- Fully paid loans are way more than compared to Charged Off
- **75%** of the loans are sanctioned for the term of 36 Months
- Loan Amount varies from **500 to 35K** and approx. **80%** of the loans are sanctioned for loan amount in range of **500 – 15K**
- **75%** of loans defaulted are in the loan amount range of **5k – 10K.**
- More than **30%** of overall interest between **9-13%**
- Around **50%** of Charged Off loans are in **13-21%** interest rate range
- Loans taken for Debt Consolidation purpose are major defaulters, followed for credit card & others. Same for Charged off loans
- **60 month** terms has higher numbers of defaulters
- Loan applications with **10+ years** of experience are most likely to default loans.

Segmented Univariate Analysis

- 60 month term loans are more chance of defaulting loans, where as 36 month term loans have higher chance of being fully paid.
- Loans availed for purpose of Debt Consolidation have highest number of loans, also have highest number of Fully Paid and Charged Off Loans.

Bivariate Analysis

- Interest Rate for Charged Off Loans is higher than the Fully Paid loans for both 36 months and 60 months term. Indicates that loans with higher interest rate are more likely to be defaulted.
- Charged Off loans are lower for Grade A, the defaulted loans steadily increase as we move from higher grades to lower grades.
- Higher Grades have lower loan amount and interest rates than lower grades, this indicates that lower the grade higher the risk and hence higher interest rates. Also, the fact that lower grades have higher loan amount sanctioned.
- Even though the loan applicants whose Source of Income is not verified is higher, the defaulted loans are higher for loan applicants whose source of income is verified.
- Higher Interest Rates for Charged Off Loans
- Fully Paid Loans have significantly lower interest Rates than the Charged Off loans

Data visualization

Visualization Approach Used	Visualization Techniques Used
• Univariate • Segmented Univariate Analysis • Bivariate Analysis	• Box Charts • Bar Charts • Histogram Charts • Pie Charts • Heat Map

Technology used

Jupyter notebook	Anaconda version 3
Numpy	version 1.21.5
Pandas	version 1.5.3
matplotlib.pyplot	version 3.5.2
Seaborn	version 0.11.2
Python	version 2.7.18

5 Recommendations

The probability of defaulting is high when:

- Loan Applicants in lower Grades seeking loan amount > 15K and interest rate > 10%.
- Loan Applicants not owing home (Mortgage or Rent) and have high Annual Income Range (60K - 70K)
- Loans with interest rates between 9% - 17%
- Loan Applicants with >10 years of experience and with loan amount > 10K or with interest rate (>10%)

Driving Factors for Loan Defaulters:

- Grades, Annual Income, DTI , Purpose of Loan, Loan Amount, Interest Rate

Credits

Respective documentations sites for numpy, panda, matplotlib, and python.

https://stackoverflow.com/


THANK YOU!