LENDING CLUB EDA CASE STUDY

SHASHANK PAWASKAR
UPGRAD STUDENT
FEB 2023

AGENDA

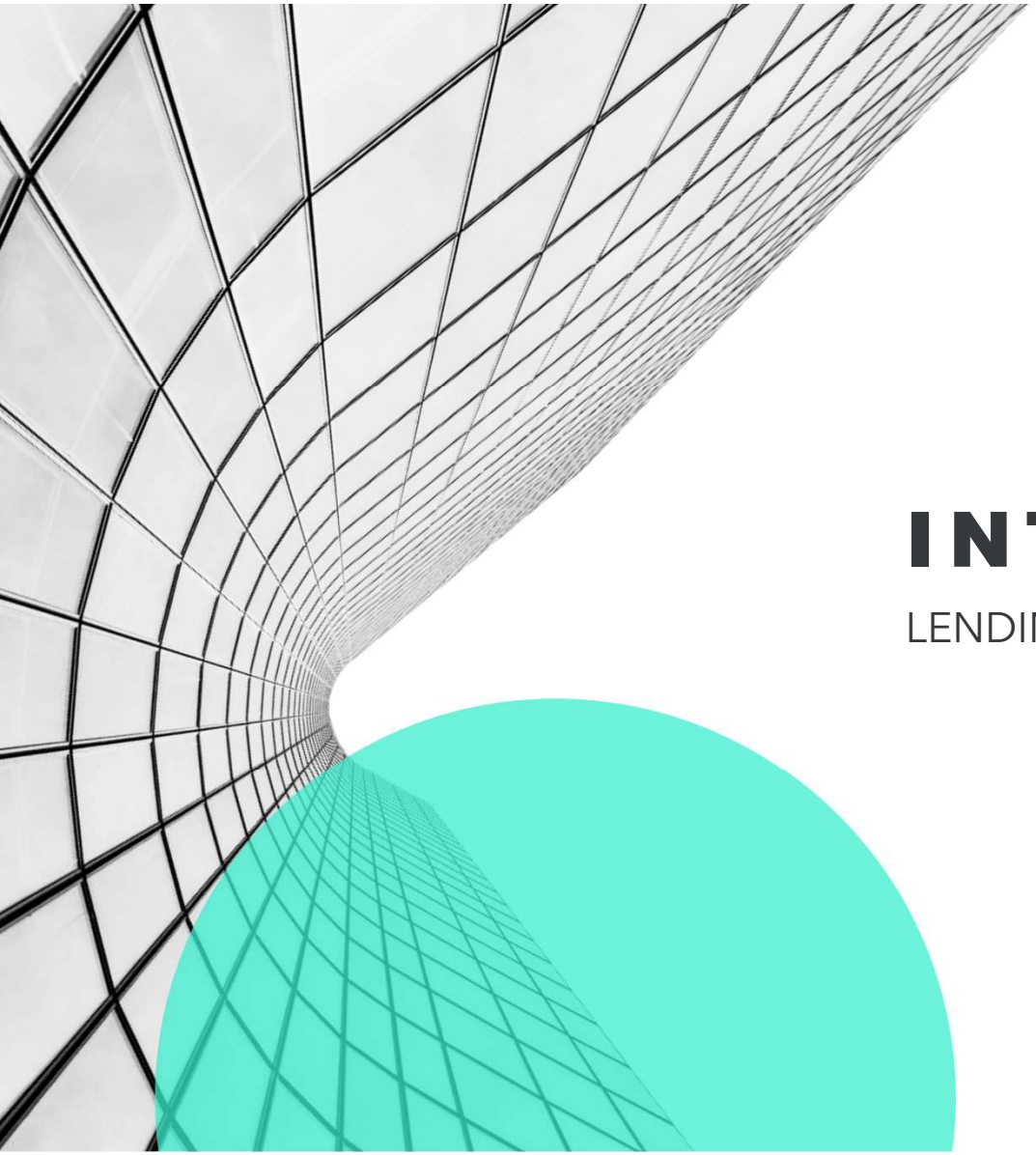
INTRODUCTION

PROBLEM STATEMENT & OBJECTIVE

PRIMARY GOALS & APPROACH

EXPLORATORY DATA ANALYSIS (EDA)

CONCLUSION - RECOMMENDATIONS



INTRODUCTION

LENDING CLUB EDA CASE STUDY

ABOUT THE CASE STUDY

Company

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.

Key Takeaway for the Student:

Solving this assignment will give an idea about how real business problems are solved using EDA.

In this case study, apart from applying the techniques learnt in EDA, also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

BUSINESS UNDERSTANDING

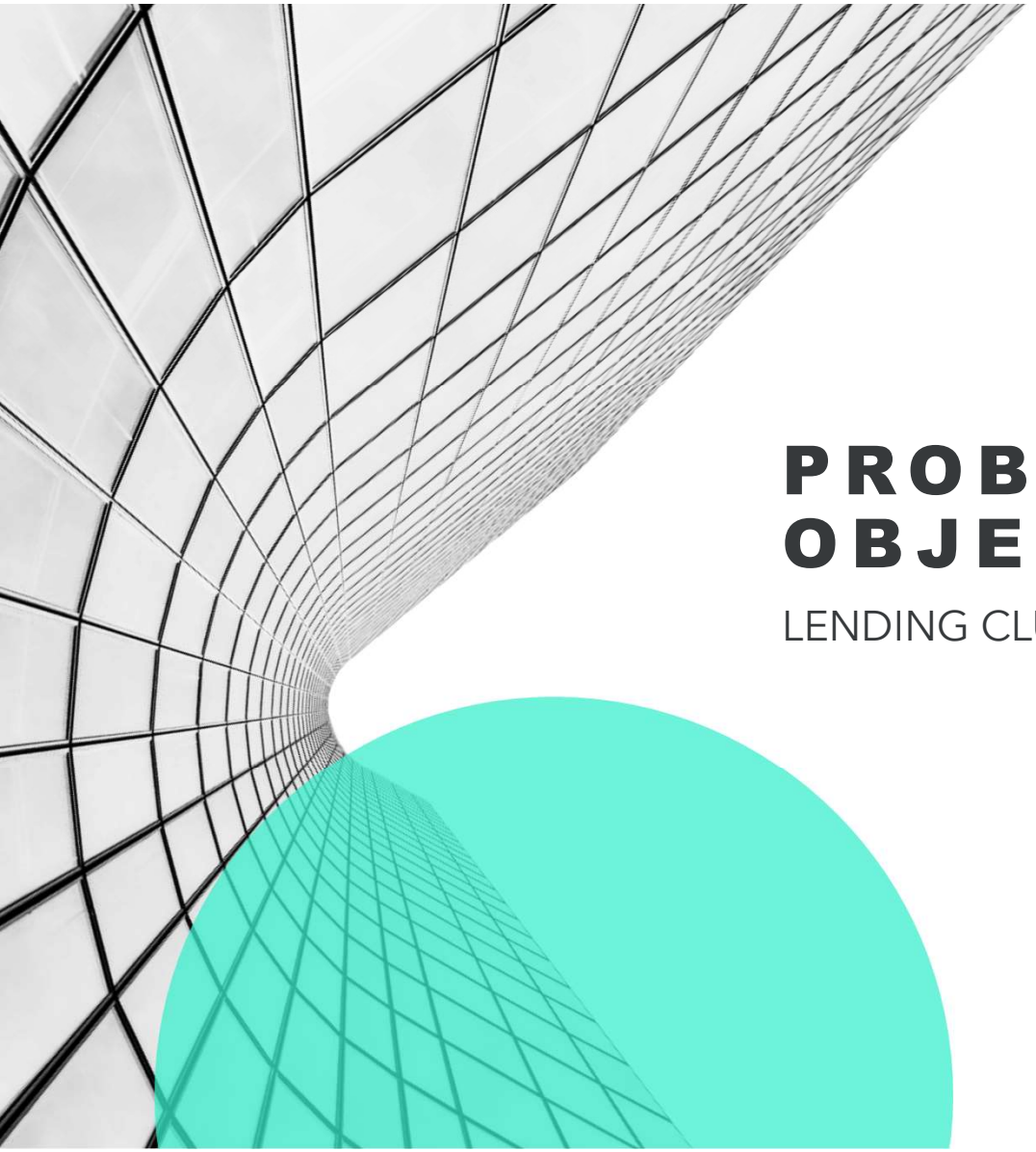
Lending Club company specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

- 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

The data given below contains information about past loan applicants and whether they 'defaulted' or not. 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.

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

1. 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
2. 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)



PROBLEM STATEMENT & OBJECTIVE

LENDING CLUB EDA CASE STUDY

CONTEXT OF PROBLEM

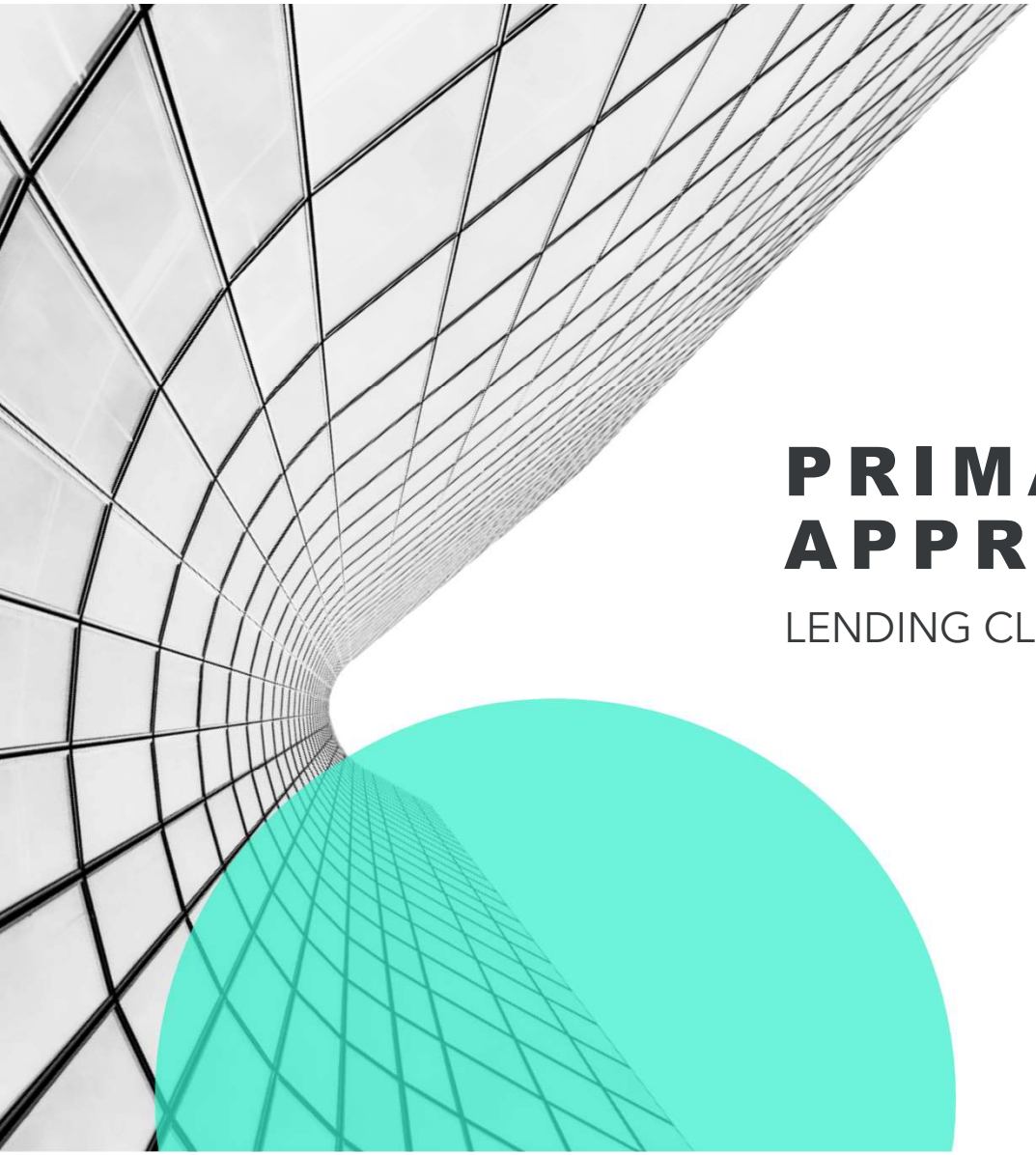
Lending Club company is the largest **online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.** Borrowers can easily access lower interest rate loans through a fast online interface.

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'**.

OBJECTIVE

Objective is to identify the risky loan applicants, then Charged Off or Defaulted 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.



PRIMARY GOAL & APPROACH

LENDING CLUB EDA CASE STUDY

PRIMARY GOAL

Primary Goal

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

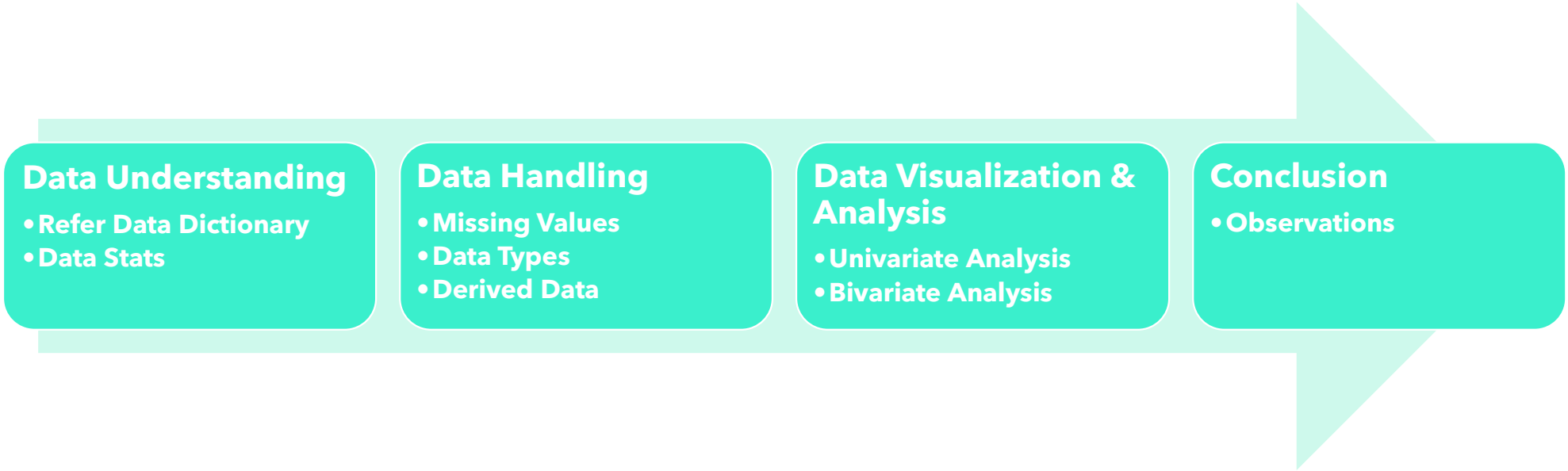
Associated Risks

- 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.

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.

EXPLORATORY DATA ANALYSIS (EDA) APPROACH



EXPLORATORY DATA ANALYSIS (EDA)

LENDING CLUB EDA CASE STUDY

DATA UNDERSTANDING

- **Loan Data Set Statistics**
  - Total Rows – 39,717
  - Total Columns – 111
- Observations:
  - Many (54 out of 111) Columns have all null values
  - Column Data Types for imported file
    - Float64 - 74
    - Int64 - 13
    - Object - 24
  - Loan Status
    - Loan status is Fully Paid, Charged Off and Current.
    - **Current** status is for the applicant who have availed loan and are repaying on-time,
    - **Charged Off** status indicates the defaulters.
    - **Fully Paid** status 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	44	39,717
4	Dropped irrelevant columns which are calculated after loan is approved thus have no relevance to the analysis	39	39,717

Excluded the rows with Loan Status = 'Current', as our goal is to see who is likely to default and this can only be said in case of either fully paid or charged off loans.

- Remaining Row – 38,577
- Remaining Columns - 39

#	Description	Remaining Columns	Remaining Rows
1	Replaced Missing values in the columns	39	38,577
2	Added Derived Columns (category datatypes) from Numeric, Date and Object Data types.	58	38,577
3	Updated the datatypes for the columns, based on the data content (e.g. Object --> Category, Object --> Numeric, Object --> Date)	58	38,577
4	Checked for Outliers and removed the data.	58	36,654

DATA VISUALIZATION

Visualization Approach Used

- Univariate
- Segmented Univariate Analysis
- Bivariate Analysis

Visualization Techniques Used

- Box Charts
- Bar Charts
- Histogram Charts
- Pie Charts
- Line 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
- plotly.express - version 5.9.0
- Python - version 2.7.18

View the above at - https://github.com/ShashankPawas/EDA-Lending_Club_CaseStudy

OBSERVATION

View the above at - https://github.com/ShashankPawas/EDA-Lending_Club_CaseStudy

Univariate Analysis

•Charged Off Loans constitute approx..15% of the overall loans sanctioned.
•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
•73% of loans defaulted (Charged Off) are in the loan amount range of 500 – 15K.
•Applicants reside in rental homes
•Applicants who have mortagaged their homes
•Applicants Applying for loan with purpose of Debt Consolidation, i.e. applicants who use the loan to clear other debts
•Around 50% of the loans are sanctioned for interest range between 5% - 13%
•Around 50% of Charged Off Loans are in 13% - 21% interest rate range.
•Applicants who have an income of range 30k - 75k
•Applicants with employement length of 10+ years
•Average Installment is range of 270 – 280.
•Majority of Loan Applicants have annual income in range of 40k – 60k
•Majority of defaulted loan applicants have income range between 50k – 75%
•Majority of Loans are sanctioned for purpose of Debt Collection and same purpose has majority of loan defaulters.
•Majority of loan applicants do not have public records
•More than 50% of the loans are sanctioned for loan applicants with Grade A or B.
•Grade B and C have higher count of loan defaulters.
•Majority of Loan applicants have 10+ years of experience, 10+ years experience has higher default loans count.
•Majority of Loan applicants do not own house, have home ownership as Mortgage or Rent.
•Majority of Loan Applicants are from CA, FL, TX and NY. With CA leading the pack.
•Majority of Loan Applicants have debt to income ratio of 10 - 15.
•Majority of defaulter have DTI ratio of 12 – 18.
•Loan sanctioned are in larger number during the last quarter of the year, also the defaulted loans show same trend.
•50% of loans were sanctioned in 2011 year.
•Loans sanctioned increases exponentially YoY, upto 2x times, between 2007 to 2011, the defaulted loans also show the same trend.

OBSERVATION

View the above at - https://github.com/ShashankPawas/EDA-Lending_Club_CaseStudy

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.

•0.25 percentile is almost same for both Full Paid and Charged Loans, but the gap is higher at 0.75 percentile, indicating that higher loan amount are more likely to be defaulted.

•Loan applicants with DTI in range of 10 - 15 are more likely to default loan.

•Loan Applicants who do not own home (Mortgage & Rent) are high probability of defaulting loans

•Loan Applicant with Annual Income below <60k are most likely to default.

•Loan Applicant with 10+ years of experience is more likely to default, but also have higher chance of Fully Paying the loan.

•Charged Off (Default) loans increase with interest rate, then gradually declines after hitting 17.5% interest rate.

•Both Fully Paid and Charged Off Loans are increase exponentially YoY from year 2007 to 2011. With 2011 accounting for around 53% of loans.

•Grade A and B have highest number of Fully Paid Loans, while Grade B and C have higher count of Charged Off Loans

OBSERVATION

View the above at - https://github.com/ShashankPawas/EDA-Lending_Club_CaseStudy

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.

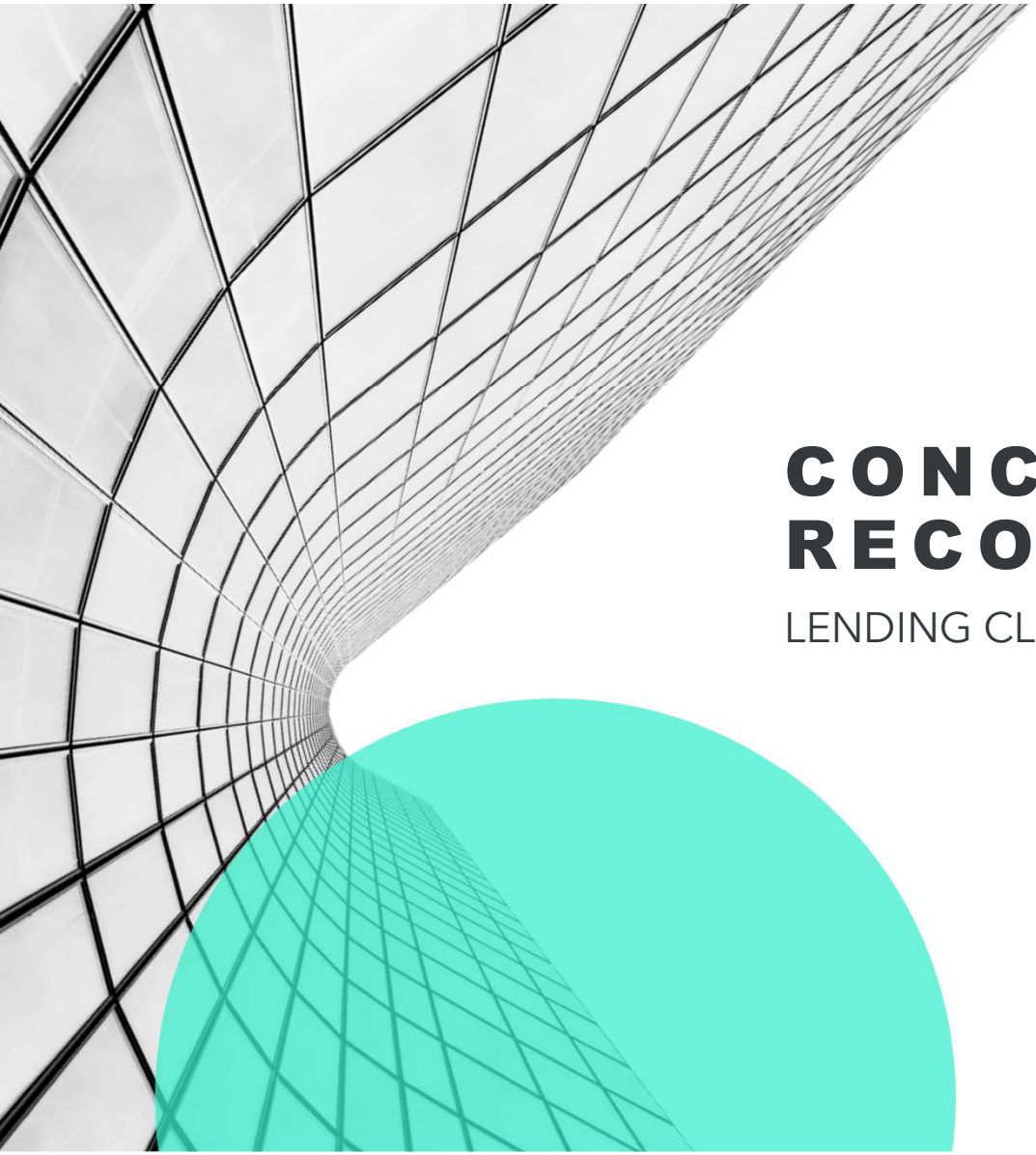
•Higher Annual Income loan applicants are higher interest rates and are also sanctioned higher loan amounts. But as the loan amount increases the number of default loans also increase. Hence it can be inferred that Higher Loan amount with high interest rates are more likely to be defaulted.

•Loan availed with a purpose of Debt Consolidation, Small Business and Credit Card are top three loan defaulters.

•Loan applicants with public records (Derogatory or Bankruptcy) and with higher loan amounts & Interest Rates are more likely to default.

•Loan Applicants whose Public Records are not known have higher defaulted loans, indicating that Public records are one of key indicator.

•Generic Trend noticed is Higher Loan amount with higher interest rates are more likely to defaulted.



CONCLUSION - RECOMMENDATIONS

LENDING CLUB EDA CASE STUDY

RECOMMENDATIONS

View the above at - https://github.com/ShashankPawas/EDA-Lending_Club_CaseStudy

Following are Recommendations based on the analysis of charged off loans. 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%
•Applicants with Annual Income range between 35k - 70k and availing loan for Debt Consolidation.
•Loan Applicants with Public Records (Derogatory / Bankruptcy) not known.
•Loan Applicants with Public records (Derogatory / Bankruptcy), with loan amount >10k or interest rate > 10%
•Loan Applicants irrespective of the Verified status, with loan sanction is >=10K and interest rate >10%.
•Loan Applicants with >10 years of experience and with loan amount > 10K or with interest rate (>10%)
•Loan Applicants with purpose of Debt Consolidation or Small Business, with loan amount >=14K or with interest rate > 10%
•Loan Applicants with high DTI (Debt to Income) ratio.
•Loan Applicants from state CA, FL, TX and NY

Driving Factors / Indicators for Loan Defaulters:

•Grades, Annual Income, DTI , Public Record (Derogatory / Bankruptcy), Purpose of Loan, Loan Amount, Interest Rate
•Note - DTI should ideally cover Home Ownership - Mortgages & Rent, if not these can be added as key indicators.

CREDITS

Shashank Pawaskar

Budding Data Scientist.. ☺

Thanks to the examples and documentation.

- https://towardsdatascience.com/
- https://www.youtube.com/watch?v=vN5dAZrS58E
- https://stackoverflow.com/
- Respective documentations sites for numpy, panda, matplotlib, plotly, and python.

THANK YOU

Shashank Pawaskar

shashanksp@msn.com