LENDING CLUB CASE STUDY

GROUP ID: 100084

AGENDA

- Problem Statement
- Risk Types
- Data Cleaning
- Data Analysis
  - Univariate
  - Segmented
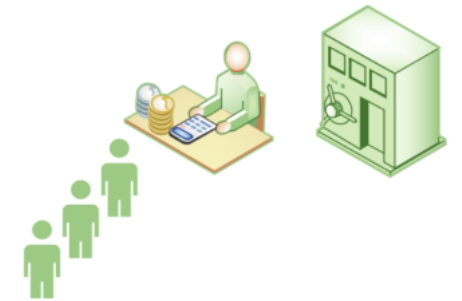 - Bivariate
- Recommendations

PROBLEM STATEMENT

- 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

- 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.

- 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.

RISK TYPES

- Risk associated with bank for approving the loan

Start

Loan Approval

No

Yes

Applicant Likely to repay the Loan

Applicant not likely to repay Loan DEFAULT

Loss of Business

Financial Loss
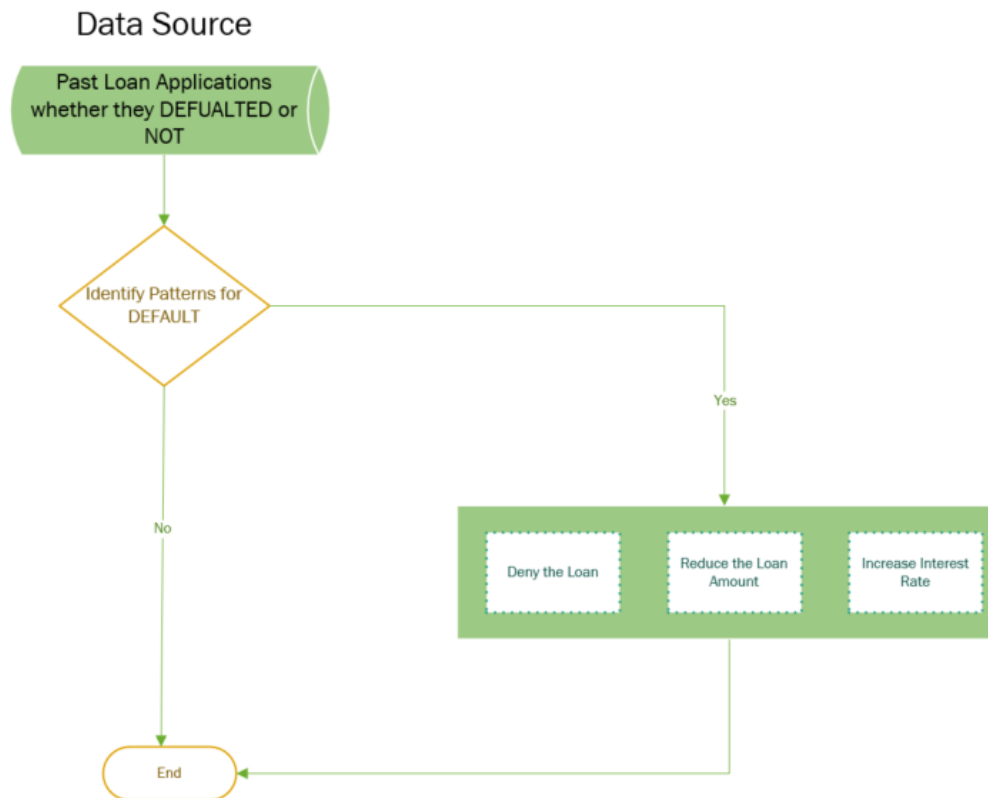
End
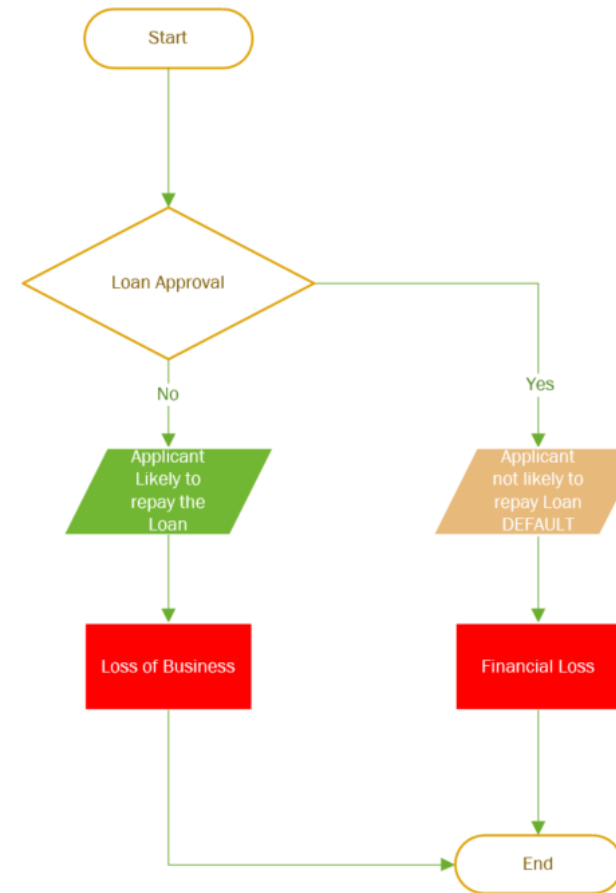
Risk Types Associated to Bank

AIM OF USE CASE

Identify Risky Applicants

Identify Driving Factors or Driver Variables for Risky Applicants

Lending Loans to Risky Applicants leads to Credit Loss(Financial Loss)

AIM OF USE CASE

Data Source

Past Loan Applications whether they DEFUALTED or NOT

Identify Patterns for DEFAULT

Yes

No

Deny the Loan

Reduce the Loan Amount

Increase Interest Rate

End

DATA ANALYSIS

- From the loan data set, there are **3** possible loan statuses

  - **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
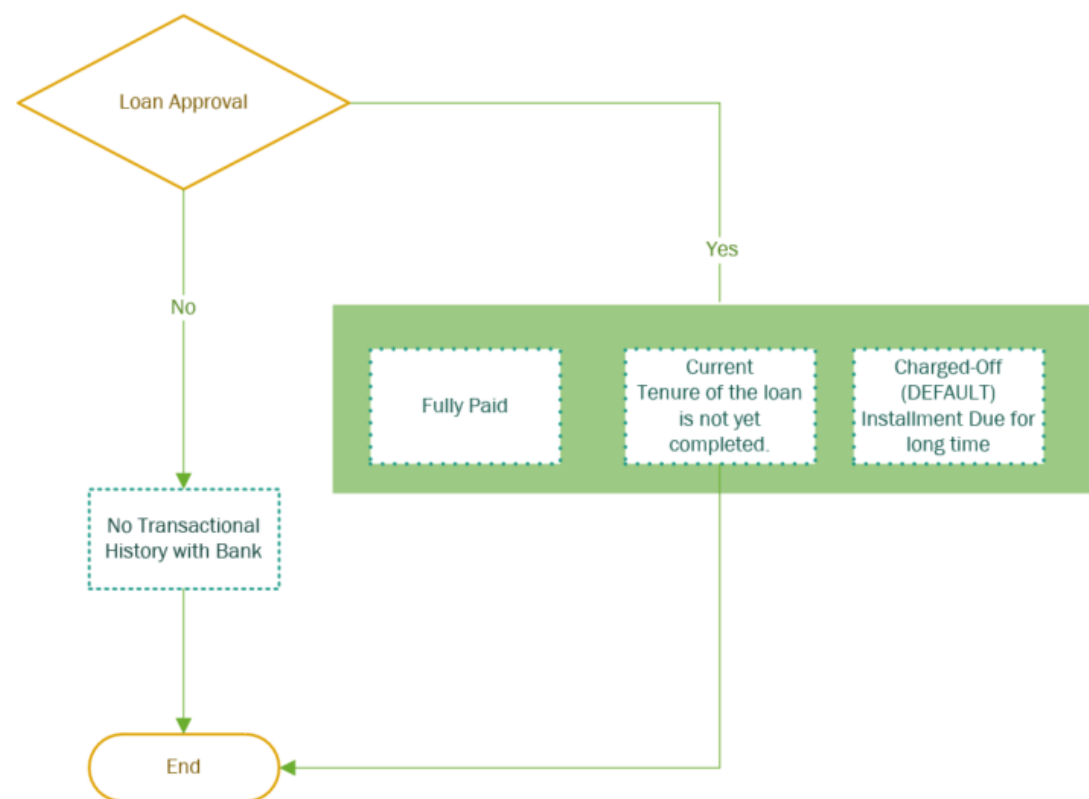
DATA ANALYSIS

There are broadly three types of variables –

1. Related to the applicant (demographic variables such as age, occupation, employment details etc.),
2. Loan characteristics (amount of loan, interest rate, purpose of loan etc.) and
3. Customer behavior variables (those which are generated after the loan is approved such as delinquent 2 years, revolving balance, next payment date etc.).

Now, the customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.

DATA CLEANING

❖ Keep only columns where 80% or more valid data is available
❖ Keep only rows where 70% or more valid data is available
❖ There is no duplicate content in data frame
❖ There is no Incorrect rows - Header rows, footer rows
❖ There is no Summary rows, total and subtotal rows
❖ There is no Extra rows, Column numbers, indicators, blank rows
❖ Missing Column Names Column names as blanks, NA, XX etc.
❖ There is no Inconsistent column names ,X1, X2,C4 which give no information about the column
❖ No Unique Identifier E.g. Multiple cities with same name in a column
❖ Misaligned columns Shifted columns
❖ There is no Columns containing Multiple data values E.g. address columns containing city, state, country

DATA CLEANING

❖ Customer behavior variables are not available at the time of loan application hence they are dropped.
❖ Renamed issue_d to loan_date
❖ Renamed Emp_length to emp_exp
❖ Renamed Dti to debt_to_income
❖ Converted date column to panda date and time format
❖ Interest rate columns are cleaned by removing %
❖ Amount columns have been rounded off to 2 decimals
❖ Employee experience has been adjusted with range of values (0-10)
❖ The ones marked 'current' are neither fully paid not defaulted, so get rid of the current loans.

FEATURES USED FOR ANALYSIS

- ❖ Loan_amt
- ❖ Term
- ❖ Interest_rate
- ❖ Grade
- ❖ SubGrade
- ❖ Annual Income
- ❖ Purpose of loan
- ❖ DTI (Debt to Income)
- ❖ Emp_Lenght
- ❖ Loan_Date(Month)
- ❖ Home_Ownership
- ❖ Verification_Status

TARGET FEATURES

- ❖ Loan_status()

DATA ANALYSIS UNIVARIANT

- ❖ Loan Applicants percentage is high for 36 months term
- ❖ Number of Loan applicants are higher with interest rate % for 7.5 and between (10 -12)
- ❖ Applicants with Loan Grade E,F,G are lesser in number and Highest Applicants are with loan grade B
- ❖ Loan Applicants with 10+ years experience get more loans
- ❖ Applicants staying in Rented Houses and Mortagage get high chances of Loans
- ❖ Most of Loans are approved without been NOT VERIFIED
- ❖ % of Fully Paid ratio is higher compared to Charged off
- ❖ Applicants with purpose of Debt_consolidation have high chance of Loan been approved.
- ❖ Loan Applicants with debt to income ratio of 15 are the highest loan receivers.
- ❖ Most loans are approved between 3k and 15k
- ❖ Loans applicants with approved amount of 20K and above are less in number.

Column	Type of Variable
term	Ordered Categorical
int_rate	Ordered Categorical
grade	Ordered Categorical
sub_grade	Ordered Categorical
emp_exp	Ordered Categorical
home_ownership	UnOrdered Categorical
verification_status	UnOrdered Categorical
loan_status	Ordered Categorical
purpose	UnOrdered Categorical
loan_amnt	Quantitative
funded_amnt	Quantitative
funded_amnt_inv	Quantitative
debt_to_income	Quantitative

DATA ANALYSIS SEGMENTED UNIVARIANT

- ❖ Applicants applied for 3 years term: 29096
- ❖ Applicants applied for 5 years term: 9481
- ❖ Percentage of loan applicants applied for 3 years term: 75.42
- ❖ Percentage of loan applicants applied for 5 years term: 24.58

DATA ANALYSIS BIVARIANT

- ❖ Charged off Ratio is higher when analysed with Interest rate.
- ❖ Loan Grades with B and C are defaulting.
- ❖ Applicants with Purpose of loan as Debt consolidation have high chances of Defaulting loan.

DATA ANALYSIS DERIVED METRICS

❖ We could derive new column say Risk from Loan status assuming Fully paid as Low risk and Charged of as High risk

❖ We could derived member id from URL but since explicitly member id column exists we are not deriving it again.

❖ We would derive new column say Year and get insights of loan defaults w.r.t year.

❖ Loans issued in Year 2011 has Highest number of Defaulted.

RECOMMENDATIONS

- ❖ Applicants with Verification status as Source Verified has higher chances of Loan Defaults
- ❖ Applicants with Home ownership MORTAGAGE are taking higher amount of loans and have high chance of Defaulting.
- ❖ Applicants with Home ownership RENT and OWN are having higher Loan Default Ratio compared Loan Closed Ratio.
- ❖ For tenure of 36 months, interest rates with 10.37% and higher would always have high chance of Defaults.
- ❖ For tenure of 60 months, interest rates with 13.06% and higher would always have high chance of Defaults.
- ❖ Applicants with Purpose of loan as Debt consolidation have high chances of Defaulting loan.
- ❖ Loan Grades with B and C are defaulting.

THANK YOU

- Ravikiran Kongati(ravikiran.kongati@sap.com)
- Arjun J Reddy(arjun.janardhan.reddy@sap.com)