upGrad

#LifeKoKaroLift

Lending Club Case Study: Pre-Assignment Session

upGrad

Course : ML/AI

Lecture On : Case Study

Instructor : Dr Reena Duggal

upGrad

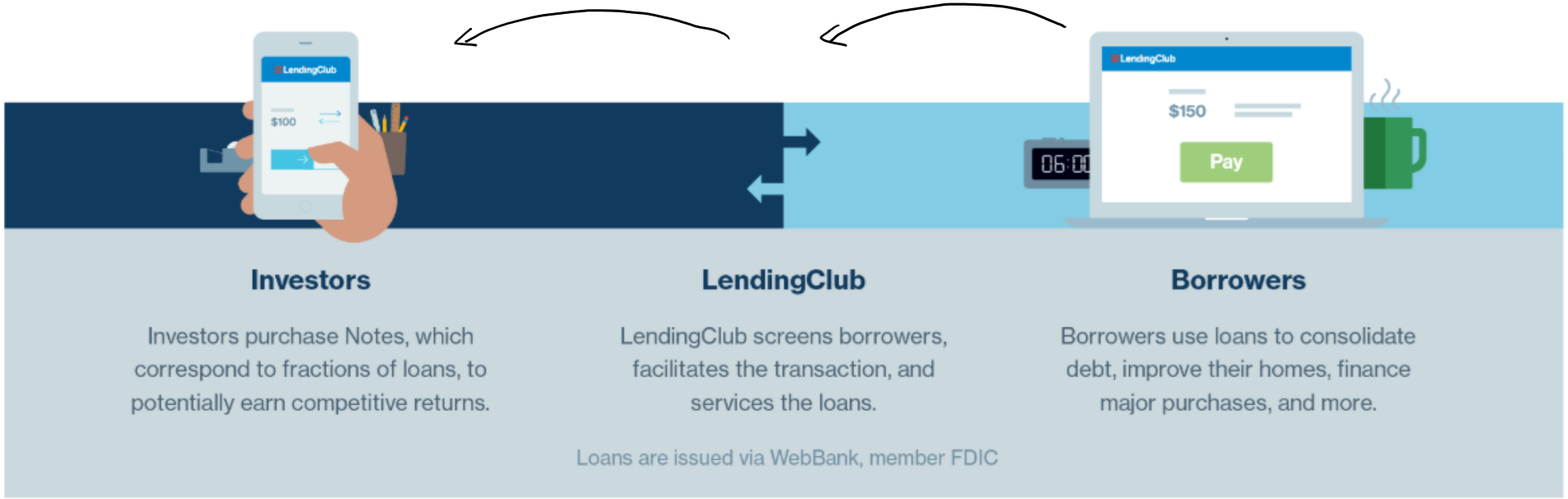
What we will cover in this session?

1. Problem statement
2. Assignment walkthrough
3. QnA

Online bank

What is Lending Club?

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

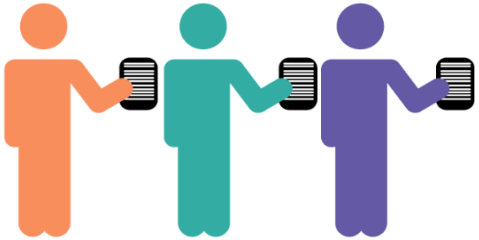


How Lending Club works?

1. Customers interested in a loan complete a simple application at LendingClub.com.
2. Lending Club evaluates each borrower's credit score using their data science process which uses past historical data and assigns an interest rate to the borrower.
3. Qualified applicants receive loan offers in just minutes.
4. Investors select the loans they want to invest in based on their own risk tolerance, investment portfolio goals, and time horizon.

Patterns in the past data

LOAN DATASET



Loan Accepted → Default / Non-Default

Loan Rejected

(Not considered in dataset)

100 columns

What is loan_amnt, funded_amnt, funded_amnt_inv ?

The loan_amnt is the amount applied by potential borrowers, funded_amnt is the amount recommended/approved by Lending Club, and the funded_amnt_inv is the amount funded by investors.

95%

loan-amt → Amt asked by the borrower — $100

funded-amt → Amt suggested by lending club to the invester → $95

funded-amt-inv → Amt given by the invester to the borrower → $90

$100	$100
$100	$90
$100	$100

Steps to proceed with the Case Study

There are four major parts that are needed to be done for this case study:

1. Data understanding
2. Data cleaning (cleaning missing values, removing redundant columns etc.)
3. Data Analysis
4. Recommendations

Data Understanding

object
int
float

1. Read the data to Python dataframe
   - Load
   - shape
   - Describe
   - head( )
2. Check the datatype of various columns
3. **Correct the datatype of the column if required**
   - Check columns where you may require to extract numerical data.
4. Identify the target column.
   - Default
   - Non-default

emp_length: 5 years, 6.5+ years

Interest_rate: 5% → Numeric

Outliers Detection

- 
- Boxplot
- → Mean = Median
- → 75% per , Max

Data Cleaning

Information Content

1. Check the percentage of missing values. >90%
2. **Remove all those with very high missing percentage.**
3. **For columns with less missing percentage: perform data cleaning steps for both columns and rows**
   a. You don't need to impute the data, you can just identify the correct metric to impute the column.
   b. You can drop rows where the missing percentage is quite high.

→ Mean ✓
→ Median → Outliers
→ Mode → Categorical Column

100

→ 10

KNN → Nearest Neighbour

Age is missing

Emp-length

Data Analysis

- The objective is to identify predictors of default so that at the time of loan application, we can use those variables for approval/rejection of the loan.

  There are broadly three types of variables -

  1. those which are 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 behaviour variables (those which are generated after the loan is approved such as delinquent 2 years, revolving balance, next payment date etc.).

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

- The ones marked 'current' are neither fully paid not defaulted, so get rid of the current loans. Also, tag the other two values as 0 or 1 to make your analysis simple and clean.

Target Var
- Fully paid → Non-defaulter (0)
- ~~Current~~ Drop
- Charged Off → Defaulter (1)

Few Important Variables

upGrad

- Loan-amt
- Term
  - 36 months
  - 60 months
- Interest Rate
- Grade
  - A
  - B
  - C
  - D

  A → Less Risky → Most Risky

  Subgrades: 1, 2, 3, 4

  Interest Rate: 5%, 6%

- Verification Status
  - Income verified (Pay slips,
  - Income Source Verified (3rd Party)
  - Not verified
- Home ownership
- Year (Loan-Date)

- Annual Income
- Purpose of Loan
- DTI (Debt to Income)
- Emp length
- Description

2 Debts
$1000
$1500

Income/Salary
$2000

$\frac{1500}{2000} =$

$\frac{4000}{2000}$ 50%

Data Analysis: Univariate Analysis

- For univariate analysis, you may check the default rate across various categorical features.
- For continuous features, you may perform binning and then you may perform univariate analysis.

Data Analysis: Bivariate Analysis

- Here you may choose two or more feature to understand the default variable.

Univariate Analysis examples

upGrad

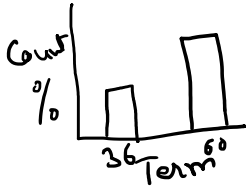
sns.barplot(x = term, y = target_var)

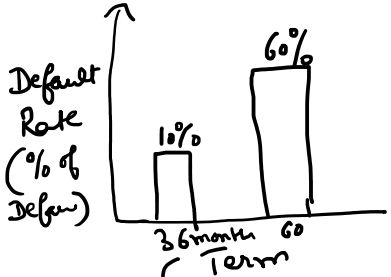
x → Categ.

y → numerical (avg. of target var)

User-defined-funct.

value_count 20% 80%





	Term	Target
P1	36	0
P2	36	0
P3	36	1
P4	36	0
P5	60	1
P6	60	1
P7	60	0

Default Rate = $\frac{3}{7} = 0.43 = 43\%$

36 months = $\frac{1}{4} = 25\%$

60 months = $\frac{2}{3} = 66\%$

Cat. Data

- Grade
- Sub-grade
- Purpose of Loan

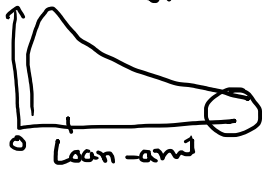
- Loan-amt
- Interest-Rate
- Emp. length
- Annual Income

If loan-amt < 10000 Small

10000 50000 Medium

> 50000 Large

Dist. Plot



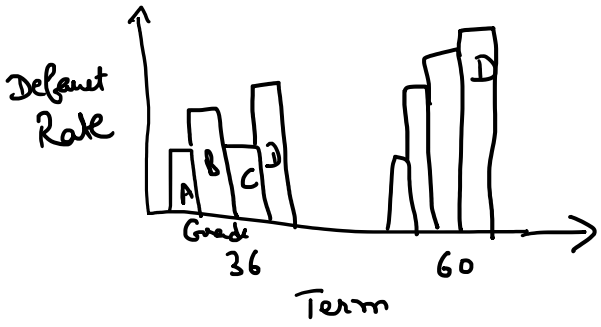
Year (Loan Date)

Line Plot

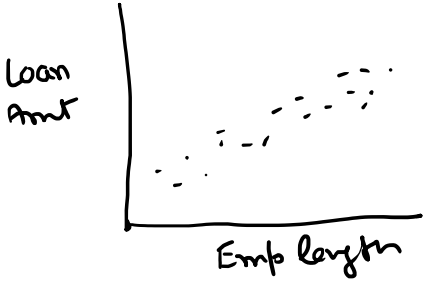
Bivariate Analysis examples

Top 5 purpose of loans

upGrad

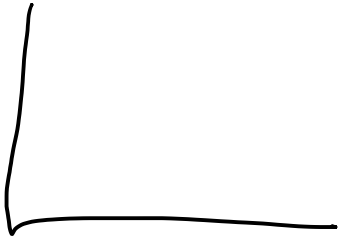


Scatter Plot



Pair plot

Heatmap



Another technique to find Important Predictors

upGrad

Grade	Default Rate
A	0.01
B	0.20
C	0.30
D	0.50

Information Content is high

Emplength(1,20)	Default Rate
1-5	0.03
6-10	0.07
10-15	0.13
>15	0.15

is less

Interest Rate	Default Rate
Low	0.20
Medium	0.20
High	0.20

is zero

Grade = Important feature

Diff = Max - Min = 0.50 - 0.01 = 0.49
0.12
0.00

Sort it index order

User-defined-function

Best Predictors

Recommendations

- Remember this is an important part of the case study. After performing your analysis, you need to recommend some points to the investors. You need to emphasise on how they can reduces the chances of funding a likely defaulter.

- This is need to be done for both PPT and the Jupyter Notebook

Presentation and Points to remember

- Remember in this case study we are trying to figure out the important features that contributes toward default.
- Any assumption taken is fine, until it is clearly mentioned on your jupyter notebook.
- PPT is needed to be drafted for investors, so it should not have any code. You can include plots with the explanation and recommendation to the investor.
- You need to submit a PDF. You can convert the PPT to a PDF and then submit it.
- A single ZIP file is needed to be submitted with one Jupyter Notebook and a PDF file.
- Don't forget to comment the code properly as it carries separate marks.

Poll Questions

Q-1: When should we use Median to impute missing values rather than mean?

a) When there are so many data points
b) When you have so many missing values
c) When the data is having outliers
d) Can use both mean and median

Q-2: Should we drop the variable if we have 30% missing values

a) Yes
b) No
c) Depends how critical the variable is

Q-3: If my target column in the data has the number of 0s as 80 and the number of 1s as 20, then what is the imbalance percentage of the given target column?

a) 50%
b) 20%
c) 80%

Poll Questions

Q-4: Suppose there is a data with 1000 rows, We have a categorical column with two categories, One of the category has 950 observations while the other have 50 observation. Is this fine to keep this variable if we need to build a model using this data?

a) Yes
b) No

Q-5: If we have a huge amount of data, is it ok to process it in local machine or should we look for any alternative?

a) Fine with local machine
b) Look for alternative

1 c) 2 c) 3 b) 4 b) 5 b)

upGrad

#LifeKoKaroLift

Thank You!