upGrad

#LifeKoKaroLift

Lending Club Case Study: Pre-Assignment Session

upGrad

Course : Machine Learning

Lecture On : Case Study

upGrad

Agenda

- Problem Statement
- Discussion over solution approach
- Q & A

Lending Club: EDA Case Study

upGrad

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.

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

Lending Club: EDA Case Study

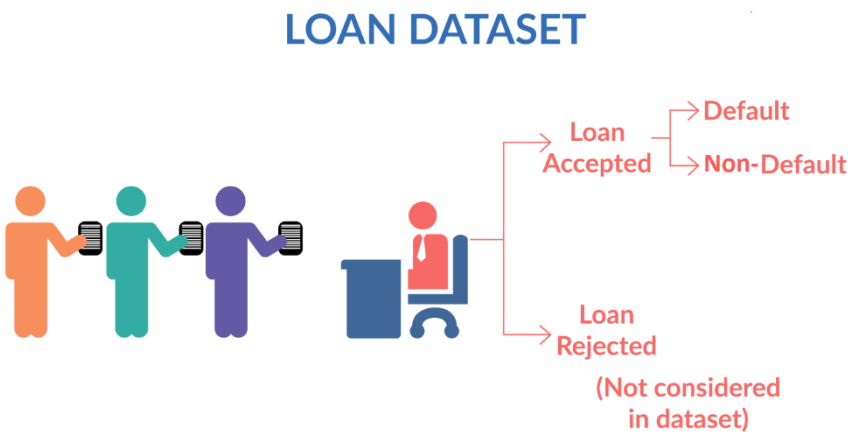
upGrad

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 utilize this knowledge for its portfolio and risk assessment.

Lending Club: EDA Case Study

upGrad



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

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.

Lending Club: EDA Case Study

upGrad

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 Cleaning

1. Check the percentage of missing values
2. Remove all those with very high missing percentage
3. For columns with less missing percentage: perform Imputations
    - You don’t need to impute the data, you can just identify the correct metric to impute the column.
4. You can drop rows where the missing percentage is quite high

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 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.

- 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.

Lending Club: EDA Case Study

upGrad

Few Important Variables

Important variables

- Loan_amt
- Term → 36 Months, 60 Months
- Interest_rate
- Grade → A, B, . , .
- SubGrade → 1, 2, . , .
- Annual Income
- Purpose of loan
- DTI (Debt to Income)
- Emp_Lenght
- Loan_Date(Month)
- Home_Ownership
- Verification_Status → Income Verified (Pay Slips, Tax Returns); Income Source Verified (3rd Party vendors has checked); Not Verified

Customer behaviour variables

delinq_2yrs
earliest_cr_line
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_d
last_pymnt_amnt
last_credit_pull_d
application_type

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

variables such as acc_now_delinquent, chargeoff within 12 months etc. (which are related to the applicant's past loans) are available from the credit bureau.

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 features to understand the Default variable

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 emphasize on how they can reduce the chances of funding a likely defaulter.

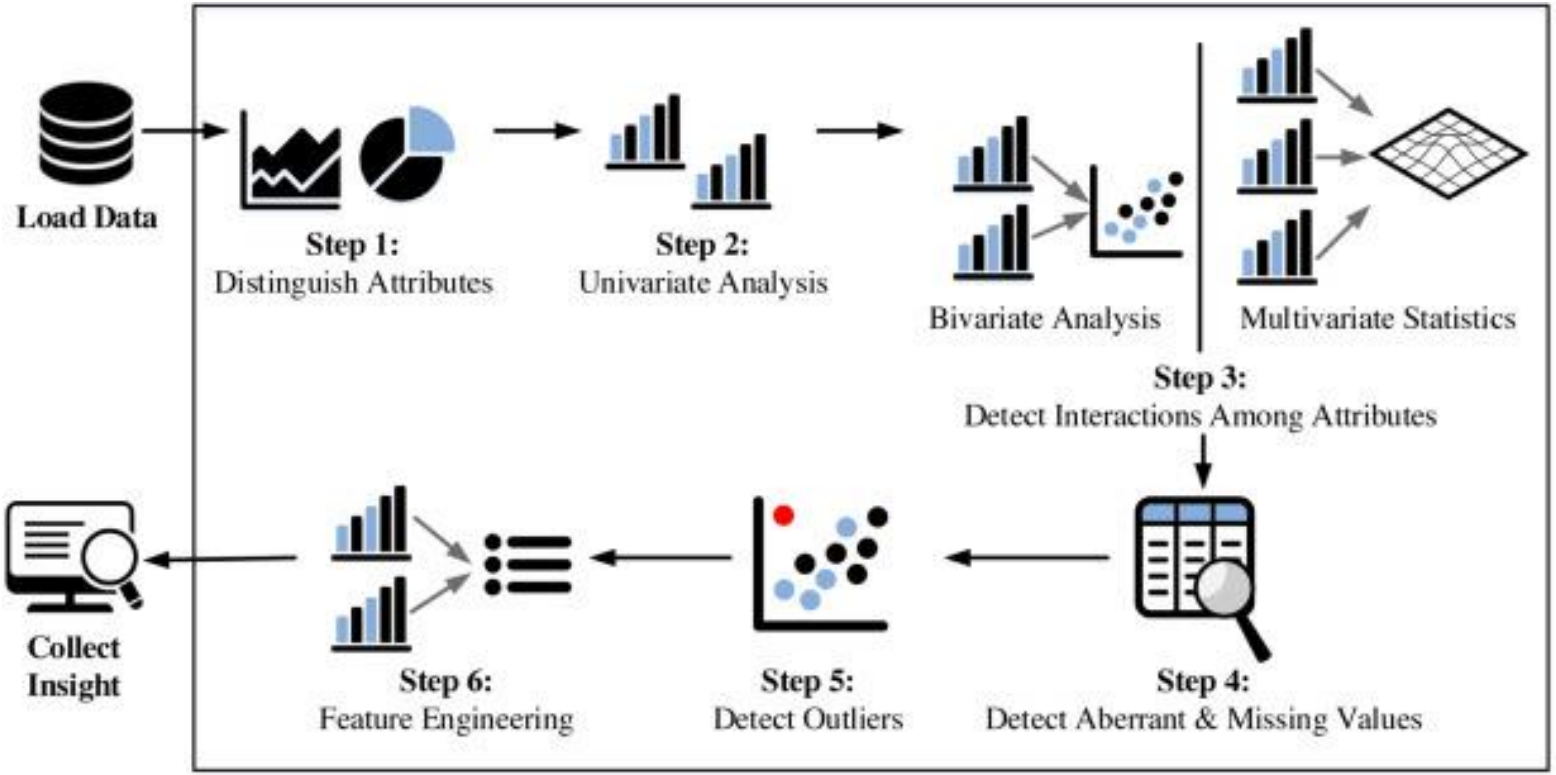
- This is need to be done for both PPT and the Jupyter Notebook

Presentation and Points to remember

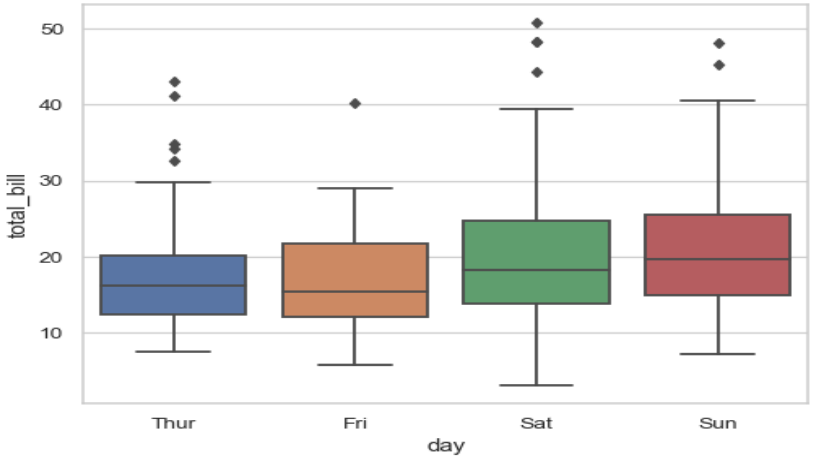
- Remember in this case study we are trying to figure out the important features that contribute toward default.
- Any assumption taken is fine, until it is clearly mentioned on your jupyter notebook.
- PPT is needed to be drafted for investors, so it should not have any code. You can include plots with the explanation and recommendation to the investors. 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.
- Please make sure to rename your Python notebook "Group_Facilitator_Name.ipynb".

Fundamental steps of EDA process

upGrad



Categorical & continuous



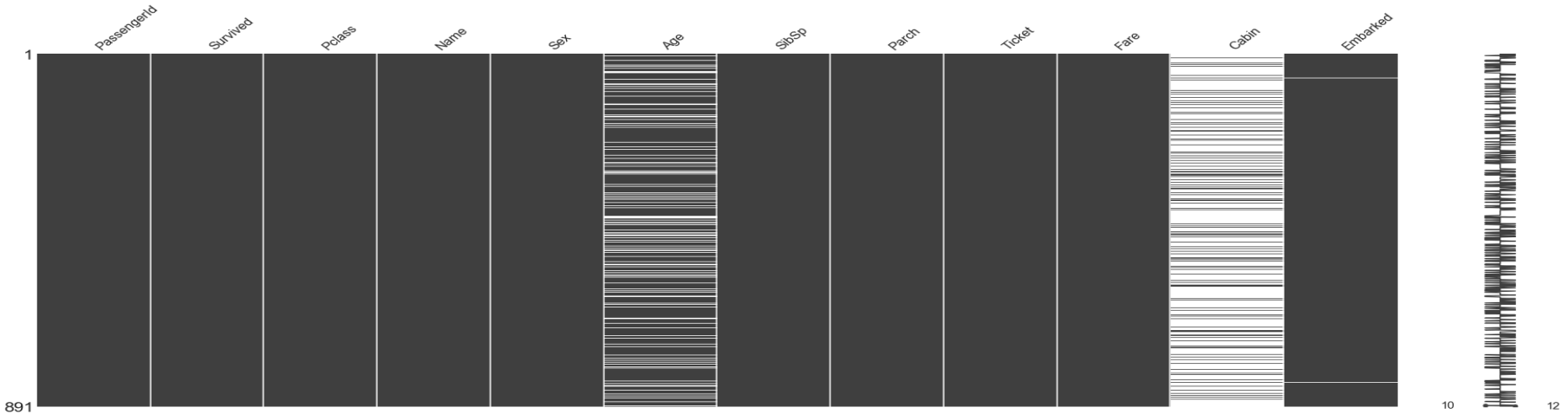
While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables

Missing Value Treatment

upGrad

1.Data Extraction: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.

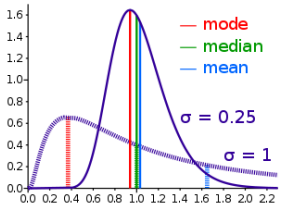
2.Data collection: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:



Methods to treat missing values

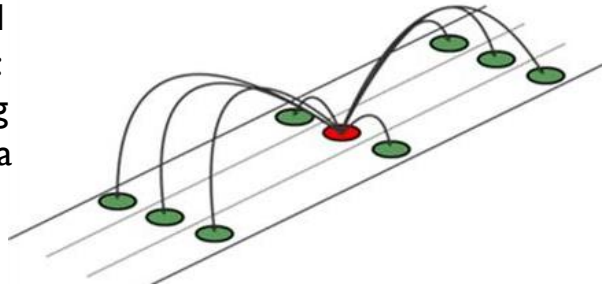
upGrad

Data – deletion Deletion methods are used when the nature of missing data is "**Missing completely at random**" or we have good amount of data and the data loss would be really low ,else non-random missing values can bias the model output



Mean/ Mode/ Median Imputation Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

Prediction Model we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable.



upGrad

#LifeKoKaroLift

Thank You!