upGrad

#LifeKoKaroLift

Credit EDA Case Study: Doubts Session

upGrad

Course : Data Science

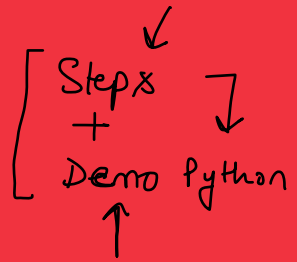
Lecture On : Credit EDA CS

Instructor : Sumit Shukla

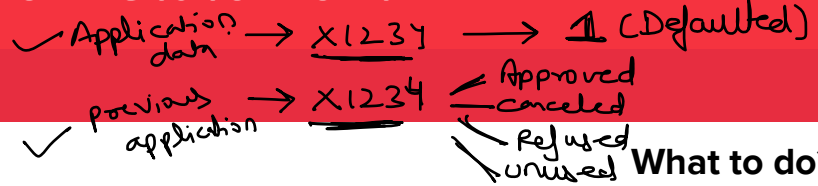
upGrad

What we will cover in this session?

1. How to start with the “Credit EDA Case Study”
2. What are the important steps that should be included
3. Points to remember
4. Python Demo
5. QnA



Problem Statement



What to do?

Bank sector — Loan

If the person is not able to repay the loan back to the bank in the given time frame then the customer is said to be defaulted.

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 company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default

i) Application data :— current
ii) Previous application:

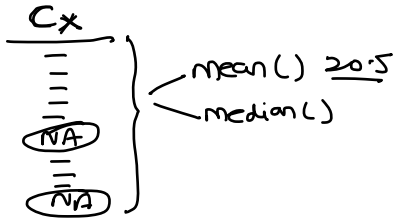
Problems
- i) If the bank approves the loan of that person who is not capable of paying back the loan. → Credit loss
- ii) If the bank dis-approves the loan of the person who is capable of paying back → Credit loss

How to Solve?

code-demo to overcome memory issue ✓

upGrad

Those numerical columns
No/Less outliers — mean()
outliers — median()



Primary file

What to do?

1. Start by importing the 'application_data.csv'.
2. Check the structure of the data(Normal routine check). — df.shape, df.info, df.describe()
3. Data Quality Check and Missing values
   - a. Find the percentage of missing values for all the columns.
   - b. Remove columns with high missing percentage.(≥ 50%)
   - c. For columns which has less percentage(around 13% or so), you need to check what will be the best metric to impute the missing values? Like if the column you are checking is a categorical column check, which category you can use to fill the nulls. For others check does mean or median can be imputed or not. Others cases may be imputing with 0. You need to do this task for some variables and not for all, say 5-6

you don't need to really impute the column but you need to report the best metric that can be used to impute a missing data in a column

categorical: object / float / int → typecast → object

load date into (Pandas datetime format

   - d. Check the datatypes of all the columns and change the datatype like negative age and date.
   - e. For numerical columns check for outliers and report them for at-least 5 variables. Add observations and reasoning. + comments
   - f. Binning of continuous variables. Check if you need to bin any variable in different categories. Do this for at least 2 variables.

✓ amount (continuous) 0 ↕ 100 → Binning → Categorical: 0–25 : Low; 25–50 : Avg; 50–75 : Good; 75–100 : Best

How to solve..Continue.

upGrad

application train → Target ← 1 (Defaulted) / 0 (Not defaulted)

What to do?

4. Analysis

25-30 columns

- Check the Imbalance percentage. No balancing technique required.
- Divide the data into two sets, i.e. Target=1 and Target=0.
- Perform univariate analysis for categorical variables for both 0 and 1. Compare the target variable across categories of categorical variables.
- Find correlation for numerical columns for both the cases, i.e. 0 and 1.
- Check the variables with highest correlation are the same in both the files or not?
- Perform univariate for numerical variables for both 0 and 1. Compared the target variable across categories of continuous variables.
- Perform bivariate analysis for numerical variables for both 0 and 1.

5. Read "Previous Application" data.

- You can merge the files, but it's a challenge.
- Perform univariate and bivariate analysis to find some pattern.

6. Final words

- Based on your analysis, define the results and conclusion.

What to keep in mind

upGrad

major analysis + Recommendation + conclusion } Max-15

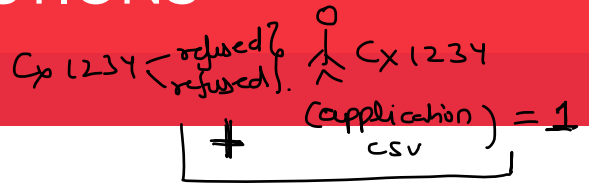
PPT → PDF

zip → Jupyter notebook, PDF

- Keep in mind “There is no correct or incorrect solution”.
- Every approach is correct if you are able to answer all the questions asked.
- The main objective of this case study is to learn and implement EDA techniques. So don’t focus so much on columns and their descriptions. (80 + columns)
- Remember you need to use plots and then understand the pattern. Then report your analysis in your notebook or PPT. No marks will be awarded if you have just plotted so many variables and have not explained the pattern. 10 plots + explanation
- It’s not possible to cover all the columns, so try to cover some of them. Based on the plots you get, try to identify the important variables.

POLL QUESTIONS

upGrad



My data contains two classes 0 and 1. The total number of rows in my data is 150. The count of rows for class 1 is 40. What is my Imbalance percentage for class 1?

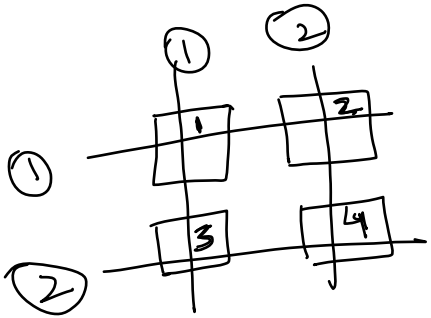
A. 47%
B. 38%
C. 27%
D. 26%

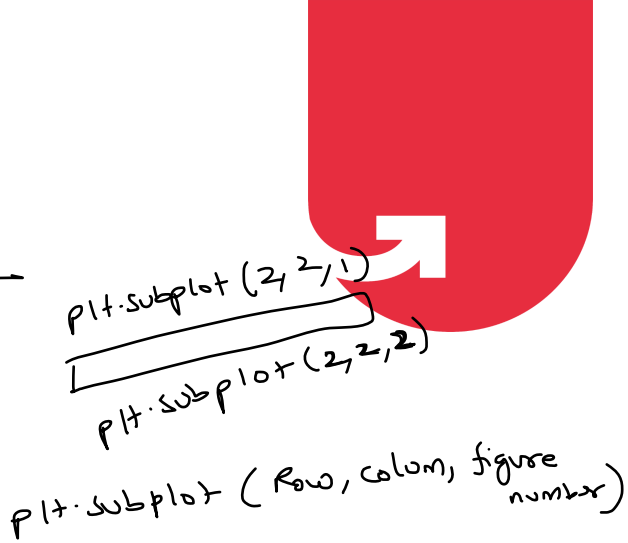
I have two files df1 and df2, both files have a common primary column as “ID”. I need all rows from my df1 data and only the common rows from the df2 data. Which join should I use?(df1.merge(df2))

A. Right
B. Inner
C. Left

upGrad

#LifeKoKaroLift

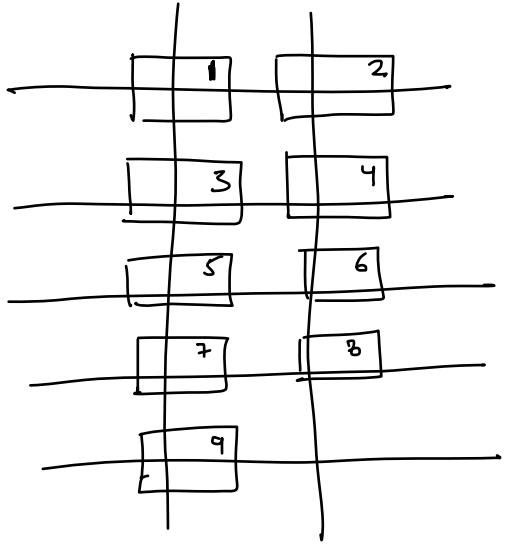




Python Demo. Let's Code

upGrad

#LifeKoKaroLift



i[0]+1

plt.subplot(5,2,i[0]+1)

i[1]

i → (0, gender)

i[0] = 0

Thank You!