Iconiq Inc / Iconiq Skills India Pvt. Ltd
601 Allerton Street, Redwood City, CA 94110

Loan Eligibility Prediction

Problem Statement:

Dream Housing Finance company deals in all kinds of home loans. They have a presence across all urban, semi-urban, and rural areas. Customer-first applies for a home loan, and after that, the company validates the customer's eligibility.

The company wants to automate the loan eligibility process (real-time) based on customer detail provided while filling online application forms. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have provided a dataset to identify the customer segments that are eligible for loan amount so that they can specifically target these customers.

Data Description

The data contains 13 variables and of which **Loan_Status** is the Target variable. Complete data dictionary can be found here.

Kindly download the data from here.

Tasks

1. Hypothesis based EDA:
    - Does being a ***Graduated Married Male*** have any positive impact on the Loan Status compared to the ***Non-Graduated Married Male***?
    - Does having a ***CoapplicantIncome*** increases the Loan Approval Percentage? (*Loan_Approval_Percentage = Total_Yes / Total_Yes+Total_No)*
    - What is the Loan Approval Percentage of users who request for a higher loan amount (Loan Amount >=$200) across different Property Area?
2. Create new composite features:
    - ***Gender+Married+Education***
    - ***Self_Employed+Property_Area***
3. Use Target Encoding to encode categorical variables (Link)

4. Detect and remove outliers using IQR (Inter quartile range)
5. Use Recursive Feature Elimination (RFE) technique to select top features
6. Build logistic regression algorithm and compare the performance between them:
    - On data with selected features
    - On data with selected features + downsampling
    - On data with selected features + upsampling