Explanatory analysis

Steps for explanatory analysis

1. Define the imbalance of the dependant variable.

2. Clean the data from missing values

3. Plot and describe distribution

- a. Identify similar patterns
- b. Explore how the ratios behave differently from feautures

4. Calculate different group means

This way, we check weather each variables differs systematically between the two groups y=0 and y=1. For example if the group mean differ substantially, it indicated that x_j is associated with the response variable y and it could potentially be a useful predictor in a logistic regression, tree or any classification model.

To determine the difference in the group mean, we conduct two-sample t-test

H0: $E[x_j|y = 0] = E[x_j|y = 1]$

H1: $E[x_j|y = 0] \neq E[x_j|y = 1]$

We create a table and order the features by the lowest p-value. It means that its group means differ the most relative to the within-group variability.

5. Construct contigiency table for categorical variables

It helps us understand whether the categorical variable is associated with default.For each categorical predictor, construct a contingency table against the binary outcome (default vs non-default) and performed a chi-squared test of independence. Variables with p-values below 0.05 were considered significantly associated with default, indicating that the distribution of default events differs across their categories.

To check whether category distributions differ between groups, we conduct a Chi-squared test

From the results (p-values being < 0.5) we see that there's strong association of categorical variables and the outcome y. Distribution differs significantly between the two groups.

6. Correlation matrix

7. Identify patterns using scatteplots