EDA and Feature Engineering Assignment

STA 511 - Foudations of Data Science

Contents

1 Data Set

Choose a data set that has at least four categorical variables and four numerical variables. The sample size should be at least 200. You can find a data set either from my teaching data repository or other data sources. The data set should be cross-sectional (i.e., each of the data points must be observed/collected/generated at the same time).

2 Description of Data

The following information of the data should be provided in the report:

- A brief description of the data source.
- How the data set is generated or collected.
- Number of variables and their type (categorical or numerical) and size of the data set.
- List the variable names and their description/definitions.

3 Exploratory Data Analysis and Feature

Perform the standard EDA such as distribution for categorical and numerical variables respectively, the relationship between two variables (combinations of categorical and numerical variables), and pairwise relationship. Keep in mind that the pairwise scatter plot is only meaning for numerical variables.

For each EDA and associated representation, you should

- interpret what you observed and the implication of potential feature engineering
- perform feature engineering based on EDA by writing an R/Python function.
- Write a main function to wrap individual feature engineering functions.
- Test the main function with different patterns in the components and sure it produces the expected result.

4 Reporting and format

The format of the report should be similar to that of the report of the case study. You can use the report template that is used for the case study.