Week 8 Deliverables

Group Name: Starbucks

Name: Shuran Fu

Email: fushr@outlook.com

Country: United States

College: USC

Specialization: Data Science

Problem Description

Develop a machine learning model to estimate whether a particular customer will buy a specific term deposit product or not based on the customer's past interaction with bank or other Financial institution.

Data Understanding

The dataset contains information for each customer, and it also reveals whether they subscribed the term deposit. The dataset includes the basic information, such as age, education and marital status of the customer, and also financial status data, such as loan and the communication between the bank and person.

Data Type

It can be divided into 2 parts, one of which is categorical data and the other is numerical data. At the beginning of the analysis, we might assume that financial status and the frequency of communication between the bank and customer might have significant influence on whether the customer subscribed the term deposit.

Data Problems

1.missing data: contact, poutcome

2.skewed distribution: duration, campaign, pdays, previous

3.outliers: duration

4.imbalanced categorical distribution: job, education, default, loan

Approach

1.missing data: fill with mean value; group by different entities several times to find the mode; fill the left values with unknown if the data amount is small

2.skewed distribution: use log transformation to create new variables

3.outliers: remove outliers from train set

4.imbalanced categorical distribution: use statistical methods to resample the data