Data Glacier Final Project weekly report

Data Glacier Internship

Team Member Detail

Name: Zeru Zhou

Group Name: Zeru Zhou’s Group
Email: zeruzhou9@gmail.com

Country: United States

College: University of Southern California

Specialization: Data Science

Problem Description

We need to preprocess the client profile data and institution data with cleaning and imputing strategies. After that, we need to build machine learning pipelines and optimize the models with cross validation. Our task is to build a classification model that could predict if a potential client will subscribe the term deposit or not and put more attention on those who are more likely to make the purchase. The intention of the institution is to create a shortlist of the potential clients for future marketing strategies.

In this specific step, I need to clean the data. First, I need to merge the different tables containing client information. Next, I need to drop the duplicated values with specific subset of columns. Third, I need to impute the data with different methods. Here I tried simple imputer, which is not that good for large scale datasets. I also tried KNN imputer and Iterative imputer that are great methods if we have large datasets. We need to keep in mind to adjust hyperparameter because it will take forever if we use the whole dataset for round robin algorithm!