Data Glacier Final Project weekly report

Data Glacier Internship

Team Member Detail

Group Name: Zeru Zhou's Group

Name: Zeru Zhou

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.

Data Understanding

The data includes many features that covering both numerical and categorical features. Due to the fact that we need to clean data and impute the missing value in the next steps, we need to encode the categorical features like job, marital, housing. I designed a function to check the dtype of the features and encode those variables that are not numerical.

The problem of the data and the strategies to overcome:

1. The data has duplicated rows that need to be found and removed.
2. The data has missing values. We need to use imputing strategies like round-robin or KNN to impute the missing values for future model construction. We are not using simple imputer because the data is relatively large-scale, those non-missing values are quite powerful enough for us to impute the missing value. Just use mean/median will lead to huge bias if the dataset is large.
3. There are non-numerical variables that needed to be encoded. For different kind of features, we have different strategies to encode them including label encoding, one hot encoding, and count encoding.
4. There are outliers in the dataset. For each feature, I only keep the 99% data on the middle of it. other widely used methods are the cook distance and 1.5 IQR, but I want to keep most of the data, so I only removed the most extreme data points.
5. The features need to be scaled if we want to use distance-based methods
6. There is a severe class-imbalance problem. We used SMOTE in each step of cross-validation to over-sample the dataset and make it balanced.