Assignment 1

Note: "left" is the target.

Q1. How many rows are there ?

Solution. As shown in `Answer.jpynb` ■

Q2. How many columns are there ? What are they ?

Solution. As shown in `Answer.jpynb` ■

Q3. How many features? What are they?

Solution. As shown in `Answer.jpynb` ■

Q4. How many duplicates if any?

Solution. As shown in `Answer.jpynb` ■

Q5. Remove duplicates if there are any.

Solution. As shown in `Answer.jpynb` ■

Q6. Print the distributions of each features. What do you see?

Solution. The distributions for each feature are shown in `Answer.jpynb`. From the histogram, we can easily recognize **outliers** because they usually are separated from the main group. For example, from figure 1, we may predict that the outliers may be values around 0.4.

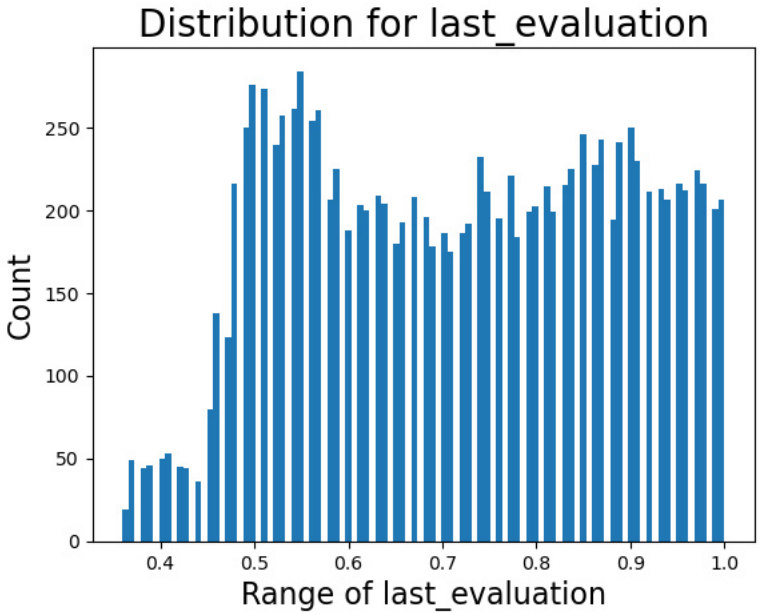

Figure 1: Distribution for `last_evaluation`

■

Q7. How many missing values are there ? What features have missing values?

Solution. As shown in `Answer.jpynb` ■

Q8. Drop the features with largest number of missing values and fill up the rest of the features with missing values with mean. Can you explain why did we drop the feature with largest portion of missing values in this case?

Solution. We want to drop columns with the largest portion of missing values because they do not contain a significant amount of data. ■

Q9. Show that there are no missing values in the data at this point.

Solution. As shown in `Answer.jpynb` ■

Q10. Feature "left" has values 'yes' and 'no'. Convert those values into integer values, 1 and 0.

Solution. As shown in `Answer.jpynb` ■

Q11. Save the resulting data into a file.

Solution. As shown in `Answer.jpynb` ■

Q12. What else do you observe in in this data set?

Solution. As shown in `Answer.jpynb` ■

Q13. Divide the data into training set(80%) and test set(20%) and show the number of data points in each by following

1. Uniform sampling

   Solution. As shown in `Answer.jpynb` ■

2. Stratified sampling based on the ratio of "yes" and "no" values of feature "left"

   Solution. As shown in `Answer.jpynb` ■

3. How would you do the test set and training set sampling and why?

   Solution. In uniform (or random) sampling, there is no bias. From a dataset of size n, we will randomly choose a sample of size $0.8n$ with **replacement** for the *training set*. Sampling with replacement means that each item in the dataset has an equal chance of being selected - it's possible that one element may be picked more than once. We also do sampling with replacements for *testing set* but now the size of sample is $0.2n$.

   With stratified sampling, there is a bias. The original data D will be classified or divided into subgroups, $D_1, D_2, D_3, ..., D_n$. For each D_i, we can get a training set T_i and a test set V_i by applying random sampling, as mentioned earlier. At last, the entire training set is $T = T_1 \cup T_2 \cup ... \cup T_n$; and the entire testing set is $V = V_1 \cup V_2 \cup ... \cup V_n$.

   Random sampling is best when the population doesn't have clear groups, and we just want a simple, unbiased sample that represents everyone. It's easy to do and works well for general studies. Stratified sampling is better when the population has different groups, like age or income levels. It makes sure each group is properly represented, which helps get more accurate results, especially if we want to compare the groups. ■