(수안)

Hello everyone, We are group 3. And I am Suahn Lee, and my team members are Minjeong Kim and Yoonji Lee. Today we will have a brief presentation about what we have figured out so far, and what we are planning to do from now on.

This is the table of contents.

—-------------------------------------------------------------------------------------------------------------

(윤지)

First and foremost, we actually found out that the given data set is from Portugal.

It all started with exploring the variables of the data set. To be specific, we focused on the fact that the data set includes a variable called ‘euribor3m’. This means that this country is a member of the European Union. Also, there is another variable named ‘education’ and this category indicates that this country has a basic 4, 6, 9 years of education before high school. Surprisingly, among the 27 countries in the EU, there is only one country that has this system, which is Portugal.

For example, this is education system of Spain,

the left side is from Netherlands, and the right side is Germany, and this shows education system of France. None of them has 4, 6, 9 years of basic education system.

To be sure, we checked if the actual Consumer Price Index of Portugal matches the given data set. We looked for the exact same value in that graph, and every value had a matching real value. By doing this, we were able to ensure that this data set is from Portugal, and also find the year of each consumer price index value, which was all measured between the years of 2008 to 2010.

This led us to make a new column called ‘year’.

----------------------------------------------------------------------------------------------------------

(수안)

Then, we moved on to data visualization and preprocessing.

This is the boxplot we created. Thanks to this boxplot, we noticed the values of ‘pdays’ were abnormal. As you can see, there’s a huge gap between twenty-seven and 999. So we deleted that column.

Then we checked the existence of duplicated values, and fortunately, nothing was found.

For the next step, data smoothing, we labeled individuals who were over 65 years old as ‘retired’, referring to the average retirement age in Portugal from 2008 to 2010. Furthermore, we filled the unknown data from job, marital, and loan variables with the mode value of each variable. We also changed the ‘month’ variable from a categorical to a discrete variable.

—--------------------------------------------------------------------------------------------------------

(민정)

We also created a heatmap to figure out the correlation between variables, and we found that these three variables, employment variation rate, euribor of 3 months, and number of the employed have high Pearson correlation coefficients. As you can see, the correlations of each variable are 0.95, 0.91, and 0.97, respectively, and this leads to a multicollinearity problem, so we removed two of those variables.

Based on our EDA, we decided to use tree-based models with ensemble techniques since tree-based models and ensemble techniques tend to show better performance in categorical variables. Lastly, we noticed that the target ratio of “No” is around 11%, so we will solve this target imbalance problem by introducing both undersampling and oversampling.

This is the end of our presentation, thank you for listening.