TEAM MEMBER’S DETAILS

Group Name: Banking Insights Squad

Group: Members: Canmert Demir & Joseph Pang

Names: Canmert Demir-Bank Marketing (Campaign) -- Group Project

Email: canmertdemir2@gmail.com

Country: Turkey

College/Company: Msc Bartin University-Applied Mathematics / Data Glacier

Specialization: Data Science

Name: Joseph Pang-Bank Marketing (Campaign) -- Group Project

Email: joseph302156@gmail.com

Country: United States

College/Company: University of California, Berkeley/ Data Glacier

Specialization: Data Science

Github Repository: https://github.com/Canmertdemir/Canmertdemir-DataGlacierWeek9

Problem Description

Handling NaN values in a dataset involves strategies to manage missing or undefined values within the data. NaN values can adversely affect analysis and modeling processes, leading to biased results or model inaccuracies. One approach is imputation, where missing values are replaced with estimations such as mean, median, or mode values of the respective columns. Alternatively, model-based imputation involves predicting missing values using other features in the dataset. Another technique, Weight of Evidence (WOE), calculates the information value of variables and replaces NaN values based on their respective information values. Moreover, NaN values can sometimes carry information themselves, representing a distinct category or indicating an absence of data. Therefore, analyzing and handling NaN values appropriately is crucial to maintaining data integrity and ensuring accurate analysis and modeling outcomes.

Problem Approach

Joseph and I conducted this study using different techniques and analyzed the data. Through these analyses, we reached significant conclusions. Initially, Joseph built a model by dropping NaN values, while I filled categorical variables with mode and numerical variables with median values to create a model. Joseph worked with RandomForest, whereas I used

CatBoostClassifier. According to the AUC metric, CatBoostClassifier yielded the best results. We performed hyperparameter optimization on this model and achieved an even higher AUC score. Finally, we merged our codes to take higher score.

Correlation Analysis:

Identify correlations among numerical features using the `cor_analiz_cardinals()` function.

- Take note of significant correlations, particularly those associated with the target variable (`y_bool`).

Variable Categorization:

Use the `grab_col_names()` function to categorize variables into different types:

- Categorical Variables: Encompassing both categorical and numerical categorical variables.
- Numerical Variables: Pure numerical variables excluding categorical representation.
- Categorical but Cardinal Variables: Categorical variables exhibiting high cardinality.
- Numerical but Categorical Variables: Numerical variables that behave similarly to categorical ones.

Analysis of Variable Distributions and Missing Values:

This analysis focuses on studying the attributes "**cons.conf.idx**", "**cons.price.idx**" ,"**emp.var.rate**", "**euribor3m**", "**nr.employed**", "**duration**", "**pdays**", "**previous**", and "**y_bool**."

Addressing Missing Values (value replacement, drop):

Last week we found missing data in our dataset:

One missing value is observed in the "**duration**" variable; "**pdays**" has two missing values, "**previous**" variable contains 3523 missing values, and "**y_bool**" records 3668 missing values.

Tested two methods in addressing these missing values: replacing the missing values with the median or dropping the rows/columns of the missing values with dropna(): Through ROC testing we came to a test set accuracy of 0.901699 for when we replaced the missing values with the median and for dropping the missing values the test set accuracy was 0.900759. The difference in the test set accuracy of both methods is very minimal and does not seem to prove whether one is better than the other.