ML Project - Bank Marketing Prediction

Tasks to perform

Read in the file and get basic information about the data, including numerical summaries.

- Describe the pdays column, make note of the mean, median and minimum values. Anything fishy in the values?
- Describe the pdays column again, this time limiting yourself to the relevant values of pdays. How different are the mean and the median values?
- Plot a horizontal bar graph with the median values of balance for each education level value. Which group has the highest median?
- Make a box plot for pdays. Do you see any outliers?

The final goal is to make a predictive model to predict if the customer will respond positively to the campaign or not. The target variable is “response”.

First, perform bi-variate analysis to identify the features that are directly associated with the target variable. You can refer to the notebook we used for the EDA discussion.

- Convert the response variable to a convenient form
- Make suitable plots for associations with numerical features and categorical features’

Are the features about the previous campaign data useful?

Are pdays and poutcome associated with the target?

If yes, and if you plan to use them – how do you handle the pdays column with a value of -1 where the previous campaign data is missing? Explain your approach and your decision.

Before the predictive modeling part, make sure to perform –

- The necessary transformations for the categorical variables and the numeric variables
- Handle variables corresponding to the previous campaign
- Train test split

Predictive model 1: Logistic regression

- Make a predictive model using logistic regression
- Use RFE to select top n features in an automated fashion (choose n as you see fit)
- Using p values and VIF, get rid of the redundant features
- Estimate the model performance using k fold cross validation
- What is the precision, recall, accuracy of your model?
- Which features are the most important from your model?

Predictive model 2: Random Forest

- Make a predictive model using random forest technique
- Use not more than 50 trees, and control the depth of the trees to prevent overfitting
- Estimate the model performance using k fold cross validation
- What is the precision, recall, accuracy of your model?
- Using the feature importance values from the Random Forest module, identify the most important features for the model

Compare the performance of the Random Forest and the logistic model –

- Evaluate both models on the test set
- Which metric did you choose and why?
- Which model has better performance on the test set?
- Compare the feature importance from the different models – do they agree? Are the top features similar in both models?