About Dataset

A Portuguese Bank found that many of their customers are not investing in long term deposit accounts. The bank would like to identify customers that have a higher chance of subscribing for a long term deposit account and focus their marketing efforts on such customers.

Data Preprocessing

Check for null/missing or duplicate values.
View data types and counts of the target variable.

Data Analysis

View statistical information on the numerical data:

Avg Age = 41
Ave Deposit Amount = $1362.27
Min Deposit Amount = -$8019
Max Deposit Amount = $102127

Blue Collar, Management, and Technicians contacted most.
Married people contacted most.
Secondary/high school contacted most.
Most contact in the month of May and summer months.
Retired, married, and college graduates have the highest average balance.
Age 84 has the highest average balance - likely from an outlier.

More Preprocessing

Encode/transform categorical data into numeric labels so the model can read and use the data.

Logistic Regression Model

Split the dataset into 70% training data and 30% test data.

lr_model.fit - input the training data to the logistic regression model - "teach" the model *how* to make predictions using the training data.

Coefficients

Coefficients represent the log odds that an observation is in the target class (subscribed). Apply the exponential function to convert to "regular" odds ratio to compare the odds of the event (subscribing) occurring for each category of the predictor relative to the reference category.
Negative coefficients means less likely to open a deposit account.

As age increases by one year, the odds of that aged person will be in the target class is nearly 100%. As the number of contacts made during this campaign increases by one, the odds of being in the target class is 47.6% or nearly 50%.

Intercept is log odds of the outcome.

Predictions

Make predictions on the test data.

accuracy_score - measures the correct predictions out of the total number of predictions made.

Classification Report

F1-Score is the harmonic mean of precision and recall. Closer to 1, the better the model.

hyperparameter -> max_iter = 20000

KNeighbors Classification Model

Find a predefined number of training samples closest in distance to the new point and predict the subscribed label from these. Classify based on nearest data points.