BAN 5753

Mini Case 2

Due: November 16th, 2022

By Zachary Vann, Divine Bennett, and Kailey Wilson

zachary.vann@okstate.edu

[divine.bennett@okstate.edu](mailto:divine.bennett@okstate.edu)

kailey.r.wilson@okstate.edu

**Background**

The objective of this case is to use Spark to explore a dataset and identify clients who will subscribe for a term deposit. For this, the Spark engine was used in Python (PySpark) to run analysis to find out more information about the different variables in the dataset and run multiple models to find which performs the best.

**Data Description and EDA**

The provided data set is from telephone calls from XYZ bank’s direct marketing campaign and consists of 21 columns (including the target variable Y) and 41188 observations. The dataset can be broken into four different sections, the first being demographics (variables related to age, marital status, housing, and loan status), the second pertaining to the latest contact for the campaign (contact type, month, day, and duration), the third related to other attributes (such as the number of contacts performed, the previous number of contacts performed, and outcome of prior campaigns), and the final pertaining to social and economic attributes (such as employment variation rate, consumer price index and the number of employees).

To better understand the dataset, an exploratory analysis was performed. Figure 1 shows the number of distinct (cardinality) and missing values for each variable (none of the variables are missing data). Variables such as duration, euribor3m, age, and campaign rank among the highest with binary variables such as y (target variable) and contact having the least. Figure 2 shows the summary statistics, skewness, kurtosis, and distribution of all the numeric variables. These figures show that the variables with the most skew are duration, campaign, and previous (all highly skewed to the right), and pdays (highly skewed to the left).

Also, Figure 3 shows a Pearson Correlation matrix that was completed before any transformations on the categorical variables. Based on this matrix, there are several variables where the correlation is greater than .5 (strong positive) and -.5(strong negative).

**Methods**

After the EDA was conducted, the data was then transformed to be able to use in modeling. The first step was to create an index value for all categorical variables. After this, all the categorical variables were transformed using one-hot encoding as well as encoding the target variable. Next, all the variables were put into a vector assembler to place everything into one data frame column (everything except the target variable). The vectorized column was then again transformed by a standard scaler to produce similar ranges.

After these transformations, the data was then split into a train (.70 with 28786 observations) and test (.30 with 12402 observations), and the data was then put into several different models. The first model that was run was a logistic regression, the second model was a decision tree, the third a random forest, the fourth a gradient boosting, the fifth a neural network, and the sixth a gradient boost with tuned parameters.

**Results**

|  |  |  |
| --- | --- | --- |
| **Model** | **Area under ROC** | **Area under PR** |
| **Logistic Regression** | **.9328** | **.5588** |
| **Decision Tree** | **.8360** | **.5578** |
| **Random Forest** | **.9219** | **.5839** |
| **Gradient Boosting** | **.9468** | **.6700** |
| **Neural Network** | **.5000** | **.1140** |
| **Gradient Boosting (with parameter tuning)** | **.9472** | **.6700** |

**Table 1: Results of Modeling**

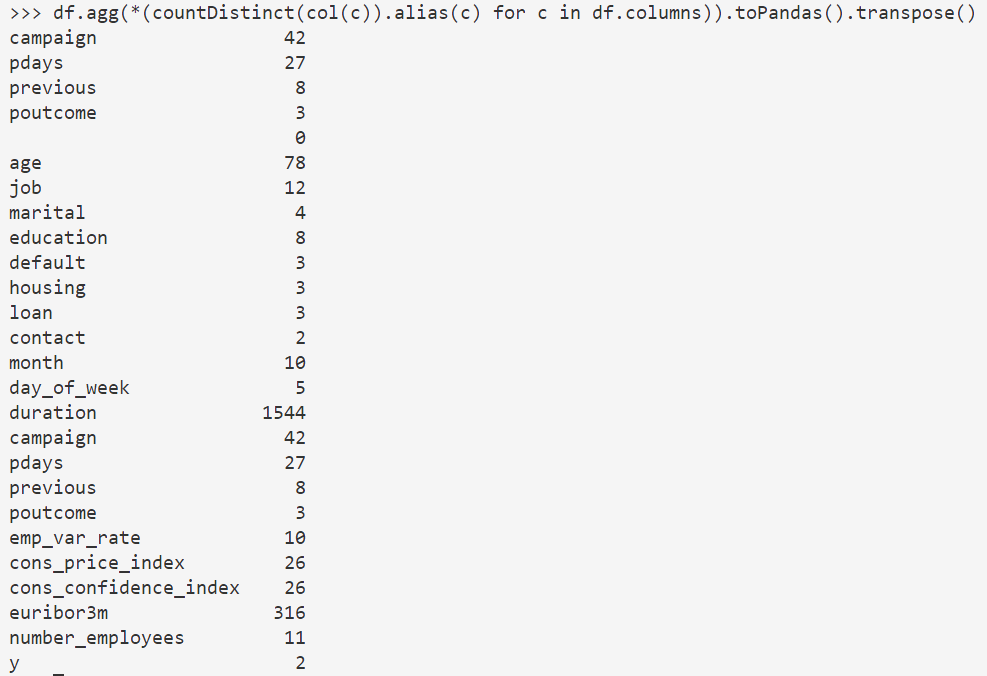
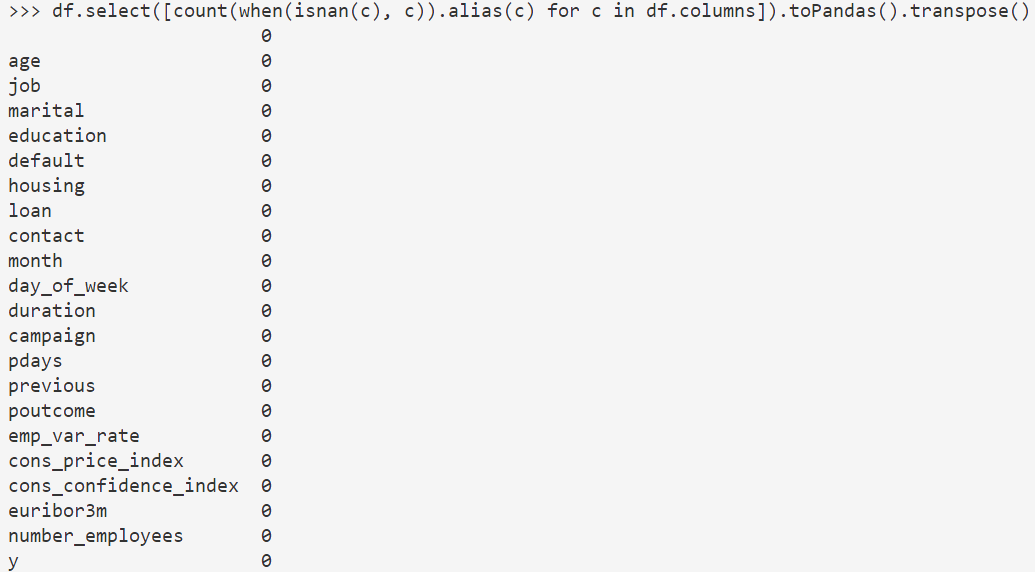
The results of all the modeling is shown in Table 1. Based on the table the models that performed the best are the gradient boosting models with or without the parameter tuning. These models have the best area under ROC and under PR curves with .9468 (without tuning) and .6700 respectively. With tunning, the gradient boosting model performed slightly better with the area under ROC improving to .9472.

To get a feel of which variables were important, feature importance was pulled from the logistic model (Figure 4) and the gradient-boosted model (Figure 5). They have similar results but the top 5 most important features for the logistic regression are age, duration, campaign, pdays, and previous while for the gradient-boosted model they are duration, number\_employees, euribor3m, cons.confidence.index, and age.

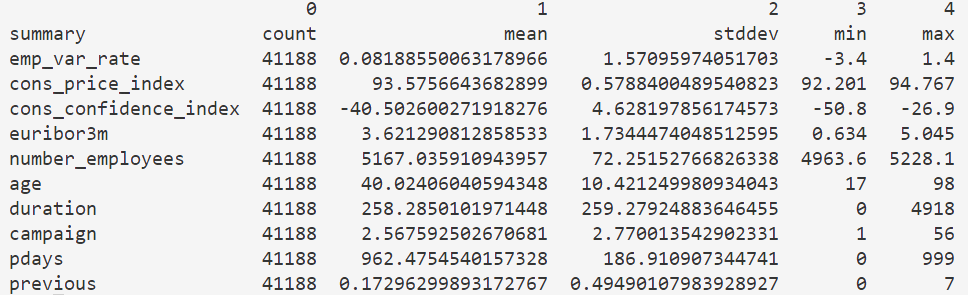
**Insights**

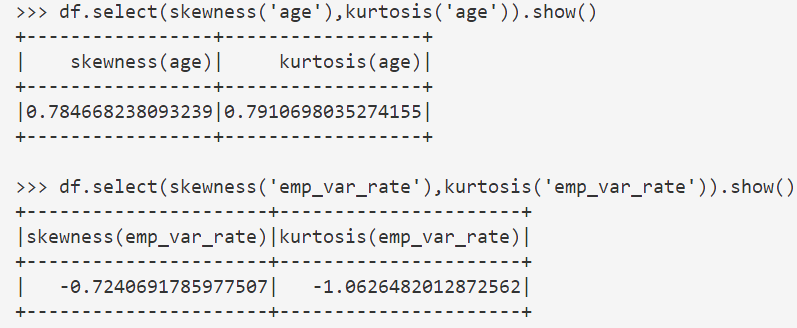
Based on the work completed, what we recommend that the bank do is to keep potential customers engaged. This can be done by having longer conversations with them and ensuring the bank is staffed adequately. Factors outside of the bank's control such as euribor3m and cons.confidence.index is also very important and should be monitored but nothing much else can be done to control these.

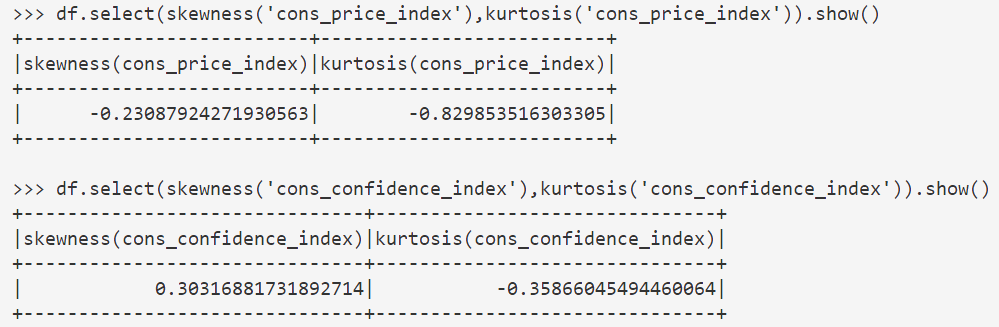
**Appendix**

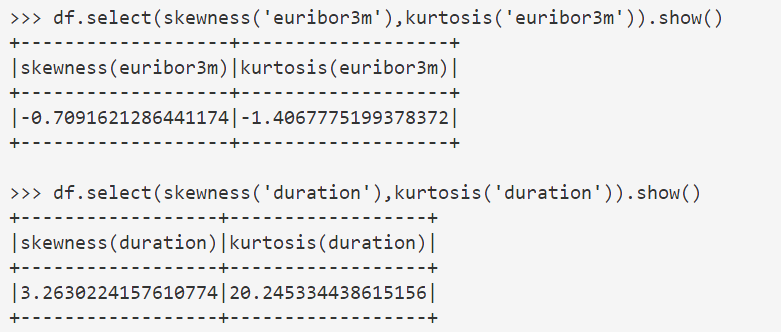
****

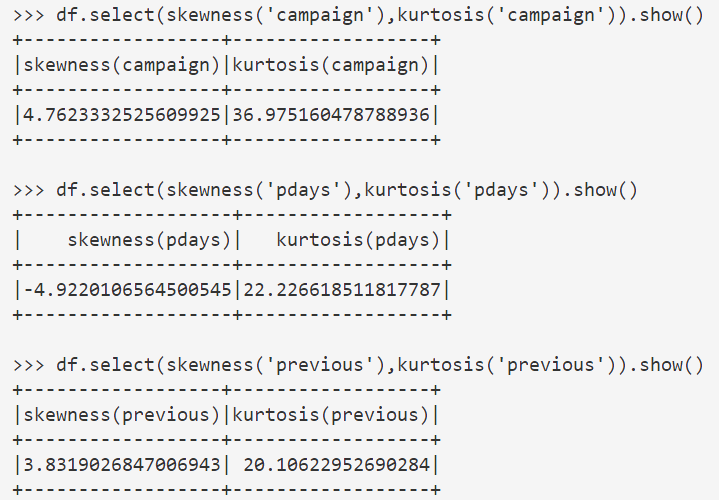
**Figure 1: EDA Part 1 Distinct and Missing Data**

****

****

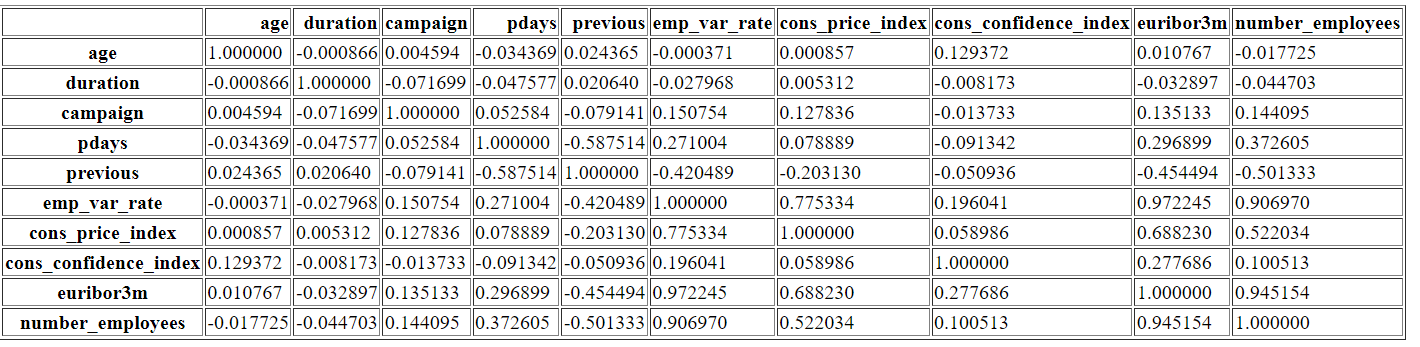
****

****

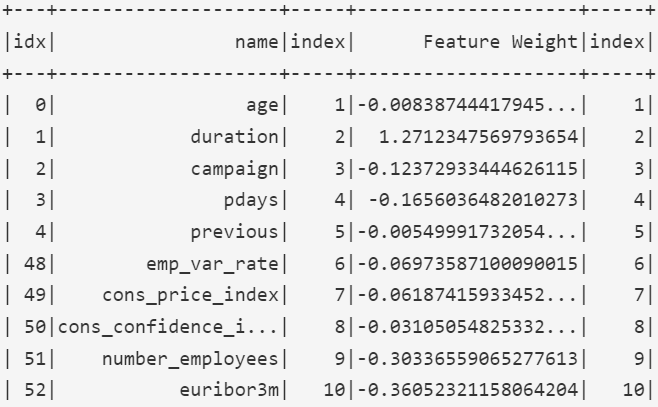
****

****

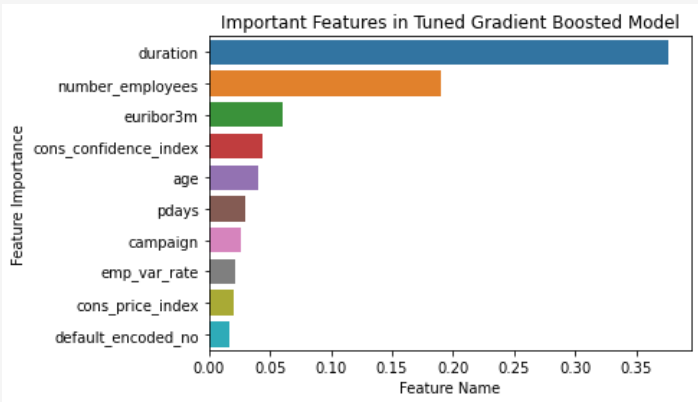
**Figure 2: Summary Statistics, Skewness, and Kurtosis of Numeric Data**

****

**Figure 3: Pearson Correlation Without Transformations of Cat Variables**

****

**Figure 4: Logistic Regression Important Features**

****

**Figure 5: Gradient Boosted Important Features**