BAN 5753

Advanced Business Analytics

Fall 2023

Mini Project 2

SPARK

Team 10 (DDKM)

Divya Pamu

Dylan Stiles

Kourtni Conner

Manik Mogali

Content

➢ Executive Summary

This project is a strategic collaboration with XYZ Bank to provide them with insights that will enable them to better align their marketing initiatives with the ever-changing needs of their customer base. XYZ Bank sought to enhance its direct marketing campaigns by accurately predicting customer subscriptions to term deposits.

Our diverse approach—which includes both advanced modeling techniques and in-depth data exploration—allowed us to provide a comprehensive solution that goes beyond outcome prediction to influence future tactics. Leveraging logistic regression and a subsequent random forest model, we achieved accuracy rates of 90.45% and 89.75%, respectively.

Key features influencing predictions, such as "duration_group_index," "euribor3m," and "pdays_missing," were identified through feature importance analysis. The outcome allowed us to provide actionable recommendations, urging XYZ Bank to focus on previously contacted customers, limit contact durations, and time campaigns with favorable Euribor 3-month rates, optimizing their marketing strategy for higher enrollment.

In conclusion, our analysis provides XYZ Bank with a robust predictive model and strategic insights. By tailoring future campaigns based on our recommendations, the bank can effectively target customers with a higher propensity to subscribe to term deposits, thereby improving overall campaign success. These recommendations will help XYZ Bank improve their marketing tactics and better target audiences with their campaigns.

➢ Introduction

The banking sector is going through a transitional period, and companies like XYZ Bank are looking for creative ways to improve their marketing approaches. In this project, the focus of our analysis is to address a critical business issue for XYZ Bank: identifying the customers who are most likely to sign up for a term deposit. The efficiency of these forecasts has a big impact on how well the bank allocates its resources and runs its marketing initiatives.

➢ Project Objectives

The primary objective of our project is to create a prediction model that can determine if a customer will sign up for a term deposit ('yes') or not ('no'). As the basis for well-informed decision-making is Exploratory Data Analysis (EDA), our goal is to thoroughly examine the dataset to identify relationships and trends in data. With an emphasis on the champion model, our aim is to investigate a range of methods and offer insights into the advantages and interpretability of the selected strategy. Beyond forecasting, our objective is to offer practical suggestions derived from our research.

➢ Data Introduction/Sourcing

The "XYZ_Bank_Deposit_Data_Classification.csv" dataset that was used in this analysis comes from XYZ Bank's direct marketing campaign. The data, which has 20 columns and 41,188 rows, covers the period from May 2008 to November 2010. The dataset is primarily focused on telephone-driven marketing campaigns, in which it was frequently necessary to make several contacts with the same client in order to determine whether or not they would subscribe to a term deposit.

➢ Data Cleaning/Preparation

Steps taken:

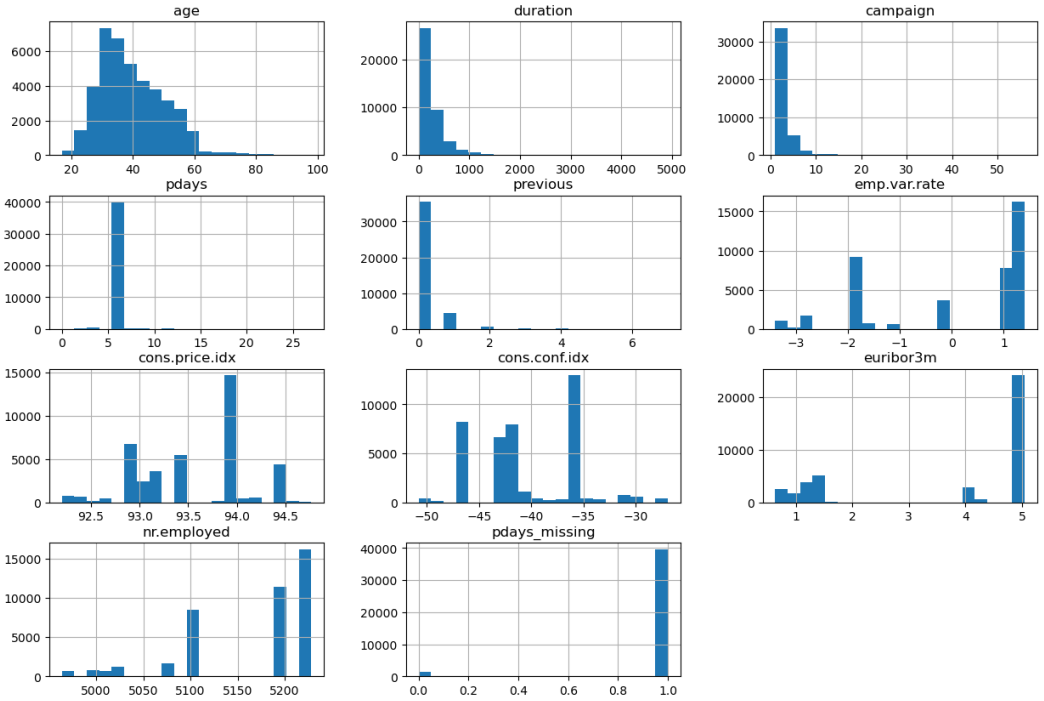
- Handling the "999" variable in the "Pdays" column.
- Dropping columns "nr.employed" and "emp.var.rate" to address multicollinearity.
- Normalizing the "duration" column using standard scaler.
- One-hot encoding of string variables.
- Binning for the "age" and "duration" columns.
- Splitting the data into training and test sets (70/30 split).

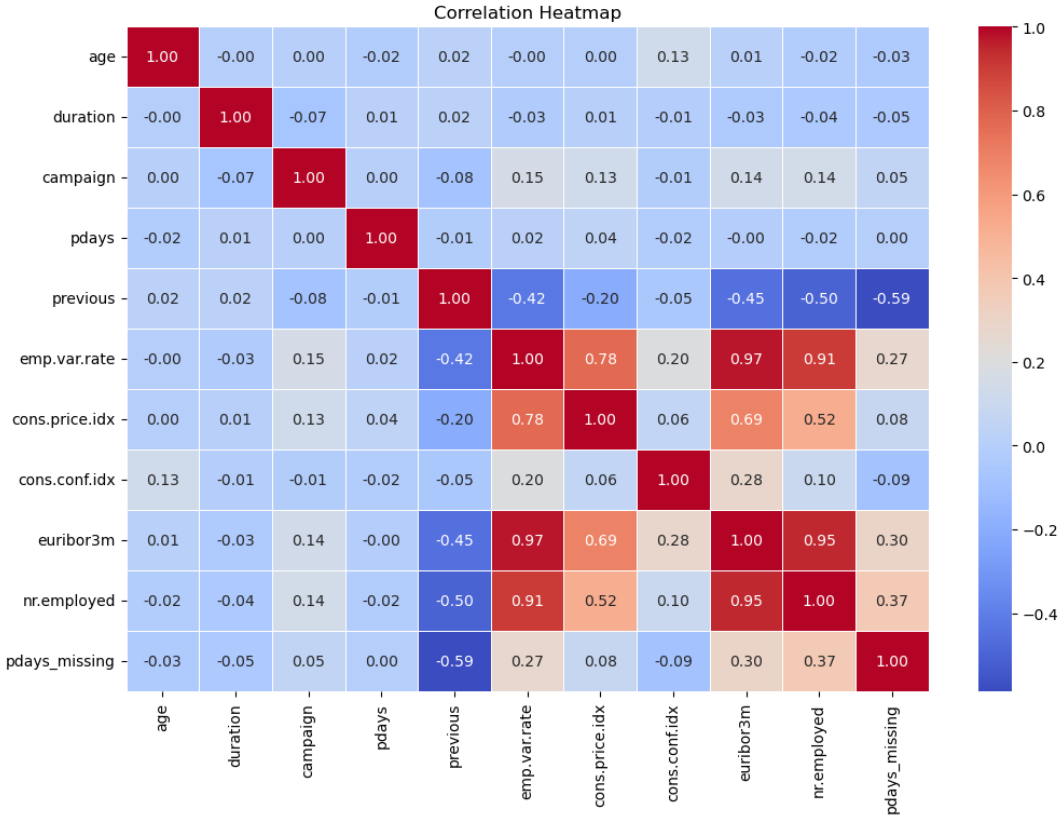
To guarantee the caliber and efficacy of our prediction models, data preparation was an essential stage in our project. The "999" variable in the "Pdays" column was handled to address any missing or ambiguous values. The column mean was substituted for the corresponding values in a newly created column that indicated the presence of "999." The columns "nr.employed" and "emp.var.rate" were removed in order to prevent multicollinearity.

Furthermore, because of its large range, the "duration" column was normalized using a standard scaler. To speed up our models' learning process, one-hot encoding was applied to string variables. The "age" and "duration" columns underwent binning in order to simplify the data and minimize the possibility of overfitting. To prepare for model training, the dataset was split into training and test sets using a 70/30 split. The model was then trained using logistic regression and predictions were made on the test set. After checking for accuracy, our model predicted the test set with 90.45% accuracy.

➢ Data Visualization/Analysis

Conducted Exploratory Data Analysis (EDA) to identify: Relationships and trends in the data, Correlations, Bivariate analysis of target versus input variables, Univariate patterns, and Missing data.


age
duration
campaign
pdays
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
pdays_missing


Correlation Heatmap
age 1.00 -0.00 0.00 -0.02 0.02 -0.00 0.00 0.13 0.01 -0.02 -0.03
duration -0.00 1.00 -0.07 0.01 0.02 -0.03 0.01 -0.01 -0.03 -0.04 -0.05
campaign 0.00 -0.07 1.00 0.00 -0.08 0.15 0.13 -0.01 0.14 0.14 0.05
pdays -0.02 0.01 0.00 1.00 -0.01 0.02 0.04 -0.02 -0.00 -0.02 0.00
previous 0.02 0.02 -0.08 -0.01 1.00 -0.42 -0.20 -0.05 -0.45 -0.50 -0.59
emp.var.rate -0.00 -0.03 0.15 0.02 -0.42 1.00 0.78 0.20 0.97 0.91 0.27
cons.price.idx 0.00 0.01 0.13 0.04 -0.20 0.78 1.00 0.06 0.69 0.52 0.08
cons.conf.idx 0.13 -0.01 -0.01 -0.02 -0.05 0.20 0.06 1.00 0.28 0.10 -0.09
euribor3m 0.01 -0.03 0.14 -0.00 -0.45 0.97 0.69 0.28 1.00 0.95 0.30
nr.employed -0.02 -0.04 0.14 -0.02 -0.50 0.91 0.52 0.10 0.95 1.00 0.37
pdays_missing -0.03 -0.05 0.05 0.00 -0.59 0.27 0.08 -0.09 0.30 0.37 1.00
age duration campaign pdays previous emp.var.rate cons.price.idx cons.conf.idx euribor3m nr.employed pdays_missing
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

➢ Results

Logistic regression model:

- Training the model.
- Making predictions on the test set.
- Accuracy of the model on the test set (90.45%).

Random forest model:

- Rationale for choosing random forest: Due to its ability to interpret feature importance for our prescriptive analytics as well as its ability to handle categorical data. In addition to those benefits, random forest also can test for uncertainty in our predictions that can test the reliability of our model for XYZ Bank.
- Application of the vector assembler.
- Training the model.
- Accuracy of the final model on the test set (89.75%).

With our final model running an acceptable accuracy score of 89.75%, we turned to the prescriptive portion of our analysis by running code for feature importance. This helped us determine some of the key features our random forest utilized to make its predictions. Based on the outcome of that code, our model identified the "duration_group_index", "euribor3m", and "pdays_missing" entries as the three most important features.

In order to determine how those features factored into these models' decision we ran further code that grouped the "y_index" by its outcome and then aggregated for the mean output of each of those features. With the mean of both groups in the "pday_missing" entry being close to 1, it is safe to assume that while the feature may have been important in our model's selection, a majority of the customers in the analysis had been contacted previously.

Feature	Importance
duration_group_index	0.34459839339470066
euribor3m	0.20079082748395233
pdays_missing	0.10952086725162195
poutcome_index	0.08347786712956108
pdays	0.08062445572947462
cons_conf_idx	0.059005192273336185
month_index	0.05838095904045756
cons_price_idx	0.014888883087492047
day_of_week_index	0.010147820202797015
contact_index	0.009644559824238886
previous	0.008335559778918266
age_group_index	0.005629240300588008
job_index	0.004767504221643915
education_index	0.002933736912124104
campaign	0.002278275613033...
default_index	0.002151659577467...
marital_index	0.002096486103669356
loan_index	4.676920435374125E-4
housing_index	2.600200313848912...

y_index	avg(euribor3m)
1.0	2.1231351293103313
0.0	3.8114911623072825

y_index	avg(duration_group_index)
0.0	0.09201597898653825
1.0	0.6176724137931034

y_index	avg(pdays_missing)
1.0	0.7915948275862069
0.0	0.9850060194812301

➢ Recommendations

- Based on the results of our model, feature analysis, and breakdown of the top three features we were now able to make a recommendation to XYZ Bank to help identify and improve their enrollment.
- Given the average duration group index of 0.09 we can determine that the majority of the calls our model predicted to be positive enrollment were under 500 seconds and primarily in our “0” group.
- We can also conclude, based on the average, that the majority of positive enrollment predictions had an average Euribor 3-month rate of 2.12%.
- Therefore, our recommendations to XYZ Bank, based on our models’ predictions, would be to focus their next campaign towards previously contacted customers, to keep their contact durations under 500 seconds, approximately 8 minutes, and to focus on starting the campaign when rates are closer to 2.12% or lower. By doing so, they should be able to target customers who are more susceptible to enrolling.