



Data Glacier

Your Deep Learning Partner

Bank Marketing Campaign Classification

Week 11: Deliverables - EDA Presentation & Model Recommendation

Name: Elissa Kuon

Email: e.kuon491@gmail.com

Country: United States

College: University of Houston

Specialization: Data Science

Batch Code: LISUM17

Date: March 16, 2023

Submitted to: Data Glacier



INTRODUCTION AND MOTIVATION

Even though the banking industry spends a lot of money on marketing these days, banks must improve the efficiency of their marketing plans. Traditional marketing strategies have not helped banks expand their operations. They used direct marketing to offer long-term deposits at competitive rates of interest to the general public, despite the process being time-consuming and the chance of success being low. By understanding consumer wants can result in more intelligent product design, more successful marketing strategies, and higher levels of customer happiness. The bank will be able to forecast consumer saving behaviors and determine which customers are most likely to make term deposits by looking at customer attributes like demographics and transaction history. Following that, the bank can concentrate its marketing efforts on such clients. As a result, the bank will be able to safeguard deposits more effectively and improve customer satisfaction by omitting campaigns that are inappropriate for particular clients. The Portuguese Banking Institution supplied data on marketing initiatives that were based on phone calls. This data will be used to assist the banking industry in determining which clients will sign up for a term deposit.

The purpose of this project is to use machine learning approaches to discover previously undiscovered patterns, maps, and various input variables that can be used to categorize whether or not customers will subscribe for longer deposits. We think this is significant because it will help banks better understand their customer base, predicting how customers will react to their telemarketing campaign, and create a target customer profile for the next marketing initiatives.

PROJECT PLAN LIFECYCLE

To keep track of our progress on this project, we established a timeline with important deadlines and a working plan in mind for each week prior to the deadline.

Weeks	Dates	Plan
Week 07	February 19, 2023	Problem Statement, Data Collection, Data Report
Week 08	February 26, 2023	Data Understanding
Week 09	March 02, 2023	Data Preprocessing
Week 10	March 09, 2023	Exploratory Data Analysis
Week 11	March 16, 2023	Building the Model
Week 12	March 23, 2023	Model Result Evaluation
Week 13	March 30, 2023	Final Submission (Report + Code + Presentation)



DATA DESCRIPTION

The Portuguese Banking Institution donated four separate datasets of marketing data to the UCI Machine Learning Repository that range in time from May 2008 to November 2010. Due to the fact that these two of the four datasets provided contain the institution's most recent marketing data, we will focus on them. Fortunately, these two datasets originate from the same sample, and we will utilize one of them (bank-additional-full) for training the model and the other dataset for testing the model (bank-additional). The bank-additional dataset only includes 10% of the inputs from the bank-additional-full dataset, which has 41188 observations (client inputs) and 20 variables (client demographic and transaction history, consisting of a mixture of numerical and categorical types). To avoid any confusion between the bank-additional-full dataset and the bank-additional dataset, we will refer to them as the original dataset and testing dataset, respectively.

This dataset was still in its raw state, so we had to clean it up before creating the proper data visualization and classification models to comprehend the relationships between the features and ascertain whether the client will sign up for a term deposit.

DATA PREPROCESSING

Extracting Observations from Original Dataset Present in Testing Dataset

The inputs from the original dataset that are currently present in the testing dataset must be removed in order to prevent the same examples from appearing for both datasets. In order to get a more accurate result from our models, we must avoid our model from already learning from the "unknown" inputs.

Checking for Missing and Duplicate Values

Fortunately, none of the values that were currently available in the raw format were missing. In order to make the original data more generalizable, we dropped 11 duplicate indexes that we discovered when we searched for any duplication.

Checking for Skewness and Kurtosis

We took notice of the high skewness and kurtosis values for the variables duration, campaign, pdays, and previous, which could be signs of outliers. There were outliers, as can be seen by carefully examining the boxplot distributions for these variables as well as the 5-number summary (min, lower quartile, median, upper quartile, and max). Although there were far more outliers in the duration variable than in the other variables, we felt that capping the upper limit for this variable would be the best solution to this problem. Although the mean values are heavily influenced by outliers, we will substitute the median values for the outliers for the other variables. After managing the outliers, these modifications greatly improved the skewness and kurtosis for each variable. However we saw that pdays and previous have 0 skewness and kurtosis, indicating that there is only one value in the data for those variables. As the data only



contains one value, we can safely ignore these two variables and conclude that the clients were not contacted by a previous campaign and that there were no previous contacts made for this client before this campaign.

We must now proceed to make the identical adjustments to the testing dataset as we did for the original dataset.

Checking Classes Within the Categorical Variables

We discovered that certain classes have an "unknown" class when we examined the individual categorical variables and their classes. In order to prevent our model from detecting trends that do not exist, we took a closer look at the counts of these variables with the 'unknown' class, where we replace them with NaN.

As we looked at the counts, we noticed that the variable default had an unusually large number of NaN values; yet, only two observations supported the other class ('yes'), while the bulk of observations were in favor of the former ('no'). If we had more time, we could have used a classification classifier to sort the NaN into the 'yes' and 'no' categories. But, due to time restrictions, we will remove this variable, leading us to the conclusion that none of our clients have a client default. We move on to replace the other NaN values in the other variables with the most common classes with regard to those variables. Next, we recategorize some of the categorical variables which helps us simplify our results.

Similarly, we must now proceed to make the identical adjustments to the testing dataset as we did for the original dataset.

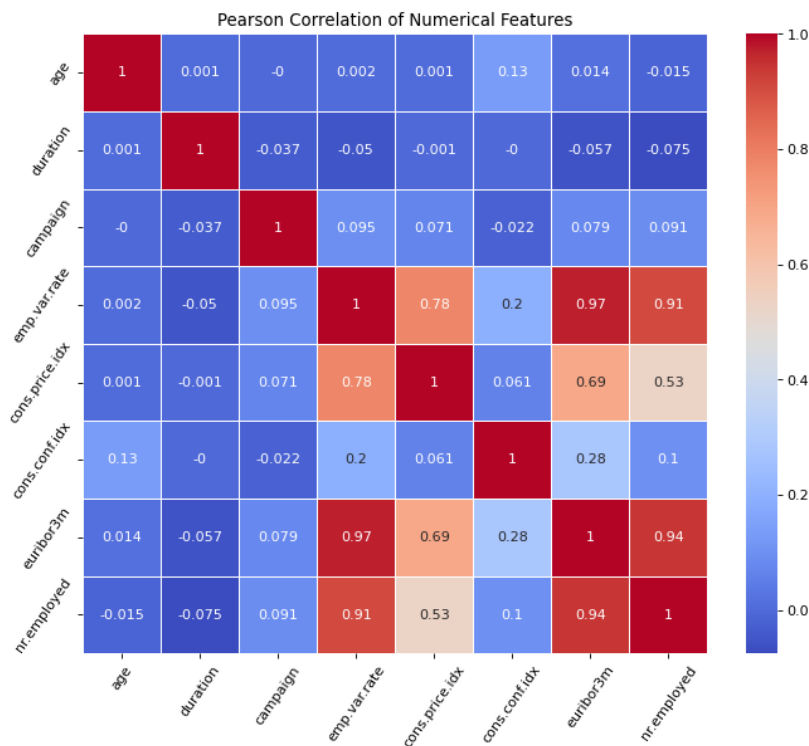
DATA VISUALIZATIONS AND INSIGHTS

Data visualizations are often used to give us a glimpse of the distribution and understatement of the clients' characteristics and their banking history. This can help banks in several ways.

Marketing plans aimed at certain customer segmentation can be used to analyze its approach in their banking campaigns. If such plans include elements that could boost profitability, banks will be better able to control the market by relying on the features and making future changes to their marketing efforts to retain their current customers and draw in new ones with similar traits. The relationship between the customers' banking history and visualization can also be studied in further detail. While beginning their marketing strategy, banks may be able to frame data models with much more clarity thanks to these insights.

Exploratory Data Analysis

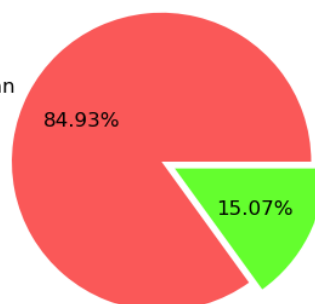
To check if there was any relationship between these predictors that we should be aware of, we will first look at the correlations for the numeric variables. According to the correlation heatmap, the more positively connected these two variables are, the redder the box is, and the more negatively correlated these two variables are, the bluer the box is. One of the key findings from the correlation heatmap we created is that, with the exception of the consumer confidence index, the social and economic traits are all positively connected with one another.



Following this, we developed an interest in finding out how the client's loan history was distributed. Most of our clients, as far as we can tell, have no personal or house loans. A substantial majority of clients—roughly 84.93 percent—did not have any kind of personal loan. Yet, just 54.73% of consumers were without any form of home loan, indicating a roughly equal distribution of those with and without mortgages.

% of Condition of Loans

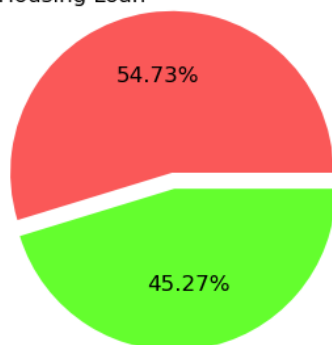
Did NOT Have Personal Loan



Did Have Personal Loan

% of Condition of Housing Loan

Did NOT Have Housing Loan



Did Have Housing Loan

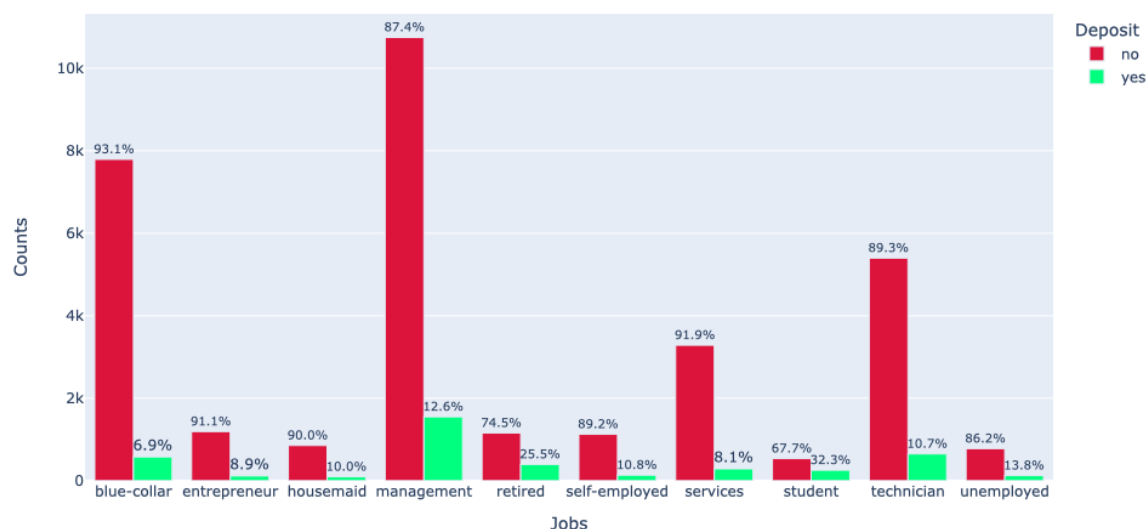


To get a better picture of our client's chances of using bank services, we looked at their age distribution based on occupation. The bank serves a varied variety of consumers from young to old; the oldest customer to use bank services is 98 years old, while the youngest user is 17. Together with providing services for a broad spectrum of clients' occupations, including employed clients, students, retirees, and those without a job. Overall, the client age distribution appears to be rather normal, and the bulk of bank clients are between the ages of 30 and 40. It's interesting to note that the majority of the clients come from the management sector, then blue-collar workers, service providers, and technicians. The group most unlikely to use bank services is students. Using the client's occupation, we examined the distribution of individuals who subscribed to a term vs those who did not. Most customers across all professions did not commit to a term. The difference between those who subscribed to a deposit and those who didn't is over half for retirees and students. Apart from these two, every other occupation has a significant difference.

Age Distribution by Occupation



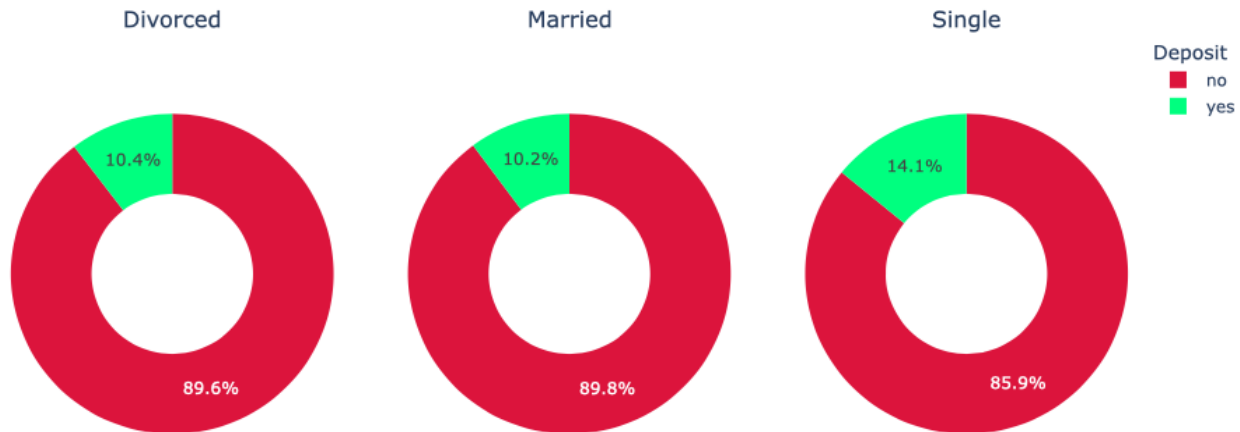
Term Depositors Based on Job Type





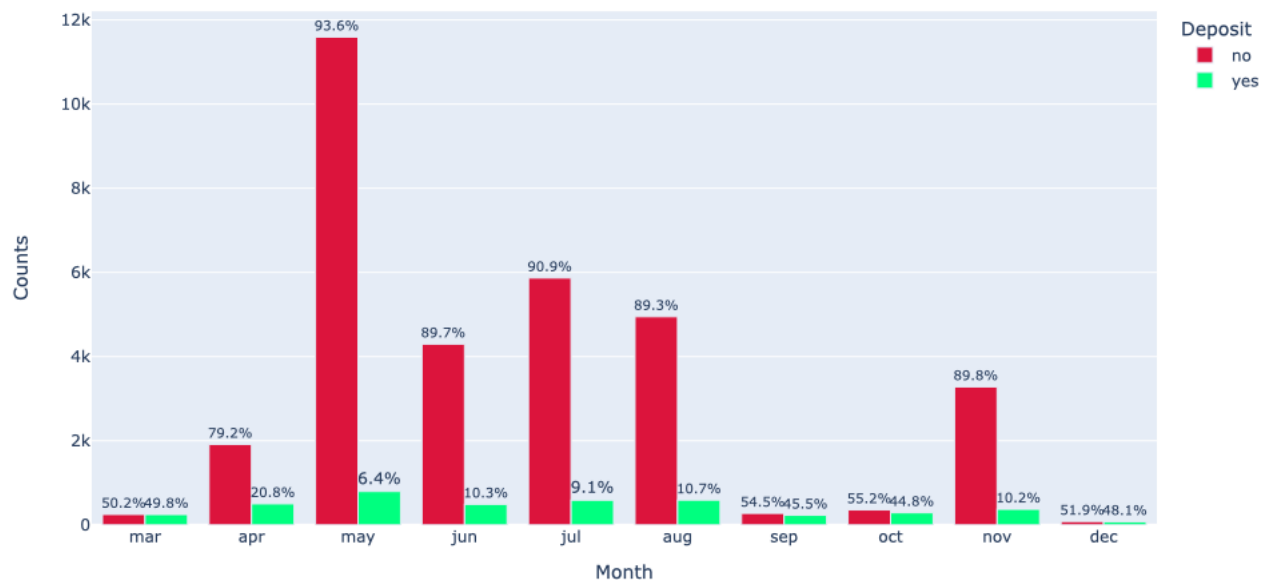
In accordance with the term deposit, we also examined the distribution of marital status. According to the distribution, most of the clientele are married, then single, then divorcees. We can observe that the three statuses, by a wide margin, did not agree with a sentence. Less than 2500 customers actually subscribed to a deposit for each marital status. To put it into perspective, despite being married, at least 85% of clients did not sign up for term deposits.

Term Deposits Based on Marital Status



Furthermore, we investigated the date of the client's most recent monthly communication with the bank. According to the plot, the bank contacts the majority of its clients in the months of May, June, July, and August. In the months of March, September, October, and December, very few of the clients are contacted. However, when we look at the differences we see that there is a better chance for subscribers to subscribe to a term deposit in these months as there is roughly a 50% chance of them subscribing. Despite the bank's interaction, there are still significantly more people who did not sign up for a term deposit than those who did.

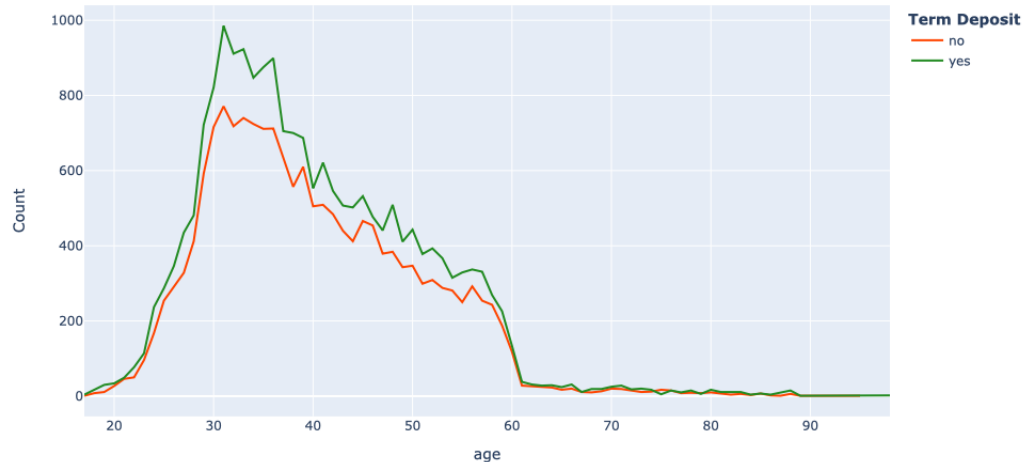
Deposits Based on Last Contact



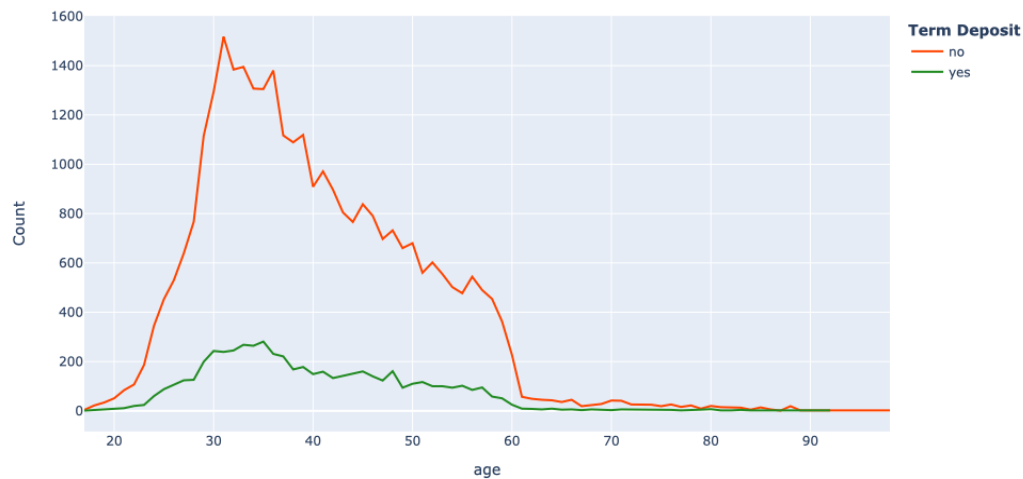


Returning to the client's age, we look at the changes based on their loaning history (personal and housing), as well as the rate of the term deposits. According to the data, none of the clients increased their personal loan subscriptions across all age groups. Along these lines, we also observe a distinct majority of customers across all age categories who do not subscribe for term deposits. Yet when we examine the study of housing loans, we can observe that there is a predominance of people in the mid-age group who did subscribe to a deposit. In contrast to the mid-age groups, where we do notice significant changes, there does not appear to be a distinct dynamic between those who deposited and those who did not during the past 60 years.

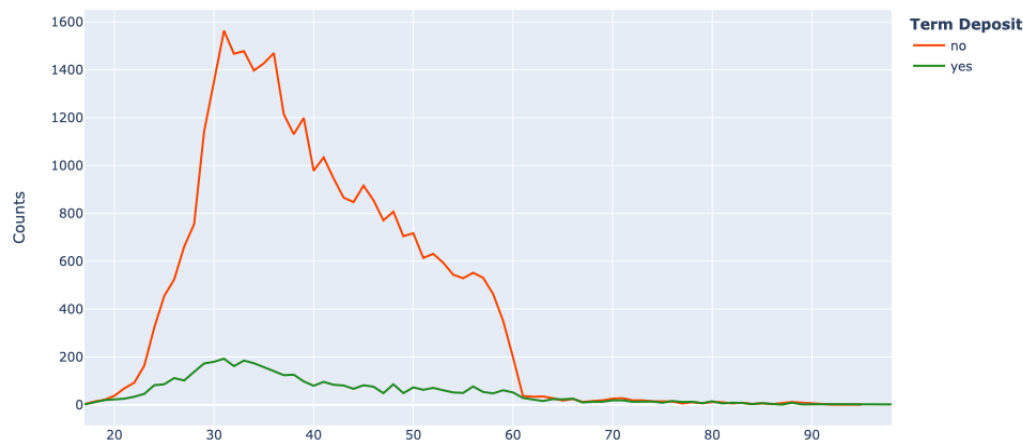
Effect of Age on Housing Loan



Effect of Age on Personal Loan



Effect of Age on Term Deposits

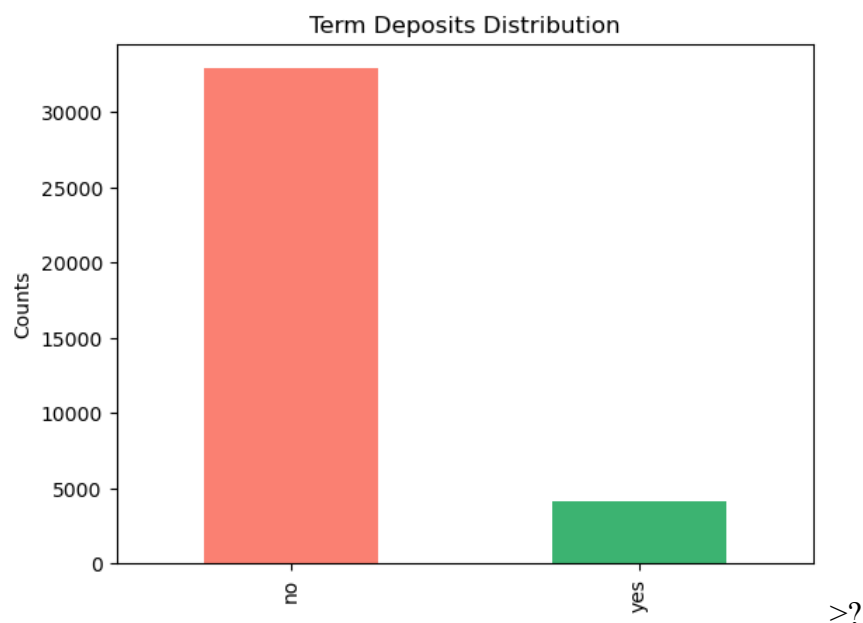




Then, using the number of contacts made during the campaign, we looked at the rate of change between those who did and did not subscribe to a term deposit. Our data shows that clients are less likely to sign up for term deposits the more frequently the banks contact them during the campaign.



There is a glaring difference between those who did and did not subscribe to a term deposit as we examined the correlation between the variables in our original dataset. Due to the large majority of clients who didn't subscribe to a term deposit, we can't really identify a distinct trait that distinguishes those of the bank's clients who did. As a result, we examined the term deposit distribution and found a glaring disparity between the two classes, as indicated. Before we start training the categorization models, we need to address this and make the necessary improvements.





Even while we were able to gather a lot of data from our investigation, we also learned a lot about the bank's customers, particularly those who did not sign up for a term deposit. Unfortunately, because the courses were unbalanced, we did not really get a meaningful analysis because the judgments we made were heavily centered on those who did not subscribe to a term deposit instead of examining important traits of those who did.

We will not undertake exploratory data analysis on our original dataset after 'balancing' the classes because of the time limits for this project.

Imbalancing the Classes

For this project, we will concentrate on balancing our target classes by employing the undersampling method, which reduces the number of observations from our original dataset to 8,376.

```
#UNDERSAMPLING
from collections import Counter
from imblearn.under_sampling import RandomUnderSampler

# summarize class distribution
print("Before undersampling: ", Counter(df['y']))

# define undersampling strategy
undersample = RandomUnderSampler(sampling_strategy='majority')

# fit and apply the transform
X_train_under, y_train_under = undersample.fit_resample(df.drop('y', axis=1), df['y'])

# summarize class distribution
print("After undersampling: ", Counter(y_train_under))

Before undersampling: Counter({'no': 32869, 'yes': 4188})
After undersampling: Counter({'no': 4188, 'yes': 4188})

X_train_under.shape
(8376, 17)
```

We can see that our classes are considerably more evenly distributed. We will now go ahead and fit our data into the classification models.

NOTE: Other methods, such as oversampling or a mix of undersampling and oversampling, can be used to balance our target classes. Due to timing, we will, however, be limited to undersampling because it is a little more less costly in terms of resources. However, we should be mindful of the drawbacks of utilizing undersampling techniques, which we can later study or choose to use a different strategy to balance the classes.

TRAINING AND TESTING DATASETS

We do not need to bother about dividing our dataset into training and testing sets, as was previously explained. But, in order to create our classification models smoothly and without errors, we must make certain changes to our dataset.

Numerical Variables - Scaling

The next step after balancing the target classes is to standardize the numerical variables, which enables us to scale various variables (or measures) uniformly. To achieve this, we first calculate the mean and standard deviation for the training set before standardizing the testing set with these precise numbers.



Categorical Variables - Encoding

Our category variables are then converted to numerical values, which will make modeling easier to understand and enable the extraction of any useful data. Label encoding and one hot encoding are the two types of encoding techniques that we will employ. For categorical variables with more than three distinct values or if they were ordinal, we utilized label encoding (e.g., education and jobs). We'll use one hot encoding for all the other categorical variables.

Creating a Dataset – Removing ‘Duration’ Variable

It's interesting to note that the variable duration has a significant impact on the target variable. Before making a call, the duration cannot be determined. The target variable is obviously known once the call is made. For a more accurate classification model, we will thus remove the duration variable from our training and testing sets. To serve as a benchmark and to contrast the datasets with and without the duration variable, we will also use the original training and testing set (with the duration variable).

BUILDING THE CLASSIFICATION MODELS

In this project, we want to use machine learning models to shortlist any bank customers by predicting whether they will sign up for a term deposit based on their previous interactions with banks. Methods that were considered in this study are Logistic Regression, Random Forest, and Boosting.