Data Glacier

Your Deep Learning Partner

Bank Marketing Campaign Classification

Week 8: Deliverables - Data Understanding and Starting Data Preprocessing

Name: Elissa Kuon
Email: e.kuon491@gmail.com
Country: United States
College: University of Houston
Specialization: Data Science
Batch Code: LISUM17
Date: February 26, 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.

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.

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