Bank Marketing Campaign Classification

Week 13: Deliverables - Final Report

Name: Elissa Kuon
Email: e.kuon491@gmail.com
Country: United States
College: University of Houston
Specialization: Data Science
Batch Code: LISUM17
Date: March 30, 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.

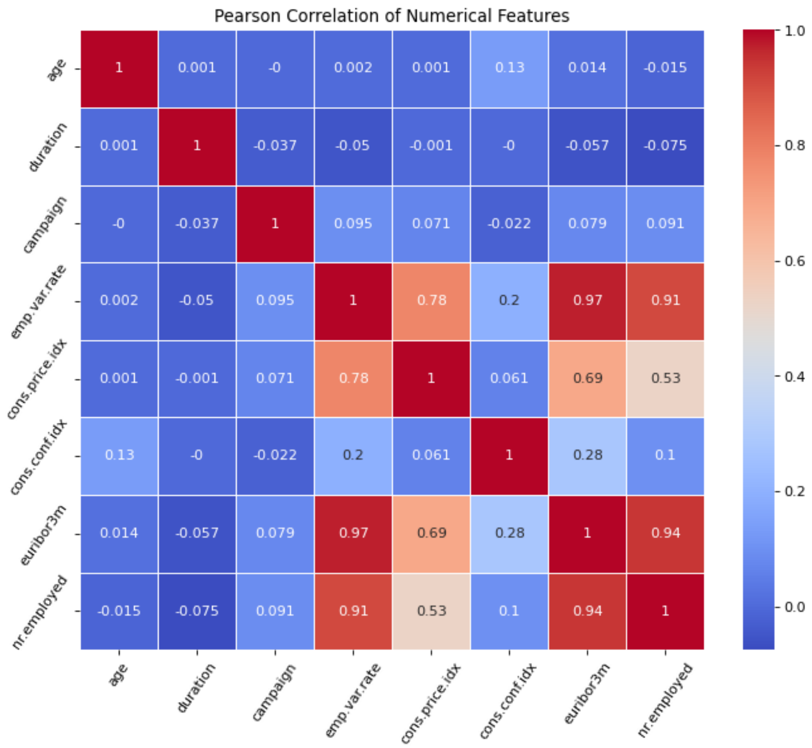
Similarly, we must now proceed to make the identical adjustments to the testing dataset as we did for the original dataset.

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

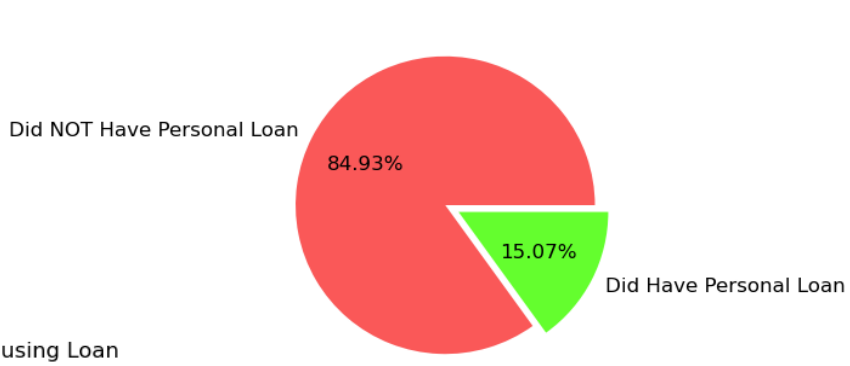
To check if there was any relationship between these predictors that we should be aware of, we will first look at the correlations for the numeric variables. According to the correlation heatmap, the more positively connected these two variables are, the redder the box is, and the more negatively correlated 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.



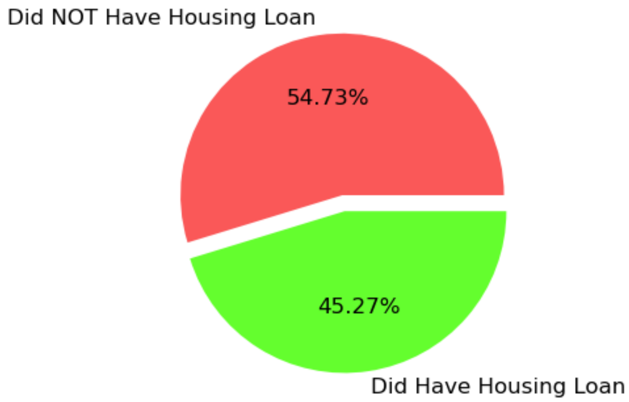
	age	duration	campaign	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
age	1	0.001	-0	0.002	0.001	0.13	0.014	-0.015
duration	0.001	1	-0.037	-0.05	-0.001	-0	-0.057	-0.075
campaign	-0	-0.037	1	0.095	0.071	-0.022	0.079	0.091
emp.var.rate	0.002	-0.05	0.095	1	0.78	0.2	0.97	0.91
cons.price.idx	0.001	-0.001	0.071	0.78	1	0.061	0.69	0.53
cons.conf.idx	0.13	-0	-0.022	0.2	0.061	1	0.28	0.1
euribor3m	0.014	-0.057	0.079	0.97	0.69	0.28	1	0.94
nr.employed	-0.015	-0.075	0.091	0.91	0.53	0.1	0.94	1

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.

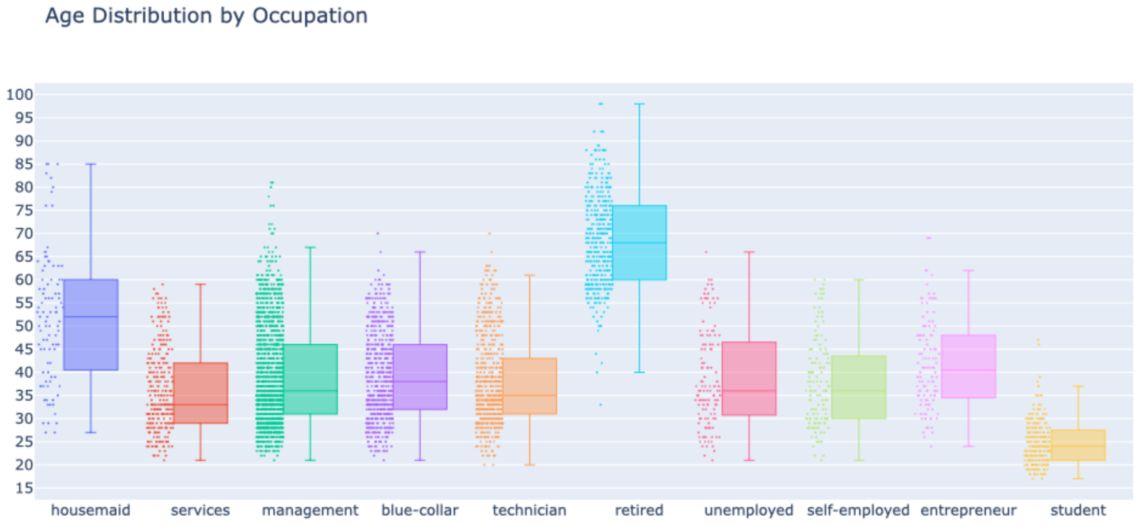
% of Condition of Loans

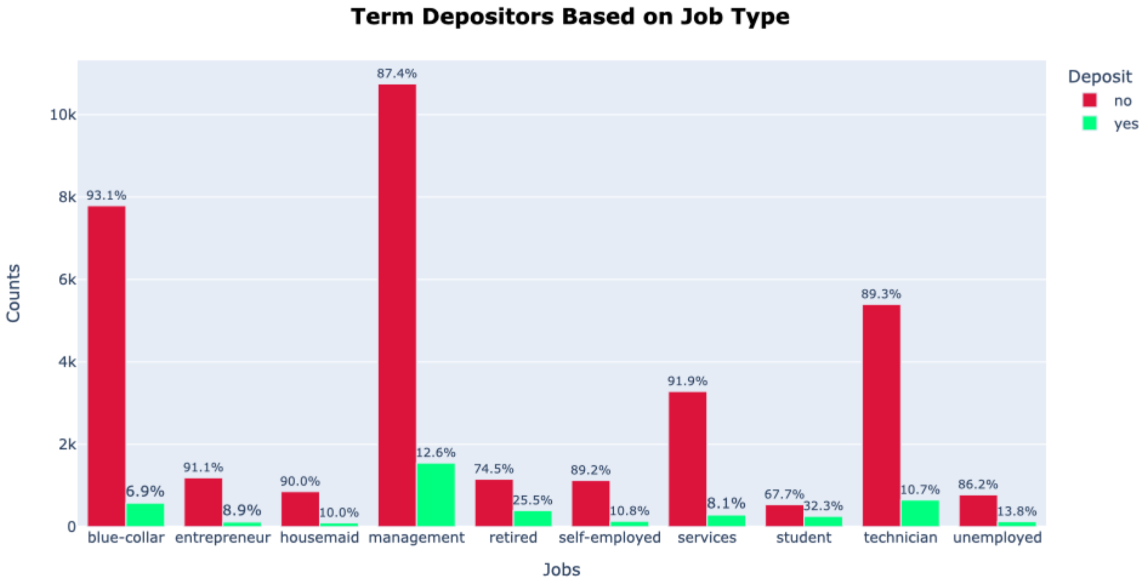


% of Condition of Housing Loan

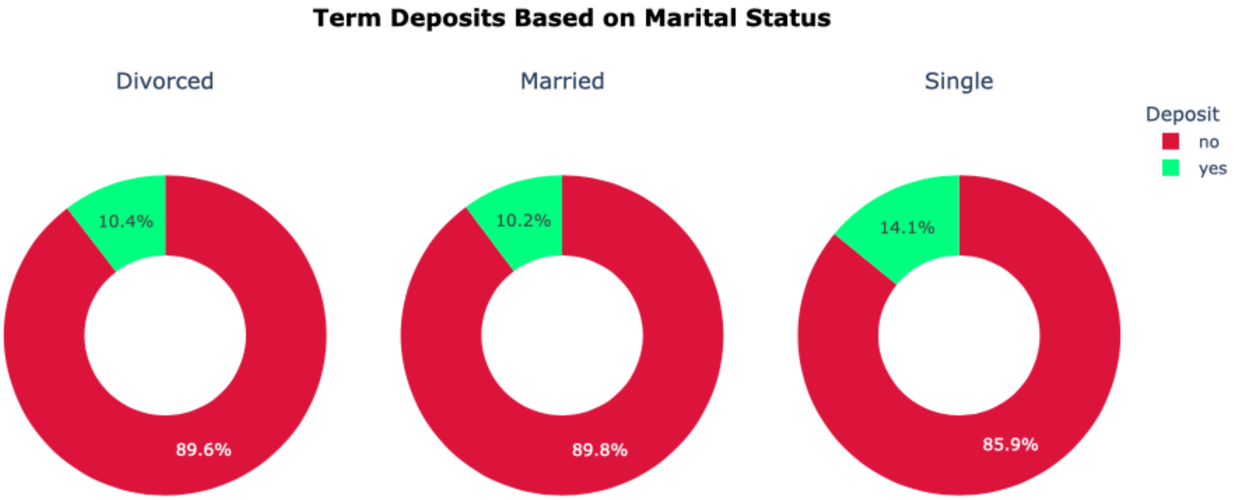


To get a better picture of our client's chances of using bank services, we looked at their age distribution based on occupation. The bank serves a varied variety of consumers from young to old; the oldest customer to use bank services is 98 years old, while the youngest user is 17. Together with providing 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.

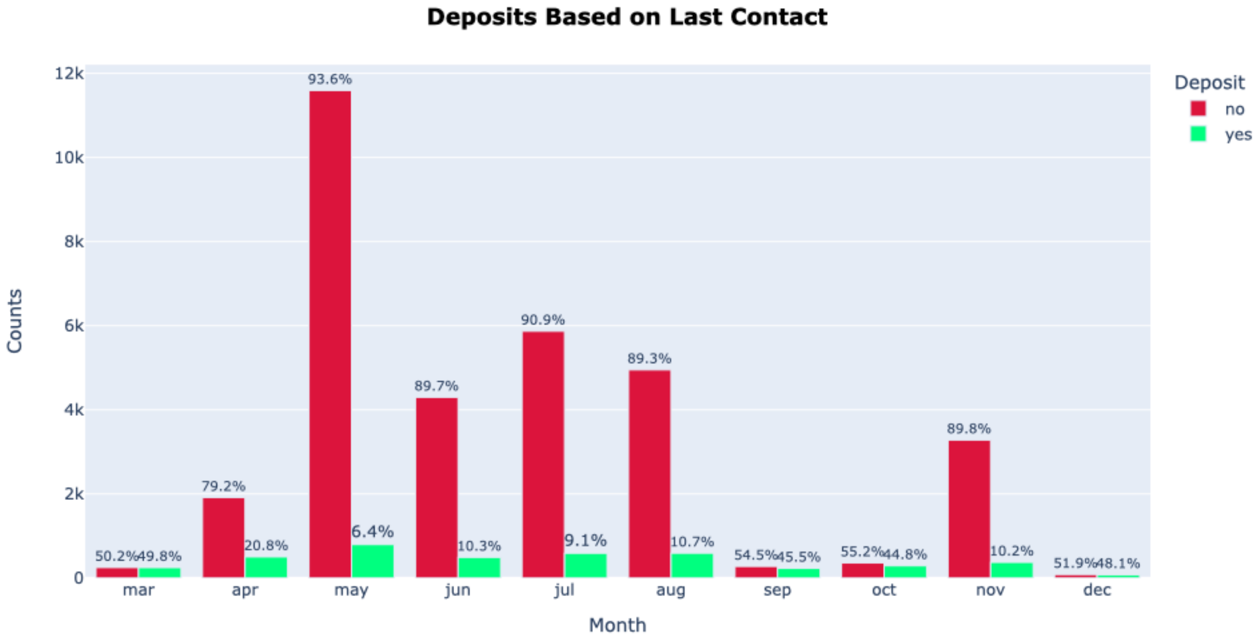




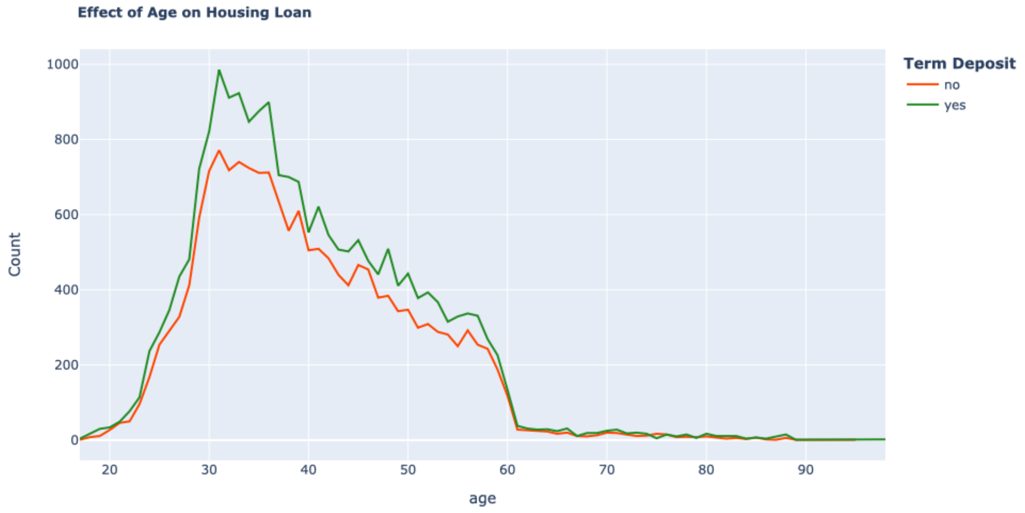
In accordance with the term deposit, we also examined the distribution of marital status. According to the distribution, most of the clientele are married, then single, then divorcees. We can observe that the three statuses, by a wide margin, did not agree with a sentence. Less than 2500 customers actually 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.

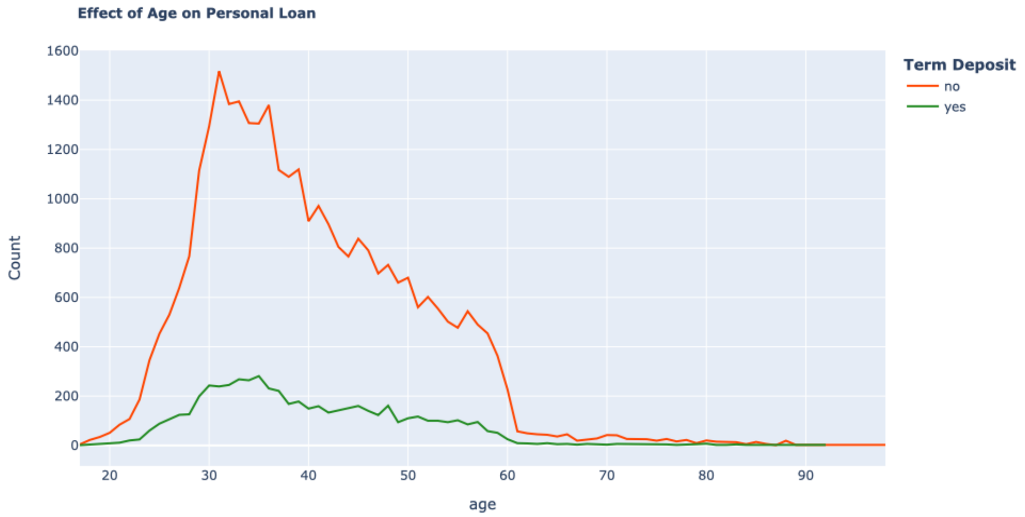


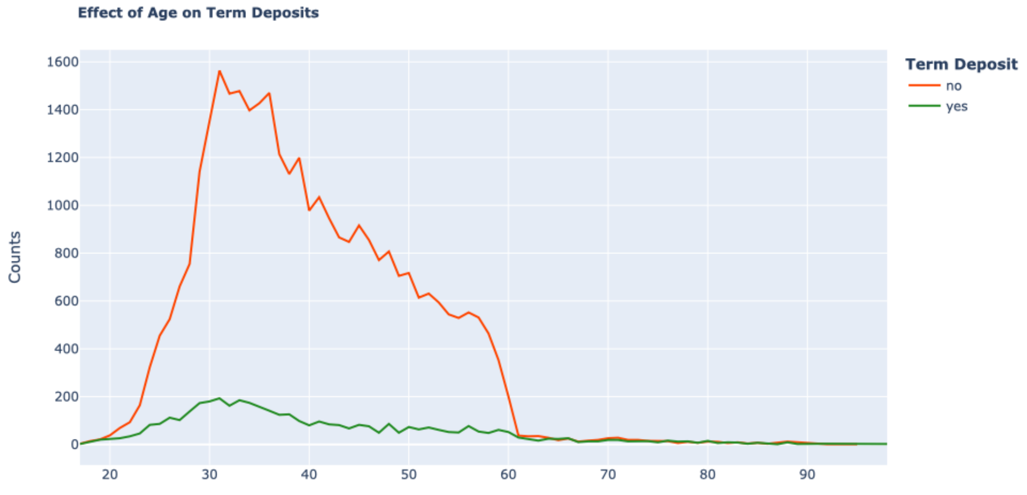
Furthermore, we investigated the date of the client's most recent monthly communication with the bank. According to the plot, the bank contacts the majority of its clients in the months of May, June, July, and August. In the months of March, September, October, and December, very few of the clients are 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.



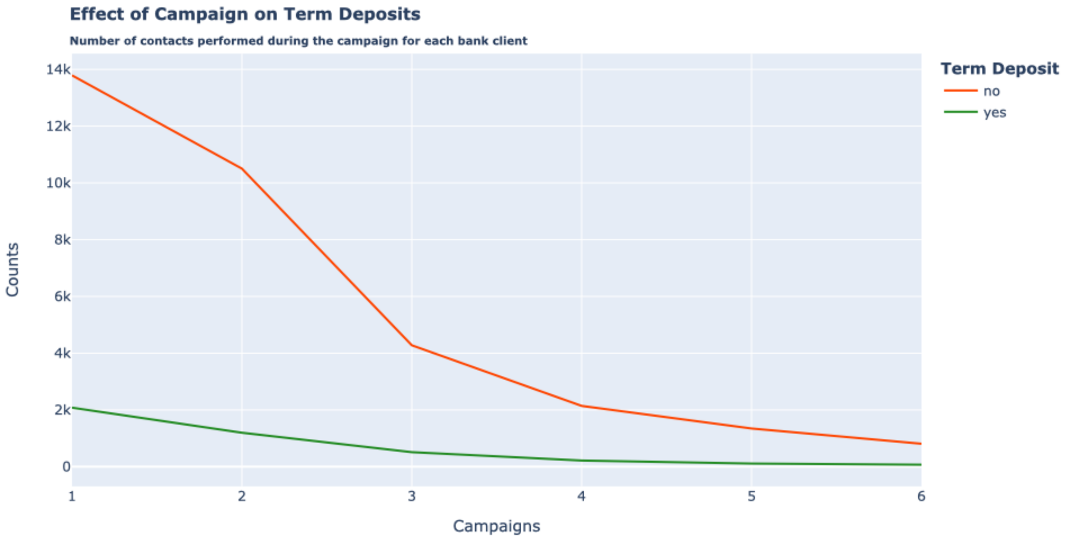
Returning to the client's age, we look at the changes based on their loaning history (personal and housing), as well as the rate of the term deposits. According to the data, none of the clients increased their personal loan subscriptions across all age groups. Along these lines, we also observe a distinct 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.



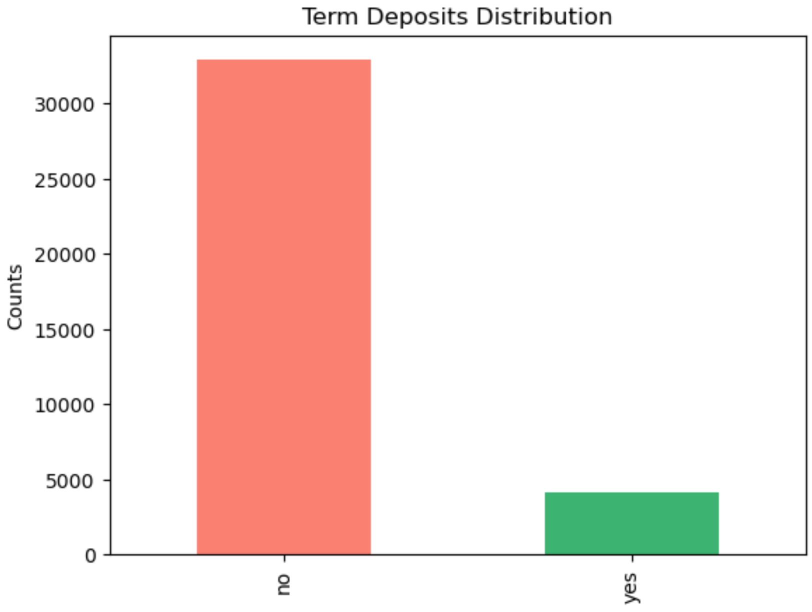




Then, using the number of contacts made during the campaign, we looked at the rate of change between those who did and did not subscribe to a term deposit. Our data shows that clients are less likely to sign up for term deposits the more frequently the banks contact them during the campaign.



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.



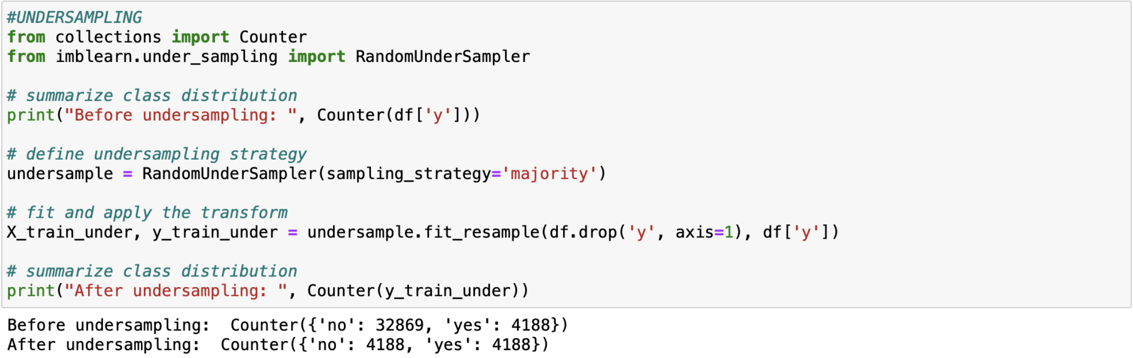
>?

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.

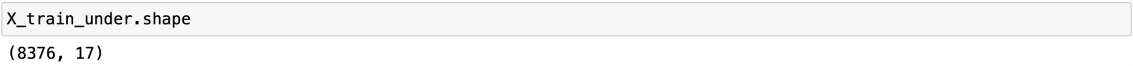
Imbalancing the Classes

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, Support Vector Machine, Random Forest, and Gradient Boosting.

Understanding of the Models

To better comprehend how to analyze the models we are developing, we must first have a basic comprehension of them before we start our study.

A statistical analysis technique called logistic regression (baseline model) is used to forecast a binary outcome. It functions similarly to a linear regression model for classification in that it predicts the target variable by examining the connection between the predictor variables. Another model under consideration is the support vector machine (SVM), which is a linear model that may be used for classification as well as regression (in our case we are using it for classification). The support vector classifier draws a straight line between the two classes, with all of the data points on one side of the line designating one category and all of the data points on the other side of the line designating a different category. Another approach is a decision tree ensemble learning technique known as the random forest classifier, which employs many decision trees. Last but not least, gradient boosting is a type of machine learning boosting that combines a number of weak learning models to create a powerful predictive model; decision trees are the most widely used model for this technique. In random forest and gradient boosting, the decision trees are constructed and aggregated in various ways. With gradient boosting, each decision tree is built one after the other, whereas for random forest, each decision tree is built individually.

Understanding the Evaluation Metrics

Five evaluation metrics we will be looking at are the precision score, recall score, f1 score, AUC-ROC score, and accuracy rate.

The precision score, which measures how well the model predicts, is calculated as the ratio of correct predictions to all predictions. The precision, for instance, is 0.90 (9/10) if the model correctly predicts ten things but wrongly predicts one. Recall score, which measures how well the model selects the right things, is calculated as the ratio of correct predictions to all of the correct items in the collection. If there are 20 items in the collection, for instance, and the model successfully identified 5 of them, then the recall is 0.25 (5/20). With these two factors together, let's say there were 20 things total, and the model correctly predicted 10 of them. Six of the ten items were correctly predicted, while four were incorrect. Following that, the recall is 0.30 (6/20), and the prediction score is 0.60 (6/10). The F1 score is another evaluation metric that computes the harmonic mean of the precision and recall scores of a model. The AUC-ROC score indicates the model's effectiveness by indicating how well it can distinguish between the target classes. Finally, the accuracy score represents the percentage of correct predictions made by our model.

The quality of the model is determined by how near the scores are to one for each metric evaluation; conversely, the closer the scores are to zero, the lower the model's capability to predict classes is.

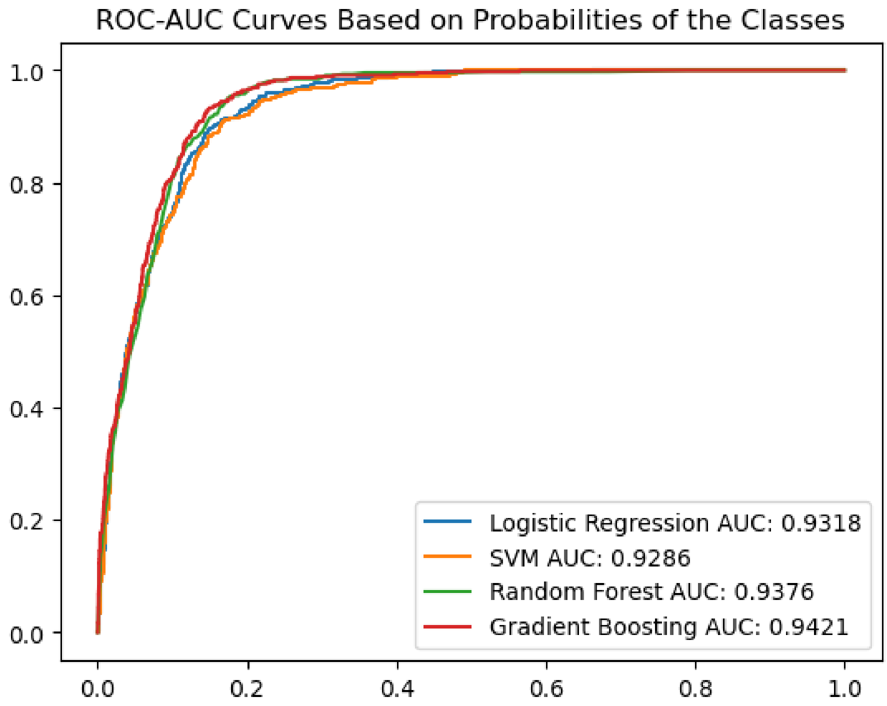
The following assessment metrics were attained for each model by training it with the training set and evaluating its performance using the testing data.

Including ‘Duration’ Column

The evaluation metric scores for the dataset with the variable "duration" are listed below:

Classifier	**Precision Score**	**Recall Score**	**F1 Score**	**AUC-ROC Score**	**Accuracy Rate**
Logistic Regression	0.3874	0.9157	0.5445	0.8689	0.8322
Support Vector Classifier	0.3660	0.9180	0.5234	0.8612	0.8169
Random Forest	0.4113	0.9357	0.5714	0.8855	0.8463
Gradient Boosting	0.4149	0.9401	0.5757	0.8886	0.8483

Additionally, we examined the ROC-AUC curves for each model:



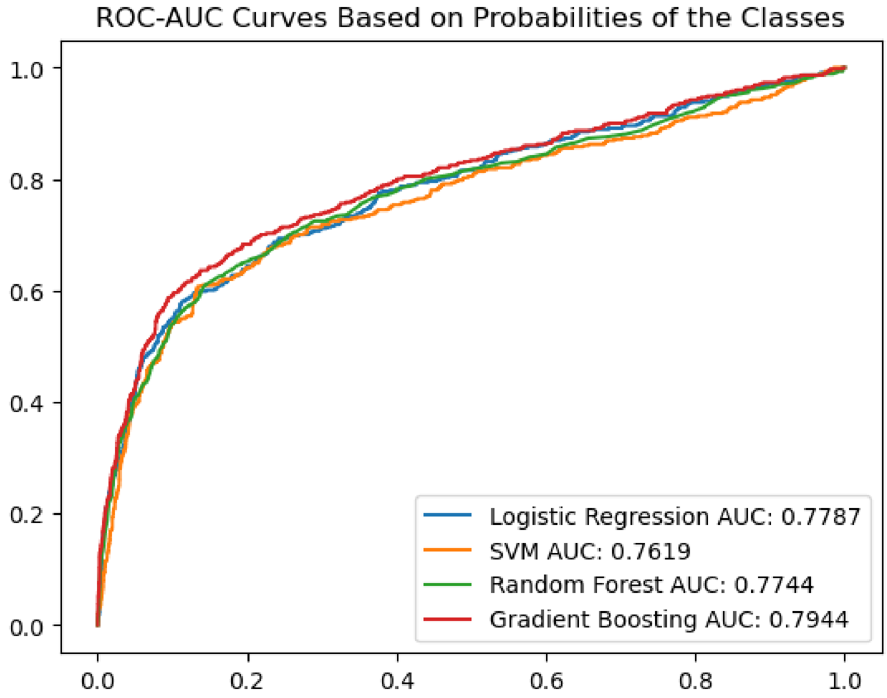
NOTE: Keep in mind that the AUC in the graph above and AUC-ROC scores differ from each other. This is because, when using the model to predict testing data, we are actually predicting the classes that would provide the specified AUC-ROC scores, as opposed to doing so for the graph, where we are predicting class probabilities.

Excluding ‘Duration’ Column

The evaluation measure scores for the absence of the variable "duration" from the dataset are shown below:

Classifier	**Precision Score**	**Recall Score**	**F1 Score**	**AUC-ROC Score**	**Accuracy Rate**
Logistic Regression	0.3011	0.6208	0.4055	0.7218	0.8007
Support Vector Classifier	0.2665	0.6718	0.3816	0.7222	0.7616
Random Forest	0.2724	0.6608	0.3858	0.7219	0.7696
Gradient Boosting	0.3500	0.6364	0.4516	0.7455	0.8308

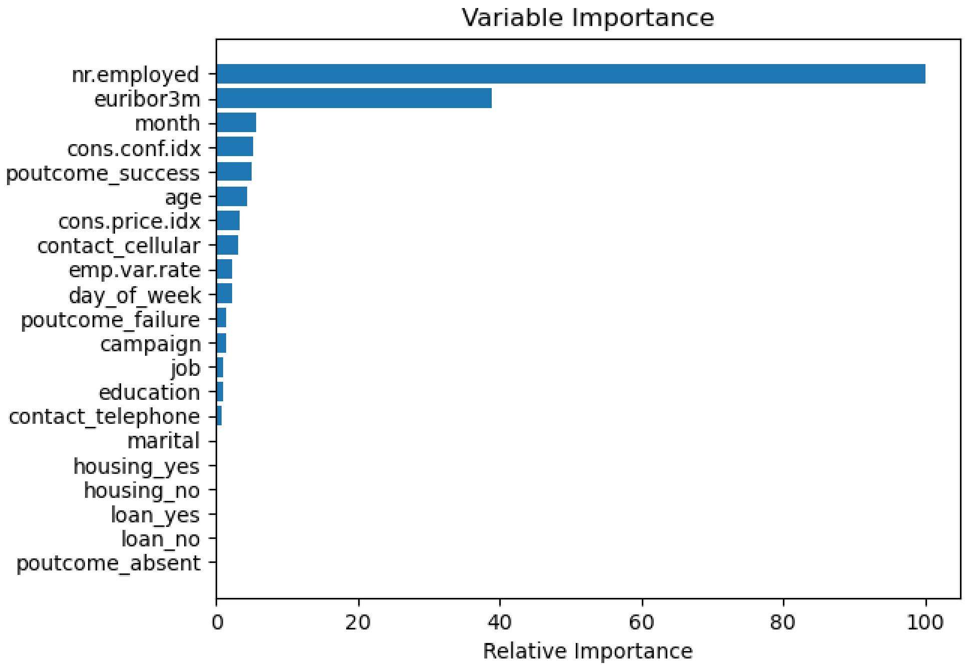
In similar fashion, we also examined the ROC-AUC curves for each model:



Final Comment: Based on the model evaluation, it is clear that gradient boosting is the most effective model because it performs best on all metrics both with and without the duration variable.

FEATURE IMPORTANCE

We will be examining the variable relevance to see which features were relevant with gradient boosting for a realistic approach utilizing our best model.



Based on the figure above, we can see that social and economic traits are important to our gradient boosting model. As we can see, while predicting the target classes, the model gives the largest relative weight to the variables nr.employed (number of employees) and the euribor3m (euribor 3 month rate - daily indicator). It's interesting to note that the model seems to give little weight to the clients' past loaning behavior.

CONCLUSION

By examining their banking history using four machine learning models—Logistic Regression, Support Vector Classifier, Random Forest, and Gradient Boosting—we want to help banks forecast customers' (and potentially prospective customers') likelihood to subscribe to term deposits. In conclusion, we suggest that gradient boosting is the best model for foretelling whether or not customers will sign up for term deposits.

Future Considerations (Technical Users)

Further investigation reveals that the social and economic characteristics have a significant impact on the customer prediction model. We should think about eliminating these traits and re-doing the simulation, though, as these traits are social and economic traits rather than individual traits. Although gradient boosting is now our best model, we may wish to use alternative machine learning models (or consider hypertuning the models) in order to provide a more thorough examination of our study.

We discovered a great deal about the bank's clients, especially those who chose not to sign up for a term deposit. Unfortunately, because the courses were not balanced, we did not really gain any useful study because the judgements we made disproportionately leaned toward individuals who did not subscribe to a term deposit rather than looking at significant characteristics of those who did. Reevaluating the decisions we took throughout the course of this project is another thing to think about. For instance, we may want to oversample rather than undersample in order to balance the classes because this would allow us to analyze our study more thoroughly and keep our options open. If given the chance, we might want to conduct a different study for a different sample of clients using the same decisions we made for this project in order to compare and contrast the outcomes.

REFERENCES

UCI Machine Learning Repository: Bank Marketing Data Set, https://archive.ics.uci.edu/ml/datasets/bank+marketing

Holtz, Yan. “Python Graph Gallery.” *The Python Graph Gallery*, https://www.python-graph-gallery.com/.

guest_blog. “10 Techniques to Solve Imbalanced Classes in Machine Learning (Updated 2023).” *Analytics Vidhya*, 3 Mar. 2023, https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/.

Bachmann, Janio Martinez. “Bank Marketing Dataset.” *Kaggle*, 12 Nov. 2017, https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset.