Analysis of Portuguese Bank Marketing: Data Analysis and Recommendations

By: Richard Broyles

INTRODUCTION

This project was about classifying whether or not a customer will subscribe to a deposit program at a bank in Portugal. The bank was using a marketing campaign to get potential customers to open a new savings account with the bank. The dataset for this project was obtained from the University of California - Irvine public data repository (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#).

This dataset features sixteen input variables and one unknown output variable, which are described below:

1) Age - numeric
2) Job - type of job (Admin, Unknown, Unemployed, Management, Housemaid, Entrepreneur, Student, Blue-collar, Self-employed, Retired, Technician, and Services); categorical
3) Marital - Categorical (Married, Divorced, Single. Note: Widowed falls under the divorced category)
4) Education - Categorical (Unknown, Primary, Secondary, Tertiary)
5) Default - Binary (Has credit in default? Yes/No)
6) Balance - Numeric (Average yearly balance in Euros)
7) Housing - Binary (Has a housing loan? Yes/No)
8) Loan - Binary (Has a personal loan? Yes/No)
9) Contact - Categorical (Contact communication type: Unknown, Cellular, Telephone)
10) Day - Numeric (Day of month of the last contact)
11) Month - Categorical (Month of last contact)
12) Duration - Numeric (Duration of last contact, in seconds)
13) Campaign - Numeric (Number of contacts performed during this campaign and for this client; also includes last contact.)
14) Pdays - Numeric (Number of days that have passed after the client was last contacted from a previous campaign; -1 means that the client was not previously contacted.)
15) Previous - Numeric (Number of contacts performed before this campaign and for this client.)
16) Poutcome - Categorical (Outcome of the previous marketing campaign: Unknown, Other, Failure, Success)

The output variable, “y”, is the variable that we are trying to predict whether or not a customer will open a new account with this bank. There are a total of 45,211 data points in this dataset. The overall goal of this project is to answer the question: What type of customer will open a new account given these attributes and what attributes will likely determine whether a customer will open an account?

Data Cleansing

Several changes were made to the data to prepare it for analysis:

1) Drop all ambiguous values, such as “others” and “unknown”.
2) Drop all outliers (which are more than 3 standard deviations away from the mean) to capture the general trend.
3) Change the ‘response’ variable (yes/no) to binary values (1/0).

Exploratory Data Analysis

The key variables in the dataset are age, balance, duration, and number of campaigns. The distribution of two of these key variables are shown below in Figure 1.

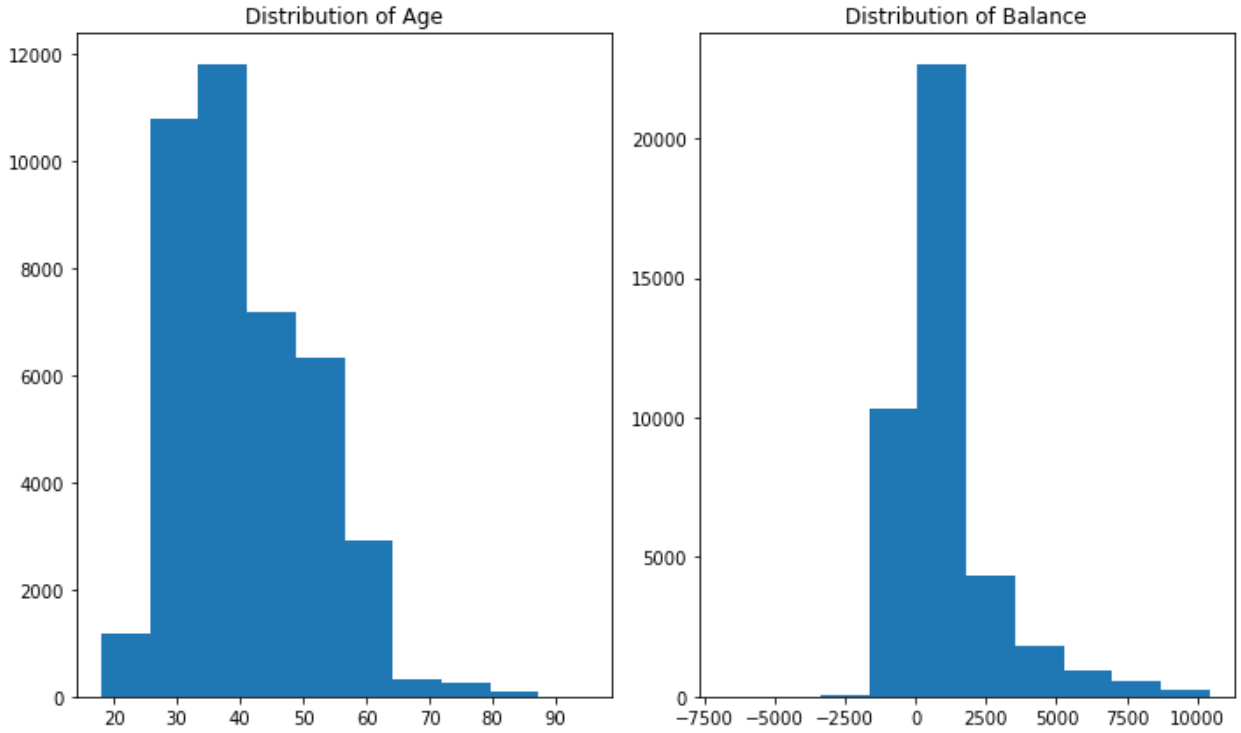

Figure 1 - Distributions of Age and Balance

There is an extensive age range, from 18 to 95 years old. The majority of the customers called was in the age range of 30s to 40s. The IQR for this column is from age 33 to age 48. The distribution of customer age is fairly normal with a very small standard deviation.

After all of the outliers from the balance column were removed, the range of the column was still large (min = -6847, max = 10443, range = 17290). The distribution of balance has a huge standard deviation relative to the mean, suggesting that there are huge variabilities in the customer’s balance.

Age and balance are two key variables, does a relationship exist between these two? The answer to this question is no, as shown in Figure 2.

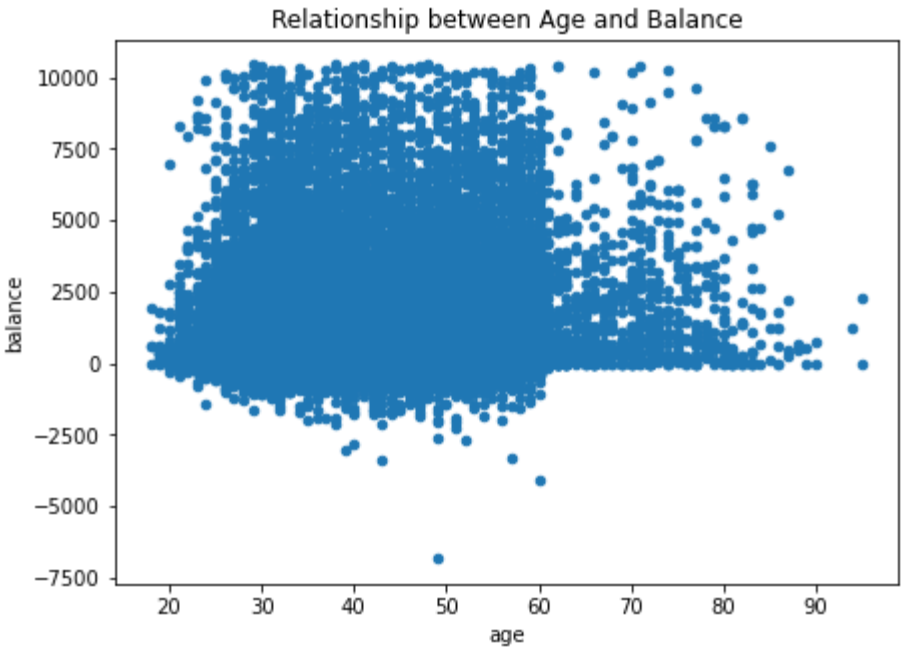

Figure 2 - Relationship between Age and Balance

The plot above shows that no relationship exists between Age and Balance. Note that customers who are above the age of sixty have low balances (< 5000 euros) because most people who are 60+ have retired from their jobs and no longer have a steady income.

Call Duration and Number of Campaigns

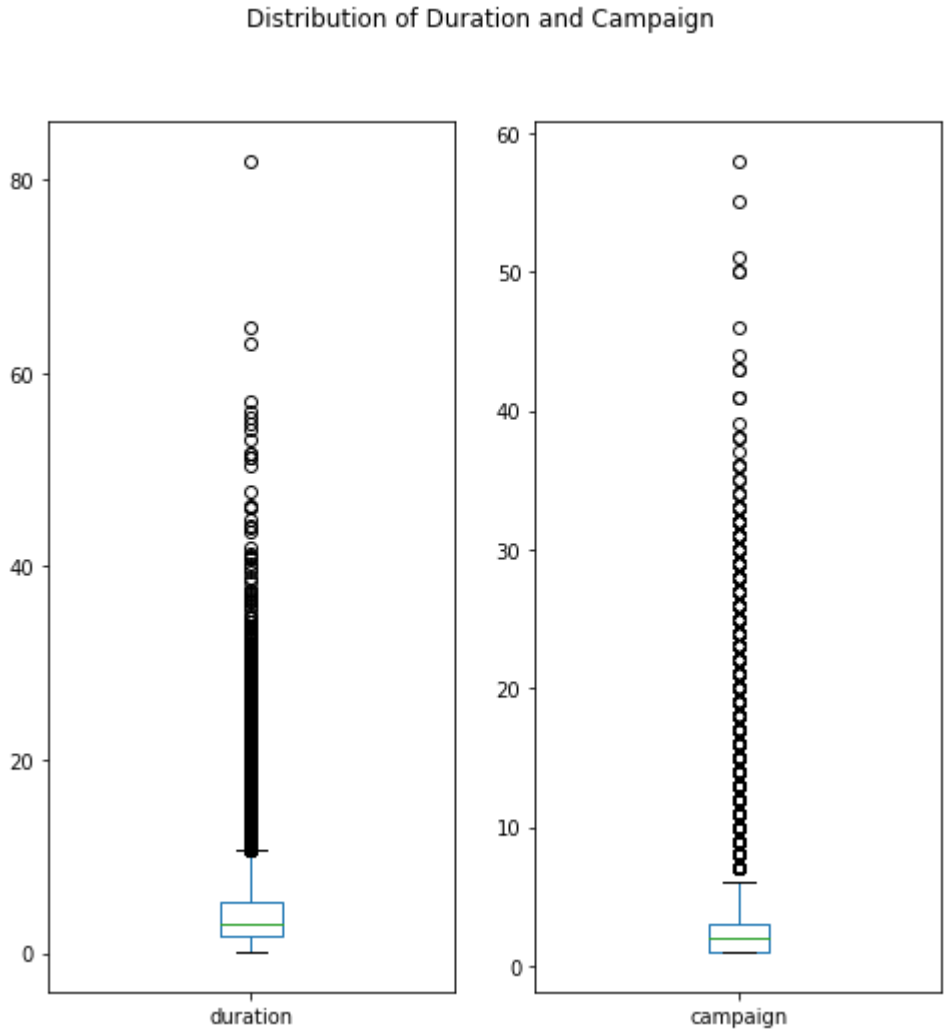

Figure 3 - Boxplots of Duration and Number of Calls

The duration of a contact has a median of about 3 minutes, with an IQR of 1.73 to 5.3 minutes. This shows that most of the calls are short; there is a large number of outliers ranging from 10 to 40 minutes, which will be investigated later.

Most of the clients have been contacted at least twice, with a majority of clients having been contacted between 1 and 3 times. The outliers in the range of 40-60 times are not normal and could potentially bother customers.

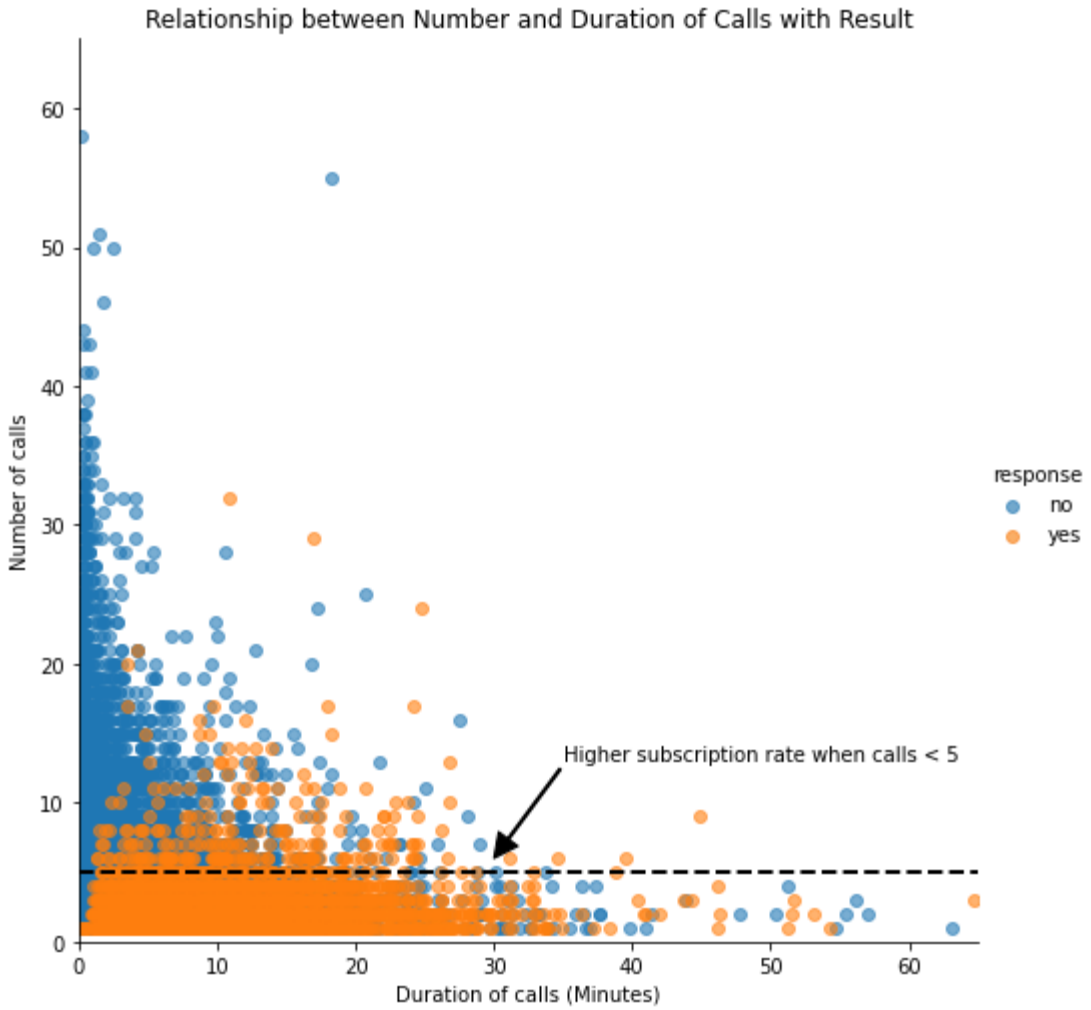

Figure 4 - Relationship between Number and Duration of Calls

Figure 4 shows the relationship between the number of calls and the duration of those calls. There are two distinct groups, those who said "yes" to the deposit and those who said "no". Compared to the "no" group, the "yes" clients were contacted fewer times and had a longer call duration. Clients were likely to reject the deposit after 5 campaign calls, unless the duration of those calls were very high. A majority of the "yes" clients were approached less than 10 times. This suggests that the bank should resist calling a client more than 5 times. As the number of calls increases, the likelihood that the customer would get angry or disturbed at the bank will also increase.

Correlation between Age, Balance, Duration, and Campaign number.

There was no clear relationship between the four key variables. See Figure 5.

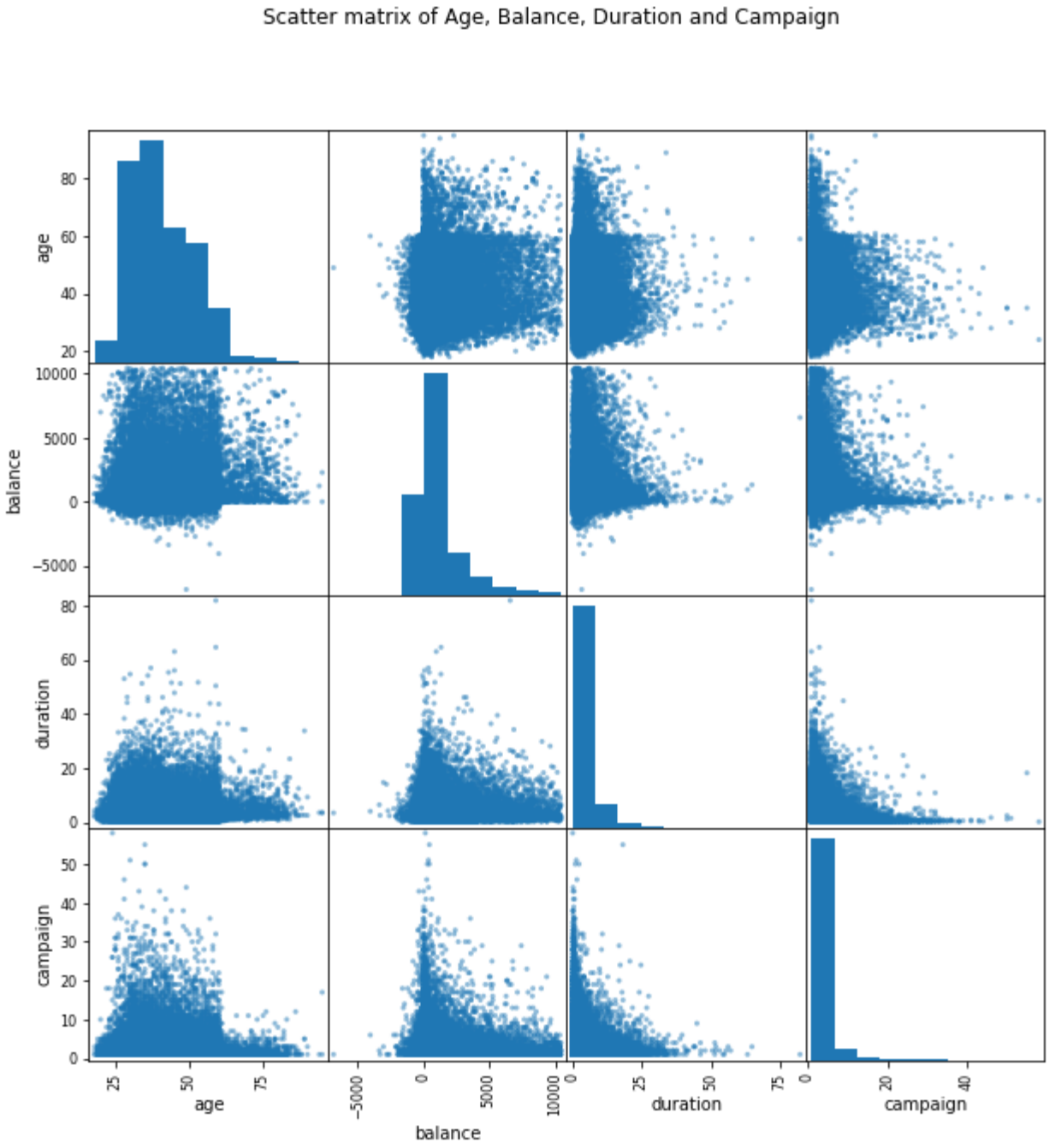

Figure 5 - Matrix of the four key variables. No relationship is discernible between these variables.

A correlation matrix was created with all of the qualitative variables. See Figure 6.

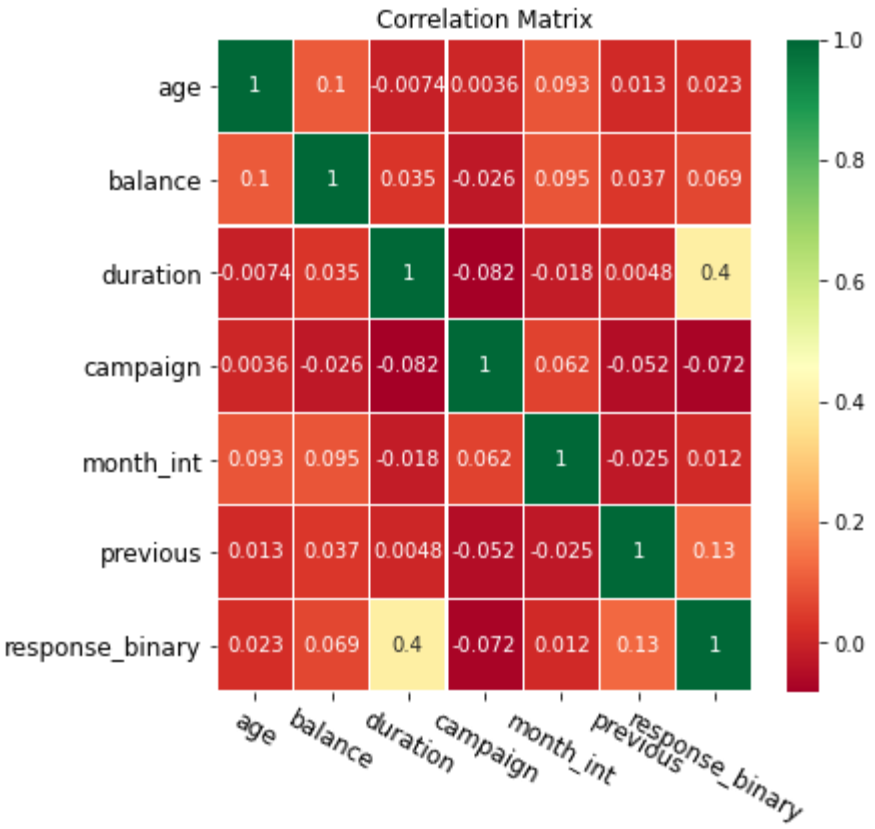

Figure 6 - Correlation Matrix of all qualitative values.

'Response_binary' has a strong correlation with duration, a moderate correlation with 'previous contacts', and a mild correlation between 'balance', 'month of contact', and 'number of calls'. Their influence on the campaign outcome will be looked at in the machine learning section. It appears there's some potential mulit-collinearity between "number of calls" and month, but other than that it appears there isn't too much correlation between the features.

Data Visualization

Now that we have a clear understanding of the distribution of the key variables, we will now investigate the other characteristics and their effect on the likelihood of a subscription.

Contact Rate and Customer Age

Figure 7 shows how a customer's age might have an impact on the likelihood of a subscription.

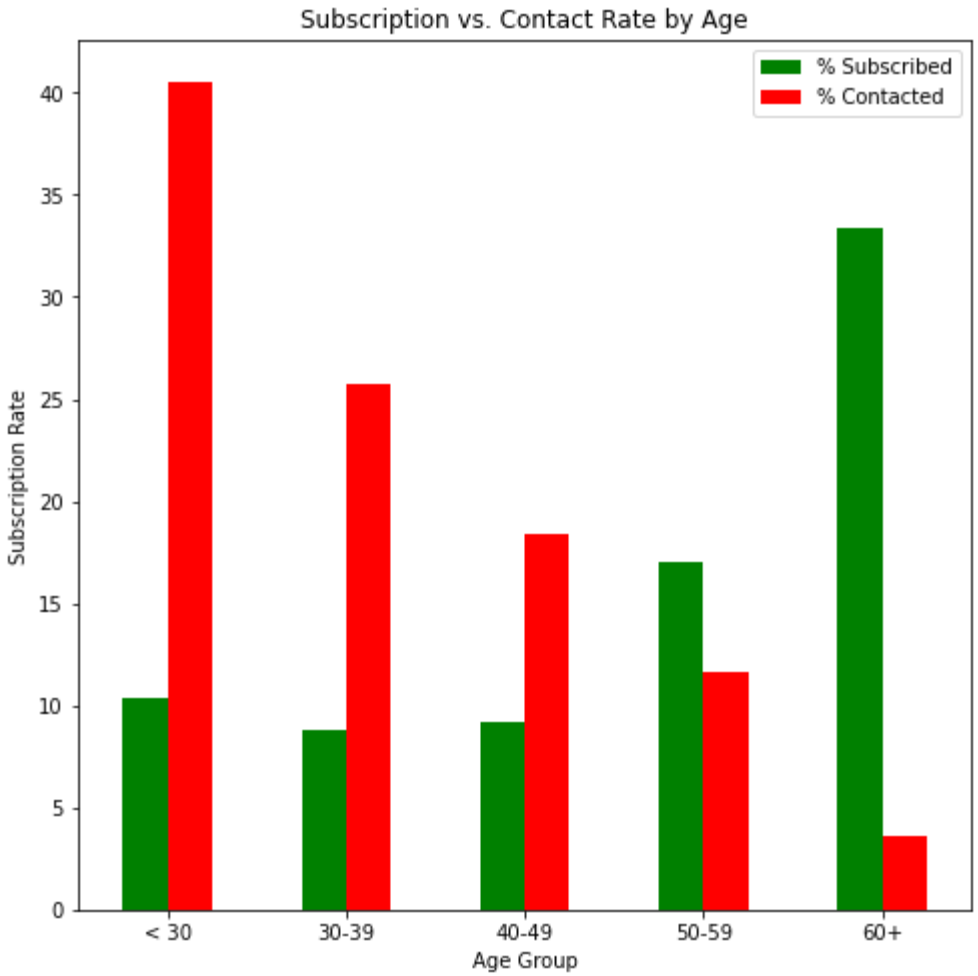

Figure 7 - Subscription vs. Contact % by Age

The chart above shows two distinct items. First, people over the age of 60 are the most likely to accept a term deposit subscription followed by people aged 50-59. Second, a vast majority of the potential clients were in the under 30 group (more than 40%). However, this approach the bank is currently using is not working. The bank should be focusing on both the group under 30 and the group over 60 for two reasons:

1) The main investment objective of older people is to save for retirement. Term deposits are the least risky investment for this age group. Middle age people are going to go for the riskier type of investments to generate more retirement income faster.
2) The youngest group are most likely students, who are likely inexperienced in investing their money and are unlikely to have a lot of money to invest. Term deposits give them liquidity and generate interest incomes that are at a higher interest rate than regular saving accounts. This makes term deposits ideal for students.

The red bars in the chart show that the bank is currently focused on marketing to the middle-aged group, which have returned lower subscription rates. To make the marketing campaign more effective, the bank should only focus on the younger and oldest clients.

Subscription Rate vs. Balance

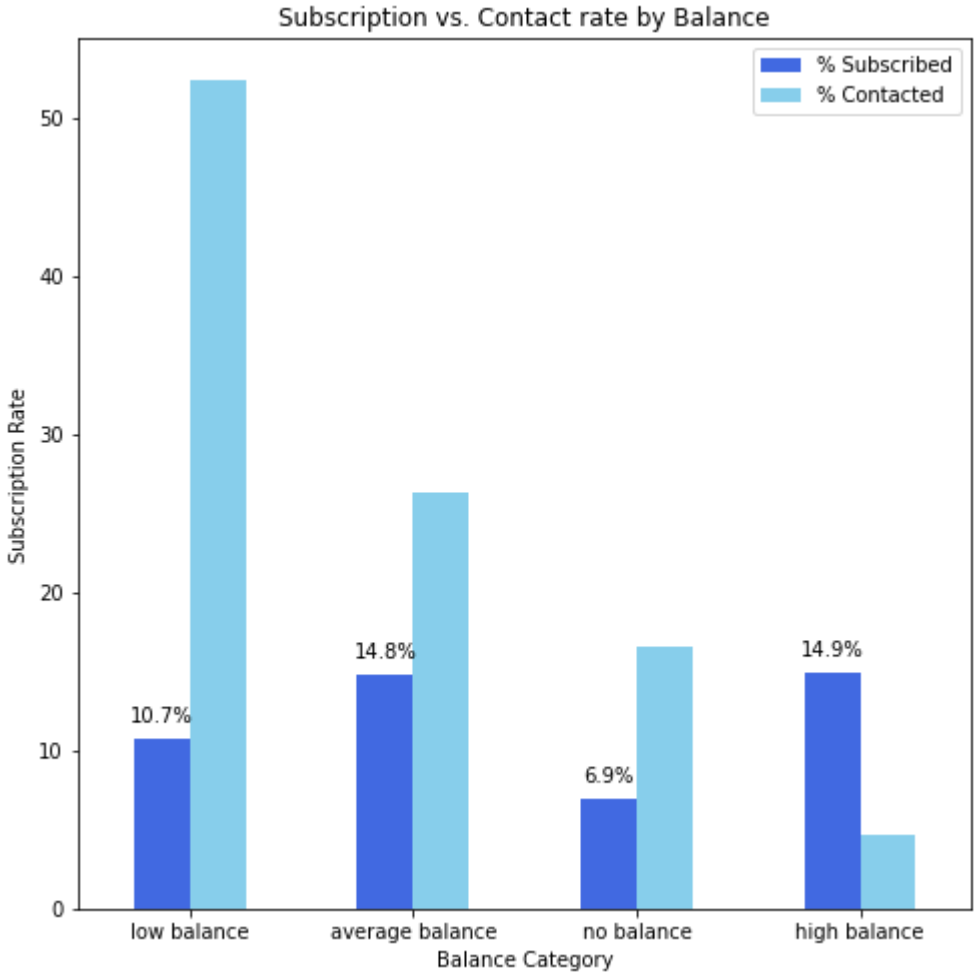

Figure 8 - Subscription Rate vs. Balance Figure 8 shows a positive correlation between the client's balance and the subscription rate. Clients who have an average balance (amount between 1000 and 5000 euros) and a high balance (> 5000 euros) have a subscription rate of 15% each. The bank however is focusing its efforts and those people who have a low balance (< 1000 euros) or no balance. The bank needs to focus its efforts on people who have an average balance or high balance. The more people that have money available to invest, the more likely they are to invest.

When you combine these two factors, you narrow down the group that the bank should focus its efforts on. See Figure 9.

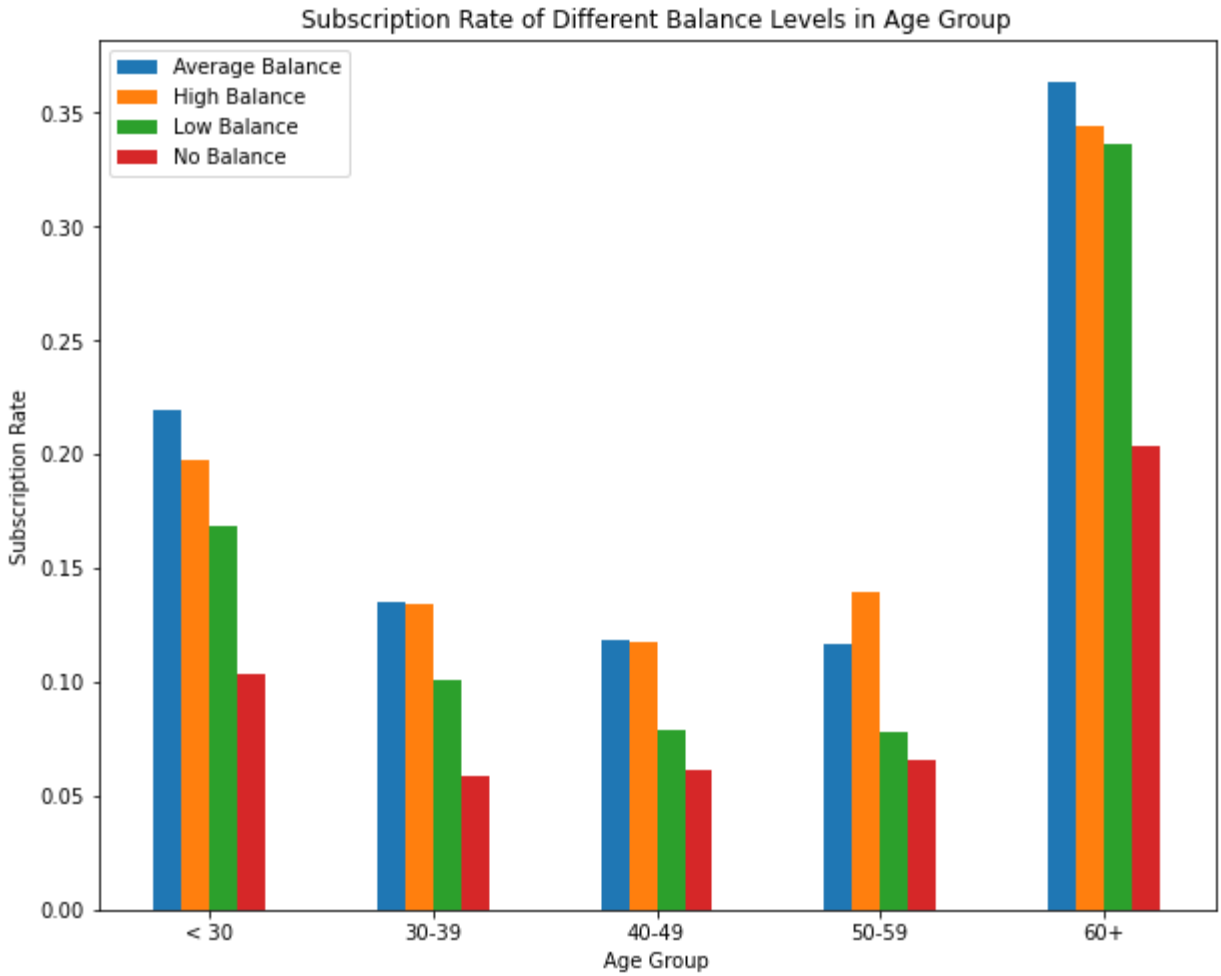

Figure 9 - Age and Balance combined together.

There is a common trend across all of the age groups when we look at age and balance jointly. People under the age of 30 and those over the age of 60 are likely to subscribe. Age represents a person's stage in life. Balance represents a person's financial strength. The amount of each group's balance has an impact on the subscription rate. The bank should prioritize its marketing to clients who are 60+ and have high balances and people who are under 30 that have a positive balance.

Figure 10 supports this argument even further. The highest job classifications that have the most subscriptions are students and people who are retired. Those two job types combined represent almost 50% of the total subscriptions.

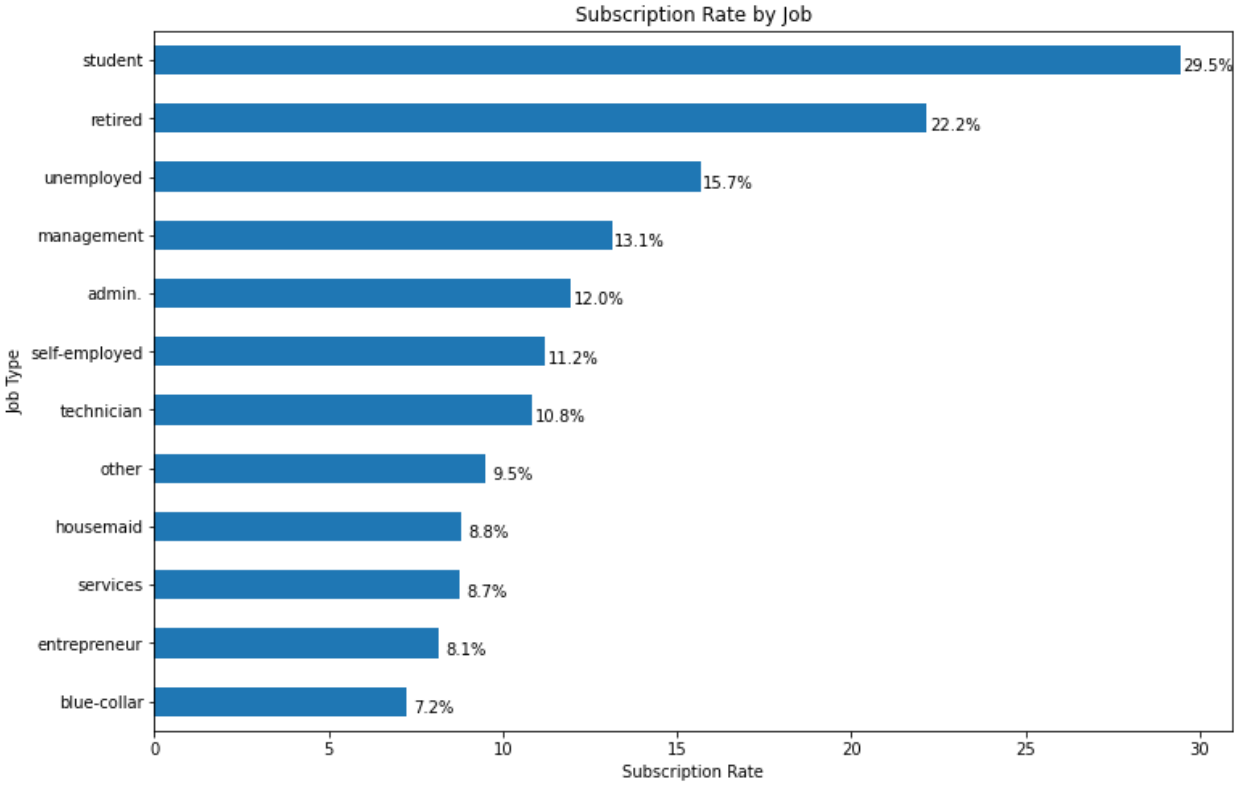

Figure 10 - Subscription Rate by Job Type

Subscriptions by Month

There are some external factors that will impact the subscription rate. One of those is the time of year of when a campaign contacts a potential client. See Figure 11.

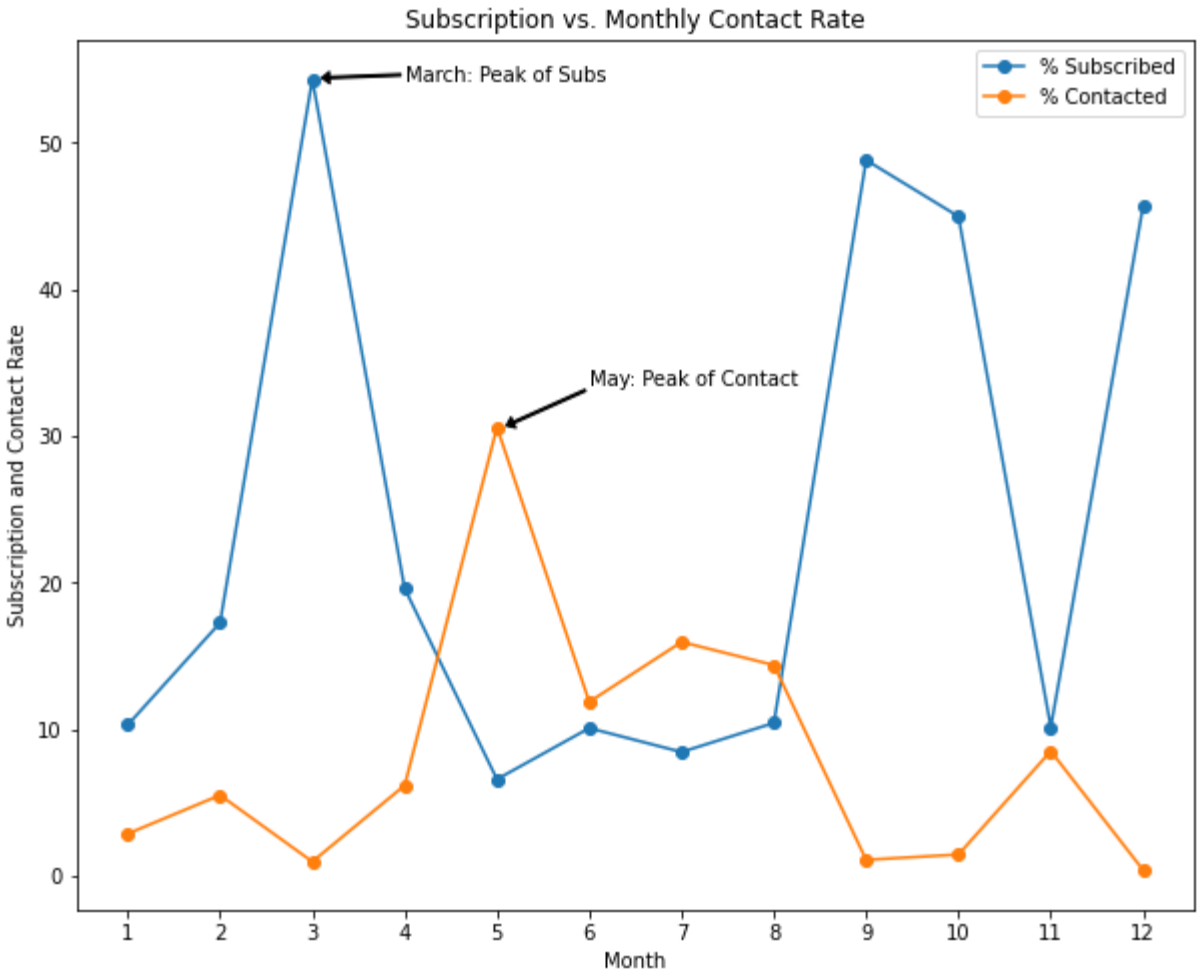

Figure 11 - Time of Year when Customer Was Contacted

From the chart, the bank contacted the most customers between the months of May and August, with May having the highest contact percentage at 30%. However, the subscription rate was highest during the months of March, September, October, and December, despite these months having little to no contacts. March had the highest percentage of new subscribers with over 50% of contacts subscribing.

The problem with this is that the timing of the bank's campaign is misaligned. These two values should match if the bank wants to have an effective marketing campaign. When looking at other external factors that can impact a campaign, there is not enough data to make a valid conclusion so the bank must be careful when weighing other factors into the timing of their campaigns.

Modeling

The main objective of this project is to find the most responsive customers before the marketing campaign so the bank will be able to effectively reach out to them which will save time and resources. To align with this objective, we will be using classification algorithms to

analyze the customer's statistics. A classification model was built to classify the clients into two groups: those who said "yes" to signing up and those who said "no". The data was prepared in the following ways:

1) Select the most relevant customer information columns. These columns are job title, education, age, balance, default record, housing loan records, and personal loan records.
2) All five categorical variables (job, education, default, housing, and loan) were transformed into dummy variables. Dummy variables give us a way of representing categorical variables numerically.
3) Feature selection: all of the customer's statistics were selected as features while the outcome was selected as the target. The dataset was split into 80% for training the model and 20% for testing the model.

Seven different classification algorithms (Logistic Regression, K-Neighbors Classifier, Random Forest Classifier, Gaussian Naive Bayes, Gradient Boosting, Linear Support Vector Machine, and Neural Network) were run on the dataset and the model that scored the highest was used on the test set. See Figure 12.

Classifier	Cross Validation Mean Score
Gradient Boosting	0.906744
Logistic Regression	0.904261
Random Forest	0.903386
SVC	0.903036
Neural Network	0.902896
K-Nearest Neighbors	0.896250
Naïve Bayes	0.858647

Figure 12 – Table Containing the Highest Model Scores

The model that had the best performance was the Gradient Boosting model, with an accuracy score of 90.6%, so this model will be used to predict the customers' responses given all of the defined customer features. To verify the performance of this model, a confusion matrix was created. See Figure 13.

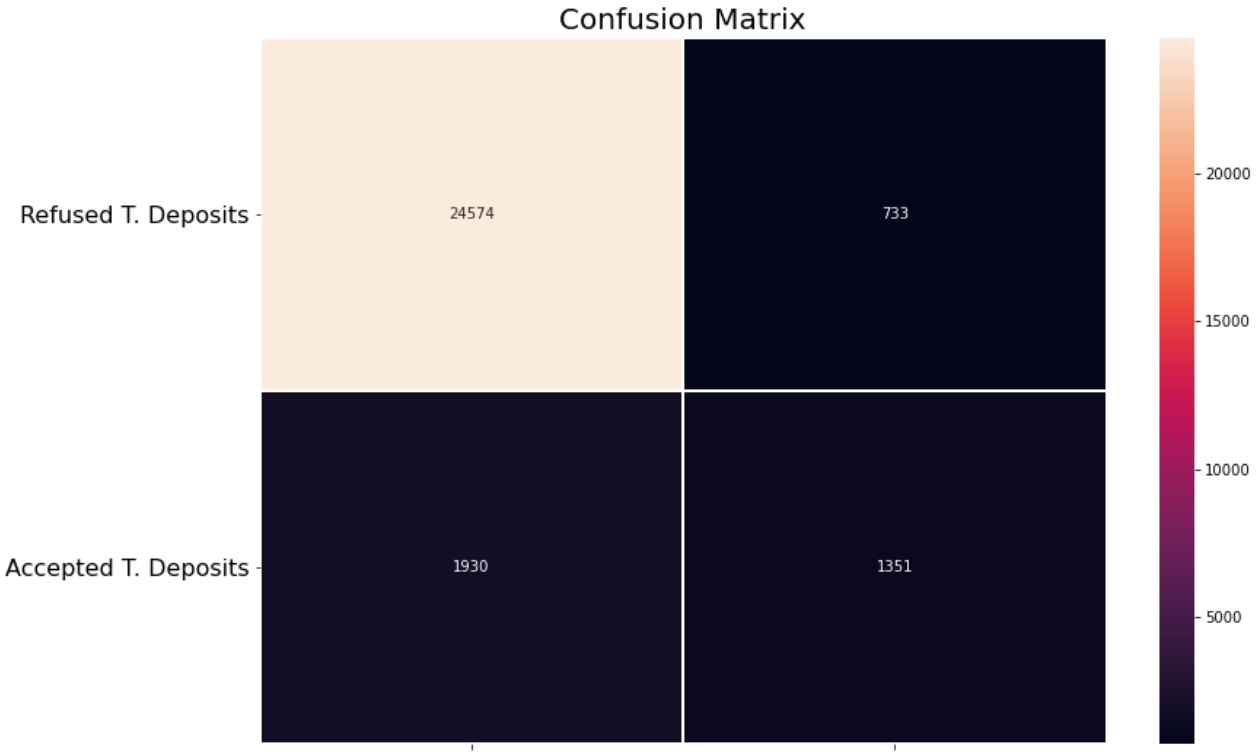

Figure 13 - Confusion Matrix Evaluating the Model

There are two problems that this confusion matrix shows.

1) The accuracy score is misleading because the number of observations in the different classes vary largely. A confusion matrix gives a breakdown of the prediction result and the error types. In the test set, the matrix proves that the algorithm performed well with 7278 True Positive predictions, which are correct. There were 1351 False Negative predictions, which means that the model predicted that 1351 people would subscribe to the term deposit, but in reality they did not.
2) This dataset is extremely unbalanced, with nearly all of the potential clients declining to subscribe. This means that the accuracy score is biased and further evaluation is required to determine the true accuracy score.

Precision Vs. Recall

Precision is defined as the ratio between the True Positives and all the Positives. This is a measure of how many people would sign up for the deposit out of all of the people who were contacted. It is calculated by the following formula:

$$\text{Precision} = TP / (TP + FP)$$

Recall is defined as the measure of our model correctly identifying the true positives. In our case, the recall tells us how many people the model correctly identified making a deposit.

This is calculated by the formula:

$$\text{Recall} = TP / (TP + FN)$$

The precision score for this model is 0.6482 and the recall score for the model is 0.4117 with an F1 score of 0.5036. Based on these scores, we are unsure if the model is correctly identifying the true positives or the false positives. When we analyze the tradeoff between the two, we are given the plot in Figure 14.

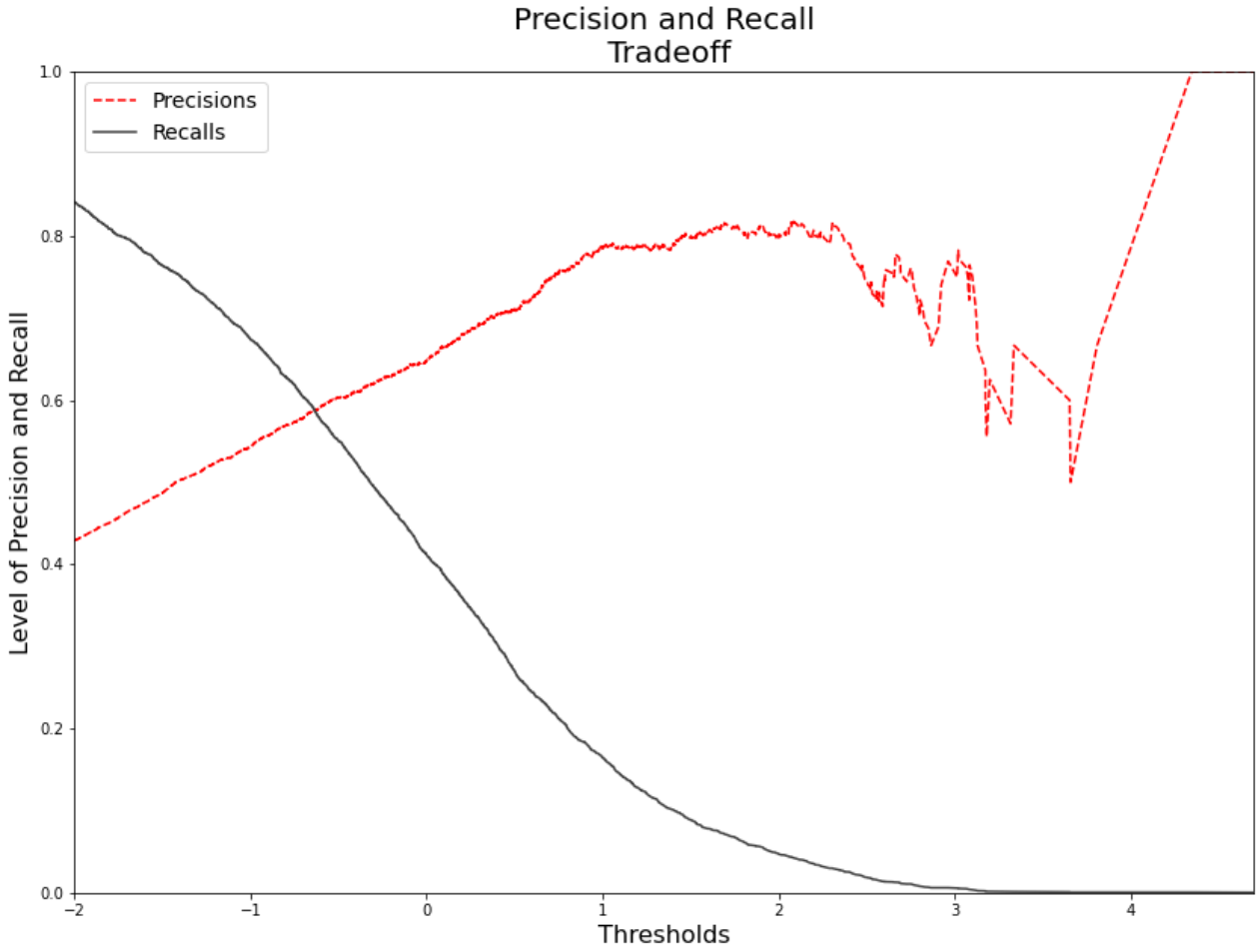

Figure 14 – Analysis of the Precision-Recall Tradeoff

A test used to gauge the effectiveness of the model as well as the tuning of the hyperparameters was the Youden's J statistic, which measures the performance of a dichotomous diagnostic test. This statistic tests whether an informed decision was made about setting up an account. The statistic is defined as:

$$J = \text{sensitivity} + \text{specificity} - 1$$

This value of this model ranges from 0 to 1, with 0 meaning that the test has the same proportion of positive and negative results, meaning that the test is useless. A value of 1 means that the test had no false positives or negatives. This index gives equal weight to the false negatives and false

positive values. While it is possible to have a value of less than 0, this means that positive and negative labeling was switched, and the test needs to be corrected. Running this statistic on our test data yielded a value of 0.09, which means that this test has some value, but not a lot.

We next looked at the ROC curves for the top three models (Gradient Boosting, Logistic Regression, and Random Forest). After doing some cross-validation on the models, we analyzed the ROC scores for those models as shown in Figure 15.

Classifier	ROC Score
Gradient Boost	0.9208
Logistic Regression	0.9099
Random Forest	0.8994

Figure 15 – ROC Scores for Top 3 Models

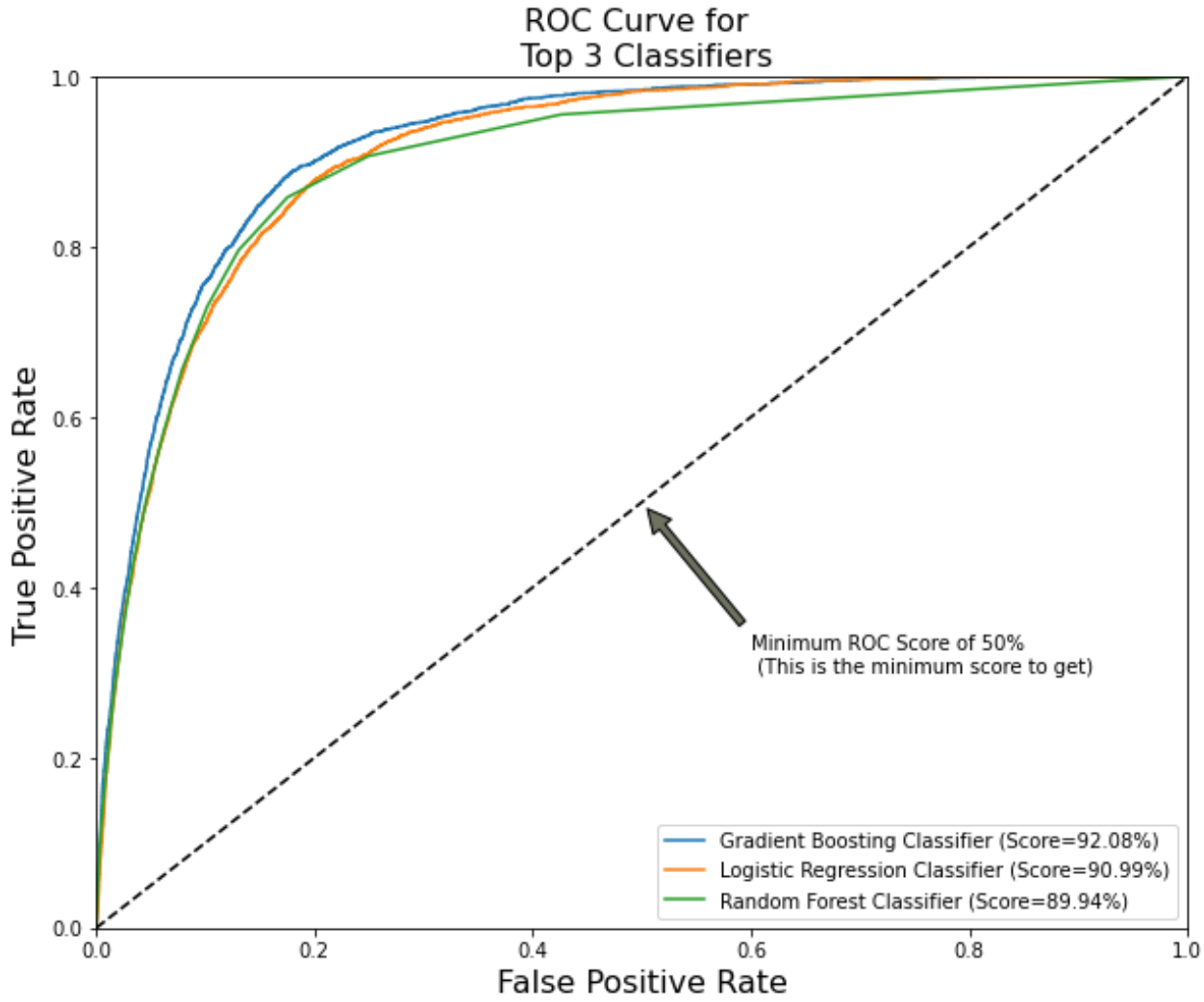

Figure 16 – ROC Curves for Top 3 Classifiers

Conclusions The overall goal of this project was to increase the effectiveness of the bank's telemarketing campaign. This was achieved using the Data Science process from analyzing the data, visualizing the data, and building a model.

The target customer profile was established with the following features:

- Feature 1: Age (Age < 30 or Age > 60)
- Feature 2: The job type (Student and Retired people)
- Feature 3: A balance in the bank of more than 5000 euros.

By applying a classification and regression model, the bank will be able to predict a customer's response to a telemarketing campaign before calling the customer. The bank can direct its resources towards customers who are most likely to open a term deposit account and move away from customers who are not likely to open an account.

In addition, predicting the duration of a call and adjusting the marketing plan benefits the bank and their clients. It will increase the efficiency of the bank's telemarketing campaign and it will save the bank money in advertising costs by allowing the bank to pick and choose where to send their advertisements, which will make customers happy because they will no longer receive advertisements they don't want.

Recommendations

1. Improve the timing of the campaign

When starting a marketing campaign, there are external factors, such as the time of year when starting the campaign, that must be considered. In the month's analysis, it was shown that March, September, October, and December had the highest success rates. More data should be collected and analyzed to see if the seasonal effect is constant over time. If this trend continues, the bank should consider starting the campaign in fall or spring.

2. Smarter marketing design

The bank needs to start targeting the right customers who have the right attributes to generate more subscriptions. The classification algorithms used in the modeling process will fix the imbalance of the dataset, which will give the bank more accurate information in improving their subscriptions. The current telemarketing campaign needs to be re-evaluated to appeal to more customers.

3. Better services provision

This project gives the bank a more granular understanding of their customer base. This will give the bank the ability to provide better banking services to a variety of customers. This will give the bank more flexibility in attracting new clients or managing risk.

Future Works

The following is purely speculative in nature as there is no data in the original dataset to support this idea; creating a model based on the ideas in this section will be highly beneficial for the bank to improve the campaign and make it more profitable.

The data does not tell us the amount of money the customers who did invest money into the term deposit. However, a model can be created based on the balance of the customers who did invest. A profitability analysis can be done for the bank; however, the following questions must be accurately answered for the analysis to have any merit:

1) How much money was invested per customer in the deposit?
2) What interest rate was offered to the customer and how is this interest collected (simple vs. compound)?
3) What is the term (length of time) of the deposit (1 year, 18 months, 2 years, etc.)?
4) What is the interest rate being offered to other banks who borrow these deposits? What happens to the deposits if the bank default on the loan?
5) What is the minimum amount of money required for the initial investment? What is the minimum balance the customer must have to invest?
6) What is total cost of each campaign the bank absorbs?