Introduction

A common practice for the investment banking division (IBD) of a bank is to identify clients who are willing to invest their funds with the financial institution. Marketing managers of these divisions are now carefully tuning their directed campaigns in an effort to predict which clients are most likely to deposit funds. This enables the managers to prioritize contacting those clients over others based on available data.

For this project, we will consider the data set *Bank Marketing* from UCI Machine Learning Repository. This data is related with direct marketing campaigns of a Portuguese banking institution based on client phone calls. Direct marketing campaigns are a type of advertising campaign that seeks to achieve a specific action in a selected group of customers.

In this case, the action is whether the client will subscribe to the bank's term deposits. A term deposit is a fixed-term investment that includes the deposit of money into an account. Term deposit investments usually carry short-term maturities and will have varying levels of required minimum deposits. Examples of term deposits include certificates of deposit (CDs) and time deposits.

The original data set contained 22 variables with a total of 41,188 observations. There were 21 features which include both quantitative and qualitative data types. The target is a binary classifier which indicates whether the client subscribed a term deposit. In other words, the target refers to whether or not the client will commit to invest in a term deposit with the bank.

Data Description

An initial challenge with this data set is that some clients required more than one contact from the financial institution to assess whether the bank term deposit would result in a subscription or not. To circumvent this complication, we can focus this project's analysis on whether a client will subscribe a term deposit based solely on the initial contact and disregard the records pertaining to subsequent calls. After the removal of this data, the resulting data set contained 17,642 observations, which is approximately 43% the size of the original data set.

There are a total of 11 quantitative variables and 11 qualitative variables in the data set, including the target, shown in the table below.

Attribute Information		
Client Data	**Social and Economic**	**Other**
Age	Employment Variate Rate	Campaign
Job	Consumer Price Index	Pdays
Admin	Consumer Confidence Index	Previous
Marital Status	Euribor 3 Month Rate	Poutcome
Education	Number of Employees	
Default		
Housing		
Loan		
Contact Type		
Month of Contact		
Day of Week of Contact		
Duration of Contact		
Subscription		

We can divide the variables into three categories. The first category is comprised of the data pertaining to the banking client. All of these attributes are qualitative other than age and duration of contact which are quantitative. After the bank has made contact with the client it is known whether the person subscribed for a term deposit or not, so the *subscription* variable is the binary target. Note that the duration of contact (in seconds) highly affects the target, and is therefore not included in the predictive modeling. This duration can only be known after the call has been performed, and is really only useful for benchmark purposes.

The second category consists of social and economic variables. The employment variation rate is a quarterly indicator, the consumer price index and consumer confidence index are monthly indicators, the euribor 3 month rate is a daily indicator, and the number of employees at the bank is a quarterly indicator.

The third category contains all other variables. The *campaign* variable corresponds to the number of contacts performed during the campaign, which for this analysis is always equal to one. The *pdays* variable is the number of days that have passed by after the client was last contacted from a previous

campaign. The *previous* variable corresponds to the number of contacts performed before this campaign for a given client. Lastly, the *poutcome* variable corresponds to the outcome of the previous marketing campaigns. From this third category, only *poutcome* was used in the modeling. The other variables were either not necessary to include or did not improve model performance.

Data Preparation

To prepare the data for analysis, we can convert the qualitative variables into factor data types which, by default, R implements 1-hot encoding. The quantitative variables are stored as numeric data types. Then we can check for any missing values in the data set, but find there are no missing values. Therefore, imputation is not required.

Exploratory Analysis

Initially, we can check how the response is distributed to ensure we have an adequate amount of observations in each class. The frequency distribution is shown below. Note the class imbalance in the target's frequency distribution. There are more than 7 times the number of observations in the "No" class than there are in the "Yes" class. This will be shown to impact the accuracy of the models, but there are enough observations in each class to still perform a useful analysis.

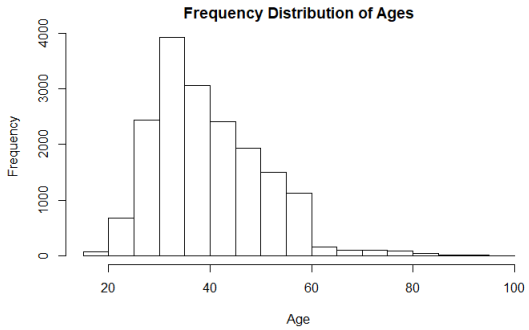
Frequency Distribution Table of Target	
No	**Yes**
15,342	2,300

A frequency distribution which suggests a pattern in the data is the month of the year that the client was contacted against the target. The frequency distribution is shown on the following page. We can note that there are no calls in January and February. At this time, it is not clear why this is the case. We can also see that there is increased call volume during the summer months. The month of the year that the client is contacted is later shown to be a significant predictor of the target, which is not a surprising result based on the frequencies. In comparison, when we examine the frequency distribution that corresponds to the day of the week the client was contacted

against the target, we find an approximately evenly distribution which does not seem to be very significant for modeling.

Frequency Distribution Table of Months by Target		
Month	**No**	**Yes**
January	0	0
February	0	0
March	111	142
April	1,017	308
May	5,351	415
June	1,792	251
July	2,427	247
August	2,302	294
September	181	147
October	245	210
November	1,872	247
December	44	39

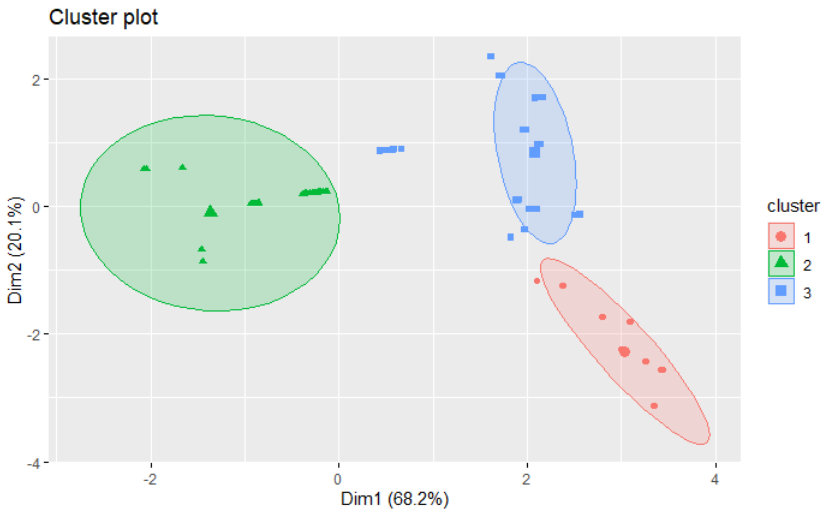
We can also plot the frequency distribution of the client ages using a histogram. The histogram is shown below with the corresponding five-number summary statistics and the mean. We can note that the distribution is right skewed, but that the average client age appears to lie around the median age of 38.



Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
17.00	32.00	38.00	40.08	47.00	98.00

Dimension Reduction

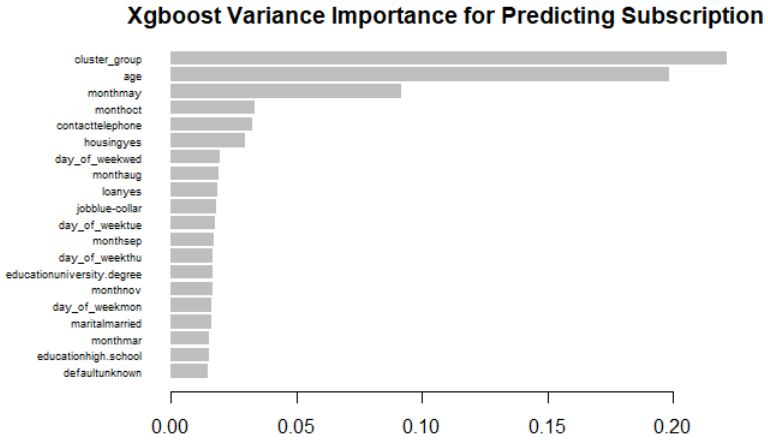
In an effort to reduce the dimensions of the feature set and improve model performance, we can use kmeans clustering on the scaled quantitative variables listed under the social and economic attributes category. Comparing the silhouette and elbow methods to determine the optimal number of clusters, we found that the optimal number of clusters was 3 using the kmeans clustering algorithm. This new qualitative variable can then be thought of as an indicator of the market condition at the time the customer is contacted. This allows us to convert 5 quantitative variables into one categorical variable with 3 levels. The clusters can be visualized in the plot below.



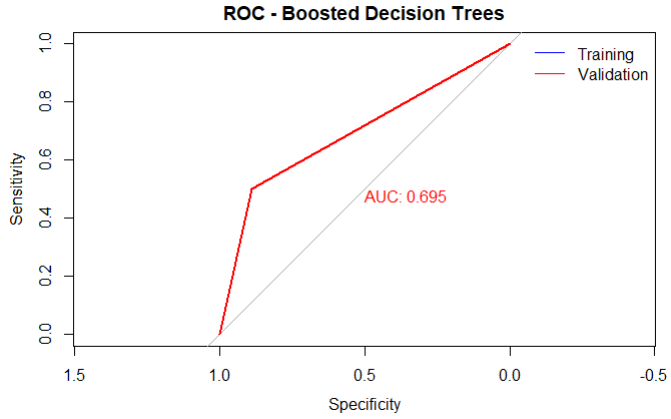
Statistical Analysis

Prior to modeling, the data was randomly split into 80% training and 20% testing data sets. Recall that the goal of this analysis was to predict whether a client will subscribe a term deposit after one contact from the bank. Since this is a classification problem with a binary target, we can implement various supervised learning techniques in an attempt to learn from the provided labels. Then we can compare the different algorithms and choose the most appropriate technique to model this data. We choose to use gradient boosted trees, logistic regression, SVM with a Gaussian kernel, and naive Bayes to model the response.

First, let's consider the boosted decision trees model. This is a good initial technique to use because we can identify, as well as visualize, some of the key features for predicting the response. After performing 10-fold cross validation to obtained the tuned hyperparameters, we found the boosted decision trees model had a training accuracy of 93.59% and a test accuracy of 86.40%. A variable importance plot for predicting the target is shown below. We can note that the cluster group we created as an indication of current market condition, as well as age and the month of contact, seem to be important variables for predicting the response.

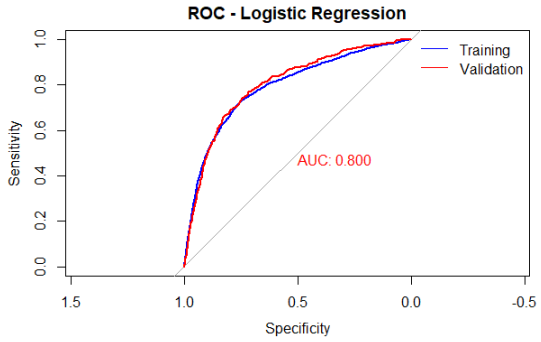


We can also plot the ROC curves for the training and test sets shown below. We find the AUC for the test set is equal to 69.5%.

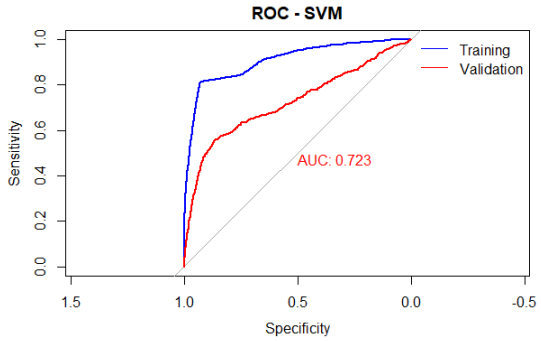


Note that the high accuracy is partly due to class imbalance which can be seen in the ROC curves for the boosted decision trees and all subsequent ROC curves. The limited samples corresponding to client subscribing to term deposits impacts the analysis because if the classifier predicts no subscription all the time, we will still get a high accuracy when we analyze the confusion matrix. This is an aspect of the data we will need to keep in mind during model selection.

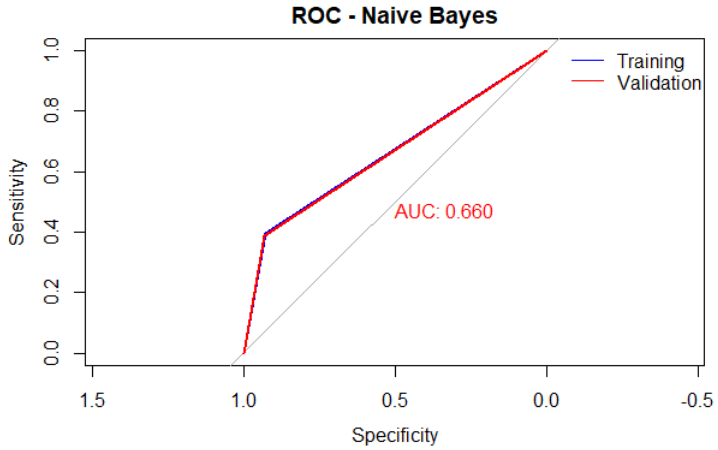
The second modeling approach we will consider is a generalized linear model with a logit link function. In other words, a logistic regression model. We found that the glm model had a training accuracy of 87.18% and a test accuracy of 86.31%. Shown below are the ROC curves for the training and test sets as well as the AUC for the test set which was equal to 80.0%.



Then we can compare the previous methods with an SVM model using a Gaussian kernel. Using 10-fold cross validation to find the gamma and cost hyperparameters, the SVM had a training accuracy of 90.97% and a test accuracy of 87.16%. Shown below are the ROC curves for the training and test set as well as the AUC for the test set which was equal to 0.723.



Lastly, we can use Naive Bayes as a classification technique. We will loosen the assumption of independence among predictors since this is clearly violated, but the algorithm still shows strong predictive power given its simplicity. We assumed that the continuous variables were normally distributed and applied Laplace smoothing to accounting for the zero frequency class in the *default* variable. This approach had a training accuracy equal to 86.09% and a test accuracy of 85.66%. Shown below are the ROC curves for the training and test set as well as the AUC for the test set which was equal to 66.0%.



Model Comparison

The table below lists the training and test accuracy as well as the test set AUC for each of the models under consideration.

Model	Training Accuracy	Test Accuracy	AUC
Boosted Decision Trees	93.59%	86.40%	69.5%
Logistic Regression	87.18%	86.31%	80.0%
SVM	90.97%	87.16%	72.3%
Naive Bayes	85.58%	85.89%	66.0%

From the comparison table, we see that the models have comparable test accuracies which are between 85% and 87%. Due to the class imbalance, this may be a misleading metric and should not solely be used for model selection. The AUC metric indicates that the logistic regression model is the best model, SVM as second best, boosted decision trees as third best, and

Naive Bayes as the fourth best option. When determining the model to select, consideration may also be given to the models' prediction power, the time it takes to train the models, as well as the complexity and interpretability of the models. These factors leave room for different models to be chosen at different times given the requirements.

Final Remarks

As mentioned previously, one issue with this data set is the class imbalance of the target which led to a misleading high overall accuracy. The AUC values highlighted this issue and shed light on the differences between the models. If more data was available, or we implemented a resampling technique, to account for the class imbalance, this issue may have the potential to resolve itself but further analysis will be required. At a later time, we may also consider using the entire data set and include multiple contacts to the client in the model.