Predictive Analytics for Banking and Financial Services

Hayden D. Hampton, STA 6704 Class Project
Department of Statistics and Data Science, University of Central Florida

***Abstract*—In banking, a common practice for the investment banking division (IBD) 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. The aim of this project is to use predictive analytics and machine learning techniques to predict whether clients will subscribe to a term deposit with a Portuguese bank using data from a direct marketing campaign.**

***Index Terms*—classification, machine learning, predictive analytics, banking, financial services, gradient-boosted decision trees, neural networks, naive Bayes, generalized linear models.**

I. Introduction

In banking, a common practice for the investment banking division (IBD) 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 [1]. 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 includes 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.

II. 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 Table 1.

Table I
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 client-specific 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 *subscription* 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 length of time 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.

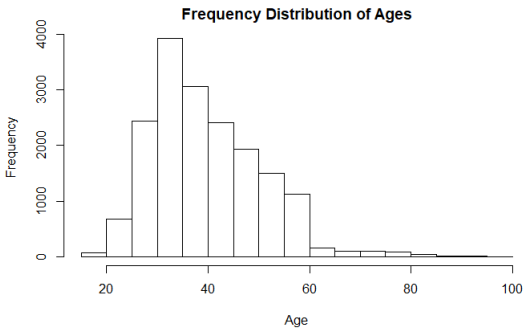

Figure 1. Frequency Distribution of Client Age

III. Exploratory Analysis

Initially, let's consider the distribution of the response shown in Table 2. Note the severe class imbalance in the target's frequency distribution. There are more than six times the number of observations in the "No" class than there are in the "Yes" class. This class imbalance will need to be addressed if conventional classification algorithms are to be used.

Table II
Frequency Distribution Table of Target

No	Yes
15,342	2,300

Next, we can plot the frequency distribution of client age using a histogram. The histogram is shown in Figure 1 and the corresponding five-number summary statistics and the mean are shown in Table 3. Note that the distribution is right skewed, but that the average client age lies close to the median age of 38.

Table III
Five Number Summary Statistics and Mean

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

A frequency distribution which suggests a pattern in the data is the month of the year that the client was contacted against the target. This frequency distribution is shown in Table 4. We can note that there were no calls in January and February. At this time, it is not clear why this is the case. Also note 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.

IV. Data Preparation

The imbalance of the response arises from the class distribution because most clients prefer not to invest in a term deposit when contacted by the bank. This results in a problem when using the original data with conventional classification algorithms. Bias towards the majority class is common since the loss function does not take the data distribution into account during optimization. This leads to issues when implementing threshold-dependent evaluation metrics, such as accuracy, which require a confusion matrix to be calculated using a hard cutoff on predicted probabilities. Therefore, it would be ill-advised to implement these algorithms without first accounting for the imbalance.

Table IV
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

Many techniques have been proposed to learn from class-imbalanced data [2]. There are various short-comings to randomly oversampling the minority class as well as randomly undersampling the majority class. Random oversampling can lead to overfitting while random undersampling has the potential to remove important examples from the analysis. The effect of imbalance in a data set was assessed given the aforementioned concerns, and it was found that random resampling was effective. It was also observed that using sophisticated sampling techniques did not provide a clear advantage in the domain considered [3].

Therefore, to account for the imbalanced data we can implement a random oversampling strategy on the minority class and later use a stratified sampling technique for the training and test partition. Lastly, we can check for any missing values in the data set, but find there are none present. Therefore, imputation is not required.

Dimension Reduction

In an effort to reduce the dimensions of the feature set and improve model performance, we can use K-means 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 three clusters were appropriate. We can characterize the three clusters as a nominal variable where each of the three heterogeneous levels loosely corresponds to a given state of the financial markets at the time the customer is contacted. This allows us to convert 5 quantitative variables into one qualitative variable with 3 levels. The clusters can be visualized in Figure 2.

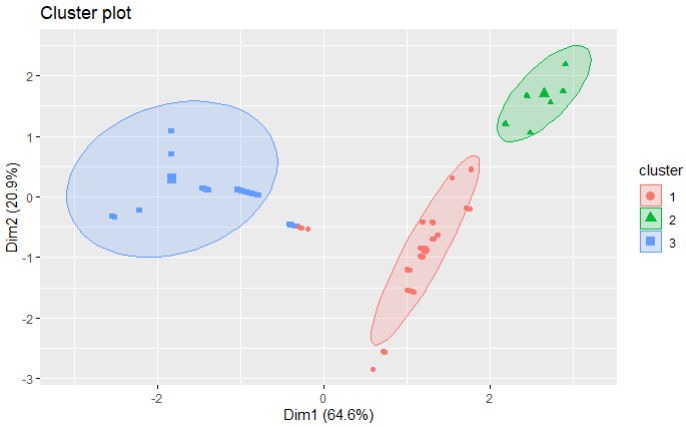

Figure 2. Visualization of the K-means clusters of the social and economic variables

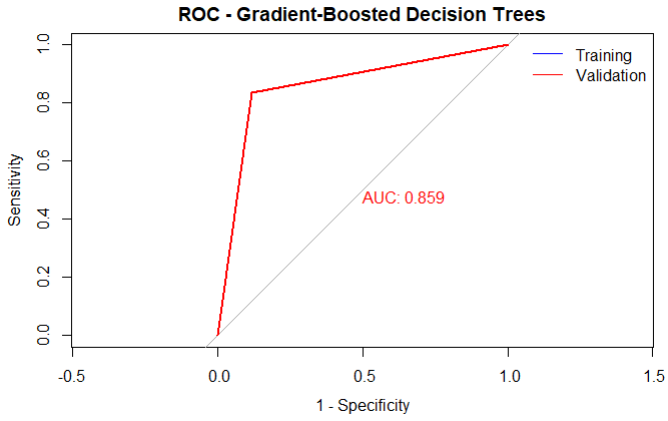

Figure 4. Gradient-Boosted Decision Trees ROC Curves

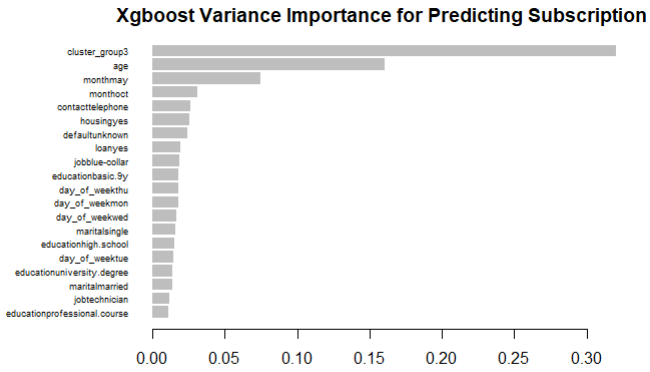

Figure 3. Gradient-Boosted Decision Trees Variable Important Plot

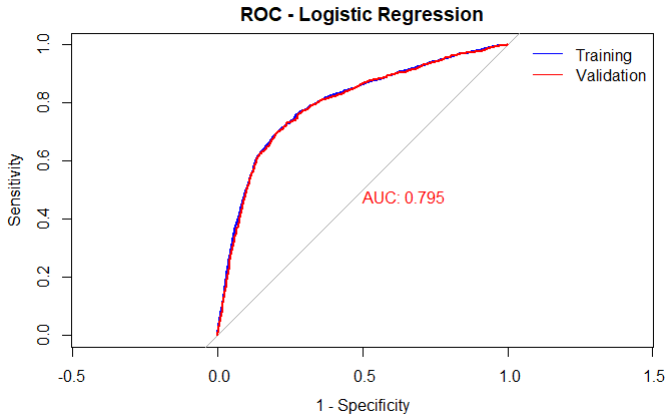

Figure 5. Logistic Regression ROC Curves

Statistical Modeling

Prior to modeling, the data was randomly split into 80% training and 20% testing data sets using stratified sampling based on the target class. 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 the data.

First, let's consider gradient-boosted decision trees algorithm. 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 obtain the tuned hyperparameters, we found the gradient-boosted decision trees model had a training accuracy of 90.15% and a test accuracy of 85.70%. A variable importance plot for predicting the target is shown in Figure 3. Note that the cluster groups previously created, as well as client age and month of contact, seem to be important variables for predicting the response.

Figure 4 displays the receiver operator characteristic (ROC) curves for the training and test sets as well as the area under the curve (AUC) for the test set which was equal to 85.9%. We can note that this model does not seem to be overfitting the training data and generalizes to the test set well.

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 74.14% and a test accuracy of 74.11%. Shown in Figure 5 are the ROC curves for the training and test sets as well as the AUC for the test set which was equal to 79.5%. We can note that this model seems to be slightly overfitting the training data and generalizes to the test set fairly well.

The third modeling approach we will consider is Naive Bayes. 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 73.65% and a test accuracy of

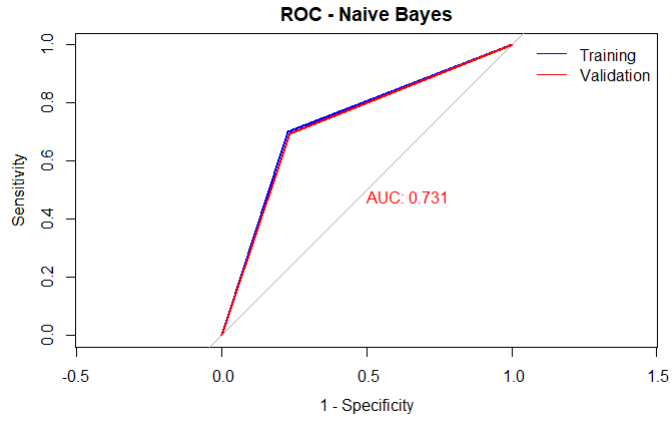

Figure 6. Naive Bayes ROC Curves

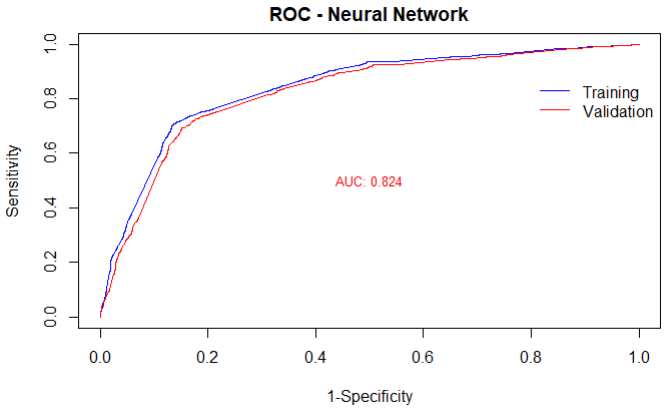

Figure 7. 5-Layer Neural Network ROC Curves

73.05%. Shown in Figure 6 are the ROC curves for the training and test set as well as the AUC for the test set which was equal to 73.1%. We can note that this model seems to be slightly overfitting the training data and generalizes to the test set fairly well.

Lastly, we will consider using a neural network as the final classification technique. After tuning the hyperparameters, we found that a 5 layer neural network had a training accuracy of 78.57% and a test accuracy of 76.96%. Shown in Figure 7 are the ROC curves for the training and test sets as well as the AUC for the tet set which was equal to 82.4%. We can note that this model has the most overfitting out of all the models, but still generalizes to the test set well.

Model Comparison

Table 5 lists the training accuracy, test accuracy, and AUC for the test set for each of the models under consideration.

From the comparison table, we see that the gradient-boosted decision trees model scored the highest in all three categories. The second best model based off of the test accuracy and AUC is the neural network. Third is the logistic regression model and fourth is the naive Bayes model.

Table V
Model Comparison

Model	Training Accuracy	Test Accuracy	AUC
Gradient-Boosted Decision Trees	90.15%	85.70%	85.9%
Logistic Regression	74.14%	74.11%	79.5%
Naive Bayes	73.65%	73.05%	73.1%
Neural Network	78.575%	76.96%	82.4%

Final Remarks

When determining the best model to select, consideration may be given to the predictive power of the model, the time it takes to train the model, as well as the complexity and interpretability of the model. These factors leave room for different models to be chosen at different times given the requirements, but overall the gradient-boosted decision trees model seems preferable for the data.

Instead of only considering the data from the initial call, future consideration may be given to using the entire data set which would include multiple contacts to a given client. Incorporating this additional information into the modeling may provide additional insight into client behavior and improve the prediction power when determining whether clients will subscribe to a term deposit when contacted by the financial institution.

References

[1] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[2] Guo Haixiang et al. “Learning from class-imbalanced data: Review of methods and applications”. In: *Expert Systems with Applications* 73 (2017), pp. 220–239. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2016.12.035. URL: https://www.sciencedirect.com/science/article/pii/S0957417416307175.

[3] Nathalie Japkowicz. “The Class Imbalance Problem: Significance and Strategies”. In: *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI*. 2000, pp. 111–117.