STA 6933 Advanced Topics in Statistical Learning

Project 2
Spring 2022
Marc Sandoval, Yahui Peng, Sanbrina Fautheree

5/10/2022

Abstract

A bank compiled customer information during a campaign to identify the cause for revenue decline, which they concluded was caused by existing customers not obtaining long-term deposit accounts. In the project, we attempt to identify which customers are likely to obtain long-term deposit accounts based on customer information and demographics. Since the compiled data set mainly consists of categorical predictors, we will focus on tree-based methodologies in this project.

Overview

For this project, we are going to be looking at a banking data set to try and predict the likelihood of a customer obtaining a long-term deposit account based on their information and demographics. This data set has a total of 32,950 observations within it. We are looking at the data set from a business perspective. Meaning we are wanting to try a figure out which customers we have that are more likely to have a long-term deposit account. We will be examining the data through various tree-based modeling techniques considering a majority of predictor variables are categorical. Our analysis of data includes exploratory data analysis, pre-processing, feature selection, modeling, and performance evaluation of five different tree-based model types. These various techniques are what lead us to our conclusions.

Data Structure

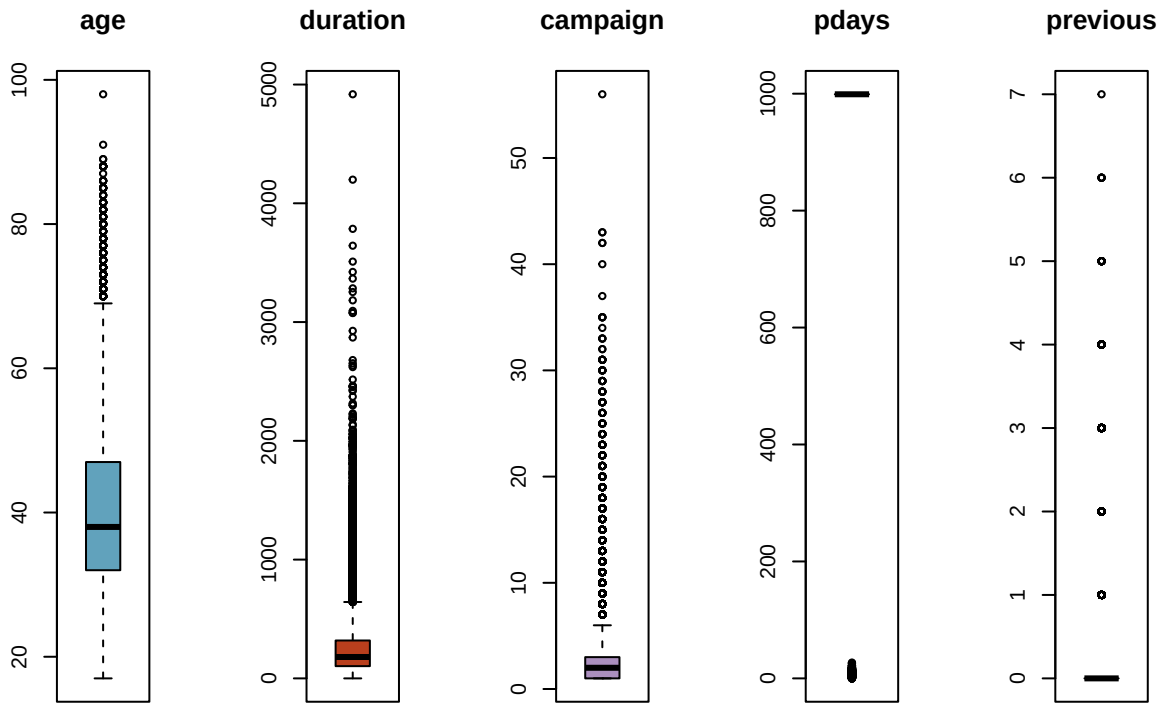
The banking data consists of 15 predictors and 1 binary response variable, among which 10 predictors are nominal and 6 predictors contain nulls.

Table 1: Predictor Variables

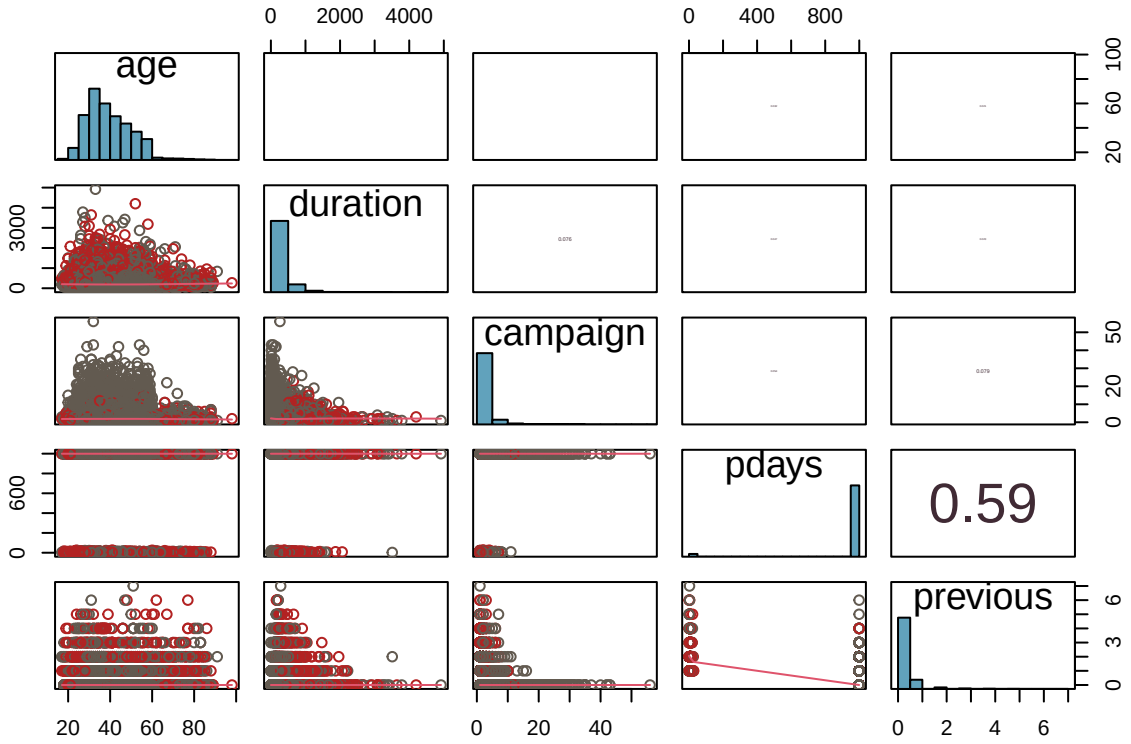
Predictor	Data Type	Contains Nulls
age	numeric	no
job	nominal	yes
marital	nominal	yes
education	nominal	yes
default	nominal	yes
housing	nominal	yes
loan	nominal	yes
contact	nominal	no
month	nominal	no
dayofweek	nominal	no
duration	numeric	no
campaign	numeric	no
pdays	numeric	no
previous	numeric	no
poutcome	nominal	no

As **pdays** represents the number of days that passed by after the client was last contacted from a previous campaign (999 means the client was not previously contacted). Therefore, for better visualization and interpretation, we encode **pdays** == 999 to NA for EDA purposes only, but they are immediately encoded back to 999 after EDA.
Summary statistics and Box-plots suggest outliers exist for all predictors, however, tree-based models are the models of interest for this project and are outlier-insensitive. Therefore, we do not resolve the outliers.

```
##       age           duration         campaign          pdays      
##  Min.   :17.00   Min.   :   0.0   Min.   : 1.000   Min.   :  0.0  
##  1st Qu.:32.00   1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.:999.0  
##  Median :38.00   Median : 180.0   Median : 2.000   Median :999.0  
##  Mean   :40.01   Mean   : 258.1   Mean   : 2.561   Mean   :962.1  
##  3rd Qu.:47.00   3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.:999.0  
##  Max.   :98.00   Max.   :4918.0   Max.   :56.000   Max.   :999.0  
##     previous        
##  Min.   :0.0000  
##  1st Qu.:0.0000  
##  Median :0.0000  
##  Mean   :0.1747  
##  3rd Qu.:0.0000  
##  Max.   :7.0000
```



Pairwise scatterplots, correlation matrix, and histograms for numeric preditors are shown below. First, in pairwise scatterplots, points are colored in red and grey for y = 1 and 0, respectively. We observe that clients with longer **duration** are more likely to have subscribed to a term deposit (y = 1). No obvious trend is observed between y and other numeric predictors. Second, The font size of the correlation coefficients displayed in the upper panel is proportional to the magnitude of the value. Therefore, no strong correlation is present between each pair of numeric predictors. Last, histograms show all numeric predictors are skewed.



Missing completely at random (MCAR) is the desirable scenario in case of missing data. Assuming data is MCAR, too much missing data can be a problem, too. Usually, a safe maximum threshold is 5% of the total for large datasets. The aggregation plot helps us understanding that almost 74% of the samples are not missing any information, 21% are missing the **default** value, and the remaining ones show other missing patterns. We probably should leave the feature **default** out.

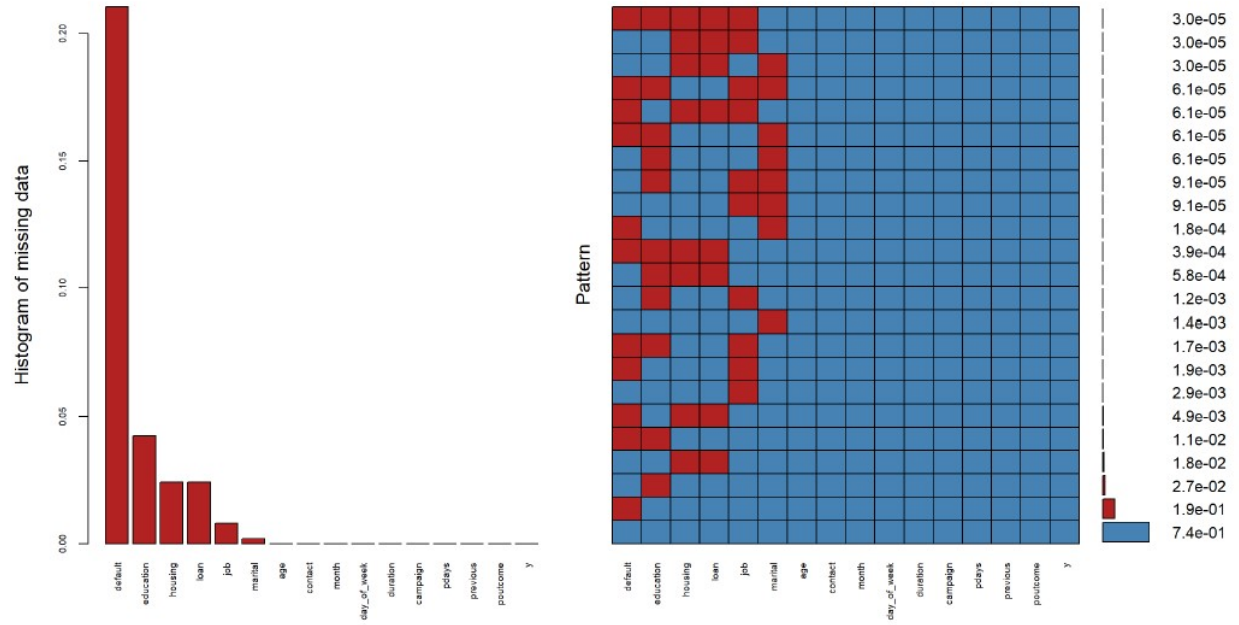


Figure 1: Missing Data Pattern

Data Imputation

Our objective in imputing missing data for our categorical predictors was to retain the distribution of the variable after imputation. To achieve this, we apply the *mice()* function with the *random* method, which imputes the missing observations by replacing them with a random sample from the observed values.

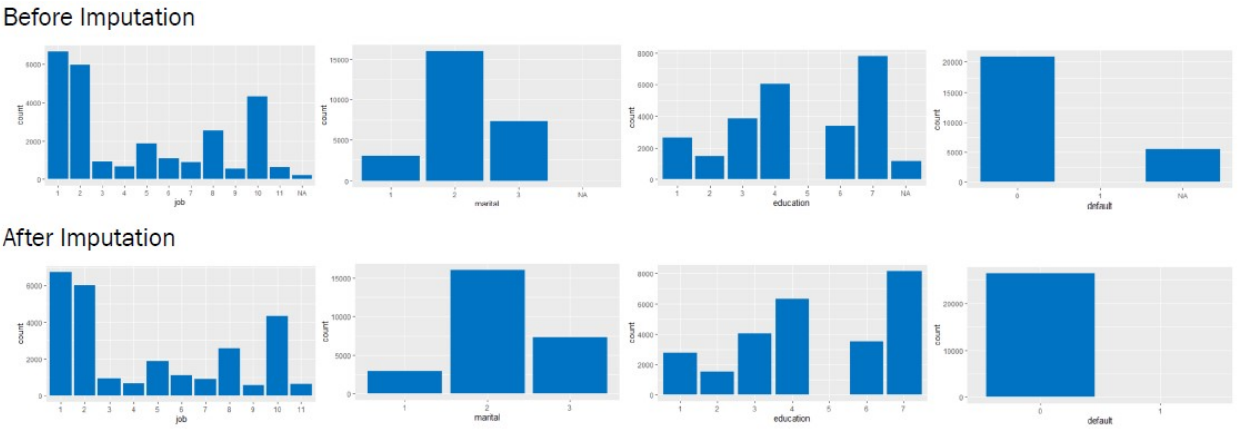
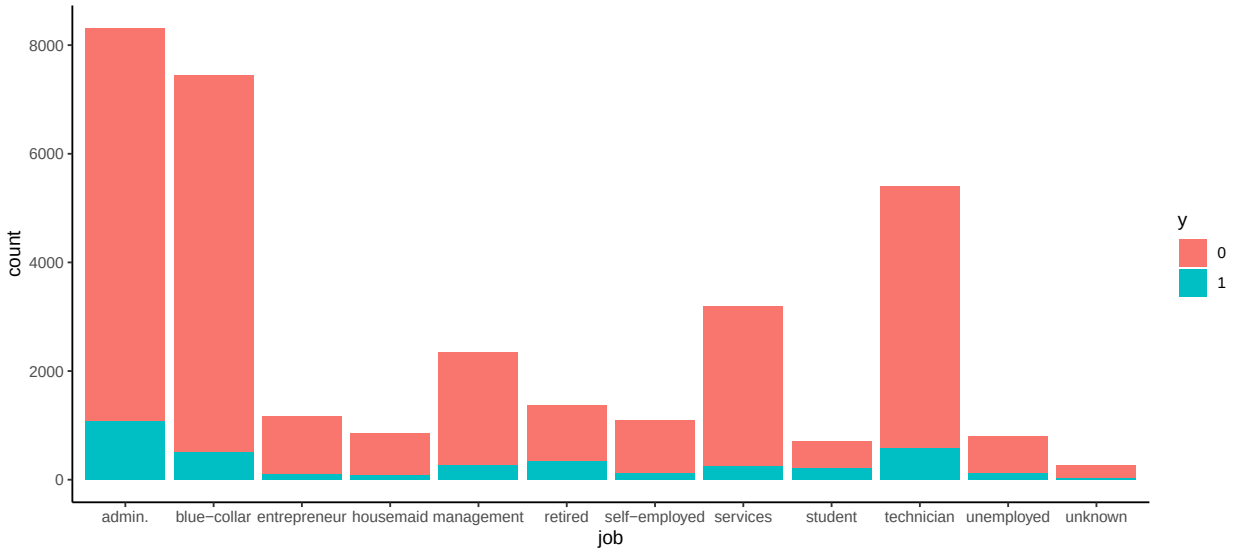
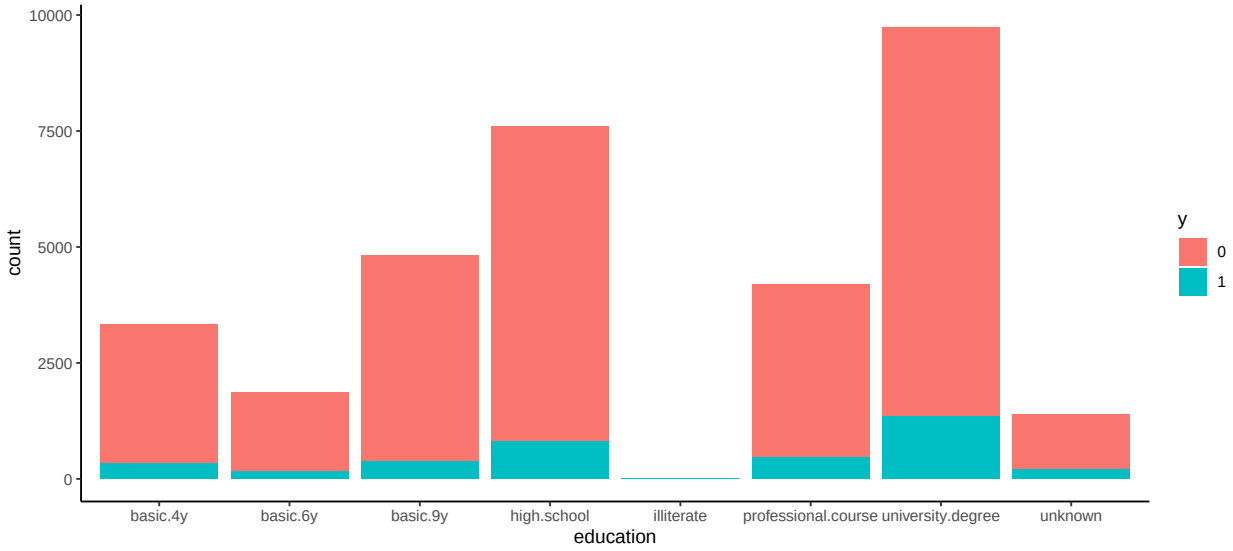
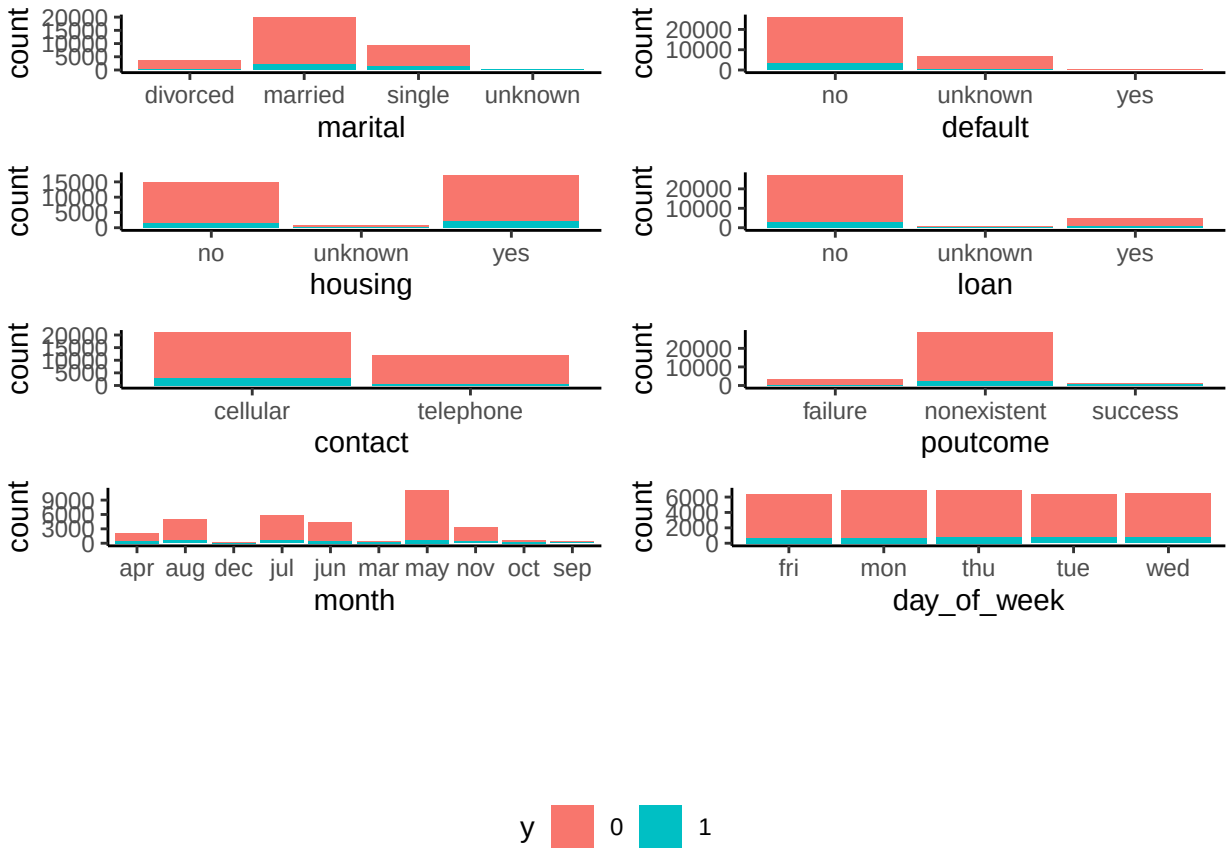


Figure 2: Distribution on Predcitor Variables

Bivariate analysis for ten categorical predictors is shown below. It is observed that **month** dec, mar, oct, and sep show a noticeable probability of y=1, which is nearly 50 %, indicating a possible lack of impact of these month categories on the response.


8000
6000
4000
2000
0
count
admin.
blue−collar
entrepreneur
housemaid
management
retired
self−employed
services
student
technician
unemployed
unknown
job
y
0
1


10000
7500
5000
2500
0
count
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
education
y
0
1



Imbalanced Data

We observe that our response variable is imbalanced, which may cause poor model performance in the minority class. We applied the *SMOTE()* algorithm to create new observation in the minority class, which uses the k-nearest neighbor algorithm to generate new observation. The resulting training data set is more balanced.

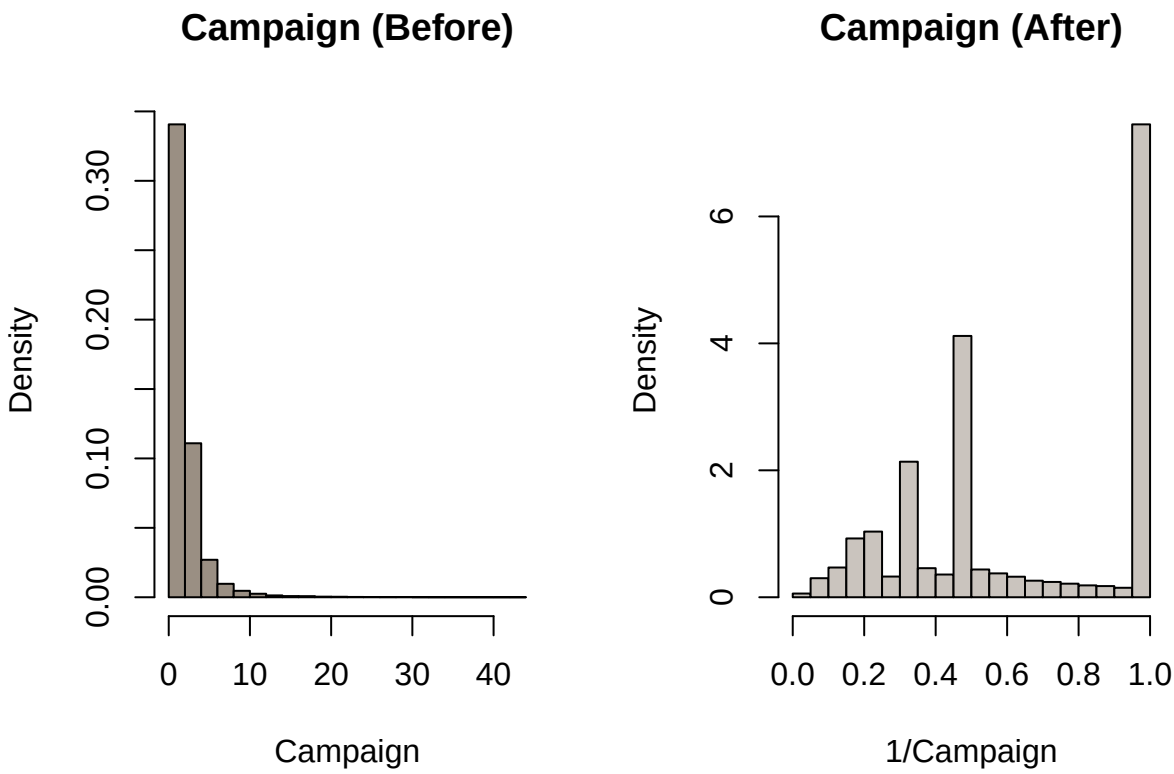
Table 2: Response Variable Proportion

Class	Count	%
Y=0	29,238	0.887
Y=1	3,712	0.113

Table 3: Response Variable Proportion after Balancing

Class	Count	%
Y=0	23,391	0.568
Y=1	17,820	0.432

The skewness of numeric predictors are checked, and Box-Cox transformation suggests a reverse on campaign. Comparative histograms show that skewness is relieved after the transformation.



Near-zero variance predictors are further obtained. Considering that default have 21% missing in the original data, and has near-zero variance after imputation, we should consider drop it out. Additionally, job 3, 4, 6, 7, 9, 11, education 2 and 5, month 3, 9, 10, and 12 show near-zero variance.

```
##  [1] "default"    "pdays"      "job3"       "job4"       "job6"
##  [6] "job7"       "job9"       "job11"      "education2" "education5"
## [11] "month3"     "month9"     "month10"    "month12"
```

Analysis

Feature Selection

We utilized a logistic regression model to identify which predictors seem reasonable and used these predictors to construct our tree-based models. Based on multiple modeling iterations, we observed that the following combination of predictors resulted in a logistic regression model with all predictors significant at an $\alpha = 0.001$ level: age, education, job, and marital status. In particular, the dummy variables for 9th grade education and university degrees were kept due to their significance. For job, the dummy variables for blue collar, retired, services, and student responded the best and will be retained. For martial status, the dummy variable for single was retained due to this analysis. In other words, this analysis provides evidence that 9th-grade or university educated, blue-collar/retired/service/student employed, and single customers appear to be more likely to open long-term deposit accounts.

Models

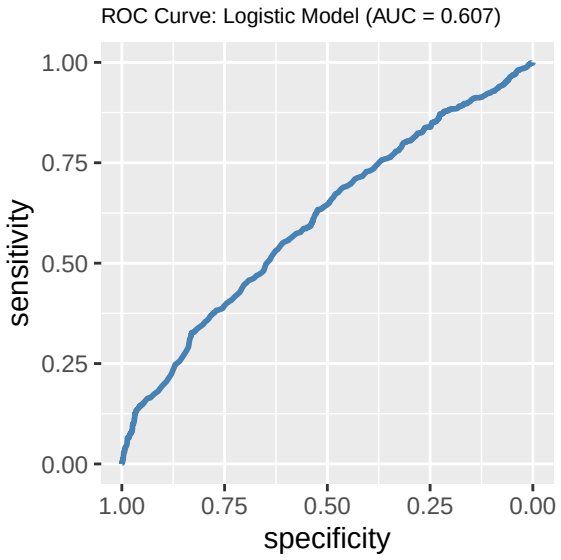
Both summary of the baseline logistic regression model and variable importance of the random forest model suggest that age, education3, education7, job2, job6, job8, job9, and marital3 are informative predictors.

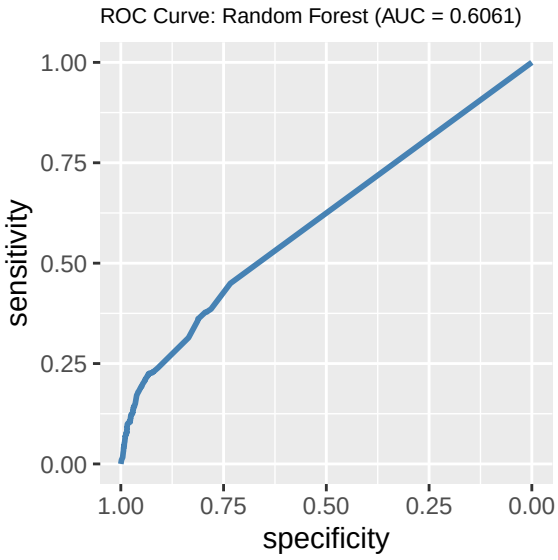
Boosting model is further fitted using n.trees = 1000. Moreover, we fit BART and C5.0 classification models. The training and test errors as well as ROC curves are presented below.

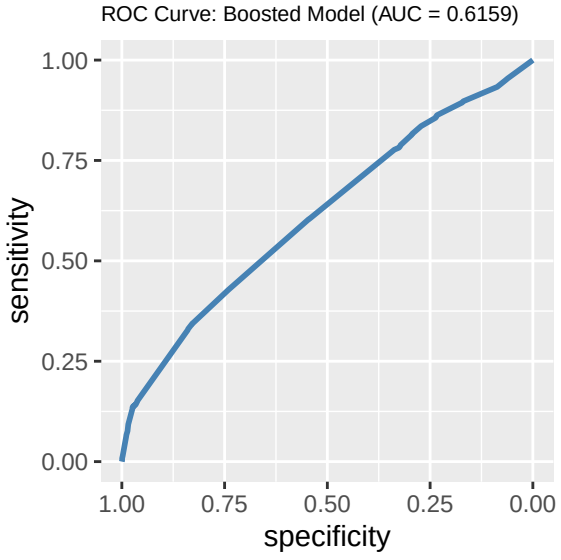
Table 4: Model Performance–Misclassification Error

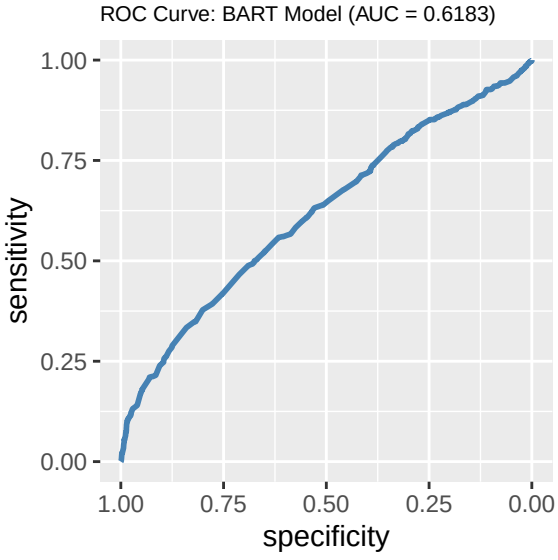
Model	Training Error	Testing Error
Logistic Model	0.3262	0.1490
Random Forest	0.1466	0.1287
Boosted Model	0.2182	0.1283
BART Model	0.1421	0.1279
C5.0 Model	0.1267	0.1306

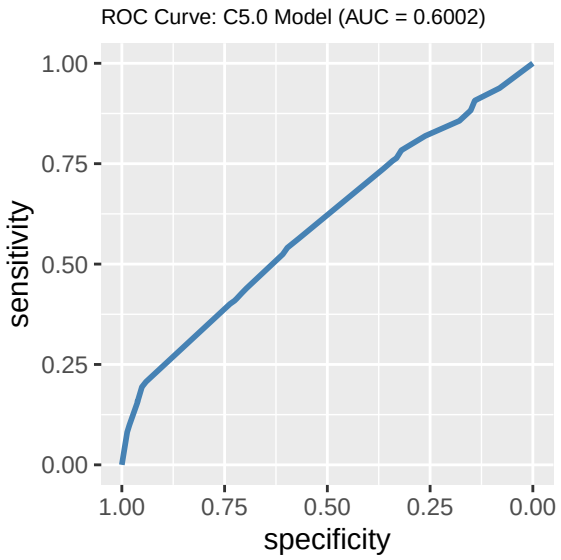
ROC Curves











Conclusion

In this project, we identified the customer demographics that were likely to result in long-term deposit accounts. In order to model the data, we performed data understanding and data pre-processing, such as imputation and balancing. We utilized tree-based methodologies since the majority of our predictors were categorical. We observed that the testing errors and AUC measures were similar across the 5 models we built, but the logistic regression model and boosted tree model had the highest training error, indicating these models were less flexible than the other 3 models. Based on the misclassification errors and AUC, it appears that the BART model performed the best for this data set.