MA468: Predictive Analytics & Data Mining

Project 1 | Bank & Loan Classification

Nathan Gonyo, Aaron Neiger, Jack Michalowski

Dr. Rasitha Jayasekare

Table of Contents

Introduction & Background Information on Dataset

The dataset for this project is a “Bank Marketing Dataset” from UCI Machine Learning data repository. The dataset has 45211 records and 17 variables. The data surrounds the marketing campaign of a Portuguese banking institution, with the response variable indicating whether the customer invested in a bank term deposit.

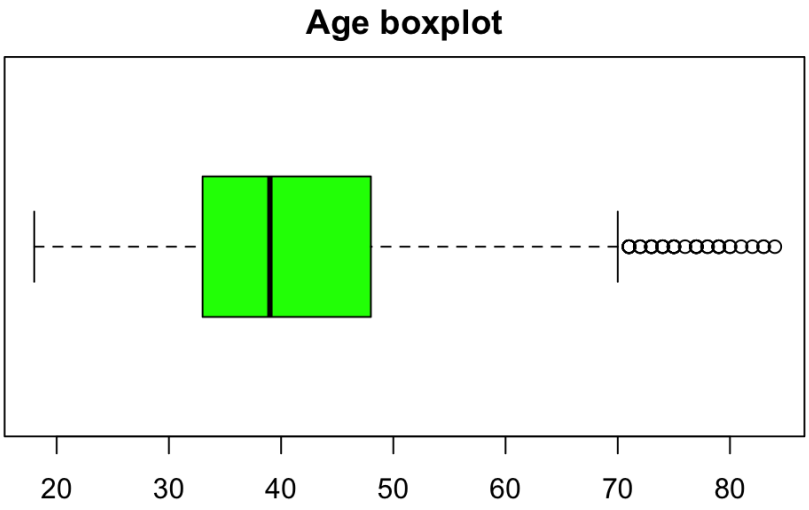
Variables:

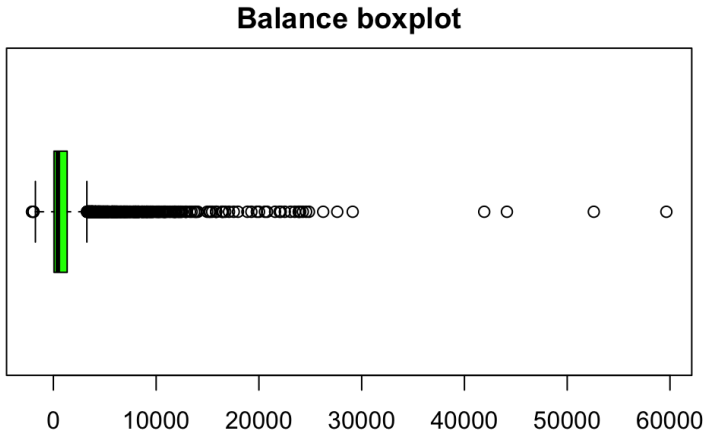
1: age - age
2: job - type of job (categorical)
3: marital - marital status (categorical: 'divorced','married','single','unknown')
4: education (categorical)
5: default - does customer have credit in default
6: balance - balance of loan***
7: housing - does customer have housing loan
8: loan - does customer have personal loan
9: contact - contact communication type (categorical: 'cellular','telephone')
10: day - last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11: month - last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
12: duration - last contact duration, in seconds (numeric).
Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no')
13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

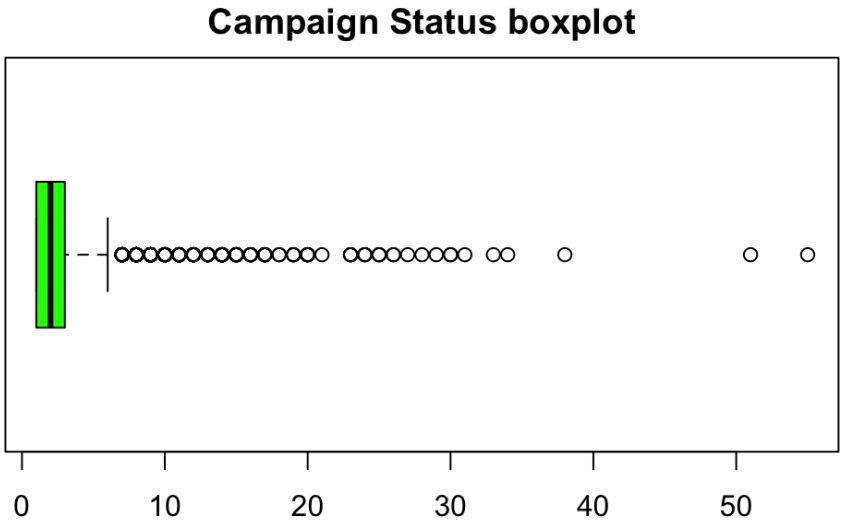
Response variable:
17 - y - has the client subscribed to a term deposit? (binary: 'yes','no')

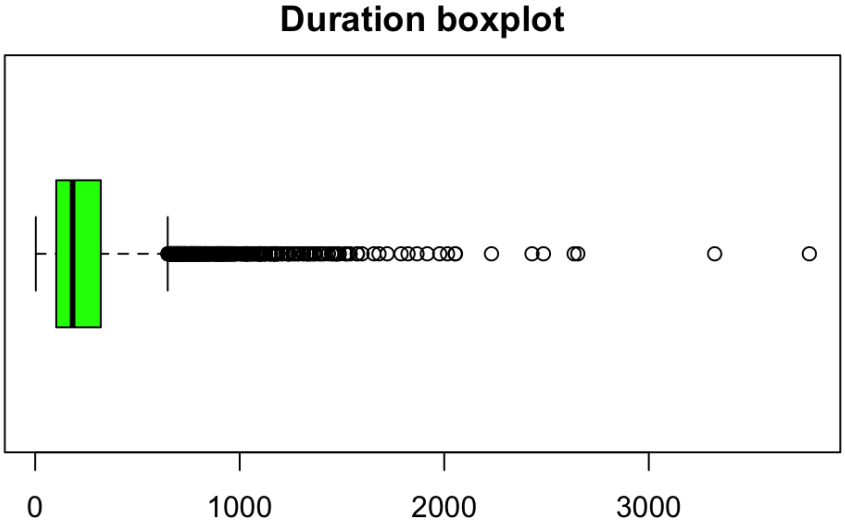
Dataset Preprocessing

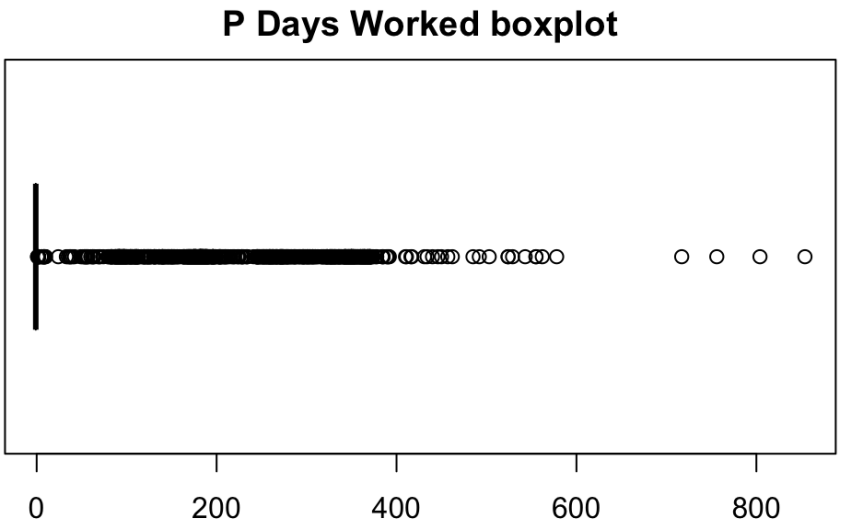
For this project, we are going to use a random sample of 5000 records rather than the entire dataset of more than 40,000 records. The first step of preprocessing was to randomly select 5000 records and use this dataset going forward. The next step is to remove any rows within the random sample which contain missing values, of which there were none found. Now, we began outlier removal within the quantitative variables. For this step, we first generated boxplots for each quantitative variable to check for the presence of outliers. The variables which we have to check for outliers are age, balance, duration, campaign, pdays, previous.

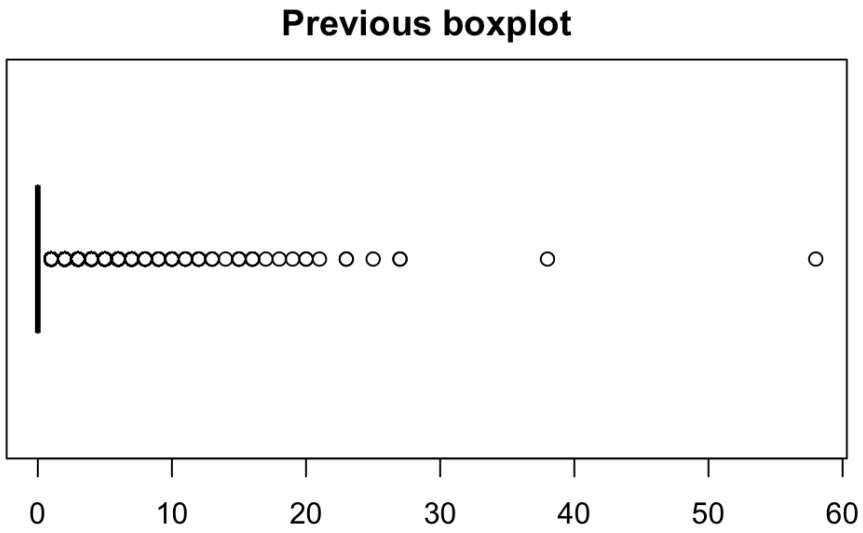




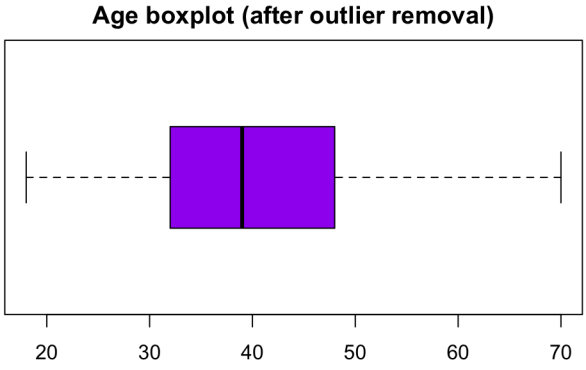

Campaign Status boxplot
0
10
20
30
40
50

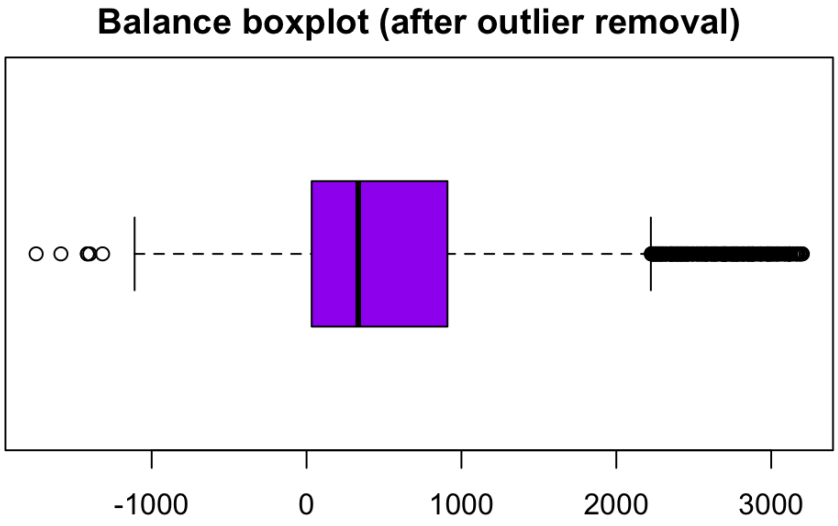

Duration boxplot
0
1000
2000
3000

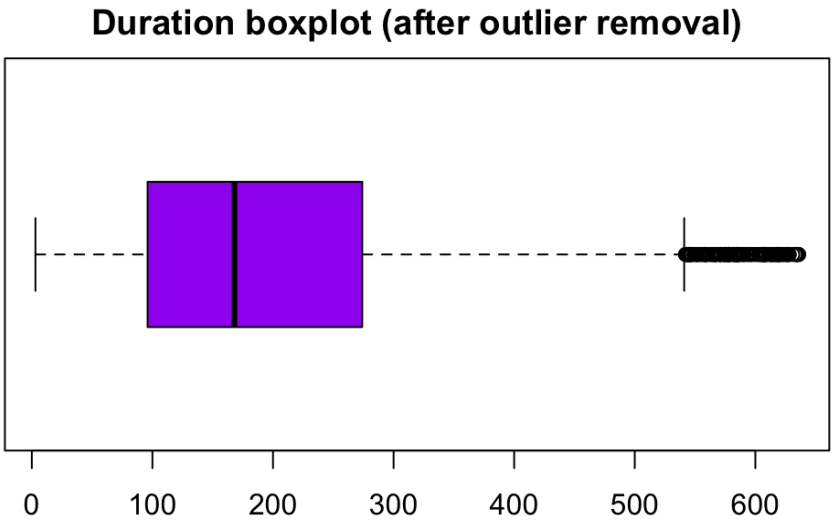


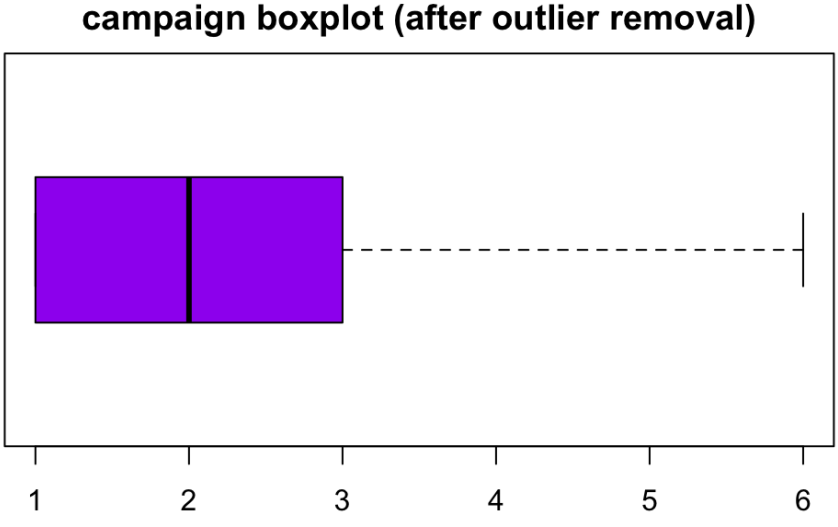


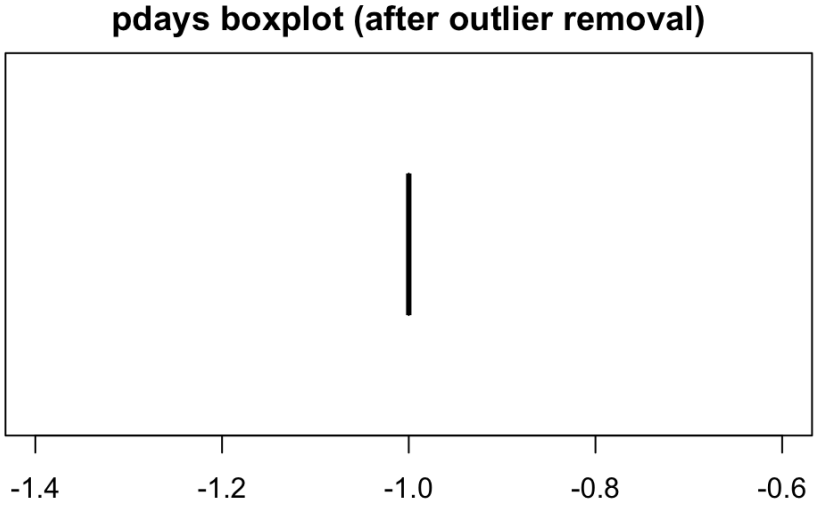
We can see that all the quantitative variables contain outliers, so we will remove the outliers for each variable individually, then create boxplots again to see how the presence of outliers changed.

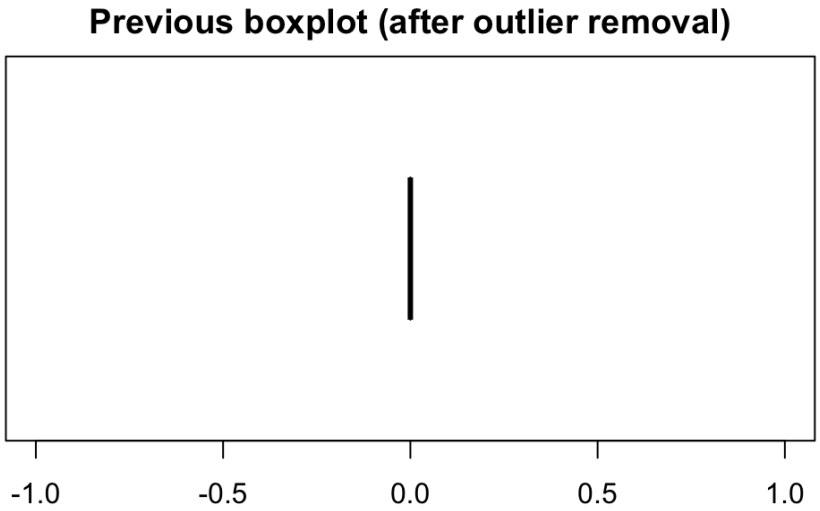



Balance boxplot (after outlier removal)
-1000
0
1000
2000
3000


Duration boxplot (after outlier removal)
0
100
200
300
400
500
600


campaign boxplot (after outlier removal)
1
2
3
4
5
6





We have successfully removed outliers for all the quantitative variables. We can now move onto data description and visualization.

Note that for pdays and previous, because the vast majority of observations were a certain value, after removing outliers all the observations were that heavily observed value. This makes sense but now that these variables are essentially “constants”, they will be removed from the dataset before cluster analysis is performed.

Numeric Description of Dataset

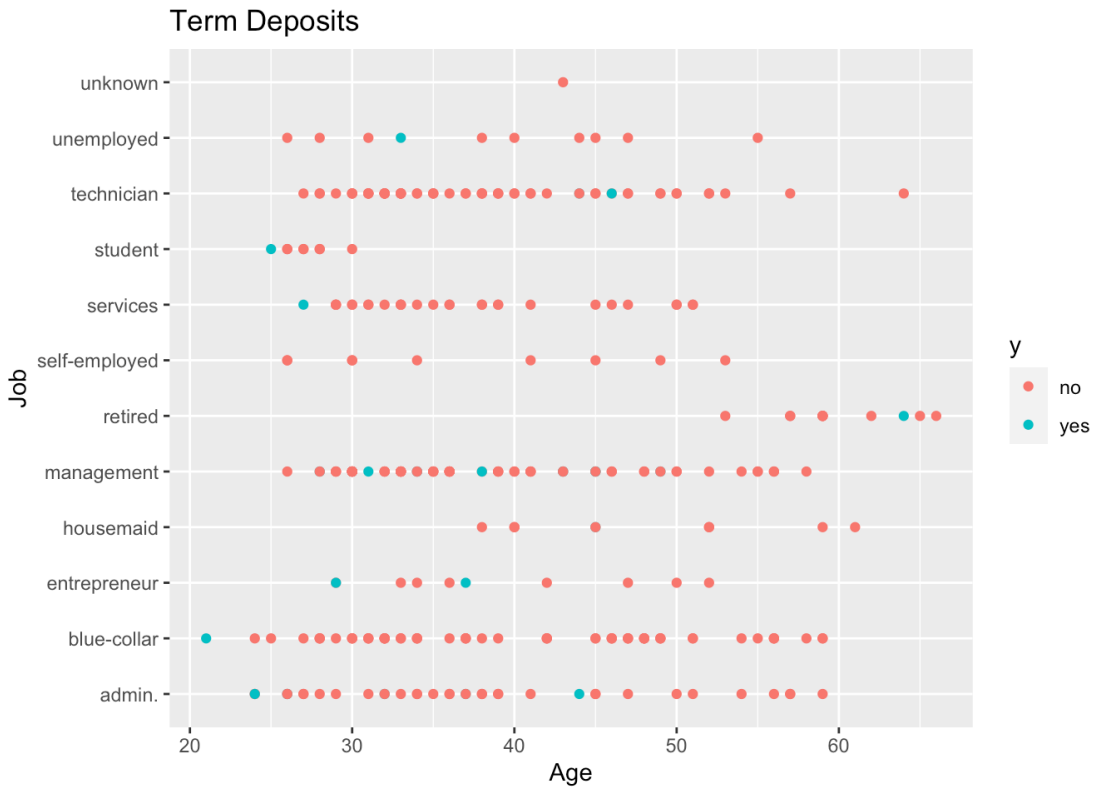
For numeric description of the dataset, we will need summary statistics for each quantitative variable contained in the dataset. Recall that after outlier removal, the variables pdays and previous contained only one value in all entries, so we only need to summarize the quantitative variables age, balance, duration, campaign.

Numeric Description of Dataset (Quantitative Variables)

	Min	Mean	Median	Max	Std. Dev.
Age	18.00	40.33	39.00	69.00	9.82
Balance	-1746.00	569.60	310.00	3201.00	780.59
Duration	3.00	205.50	173.00	636.00	136.69
Campaign	1.00	2.17	2.00	6.00	1.32

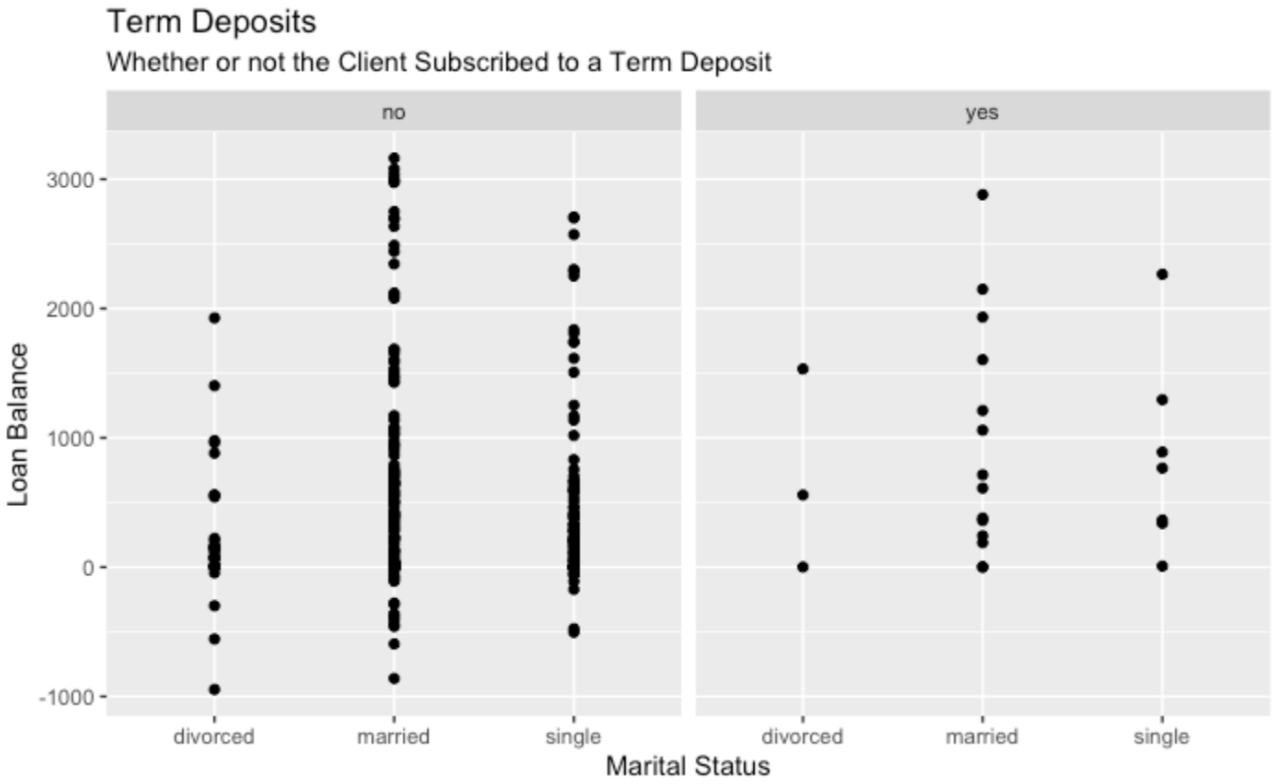
Visual Description of Dataset

Graph 1



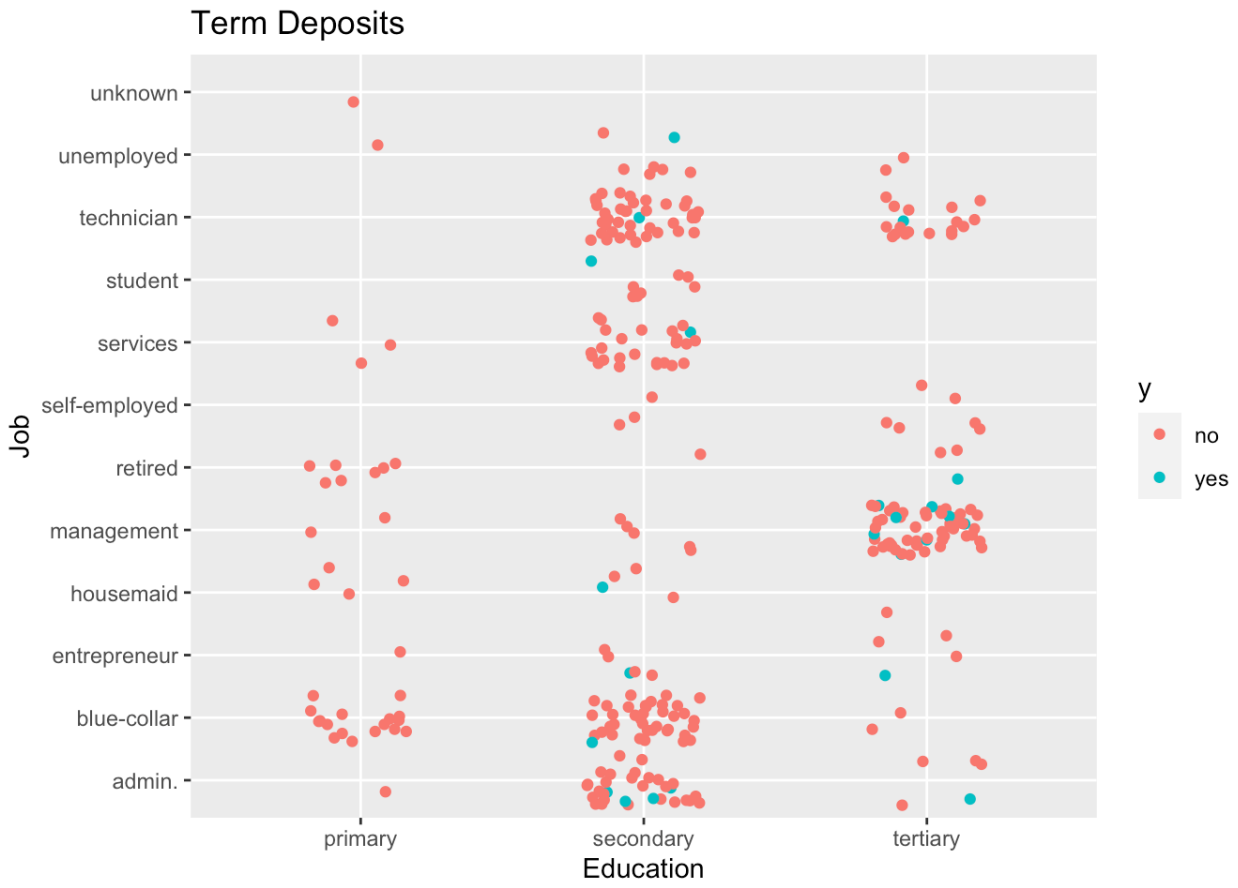
This multivariate ggplot graph shows the age and job type of the clients and whether or not they subscribed to a term deposit. This graph is useful to tell which job type and which age are more likely to subscribe to a term deposit.

Graph 2



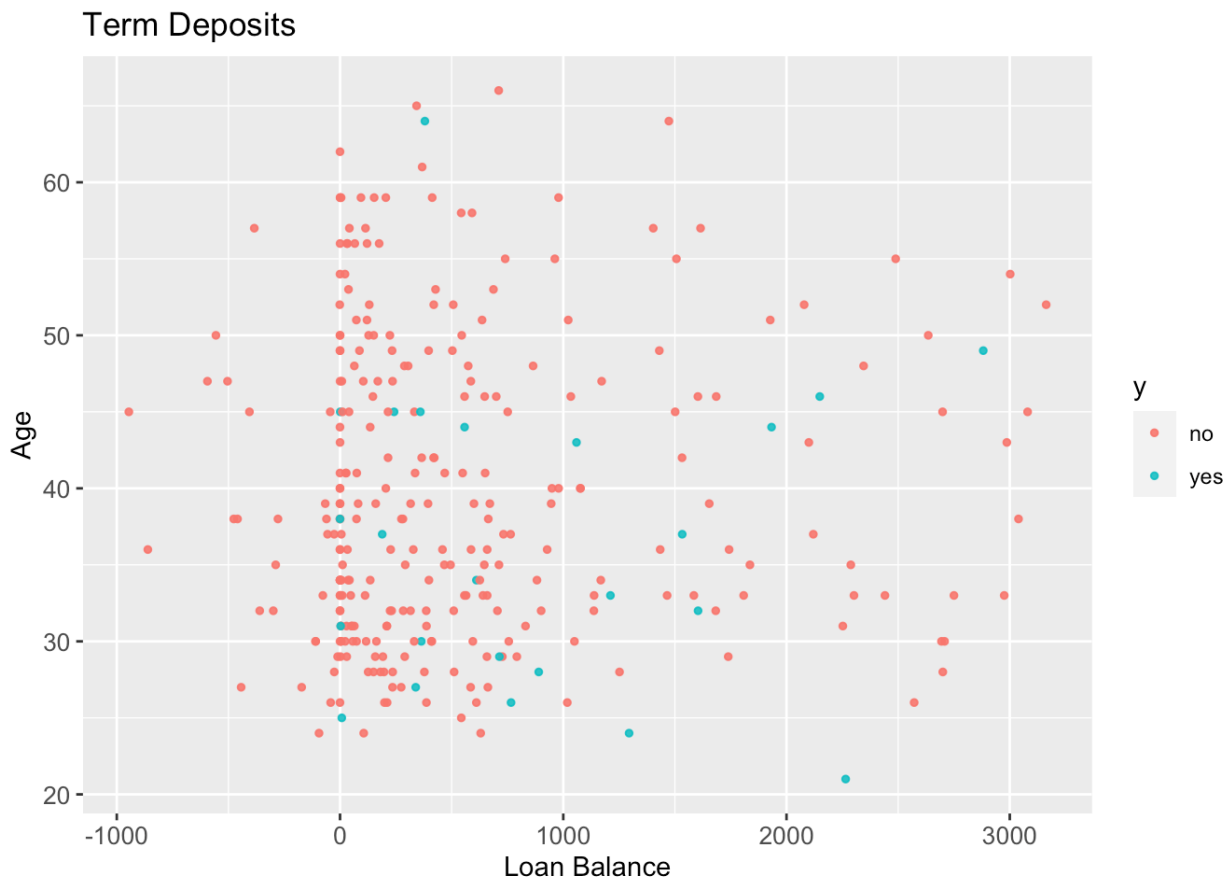
This multivariate ggplot graph shows the marital status and balance of the loan of the client. Then they are split into two partitions of whether or not they subscribed to a term deposit. This graph is useful to tell if a person is more likely to subscribe to a term deposit based on their marital status and loan balance.

Graph 3



This multivariate ggplot graph shows the job type and education level of the client and whether or not they subscribed to a term deposit. This graph is useful to tell if a person is more likely to subscribe to a term deposit based on their job type and level of education.

Graph 4

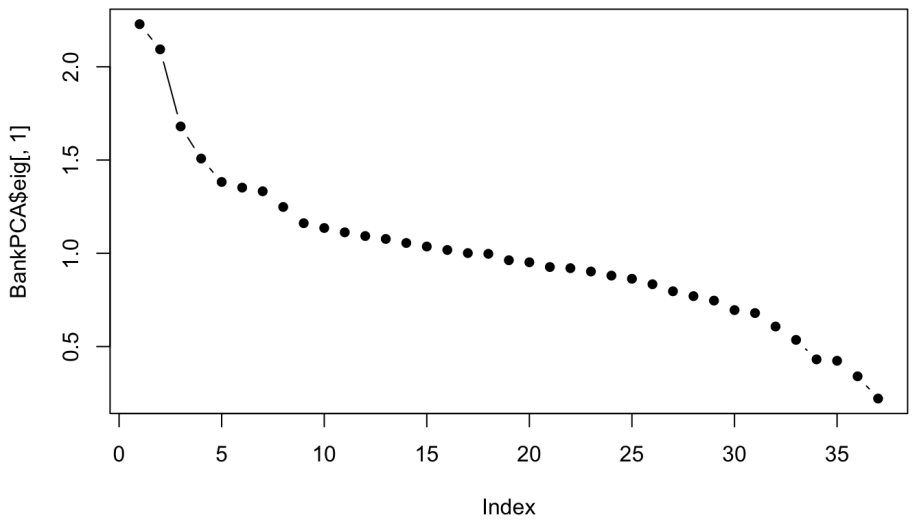


This multivariate ggplot graph shows the age and loan balance of the client and whether or not they subscribed to a term deposit. This graph is useful to tell if a person is more likely to subscribe to a term deposit based on their age and loan balance.

Creation of Train / Test Dataset, Addressing Class Imbalance, Mixed PCA

The final steps of preprocessing before the application of classification algorithms begin with the creation of training and testing datasets. For this we randomly selected 80% of the entries from our dataset at this point in the project to go into the Train dataset. The remaining 20% go into the testing dataset. After the creation of the testing dataset, we had to address a class imbalance problem as the response variable “y” contained a vast majority of responses of “no”, with only a small portion “yes”. To make up for this, we applied oversampling of the “yes” variable and undersampling of the “no” variable. Until they both were equal. Before addressing class imbalance, there are only 135 instances of the response variable "y" as yes and 2332 instances of the response variable "y" being no. After, there were 405 of each.

The last step of preprocessing is principal component analysis. Based on the eigenvalue approach, we select PC's with eigenvalues greater than 1, and we ended up selecting the first 17 PC's. Based on the cumulative proportion approach, we ended up selecting the first 25 PC's assuming we wanted to preserve 80% of the dataset's variance. Finally, based on the based on the screeplot, it seems that after 9 it levels out, so we select first 9 PC's.



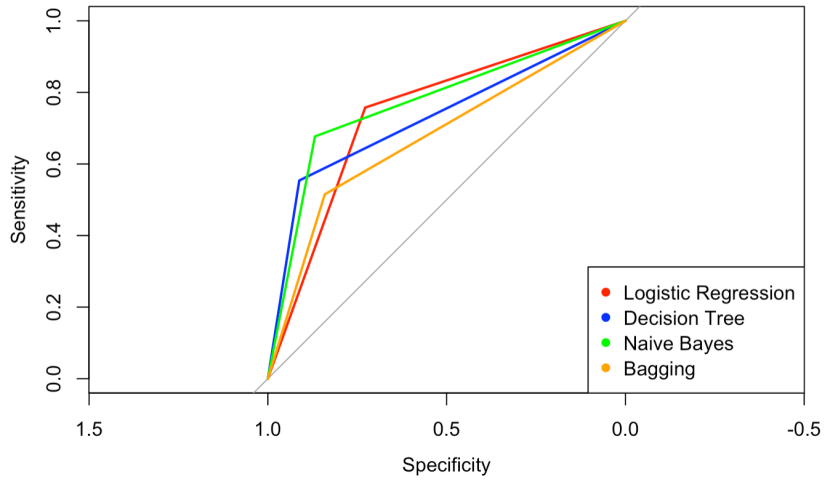
```
Squared loadings :
           dim 1 dim 2 dim 3 dim 4 dim 5 dim 6 dim 7 dim 8 dim 9 dim 10 dim 11 dim 12 dim 13
age         0.00  0.60  0.01  0.01  0.00  0.03  0.03  0.00  0.01   0.00   0.00   0.00   0.01
balance     0.01  0.01  0.09  0.00  0.16  0.02  0.01  0.11  0.03   0.01   0.05   0.00   0.02
day         0.02  0.01  0.23  0.03  0.28  0.01  0.05  0.05  0.01   0.01   0.00   0.02   0.00
duration    0.00  0.01  0.01  0.01  0.01  0.06  0.00  0.00  0.01   0.00   0.27   0.11   0.00
campaign    0.03  0.01  0.03  0.05  0.00  0.27  0.01  0.00  0.00   0.02   0.03   0.00   0.00
job         0.33  0.47  0.25  0.55  0.13  0.37  0.34  0.22  0.37   0.37   0.24   0.30   0.42
marital     0.00  0.38  0.05  0.05  0.05  0.09  0.04  0.06  0.01   0.08   0.01   0.06   0.12
education   0.25  0.30  0.27  0.29  0.08  0.08  0.16  0.11  0.32   0.02   0.10   0.01   0.03
default     0.00  0.00  0.01  0.02  0.08  0.00  0.01  0.19  0.01   0.00   0.01   0.00   0.01
housing     0.33  0.07  0.00  0.11  0.03  0.01  0.05  0.02  0.02   0.01   0.01   0.00   0.00
loan        0.00  0.00  0.15  0.01  0.06  0.05  0.07  0.06  0.00   0.00   0.03   0.03   0.01
contact     0.55  0.10  0.09  0.07  0.03  0.00  0.13  0.05  0.02   0.17   0.05   0.02   0.06
month       0.70  0.13  0.48  0.31  0.47  0.37  0.43  0.37  0.35   0.45   0.32   0.55   0.39
```

Because we wish to maintain the maximum amount of variability in our dataset to perform high end analytics, we will select 25. Now we look at the squared loading matrix and keep variables with an |r| above .5, so we will select the variables with a squared r above .25. After this we ended up keeping the following variables: age, day, duration, campaign, job, marital, education, housing, contact, month. We dropped the variables balance, default, and loan.

Algorithm Results

We are going to run through four varying algorithms here, but we will not be implementing the kNN classification algorithm. The reason for this is that it has some limitations which are not suitable for the data in this project. The kNN algorithm works for numerical data, which some of our data is not. Also it is a lazy learner, meaning the model is not used for prediction, and the predictions are made from local information only.

Algorithm	Accuracy	Sensitivity	Specificity	Precision	AUC of ROC
Logistic Regression	72.93%	75.76%	72.77%	13.59%	0.743
Decision Tree	76.28%	81.82%	64.04%	11.39%	0.733
Naïve Bayes	74.23%	81.82%	73.80%	15%	0.778
Bagging	80.71%	54.55%	83.90%	16.19%	0.701



Here the bagging algorithm seems to give us the highest amount of accuracy across all the algorithms, but there does seem to be some off putting sensitivity in predicting true positives. The algorithm with the second highest accuracy, Decision Tree, also has a lower specificity percentage. The other two even out in that regard, but do not have the best accuracy percentages. Also considering the precision and AUC of ROC numbers, Bagging looks as if it would be the best choice in this scenario.