MSDS 422 Assignment 3

In order to predict the response of a bank marketing promotion, classification models were employed and evaluated using the area under the receiver operating characteristic (ROC) curve as an index of classification performance. Data was collected for each of the 4521 phone calls, which included the following 17 attributes:

Table C.2. Bank Marketing Study Variables

Variable Name	*Description (Possible Values)*
	Client Demographics
age	Age in years
job	Type of job (admin., unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services)
marital	Marital status (married, divorced, single) [Note: "divorced" means divorced or widowed]
education	Level of education (unknown, secondary, primary, tertiary)
	Client Banking History
default	Has credit in default? (yes, no)
balance	Average yearly balance (in Euros)
housing	Has housing loan? (yes, no)
loan	Has personal loan? (yes, no)
	Data from Most Recent Marketing Contact/Call
contact	Contact communication type (unknown, telephone, cellular)
day	Last contact day of the month
month	Last contact month of year (jan, feb, mar, . . ., nov, dec)
duration	Last contact duration (in seconds)
	Data from All Marketing Contacts/Calls
campaign	Number of contacts performed during this campaign for this client (includes last contact)
pdays	Number of days that passed since the client was last contacted from a previous campaign (-1 means client was not previously contacted)
previous	Number of contacts performed before this campaign for this client
poutcome	Outcome of the previous marketing campaign (unknown, other, failure, success)
	Response to Most Recent Marketing Contact/Call
response	Has the client subscribed to a term deposit? (yes, no)

The goal is to be able to predict the response of the marketing campaign, so that the bank can fine tune the direction of their campaign. Some preliminary statistics and graphs were

produced in order to better understand the data. Also, those who agreed to the campaign were observed.

Descriptive Statistics of All Data

	age	balance	day	duration	campaign	pdays	previous
count	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000
mean	41.170095	1422.657819	15.915284	263.961292	2.793630	39.766645	0.542579
std	10.576211	3009.638142	8.247667	259.856633	3.109807	100.121124	1.693562
min	19.000000	-3313.000000	1.000000	4.000000	1.000000	-1.000000	0.000000
25%	33.000000	69.000000	9.000000	104.000000	1.000000	-1.000000	0.000000
50%	39.000000	444.000000	16.000000	185.000000	2.000000	-1.000000	0.000000
75%	49.000000	1480.000000	21.000000	329.000000	3.000000	-1.000000	0.000000
max	87.000000	71188.000000	31.000000	3025.000000	50.000000	871.000000	25.000000

Descriptive Statistics of those who said Yes

	age	balance	day	duration	campaign	pdays	previous
count	521.000000	521.000000	521.000000	521.000000	521.000000	521.000000	521.000000
mean	42.491363	1571.955854	15.658349	552.742802	2.266795	68.639155	1.090211
std	13.115772	2444.398956	8.235148	390.325805	2.092071	121.963063	2.055368
min	19.000000	-1206.000000	1.000000	30.000000	1.000000	-1.000000	0.000000
25%	32.000000	171.000000	9.000000	260.000000	1.000000	-1.000000	0.000000
50%	40.000000	710.000000	15.000000	442.000000	2.000000	-1.000000	0.000000
75%	50.000000	2160.000000	22.000000	755.000000	3.000000	98.000000	2.000000
max	87.000000	26965.000000	31.000000	2769.000000	24.000000	804.000000	14.000000

```
****Job distibution:
management       969
blue-collar      946
technician       768
admin.           478
services         417
retired          230
self-employed    183
entrepreneur     168
unemployed       128
housemaid        112
student           84
unknown           38
Name: job, dtype: int64

****Marital status distibution:
married     2797
single      1196
divorced     528
Name: marital, dtype: int64

****Education distibution :
secondary    2306
tertiary     1350
primary       678
unknown       187
Name: education, dtype: int64

****Default distibution:
no     4445
yes      76
Name: default, dtype: int64

****Housing distibution :
yes    2559
no     1962
Name: housing, dtype: int64

****Loan distibution:
no     3830
yes     691
Name: loan, dtype: int64

****Contact distibution:
cellular     2896
unknown      1324
telephone     301
Name: contact, dtype: int64

****Month distibution :
may    1398
jul     706
aug     633
jun     531
nov     389
apr     293
feb     222
jan     148
oct      80
sep      52
mar      49
dec      20
```

```
****Job distibution for those who responded:
management       131
technician        83
blue-collar       69
admin.            58
retired           54
services          38
self-employed     20
student           19
entrepreneur      15
housemaid         14
unemployed        13
unknown            7
Name: job, dtype: int64

****Marital status distibution for those who responded
married     277
single      167
divorced     77
Name: marital, dtype: int64

****Education distibution for those who responded:
secondary    245
tertiary     193
primary       64
unknown       19
Name: education, dtype: int64

****Default distibution for those who responded:
no     512
yes      9
Name: default, dtype: int64

****Housing distibution for those who responded:
no     301
yes    220
Name: housing, dtype: int64

****Loan distibution for those who responded:
no     478
yes     43
Name: loan, dtype: int64

****Contact distibution for those who responded:
cellular     416
unknown       61
telephone     44
Name: contact, dtype: int64

****Month distibution for those who responded:
may    93
aug    79
jul    61
apr    56
jun    55
nov    39
feb    38
oct    37
mar    21
sep    17
jan    16
dec     9
```

```
****Job percentage for those who responded:
admin.           12.133891
blue-collar       7.293869
entrepreneur      8.928571
housemaid        12.500000
management       13.519092
retired          23.478261
self-employed    10.928962
services          9.112710
student          22.619048
technician       10.807292
unemployed       10.156250
unknown          18.421053
Name: job, dtype: float64

****Marital status percentage for those who responded:
married      9.903468
single      13.963211
divorced    14.583333
Name: marital, dtype: float64

****Education percentage for those who responded:
secondary    10.624458
tertiary     14.296296
primary       9.439528
unknown      10.160428
Name: education, dtype: float64

****Default percentage for those who responded:
no     11.518560
yes    11.842105
Name: default, dtype: float64

****Housing percentage for those who responded:
no     15.341488
yes     8.597108
Name: housing, dtype: float64

****Loan percentage for those who responded:
no     12.480418
yes     6.222865
Name: loan, dtype: float64

****Contact percentage for those who responded:
cellular     14.364641
unknown       4.607251
telephone    14.617940
Name: contact, dtype: float64

****Month percentage for those who responded:
apr    19.112628
aug    12.480253
dec    45.000000
feb    17.117117
jan    10.810811
jul     8.640227
jun    10.357815
mar    42.857143
may     6.652361
nov    10.025707
oct    46.250000
sep    32.692308
```

Note: Data visualization can be found in attached code PDF

As one can notice, 521 of the clients agreed to the promotion. If one were to describe the average person who said “yes” to the campaign based on the highest percentage of each attribute,

he/she would be about 29-55 years old, have a balance around 1500 euros, be a student or retired, and would not have a personal or housing loan. Furthermore, the best method would be to have a phone call duration around 550 seconds and to call two times with a 68 day gap from the previous campaign. However, due to the large standard deviations and wide range of responses, one can recognize that this is not the best way for the bank to base their marketing campaign off of.

Next, classification techniques were used to predict whether the response would result in a 'yes' or 'no.' The two method used were Logistic Regression and naïve Bayes classification, and were analyzed using the area under the receiver operating characteristic (ROC) curve. The explanatory variables of housing, loan and default would be used in order to predict the response. In order to complete this, training and test data sets were made by allocating 1 data point to the data set for every 9 training data points (1:10 ratio). Furthermore, 10 trials were taken with different test and training data sets. The results are as follows:

```
********************************
Average from 10 folds
Method                 Area under ROC Curve
Logistic_Regression    0.611733
Naive_Bayes            0.611060
dtype: float64

Standard Deviation
Logistic_Regression    0.052946
Naive_Bayes            0.053606
dtype: float64
```

As one can see, after averaging the results of the 10 trials Logistic Regression has a slightly higher area under the ROC curve with lower standard deviation. This indicates that it will be better at predicting the response with less variance.

In conclusion when it comes to management, I would recommend Logistic Regression. However, naïve Bayes classification produces very similar results, so I would not disregard it completely. Therefore, I believe it is important to gather more data in order to determine whether these two models can be further distinguished. Furthermore, I would look more into clients who are retired and those who are students. Although, those in management have the highest sum, retired clients and students had larger response rates.