INTRODUCTION

A Portuguese bank conducted seventeen telephone marketing campaigns between May 2008 and November 2010. The bank recorded client contacts information for each telephone call. Client characteristics include demographic factors: age, job type, marital status, and education. The client's previous use of banking services is also noted. Current contact information shows the date of the telephone call and the duration of the call. There is also information about the call immediately preceding the current call, as well as summary information about all calls with the client.

SUMMARY AND PROBLEM DEFINITION FOR MANAGEMENT

The bank wants its clients to invest in term deposits. A term deposit is an investment such as a certificate of deposit. The interest rate and duration of the deposit are set in advance. A term deposit is distinct from a demand deposit. The bank is interested in identifying factors that affect client responses to new term deposit offerings, which are the focus of the marketing campaigns. What kinds of clients are most likely to subscribe to new term deposits? What marketing approaches are most effective in encouraging clients to subscribe.

MEASUREMENT AND STATISTICAL METHODS

The data used in this evaluation was obtained from a previous telephone marketing campaign.

The dataset consists of 4521 respondent's answers (rows) to 17 marketing questions (columns).

```
dataset dimensions (4521, 17)
```

Non-values were dropped if present in data. However, removal has not change the dataset dimensions.

```
dataset shape after dropna (4521, 17)
```

Model data consists of three explanatory variables -**default, housing, and loan**, were used and one **response** variable

The response is binary variable and the research two classification models – Naïve Bias and Logistic Regression have been utilized for the training and prediction.

Linear logistic regression is solved by maximizing the conditional likelihood of G given X: Pr(G = k | X = x), while LDA maximizes the joint likelihood of G and X: Pr(X = x, G = k).

Naive Bias is very simple, easy to implement and fast. If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression. Even if the NB assumption doesn't hold, it works great in practice. Need less training data. Highly scalable. It scales linearly with the number of predictors and data points. Can be used for both binary and mult-iclass classification problems. Can make probabilistic predictions. Handles continuous and discrete data.

EXPLORATORY DATA ANALYSIS METHODS

Dimensions

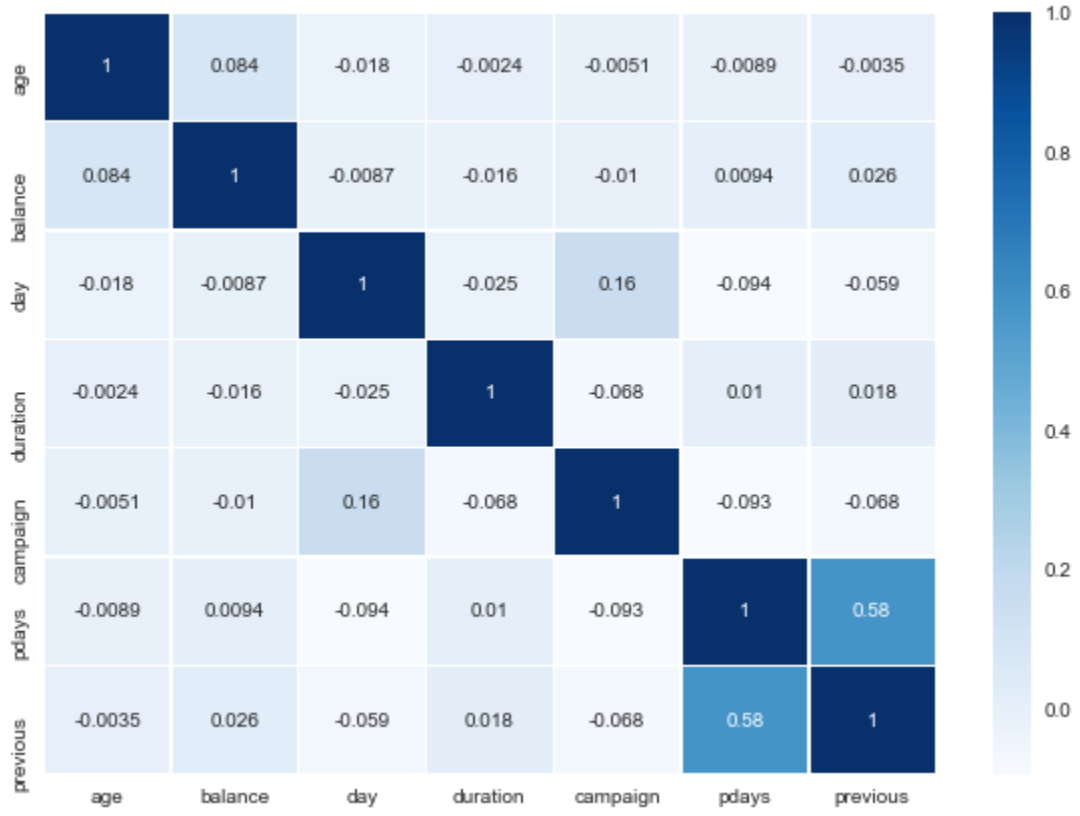
The dataset consists of 4521 respondent's answers (rows) to 17 marketing questions (columns).

```
dataset dimensions (4521, 17)
```

Dataset Basic Descriptive Statistics

	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

Correlation



Transformation

In order to apply classification model, the categorical variable should be converted to their binary representation

```
1.  bank['response_ind'] = le.fit_transform(bank['response'].astype(str))
2.  bank['default_ind'] = le.fit_transform(bank['default'].astype(str))
3.  bank['loan_ind'] = le.fit_transform(bank['loan'].astype(str))
4.  bank['housing_ind'] = le.fit_transform(bank['housing'].astype(str))
```

Double check that response, default, loan, housing are binary

```
1.  response (array([0, 0, 0, ..., 0, 0, 0], dtype=int64), array(['no', 'yes'], dtype=object))
2.  default (array([0, 0, 0, ..., 0, 0, 0], dtype=int64), array(['no', 'yes'], dtype=object))
3.  loan (array([0, 1, 0, ..., 0, 0, 1], dtype=int64), array(['no', 'yes'], dtype=object))
4.  housing (array([0, 1, 1, ..., 0, 0, 1], dtype=int64), array(['no', 'yes'], dtype=object))
```

Head of the research dataset

	default_ind	loan_ind	housing_ind
4383	0	0	0
502	0	0	0
4052	0	0	1
3634	0	1	1
3088	0	0	

Overview of Programming Work

Python was used exclusively for the analysis of the telephone direct marketing data, including the use of the following packages: Pandas and Numpy for data handling, and Scikit Learn for machine learning and model evaluation metrics. Telephone direct marketing campaign data was presented to us as a CSV file and loaded into the program using Pandas. K fold cross validation design, with ten folds using the AUROC as index for classification performance, was used for both logistic regression and naives Bayes classification machine learning models and was implemented using the Python Scikit Learn environment.

Results and Recommendations

	Logistics Regression	Naïves Bias
score	0.886891	0.879977876
Accuracy	0.876243	0.8629834
Training accuracy	87.6243 %	86.2983 %
Training AUC	58.4517 %	0.5001 %

Confusion matrix	[[793 0] [112 0]]	[[779 14] [110 2]]

Examining the average AUROC for the Logistic Regression classification method 58.4%, and the Naives Bayes classification method, 50%, the Logistic Regression model performs better for predicting customers that will participate in term deposits when using three explanatory variables, default, loan, and housing. Those most likely to participate also have no defaults. The recommendation is the Logistic Regression method to direct t marketing campaigns towards those with no defaults.