```
---
title: "Capstone Project Overview"
date: November 5, 2019"
output:
  pdf_document: default

---
```

INDEX

 # 1. Introduction

 # 2.Executive Summary Section

*# 2.1 **Dataset:*

*# 2.2 **Goal of the project*

*# 2.3 **Key steps that were performed*

 # 3. Methods/analysis section

*# 3.1 **Process and techniques used,*

*# 3.2 **Data cleaning*

*# 3.3 **Data exploration*

*# 3.4 **Visualization and insights gained*

*# 3.5 **Modeling approach*

*# 4. **Results section ****

*# 4.1 **Modeling results and discusses*

*# 4.2 **Model performance*

 # 5. Conclusion section

*# 5.1 **Summary of the report*

*# 5.2 **Limitations*

*# 5.3 **Future work*

1.Introduction

```
Hello, fellow students.
I am submitting my work for your evaluation.
I hope the work could be clear enough for you to understand it without
effort, at least, I'll do my best to get this objective.

As it is said by the staff,
```

"The ability to clearly communicate the process and insights gained from an analysis is an important skill for data scientists.

The project itself is based on the postulates that were given to us at the capstone module of the Data Science Professional Certificate.
My submission for this project is three files:

    A report in PDF format: Report.pdf
    A report in Rmd format: Report.Rmd
    A script in R format: Script.R

For this project, I have chosen the public use Bank Marketing Dataset.
It is available at the UCI site
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#
Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.
The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Set Characteristics: Multivariate

Number of Instances: 41189

Area: Business

Attribute Characteristics: Real

Number of Attributes: 20

Missing Values? N/A

Number of Web Hits: 1018726

Files are expected to be downloaded from the uploaded files using the edx submision form.
As the staff recommend also providing a link to a GitHub repository containing the three files above, I also include links to a repository.

If one or more of these files are damaged or not available,an alternative way to download them is:
All the files are in the public access repository:

https://github.com/ubatifce/capstone

or

https://github.com/ubatifce/capstone/blob/master/Report.Rmd

https://github.com/ubatifce/capstone/blob/master/Report.pdf

https://github.com/ubatifce/capstone/blob/master/Script.R

https://github.com/ubatifce/capstone/blob/master/bank-additional-full.csv

Clone with HTTPS: Use Git or checkout with SVN using the web URL.

https://github.com/ubatifce/capstone.git

#2. Executive Summary Section

Dataset:

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution.
The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required,
in order to access if the product (bank term deposit) would be (or not) subscribed.

The dataset is:
1) bank-additional-full.csv with all examples (41188 rows) and 20 inputs, ordered by date (from May 2008 to November 2010)

Attribute Information:

Input variables:

bank client data:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - 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)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

 # social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

Missing Attribute Values:

```
There are several missing values in some categorical attributes, all
coded with the "unknown" label.
These missing values can be treated as a possible class label or using
deletion or imputation techniques.
```

The data is enriched by the addition of five new social and economic features/attributes (national wide indicators from a ~10M population country),
published by the Banco de Portugal and publicly available at: https://www.bportugal.pt/estatisticasweb.
This dataset is almost identical to the one used in [Moro et al., 2014] (it does not include all attributes due to privacy concerns).

Goal of the project

The classification goal is to Predict the Success of Bank Telemarketing, that is to predict if the client will subscribe (yes/no) a term deposit (variable y).
One of the goals is to understand the **contact** media that was used, that is to say the use of telephone calls (fix line calls) , agains cellular calls (mobile line calls).
(Do people react better when the fix line is used?)
Education level is another field to understand, in special.
(Is there a trend correlatd with the education level of clients?)
Combined use of calls and education is important to see if some education levels prefer fix or mobile contacts.

Key steps that were performed

The first step performed was to create a sub set for train , called train_set
The second step performed was to create a sub set for test, called train_set

#3. Methods/analysis section

Process and techniques used,

Being **contact** a column that shows attemps on telephone calls (fix line calls) , agains cellular calls (mobile line calls)

1 Baseline: calculated as a sample 0,1. mean(y_hat ==test_set$y)

2 clients who subscribed a term deposit by contact

3 clients who subscribed a term deposit using contact,education

4 clients who subscribed a term deposit using LDA

5 clients who subscribed a term deposit using QDA

6 clients who subscribed a term deposit using GLM y ~ contact+education+job+age

Data cleaning

The first data cleaning/wrangling was to convert the column y
from "Has the client subscribed a term deposit yes/no?"
to "Has the client subscribed a term deposit 1/0"
bank <- mutate(y = ifelse(y=="yes"), 1, 0))
***I could used "mutate" or a direct replacement. The second method was used.
bank$Y<- 1
bank$Y[bank$y=="no"]<- 0 # ("no"=0, so now only "yes"=1). No NA's found in the set.

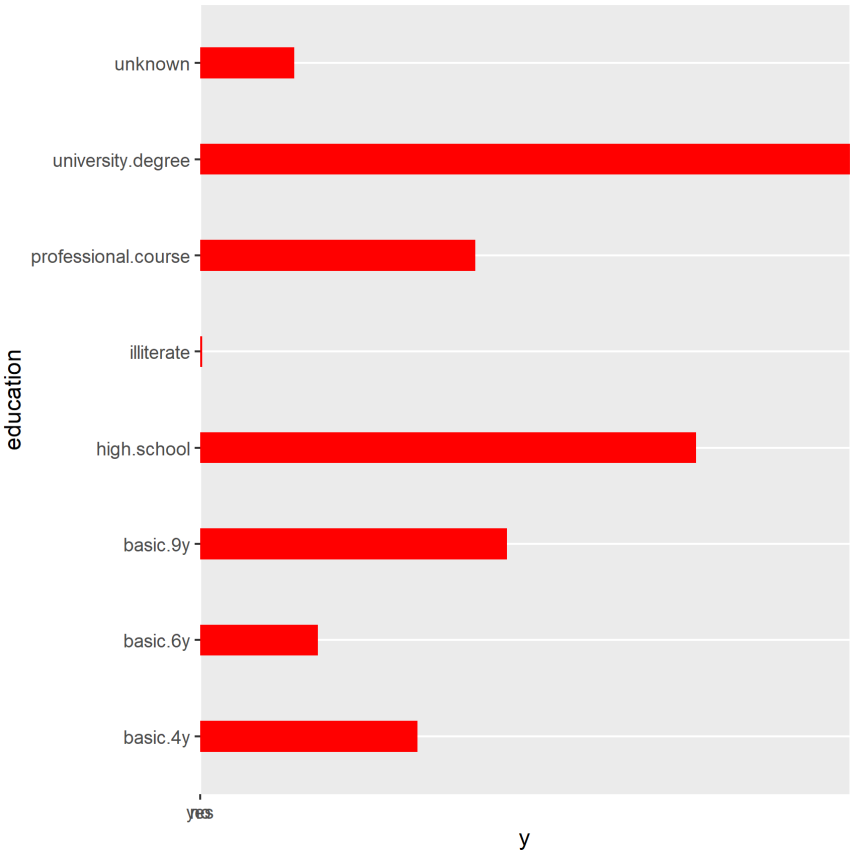
Data exploration

the objective is then to predict:

*y <- bank$Y

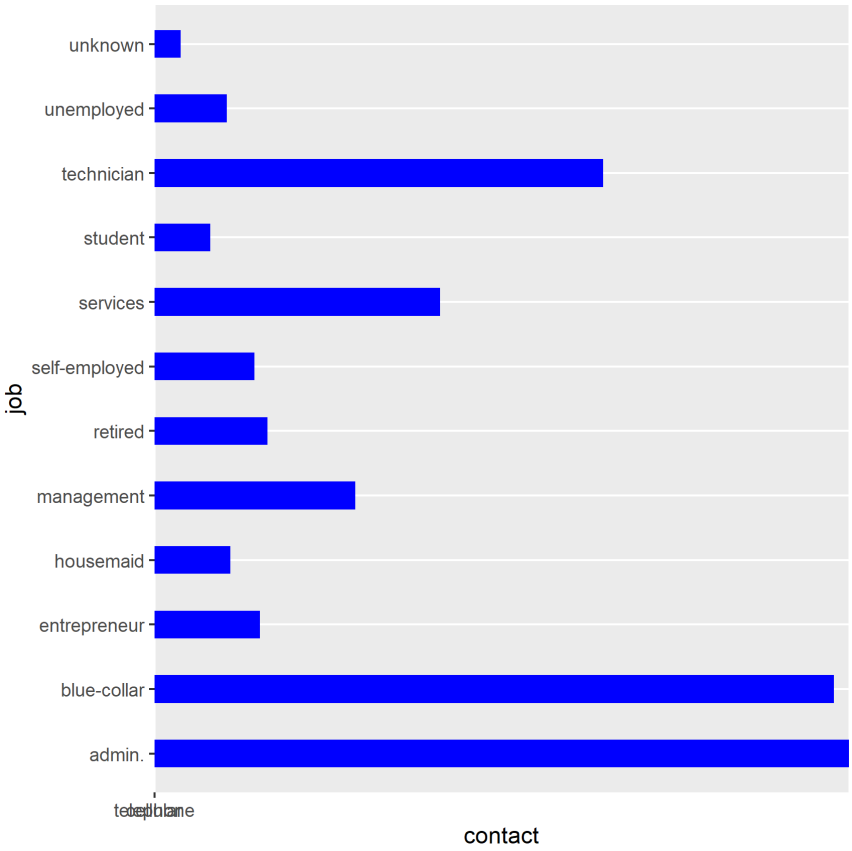
**Visualization and insights gained
I will be using knitr::kable() to visualize the results

```{r}
 bank %>% ggplot(aes(education,contact)) +
   geom_bar(width = 0.4, stat = "identity", color = "yellow") +
   coord_flip()
```



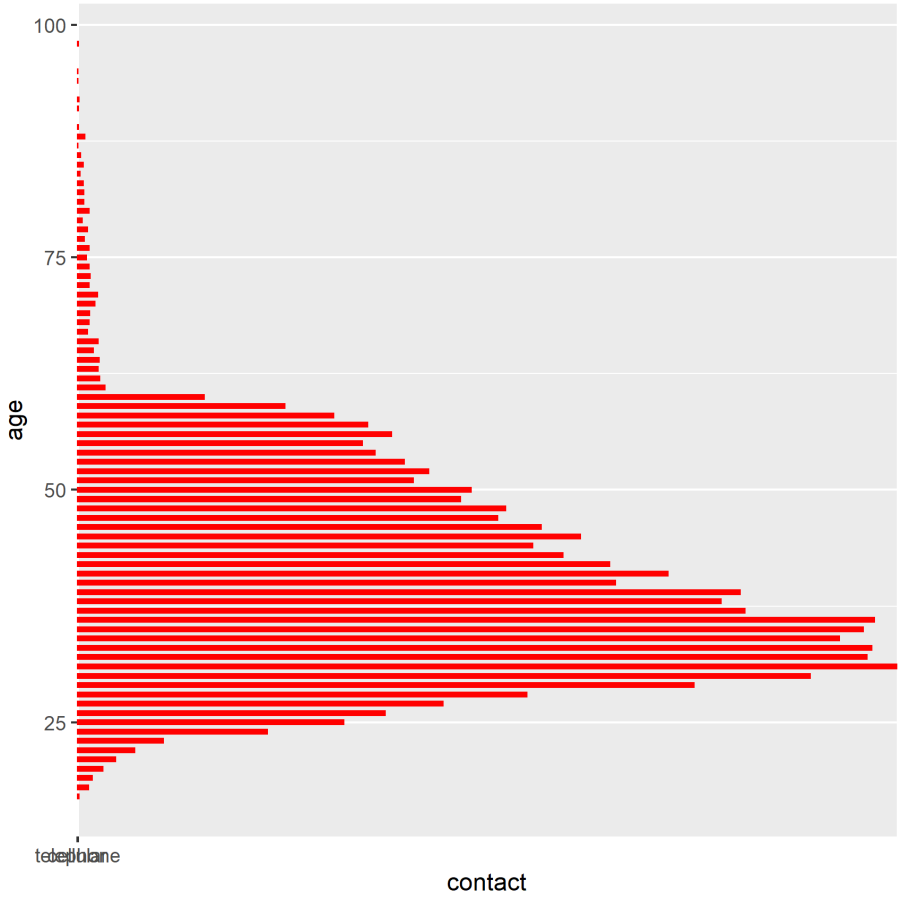
We see that many contacts were made to "universitary degree" or "high school" education

```{r}
graph<- bank %>% ggplot(aes(job,contact)) +
   geom_bar(width = 0.4, stat = "identity", color = "blue") +
   coord_flip()
```



On the other side, we see that many contacts were made to people who worked as "blue-collars" or "admins"

```{r}
 bank %>% ggplot(aes(age,contact)) +
   geom_bar(width = 0.4, stat = "identity", color = "red") +
   coord_flip()
 ```



Finally, we see that many contacts were made to people in their 30's (30-40 yrs)

training and test sets names

In order not to use validation set for training, I created to experiment:

```{r}
set.seed(42, sample.kind="Rounding")
test_index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)
train_set <- bank %>% slice(-test_index)
```

baseline model

The baseline model was to create a sample with the values 0-1, meanning suscribe/not suscribe.

#Baseline: A sample 0,1

```
y_hat <- sample(c(0,1),nrow(test_set),replace=TRUE)
#Baseline Accuracy
mean(y_hat ==test_set$y)
```

Models/calculations used

#Accuracy

method	Accuracy
Baseline: A sample 0,1	0.4956300
clients who subscribed a term deposit using contact	0.5623938
clients who subscribed a term deposit using contact,education	0.7808934
clients who subscribed a term deposit using LDA	0.8882010
clients who subscribed a term deposit using QDA	0.8402525
clients who subscribed a term deposit using GLM y ~ contact+education+job+age	0.8946346

4. Results section

4.1 Modeling results and discusses

I was able to improve prediction from a **0.4956300** of the baseline model to 0.8946346
The selected Accuracy was then **0.8946346**

The model chosen was GLM (y ~ contact+education+job+age)

```
 y_hat <- predict(train_lda, test_set)
 # Accuracy
 mean(y_hat ==test_set$y) 0.8946346
```

At the bottom of this report, the algorithm is exposed.

4.2 Model performance

In this case, the method GLM (y ~ contact+education+job+age) seemd to be the best alternative.
set.seed(42, sample.kind="Rounding") was chosen as standard seed for the project.

5. Conclusion section

5.1 Summary of the report

The goal of the project , in the sense of predict if the client will subscribe a term deposit (variable y) and print the values was reached.
As it was pointed earlier, The final Accuracy calculated was 0.8946346
The model chosen was Using **GLM (y ~ contact+education+job+age)

5.2 Limitations

These models are very difficult to work when you are a student and you do not have access to the ultimates PC.
Using PCs with low memory ans low processors is possible, but consumes a lot of time.

5.3 Future work

This is a very nice dataset to work with.
It is very important to realize what factors are important while you want to persuade customers to pay for your service.
I'd rather investigate further after finishing my course.

Algorithm. Function used:

```{r}
################################
# INIT LOADING DATASETS. SKIP IF DATA IS ALREADY LOADED
################################

if(!require(tidyverse)) install.packages("tidyverse", repos =
"http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-
project.org")
if(!require(data.table)) install.packages("data.table", repos =
"http://cran.us.r-project.org")
if(!require(dslabs)) install.packages("dslabs", repos =
"http://cran.us.r-project.org")
if(!require(readxl)) install.packages("dslabs", repos =
"http://cran.us.r-project.org")

#Set default directory to C:/ or change the path in setwd(...)
setwd("C:/")
bank<-read.csv2("bank-additional-full.csv")
 ncol(bank)
 nrow(bank)

# clean the data


#From "Has the client subscribed a term deposit yes/no?"
#To      "Has the client subscribed a term deposit 1/0"

  bank$Y<-  1
  bank$Y[bank$y=="no"]<-  0
```

```
#Has the client subscribed a term deposit
y <- bank$Y

#Set the seed to 42, then using the caret package to create a 20% data
partition based on the "y" data.

#Assign the 20% partition to test_set and the remaining 80% partition to
train_set.

set.seed(42, sample.kind="Rounding")
test_index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)
train_set <- bank %>% slice(-test_index)
test_set <- bank %>% slice(test_index)
print("Train SET")
nrow(train_set)
ncol(train_set)
print("test SET")
nrow(test_set)
ncol(test_set)


# Proportion of clients who subscribed a term deposit

testy<-subset(test_set,test_set$y=="yes")

nrow(testy)/nrow(test_set)


#Baseline: Using a  sample 0,1
y_hat <- sample(c(0,1),nrow(test_set),replace=TRUE)
#Baseline Accuracy
mean(y_hat ==test_set$Y)

results <- data_frame(method="Baseline: A sample 0,1 ", Accuracy =
mean(y_hat ==test_set$Y)
)

#Proportion of clients who subscribed a term deposit contacted by type of
phone

train_set  %>% filter(contact=="cellular") %>% summarize(mean(Y==1))
train_set  %>% filter(contact=="telephone") %>% summarize(mean(Y==1))

#Predict clients who subscribed a term deposit  using contact on the test
set:
# Accuracy of this contact-based prediction method on the test set
y_hat <- if_else(test_set$contact=="telephone",1,0)
# Accuracy
mean(y_hat ==test_set$Y)

results <- bind_rows(results,
                          data_frame(method="clients who subscribed a
term deposit  using contact ",
                           Accuracy = mean(y_hat ==test_set$Y)
))
```

```
#Predicting clients who subscribed a term deposit  by contact & education
#Accuracy of this contact & education-based prediction method on the test
set



# Accuracy

train_set  %>% group_by( contact,education) %>% summarize(mean(Y==1))
y_hat <- if_else(test_set$contact=="telephone" & test_set$education %in%
c("professional.course","university.degree") ,1,0)
mean(y_hat ==test_set$Y)

# see a graph
graph<-bank %>% ggplot(aes(education,contact)) +
   geom_bar(width = 0.4, stat = "identity", color = "green") +
   coord_flip()
graph


results <- bind_rows(results,
          data_frame(method="clients who subscribed a term deposit  using
contact,education ",
           Accuracy = mean(y_hat ==test_set$Y)
))

#Confusion matrices for the contact  model, education model, and combined
contact & #education model.

# see a graph
graph<-bank %>% ggplot(aes(age,contact)) +
   geom_bar(width = 0.4, stat = "identity", color = "red") +
   coord_flip()
graph


set.seed(42, sample.kind="Rounding")


#Models:

#Training  a model using linear discriminant analysis (LDA) with the
caret lda method using job as the only predictor.
#LDA y ~ contact+education+job+age
train_lda <- train(y ~ contact+education+job+age, method = "lda", data =
train_set)
y_hat <- predict(train_lda, test_set)
# Accuracy

cm<-confusionMatrix(data = y_hat, reference =
test_set$y)$overall["Accuracy"]
```

```
cm

results <- bind_rows(results,
       data_frame(method="clients who subscribed a term deposit  using
LDA ",
        Accuracy =cm)
)



#Training  using quadratic discriminant analysis (QDA) with the caret qda
method using job as the only predictor.
#QDA y ~ contact+education+job+age
train_qda <- train(y ~  contact+education+job+age, method = "qda", data =
train_set)
y_hat <- predict(train_qda, test_set)
# Accuracy

cm<-confusionMatrix(data = y_hat, reference =
test_set$y)$overall["Accuracy"]
cm

results <- bind_rows(results,
       data_frame(method="clients who subscribed a term deposit  using
QDA ",
        Accuracy = cm)
)



#GLM y ~ contact+education+job+age

# see a graph
graph<-bank %>% ggplot(aes(job,contact)) +
   geom_bar(width = 0.4, stat = "identity", color = "blue") +
   coord_flip()
graph



train_glm <- train(y ~ contact+education+job+age, method = "glm", data =
train_set)
y_hat <- predict(train_glm, test_set)
# Accuracy
mean(y_hat ==test_set$y)
cm<-confusionMatrix(data = y_hat, reference =
test_set$y)$overall["Accuracy"]
cm

results <- bind_rows(results,
data_frame(method="clients who subscribed a term deposit  using GLM Y ~
contact+education+job+age",
        Accuracy = cm)
)
```

```
results %>% knitr::kable()
max(results[2])
```

```
```

**Citation Request:

The dataset is public available for research.
Please include this citation if you plan to use this database:
[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014