## DESCRIPTION

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

## SUMMARY

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 ('yes') or not ('no') subscribed.

There is one dataset:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

## MAIN OBJECTIVE OF THIS ANALYSIS:

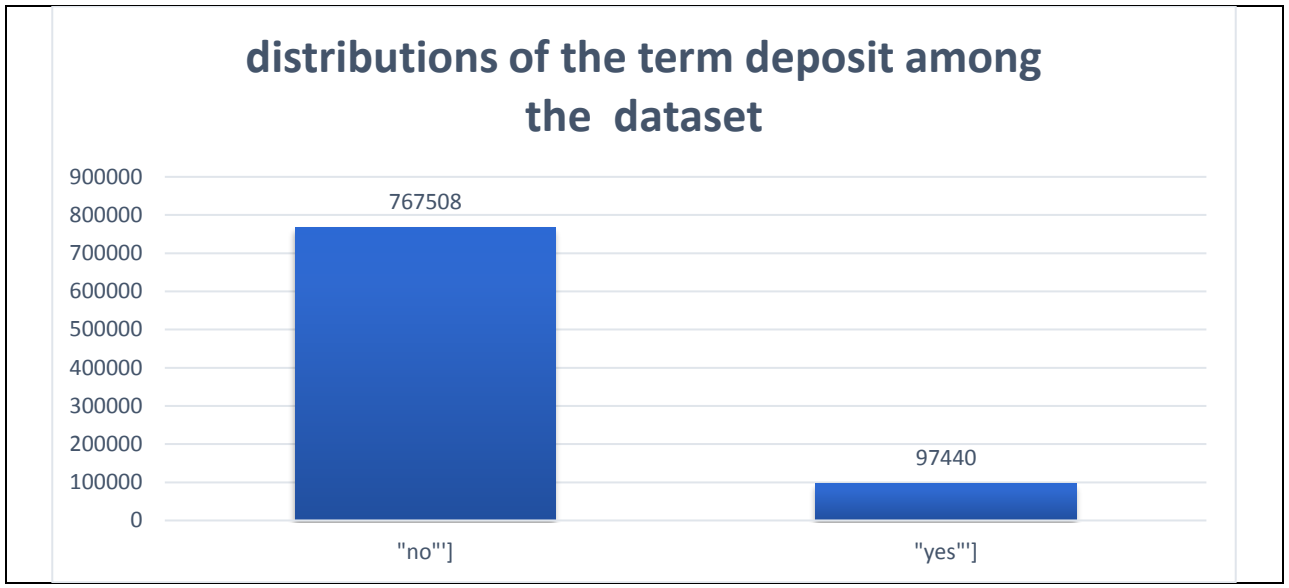
The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). through these main stages:

- Gathering data
- Preparing data
- Choosing a model
- Training the model
- Evaluation the model
- Hyper parameter Tuning
- Prediction and verifications

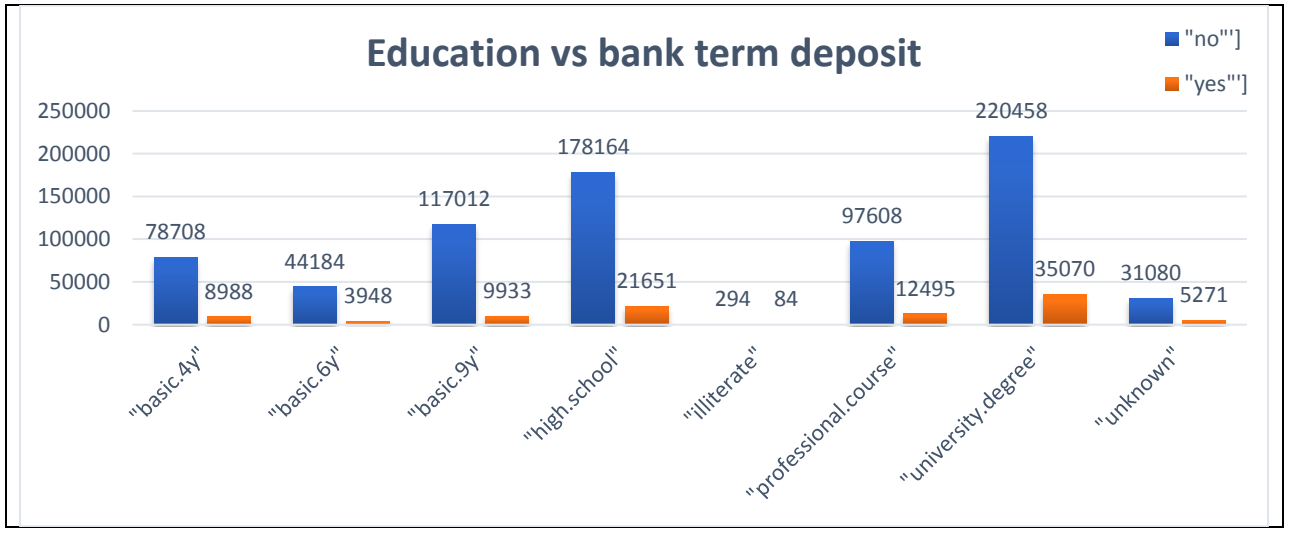
Additionally,  new features may be used to enhance the accuracy of the model

For example:

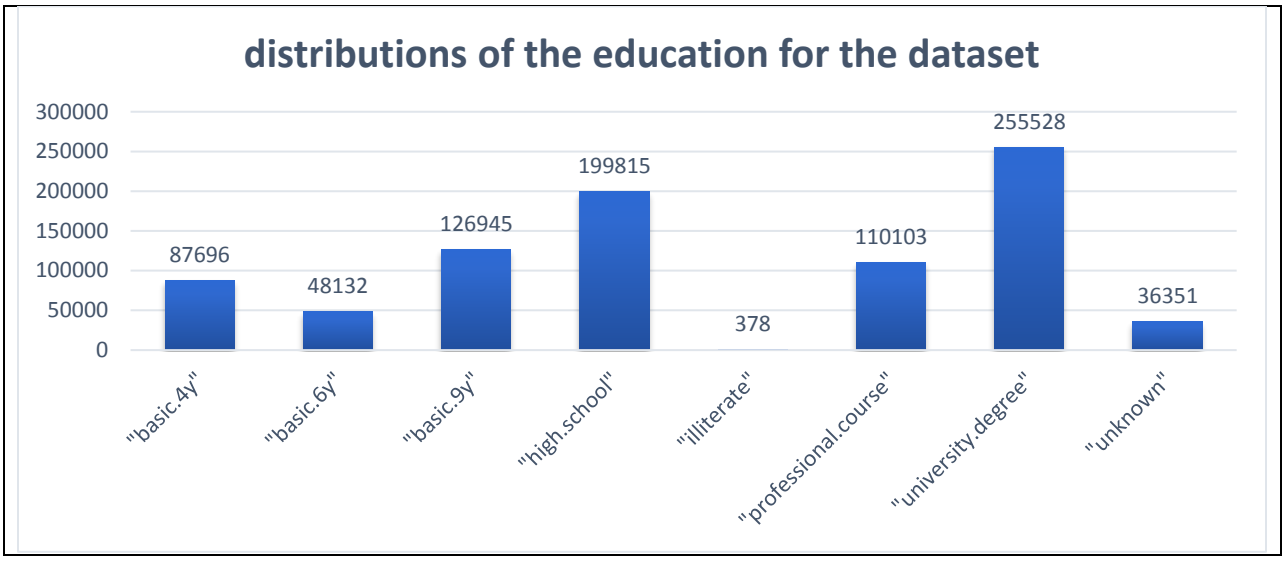
The chart below illustrates the distributions of the term deposit among the whole dataset:




The charts below show the relationship between the relationship between education and “term deposit”




The charts below show the relationship between the relationship between education and “term deposit”




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')

Source: UCI

http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Analysis summary:

The whole data set is been splitting into 3 categories as below:

- 70% training data
- 15% cross validation
- 15% testing data

1- Using Logistic Regression:

A logistic regression classifier is been used on data training and the results were as below:

- Accuracy on training set: 0.9088701845257365
- Accuracy on Cross validation set: 0.9058684794672586
- Accuracy on testing set: 0.9137069922308546

2- Using Decision Tree Classifier:
   a. Accuracy on training set: 1.0
   b. Accuracy on Cross validation set: 0.8818673695893452
   c. Accuracy on testing set: 0.8900527192008879

3- K Neighbors Classifier using n_neighbors=5:
   a. Accuracy on training set: 0.928617675623179
   b. Accuracy on Cross validation set: 0.9026775804661488
   c. Accuracy on testing set: 0.9089900110987791

4- KNeighborsClassifier using n_neighbors=6:
   a. Accuracy on training set: 0.9234380058271285
   b. Accuracy on Cross validation set: 0.9037180910099889
   c. Accuracy on testing set: 0.9097530521642619

5- KNeighborsClassifier using n_neighbors=7:
   a. Accuracy on training set: 0.9248947879572678
   b. Accuracy on Cross validation set: 0.9037874583795783
   c. Accuracy on testing set: 0.910169256381798

6- KNeighborsClassifier using n_neighbors=8:
   a. Accuracy on training set: 0.9209291032696666
   b. Accuracy on Cross validation set: 0.9044117647058824
   c. Accuracy on testing set: 0.9104467258601554

7- KNeighborsClassifier using n_neighbors=9:
    a. Accuracy on training set: 0.9222240207186791
    b. Accuracy on Cross validation set: 0.9033018867924528
    c. Accuracy on testing set: 0.9118340732519423

Although Decision Tree Classifier shows higher accuracy in training set but the accuracy on testing set was the lowest value among the 3 used classifiers.
The best model regarding the accuracy is the logistic regression classifier

Also, for the next steps, some techniques may be used as feature engineering , combining classifiers together or adding more features to increase the accuracy