A. Attributes Description:

Input variables:

Ai. 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: primary, secondary, tertiary and 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')  
8 - **balance:** Balance of the individual.

Aii. 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:** 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.

Aiii. 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')

Output variable (desired target):  
21 - **y** - has the client subscribed a term deposit? (binary: 'yes','no')

정기예금을 가진 사람들 중에서

직업에 따라 자산의 balance를 조사

교육수준에 따라 나누어도 divoced그룹의 mean balance가 다른 그룹에 비해 현저히 낮다

빚이 있냐 없냐에 따라서 이혼,결혼에 영향을 미치고 동시에 balance에도 차이가 존재

(loan이 없는 그룹일수록 결혼 상태이며, balance값이 크다)

## Classification Model:

1. Stratified sampling

빚을 가지고 있느냐 없느냐에 따라 정기예금을 할지 말지에 영향을 미침

* Correlation matrix를 확인해보면 빚의 유무와 정기예금을 할지 말지의 상관관계가 -0.11이라는 것을 확인할 수 있음 (빚이 있을수록 정기예금을 하지 않는 경향이 있음)
* Stratified sampling을 하기 전에 어떠한 조치를 취해야 한다

1. 데이터 분포를 살펴보면, 빚이 없는 경우(“no”)가 87%, 있는 경우(“yes)가 13%임
2. regression, classifiaction을 수행하기 전에,

train/test set이 “yes” “no”의 비율을 동등하게 50%씩 갖도록 설정해야함

* stratified sampling
* we must make sure that at least the features that have the greatest influence on our label (whether a potential client will open a term deposit or not) is equally distributed

결과값(정기예금을 할지 말지)에 지대한 영향을 미치는 요인들이 동일하게 분포되어있도록 해야한다는 뜻!

1. 각각의 모델에 따른 train\_score와 training\_time 비교
2. Overfitting문제

- overfitting이란? Training set score는 매우 높지만(awesome) Test set score는 awful

- 즉, 학습이 너무 training set에만 초점이 맞춰진 경우

* 데이터의 전반적인 패턴을 파악해야 함
* 2번 결과에 따르면 Decision Tree Classifier와Random Forest classifiers가 overfitting 이라고 할 수 있음(since they both give us nearly perfect scores***)***
* ***따라서!! Overfitting을 피하기 위해서***

***cross validation을 실행한다 !!!***

1. Cross-validation

Taking the training test and splitting it. For instance, if we split it by 3, 2/3 of the data or 66% will be used for training and 1/3 33% will be used or testing and we will do the testing process three times. This algorithm will iterate through all the training and test sets and the main purpose of this is to grab the overall pattern of the data.

* 모델별로 Crossval mean score를 비교

1. Confusion Matrix

* 설정한 모델이 “얼마나 classification을 잘 수행하냐” 를 측정
* train dataset으로 y 값을 cross validation predict
* 이 값(y\_train과 y\_train을 예측한 y\_train\_pred)으로 confusion matrix 구함

1. Precision and Recall

* Trade-off 관계가 존재
* As the precision gets higher the recall gets lower and vice versa. For instance, if we increase the precision from 30% to 60% the model is picking the predictions that the model believes is 60% sure. If there is an instance where the model believes that is 58% likely to be a potential client that will suscribe to a term deposit then the model will classify it as a **"No."** However, that instance was actually a **"Yes"** (potential client did suscribe to a term deposit.) That is why the higher the precision the more likely the model is to miss instances that are actually a **"Yes"!**

\*\* 수많은 변수 중에서 어느 변수가 결과(정기예금 가입)에 영향을 미치는가?

* DecisionTree Classifier

(근데 왜 이걸로 변수를 뽑아내는거지…? 가장 강력한 모델은 GBC인데..)

(그냥 확인 차원에서 한건가….)