# 

# Dataset:

Input variables:

**# bank client data:**

1 - ID : ID of client

2 - age (numeric)

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

4 - marital : marital status (categorical: "divorced","married","single"; note: "divorced" means divorced or widowed)

5 - education (categorical: "primary","secondary","tertiary,"unknown")

6 - default: has credit in default? (categorical: "no","yes")

7 - balance : has money in account? (numeric)

8 - housing: has housing loan? (categorical: "no","yes")

9 - loan: has personal loan? (categorical: "no","yes")

10 - contact: contact communication type (categorical: "cellular","telephone")

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

12 - day: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

13 - 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.

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

15 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)

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

17 - poutcome: outcome of the previous marketing campaign (categorical: "failure","unknown","other","success")

**Output variable (desired target):**

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

(We can't use test.csv because of we don't have y\_test values.)

Train shape: (31647, 18)

Test shape: (13564, 17)

Link: <https://www.kaggle.com/berkinkaplanolu/banking-analyze>

1. Fill unknown values with mode value/ eduRatio(job/ edu degree)
2. Encoding (categorical → numeric)
3. Training (Logistic Regression): confusion matrix and acc result are different cuz imbalance dataset
4. Data Augmentation with SMOTE
5. Training again
6. Logistic Regression
7. Support Vector Classifier
8. KNN
9. Random Forest Classifier

* **Logistic Regression without SMOTE:**

Accuracy: 89.18099089989889

Sensitivity :99.19885550786839

Specificity : 13.232104121475054

* **Logistic Regression with SMOTE**

Accuracy:71.62394386366891

Sensitivity :69.81484139514856

Specificity : 73.42477496785256

* **Support Vector Classifier with SMOTE**

Accuracy:60.75468996133468

Sensitivity :84.29740203818

Specificity : 37.319617088155454

* **K Nearest Neighbors Classifier with SMOTE**

Accuracy:80.99670628669627

Sensitivity :70.76216448973733

Specificity : 91.18445492213174

* **Random Forest Classifier with SMOTE**

Accuracy:87.97794644135757

Sensitivity :88.54600258360844

Specificity : 87.41248749821403