Steps for a Machine learning model

Carlos Montenegro

1. Data Summary: Variables

```
'data.frame':	7000 obs. of  21 variables:
 $ client_id     : int  2 3 4 5 6 7 8 9 14 15 ...
 $ age           : int  29 39 49 32 29 51 34 52 52 29 ...
 $ job           : Factor w/ 12 levels "admin.","blue-collar",..: 4 11 2 7 1 7 2 8 1 1 ...
 $ marital       : Factor w/ 4 levels "divorced","married",..: 3 2 2 3 3 2 2 2 2 3 ...
 $ education     : Factor w/ 8 levels "basic.4y","basic.6y",..: 4 3 2 7 4 7 1 4 7 7 ...
 $ default       : Factor w/ 2 levels "no","unknown": 1 2 2 1 2 2 1 1 1 1 ...
 $ housing       : Factor w/ 3 levels "no","unknown",..: 1 3 1 3 3 3 3 3 3 3 ...
 $ loan          : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 1 1 1 2 1 1 1 1 ...
 $ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 5 8 7 4 5 8 8 8 5 ...
 $ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 2 1 4 2 1 4 4 4 3 2 ...
 $ campaign      : int  3 6 2 3 2 1 1 1 3 1 ...
 $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
 $ previous      : int  0 0 0 1 0 0 0 0 0 0 ...
 $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 1 2 2 2 2 2 2 ...
 $ emp.var.rate  : num  1.1 1.4 -0.1 -1.8 1.4 1.4 -0.1 -0.1 -0.1 -2.9 ...
 $ cons.price.idx: num  94 94.5 93.2 92.9 93.9 ...
 $ cons.conf.idx : num  -36.4 -41.8 -42 -46.2 -42.7 -41.8 -42 -42 -42 -40.8 ...
 $ euribor3m     : num  4.86 4.96 4.15 1.3 4.96 ...
 $ nr.employed   : num  5191 5228 5196 5099 5228 ...
 $ subscribe     : int  0 0 0 0 0 0 0 0 0 0 ...
```

Both data sets, *bank_mkt_train* and *bank_mkt_test* contain the same variables. The first dataset contains 7000 observation and the latter contains 3000. Each of them contain 6 variables that are defined as integers, 10 defined as factors and 5 defined as numeric.

1. Data Summary: Distribution & missings

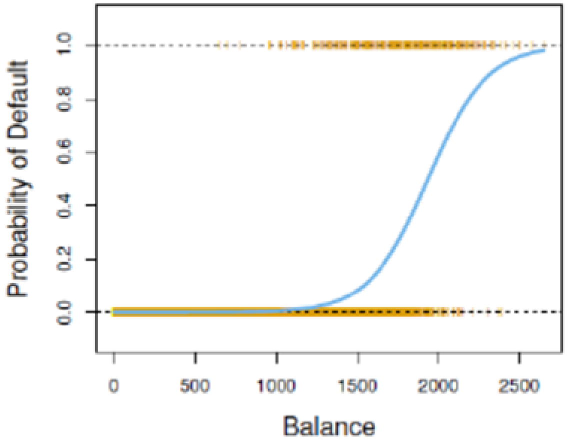
Missing values	
client_id	0
age	0
job	0
marital	0
education	0
default	0
housing	0
loan	0
contact	0
month	0
day_of_week	0
campaign	0
pdays	0
previous	0
poutcome	0
emp.var.rate	0
cons.price.idx	0
cons.conf.idx	0
euribor3m	0
nr.employed	0
subscribe	0

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
client_id	1	7000	5002.60785714	2891.9901535	5018.500	5001.87285714	3725.0325000	2.000	9998.000	9996.000	0.004286773	-1.2005890
age	2	7000	40.34357143	10.6024871	38.000	39.60107143	10.3782000	18.000	98.000	80.000	0.822653018	0.9674292
job*	3	7000	4.74214286	3.5794329	3.000	4.50375000	2.9652000	1.000	12.000	11.000	0.438387413	-1.3953122
marital*	4	7000	2.16942857	0.5996925	2.000	2.21035714	0.0000000	1.000	4.000	3.000	-0.052613827	-0.3175492
education*	5	7000	4.77185714	2.1353225	4.000	4.91214286	2.9652000	1.000	8.000	7.000	-0.257946129	-1.1959371
default*	6	7000	1.20414286	0.4031027	1.000	1.13017857	0.0000000	1.000	2.000	1.000	1.467689986	0.1541360
housing*	7	7000	2.06442857	0.9856770	3.000	2.08053571	0.0000000	1.000	3.000	2.000	-0.129067394	-1.9583801
loan*	8	7000	1.31671429	0.7132775	1.000	1.14589286	0.0000000	1.000	3.000	2.000	1.869634180	1.5857008
contact*	9	7000	1.35971429	0.4799509	1.000	1.32464286	0.0000000	1.000	2.000	1.000	0.584500147	-1.6585965
month*	10	7000	5.24871429	2.3338330	5.000	5.32642857	2.9652000	1.000	10.000	9.000	-0.299013583	-1.0127403
day_of_week*	11	7000	3.02014286	1.3988831	3.000	3.02517857	1.4826000	1.000	5.000	4.000	-0.012978591	-1.2749168
campaign	12	7000	1.55500000	2.6315055	1.000	0.99910714	1.4826000	0.000	32.000	32.000	4.160547700	26.5522882
pdays	13	7000	962.26442857	187.4276625	999.000	999.00000000	0.0000000	0.000	999.000	999.000	-4.904734224	22.0600271
previous	14	7000	0.17614286	0.4965686	0.000	0.04982143	0.0000000	0.000	6.000	6.000	3.767054784	19.4431201
poutcome*	15	7000	1.92614286	0.3666353	2.000	1.99142857	0.0000000	1.000	3.000	2.000	-0.886187415	3.7809451
emp.var.rate	16	7000	0.04881429	1.5878130	1.100	0.23405357	0.4447800	-3.400	1.400	4.800	-0.694219113	-1.1032336
cons.price.idx	17	7000	93.56868986	0.5829994	93.444	93.57442946	0.8154300	92.201	94.767	2.566	-0.228388905	-0.8438292
cons.conf.idx	18	7000	-40.47275714	4.6829851	-41.800	-40.59971429	6.5234400	-50.800	-26.900	23.900	0.333920059	-0.3153432
euribor3m	19	7000	3.58687614	1.7488719	4.857	3.76443143	0.1601208	0.634	5.045	4.411	-0.669202310	-1.4640703
nr.employed	20	7000	5165.42015714	73.3050028	5191.000	5176.93112500	55.0044600	4963.600	5228.100	264.500	-1.009506512	-0.1018737
subscribe	21	7000	0.11742857	0.3219533	0.000	0.02178571	0.0000000	0.000	1.000	1.000	2.376225422	3.6469683

We observe that the variables *default, loan, campaign, pdays, previous, nr.employed and subscribe* are variables that are highly skewed (skewness less than -1 or greater than 1). Also we observe that some of this variables also have a leptokurtic distribution (kurtosis higher than 3): *campaign, pdays and previous.* This two indicators suggest the presence of outliers or a heavy concentration of values.

2. Model: Logistic regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



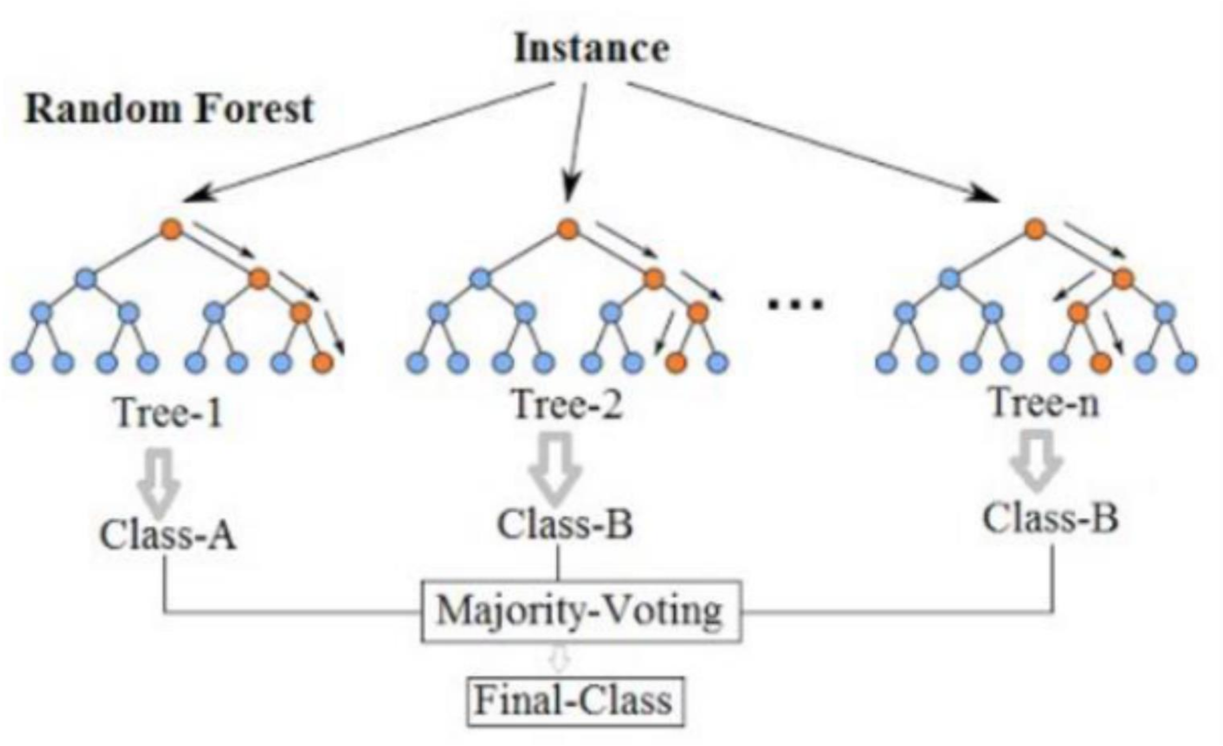
Target: maximize the likelihood function

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

The logistic regression is more appropriate in this case that we have a binary variable. This regression return us a number between 0 and 1 that is the probability to be 1.

The parameters of a logistic regression model are estimated by the probabilistic framework called maximum likelihood estimation.

2. Model: Random Forest



A decision tree splits the observations in each node in order to increase the purity of the subgroup. This means that the resulting groups are as different from each other as possible.

The Random Forest classifier consists of a number of decision trees that operate as an ensemble. Each Decision tree predicts a class and the majority class becomes our model's prediction.

2. Model: Naive Bayes

Bayes theorem

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

The Naïve Bayes classifier is a probabilistic machine learning model that is based on the Bayes Theorem. We can find the probability of A happening, given that B has occurred.

In this model we want to maximize the probability of finding one of the classes of the target variable (y).

2. Model: Linear discriminant analysis (LDA)

- Within-class variance:

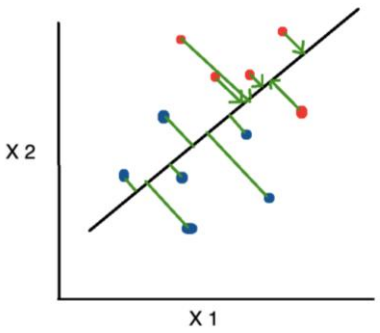
$$s_k^2 = \sum_{n \in \mathcal{C}_k} (z_n - m_k)^2, \quad k = 1, 2$$

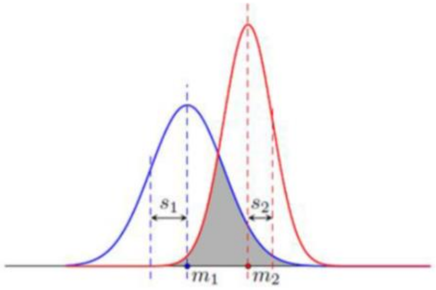
- Between-class variance:

$$m_1 - m_2 = \frac{1}{N_1} \sum_{i \in \mathcal{C}_1} z_i - \frac{1}{N_2} \sum_{j \in \mathcal{C}_2} z_j = \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2)$$

- Maximize the objective function:

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$



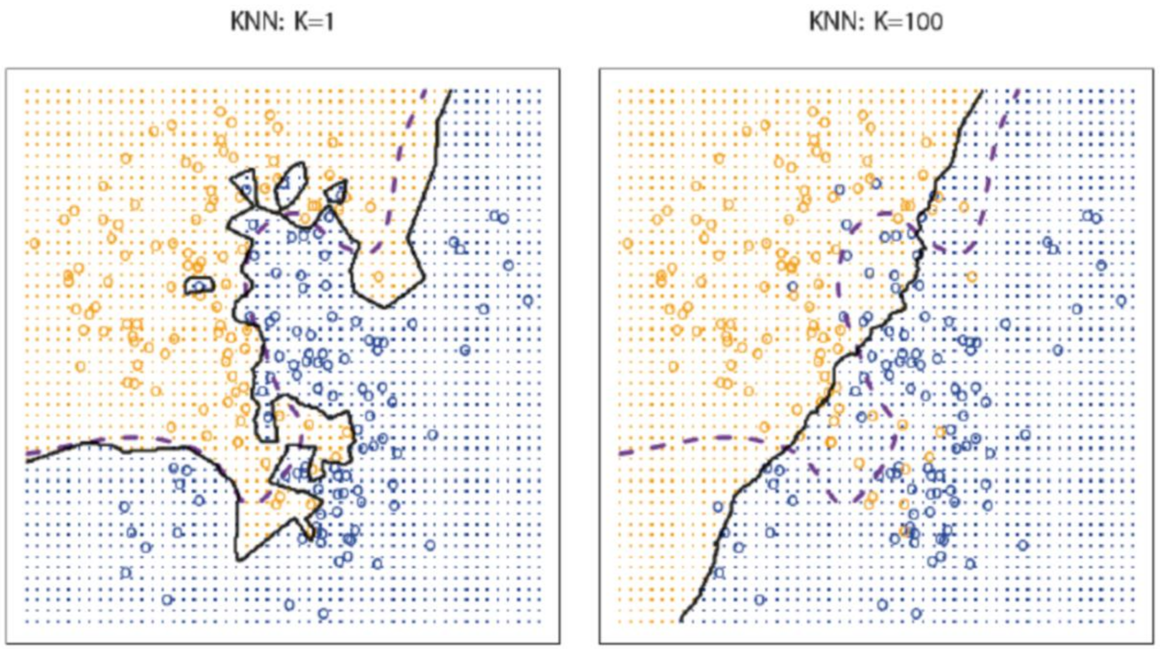


b) Small $(m_1 - m_2)^2$, small $s_1^2 + s_2^2$

LDA is a dimensionality reduction technique that looks for a linear combination between features that best separates the classes of the target variable.

With this model we rescue the information coming from different dimensions and we find a new dimension that minimizes the variance and maximizes the distance between the means of the two classes.

2. Model: KNN



The KNN is a supervised classification algorithm. It calculates the distances between a data point and its K nearest neighbors (based on distance) and based on their classes we assign the data point a class. Due to this, when we increase the number of neighbors (K) the separation becomes smoother.

3. Models summary: AUC

Logistic regression: 0.8067

Pro: easy to interpret
Cons: can solve non linear problems

LDA: 0.8098

Pro: No hyperparameter tuning
Cons: can't reduce dimensions to more than the number of classes

Naïve bayes: 0.7929

Pro: Performs well in multiclass prediction
Cons: Has the assumption of independent predictors.

Random forest: 0.7368

Pro: Interpretability
Cons: Unstable

KNN: 0.6958

Pro: Its consistency increases with + data
Cons: Time consuming

Cross Validation
We run a cross validation with the random forest mode to tune its number of trees and number of available variables to split the tree nodes.

Performance
The best performing model was the logistic model followed by the LDA and Naïve Bayes.