




# ANALYSE STATISTIQUE & LANGAGE R

Modélisation et évaluation

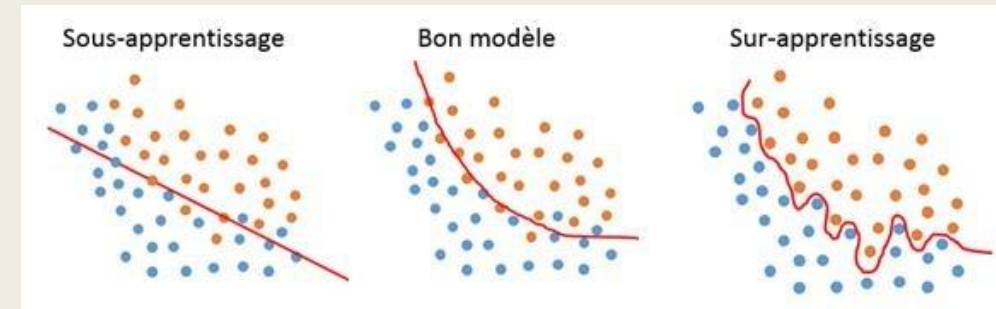
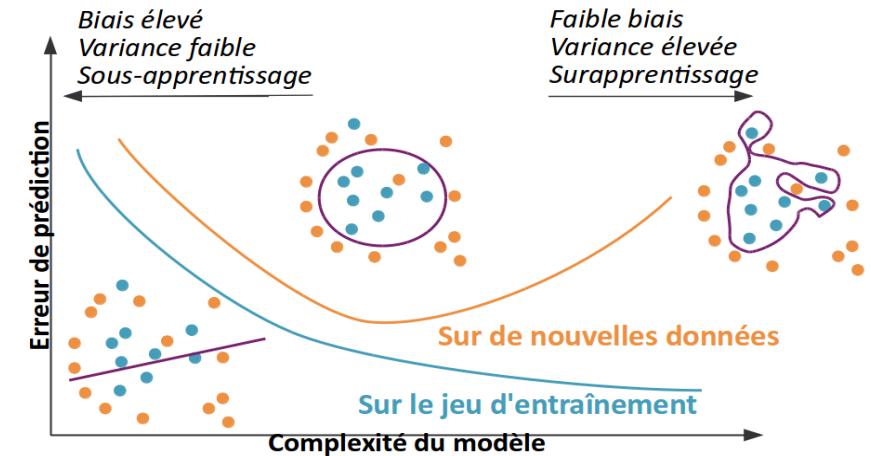


# Déroulement

- Biais / variance
- Régression logistique
- Métriques
  - *Matrice de confusion*
  - *Courbes ROC / précision*
  - *AUC*
  - *Probabilité d'appartenance à une classe*
- Exercice
  - *Régression logistique sur une variable (avec affichage de graphiques)*
  - *Régression logistique multivariées*
- Analyses factorielles
  - *PCA*

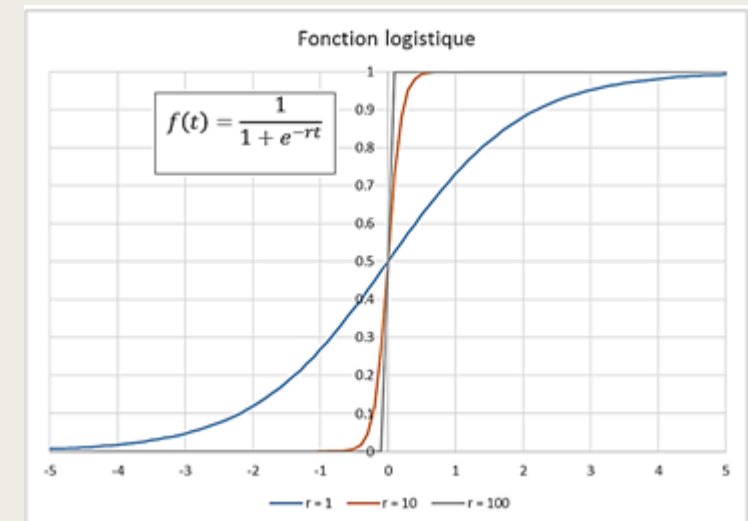
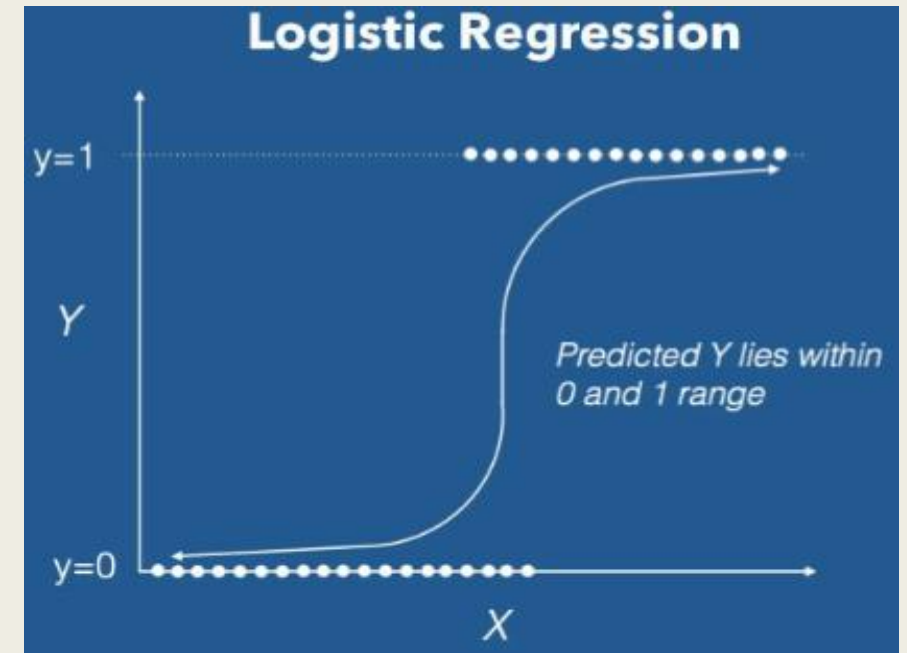
# Biais / variance

- **Biais faible et variance élevé :**
  - *peu d'erreurs en apprentissage*
  - *modèle complexe*
  - *modèle sur-entraîné*
  - *adaptation moindre aux nouvelles entrées*
- **Biais élevé / variance faible :**
  - *beaucoup d'erreurs en apprentissage*
  - *modèle sous-entraîné*
  - *meilleure adaptation aux nouvelles entrées*



# La régression logistique

- Problème de ~~régression~~ / classification
- Fonction logistique = fonction sigmoïde [0 ; 1]
- Régression logistique binomiale : prédire la variable à expliquer 0/1
- N variables explicatives X
$$\frac{e^{\beta_0 + B_1 * x_1 + B_2 * x_2}}{1 + e^{\beta_0 + B_1 * x_1 + B_2 * x_2}}$$
- Ajuster l'équation en fonction des données, maximiser la probabilité d'observer l'échantillon



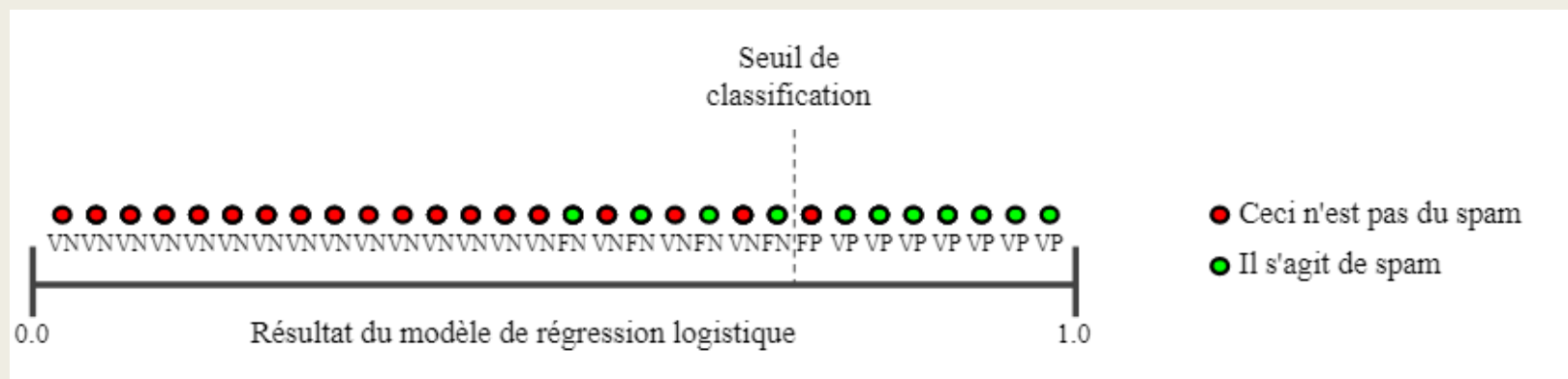
# Matrice de confusion

	FALSE	TRUE	
0	7076	3929	Vrais négatifs (VN) Faux positifs (FP)
1	358	994	Faux négatifs (FN) Vrais positifs (VP)

- **Sensibilité (VP)** =  $TP / (TP+FN) \sim 73\%$ 
  - *Proportion de positifs prédits parmi tous les positifs avérés*
- **False Positive Rate (FP)** =  $FP / (FP+TN) \sim 36\%$ 
  - *Proportion de faux positifs, parmi tous les négatifs avérés*
- **Precision** =  $TP / (FP+TP) \sim 20\%$ 
  - *Précision du modèle avec les vrais positifs prédits, sur tous les positifs prédits (bon ou mauvais)*

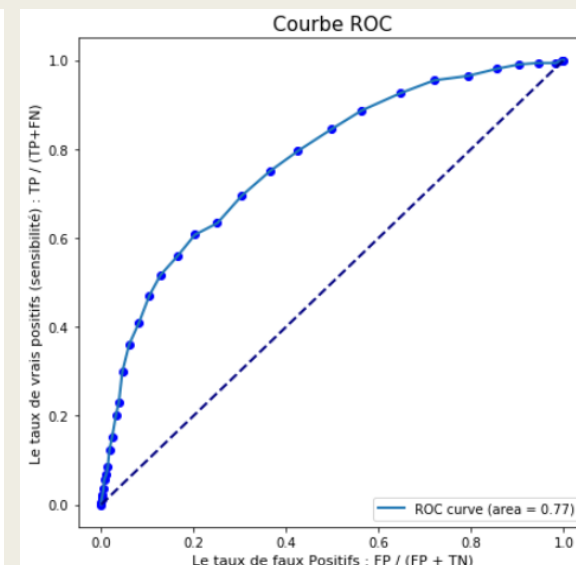
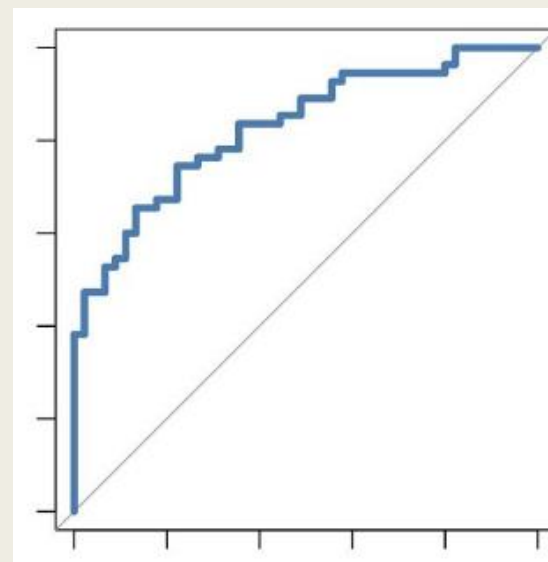
# AUC – area under curve

## ■ Seuil de classification



ID	Proba	Class
1	0,98	1
2	0,96	1
3	0,96	1
4	0,95	1
5	0,89	1
6	0,88	1
7	0,84	1
8	0,82	0
9	0,78	1
10	0,77	0

- Aire sous la courbe ROC
  - Aire à 0,5 => algo inefficace
  - Aire à 1 => algo parfait
- Abscisse : taux de faux positif
- Ordonné : taux de vrais positif

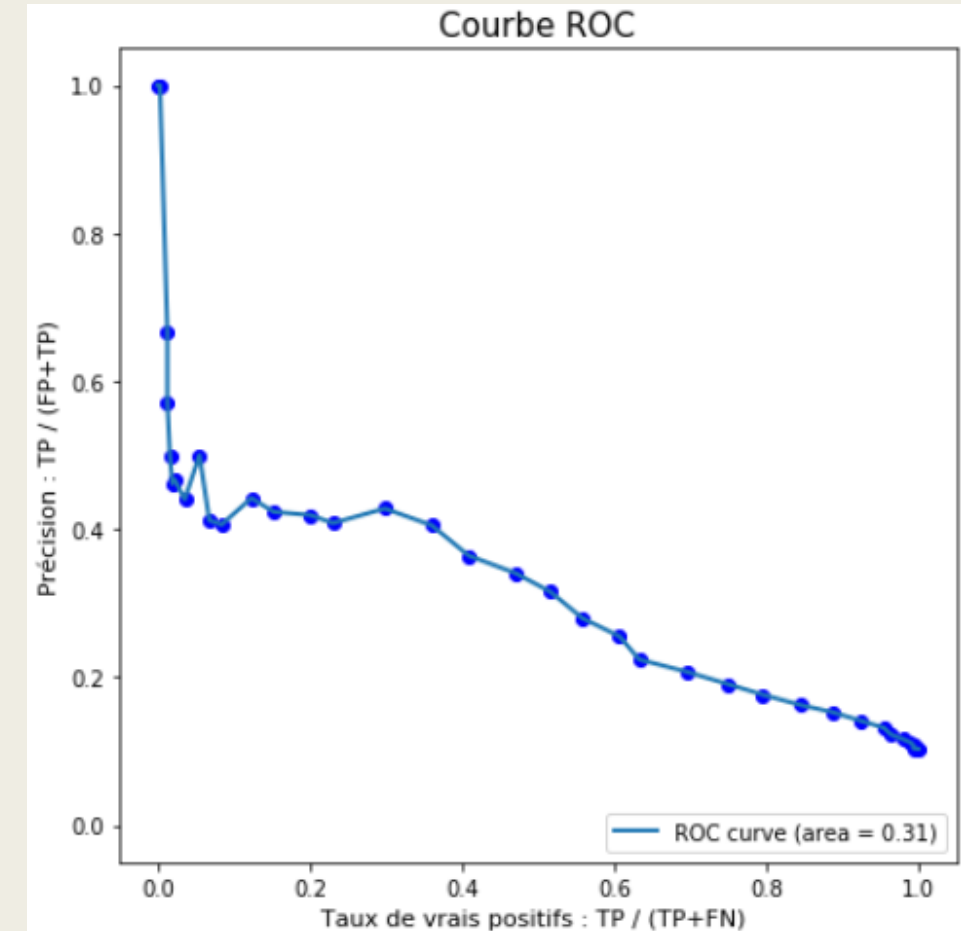


# Courbe de précision

- Ordonné : la précision  $TP / (FP+TP)$
- Abscisse : recall/sensibilité  $TP / (TP+FN)$

Ici la précision diminue au détriment de la sensibilité, on cherche donc à classer le plus de positifs possibles, quitte à classier des FP.

Si l'on dispose de peu de VP, cette métrique est plus révélatrice



# Exercice de régression logistique

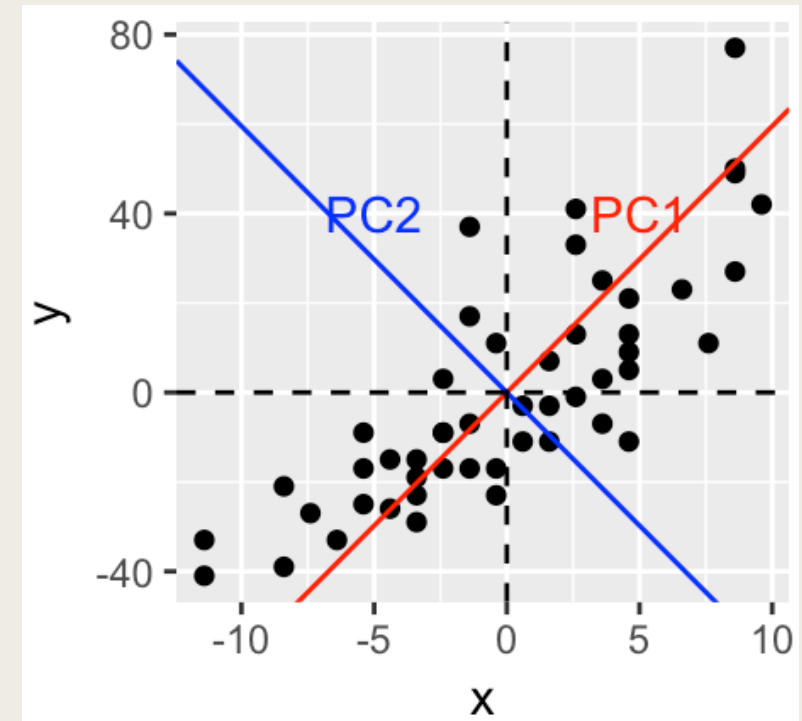
- Jeu de données : <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>
- Réaliser une régression logistique une variable
  - *Split train / test : **sample()***
  - *Entraîner le modèle : **glm()***
  - *Tester le modèle : **predict()***
  - *Tracer la régression*
  - *Matrice de confusion, modifier le seuil de classification*
- Réaliser une régression logistique avec l'**ensemble des variables**
  - *Matrice de confusion*
  - *Afficher une courbe ROC*
  - *Afficher une courbe de précision*
  - *Afficher l'aire sous la courbe ROC*

=> Librairie **ROCR** pour la courbe ROC



# Analyses factorielles

- Chercher des éléments représentant la diversité des données, et identifier les facteurs descriptifs principaux
- Créer de nouveaux axes principaux, dits « synthétiques », expliquant au mieux la variance des données
- Calculer des combinaisons linéaires, matrices de variance / covariance, diagonalisation des matrices
- PCA : Principal Component Analysis
  - *Matrice de distances euclidiennes :  $\sqrt{\sum (x_i - y_i)^2}$*
  - *Variables quantitatives*
- Objectifs :
  - *Réduire le nombre de variables*
  - *Combiner des variables corrélées, pour créer de nouvelles variables indépendantes*



# Cercle des corrélations

- Mettre en évidence la ou les variables les plus discriminantes, celles expliquant le mieux la distribution / l'inertie des données (variance...)
- Déterminer les variables les plus contributrices
- Visualiser les variables positivement / négativement corrélées, qui sont regroupées.
- Les axes partent de l'ordonnée à l'origine (0,0), ensuite chaque axe possède une coordonnée

