Betclic

Théophile Ravillion – NBA Challenge

Test technique recrutement Data Science Nov-Dec 2020

NBA Challenge – énoncé

A partir des données fournies, proposer une côte pour la victoire finale de ***chacune des 30 équipes*** *de la ligue* ***avant le début de la saison*** *2018-2019 puis* ***avant les playoffs*** *(à la fin de la saison régulière) de cette même saison.*

Une approche de travail en 3 étapes

1 Travail préparatoire

Comprendre et retraiter les données mises à disposition pour mettre au point un modèle performant de ML

1a

Pre-processing

Compréhension de la problématique
Récupération des données
Traitement & consolidation des données
Définition et extraction des variables

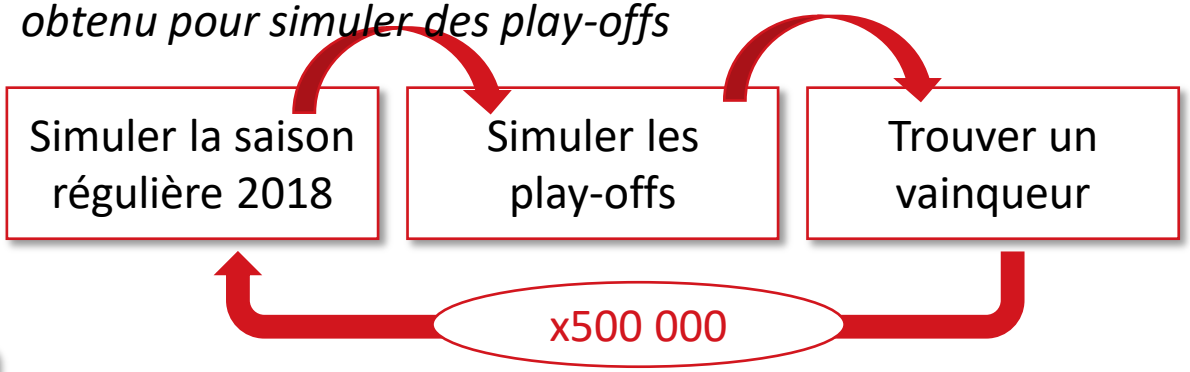
1b

Entraînement algo & mesure de la performance

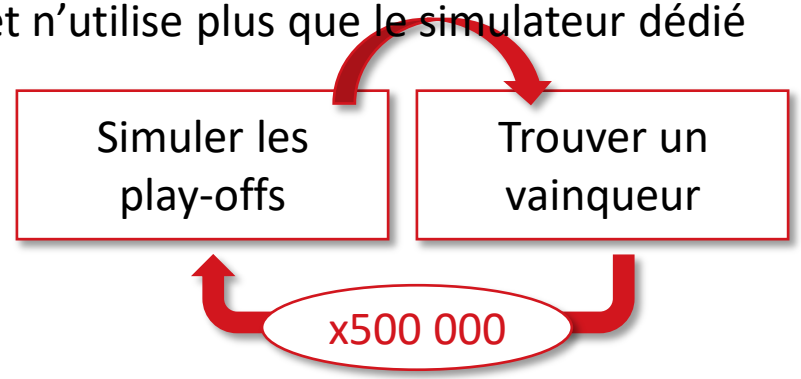
Recherche d'un modèle Machine Learning
Entraînement de l'algorithme
Evaluation des performances du modèle

2 Simulation de la NBA

On simule la saison régulière et utilise le classement obtenu pour simuler des play-offs



Si la saison régulière a déjà eu lieu, on entraine un model avec des données à jour et n'utilise plus que le simulateur dédié



3 Etablissement des cotes

Le pourcentage de victoire obtenu par chaque équipe converge vers sa probabilité de victoire et permet d'établir sa cote

1. Travail préparatoire (1/2)

Identification de **l'équipe de chaque joueur pour chaque match**

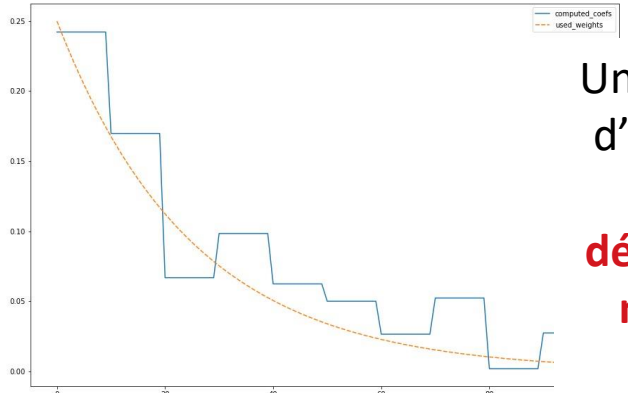
Game	Player	Player Team	Home Team	Away Team
201905250TOR	Kawhi Leonard	?	milwaukee bucks	toronto raptors
201905300TOR	Kawhi Leonard		golden state warriors	toronto raptors
201905210TOR	Kawhi Leonard		toronto raptors	golden state warriors

Calcul du **Game Score par joueur et par match**: Kpi haut level du statisticien John Hollinger

Game Score = PTS + 0.4 * FG - 0.7 * FGA
+ 0.4*(FTA - FT) + 0.7 * ORB + 0.3 * DRB
+ STL + 0.7 * AST + 0.7 * BLK - 0.4 * PF - TOV

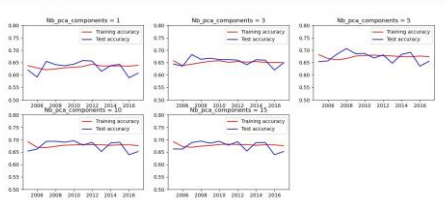
Pour chaque équipe, chaque joueur et chaque match : Calcul de la **moyenne pondérée de chaque KPI disponible des 150 derniers matchs**

Calcul pour chaque équipe de la **moyenne des kpi's de chacun de ses joueurs**



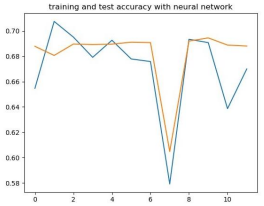
Une régression logistique a permis d'identifier le coefficient de la **loi de puissance qui décrit la décroissance de l'importance des matchs passés pour prédire le présent**

Synthèse des nombreuses variables créées via un PCA qui ne garde que **5 variables par équipe**



5 composants est le meilleur compromis biais variance

Recherche du modèle approprié



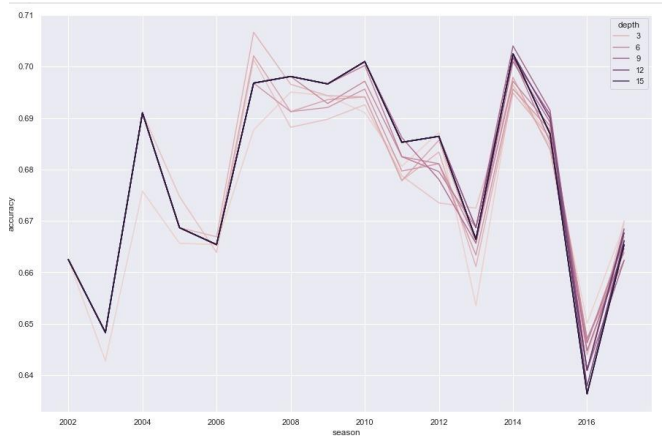
Neural Network et **Logistic Regression** parviennent à des résultats sensiblement identiques : accuracy=~68%. **On préfèrera la simplicité de la deuxième solution.**

2. Simulation de la NBA

Entraînement d'une **régression logistique** & **recherche des hypers paramètres**

Grid Search permet de conclure qu'**une absence de régularisation sera ici le plus efficace**. Nous obtenons un modèle d'une grande simplicité.

Recherche des hyper paramètres : **On utilise seulement les 12 années précédentes**



Le Basket évolue…

Utiliser 12 ans de données maximum pour entrainer le modèle est optimal

Evaluation des résultats des **870 affrontements possibles et de la saison régulière**

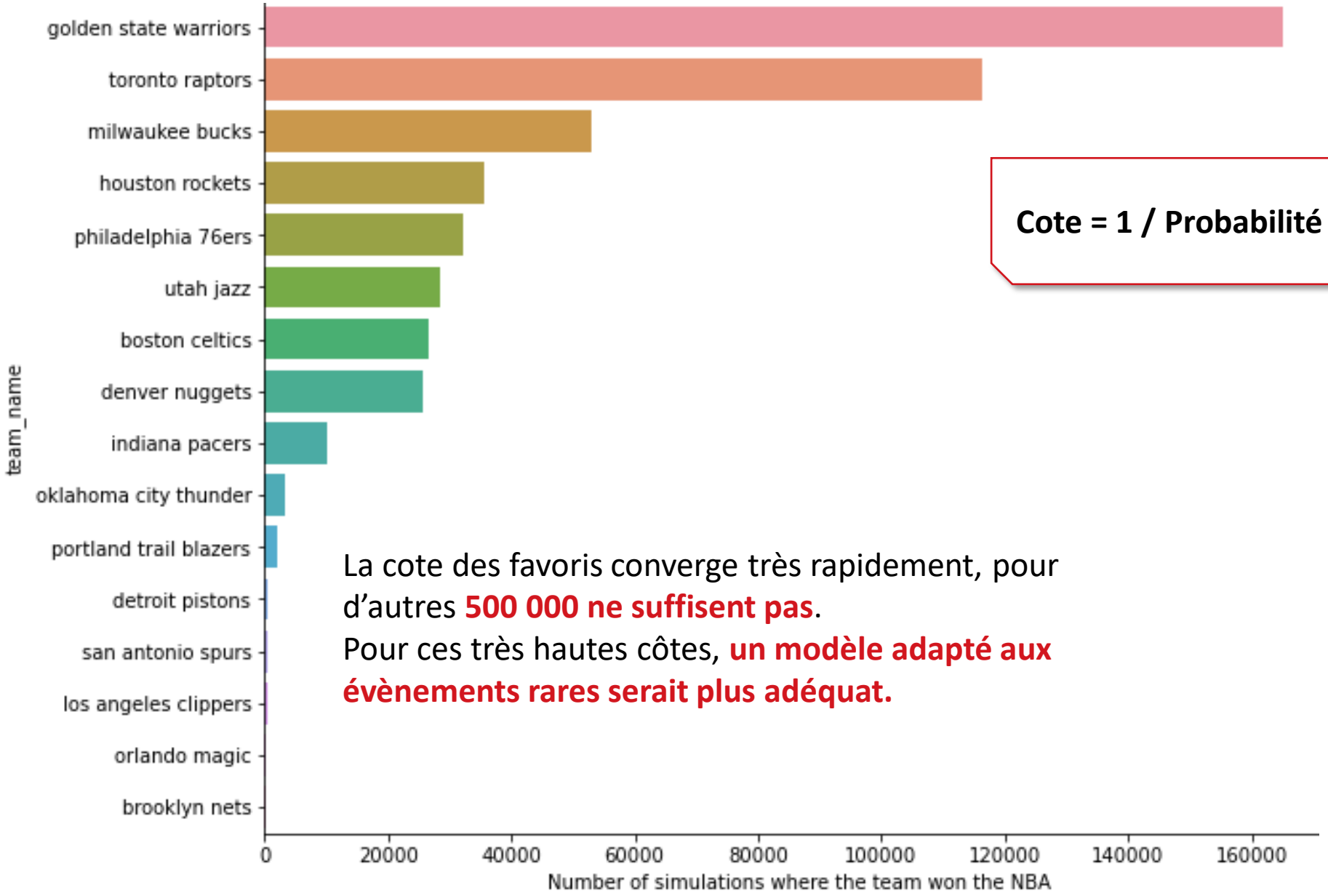
Simulation de la saison régulière : **Pour chaque match, on lance un dé**

Le Grizzly n'a plus qu'a espérer un double six !

Simulation des play-offs : **Pour chaque « Best Of 7 », on lance un dé**

Une exploitation de la loi binomiale nous permet de passer des probabilités de gagner à domicile et à l'extérieur à celle de gagner un BO7

3. Prédiction des cotes



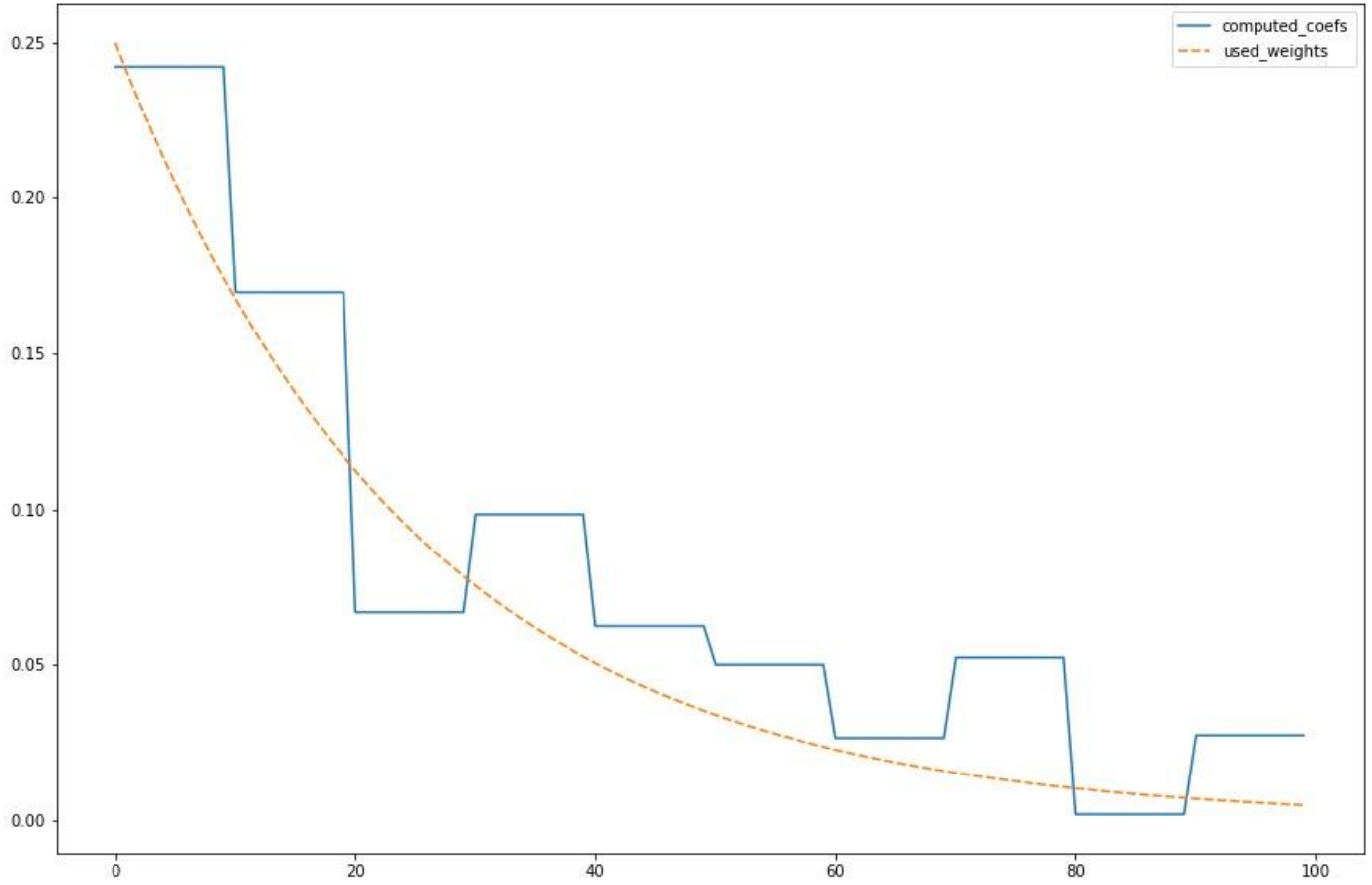
Cote = 1 / Probabilité estimée

La cote des favoris converge très rapidement, pour d'autres **500 000 ne suffisent pas**.
Pour ces très hautes côtes, **un modèle adapté aux évènements rares serait plus adéquat.**

ANNEXES

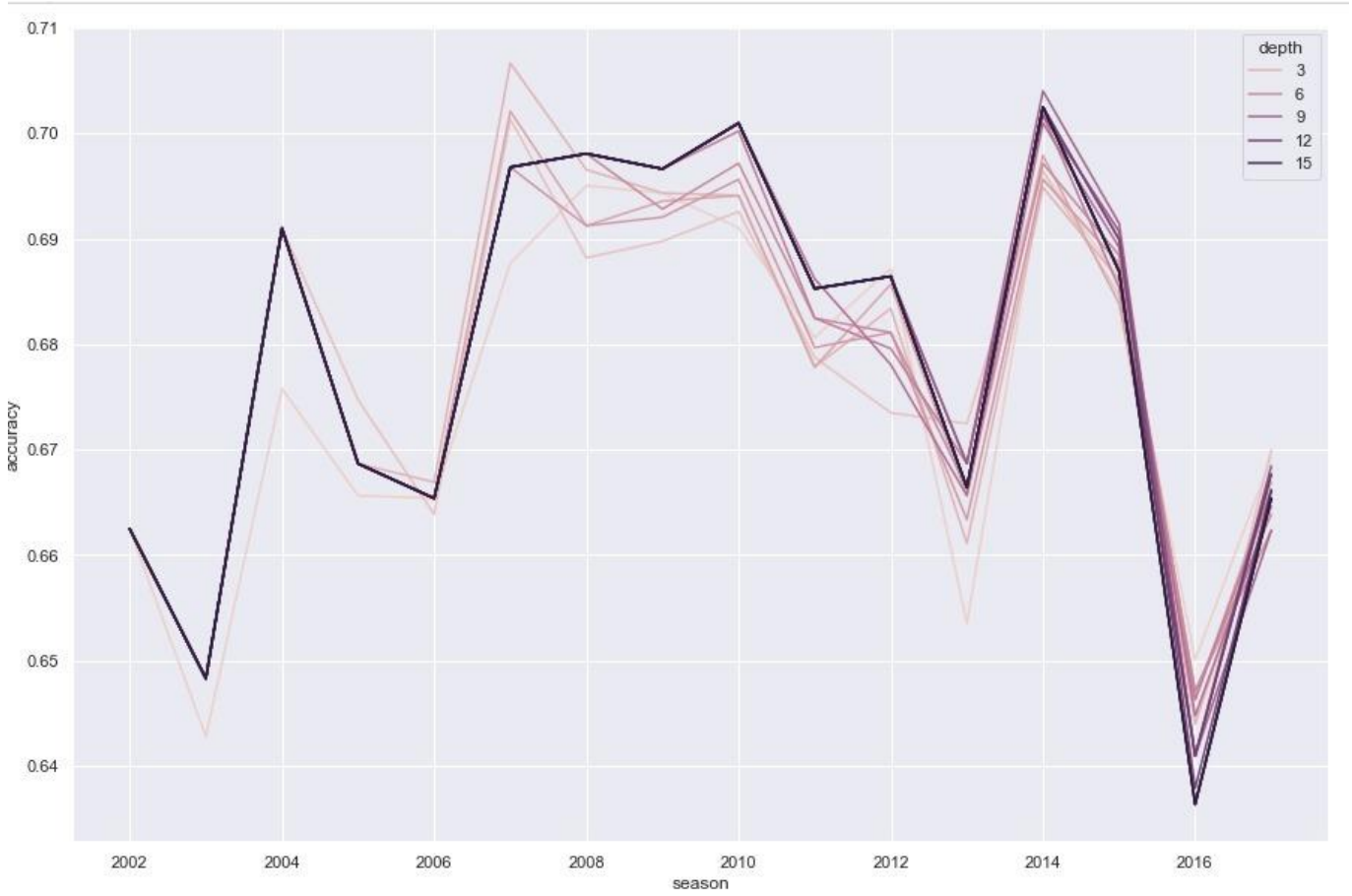
Loi décrivant l'oubli des performances passées

- Une régression logistique prenant en paramètre le % de victoire des 100 derniers matchs regroupé en 10 moyennes (1 à 10, 11 à 20, …) affecte des coefficients plus important aux matchs récents.
- L'importance décroit suivant une loi de puissance paramétrée par -0,03

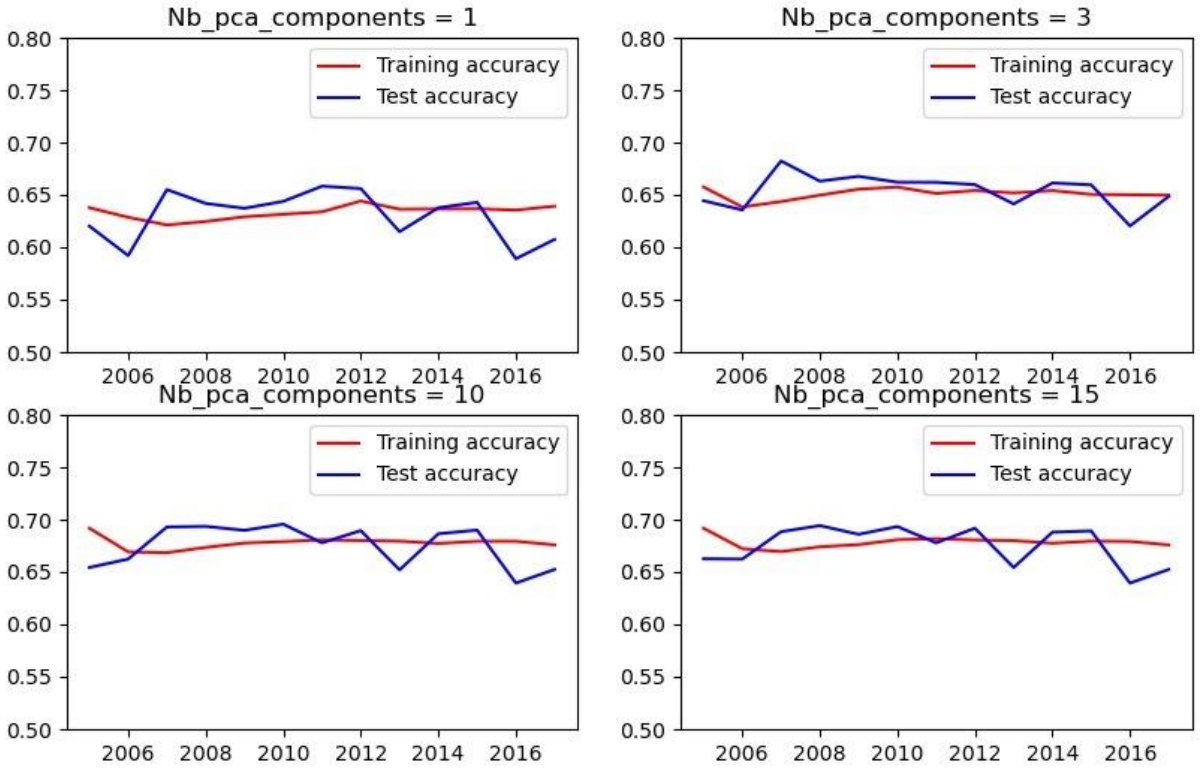


Précision du modèle par an et par profondeur mnésique

- La précision du modèle varie avec les années.
- Elle varie aussi suivant le nombre d'années (profondeur) utilisées pour entrainer le modèle.



Choix du nombre de composants issus du PCA



- Un PCA nous permet de synthétiser l'ensemble de l'historique des kpis d'une équipe.
- 5 Composants offrent le meilleur compromis biais variance.

Précision du prédicteur de match

- Un réseau de neurone n’aurait pas permis d’obtenir des résultats significativement meilleurs.