RICHARD Paul-Antoine, TERNAT Clément M2 Data-Sciences

Économétrie des choix discrets - COMPAIRE Philippe

SUJET 15 : Arbres de décision et Nested Logit (5 boules + 2 étoiles)

1. Préambule

Notre consigne était de réaliser un Arbre de décision ou un Nested Logit, sous contrainte d'utiliser les 5 boules ainsi que les 2 étoiles. Nous avons fait le choix de réaliser des arbres de décision.

Nous avons choisi de travailler sur Python, par habitude de travail et parce que les rendus visuels du logiciel pour ce qui concerne les arbres de décision sont plus poussés que ceux proposés par SAS.

Notre problématique est la suivante :

"Quelle est la stratégie de jeu la plus efficace pour maximiser ses chances d'être gagnant au rang 1 ?"

2. Traitement des données

Pour commencer, à l'aide de la librairie "pandas" nous avons récupéré les données du document Excel qui nous avait été fourni, et étant donné que nous avions pour contrainte de nous concentrer sur l'entièreté des boules, nous avons demandé à notre programme de ne plus s'occuper du reste, excepté notre variable à expliquer : 'GAGNANT RANG 1'.

```
features = data[['B1', 'B2', 'B3', 'B4', 'B5', 'E1', 'E2']]
labels = data['GAGNANT RANG 1']  # Cible
```

Il faut noter que pour gagner au rang 1 il faut avoir trouvé toutes les boules et toutes les étoiles.

Ensuite nous avons regardé s'il manquait des données dans certaines lignes :

```
Nombre de lignes avant suppression : 1685
Nombre de lignes après suppression : 1685
```

On constate qu'il n'y a pas de données manquantes concernant les 8 colonnes que nous étudions.

Nous avons voulu vérifier s'il pouvait y avoir plusieurs gagnants de rang 1 et le résultat est le suivant :

```
array([1, 0, 2, 4, 3, 5])
```

Il est déjà arrivé qu'il y ait jusqu'à 5 gagnants de rang 1 pour un seul tirage d'euromillions.

On a aussi calculé le pourcentage de gagnant au rang 1 parmi tous les gagnants :

```
0.24154302670623146
```

24,15% des gagnants sont des gagnants au rang 1.

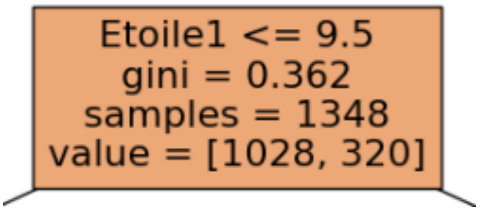
Pour finir, nous avons simplifié les étiquettes gagnantes. En effet notre objectif étant de chercher la meilleure stratégie gagnante, peu nous importait s'il y avait un ou plusieurs gagnants par tirage. Nous avons donc pour simplifier notre analyse affecter une étiquette égale à 0 s'il n'y avait pas de gagnants au rang 1, et une étiquette égale à 1 s'il y avait un ou plusieurs gagnants au rang 1 lors d'un tirage.

Nos données sont donc prêtes à être utilisées dans notre étude !

3. Arbre de décision

3.1 - Explication d'une feuille

Pour démarrer avec les arbres de décision, nous allons expliquer quelles sont les informations présentes sur un noeud avec un exemple :



- **Etoile1 >= 9.5 :** C'est la condition que l'arbre utilise pour séparer les données. Si la condition est validée, alors on suivra la branche de gauche, sinon celle de droite.

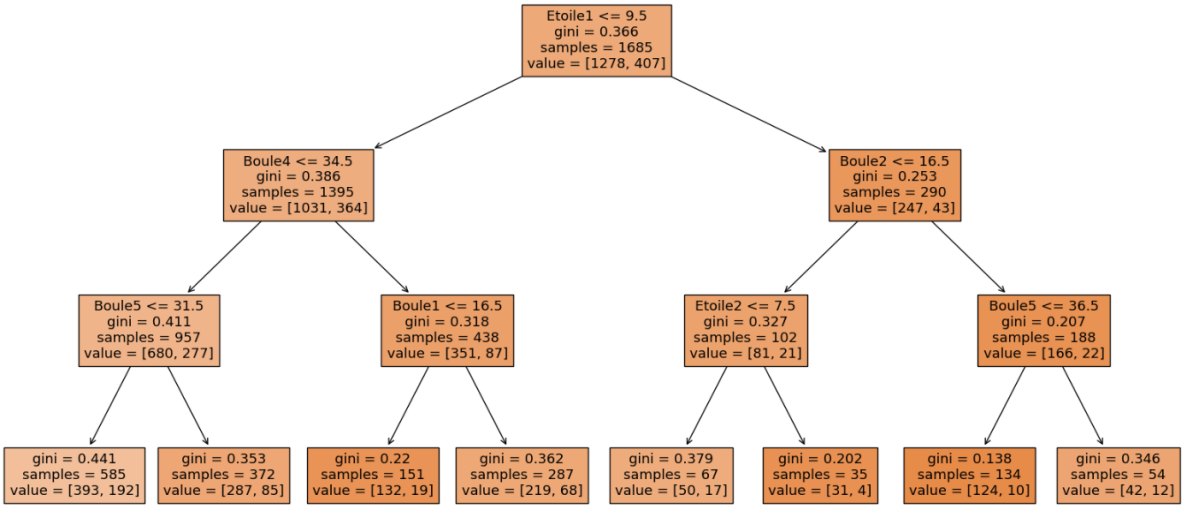
- **gini = 0.362 :** Il s'agit d'une mesure de pureté comprise entre 0 et 1, où 0 indique une séparation parfaite entre les caractéristiques (par exemple ici gagnant ou perdant). Ici le niveau d'impureté est intermédiaire.

- **samples = 1348 :** Il s'agit du nombre d'observations qui sont dans ce noeud.

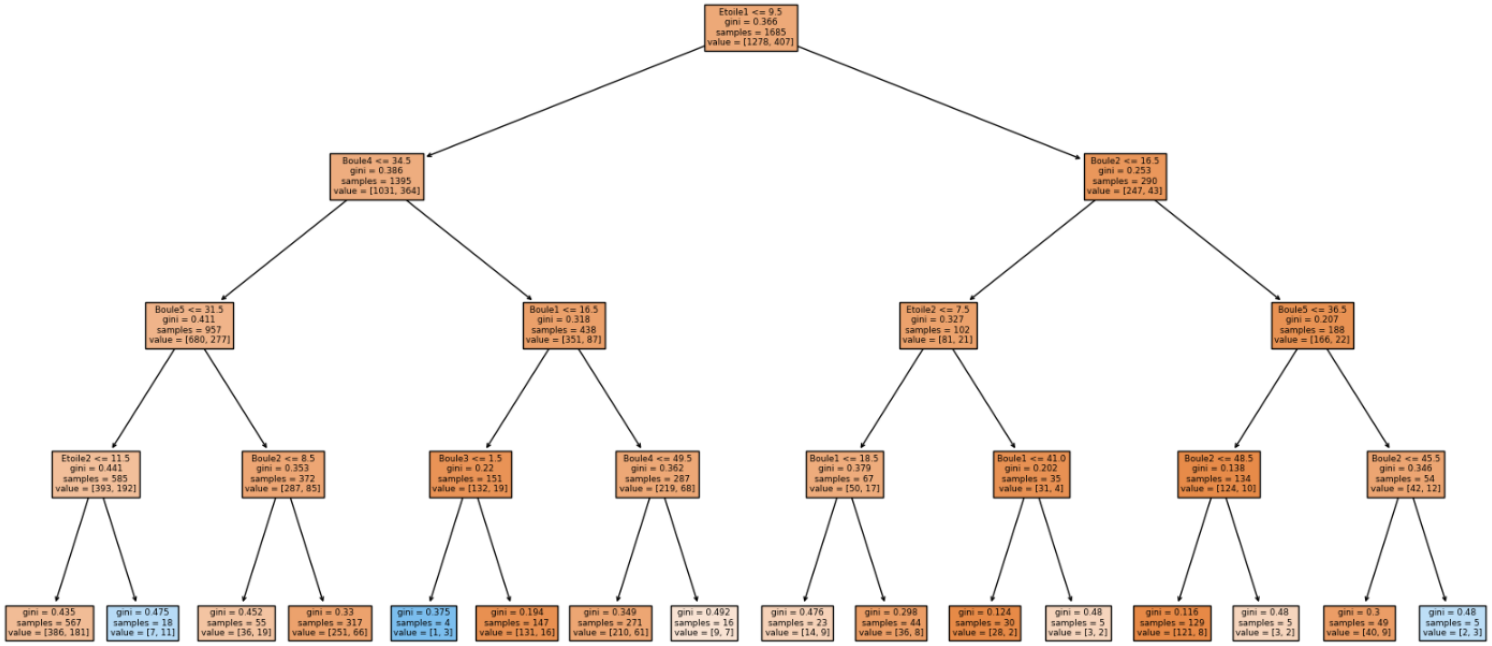
- **value = [1028, 320] :** Il s'agit de la distribution entre non-gagnants et gagnants parmis ce qui est contenu dans 'samples'.

3.2 - Arbres de décision avec profondeurs définies

Pour démarrer nous traçons un arbre de décision de profondeur maximale 3, afin de visualiser comment fonctionne notre arbre et de voir si des feuilles nous montrent déjà un chemin potentiel vers les gains :

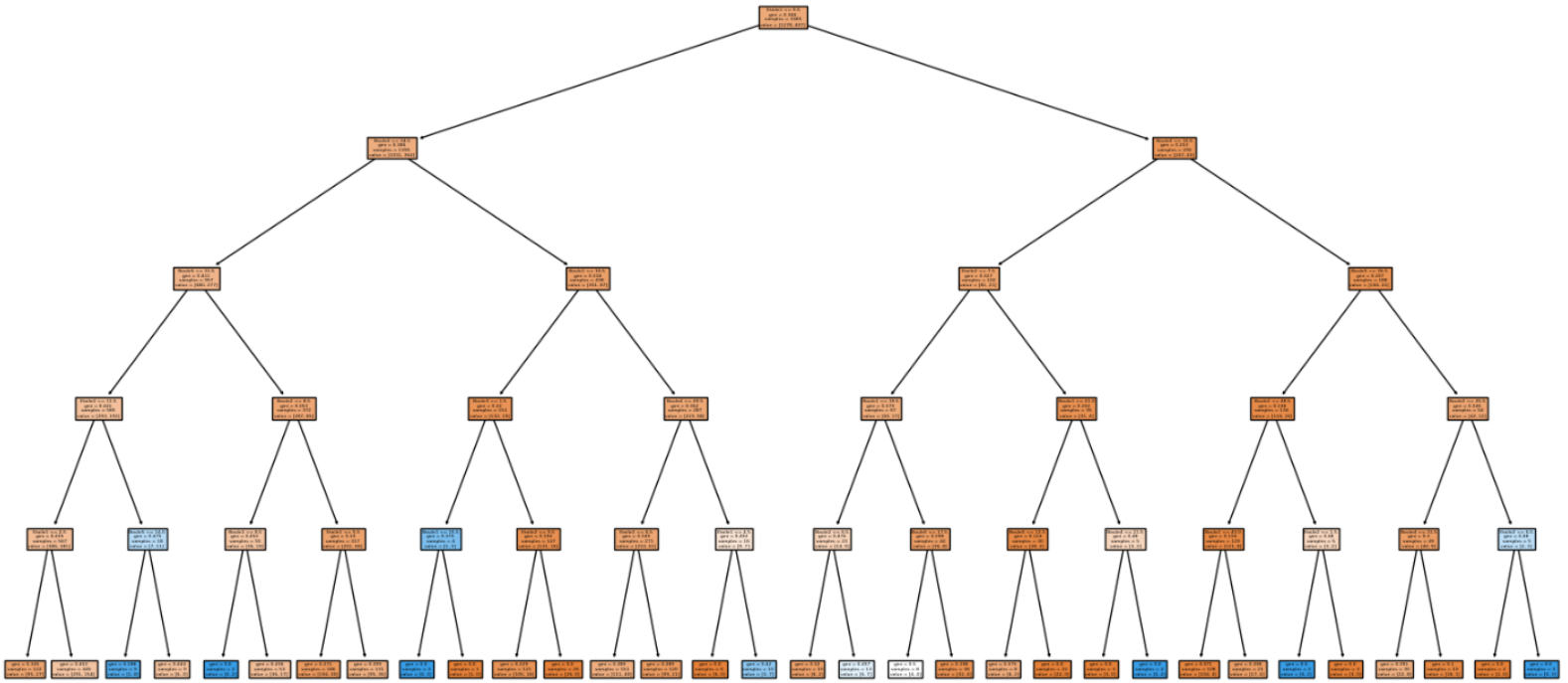


On ne constate aucune feuille où les valeurs sont en faveur du gain, il faut donc continuer avec un arbre de profondeur 4.



Dans cet arbre plus profond, nous voyons apparaître des cases bleues qui laissent sous-entendre qu'il y a des gains en majorité, cependant il n'y a toujours pas de chemins qui mènerait à coup sûr à un gain (évidemment qu'aucun chemin n'est sûr, on se base sur les données passées).

On passe alors à un arbre plus profond :

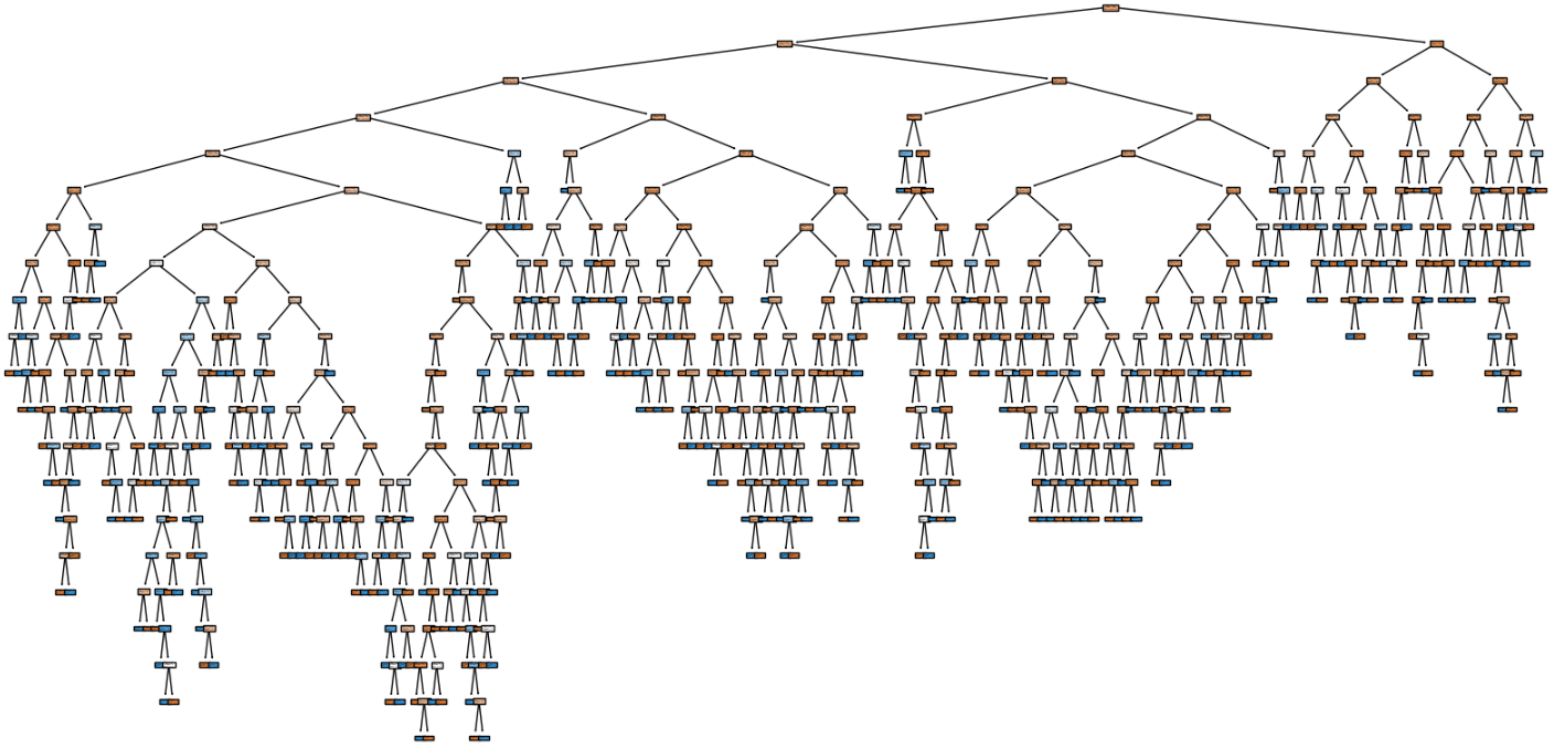


Cette fois, des cases bleues foncées apparaissent, ce qui signifie que des chemins qui ne mènent qu'à des réussites sont apparus, en l'occurrence ici 6 différents. On constate aussi évidemment que lorsqu'on ajoute 1 de profondeur, les mêmes arbres apparaissent avec un étage supplémentaire, ce qui montre que les arbres sont bien cohérents. Cependant il devient presque impossible de lire ce qu'il est écrit sur les feuilles.

Maintenant que nous savons comment fonctionnent nos arbres et que nous avons la preuve qu'ils sont cohérents nous allons pouvoir démarrer notre analyse grâce à un arbre de décision global.

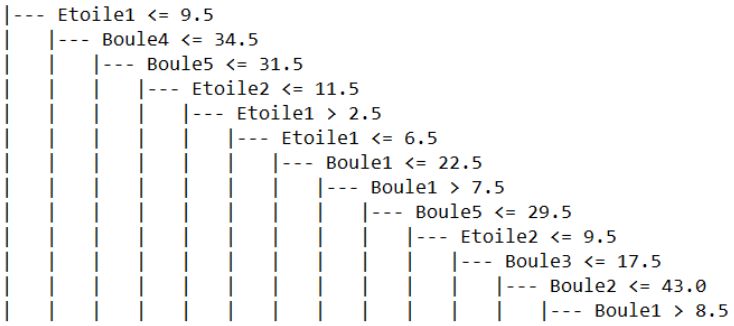
3.3 - Arbre global et Analyse

On trace l'arbre de décision le plus complet :

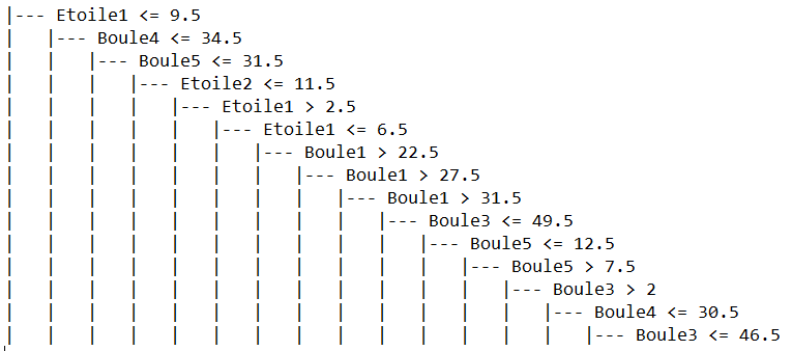


La profondeur maximale de cet arbre est 21, et il y a 90 moyens d'atteindre des prédictions qui laissent présager un gain de rang 1.

On cherche alors à savoir quels sont les chemins qui ont menés vers le plus de gagnants par le passé. Pour cela nous analysons l'arbre ci-dessus. Les deux chemins qui ont mené à un maximum de victoire sont les suivants :



Chemin [0, 13]



Chemin [0, 13]

Nous avons deux chemins différents qui ont tous les deux menés à un total de 13 gains.

Analysons ces chemins :

- On sait que les étoiles peuvent prendre des valeurs de 1 à 12 et que les boules peuvent prendre des valeurs entre 1 et 50.
- Les valeurs optimales obtenues pour le premier sont dans ce tableau:

	B1	B2	B3	B4	B5	E1	E2
MIN (>=)	9	1	1	1	1	3	1
MAX (<=)	22	43	17	34	29	6	9

- Pour le deuxième :

	B1	B2	B3	B4	B5	E1	E2
MIN (>=)	32	3	1	1	8	3	1
MAX (<=)	50	50	46	30	12	6	11

On a donc les bornes de chacun des chemins, on va donc pouvoir obtenir les bornes les plus efficaces en fusionnant ces tableaux :

	B1	B2	B3	B4	B5	E1	E2
MIN (>=)	9	3	1	1	8	3	1
MAX (<=)	22	43	17	30	12	6	9

Nous avons donc obtenu la stratégie qui est statistiquement la plus efficace. Pour avoir le plus de chance de gagner à l'Euromillion il faut donc suivre ces contraintes.

On remarque par ailleurs que pour les deux chemins on a les 6 premières étapes qui sont similaires, ce qui semble être un bon début pour gagner.

4. Conclusion

Pour conclure, on a pu constater grâce à un arbre de décision comme ceux présentés ci-dessus qu'il reste quasiment impossible de prévoir un gain de rang 1. En effet, nous nous sommes concentrés sur les combinaisons qui fonctionnent le plus, mais au final elles ne représentent que 6,39% des gains de rang 1, ce qui est très faible, on ne peut pas faire de ces résultats une généralité.