SD701 : Prédiction de la « Draft » NBA

1. Description du projet

Aux États-Unis, pour qu'un athlète puisse devenir professionnel, il doit s'inscrire dans une université. Après la validation de la première année, il est éligible à passer professionnel. Pour faire ses preuves, il participe, avec son équipe d'université, à la compétition organisée par la NCAA (*National Collegiate Athletic Association*). Le sport choisi pour ce projet est le basketball et sa ligue, la NBA (*National Basketball Association*). Quand il souhaite faire le grand saut, il s'inscrit à la « draft ».

La draft est un évènement annuel où sont réunis le commissionnaire de la ligue et les dirigeants des trente équipes. Chaque équipe va sélectionner, selon un ordre précis, un joueur s'ayant inscrit à la draft. Au total, seulement 60 joueurs rejoignent la ligue chaque année sur environ 3500 inscrits.

Les 30 premiers journées draftés lors de la NBA Draft de 2019

L'objectif de ce projet est de déterminer si un universitaire peut potentiellement rejoindre la NBA suivant ses statistiques et ses performances dans la compétition NCAA.

Pour répondre, on réalisera une première phase de « web scrapping » des statistiques des joueurs. La deuxième phase consistera à l'analyse des statistiques et la troisième phase sera consacrée à trouver le meilleur modèle de classification de Machine Learning.

2. Collection de données

2.1. Extraction des données

La collecte des données était assez complexe car il n'existe pas de datasets répondant à cette problématique. Il existe une compétition Kaggle organisée par la NCAA et Google Cloud mais le dataset, bien que fourni, donne des informations que sur les salles de sport, l'affluence et sur les matchs.

Le choix s'est porté sur la référence en statistiques, Basketball Reference. Cependant, ce site ne met pas à disposition une API. L'extraction de la donnée a été faite par *Beautiful Soup* complété par une API trouvé sur Github.

2.2. Nettoyage des données

Un autre problème encouru est le mapping entre les joueurs NCAA et les joueurs draftés. En effet, la base de données des joueurs NBA a été faite des années avant celle des joueurs universitaires. De ce fait, les joueurs se trouvant à la fois dans la base de données NBA et NCAA ne possèdent pas le même « id ». Pour créer dans notre dataset la colonne « y » à prédire (« drafté » : 0 non, 1 oui), il a fallu faire une jointure entre les noms des joueurs qui a résulté un nouveau problème : les joueurs homonymes.

La présence de données dupliquées devait être traitée. Prenons le cas d'un joueur X qui décide de ne s'inscrire à la draft qu'au bout de sa quatrième année (soit à la fin du cursus universitaire). Le joueur X possèdera dans le dataset 8 lignes ! En effet, il aura les 4 lignes dédiées à ses performances sur ses 4 années puis il aura des lignes appelées « carrière » qui est la somme des statistiques après chaque année :

1ère ligne carrière : Somme des stats de la première année universitaire
2ème ligne carrière : Somme des stats de la 1ère et 2ème années universitaires
3ème ligne carrière : Somme des stats de la 1ère, 2ème et 3ème année universitaire
4ème ligne carrière : Somme des stats de la 1ère, 2ème, 3èmes et 4èmes années universitaires

Pour la suite du projet, un joueur universitaire sera assigné à ses statistiques en carrière de sa dernière année universitaire.

Dernier problème, certaines métriques que l'on appelle « stats avancées » ne sont pas présentes pour tous les joueurs et notamment les joueurs les plus méconnus. L'une des stratégies serait de remplacer le NaN par la moyenne mais pour obtenir une meilleure analyse, il est possible de calculer ces métriques à l'aide de d'autres statistiques.

Exemple du « true shooting pourcentage » ou « pourcentage au tir réel » qui représente l'efficacité réelle d'un tireur. Cette métrique peut être utilisé à l'aide de PTS (moyenne de points marquées par match), FGA (moyenne de tirs tentées par match) et le FTA (moyenne de lancer francs tentées par match)

$$TS\% = \frac{PTS}{2(FGA+(0.44\times FTA))} \times 100\%$$