PROJET 8

La course à l'Europe fait vibrer chaque saison notre cœur de fan de foot dans l'espoir de voir notre club de cœur y figurer. Ces dernières saisons, le Montpellier Hérault Sport Club est toujours bien placé mais jamais récompensé, est-ce que cette saison est la bonne ?

Le MHSC européen ?

Table des matières

Introduction

Chaque année, le débat est là. QUI va être européen à la fin de la saison ? Quelles sont les clubs qui ont les moyens d'avoir cet objectif en tête ? Va-t-on assister à des surprises ? Ces courses à la Ligue des champions, à la Ligue Europa, sont souvent épiques et pleines de suspenses. Mais peut-on distinguer les clubs qui vont arriver à se qualifier à la fin du mercato ? C'est ce que nous allons voir dans chacun des 5 premiers championnats européens et nous focaliserons sur la Ligue 1 avec le Montpellier Hérault Sport Club.
Le MHSC est un club fondé par l'excentrique Louis Nicollin ayant accueilli de nombreux illustres joueurs comme Valderrama, Laurent Blanc ou Eric Cantona, et remporté de plusieurs titres dont un titre de champion de France en 2012. Aujourd'hui, il est dans une situation où ils répondent présents mais ils s'arrêtent toujours aux portes de l'Europe à la fin de la saison. Est-ce que ce sera encore le cas cette année ?

Pour réaliser ce projet, nous allons nous baser sur 4 datasets provenant de 3 sites :

- Understat
- Transfermarkt
- Données Fifa

Source de données

UNDERSTAT

Understat est un site ayant développé des statistiques avancées autour du football dont les « Expected Goals ». Ces statistiques sont très utiles, elles permettent de donner une autre vision sur la réalité d'un match parfois injuste ou d'une saison. Sur le long terme, celles-ci représentent un nouvel outil précieux entre les mains des entraineurs, de managers,… Mais elles le sont également pour le public dans le but d'avoir une meilleure compréhension des événements.

Ces données suivent différents modèles statistiques notamment une régression logistique pour les Expected Goals par exemple.

TRANSFERMARKT

Transfermarkt est un site qui rassemble l'ensemble de l'actualité footballistique (compétition, mercato,…) et diverses informations autour du football.

Ce dernier est surtout connu pour avoir créé un système déterminant la valeur marchande des joueurs. Plus ou moins fiable, cela représente néanmoins un bon critère pour avoir une idée sur la qualité du joueur.

SOFIFA

Sofifa est le site abritant l'ensemble de la base de données utilisée par le jeu FIFA 07 à FIFA 21. Ces dernières sont les statistiques propres aux joueurs dans le jeu essayant d'être le plus proche possible de la réalité. Même si des erreurs reste possible, cela reste un bon indicatif sur les capacités du joueur.

Variables

Résultats de la Ligue 1

- **CLASSEMENT**
- **MATCHS GAGNES**
- **MATCHS NULS**
- **MATCHS PERDUS**
- **BUTS MARQUES**
- **BUTS ENCAISSES**
- **EXPECTED GOALS (peut être traduits par « nombre de buts attendus »):** Buts qui auraient du être marqués en fonction de la valeur de l'action (un penalty a une valeur 0.76 expected goals parce qu'un penalty est très souvent converti alors qu'un tir au milieu du terrain aura une valeur beaucoup plus basse, action beaucoup difficile à convertir). Cette variable doit répondre à la question : quelles sont les chances que le joueur marque dans cette situation (position du joueur / du gardien, distance, fatigue,….) ?
- **EXPECTED GOALS AGAINST (peut être traduits par « nombre de buts encaissés attendus ») :** Buts qui auraient du être encaissés en fonction de la valeur de l'action
- **EXPECTED POINTS (peut être traduits par « nombre de buts points ») :** Points qui auraient dû être remportés (l'équipe a eu de la réussite et n'aurait pas dû remporter ce match)
- **CLUB DU JOUEUR**
- **ANNEES**

Nous garderons le nom des variables originales pour plus de simplicité.

Valeurs marchandes des effectifs

- **ÂGE :** moyenne des âges des joueurs en fonction de leur rôle tactique (gardien, défense, milieu, attaque)

- **VALEUR MARCHANDE (SOMME) :** valeurs marchandes cumulées des joueurs en fonction de leur rôle tactique
- **VALEUR MARCHANDE (MOYENNE) :** valeurs marchandes moyenne des joueurs en fonction de leur rôle tactique
- **VALEUR MARCHANDE MAXIMALE :** valeur marchande maximale en fonction de leur rôle tactique
- **CLUB DU JOUEUR**
- **ANNEES**

Statistiques des joueurs

- **NOMBRES DE MATCHS JOUES**
- **NOMBRES DE MINUTES JOUEES**
- **BUTS MARQUES**
- **EXPECTED GOALS**
- **PASSES DECISIVES**
- **EXPECTED ASSISTS :** nombre de passes décisives que le joueur aurait dû faire
- **NOMBRE DE TIRS**
- **NOMBRE PASSES CLES :** nombre de passes précédant un tir
- **NOMBRE DE CARTONS JAUNES**
- **NOMBRE DE CARTONS ROUGE**
- **POSITION**
- **NOMBRE DE BUTS SANS LES PENALTY**
- **EXPECTED GOALS (SANS LES PENALTY)**
- **ANNEES**

Données FIFA

- **NOM DU JOUEUR**
- **CLUB DU JOUEUR**
- **POSTE**
- **TAILLE**

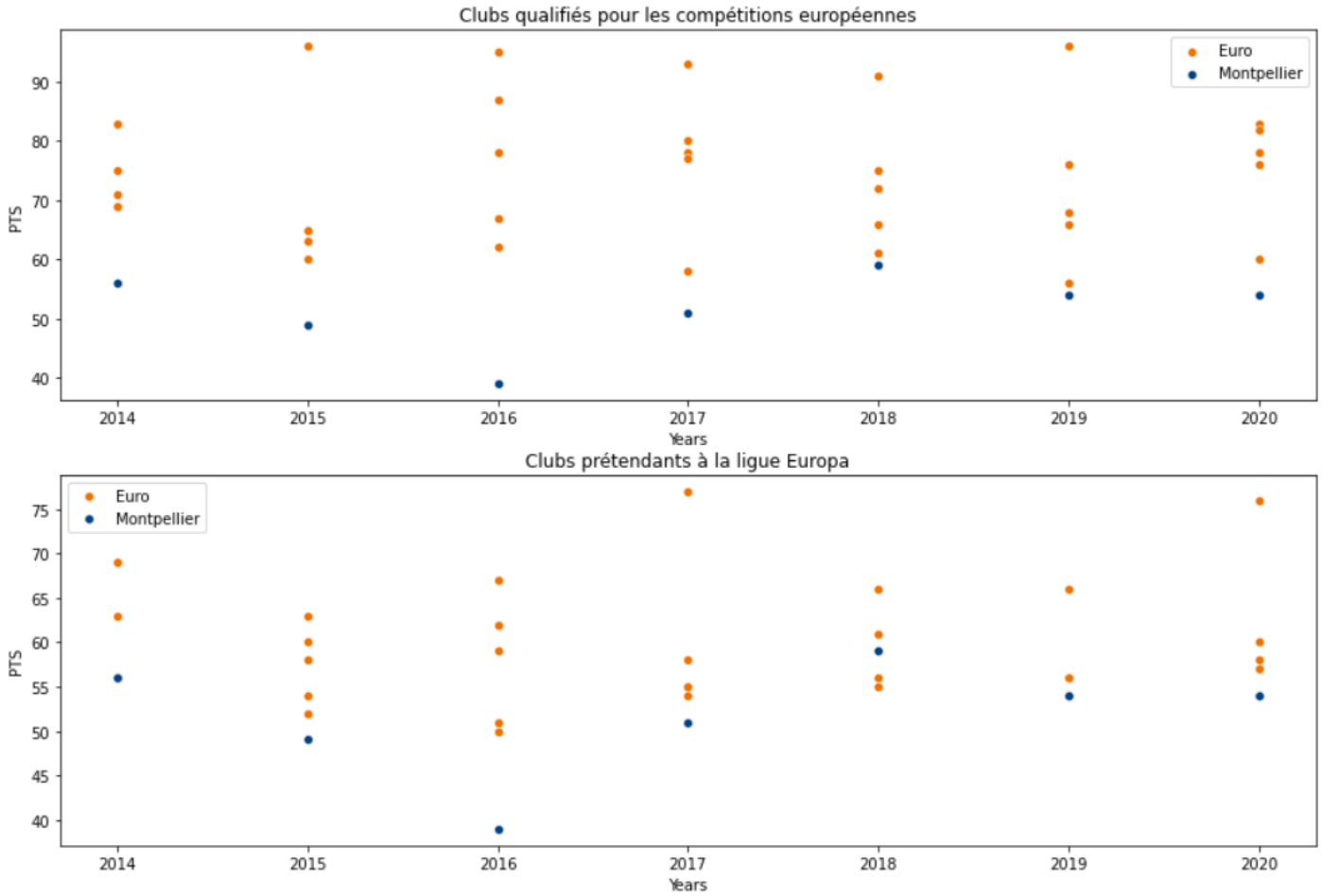
- **ACCELERATION :** la capacité à accélérer du joueur
- **VITESSE :** rapidité du joueur
- **POTENTIEL OFFENSIF :** compilation des variables offensives (centres, finition, tête, passes courtes, reprises de volée)
- **TECHNIQUE :** compilation des variables techniques (dribbles, effet, précision coup francs, passes longues, contrôle de balle)
- **MOUVEMENT :** compilation des variables de mouvement (accélération, vitesse, agilité, réactivité, équilibre)
- **DEFENSE :** compilation des variables défensives (conscience défensives, tacle, tacle glissé)
- **GARDIEN :** compilation des variables gardien (plongeon, jeu à la main, jeu au pied, placement, réflexes)
- **ANNEES**

ANALYSE DESCRIPTIVE

Dans notre analyse le critère Le plus important est le classement. C'est celui-ci qui détermine si l'équipe accède à une compétition européenne, et il est défini en fonction du nombre de points marqués.

Les équipes entre la 1ère et la 3ème place du championnat se qualifient pour la ligue des champions (barrages ou phases de groupe sous condition pour la 3ème place), la 4ème est qualificative pour la ligue europa. Enfin, la 5ème va permettre de se qualifier pour les tours préliminaires pour la nouvelle compétition (la Ligue Europa Conférence, la C4).

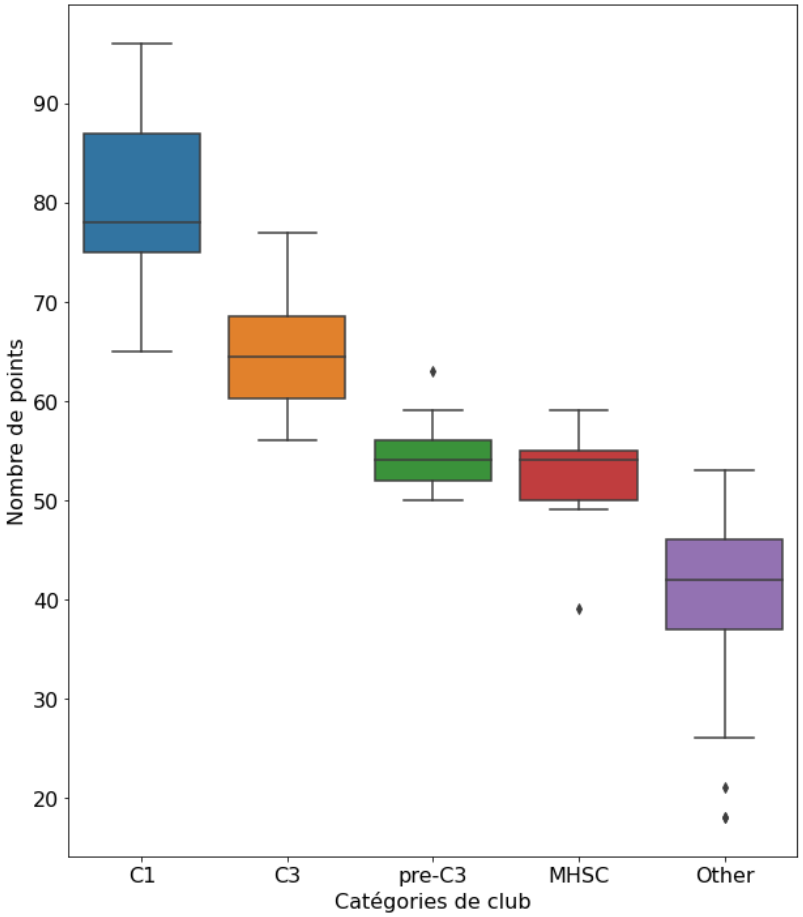
---------- Comparaison du nombre de points entre les clubs européens ----------



Montpellier est un club se situant la plupart du temps dans le ventre mou du

championnat depuis le titre en 2012. Après avoir connu une chute dans le classement jusqu'en 2016, l'équipe est revenue au contact des clubs prétendant à l'Europe.

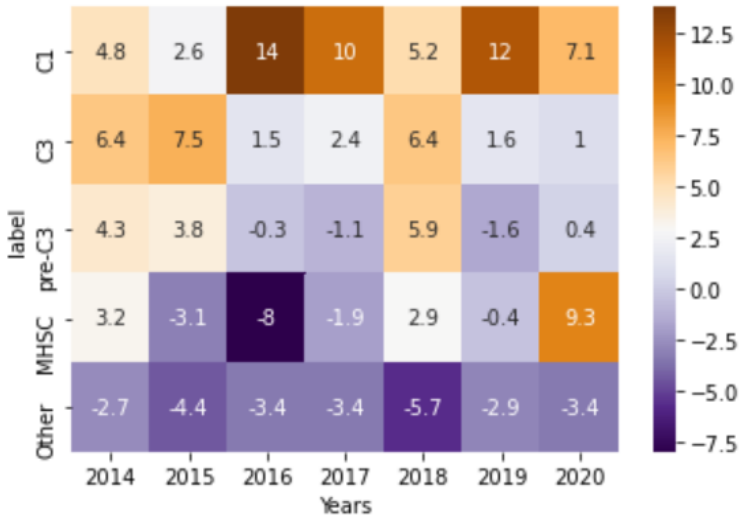
Comme on peut le voir, Montpellier est effectivement revenu dans le peloton aux prétendants à l'Europe. Cependant, il faut prendre en compte que faire partie de ce groupe n'est pas suffisant pour être européen, et la concurrence est très féroce.



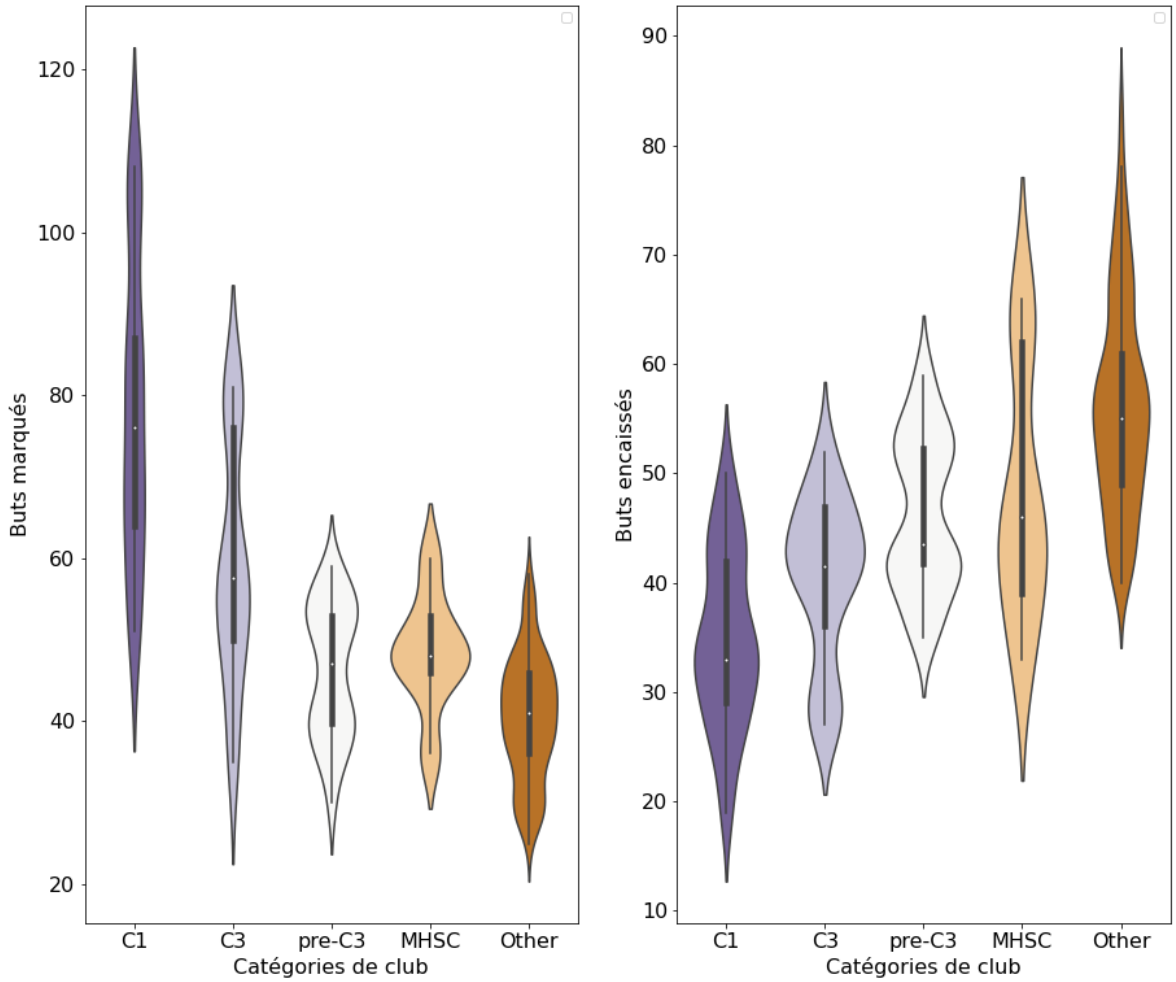
En outre, on peut constater que le gap entre les clubs en ligue des champions et ceux en Ligue Europa est très important. La différence de qualité est conséquente entre ces groupes et un gros palier est à franchir pour parvenir à ce niveau.

Différences entre le nombre de points obtenus et le nombre

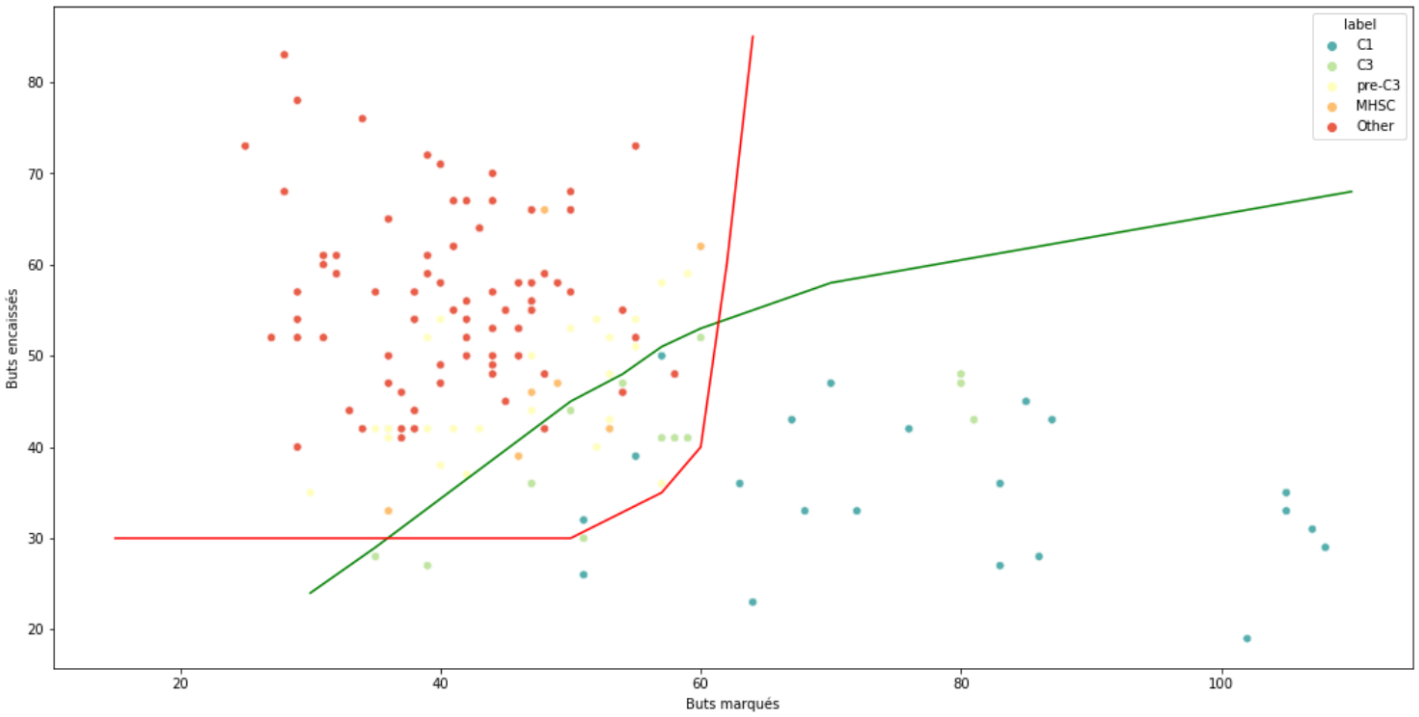
Depuis quelques années sont apparues des stats avancées, elles permettent de donner des éléments d'analyse supplémentaires. Ici, les expected points nous présentes une certaine corrélation suivant le classement du club. En effet, pour accéder au sommet du classement, il est nécessaire d'avoir de la réussite, c'est



celle-ci qui peut te faire gagner dans les moments importants. Cependant, ce n'est pas que de la chance, par exemple plus les joueurs offensifs seront capables de finir des situations dangereuses, plus ils seront capables de marquer des buts dans des positions compliquées.



Pour être européen, il est important de maintenir un haut niveau de performance sur l'ensemble de la saison, cela se caractérise par une attaque performante et / ou une défense performante. Dans le but de marquer un maximum de points, une attaque performante permet de se sortir de situations compliquées ou se mettre à l'abri d'éventuels retours de l'adversaire. Cependant, sans au minimum certaines garanties défensives, même avec une bonne attaque, l'équipe n'est jamais à l'abri d'une surprise.
Pour confirmer l'hypothèse qu'une défense avec un minimum de garantie est indispensable si l'on a des ambitions plus qu'une bonne attaque, voici un scatterplot avec les buts marqués en fonction des buts encaissés.



___ = limite des équipes européennes ___ = limite des équipes non européennes

- C1 (1ère à 3ème place)
- C3 (4ème et 5ème place)
- Pre C3 (6ème à 9ème place)
- MHSC (classement de Montpellier)
- Other (10ème à 20ème place)

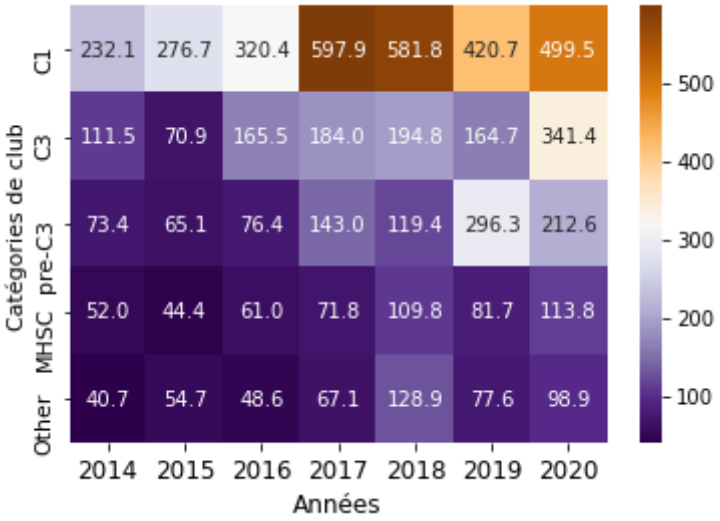
Ce graphique nous montre qu'il existe des performances minimales et maximales pour atteindre un classement. Par exemple, si l'on a une bonne défense (autour de 35 buts encaissés) mais que l'on ne dépasse pas non les 35 buts marqués alors ce sera insuffisant pour être européen. L'inverse avec la ligne verte existe également. Cependant, si l'on se trouve entre les 2 courbes, alors on se trouve dans une zone d'incertitude. Le classement de l'équipe va dépendre de nombreux autres facteurs extérieurs :

- La concurrence (si le nombre de d'équipes performances est supérieur au nombre de places disponible alors des performances permettant une qualification auparavant ne le permettront pas ici)
- La réussite

Par contre, il faut prendre en compte que notre échantillon ne se base seulement sur les 7 dernières années, il est possible que les performances minimales soient

moins importantes. Un échantillon allant jusqu'à la 1ère année de l'ère QSI (rachat du PSG) voire 1 ou 2 ans avant, pourrait nous donner un dataset beaucoup plus représentatif.

Valeurs marchandes totales moyens en fonction des classements des équipes



Ici, nous allons nous intéresser aux valeurs marchandes. Dans un premier temps, il est très intéressant de constater une hausse générale des valeurs marchandes des effectifs, notamment à partir de 2016. Ensuite, la logique est globalement respectée, sauf en 2019 où la valeur moyenne des clubs du groupe « pre-C3 » est bien supérieur aux clubs qualifiés en C3. En effet, tout d'abord, précisons que cette saison était très particulière. Ayant été le seul championnat à n'avoir pas été au bout de ses 38 journées, cela a fait le bonheur et le malheur de nombreux clubs comme par exemple Lyon qui a manqué l'Europe pour la 1ère fois depuis 23 ans ou Reims qui, à la surprise générale, retrouve l'Europe 57 ans après. Ainsi, la valeur d'un effectif ne garantit pas les résultats et des mauvaises surprises peuvent être aux rendez-vous. Cependant, celles-ci sont imprévisibles à l'aube d'une nouvelle saison, seules les données concernant l'effectif peuvent des indications concernant des résultats futurs.

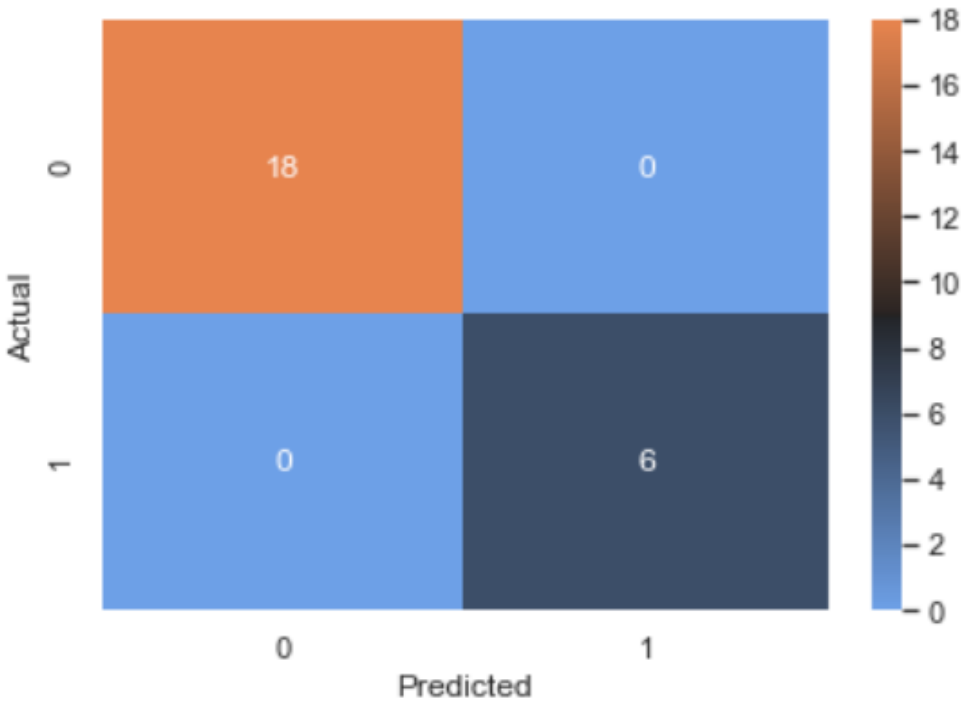
MODELISATION

Notre modélisation sera une régression logistique.
Il y a un critère qui réduit considérablement notre champ de recherche, nous avons besoin de paramètres. Sans cela, comment savoir dans quel état se trouve le club ? Une équipe à la fin de la saison, ou au début de la saison suivante, ce sont 2 choses complétement différentes. Ainsi, avoir des indications sur les clubs en début de saison est indispensable.
Et la régression logistique répond à ce critère.

Le paramètre que nous utiliserons est le Y d'un fonction affine résultant d'une corrélation entre les valeurs marchandes des effectifs des clubs et F1 (premier axe factoriel de notre analyse en composante principale, agglomération de nombreuses variables comme les points, les buts marqués ou encaissés,…). Nous obtiendrons un coefficient démontrant la capacité du club à faire de bonnes performances.

Le fonctionnement du modèle est le suivant, il va examiner chaque valeur et regarder les labels correspondants. Est-ce que pour ce montant nous avons un résultat positif ou négatif ? Pour cela, nous allons séparer notre dataset en 2 parties, 50% pour entrainer notre modèle et les 50 autres pourcents pour le tester.

L'indicateur de performance du modèle est le nombre de faux positifs ou négatifs grâce à une heatmap.



Comme on peut le voir, notre modèle n’a pas fait d’erreur avec nos données test. Il n’y a pas de faux positifs ou faux négatifs.

CONCLUSION

Les résultats sont représentés sous la forme de pourcentage, quel est le pourcentage que le club soit européen à la fin de la saison ? Ces derniers suivent une distribution croissante suivant la valeur réelle des effectifs. Cependant, ce n'est pas un classement. En fonction de celles-ci, notre modèle va également nous sortir un pourcentage de non-qualification (1 – pourcentage de qualification). Mais ce dernier ne prend pas en compte les facteurs externes. En effet, nous l'avons déjà vu auparavant, il faut d'autres paramètres si l'on veut que le modèle étoffe le nombre de facteurs explicatifs pris en compte.

Dans ce projet, nous sommes passé par toutes les étapes d'un projet data classique :

Acquisition des données (scrapping) -> nettoyage des données -> Analyse descriptive -> Modélisation -> Communication des résultats

Il nous a également apporté de bons résultats.

Cependant, des améliorations peuvent être apportées. Les facteurs explicatifs pris en compte et le modèle utilisé ne sont pas forcément optimaux. Ainsi, notre projet est imparfait, des axes d'amélioration sont présent. A méditer !

Bibliographie

Api Understat :

https://understat.readthedocs.io/en/latest/classes/understat.html#the-functions

Expected goals / goals against :

https://theanalyst.com/na/2021/06/what-are-expected-goals-xg/)