

Mathis OUDJANE

Nao MAUSSERVEY

Jonathan HAYOT

M1 Dev Manager Full Stack



TP Databricks Groupe 2 :

Lien du github : https://github.com/moudjane/tp_databricks

Professeur : Alexandre BERGÈRE

Table des matières :

Introduction.....	3
Structure et présentation des Notebooks.....	3
Présentation de notre Azure :.....	4
ds-bronze : Données brutes.....	4
ds-silver : Données nettoyées.....	6
ds-gold (delta-tables) : Données nettoyées.....	7
MCD de notre table gold :.....	10
Analyse de l'impact des performances des clubs sur la valeur des joueurs.....	11
Données Utilisées.....	11
Traitements Effectués.....	12
Visualisation des Résultats.....	14
Interprétation.....	15
Conclusion.....	16

Introduction

Dans le cadre de ce projet pour le cours de Big Data, nous avons exploré une hypothèse clé : **les performances des clubs influencent-elles la valeur marchande des joueurs affiliés ?** À travers l'utilisation de données réelles et d'une architecture de données moderne, nous avons cherché à établir une corrélation entre le succès sportif des clubs, mesuré par leur taux de victoire et la valorisation économique de leurs joueurs. Ce travail s'inscrit dans une démarche analytique rigoureuse, combinant nettoyage des données, modélisation en étoile, et visualisation des résultats pour apporter des réponses précises et exploitables à cette problématique.

Structure et présentation des Notebooks

Chaque étape du projet est documentée dans des notebooks dédiés, assurant la traçabilité des traitements appliqués aux données.

Le script `notebook_players.py` nettoie les données brutes des joueurs pour produire des fichiers Silver structurés. De la même façon, `notebook_games.py` traite les données des matchs, tandis que `notebook_player_valuations.py` s'occupe des valorisations financières des joueurs.

Enfin, `notebook_gold.py` transforme les données Silver en tables Delta dans la couche Gold et réalise l'analyse principale, examinant la relation entre les performances des clubs et la valeur des joueurs.

Présentation de notre Azure :

Nom	État	Dernière modification	Niveau d'accès anonyme	État du bali
ds-bronze	Actif	21/11/2024 16:51:29	Privé	Disponible
ds-bronze	Actif	14/12/2024 22:49:41	Privé	Disponible
ds-bronze	Actif	09/12/2024 23:20:50	Privé	Disponible
ds-bronze	Actif	09/12/2024 23:47:22	Privé	Disponible
ds-bronze	Actif	21/11/2024 16:57:30	Privé	Disponible

Nous avons trois folders bronze, silver et delta-tables (qui équivaut à gold)

Bronze = ds-bronze, Silver = ds-silver, Gold = delta-tables.

ds-bronze : Données brutes

Nom	Modifié	Niveau d'accès	État de l'archive	Type d'objet blob	Taille	État du bali
_sacventemplobidrs	09/12/2024 23:20:51	Privé	Non archivé	Objet blob	-	Disponible
games	09/12/2024 22:49:41	Privé	Non archivé	Objet blob	-	Disponible
player_valuations	09/12/2024 23:47:22	Privé	Non archivé	Objet blob	-	Disponible
players	09/12/2024 23:47:22	Privé	Non archivé	Objet blob	-	Disponible

La première couche de notre pipeline, appelée DS-Bronze, contient les fichiers bruts tels qu'ils ont été collectés depuis les différentes sources. Ces données n'ont subi aucune modification et sont stockées dans leur état original, incluant les erreurs ou les incohérences potentielles. Trois fichiers principaux sont présents dans cette couche : players.csv, player_valuations.csv et games.csv.

Le fichier players.csv contient les informations de base sur les joueurs, comme leur identifiant unique (player_id), leur nom (name) et le club auquel ils sont affiliés actuellement (current_club_id). Ce fichier constitue la base pour relier les joueurs aux clubs et analyser

leur évolution. Le fichier `player_valuations.csv` contient des informations sur les valorisations financières des joueurs. Chaque ligne correspond à une évaluation d'un joueur, identifiée par son `player_id`, et associe une valeur financière à une période donnée (`market_value_in_eur`). Enfin, le fichier `games.csv` regroupe les informations relatives aux matchs joués. Il inclut les scores des matchs (`home_club_goals` et `away_club_goals`), les clubs en compétition (`home_club_id` et `away_club_id`), ainsi que des informations contextuelles comme le nombre de spectateurs présents au match (`attendance`).

Ces fichiers bruts sont essentiels pour capturer l'état initial des données avant tout traitement. Bien qu'ils soient riches en informations, ils nécessitent un nettoyage pour corriger les erreurs et standardiser les formats avant toute analyse.

ds-silver : Données nettoyées

Microsoft Azure

Accueil >

ds-silver ...
Conteneur

Rechercher

Charger Ajouter un répertoire Actualiser

Méthode d'authentification : Clé d'accès ([Basculer vers le compte](#))
Emplacement : ds-silver

Rechercher les objets blobs par préfixe (respect de la casse)

Nom
<input type="checkbox"/> _\$azuretmpfolder\$
<input type="checkbox"/> games_cleaned
<input type="checkbox"/> games_cleaned_temp
<input type="checkbox"/> player_valuations_cleaned
<input type="checkbox"/> player_valuations_cleaned_temp
<input type="checkbox"/> players_cleaned
<input type="checkbox"/> players_cleaned_temp

Après la collecte des données brutes, celles-ci ont été transformées et nettoyées dans la couche DS-Silver. L'objectif de cette étape est d'assurer que les données soient cohérentes, standardisées et prêtes pour une analyse approfondie. Les fichiers nettoyés dans cette couche ont subi plusieurs étapes de traitement, notamment la suppression des valeurs manquantes, la correction des anomalies et la standardisation des formats. Ce processus garantit que seules les données valides et fiables sont utilisées dans les analyses suivantes.

La version nettoyée de `players.csv`, renommée `players_cleaned.csv`, contient uniquement des informations vérifiées sur les joueurs. Les colonnes inutiles ou corrompues ont été éliminées, ce qui rend les données plus pertinentes pour nos besoins analytiques. De manière similaire, le fichier `player_valuations_cleaned.csv` est une version transformée de `player_valuations.csv`. Les anomalies, telles que des valeurs financières incohérentes ou

absentes, ont été corrigées pour garantir une cohérence entre les joueurs et leurs valorisations. Enfin, le fichier games_cleaned.csv offre une version nettoyée des données des matchs. Les scores ont été vérifiés pour assurer leur cohérence, et les informations contextuelles comme l'assistance ou les formations d'équipe ont été préservées pour des analyses complémentaires.

Dans cette couche, les données ont été rendues exploitables tout en conservant leur richesse d'informations. Cela constitue une base solide pour la structuration finale dans la couche suivante.

ds-gold (delta-tables) : Données nettoyées

Microsoft Azure

Rechercher dans les ressources, services et docu

Accueil > dlkefrei91320 | Conteneurs >

delta-tables

Conteneur

Rechercher

Charger

Ajouter un répertoire

Actualiser

Renommer

Supprimer

Modifier le niveau

Acquérir le bail

Résilier le bail

Envoyer de

Vue d'ensemble

Diagnostiquer et résoudre les problèmes

Contrôle d'accès (IAM)

Paramètres

Jetons d'accès partagé

Gérer l'ACL

Stratégie d'accès

Propriétés

Métadonnées

Méthode d'authentification : Clé d'accès (Basculer vers le compte d'utilisateur Microsoft Entra)

Emplacement : delta-tables

Rechercher les objets blobs par préfixe (respect de la casse)

Nom	Modifié
<input type="checkbox"/> _azuretmpfolders\$	14/12/2024 22:50:16
<input type="checkbox"/> games	14/12/2024 22:53:01
<input type="checkbox"/> main	14/12/2024 22:49:52
<input type="checkbox"/> player_valuations	14/12/2024 22:52:22


Méthode d'authentification : Clé d'accès ([Basculer vers le compte d'utilisateur Microsoft Entra](#))


Emplacement : [delta-tables](#) / games


Rechercher les objets blobs par préfixe (respect de la casse)


Nom

☐  [..]

☐  _delta_log

☐  part-00000-877828d1-f0bd-4a23-838c-7db2d79da03f-c000.snappy.parquet

☐  part-00001-11bb5626-87f2-4c98-bece-b9cf2ea08fdd-c000.snappy.parquet

☐  part-00002-e2b00ae6-f7ca-4ae7-8256-111b3689bd29-c000.snappy.parquet


Méthode d'authentification : Clé d'accès ([Basculer vers le compte d'utilisateur Microsoft Entra](#))


Emplacement : [delta-tables](#) / main


Rechercher les objets blobs par préfixe (respect de la casse)

Nom

☐  [..]

☐  _delta_log







☐  part-00000-679247b7-c14c-448d-9433-ef6c95598319-c000.snappy.parquet

☐  part-00001-6b762028-077b-4d46-a141-cf0be68c5f74-c000.snappy.parquet

Méthode d'authentification : Clé d'accès ([Basculer vers le compte d'utilisateur Microsoft Entra](#))

Emplacement : [delta-tables](#) / [player_valuations](#)

Rechercher les objets blobs par préfixe (respect de la casse)

Nom
<input type="checkbox"/>  [..]
<input type="checkbox"/>  _delta_log
<input type="checkbox"/>  part-00000-6a67863f-ffa8-4a07-ac5a-45e1721c5636-c000.snappy.parquet
<input type="checkbox"/>  part-00001-46324f36-4180-4506-bdbc-dd482203f8cb-c000.snappy.parquet
<input type="checkbox"/>  part-00002-02a0b72f-2495-46a4-8a70-50f90596dad5-c000.snappy.parquet
<input type="checkbox"/>  part-00003-12e6c95b-6888-4028-87ca-95087737affe-c000.snappy.parquet

La couche DS-Gold, également appelée "delta-tables", représente l'étape finale de notre pipeline. Dans cette couche, les données nettoyées ont été structurées en tables Delta au format Parquet pour une analyse optimisée. Ces tables sont organisées selon un schéma en étoile, ce qui facilite les requêtes analytiques en connectant une table de faits à plusieurs tables dimensionnelles.

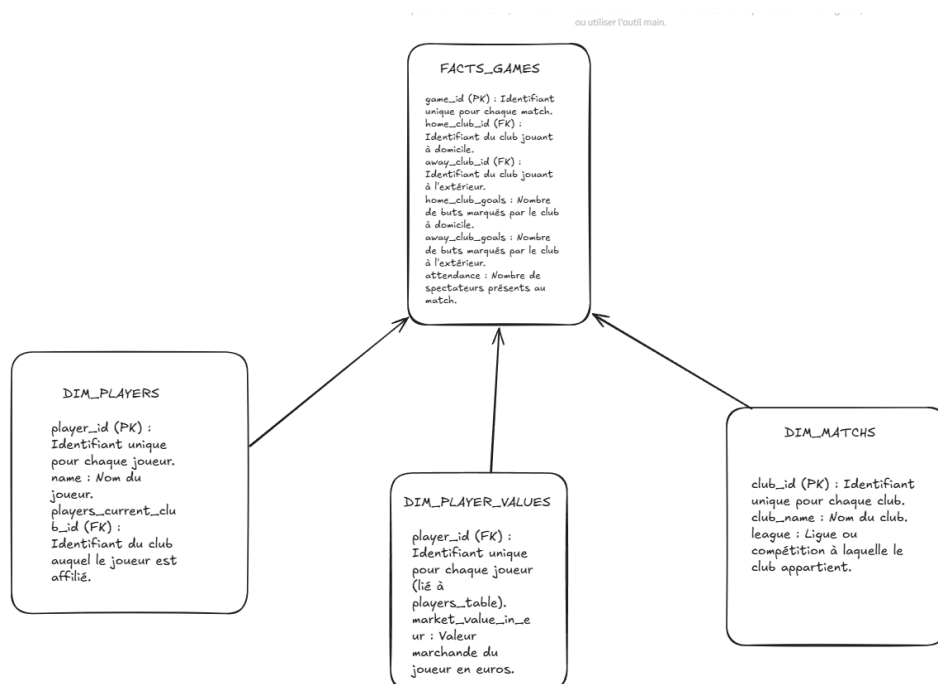
La table centrale, ou table de faits, est `games`, qui contient les détails des matchs joués. Chaque ligne représente un match unique, identifié par son `game_id`. Les colonnes principales incluent les identifiants des clubs en compétition (`home_club_id` et `away_club_id`), les scores respectifs (`home_club_goals` et `away_club_goals`), et des métriques supplémentaires comme le nombre de spectateurs présents au match (`attendance`). Cette table est au cœur de notre analyse, car elle permet de relier les performances des clubs à d'autres dimensions.

Autour de cette table centrale, nous avons deux tables dimensionnelles. La première, `main`, regroupe les informations sur les joueurs, en incluant leur identifiant unique (`player_id`), leur nom (`name`) et leur affiliation actuelle à un club (`players_current_club_id`). Cette table relie les joueurs à leurs clubs respectifs. La seconde table dimensionnelle, `player_valuations`, fournit des informations financières sur les joueurs. Elle associe chaque joueur (via `player_id`) à sa valeur marchande (`market_value_in_eur`). Ces deux dimensions enrichissent

la table de faits et permettent de croiser les données de performance avec celles de valorisation.

Les données dans la couche DS-Gold ont été partitionnées et optimisées pour l'analyse en utilisant les fonctionnalités avancées de Delta Lake. Cela offre une gestion efficace des mises à jour et une réduction significative des temps de requêtes.

MCD de notre table gold :



Analyse de l'impact des performances des clubs sur la valeur des joueurs

Dans le cadre de cette analyse, nous explorons une question fondamentale dans le monde du football professionnel : les performances des clubs influencent-elles la valeur des joueurs qui y sont affiliés ? Cette hypothèse repose sur l'idée que les clubs ayant de meilleures performances sur le terrain ont tendance à attirer ou à développer des talents plus coûteux, ou que ces performances augmentent la visibilité et la valorisation des joueurs. L'objectif principal de cette étude est donc d'identifier et de quantifier une éventuelle corrélation entre le taux de victoire (« win rate ») des clubs et la valeur financière moyenne des joueurs appartenant à ces clubs. Cette analyse s'appuie sur des données structurées provenant d'une architecture en étoile, qui permet d'organiser les informations de manière efficace pour les requêtes analytiques. Le traitement des données, les calculs effectués et les interprétations des résultats sont détaillés dans ce document afin de répondre rigoureusement à cette hypothèse.

Données Utilisées

Les données utilisées dans cette analyse proviennent de trois sources principales. La première est une table appelée "games_table", qui contient l'historique complet des matchs joués par les clubs. Cette table inclut des informations critiques telles que les scores des équipes à domicile et à l'extérieur, les identifiants des clubs impliqués dans chaque match, ainsi que des métriques supplémentaires comme l'assistance lors des matchs. Parmi les colonnes principales, on peut citer "game_id", qui identifie chaque match de manière unique, "home_club_id" et "away_club_id" pour les équipes en compétition, et "home_club_goals" et "away_club_goals" pour les scores respectifs. Ces données permettent d'évaluer les performances des clubs en termes de victoires ou de défaites.

La deuxième source est une table appelée "main_table", également référencée sous le nom de "players" dans cette analyse. Cette table décrit les joueurs et les clubs auxquels ils sont affiliés. Elle inclut des informations importantes comme "player_id", qui identifie chaque joueur de manière unique, "name", qui représente le nom du joueur, et "players_current_club_id", qui indique le club auquel le joueur appartient au moment de l'analyse. Ces données permettent de relier chaque joueur à son club et sont essentielles pour regrouper les informations à un niveau club.

Enfin, la troisième source est la table "player_valuations_table", qui contient des données financières sur la valeur des joueurs. Cette table fournit une perspective économique sur les talents, en incluant des colonnes telles que "player_id" pour établir le lien avec les joueurs et "market_value_in_eur" qui indique la valeur marchande estimée de chaque joueur en euros. Ces informations sont cruciales pour évaluer l'impact des performances sportives sur les valorisations financières.

L'ensemble des données suit une architecture en étoile, où la table centrale "games_table" agit comme table de faits, et les tables "main_table" et "player_valuations_table" servent de dimensions. Cette structure permet de relier efficacement les matchs, les joueurs et leurs valeurs, facilitant ainsi les requêtes analytiques complexes nécessaires pour cette étude.

Traitements Effectués

```
# Step 1: Calculate win rate for each club
winrate_df = games_df \
    .groupBy(col("games.home_club_id").alias("club_id")) \
    .agg(
        (sum(when(col("games.home_club_goals") > col("games.away_club_goals"), 1).otherwise(0)) / count("games.game_id")).alias("win_rate"),
        count("games.game_id").alias("total_games")
    )
```

Pour répondre à l'hypothèse posée, plusieurs étapes de traitement des données ont été effectuées. La première étape consiste à calculer le taux de victoire (« win rate ») pour chaque club. Ce calcul s'appuie exclusivement sur les données issues de "games_table". Tout d'abord, nous avons filtré les matchs afin de comptabiliser uniquement les victoires à domicile, ce qui correspond aux cas où "home_club_goals" est strictement supérieur à "away_club_goals". Ensuite, nous avons calculé le nombre total de matchs joués par chaque club en tant qu'équipe à domicile. La formule utilisée pour le taux de victoire est la suivante : Win Rate = Nombre de victoires à domicile / Nombre total de matchs joués. Ce traitement permet d'obtenir deux métriques clés pour chaque club : le "win_rate", qui représente la proportion de victoires, et "total_games", qui indique le volume total de matchs joués. Ces indicateurs constituent une première mesure des performances des clubs.

```
# Step 2: Enrich players data with valuations
players_with_valuations_df = players_df \
    .join(player_valuations_df, "player_id", "inner")
```

La deuxième étape du traitement consiste à enrichir les données des joueurs avec leurs valorisations financières. Pour cela, nous avons effectué une jointure entre "main_table" et "player_valuations_table" à l'aide de la colonne commune "player_id". Cette jointure permet d'associer chaque joueur à sa valeur marchande actuelle, ce qui est essentiel pour calculer les moyennes au niveau des clubs. Par la suite, nous avons effectué une autre jointure entre les données des joueurs enrichies et la table des taux de victoire calculée à l'étape précédente. Cette étape relie les performances des clubs, en termes de "win_rate", à la valeur financière moyenne des joueurs qui leur sont affiliés. Enfin, une agrégation finale a été réalisée pour calculer deux métriques principales par club : "avg_player_value", qui représente la valeur moyenne des joueurs en euros, et "avg_win_rate", qui est le taux de victoire moyen du club. Ce processus permet de relier directement les performances sportives et les valorisations financières.

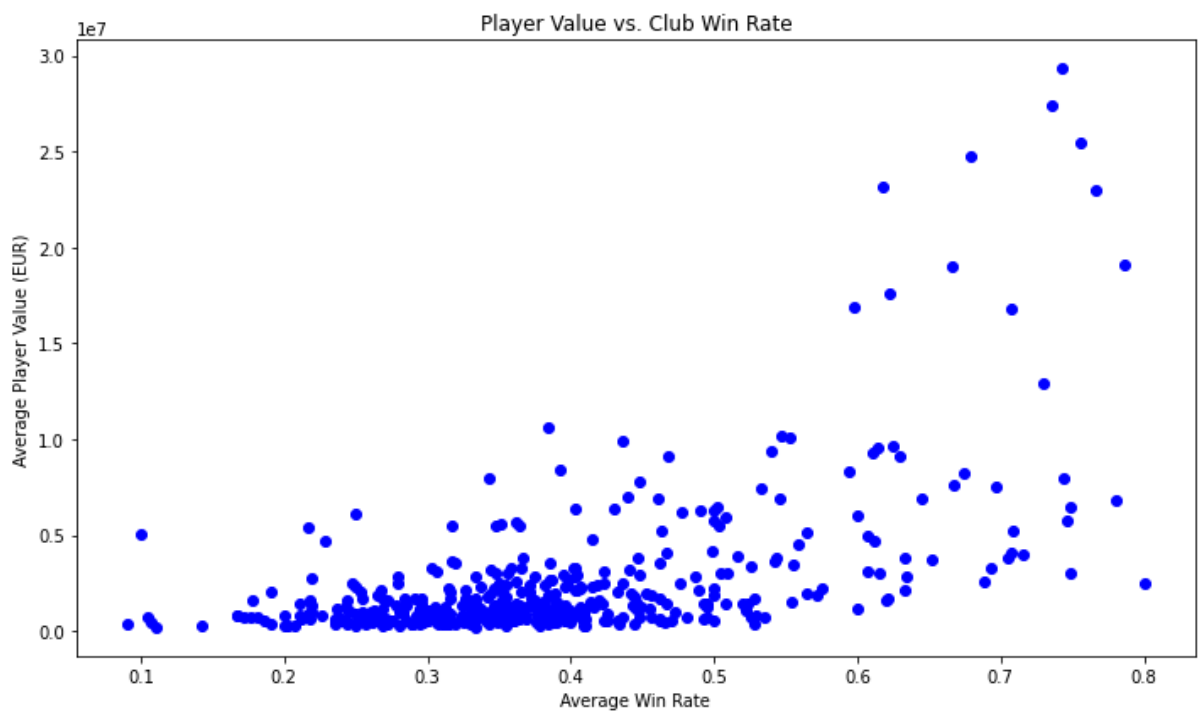
Visualisation des Résultats

```
# Step 3: Join win rate with player valuations
club_performance_df = winrate_df \
    .join(players_with_valuations_df, winrate_df["club_id"] == players_with_valuations_df["players_current_club_id"], "inner") \
    .groupBy("club_id") \
    .agg(
        avg("market_value_in_eur").alias("avg_player_value"),
        avg("win_rate").alias("avg_win_rate")
    )

# Display the results
club_performance_df.show()

# Step 4: Visualization
club_performance_pd = club_performance_df.toPandas()
plt.figure(figsize=(10, 6))
plt.scatter(club_performance_pd["avg_win_rate"], club_performance_pd["avg_player_value"], color="blue")
plt.xlabel("Average Win Rate")
plt.ylabel("Average Player Value (EUR)")
plt.title("Player Value vs. Club Win Rate")
plt.tight_layout()
plt.show()
```

Pour mieux comprendre les relations entre les performances des clubs et la valeur de leurs joueurs, nous avons généré un graphique en dispersion. Ce graphique utilise une échelle linéaire pour représenter deux variables principales : le taux de victoire moyen ("avg_win_rate") sur l'axe des abscisses et la valeur financière moyenne des joueurs ("avg_player_value") sur l'axe des ordonnées. Chaque point du graphique représente un club, offrant une vue d'ensemble des tendances dans les données.



L'analyse visuelle du graphique révèle une tendance positive entre les deux variables. Les clubs qui affichent un taux de victoire plus élevé tendent également à avoir des joueurs d'une valeur financière plus importante. Par exemple, les clubs situés dans la partie supérieure droite du graphique, avec un taux de victoire supérieur à 0,7, présentent des valeurs moyennes de joueurs dépassant fréquemment les 10 millions d'euros. En revanche, les clubs ayant un taux de victoire inférieur à 0,3 montrent des valeurs moyennes de joueurs significativement plus faibles. Ce graphique met également en évidence une certaine variabilité, notamment pour les clubs ayant des taux de victoire autour de 0,5, où l'on observe une large gamme de valeurs financières moyennes, allant de clubs avec des joueurs évalués à quelques centaines de milliers d'euros à ceux atteignant plusieurs millions.

Interprétation

L'analyse réalisée met en évidence une corrélation positive entre les performances des clubs et la valeur financière moyenne de leurs joueurs. Les résultats suggèrent

que les clubs qui gagnent fréquemment attirent ou développent des talents mieux valorisés sur le marché. Plusieurs facteurs pourraient expliquer cette tendance. D'une part, les clubs performants bénéficient souvent d'une meilleure visibilité médiatique, ce qui peut influencer la perception de la valeur de leurs joueurs. D'autre part, ces clubs ont généralement plus de ressources financières pour recruter des joueurs de haut niveau ou investir dans le développement de jeunes talents.

Cependant, il est important de noter que cette analyse présente certaines limites. Tout d'abord, les données utilisées ne permettent pas de capturer l'évolution temporelle des valeurs des joueurs, ce qui pourrait fournir des informations supplémentaires sur l'impact direct des performances sportives sur leur valorisation. Ensuite, bien que nous observions une corrélation, cela ne signifie pas nécessairement qu'il existe une relation causale. D'autres facteurs, tels que la popularité historique du club ou la qualité de la ligue dans laquelle il évolue, peuvent également jouer un rôle significatif.

Conclusion

En conclusion, cette étude confirme qu'il existe une relation positive entre le taux de victoire des clubs et la valeur financière moyenne de leurs joueurs. Les clubs les plus performants tendent à avoir des joueurs mieux valorisés, ce qui peut refléter une combinaison de facteurs sportifs, économiques et médiatiques. Ces résultats peuvent avoir des implications pratiques pour les clubs, notamment dans la gestion de leur stratégie de recrutement et de développement des talents. Par exemple, un club cherchant à augmenter sa visibilité sur le marché des transferts pourrait investir dans l'amélioration de ses performances sportives pour accroître la valorisation de ses joueurs.

Dans une perspective future, cette analyse pourrait être approfondie en intégrant des données temporelles pour examiner l'évolution des valeurs des joueurs en fonction des résultats des clubs sur plusieurs saisons. De plus, une comparaison

entre différentes ligues ou régions géographiques pourrait fournir des informations supplémentaires sur les dynamiques économiques et sportives du football professionnel.