Analyse, Modélisation et Prédiction dans le Football

Rapport de Projet

Groupe 15 : YAMEOGO, KAMBIRE, TRAORE, ZALFANI

Décembre 2024

Résumé

Ce rapport détaille l'ensemble du travail réalisé pour analyser et modéliser les performances dans le contexte du football, depuis l'importation et le prétraitement des données jusqu'à la mise en place de modèles prédictifs et l'évaluation de leurs performances. Trois axes principaux ont été explorés : (1) comprendre dans quelle mesure le temps de jeu influence la capacité d'un joueur à marquer des buts, (2) identifier les facteurs influençant la valeur marchande des joueurs et quantifier leur importance, (3) tenter de prédire des éléments complexes tels que le score d'un match.

À partir de données réelles (`appearances.csv`, `players.csv`, `clubs.csv`, `games.csv`), nous avons nettoyé et fusionné les informations, créé de nouvelles variables plus pertinentes (*feature engineering*), puis testé des modèles de régression linéaire et de forêts aléatoires (Random Forest). Nos résultats montrent que si le temps de jeu seul n'explique qu'une partie limitée des buts marqués, la valeur marchande peut être prédite avec une bonne précision grâce à une Random Forest, mettant en évidence le rôle majeur de la valeur maximale atteinte par le passé, de l'âge, et de la durée restante du contrat. Cependant, la prédiction du score d'un match reste très difficile, ce qui illustre la complexité de ce sport.

Nous intégrons également dans ce document des informations et justifications issues des rapports intermédiaires (PROJECT 2024-12-06, Project_Question1 2024-12-07), afin de fournir une vision globale et complète du processus, des choix méthodologiques, et des évolutions apportées.

Table des matières

1 Introduction **4**

2 Données et Prétraitement **4**
2.1 Sources de Données 4
2.2 Nettoyage des Données et Traitement des Valeurs Manquantes 5
2.3 Feature Engineering 5

3 Analyse Exploratoire **5**
3.1 Temps de Jeu vs Buts 5
3.2 Matrice de Corrélation 6
3.3 Buts par Poste 7

4 Modélisation et Résultats **8**
4.1 Impact du Temps de Jeu (Régression Linéaire) 8
4.2 Prédiction de la Valeur Marchande 8
4.2.1 Importance des Paramètres pour la Valeur Marchande 9
4.3 Prédiction du Score de Match 9

5 Limites et Perspectives **10**
5.1 Limites 10
5.2 Perspectives 10

6 Conclusion **10**

1 Introduction

Le football, sport mondialement suivi, est aujourd'hui un domaine où les données abondent : statistiques individuelles, performances d'équipes, données financières et informationnelles sur les joueurs et les clubs. L'exploitation de ces données permet d'aller au-delà des intuitions pour comprendre finement les facteurs qui influencent les performances. Comme rappelé dans les rapports intermédiaires (PROJECT 2024-12-06, Project_Question1 2024-12-07), l'objectif initial est multiple :

- **Comprendre l'influence du temps de jeu sur les performances offensives** : Le temps de jeu est-il un indicateur clé de la capacité d'un joueur à marquer des buts ?
- **Identifier les facteurs influençant la valeur marchande des joueurs** : Quelles variables (performances passées, âge, durée de contrat, etc.) déterminent principalement la valeur d'un joueur sur le marché ?
- **Exploration de la prédiction du score d'un match** : Peut-on anticiper le score à partir de caractéristiques préalables (club, poste, formation) ?

En toile de fond, la problématique centrale est d'expliquer et de prédire des phénomènes complexes du football avec des modèles statistiques et d'apprentissage automatique. Les travaux préliminaires (PROJECT, Project_Question1) ont souligné l'importance du prétraitement des données, de la sélection de variables pertinentes, et du choix raisonné du modèle (régression linéaire vs Random Forest).

2 Données et Prétraitement

2.1 Sources de Données

Les données proviennent de différents fichiers CSV :

- `appearances.csv` : détail des apparitions des joueurs (minutes jouées, buts, assists, cartons), crucial pour l'analyse du lien temps de jeu - buts.
- `players.csv` : caractéristiques individuelles (âge, pays, valeur marchande actuelle, valeur marchande maximale historique).
- `clubs.csv` : informations sur les clubs (identifiants, compétition).
- `games.csv` : informations sur les matchs (équipes, score final, saison).

La fusion de ces datasets via `player_id`, `club_id`, `game_id` permet d'obtenir un jeu de données global contenant à la fois des données individuelles, contextuelles, et économiques.

2.2 Nettoyage des Données et Traitement des Valeurs Manquantes

Comme expliqué dans la partie II-Analyse des données (PROJECT 2024-12-06), nous avons dû supprimer les lignes présentant des valeurs manquantes critiques (ex : `minutes_played` ou `goals`). L'encodage des variables catégorielles (`position`, `foot`, `current_club_domestic_competition_id`) a été effectué avec `LabelEncoder` et `OneHotEncoder`. Nous avons également vérifié les types des colonnes (int, float).

2.3 Feature Engineering

Afin d'améliorer la qualité des modèles, nous avons créé des variables plus pertinentes (PROJECT_Question1 et améliorations ultérieures) :

- `goals_per_minute` : indicateur de l'efficacité offensive.
- `offensive_contribution` = total_goals + total_assists.
- `contract_years_remaining` : temps restant avant la fin du contrat, impactant la valeur marchande.
- Variables d'interaction (ex : `interaction_goals_minutes`) afin de capturer des relations non linéaires.

Ces transformations, mentionnées dans les rapports intermédiaires, sont essentielles pour donner aux modèles plus d'informations exploitables et améliorer leur performance prédictive.

3 Analyse Exploratoire

3.1 Temps de Jeu vs Buts

La première question posée dans Project_Question1 concerne l'impact du temps de jeu sur la capacité à marquer des buts. Le scatter plot ci-dessous montre une relation positive, mais modérée.

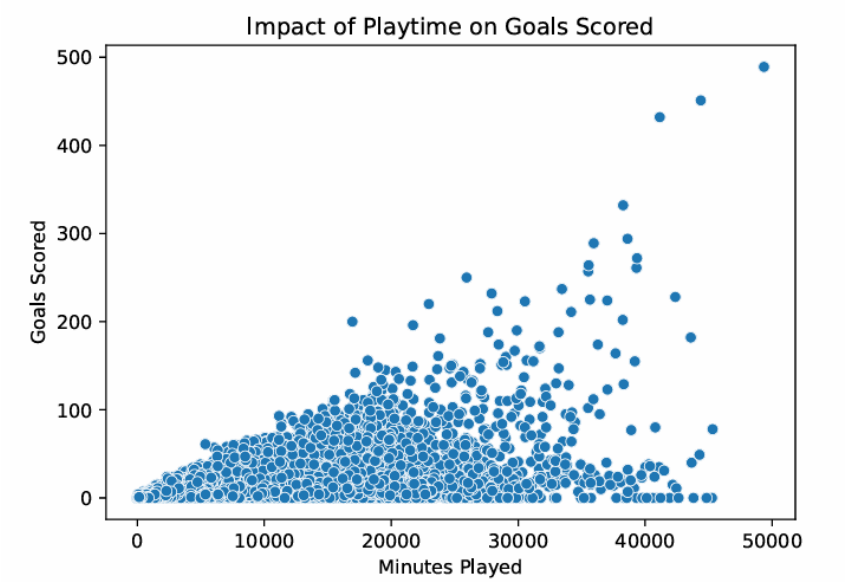

Figure 1 – Relation entre les minutes jouées et les buts marqués.

La corrélation modérée (0.57) laisse entendre que le temps de jeu n'explique pas tout. Il s'agit d'un premier constat motivant l'inclusion d'autres variables (poste, performances passées, etc.).

3.2 Matrice de Corrélation

Avant de construire nos modèles de prédiction de la valeur marchande, nous avons étudié les corrélations entre variables. L'objectif (PROJECT 2024-12-06) était d'identifier les facteurs clés. On observe par exemple que `highest_market_value_in_eur` et `market_value_in_eur` sont fortement liées, tout comme `age` et certaines mesures de performance.

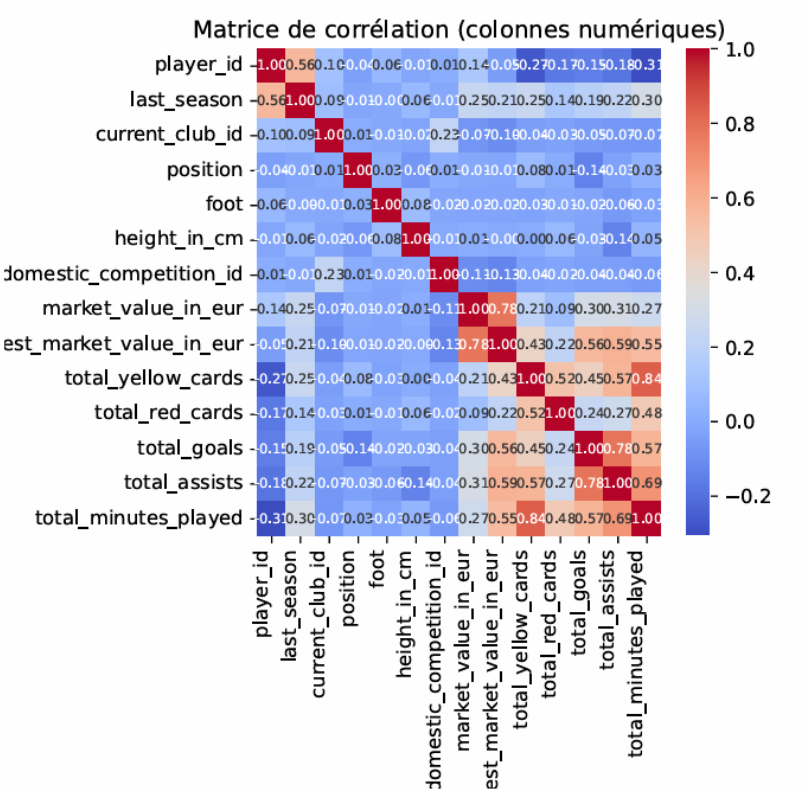

Figure 2 – Matrice de corrélation entre minutes jouées et buts marqués (exemple d'analyse).

Cette étape préliminaire a guidé la sélection des features à inclure dans les modèles prédictifs.

3.3 Buts par Poste

Comme mentionné dans les analyses intermédiaires, le poste joué par un footballeur est un déterminant de sa production offensive :

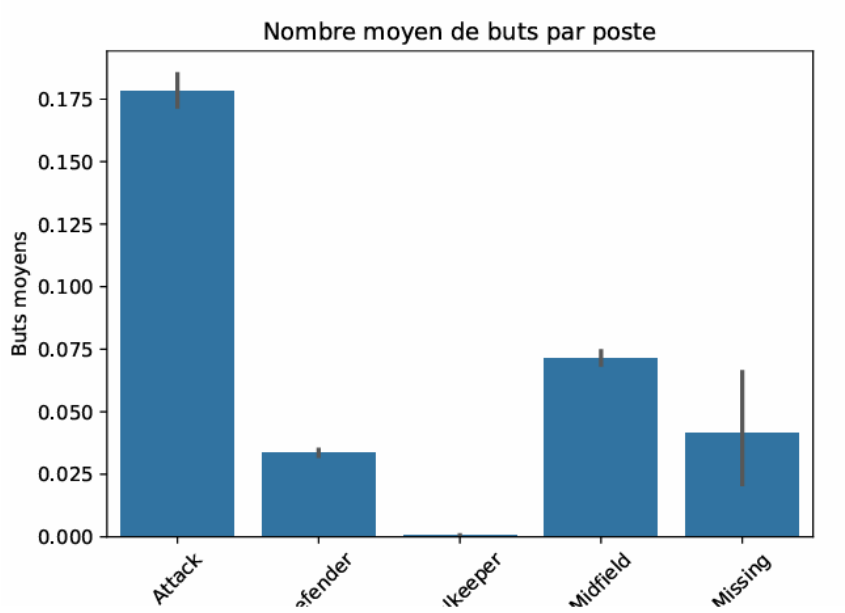

Figure 3 – Buts moyens marqués en fonction du poste.

Les attaquants marquent plus, ce qui est logique et utile à intégrer dans la réflexion, notamment pour comprendre la valeur marchande (les buteurs étant souvent plus valorisés).

4 Modélisation et Résultats

4.1 Impact du Temps de Jeu (Régression Linéaire)

La régression linéaire, comme première approche simple, montre que le temps de jeu explique environ 33% de la variance des buts marqués (R^2 0.33). La pente très faible (0.0015 but/minute) suggère qu'augmenter simplement le temps de jeu n'augmente pas drastiquement les buts. Ces résultats, déjà soulignés dans Project_Question1, confirment qu'il faut intégrer d'autres paramètres (poste, qualité du joueur, etc.).

4.2 Prédiction de la Valeur Marchande

La valeur marchande est au cœur des préoccupations énoncées (PROJECT 2024-12-06, III-Modèles prédictifs). Deux modèles ont été comparés :

- **Régression Linéaire** : $R^2 \approx 0.68$. Interprétable, mais erreurs encore importantes (MAE $>$ 1.7 M€).
- **Random Forest** : $R^2 \approx 0.88$, MAE $\approx$ 0.68 M€, nettement meilleure. La Random Forest capture les non-linéarités et interactions complexes.

Ce résultat confirme les hypothèses des rapports intermédiaires : la simplicité de la régression linéaire est utile pour interpréter les effets, mais la Random Forest offre de

meilleures performances prédictives.

4.2.1 Importance des Paramètres pour la Valeur Marchande

L'un des objectifs clés, mentionné dans les rapports, était de quantifier l'importance relative des paramètres. Après entraînement du modèle Random Forest, l'analyse des importances de variables (`.feature_importances_`) montre :

— `highest_market_value_in_eur` : 67% d'importance.
— `age` : 22%.
— `contract_years_remaining` : 3%.

Les autres variables (minutes jouées, assists, etc.) ont un rôle plus secondaire.

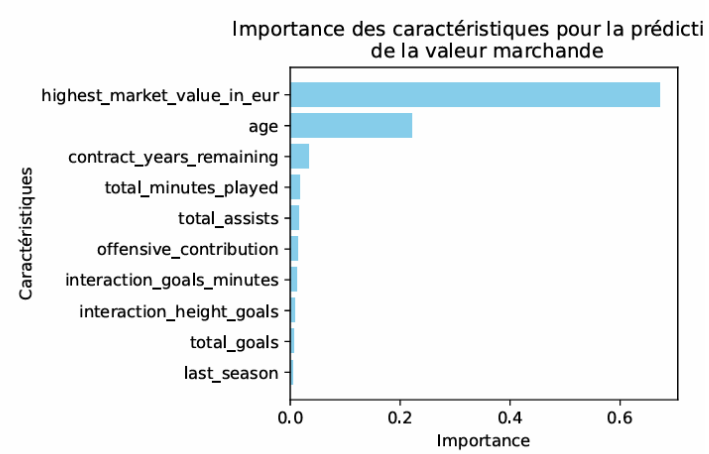

Figure 4 – Importance relative des variables pour la prédiction de la valeur marchande.

Ce graphique illustre que la valeur maximale précédente d'un joueur est le principal facteur explicatif de sa valeur actuelle, suivie de l'âge (un joueur plus jeune avec un fort potentiel est plus prisé), puis de la durée de contrat restante (un joueur sous contrat plus long possède une valeur de négociation plus élevée).

Ce résultat est cohérent avec les analyses préliminaires : la régression linéaire permettait déjà d'identifier des tendances, mais le Random Forest affine l'évaluation de l'importance de chaque paramètre et améliore la précision prédictive.

4.3 Prédiction du Score de Match

Les rapports intermédiaires ont également évoqué la prédiction du score de match. Malgré l'ajout de variables, les essais (Random Forest, autres modèles) donnent des R^2 négatifs, indiquant que le modèle n'explique pas mieux le score que la simple moyenne. Cette difficulté s'explique par la nature aléatoire et contextuelle du football : le score dépend de facteurs tactiques, psychologiques, physiques et contextuels rarement capturés dans nos données.

5 Limites et Perspectives

5.1 Limites

Plusieurs limites ressortent de nos analyses :

— Données incomplètes : Pas d'indicateurs comme les expected goals (xG), les blessures, la forme du jour, la météo.

— Modèles insuffisants pour prédire le score de match, trop complexe et multidimensionnel.

— Les variables choisies, bien que pertinentes, ne couvrent pas la totalité des aspects du football.

5.2 Perspectives

Les rapports PROJECT et Project_Question1 suggèrent diverses améliorations :

— Intégration de données plus fines, incluant des métriques avancées (xG, xA), la dynamique des équipes, des informations tactiques.

— Exploration de modèles plus complexes (XGBoost, LightGBM, réseaux neuronaux) et prise en compte de la dimension temporelle (série de matchs, tendance de performance).

— Collaboration avec des experts du domaine pour orienter la sélection des variables et l'interprétation des résultats.

6 Conclusion

Ce travail, enrichi par les informations des rapports intermédiaires, offre une vue d'ensemble complète : depuis l'identification des objectifs (analyse du temps de jeu, prédiction de la valeur marchande, tentative de prédiction du score), le prétraitement des données, la création de nouvelles variables, l'analyse exploratoire, jusqu'à la comparaison de modèles (régression linéaire, Random Forest) et l'étude de l'importance des paramètres.

Les conclusions principales sont :

— Le temps de jeu explique une partie, mais pas la majorité, des buts marqués.

— La valeur marchande est prédite de façon plus précise avec une Random Forest, mettant en évidence l'importance cruciale de la valeur maximale historique, de l'âge et du contrat.

— La prédiction du score d'un match reste un défi complexe nécessitant des données et des approches plus sophistiquées.

Ce rapport, intégrant les apports des analyses et travaux antérieurs, trace ainsi les grandes lignes d'une démarche analytique et prédictive dans le domaine du football, ouvrant la voie à de futures améliorations et de nouvelles perspectives de recherche.