Rapport du Challenge
Football, Qui va gagner ?
Apprentissage et génération par échantillonnage aléatoire

Rayane Kimouche*
rayane.kimouche@ens-paris-saclay.fr
ENS Paris Saclay
France

Mohamed Khalil Braham*
khalil.braham@telecom-paris.fr
Télécom Paris
France

1 INTRODUCTION

Dans ce challenge, le but est de construire un framework capable de traiter et de lier les statistiques de deux équipes qui s'affrontent ainsi que celles des joueurs pour prédire le score final du match. Cela nous a mené à tester de differentes approches de modélisation pour entrainer éfficacement des modèles de machine et de deep learning. L'enjeu principale est de comment peut on créer un espace partagé qui associe à la fois les données des joueurs avec celles des équipes pour faire apparaître des métriques assez informatives pour la prédictions. Dans les sections suivantes, nous décrivons l'ensemble de données fourni à travers une sélection de visualisations, puis nous présentons le pipeline adopté ainsi que tous les changements modulaires que nous avons considérés au cours de nos expériences. Enfin, nous discutons des résultats obtenus et concluons sur d'autres améliorations possibles.

2 DATA EXPLORATION

Le dataset consiste on deux sous ensembles (respectivement données de train et des test). Chacun possède des occurences correspondantes à des match differents définis par la colonne ID qui représente un identifiant unique du match et relie les occurances X (équipes + joueurs) avec leurs résultats Y.

subset	home	away
train-teams	12303	12303
train-players	237079	236132
test-teams	25368	25368
test-players	509816	504626

Table 1: Data subsets sizes

Pour chaque match, on a 25 métriques pour chacune des deux équipes, déclinées en somme cumulée, moyenne et écart-type. Pareil pour les satistiques des joueurs bien qu'elles sont un peu fines. En faisant une analyse primaire sur ces données, on percute plein d'enjeux qui rendent ce challenge intéressant.

1. On peut constater que la même équipe peut se situer dans deux ligues differentes, donc deux saison differentes, ce qui rend difficile la prise en compte de l'aspect temporel ou même la relation entre plusieurs matches pour la même équipe, en raison des changements de joueurs et des performances durant la saison

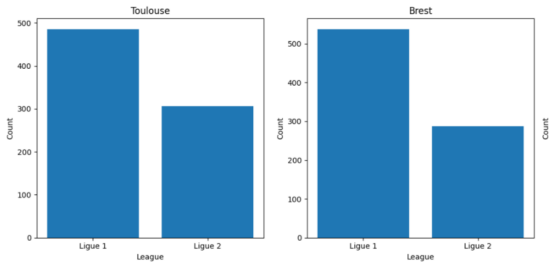

Figure 1: Répartition de quelques équipes sur différentes divisions

2. Après une projection à priori des données d'équipes via une ACP et TSNe, on constate que les données appartenant aux différentes classes sont complexes à distinguer

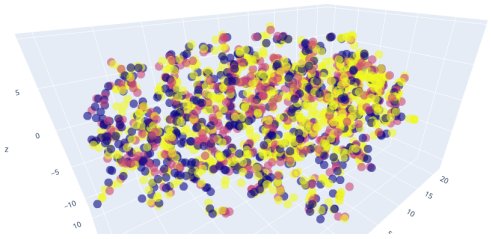

Figure 2: Projection des differentes classes (Victoire, Null et Défaite) pour les données d'équipes

3 DATA CLEANING

3.1 Gestion des données manquantes pour les équipes

Les ensembles de données relatifs aux équipes 'home' et 'away' présentent respectivement 82 407 et 82 495 valeurs manquantes. Pour pallier ce problème, nous avons procédé à une agrégation des données par équipe. Cette agrégation s'organise de manière à trier les occurrences de chaque équipe en fonction des deux variables cumulatives, à savoir **TEAM_SHOTS_TOTAL_season_sum** et **TEAM_SHOTS_INSIDEBOX_season_sum**. Cette méthode repose sur plusieurs postulats :

Postulat 1 : On suppose que pour une équipe donnée X, si deux matchs successifs se déroulent au sein d'une même saison,

*All authors contributed equally to this report.

alors un ordre croissant devrait être observé lorsqu'on se base sur ces variables cumulatives tout au long de la saison. En d'autres termes, on s'attend à ce que le nombre de tirs augmente au fur et à mesure que la saison progresse. Cette logique pourrait également s'appliquer à d'autres variables telles que le nombre de buts.

Postulat 2 : Même dans le cas où deux matchs consécutifs ne se situent pas dans la même saison, il est possible de bénéficier des corrélations existant entre les variables. Ainsi, en ordonnant les deux variables sélectionnées de façon croissante, les autres variables corrélées devraient également manifester une tendance croissante, de façon plus ou moins automatique.

3.2 Construction d'un onze de départ par équipe

Afin de rectifier les incohérences et d'améliorer la structuration des données d'entrée, une stratégie de nettoyage des données a été entreprise. L'objectif était de constituer une équipe-type pour chaque équipe en identifiant les onze meilleurs joueurs pour chaque poste. Pour ce faire, un processus de classement a été élaboré, attribuant à chaque joueur un score basé sur des critères spécifiques à sa position. Les onze joueurs les mieux classés pour chaque équipe ont ensuite été sélectionnés pour former l'équipe-type.

Par exemple, pour le poste de gardien de but, plusieurs critères ont été pris en considération pour évaluer les performances des joueurs. Ces critères incluent le nombre de tirs arrêtés, les dégagements effectués, ainsi que le nombre de buts encaissés. Les performances de chaque gardien ont été évaluées en fonction de ces critères, et un score a été attribué à chacun d'entre eux. Les onze gardiens ayant obtenu les scores les plus élevés ont été retenus pour composer l'équipe-type de gardiens de but de chaque équipe.

De manière similaire, des critères spécifiques ont été définis pour évaluer les performances des joueurs à d'autres postes tels que défenseur, milieu de terrain et attaquant. Par exemple, pour les défenseurs, les critères peuvent inclure le nombre de tacles réussis, le nombre d'interceptions effectuées, et le nombre de passes réussies. Pour les milieux de terrain, les critères peuvent inclure le nombre de passes décisives, le nombre de duels gagnés, et le pourcentage de passes réussies. Pour les attaquants, les critères peuvent inclure le nombre de buts marqués, le nombre de passes décisives, et le nombre de tirs cadrés.

Les joueurs sont ensuite classés par ordre décroissant en fonction de leur score, et les onze meilleurs joueurs pour chaque poste sont sélectionnés pour former l'équipe-type de chaque équipe.

En cas d'insuffisance de joueurs dans une position donnée, des joueurs des matchs précédents de l'équipe sont également pris en compte. Pour ce faire, les identifiants des matchs joués par chaque équipe sont stockés dans un dictionnaire, permettant d'identifier les meilleurs joueurs des matchs précédents et de les intégrer à l'équipe actuelle si nécessaire.

Cette approche garantit l'obtention de données d'entrée plus structurées et cohérentes, ce qui contribue à améliorer la qualité des prédictions effectuées par les modèles d'analyse de données.

4 SÉLÉCTION DES VARIABLES

Pour le choix des variables, on a adopté le score de **l'information mutuelle (IM)** pour les données d'équipes dans le contexte de classification. C'est est une mesure qui évalue la quantité d'informations partagées entre deux variables, cad, entre chaque statistique des deux équipes (home et away) et le résultat du match.

5 APPROCHES DE MODÉLISATION DE DONNÉES

5.1 Stratégies de Normalisation par Ratios de Performance

Dans le cadre de notre étude, nous avons adopté une approche pour traiter les ensembles de données relatives aux équipes et aux joueurs. Cette méthode consiste à normaliser les données en calculant le ratio des performances entre les équipes home et les équipes away, également pour les données des joueurs. Ce processus de normalisation par ratios offre un double avantage. D'une part, il contribue à éliminer les disparités d'échelle entre les variables, facilitant ainsi la tâche des algorithmes de classification. D'autre part, il renforce la pertinence des seuils utilisés par les méthodes de boosting basées sur les arbres de décision.

En réduisant les données à des ratios de performance comparatifs plutôt qu'à des valeurs absolues, nous permettons aux modèles prédictifs de mieux saisir et exploiter les dynamiques compétitives inhérentes à chaque confrontation sportive. Cette technique affine la discrimination opérée par les algorithmes de boosting, en leur permettant de fixer des seuils de décision plus efficaces.

Dans le cas des joueurs, chaque équipe est réduite à un vecteur moyen de ses joueurs qu'ils la composent, ensuite le ratio est calculé (home/away) en se basant sur les performances moyennes des joueurs par équipe.

5.2 Modélisation Graphique de l'Interactivité des Joueurs

La deuxième approche mise en œuvre dans notre analyse constitue une approche graphique qui se distingue par sa capacité à cartographier les interactions complexes au sein d'une équipe. Cette technique consiste à construire une représentation en réseau des joueurs, où chaque joueur est un nœud du graphe, et les connexions — ou arêtes — entre eux sont définies par divers indicateurs d'interactivité. Ces indicateurs visent à quantifier la force et la fréquence des échanges entre deux joueurs, fournissant ainsi une mesure tangible de leur cohésion.

Dans ce contexte, la cohésion ne se limite pas aux actions physiques telles que les passes effectuées pendant un match, mais englobe également les aspects moins tangibles tels que la communication et la compréhension mutuelle. Par exemple, une synergie accrue se manifeste par une augmentation de l'efficacité des passes et une meilleure anticipation des mouvements entre les joueurs qui ont fréquemment collaboré lors des matchs précédents (temps joués ensemble). L'hypothèse sous-jacente est que plus deux joueurs ont accumulé du temps de jeu ensemble, plus leur entente sur le terrain est optimisée.

Cette approche graphique offre une dimension supplémentaire à notre analyse des performances sportives, en mettant en lumière les réseaux de coopération et d'entraide qui peuvent être décisifs dans le succès collectif. Elle permet également de détecter les duos ou les trios de joueurs dont l'alchimie pourrait se traduire par une performance supérieure, ainsi que d'identifier les zones où des

améliorations dans la dynamique d'équipe pourraient être bénéfiques.

5.3 Construction des graphes

Dans ce processus de création de graphes pour chaque équipe de football, nous utilisons les données fournies dans un DataFrame pour chaque match. Chaque ligne de ce DataFrame représente un joueur participant à un match spécifique, avec des informations détaillées telles que les minutes jouées, les passes réussies, les buts marqués, etc. Notre objectif est de représenter les interactions entre les joueurs sous forme de graphes, où les nœuds représentent les joueurs et les arêtes représentent les interactions entre eux.

5.4 Détermination des liens entre les joueurs

Lorsque nous parcourons les matchs, nous extrayons les données des joueurs de chaque équipe participant à chaque match ainsi que leurs positions respectives sur le terrain. Cette information est cruciale car elle nous aide à déterminer comment les joueurs interagissent en fonction de leurs rôles spécifiques.

Ensuite, en parcourant les paires de joueurs dans chaque équipe, nous décidons si une interaction doit être représentée par une arête dans le graphe et attribuons un poids à cette arête en fonction de divers critères.

La logique pour créer des arêtes et des poids varie en fonction des positions des joueurs impliqués dans l'interaction:

- **Défenseurs (defender)** : Des arêtes sont ajoutées entre les défenseurs en fonction de certaines statistiques telles que les passes précises et les dégagements.
- **Milieux de terrain (midfielder)** : Des arêtes sont ajoutées entre les milieux de terrain en fonction de statistiques spécifiques telles que les passes précises.
- **Attaquants (attacker)** : Des arêtes sont ajoutées entre les attaquants en fonction de statistiques telles que les buts marqués et les tirs cadrés.
- **Défenseur et Milieu de terrain** : Des arêtes sont ajoutées entre les défenseurs et les milieux de terrain en fonction de statistiques telles que les interceptions et les passes décisives.
- **Milieu de terrain et Attaquant** : Des arêtes sont ajoutées entre les milieux de terrain et les attaquants en fonction de statistiques telles que les dribbles réussis et les passes clés.

En utilisant cette approche, nous créons des graphes qui capturent les différentes interactions entre les joueurs, en tenant compte de leurs performances individuelles et de leurs rôles sur le terrain. Ces graphes peuvent ensuite être utilisés pour analyser les dynamiques d'équipe, identifier les synergies entre les joueurs et prédire les performances futures de l'équipe.

5.5 Architecture du modèle

GCN: Graph Convolutional Networks Les réseaux de convolution sur graphes (GCN) sont des modèles puissants utilisés pour l'analyse de données structurées sous forme de graphes. Ils étendent les concepts de convolution des réseaux neuronaux classiques aux données de graphe. Une couche typique de GCN peut être décrite par l'équation suivante:

$$h_i' = \sigma\left(\sum_{v_j \in \mathcal{N}(v_i)} \alpha_{ij} \mathbf{W} h_j\right) \tag{1}$$

Où :

- v_i est le nœud i.
- $\mathbf{W}$ est une matrice apprenable de taille $F' \times F$.
- α_{ij} spécifie le facteur de pondération (importance) des caractéristiques du nœud j pour le nœud i.
- σ est une non-linéarité.

Les GCN permettent aux modèles d'apprendre des représentations de graphe en tenant compte de la structure locale et des relations entre les nœuds. Ils sont largement utilisés dans des domaines tels que la classification de graphe, la prédiction de liens, et la recommandation.

GAT: Graph Attention Networks

L'architecture GAT adapte l'idée d'auto-attention des transformers aux réseaux de convolution de graphes (GCN).

Les GAT calculent les facteurs de pondération en utilisant les caractéristiques des nœuds. Tout d'abord, les coefficients d'attention sont calculés comme suit :

$$e_{ij} = a\left(\mathbf{W}h_i, \mathbf{W}h_j\right) = LeakyReLu\left(\mathbf{a}\left[\mathbf{W}h_i || \mathbf{W}h_j\right]\right) \tag{2}$$

Ensuite, $a(.)$ est appelé le mécanisme d'attention et constitue un réseau feed-forward unique avec $2 \times F'$ paramètres (vecteur $\mathbf{a}$ et fonction d'activation $LeakyReLu$). Enfin, le facteur de pondération est calculé comme suit :

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in \mathcal{N}(v_i)} exp(e_{ik})} \tag{3}$$

De plus, à l'instar des transformers, nous pouvons avoir un mécanisme d'attention multi-têtes, où pour une attention à K têtes, nous avons :

$$h_j = ||_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}(v_i)} \alpha_{ij}^k \mathbf{W}^k h_j\right) \tag{4}$$

Notre modéle

5.5.1 Modèle de graphes. Au cœur de notre modèle se trouvent deux ensembles de couches de convolution GNN, l'un dédié à l'équipe jouant à domicile et l'autre à l'équipe jouant à l'extérieur. Ces couches servent à extraire des représentations significatives des nœuds du graphe pour chaque équipe, en mettant l'accent sur les interactions entre les joueurs. Grâce à l'utilisation de mécanismes d'attention, notre modèle peut pondérer différemment l'importance des connexions entre les joueurs en fonction de leurs performances individuelles et de leurs interactions collectives.

5.5.2 Utilisation des embeddings. Après l'extraction des représentations de graphe pour chaque équipe, celles-ci sont agrégées en une seule représentation globale de l'équipe à l'aide d'une opération de pooling global, qui calcule la moyenne des caractéristiques des nœuds. Cette étape permet de conserver la structure relationnelle du graphe tout en synthétisant l'information provenant de tous les joueurs.

Parallèlement à la modélisation des interactions entre les joueurs, notre modèle intègre également des caractéristiques spécifiques à chaque club. Ces caractéristiques fournissent des informations contextuelles supplémentaires sur les équipes, telles que leur historique de performance, leur style de jeu, etc., ce qui enrichit la capacité de notre modèle à prédire les résultats.

5.5.3 Sortie du modèle. Les représentations des équipes ainsi que les caractéristiques du club sont concaténées et passées à travers des couches entièrement connectées pour produire les prédictions finales. Ce processus permet à notre modèle d'apprendre des patterns complexes dans les données de graphe et de prendre des décisions éclairées sur les performances des équipes.

En somme, notre modèle se distingue par sa capacité à exploiter les informations relationnelles entre les joueurs et les équipes grâce aux mécanismes d'attention et aux couches de convolution de graphes, ce qui en fait un outil puissant pour l'analyse et la prédiction des performances dans le contexte du football.

5.6 Modèle CatBoost

Les embeddings générés par le modèle GAT, qui capturent les informations relationnelles entre les joueurs à travers les couches de convolution de graphes et les mécanismes d'attention, sont utilisés comme caractéristiques d'entrée pour un modèle CatBoost. CatBoost est un algorithme de gradient boosting spécialement conçu pour la classification multiclasse.

5.7 Modèles de régression et stacking

Dans cette partie, nous adoptons une approche en cascade pour améliorer les performances du modèle initial CatBoost. Pour ce faire, nous utilisons des modèles de régression tels que CatBoost, XGBoost et LightGBM pour estimer les résidus entre les prédictions du modèle CatBoost et les vraies étiquettes, mais nous effectuons cette opération séparément pour chaque dimension de la prédiction.

5.7.1 Résidus de la classification initiale. Plus précisément, après que le modèle CatBoost ait produit ses prédictions pour chaque match de football, nous calculons les résidus entre ces prédictions et les vraies étiquettes. Ces résidus représentent l'écart entre la prédiction du modèle et la vérité terrain pour chaque dimension de la prédiction, à savoir les probabilités de victoire à domicile, de match nul et de victoire à l'extérieur.

5.7.2 Prédiction des résidus. Nous utilisons des modèles de régression tels que CatBoost, XGBoost et LightGBM pour estimer ces résidus pour chaque dimension de la prédiction. Ces modèles de régression sont entraînés à prédire les écarts entre les prédictions du modèle CatBoost et les vraies étiquettes, en utilisant les embeddings générés par le modèle GAT comme caractéristiques d'entrée.

Une fois que les modèles de régression ont été entraînés et validés, nous les utilisons pour prédire les résidus pour chaque dimension de la prédiction sur l'ensemble de test. Ces prédictions de résidus sont ensuite ajoutées aux prédictions originales du modèle CatBoost pour chaque dimension.

5.7.3 Stacking Regressor. Enfin, un modèle de stacking est utilisé pour combiner ces nouvelles prédictions avec celles du modèle CatBoost, fournissant ainsi une prédiction finale plus précise. Le modèle de stacking peut être un modèle simple comme une régression linéaire, ou plus complexe comme un modèle de régression ensembliste. Cette approche en cascade permet d'améliorer les performances du modèle initial en exploitant les informations supplémentaires fournies par les modèles de régression.

5.8 Entrainement des modèles

5.8.1 Modèles GCN et GAT pour la Prédiction des Résultats de Matchs. Hyperparamètres :

- `num_node_features` : 272
- Nombre de Sorties (`nout`) : 3
- Dimension des Couches Cachées (`nhid`) : 256
- Nombre de Canaux Cachés du Graphe : 600
- Taille des Caractéristiques des Clubs : 16
- Taux d'Abandon (`dropout`) : 0.3

Optimiseur :

- SGD (Stochastic Gradient Descent) avec un taux d'apprentissage de 0.01, un momentum de 0.9 et une décroissance de poids de 5e-4

Nombre d'Époques : 100

5.8.2 Modèles de Régression pour l'Amélioration des Prédictions. CatBoost :

- Nombre d'arbres (`iterations`) : 1000
- Taux d'apprentissage (`learning_rate`) : 0.1
- Profondeur de l'arbre (`depth`) : 6

XGBoost :

- Nombre d'arbres (`n_estimators`) : 500
- Profondeur maximale de l'arbre (`max_depth`) : 3
- Taux d'apprentissage (`learning_rate`) : 0.01

LightGBM :

- Nombre d'arbres (`n_estimators`) : 100
- Nombre de feuilles (`num_leaves`) : 4
- Taux d'apprentissage (`learning_rate`) : 0.01

Modèle de Stacking :

- Méta-régresseur : Régression Linéaire

6 APPRENTISSAGE D'ENSEMBLE

Dans la partie d'apprentissage d'ensemble par agrégation de la moyenne pondérée, nous avons cherché à combiner les prédictions de plusieurs modèles que nous avons formés tout au long de notre analyse. L'objectif était d'améliorer la robustesse et les performances globales de notre estimateur final en exploitant la diversité des modèles individuels.

Pour ce faire, nous avons effectué plusieurs agrégations de moyennes de prédictions en utilisant des poids proportionnels aux performances individuelles des modèles. Chaque modèle a été évalué sur un ensemble de validation ou de test, et son score initial a été calculé en fonction de ses performances. Les poids pour chaque modèle dans l'agrégation de la moyenne pondérée ont été déterminés en proportion de ces scores initiaux.

Plus précisément, le processus d'agrégation s'est déroulé comme suit :

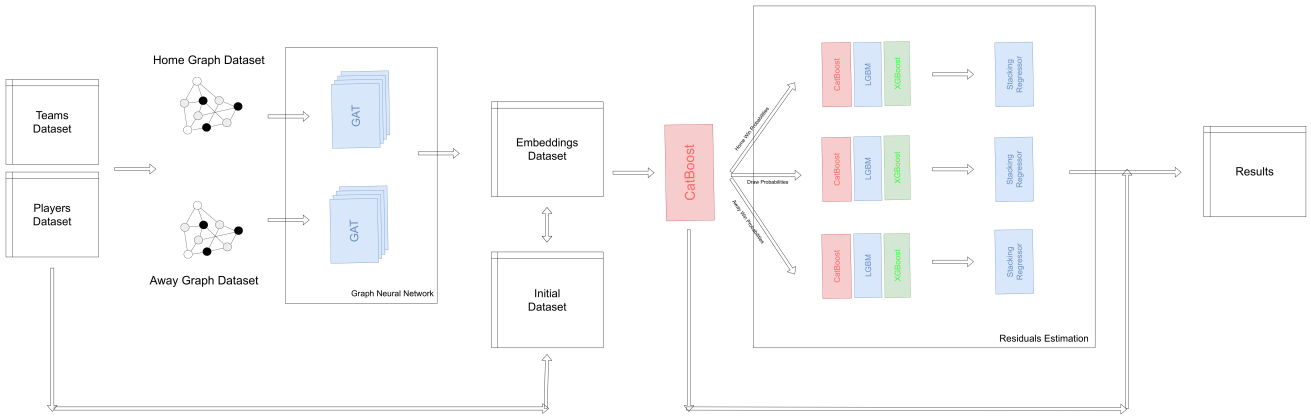

Figure 3: APPROCHE BASÉE SUR LA REPRÉSENTATION GRAPHIQUES DES JOUEURS (GAT+Bossting)

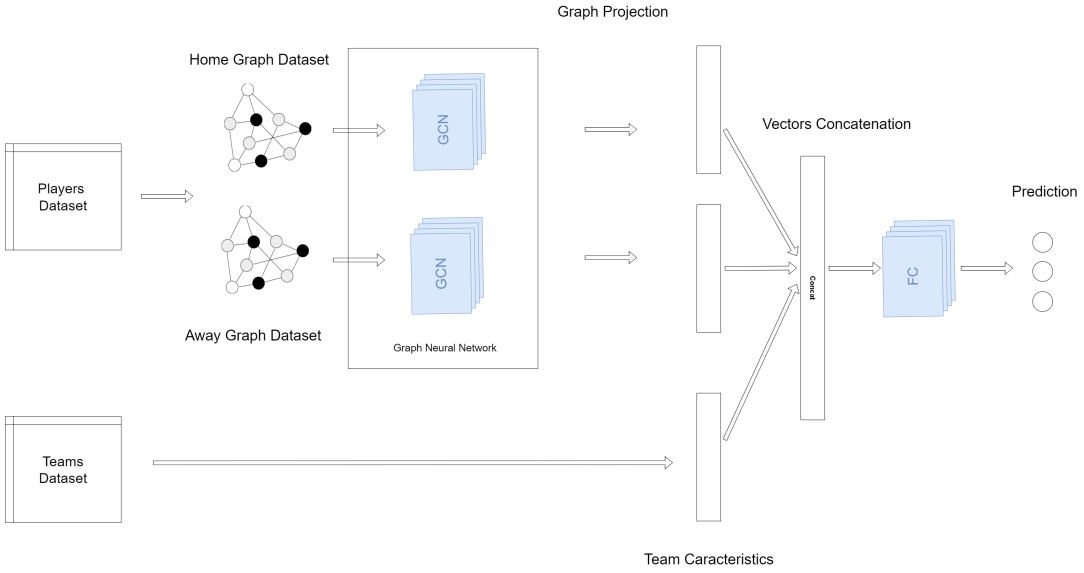

Figure 4: APPROCHE BASÉE SUR LA REPRÉSENTATION GRAPHIQUES DES JOUEURS (GCN + MLP)

(1) Calcul des performances individuelles des modèles : Chaque modèle formé a été évalué sur l'ensemble de test.
(2) Attribution de poids aux modèles : Les poids pour chaque modèle dans l'agrégation de la moyenne pondérée ont été déterminés en fonction de ses performances initiales.
(3) Évaluation de l'estimateur final : L'estimateur final, obtenu par l'agrégation de la moyenne pondérée des prédictions des modèles individuels, a été évalué sur un ensemble de validation et de test distinct. Les performances de cet estimateur final ont été comparées à celles des modèles individuels pour évaluer l'efficacité de l'approche d'apprentissage d'ensemble.

En résumé, l'agrégation de la moyenne pondérée nous a permis de tirer parti de la diversité des modèles individuels et d'améliorer la stabilité et les performances globales de notre estimateur final. En attribuant des poids aux modèles en fonction de leurs performances initiales, nous avons pu combiner efficacement les forces de chaque modèle pour produire des prédictions plus fiables et précises.

7 SOURCES D'AMÉLIORATION

Une possibilité d'amélioration pourrait être d'explorer davantage les interactions temporelles et saisonnières dans les données. Actuellement, notre framework considère chaque match de manière isolée, sans tenir compte des évolutions au fil du temps. En incorporant des caractéristiques temporelles telles que la phase de la saison, les performances récentes des équipes, ou les changements dans la composition des équipes au fil du temps, nous pourrions mieux capturer les tendances et les dynamiques à long terme qui influencent les résultats des matchs.

De plus, une analyse plus approfondie des données des joueurs pourrait également conduire à des améliorations significatives. Par

exemple, en tenant compte des performances individuelles des joueurs clés, de leur forme physique, de leur historique de blessures, ou même de leur style de jeu spécifique, nous pourrions affiner nos prédictions et identifier les facteurs qui ont le plus d'impact sur les résultats des matchs.

Enfin, l'exploration de nouvelles techniques de modélisation et d'apprentissage automatique, telles que l'apprentissage par renforcement ou l'utilisation de réseaux neuronaux plus complexes, pourrait également ouvrir de nouvelles perspectives pour améliorer les performances de notre framework. En adoptant une approche plus holistique et en intégrant de manière plus dynamique les différentes dimensions des données, nous pourrions développer des modèles encore plus précis et adaptés à la complexité du domaine du sport.

8 CONCLUSION

La construction du framework pour prédire les scores finaux des matchs de football a été un défi passionnant, mettant en lumière la complexité des données sportives et les diverses approches pour les modéliser efficacement. Nous avons exploré différentes stratégies de modélisation, en mettant l'accent sur l'intégration des données des équipes et des joueurs, ainsi que sur l'utilisation de techniques de représentation graphique pour capturer les interactions entre les joueurs.

Nos analyses ont révélé des défis intéressants, tels que la gestion des données manquantes, la construction d'un onze de départ cohérent pour chaque équipe, et la sélection des variables les plus pertinentes pour la prédiction des résultats des matchs. Nous avons également exploré l'utilisation de modèles de machine learning traditionnels ainsi que des techniques plus avancées telles que les réseaux de convolution sur graphes (GCN) et les réseaux de neurones à attention (GAT).

L'approche en cascade que nous avons adoptée, combinant les prédictions de plusieurs modèles à différents niveaux de granularité, a permis d'améliorer considérablement les performances de notre estimateur final. En utilisant des modèles de régression pour estimer les résidus entre les prédictions du modèle initial et les vraies étiquettes, nous avons pu ajuster nos prédictions de manière plus précise et fournir des estimations plus fiables des scores finaux des matchs.

En fin de compte, notre framework offre une solution robuste et polyvalente pour la prédiction des résultats des matchs de football, en exploitant pleinement les informations disponibles sur les équipes, les joueurs et leurs interactions. Bien que notre approche ait produit des résultats prometteurs, il reste encore des opportunités d'amélioration, notamment en explorant de nouvelles techniques de modélisation et en intégrant des données supplémentaires telles que les conditions météorologiques et les blessures des joueurs. En continuant à innover et à affiner notre approche, nous pourrons développer des modèles encore plus précis et informatifs pour prédire les performances dans le domaine passionnant du sport.