Football, Qui va gagner ?

Qube Research and Technologies

Mohamed Khalil Braham Rayane Kimouche

Mathématiques, Vision, Apprentissage
ENS Paris Saclay

June 14, 2024

Overview

Introduction

Introduction

Objectif principal : est de prédire qu'elle équipe va gagner dans un affrontement de deux opposants en se basant sur un historique de performances notament le contexte saisonal et les derniers statistiques.

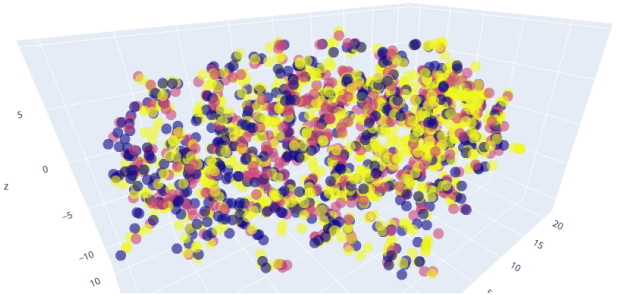
- Données d'entrainement : contenant les statistiques pour chaque équipe avec celles des joueurs plus affinés. Le résultat de chaque match est désigné par 3 variables binaires (Victoire, Null ou Défaite).

- BaseLine : qui donne une précision de 47.5% sur les données d'entraînement et 46% sur les données de tests.

Investigation

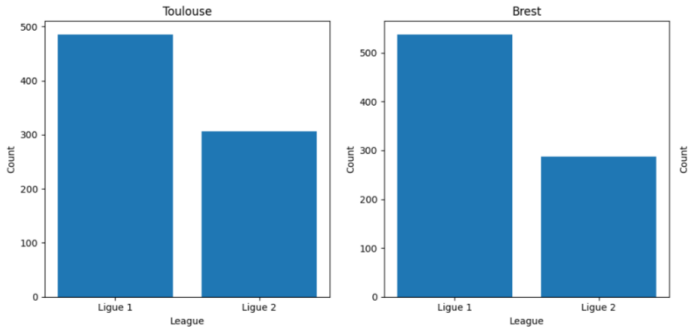
Investigation

Est-ce que les données sont discriminables ?



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

Investigation



On peut constater que la même équipe peut se situer dans deux ligues differentes, donc deux saisons differentes, ce qui rend difficile la prise en compte de l'aspect temporel = IID

Data Cleaning

Data Cleaning

Gestion des données manquantes pour les équipes

- Données manquantes pour les équipes 'home' et 'away': 82 407 et 82 495 valeurs respectivement.
- **Idée :** Agrégation des données par équipe et ordonnancement croissant via les variables cumulatives TEAM_SHOTS_TOTAL_season_sum et TEAM_SHOTS_INSIDEBOX_season_sum.

Postulats

1. Si pour une équipe agrégée, trois match successifs appartiennent à la même saison. **Choix jusdicieux** : remplir les NaNs par (ff + fb) / 2 (suivre le même rythme).
2. Si ce n'est pas le cas, on pourrait bénéficier des corrélations entre les variables, et que le restant des variables devraient se manifester par une tendance croissante, on peut remplir les NaNs par (ff + fb) / 2.

Gestion des données manquantes pour les joueurs

- Dans ce cas, on estimait une valeur manquante par la moyenne des performances de tout les joueurs pendant le match.

Séléction des Variables

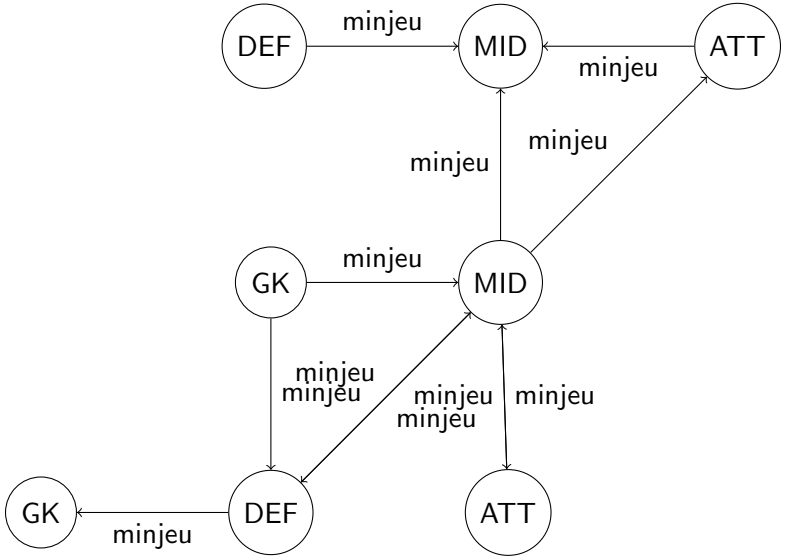
- Utilisation du score de **l'information mutuelle (IM)** pour les données d'équipes dans un contexte de classification.
- L'information mutuelle évalue la quantité d'informations partagées entre chaque statistique des deux équipes (home et away) et le résultat du match.
- L'experimentation avec d'autres mécanismes de choix comme **Analyse des Composantes Principales** et la **matrice de corrélation**.

Onze de départ

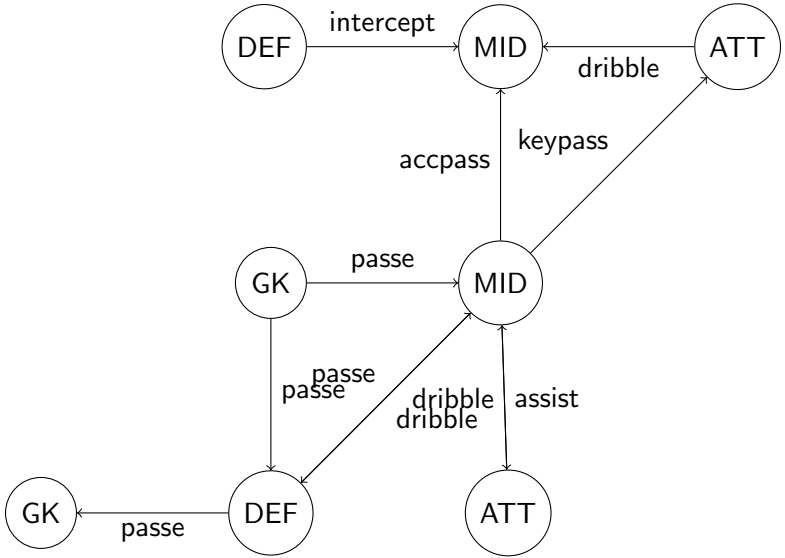
Pour trouver un 11 de départ pour chaque équipe, on considère un classement de l'effectif selon des scores calculés relativement aux positions des joueurs.

- **Goalkeeper**
  - PLAYER_SAVES_season_sum: 1
  - PLAYER_PUNCHES_season_sum: 1
- **Defender**
  - PLAYER_CLEARANCES_season_sum: 1
  - PLAYER_TACKLES_season_sum: 1
- **Midfielder**
  - PLAYER_ACCURATE_PASSES_season_sum: 1
  - PLAYER_ASSISTS_season_sum: 1
- **Attacker**
  - PLAYER_GOALS_season_sum: 1
  - PLAYER_ASSISTS_season_sum: 1

Construction des graphes: Sparse

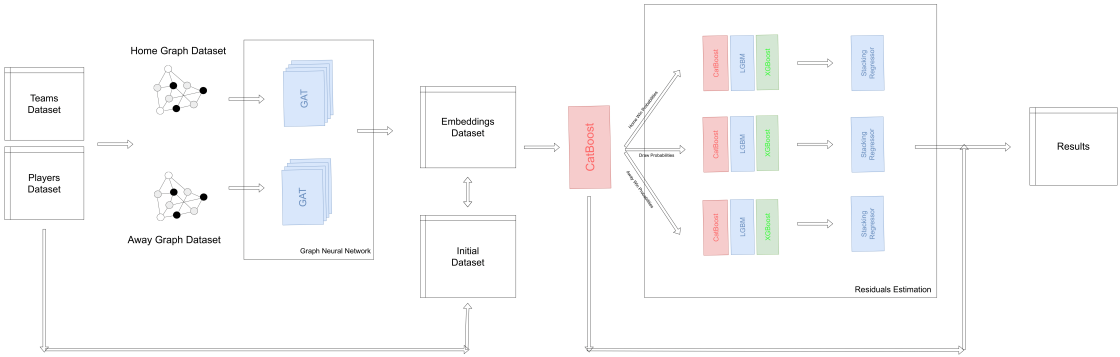


Construction des graphes

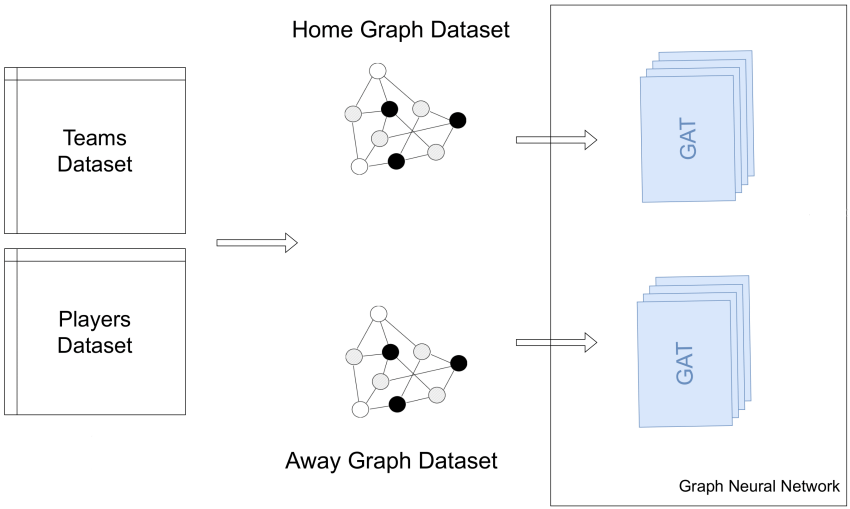


Première Approche

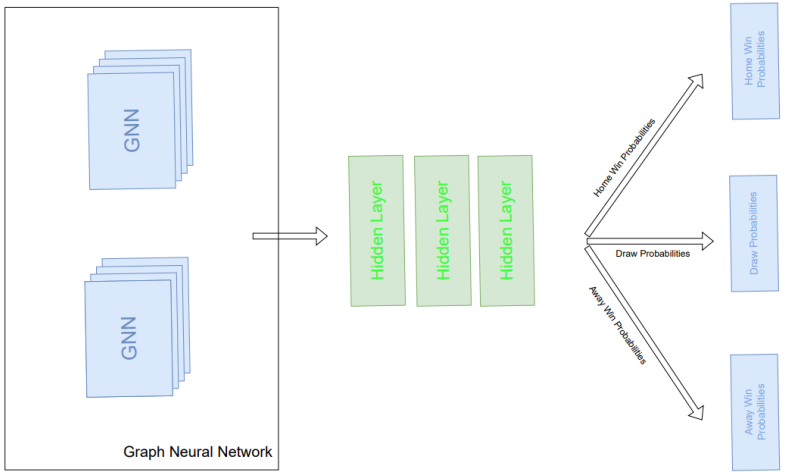
Architecture Globale



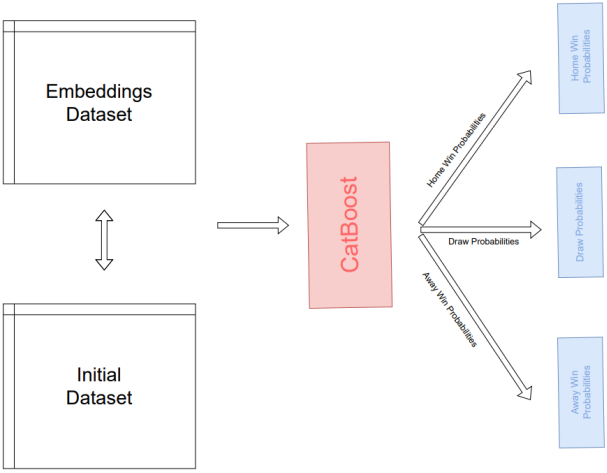
Graph Neural Networks



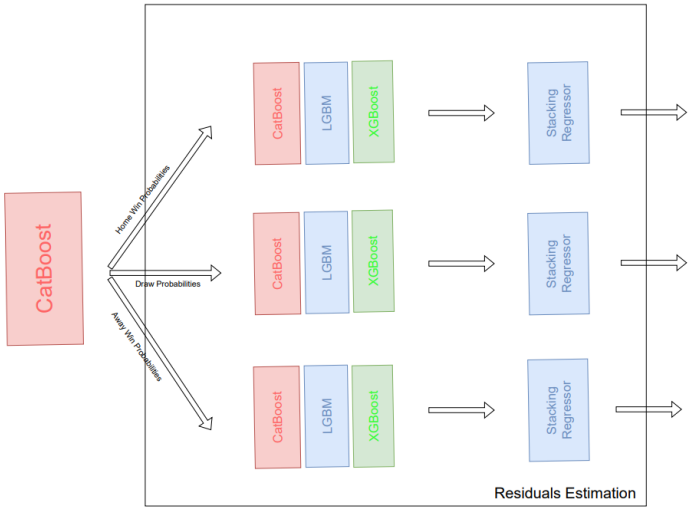
Classification Initiale



Classification Initiale



Prédiction des résidus



Entrainement

GAT 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 :

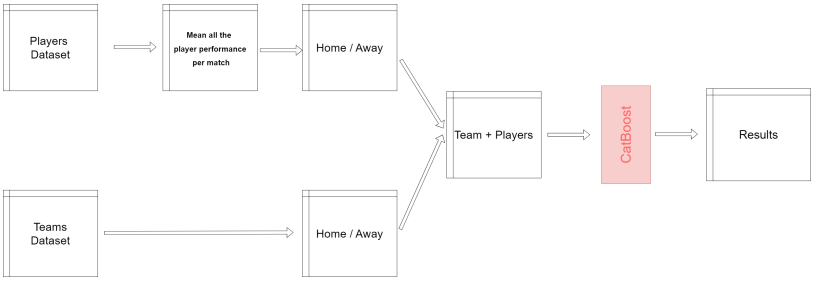
- SGD (Stochastic Gradient Descent) avec un taux d'apprentissage de 0.01, un momentum de 0.9 et une décroissance de poids de 5e-4

Nombre d'Epochs : 50
Modèle de Stacking :

- Méta-régresseur : Régression Linéaire

Deuxième Approche

Deuxième Approche : Stratégies de Normalisation par Ratios de Performance



Deuxième Approche : Stratégies de Normalisation par Ratios de Performance

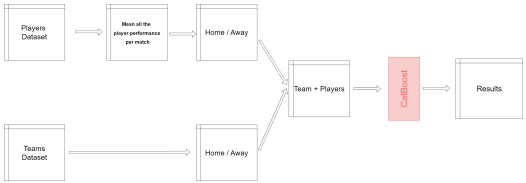


Figure: Pipeline d'entraînement

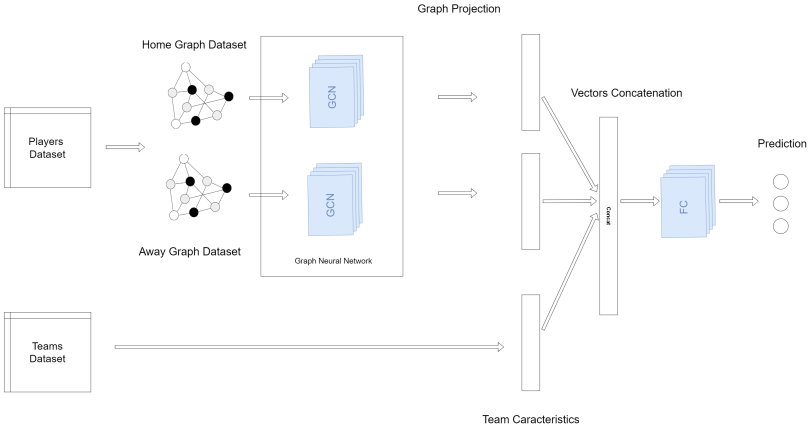
Dataset	Accuracy
Validation	59 %
Submission (test)	47.6 %

Figure: Résulats du modèle

- Réduire les données à des ratios de performance comparatifs plutôt qu'à des valeurs absolues.
- Cette technique affine la discrimination opérée par les algorithmes de boosting, en leur permettant de fixer des seuils de décision plus efficaces
- Random Search sur les parameters : 'max_depth': 3, 'n_estimators': 500

Troisième Approche

Troisième Approche



- Dans cette approche les joueurs sont représentés par un réseau, où chaque joueur est un nœud du graphe, et les connexions entre eux sont définies par la variables **PLAYER MINUTES PLAYED 5 last match sum** représentant ainsi la cohésion entre deux joueurs.

Troisième Approche

- Le modèle GCN permet de projeter les graphs dans le même espace que les vecteurs des équipes où tout sera concatené.

- C'est le modèle qui a donné les meilleurs résultat de classification (voir figure 25)

Dataset	**Accuracy**
Validation	51 %
Submission (test)	**49.05** %

Résultats du modèle

Conclusion

- Pour conclure, ce challenge a été un défi passionnant à travers lequel nous avons proposé divers frameworks basés sur des approches et des reflexions differentes.
- Bien que notre approche ait produit des résultats prometteurs, il reste encore des opportunités d'amélioration, en exploitant par exemple des données non normalisés ou en essayant de trouver d'autres associations de données plus subtiles.