Sam Vallet
Thomas Sinapi

Prédiction des issues des matchs à partir des statistiques des équipes

https://github.com/tomasnp/Predictions_Match_Outcome

Notre fichier est trop volumineux, nous n'avons pas réussi à le poster sur ecampus et comme nous avons plusieurs documents pour nos données nous n'avons pas mis de lien voici le lien vers le git ou il y a tous le code du projets ainsi que nos données.

Le football est un sport où les données sont de plus en plus utilisées pour analyser les performances des joueurs et des équipes. Ces données peuvent être utilisées pour prédire les résultats futurs, identifier les points forts et les points faibles des équipes et améliorer les performances globales. Dans ce rapport, nous allons examiner deux bases de données distinctes sur le football qui contiennent des statistiques et des informations sur les matchs.

La première base de données que nous avons utilisée contient des informations sur les matchs, telles que les noms des deux équipes et les statistiques du match (nombre de tirs cadrés, pourcentage de possession, nombre de fautes, etc…).

La deuxième base de données que nous avons utilisée est similaire, mais elle contient également des informations sur les xG (Expected Goals ou buts attendus). Les xG sont une mesure de la dangerosité des attaques de chaque équipe, basée sur la probabilité qu'un tir se transforme en but en fonction de la position sur le terrain et de la situation de jeu.

L'objectif principal de ce projet est d'explorer ces deux bases de données et de voir comment nous pouvons les utiliser pour faire des prédictions sur les résultats de matchs. Pour ce faire, nous avons utilisé des algorithmes d'apprentissage automatique pour analyser les données et identifier des tendances et des modèles significatifs.

PREPROCESSING:

Nous avons restreint notre base de données au championnat anglais "Premier League" des saisons 2016 à 2020. En fusionnant deux bases de données, nous avons créé un dataframe contenant toutes les informations sur les 1900 matchs de cette période. Nous avons ajouté une colonne "Match Outcome" qui donne l'issue de chaque match. Cependant, nous avons rencontré un problème lors de la fusion car les noms des équipes étaient différents dans chaque dataset. Pour résoudre ce problème, nous avons créé une fonction qui a comparé les noms des équipes et fusionné celles qui étaient similaires. Cette approche nous a permis de traiter les données avec précision et d'assurer que les informations de chaque équipe étaient correctement associées.

Sam Vallet
Thomas Sinapi

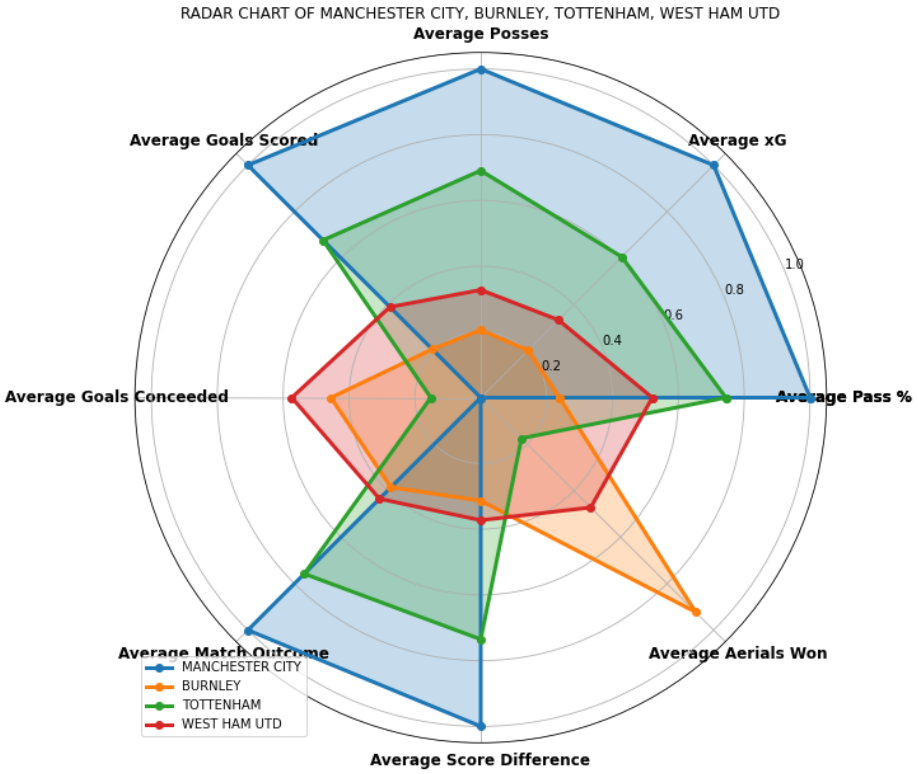
Nous avons également créé de nouvelles features à partir des données du dataframe pour obtenir des informations sur l'état de forme de chaque équipe, ce qui joue un rôle capital dans le football. En effet, une équipe enchaînant les bons résultats aura tendance à mieux aborder ses rencontres. Nous avons également ajouté des attributs donnant le nombre de points accumulés et le classement de chaque équipe avant le match, qui donnent des informations sur le niveau de l'équipe durant la saison.

Nous avons créé un second dataframe qui contient les statistiques moyennes de chaque équipe sur l'ensemble des saisons. Ce dataframe nous permettra de confronter les équipes en utilisant chacun de leurs attributs et d'essayer de déterminer le gagnant, un peu comme les combats de pokemons vus en TP.

ENCODAGE

Dans notre dataset, les noms d'équipes sont des variables catégorielles, ce qui signifie qu'elles doivent être encodées sous forme numérique avant de pouvoir être utilisées dans des algorithmes de classification. Pour ce faire, nous avons utilisé une technique appelée "encodage des labels". Cette technique consiste à attribuer un numéro unique à chaque catégorie, ce qui permet de transformer les noms d'équipes en données numériques utilisables dans nos modèles de prédiction. Nous avons également implémenté différentes fonctions afin de pouvoir retrouver facilement le nom d’une équipe à partir de son nom encodé.

Sam Vallet
Thomas Sinapi



Ci-dessus, nous avons un graphique radar qui nous montre de manière claire les statistiques de plusieurs équipes.

MODELE

Nous avons choisi l'algorithme Random Forest pour prédire les résultats de matchs de football pour plusieurs raisons. Avant de faire ce choix, nous avons testé plusieurs autres algorithmes, tels que Decision Tree et KNN, mais nous avons constaté que Random Forest nous donnait les meilleurs résultats. En effet KNN est moins performant que Random Forest pour les jeux de données de grande taille car il stocke toutes les données, ce qui peut ralentir l'algorithme. De plus, KNN ne peut pas identifier les caractéristiques les plus importantes pour prédire les résultats des matchs. Quant à Decision Tree, il est moins robuste pour les jeux de données de grande taille, car il a tendance à sur-ajuster ou sous-ajuster les données. En revanche, Random Forest est efficace pour traiter les données complexes, identifier les caractéristiques les plus importantes et fournir des probabilités de victoire, de match nul ou de défaite.

Random Forest est capable de traiter efficacement de grandes quantités de données complexes, ce qui est courant dans les données de football. De plus, il peut identifier les caractéristiques les plus importantes pour prédire les résultats des matchs, ce qui nous permet de mieux comprendre les facteurs clés qui influencent les résultats. Enfin, il permet de prédire les probabilités de victoire, de match nul ou de défaite plutôt qu'une simple prédiction binaire, ce qui nous donne une analyse plus détaillée et précise des résultats.

Sam Vallet
Thomas Sinapi

1)

Une première méthode de prédiction a été utilisée pour déterminer l'issue d'un match en utilisant toutes les statistiques disponibles pour chaque équipe. Les données ont été séparées en ensembles d'entraînement et de test pour évaluer les performances du modèle et déterminer les attributs les plus importants. Les résultats ont montré que la note attribuée à chaque équipe après le match était un critère discriminant important pour l'issue du match.

Les hyperparamètres de la fonction Random Classifier ont été optimisés pour améliorer les performances du modèle.

Nous avons effectué une analyse en composantes principales (PCA) sur nos données de prédiction des résultats de matchs de football. Cependant, les résultats de la PCA ont été limités en raison de la forte corrélation entre les variables de nos données.

Dans notre cas, les données de prédiction des résultats de matchs de football ont des variables qui sont très fortement corrélées entre elles, ce qui limite la pertinence des résultats de la PCA. Ainsi, nous ne pouvons pas tirer de conclusions significatives à partir des résultats de la PCA pour améliorer nos prédictions de résultats de matchs de football.

Nous avons ensuite effectué une validation croisée en utilisant des groupes de données chronologiques pour entraîner et tester notre modèle de classification Random Forest. Cette méthode nous permet de mesurer la précision de notre modèle sur des données de test en utilisant des données d'entraînement qui précèdent chronologiquement les données de test. Cette approche est particulièrement importante pour les données chronologiques, car elle permet de valider le modèle sur des données plus récentes et de s'assurer qu'il est capable de prédire avec précision les résultats futurs. En outre, cette méthode permet de réduire les risques d’overfitting du modèle aux données d'entraînement et de s'assurer que les résultats de prédiction sont généralisables aux données futures.

Dans notre cas, il y a plus de victoires à domicile que de victoires à l'extérieur ou de matchs nuls, ce qui peut biaiser notre modèle. Pour éviter cela, nous utilisons la méthode RandomOverSampler pour équilibrer le nombre d'observations pour chaque classe. Cela est fait en créant des copies aléatoires des observations de la classe minoritaire.

2)

Nous avons exploré une seconde approche pour prédire l’issue des matchs en utilisant le deuxième dataframe qui contenait les statistiques moyennes de chaque équipe pour chaque saison. Nous avons créé un nouveau dataframe en regroupant les noms des équipes, leurs statistiques moyennes et l'issue du match. Cependant, en comparaison avec la première approche où nous avons utilisé les caractéristiques des matchs, cette seconde approche n’a pas produit de résultats satisfaisants. En effet, les prédictions réalisées étaient moins précises et moins pertinentes que celles obtenues à partir des statistiques de chaque match. Nous avons des scores autour de 50% ce qui est faible pour la prédiction d’un match.

Sam Vallet
Thomas Sinapi

CONCLUSION/CRITIQUES

Les résultats obtenus sur les prédictions en connaissant les statistiques des matchs sont relativement corrects (70%). Cependant, il est important de souligner que le résultat d'un match ne se résume pas toujours aux statistiques telles que la possession ou le nombre de tirs effectués.

D'un autre côté, les résultats obtenus sur les prédictions sans connaître les statistiques des matchs ne sont pas suffisamment fiables (50%). Cela peut s'expliquer en grande partie par la complexité des données et la difficulté de prédire les résultats dans le monde du football. Le football est un sport où de nombreux facteurs influencent le résultat final, tels que la forme des joueurs, les blessures, les conditions météorologiques, ou encore la stratégie de jeu de chaque équipe.

Pour améliorer la précision des prédictions, il est recommandé d'avoir accès à des données plus complètes, telles que la forme récente des joueurs, la composition des équipes et d'autres facteurs clés qui peuvent influencer le résultat d'un match. En intégrant ces informations supplémentaires dans l'analyse, il est possible de se rapprocher d'une prédiction plus précise et fiable.