Description des données

Les données utilisées pour cette recherche proviennent d'un ensemble de données collecté sur le site Kaggle, qui contient des statistiques sur les performances des joueurs de football lors des matchs de la Ligue des champions de l'UEFA. Les données sont structurées sous forme d'un fichier CSV et comprennent les éléments suivants :

* Rk (Rank) : Il s'agit du classement du joueur en fonction de certaines métriques de performance.
* Player : Ce champ indique le nom du joueur.
* Squad : Ce champ indique le nom de l'équipe à laquelle appartient le joueur.
* Gls (Goals) : Ce champ représente le nombre total de buts marqués par le joueur.
* Sh (Shots) : Il s'agit du nombre total de tirs effectués par le joueur.
* SoT (Shots on Target) : Ce champ indique le nombre de tirs du joueur qui ont été cadrés, c'est-à-dire qui ont atteint le but adverse.
* SoT% (Shots on Target Percentage) : Il s'agit du pourcentage de tirs qui ont été cadrés par rapport au nombre total de tirs.
* Sh/90 (Shots per 90 Minutes) : Ce champ indique le nombre moyen de tirs effectués par le joueur pour chaque tranche de 90 minutes de jeu.
* SoT/90 (Shots on Target per 90 Minutes) : Il s'agit du nombre moyen de tirs cadrés effectués par le joueur pour chaque tranche de 90 minutes de jeu.
* G/Sh (Goals per Shot) : Ce champ représente le ratio entre le nombre de buts marqués et le nombre total de tirs.
* G/SoT (Goals per Shot on Target) : Ce champ indique le ratio entre le nombre de buts marqués et le nombre de tirs cadrés.
* Dist (Average Shot Distance) : Ce champ représente la distance moyenne en yards à partir de laquelle le joueur a tiré.
* FK (Free Kick Goals) : Il s'agit du nombre de buts marqués par le joueur sur coup franc direct.
* PK (Penalty Goals) : Ce champ indique le nombre de buts marqués par le joueur sur penalty.
* PKatt (Penalty Attempts) : Il s'agit du nombre total de tentatives de penalty effectuées par le joueur.
* xG (Expected Goals) : Ce champ représente le nombre attendu de buts que le joueur aurait dû marquer en fonction de la qualité de ses tirs.
* npxG (Non-Penalty Expected Goals) : Il s'agit du nombre attendu de buts que le joueur aurait dû marquer en fonction de la qualité de ses tirs, à l'exclusion des penalties.
* npxG/Sh (Non-Penalty Expected Goals per Shot) : Ce champ indique le ratio d'expected goals non pénalty par tir.
* G-xG (Goals minus Expected Goals) : Ce champ représente la différence entre les buts réels et les expected goals.
* np:G-xG (Non-Penalty Goals minus Non-Penalty Expected Goals) : Ce champ indique la différence entre les buts réels et les expected goals non pénalty.

**Traitement des données**

* Nettoyage des donnes :

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

Le codage des caractères est un mécanisme important pour convertir les données textuelles en un format binaire que les ordinateurs peuvent stocker et traiter. Il existe plusieurs encodages de caractères, chacun présentant des avantages et des inconvénients, selon le contexte d'utilisation.

Le codage "ISO-8859-1", également connu sous le nom de "Latin-1", est un codage de caractères largement utilisé qui prend en charge de nombreuses langues occidentales, notamment l'anglais, le français, l'espagnol et l'allemand. Il est particulièrement adapté à la représentation de documents en langues européennes et est souvent utilisé pour les fichiers texte et les documents sur le web.

Lors de l'utilisation de la bibliothèque pandas pour lire des fichiers CSV, l'encodage "UTF-8" est utilisé par défaut. Cependant, il convient de noter que tous les fichiers ne doivent pas être encodés en "UTF-8". Tenter de lire un fichier en utilisant le codage "UTF-8" peut entraîner une erreur « UnicodeDecodeError » si le fichier a été enregistré avec un codage différent. Dans de tels cas, il est nécessaire de spécifier le codage correct lors de la lecture du fichier pour éviter ces erreurs et assurer une interprétation précise des caractères.

* Transformation des données :

Les données ont été transformées pour faciliter l'analyse. Par exemple, les pourcentages ont été convertis en décimales pour faciliter les calculs.

* Normalisation des données :

Standardisez les données afin que toutes les variables soient sur la même échelle. Ceci est particulièrement important lorsque les variables ont des unités de mesure différentes, car cela peut fausser les résultats de l'analyse.

En résumé, la normalisation des données est une étape importante du prétraitement des données, qui permet de mettre toutes les variables à la même échelle, améliore les performances des algorithmes Machine Learning, prévient les problèmes numériques et facilite l'interprétation des résultats.

* Creation de nouvelles variables :

Dans le cadre du processus d'analyse des données, nous avons créé de nouvelles variables à partir des données existantes pour faciliter l'analyse et extraire des informations supplémentaires.

Dans notre exemple, nous avons créé une variable pour indiquer si un joueur a marqué un penalty ou non. Pour ce faire, nous combinons les colonnes "PK" (nombre de penaltys marqués) et "PKatt" (nombre de tentatives de penalty). Cette nouvelle variable permet d'analyser plus facilement l'efficacité des joueurs aux tirs au but et de distinguer les buts sur penalty des autres buts.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Dans ce code, nous utilisons la méthode pandas astype pour convertir les valeurs booléennes (True/False) en valeurs entières (1/0). La nouvelle colonne "Penalty\_Goal" indiquera si le joueur a marqué un penalty (1 si oui, sinon 0).

* Gestion des outliers :

Dans le cadre du processus de préparation des données pour l'analyse, nous identifions et gérons les valeurs aberrantes. Les valeurs aberrantes sont des valeurs qui sont significativement différentes des autres valeurs de l'ensemble de données. Ils peuvent être causés par des erreurs de mesure, des erreurs de saisie ou des événements inhabituels. Les valeurs aberrantes peuvent avoir un impact important sur les résultats d'analyse et peuvent fausser les tendances générales et les modèles statistiques.

Nous utilisons des techniques statistiques telles que les nuages ​​de points et les boîtes à moustaches pour identifier les valeurs aberrantes. Ces techniques nous permettent de visualiser la distribution des données et d'identifier les points en dehors de la plage des valeurs normales.

Une image contenant texte, capture d’écran, nombre, logiciel

Description générée automatiquementUne image contenant texte, Tracé, capture d’écran, ligne

Description générée automatiquementUne image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

* Segmentation des données :

Dans le cadre du processus d'analyse des données, nous divisons les données en sous-ensembles pour faciliter l'analyse et obtenir des informations plus détaillées. La segmentation des données est une technique permettant de diviser les données en groupes ou éléments plus petits et plus homogènes en fonction de certaines caractéristiques communes.

Dans notre exemple, nous regroupons les données en fonction de l'équipe dans laquelle se trouve le joueur ou de la position du joueur sur le terrain. Cette ventilation nous permet d'analyser les performances des joueurs par équipe ou par position sur le terrain et de comprendre plus précisément les tendances et les modèles dans les données.

Une image contenant texte, capture d’écran, nombre, Police

Description générée automatiquement

**Technique d’analyse :**

* Apprentissage automatique :

Nous utiliserons des modèles d'apprentissage automatique pour prédire les performances futures des joueurs en fonction des statistiques passées et actuelles. Des algorithmes de régression tels que la régression linéaire et la régression logistique seront utilisés pour établir des relations entre différentes variables et prédire les valeurs cibles, telles que le nombre de buts qu'un joueur marquera au cours de la saison à venir. Des algorithmes de classification tels que des machines à vecteurs de support et des forêts aléatoires seront utilisés pour prédire des catégories telles que la probabilité qu'un joueur soit sélectionné pour l'équipe nationale.

Compte tenu de la nature de l'ensemble de données, supposons que nous voulions prédire le nombre de buts (Gls) qu'un joueur est susceptible de marquer en fonction de ses autres statistiques :

Une image contenant texte, capture d’écran, nombre, Parallèle

Description générée automatiquement

Une image contenant texte, Appareils électroniques, capture d’écran, Police

Description générée automatiquement

Du coup le nombre de buts qu’un joueur marquera en fonction de son autre statistique c’est : RMSE : 2.467796546529555e-15

On voit que le RMSE très bas, a première vue, cela semble être une bonne chose. Cependant, cela peut également indiquer un surapprentissage. Le surajustement se produit lorsque votre modèle apprend bien sur les données d'apprentissage et peut donc ne pas fonctionner correctement sur les nouvelles données. Voici les étapes pour vérifier et corriger le surapprentissage :

* Verification :

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Les scores que nous avons obtenus de la validation croisée sont très proches de 1, ce qui indique que notre modèle de régression Ridge s'adapte très bien aux données d'entraînement. La moyenne des scores de validation croisée est également très élevée, à environ 0,9999. En général, un score de validation croisée élevé est un bon signe.

* Correction :

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Un RMSE (Root Mean Squared Error) très faible, comme celui que nous avons obtenu, signifie que notre modèle a une très bonne performance sur l'ensemble de test. Cela indique que les prédictions de notre modèle sont très proches des valeurs réelles. En d'autres termes, notre modèle est capable de prédire avec précision le nombre de buts marqués par les joueurs de l'UEFA Champions League.

En conclusion, nous avons réussi à vérifier et à corriger notre modèle de prédiction du nombre de buts. Le modèle a maintenant une performance améliorée et est moins sujet au surapprentissage. Toutefois, il est toujours recommandé de vérifier les performances du modèle avec d'autres métriques et d'examiner les résidus pour s'assurer que le modèle est robuste et bien ajusté aux données.

* Analyse des séries temporelles :

Analyse descriptive :

Une image contenant texte, capture d’écran, affichage, logiciel

Description générée automatiquementUne image contenant texte, capture d’écran, ligne, Police

Description générée automatiquement

Analyse de la performance des équipes :

Une image contenant texte, capture d’écran, ligne, Police

Description générée automatiquement

Analyse des performances individuelles :

Une image contenant texte, capture d’écran, logiciel, nombre

Description générée automatiquement

Analyse des corrélations :

Une image contenant texte, capture d’écran, Caractère coloré, nombre

Description générée automatiquement

Une matrice de corrélation montre la relation entre toutes les variables d'un ensemble de données. Les valeurs proches de 1 indiquent une corrélation positive, tandis que les valeurs proches de -1 indiquent une corrélation négative. Des valeurs plus proches de 0 indiquent des corrélations plus faibles.

* Visualiser les relations :

Une image contenant texte, capture d’écran, nombre, ligne

Description générée automatiquement

le code affiche un nuage de points pour visualiser la relation entre les buts et les tirs cadrés. Cela aide à voir si les joueurs qui prennent plus de tirs marquent également plus de buts.

* Analyse de régression :

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Le code entraîne un modèle de régression linéaire pour prédire les buts en fonction des tirs cadrés. Cela permet de quantifier la relation entre les tirs cadrés et les buts marqués. Et le modèle prédit raisonnablement bien les buts à partir des tirs cadrés.

* Analyser l’importance des variables :

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

* Effectuer des analyses supplémentaires :

Une image contenant texte, capture d’écran, nombre, Parallèle

Description générée automatiquement

Enfin, le code regroupe les joueurs en fonction des buts et des tirs au but. Cela permet aux joueurs d'être regroupés en fonction de leurs performances dans les buts et les tirs cadrés.

-\* Recommandations pour améliorer les performances sur les buts et les tirs cadrés :

1. Augmentez vos tirs cadrés : La première étape pour marquer plus de buts est d'augmenter vos tirs cadrés. Les joueurs doivent améliorer la précision de leurs tirs et tirer à partir de positions plus favorables. Les entraîneurs peuvent concevoir des exercices spécifiques pour améliorer la qualité de leurs tirs.
2. Créer des opportunités de marquer : les joueurs doivent se placer dans des positions de score en pénétrant dans des zones de danger et en faisant preuve de bon sens. L'équipe peut créer plus d'opportunités de marquer grâce à des tactiques offensives.
3. Améliorez votre capacité de tir : Les grévistes doivent pratiquer le tir de précision, en particulier dans les situations de pression. Les exercices de finition sous pression, les tirs en mouvement et les tirs de l'extérieur de la surface contribuent à améliorer la précision des tirs.
4. Utilisation de centres et de coups de pied arrêtés : les équipes peuvent utiliser des centres et des coups de pied arrêtés pour créer des occasions de marquer. Les joueurs doivent travailler sur le timing, le positionnement et la précision pour convertir ces chances en buts.
5. Analyse vidéo des coups : les joueurs et les entraîneurs peuvent analyser les clips vidéo des jeux pour évaluer les coups et identifier les domaines à améliorer. Cette analyse peut aider à comprendre la sélection de coups, la précision et la technique d'un joueur.

Exploration des données de série temporelle :

Nous allons d'abord visualiser les données pour identifier les tendances, les saisons et les motifs dans les performances des joueurs :

Dans notre fichier y’avais pas la colonne date et temps du coup j’ai ajouté une colonne « Date\_Time » aléatoire