Challenge Data

Football : Qui va gagner ?

par QRT

Connectez vous à votre compte

Description

Challenge compétitif | Sport | Classification | Tabulaires | Plus de 1Go | Niveau intermédiaire

Dates

Commencé le 10 janvier 2024

Contexte

Au cours des vingt dernières années, les professionnels sportifs du monde entier ont adopté une approche fondée sur les données pour adapter leur prise de décision. Les analyses de métriques sportives sont aujourd'hui présentes partout, dans toutes les

Fichiers

Les fichiers sont accessibles une fois connecté et inscrit au challenge

L'organisateur du challenge

Qube Research & Technologies Group is a quantitative and systematic investment manager employing around 300 people with offices in Hong Kong, London, Mumbai, Paris and Singapore. We are a technology driven firm implementing a scientific approach to financial investment. QRT's market presence is global and expands across the largest liquid electronic venues. The

retransmissions en direct, dans l'eSport et même les discussions quotidiennes. Cette évolution a été notamment alimentée par une croissance exponentielle de la richesse de ces données.

Les équipes et les joueurs, de sport et d'eSport, ainsi que les parieurs professionnels font donc appel à la Data Science et au Machine Learning pour prendre les décisions les plus éclairées. Les sites de paris sportifs sont devenus très sophistiqués dans ce domaine, et certains modèles peuvent également être utilisés par les managers d'équipes ou des recruteurs pour constituer des équipes et déployer stratégiquement des joueurs de manière à maximiser leurs chances de victoire.

Le football a été au cœur de la révolution de l'analyse sportive. Tout type de métriques, à la fois historiques et en temps réel, sont aujourd'hui disponibles. Ce challenge s'appuie sur des données de football obtenues auprès de Sportmonks, un fournisseur majeur de données sportives largement utilisées dans ce domaine. Pour plus de détails, n'hésitez pas à consulter le site **sportmonks.com**.

N'hésitez pas à visiter et à vous inscrire à notre forum dédié à **challengedata.qube-rt.com** pour plus d'informations sur le challenge, les données et QRT.

But

Dans le cadre du data challenge QRT de cette année, nous vous proposons un **challenge de prédiction de résultat de match**. Vous recevrez des données historiques réelles au niveau des équipes et des joueurs, et vous devrez prédire quelle équipe gagne ou s'il y a un match nul.

Les données couvrent de nombreuses ligues dans le monde entier ainsi que différentes divisions. Votre objectif est de construire un modèle prédictif riche qui peut fonctionner pour n'importe quelle ligue de football, quel que soit le niveau de compétition ou la situation géographique.

Description des données

Vous disposerez de données au niveau des équipes et des joueurs pour des dizaines de ligues de football.

Les données sont regroupées dans deux fichiers zip, **X_train.zip** et **X_test.zip**, ainsi que dans deux fichiers csv **Y_train.csv** et **Y_train_supp.csv**.

Les fichiers zip contiennent les données d'input, qui sont divisées en 4 fichiers csv. Les données sont séparées en **HOME** et **AWAY**, au grain des équipes et des joueurs. Toutes les métriques proviennent de matchs historiques réels, agrégés depuis le début de la saison ainsi que sur les 5 derniers matchs précédant le match à prédire.

La colonne **ID** représente un ID de match et relie les 4 tables du **X_train**, avec **Y_train** et **Y_train_supp**. Il en va de même pour les données de test.

Les ensembles de données d'input pour les équipes comprennent les trois colonnes d'identification suivantes : **ID**, **LEAGUE** et **TEAM_NAME** (notez que **LEAGUE** et **TEAM_NAME** ne sont pas inclus dans les données de test).

Vous aurez alors les 25 métriques suivantes, déclinées en somme cumulée, moyenne et écart-type :

- 'TEAM_ATTACKS'
- 'TEAM_BALL_POSSESSION'
- 'TEAM_BALL_SAFE'
- 'TEAM_CORNERS'

combination of data, research, technology and trading expertise has shaped our DNA and is at the heart of our innovation and development dynamic. The firm acts as an investment manager managing open-ended Funds used for management of third party capital.

SITE WEB DE L'ORGANISATEUR

- 'TEAM_DANGEROUS_ATTACKS'
- 'TEAM_FOULS'
- 'TEAM_GAME_DRAW'
- 'TEAM_GAME_LOST'
- 'TEAM_GAME_WON'
- 'TEAM_GOALS'
- 'TEAM_INJURIES'
- 'TEAM_OFFSIDES'
- 'TEAM_PASSES'
- 'TEAM_PENALTIES'
- 'TEAM_REDCARDS'
- 'TEAM_SAVES'
- 'TEAM_SHOTS_INSIDEBOX'
- 'TEAM_SHOTS_OFF_TARGET'
- 'TEAM_SHOTS_ON_TARGET',
- 'TEAM_SHOTS_OUTSIDEBOX'
- 'TEAM_SHOTS_TOTAL'
- 'TEAM_SUBSTITUTIONS'
- 'TEAM_SUCCESSFUL_PASSES'
- 'TEAM_SUCCESSFUL_PASSES_PERCENTAGE'
- 'TEAM_YELLOWCARDS'

Les ensembles de données d'input pour les joueurs comprennent les cinq colonnes d'identification suivantes : **ID**, **LEAGUE**, **TEAM_NAME**, **POSITION** et **PLAYER_NAME** (notez que **LEAGUE**, **TEAM_NAME** et **PLAYER_NAME** ne sont pas inclus dans les données de test).

Vous aurez alors les 52 métriques, déclinées en somme cumulée, moyenne et écart-type. Elles sont similaires à celles des équipes bien que plus fines.

Les ensembles de données de sortie sont composés de 4 colonnes :

- ID : identifiant unique de match - correspondant aux ID d'input,
- HOME_WINS,
- DRAW,
- AWAY_WINS,

Le score cible est l'accuracy de la prédiction pour les trois classes [HOME_WINS, DRAW, AWAY_WINS], Il existe donc pour un match trois outputs possibles, [1,0,0]. [0,1,0] et [0,0,1].

Nous avons fourni autant de données que possible dans l'ensemble de train, mais toutes les variables ont été normalisées et les noms des équipes, des joueurs et des ligues ont été supprimés de l'ensemble de test. N'utilisiez pas de données externes sous peine de disqualification.

Un exemple de fichier de soumission contenant des prédictions aléatoires est fourni. Il y a également un notebook benchmark que vous trouverez dans les supplementary_files.

Nous avons inclus une autre target d'entraînement **GOAL_DIFF_HOME_AWAY**, qui est la différence de buts entre l'équipe **HOME** et **AWAY**, dans le fichier **Y_train_supp**.

Disclaimer : Les données fournies sont exclusivement destinées à être utilisées dans le cadre de ce challenge, et toute utilisation de cet ensemble de données à d'autres fins est strictement interdite. Les conditions générales d'utilisation sont applicables : **conditions d'utilisation.**.

Description du benchmark

Il y a deux benchmarks pour ce challenge, le premier est de toujours prédire que l'équipe **HOME** gagne. Cela vous donne une précision d'environ 44 % sur l'ensemble d'entraînement.

Le second benchmark, qui est celui que vous verrez dans le classement, utilise **uniquement les données au niveau des équipes**. Il utilise toutes les colonnes comme

input, et entraîne un modèle de gradient boosting pour prédire si l'équipe **AWAY** gagne ou non. Cela vous donne une précision d'environ 47,5 % sur l'ensemble des données d'entraînement.

N'hésitez pas à utiliser les benchmarks, à construire vos propres modèles et à utiliser des fichiers supplémentaires comme vous le souhaitez. Bon courage pour ce challenge !