Younes OUARTASSI

Rapport Projet de Modélisation de Données
Du jeu de données Soccer

Table des matières

Introduction

Ce rapport présente les résultats d'une analyse des données sur les performances des joueurs de football. L'objectif de cette étude est d'identifier les principaux facteurs qui influent sur le rating global des joueurs et de développer un modèle de prédiction pour estimer ce rating.

Méthodologie

Les données ont été extraites de la base de données MySQL "soccer", comprenant des informations sur les joueurs et leurs attributs. Les étapes suivantes ont été réalisées :

1. **Extraction des Données** : Les données ont été extraites de la base de données MySQL "soccer". Deux tables principales, "player" et "player_attributes", ont été importées dans l'environnement R. Les packages DBI et RMySQL ont été utilisés pour établir la connexion à la base de données et pour exécuter les requêtes SQL nécessaires à l'extraction des données.

2. **Prétraitement des Données** :

   Le prétraitement des données a inclus plusieurs étapes :

   - **Fusion des Tables** : Les tables "player" et "player_attributes" ont été fusionnées en utilisant l'ID du joueur ("player_api_id") comme clé primaire. Cela a permis de combiner les informations démographiques des joueurs avec leurs attributs de performance.
   - **Conversion des Types de Données** : Les colonnes de date ont été converties en format Date et les autres colonnes pertinentes ont été converties en format numérique pour permettre les analyses statistiques et la modélisation.
   - **Calcul de l'Âge des Joueurs** : Une nouvelle variable, l'âge des joueurs, a été calculée en utilisant la date de naissance des joueurs et la date actuelle.
   - **Nettoyage des Données** : Les valeurs manquantes ont été remplacées par la moyenne des colonnes correspondantes.

3. **Analyse Exploratoire des Données (EDA)** :

L'analyse exploratoire des données a été effectuée pour comprendre les relations entre les différents attributs des joueurs et leur rating global. Voici quelques-unes des visualisations créées :

- **Scatter Plot** : Un scatter plot a été utilisé pour visualiser la relation entre la taille, le poids et le rating global des joueurs. Ce graphique a aidé à identifier les tendances générales et les éventuelles anomalies.
- **Boxplot** : Des boxplots ont été créés pour comparer les distributions de taille et de poids parmi les joueurs. Cela a permis de visualiser la médiane, les quartiles et les éventuels outliers dans les données.
- **Histogrammes** : Des histogrammes ont été utilisés pour visualiser la distribution des variables telles que la taille, le poids et l'âge des joueurs. Ces visualisations ont aidé à comprendre la répartition des valeurs et à détecter les schémas de fréquence.
- **Diagramme en Secteurs** : Un diagramme en secteurs a été utilisé pour représenter la répartition du rating global des joueurs, montrant la proportion de joueurs dans chaque catégorie de rating.

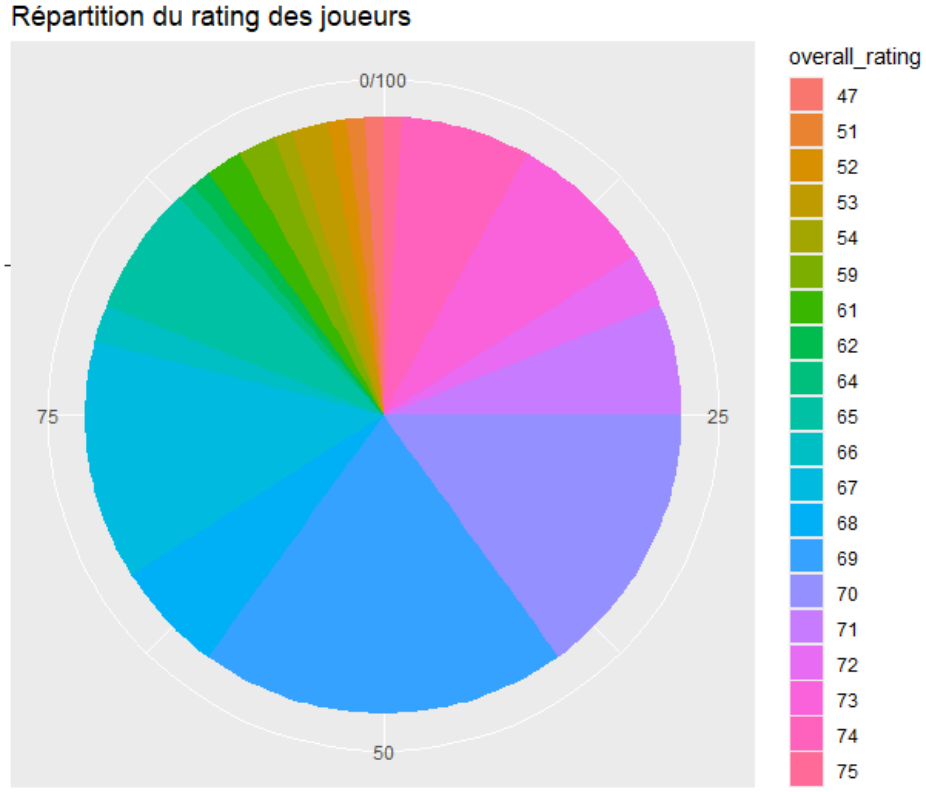

Figure 1 Répartition du rating des joueurs

- **Évolution du Rating** : Un graphique en ligne a été créé pour montrer l'évolution du rating global des joueurs au fil du temps.

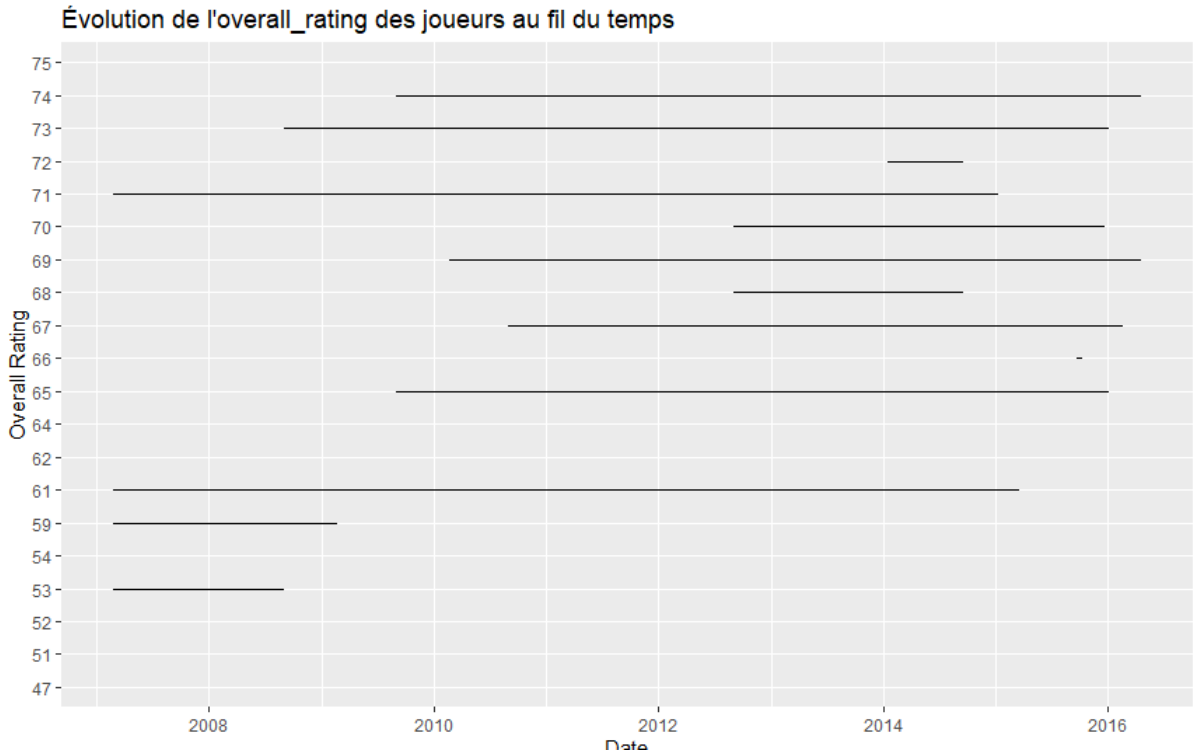

Figure 2Evolution du rating au fils du temps

4. Modélisation Prédictive :

Un modèle de forêt aléatoire a été construit pour prédire le rating global des joueurs en fonction de leurs attributs. Voici les étapes clés de la modélisation :

- **Sélection des Variables** : Les variables pertinentes pour la prédiction ont été sélectionnées, excluant les colonnes non numériques.
- **Division des Données** : Les données ont été divisées en ensembles d'entraînement et de test pour évaluer les performances du modèle.
- **Entraînement du Modèle** : Le modèle de forêt aléatoire a été entraîné sur l'ensemble d'entraînement.
- **Prédictions** : Les prédictions ont été effectuées sur l'ensemble de test, et la précision du modèle a été évaluée à l'aide de la métrique Mean Squared Error (MSE).
- **Importance des Variables** : L'importance de chaque variable dans le modèle a été calculée et visualisée pour identifier les attributs les plus influents.

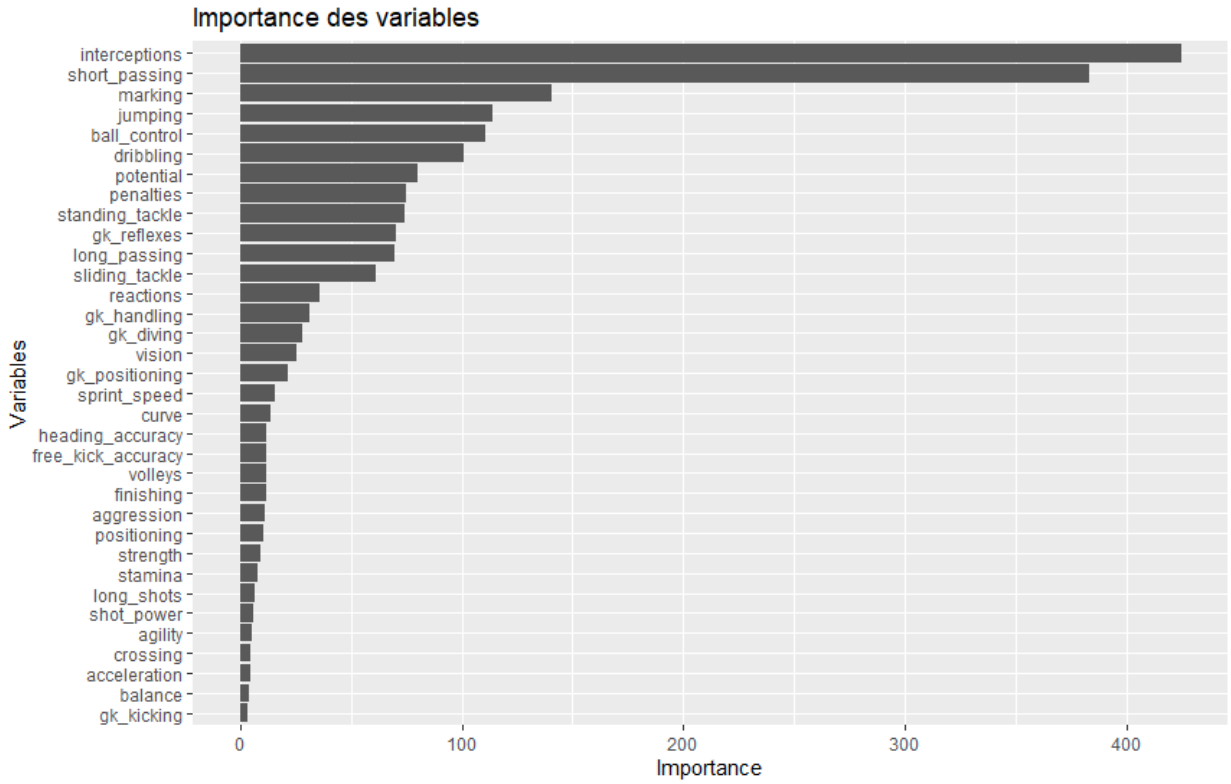

Figure 3importance des variables

Résultats Principaux

1. Analyse Exploratoire des Données :

- **Corrélation entre Variables :** Les visualisations ont révélé une corrélation positive entre certaines variables techniques et physiques des joueurs et leur rating global. Par exemple, des compétences telles que le interceptions et short_passing et le marking ont montré une forte corrélation avec le rating global des joueurs.
- **Tendances Observées :** Une tendance à la hausse du rating global a été observée avec l'augmentation de certaines capacités physiques telles que l'accélération et la vitesse de sprint. Cela suggère que des attributs comme la vitesse et l'agilité jouent un rôle crucial dans l'évaluation globale des performances des joueurs.

2. Modélisation Prédictive :

- **Précision du Modèle :** Le modèle de forêt aléatoire a été capable de prédire le rating global des joueurs avec une précision de 98,24 %. Cette précision indique que le modèle est assez efficace pour estimer le rating des joueurs en se basant sur leurs attributs.

```
>   print(paste("R-squared:", r_squared))
[1] "R-squared: 0.982395461404844"
```

Figure 4precision du modèle

Interprétation des Résultats

Les résultats de l'analyse montrent que les attributs techniques et physiques des joueurs ont une influence significative sur leur rating global. Les joueurs possédant des compétences élevées en interceptions et short_passing tendent à obtenir un rating global plus élevé. Cela met en évidence l'importance de ces compétences spécifiques dans l'évaluation des performances des joueurs.

Plan d'Action

1. Actions Concrètes Basées sur les Insights

- **Adapter les Recommandations de Contenu** : Créer des programmes d'entraînement spécifiques pour améliorer les compétences techniques identifiées comme cruciales.
- **Cibler des Segments Spécifiques** : Identifier et cibler les joueurs jeunes ou moins performants pour des programmes de développement intensif.

2. Développement du Plan pour Implémenter les Recommandations

- **Phases d'Implémentation** : Diviser le plan en phases, commencer par une phase pilote avec un groupe restreint de joueurs.
- **Suivi et Évaluation** : Mettre en place des métriques de suivi pour évaluer l'efficacité des programmes d'entraînement et ajuster en fonction des retours.

Recommandations

Sur la base des résultats de cette analyse, voici quelques recommandations :

1. **Recrutement de Joueurs :** Les clubs de football peuvent utiliser le modèle de prédiction pour identifier les joueurs ayant le potentiel de performance le plus élevé. En se concentrant sur les attributs clés identifiés par le modèle, les clubs peuvent optimiser leurs stratégies de recrutement.
2. **Développement des Joueurs :** Les entraîneurs peuvent se concentrer sur le développement des compétences clés telles que le interceptions et short_passing pour améliorer le rating global des joueurs. En investissant dans l'entraînement de ces compétences, les clubs peuvent potentiellement augmenter la performance globale de leurs équipes.

Limitations de l'Étude

1. **Représentativité des Données :** Les données utilisées dans cette analyse peuvent ne pas représenter l'ensemble de la population des joueurs de football. Par conséquent, les résultats peuvent ne pas être généralisables à tous les joueurs de football.
2. **Facteurs Non Pris en Compte :** Les performances du modèle prédictif peuvent être influencées par des facteurs non pris en compte dans l'analyse, tels que les conditions de jeu, les aspects psychologiques des joueurs, ou les stratégies spécifiques des équipes.

Conclusion

En conclusion, cette analyse a permis d'identifier les facteurs clés influençant le rating global des joueurs de football et de développer un modèle de prédiction précis. Ces résultats peuvent être utilisés par les clubs de football pour prendre des décisions éclairées en matière de recrutement et de développement des joueurs. En se concentrant sur les attributs techniques et physiques identifiés, les clubs peuvent améliorer la performance globale de leurs équipes.