Rapport de projet : Prédiction du classement final de la Premier League

Antoine Merel , **Malak Lahlou**
antoine.merel@polymtl.ca, malak.lahlou-nabil@polymtl.ca

Abstract

L'objectif de cette étude est de prédire le classement et les résultats des matchs de la saison 2021-2022 de la Premier League anglaise de football en utilisant des techniques de Machine Learning. Différents modèles de réseaux de neurones ont été utilisés pour prédire les résultats des matchs et les points cumulés par chaque équipe. Les modèles ont été entraînés sur les données des matchs des saisons 2007-2008 à 2020-2021.

Les erreurs d'apprentissage et de validation des deux modèles ont été comparées à celles de modèles naïfs prédisant toujours des victoires à domicile, des victoires à l'extérieur ou des matchs nuls. Le modèle qui a donné les meilleurs résultats est le Multilayer perceptron (MLP) et c'est donc celui qui a été utilisé pour prédire le classement final de la saison 2021-2022. Bien que les résultats ne soient pas parfaits, ils sont cohérents et suggèrent que des variables additionnelles telles que la qualité des joueurs, les blessures, les transferts, la dynamique des équipes et la forme des joueurs pourraient améliorer les prédictions.

L'étude conclut que la prédiction du classement de la Premier League anglaise de football est une tâche complexe en raison des incertitudes et des variables qui ne peuvent pas toutes être prises en compte, mais que l'intégration de données complémentaires et l'analyse de la dynamique des équipes pourraient aider à construire des modèles plus robustes et plus utiles.

1 Introduction

Dans cette étude, nous nous intéressons à la prédiction des résultats de la saison 2021-2022 de la Premier League (PL) à l'aide de techniques de Machine Learning. La Premier League est l'une des compétitions de football les plus populaires et les plus suivies au monde. Prédire avec précision les résultats des matchs et les classements peut avoir des implications importantes pour les fans, les clubs, les entraîneurs et les parieurs.

1.1 Contexte et motivation

Le football est l'un des sports les plus populaires au monde, et la Premier League anglaise est l'une des compétitions les plus regardées et les plus médiatisées. La prédiction des résultats des matchs et du classement final de la saison est un enjeu majeur pour les fans, les clubs, les entraîneurs et les parieurs. Les techniques de Machine Learning offrent une approche prometteuse pour aborder ce problème complexe, qui est soumis à de nombreuses variables et incertitudes.

1.2 Objectifs de l'étude

L'objectif de cette étude est de développer et d'évaluer des modèles de Machine Learning pour prédire les résultats des matchs de la saison 2021-2022 de la PL, ainsi que le classement final des équipes. Nous explorerons différents modèles de réseaux de neuronnes : des multilayer perceptron (MLP), et des deep neural network (DNN). Ensuite, nous analyserons la performance de ces modèles en termes de précision des prédictions, et nous sélectionnerons celui qui produit les meilleurs résultats pour prédire le classement final de la Premier League, lors de la saison 2021-2022.

1.3 Organisation du rapport

Le rapport est structuré comme suit : la section 2 présente une revue de la littérature sur la prédiction des résultats sportifs et les techniques de Machine Learning appliquées au football. La section 3 décrit la méthodologie utilisée, y compris les données, les modèles de Machine Learning et l'évaluation des modèles. La section 4 présente les résultats obtenus, incluant la performance des modèles et une analyse des prédictions. La section 5 discute des limitations des modèles, des variables additionnelles pour améliorer les prédictions, et des autres approches et modèles possibles. Enfin, la section 6 conclut le rapport et propose des perspectives pour des travaux futurs.

2 Revue de la littérature

2.1 Prédiction de résultats sportifs

La prédiction de résultats sportifs a toujours suscité un grand intérêt, notamment en raison de son importance dans les domaines des paris sportifs et du divertissement. De nombreux travaux ont été menés pour développer des modèles capables de prédire les résultats dans la plupart des sports. Les méthodes de prédiction sont diverses et ont évolué au fil du

temps, passant des approches statistiques traditionnelles aux techniques de machine learning et d'intelligence artificielle.

2.2 Techniques de Machine Learning appliquées au football

L'application de techniques de machine learning à la prédiction des résultats de football est un sujet de recherche en pleine croissance. Plusieurs approches ont été utilisées pour prédire les résultats des matchs de football, notamment les réseaux de neurones, les forêts aléatoires, les machines à vecteurs de support (SVM) et les modèles de régression. Les modèles de machine learning sont généralement entraînés sur des ensembles de données historiques contenant des informations sur les performances passées des équipes, les joueurs, les matchs et d'autres facteurs pertinents.

Herbinet [Herbinet, 2018] a utilisé des techniques de machine learning pour prédire les résultats de matchs de football professionnels. Il a montré que les modèles de machine learning peuvent surpasser les modèles statistiques traditionnels et a proposé des améliorations pour les modèles existants.

Petterson et Nyquist [Petterson and Nyquist, 2017] ont utilisé l'apprentissage profond pour prédire les résultats des matchs de football de la Premier League anglaise. Leur étude a montré que les réseaux de neurones profonds peuvent être efficaces pour prédire les résultats des matchs, bien qu'ils soient sensibles aux variations des ensembles de données d'entraînement.

2.3 Travaux liés à la prédiction de la Premier League

Plusieurs études ont été spécifiquement consacrées à la prédiction des résultats des matchs de la Premier League anglaise. Ulmer et Fernandez [Ulmer and Fernandez, 2014] ont utilisé des modèles de régression pour prédire les résultats de la Premier League et ont constaté que les modèles de régression linéaire multiple et de régression logistique étaient performants dans la prédiction des résultats des matchs.

Sushant [Sushant, 2019] a appliqué des techniques de machine learning pour prédire les résultats des matchs de la Premier League. Il a utilisé un ensemble de données comprenant des statistiques sur les équipes, les joueurs et les matchs pour entraîner des modèles de machine learning et a comparé leurs performances. Ses résultats ont montré que les modèles de machine learning peuvent être efficaces pour prédire les résultats des matchs de la Premier League, avec une précision allant jusqu'à 60 %.

3 Méthodologie

3.1 Données utilisées

Acquisition des données

Nous avons obtenu les données à partir de la source en ligne Football-Data [Football-Data, 2020], qui fournit des données historiques sur les matchs de la Premier League anglaise, les résultats des matchs, les cotes des bookmakers, les statistiques sur les équipes et les joueurs, et bien plus encore. Les données couvrent plusieurs saisons passées, ce qui permet d'entraîner des modèles de machine learning sur un échantillon de données suffisamment large. Pour ce projet, les données des 15 saisons précédentes (2007-2008 à 2021-2022) ont été extraites.

Traitement des données

Une fois les données recueillies, nous avons dû les préparer et les nettoyer avant de pouvoir les utiliser pour l'entraînement et l'évaluation des modèles. Nous avons commencé par identifier et supprimer les données manquantes, inexactes ou incohérentes. Par exemple, certaines valeurs étaient manquantes ou mal saisies pour certains matchs, que nous avons corrigées ou supprimées si nécessaire.

Ensuite, nous avons sélectionné les variables pertinentes pour la prédiction des résultats des matchs, en nous concentrant sur les données telles que les statistiques de performance passées des équipes, les cotes des bookmakers et les informations sur les joueurs. Nous avons également dérivé certaines variables à partir des données existantes, comme la forme récente d'une équipe ou les différences de classement, afin d'améliorer la précision des modèles de prédiction.

Enfin, nous avons normalisé les variables numériques pour garantir qu'elles étaient sur la même échelle et qu'elles contribuaient également au modèle de prédiction. Après ce traitement des données, l'ensemble de données final était prêt à être utilisé pour l'entraînement et l'évaluation des modèles de Machine Learning.

3.2 Modèles de Machine Learning

Nous avons testé plusieurs modèles de réseaux de neuronnes, et pour chacun, les hyperparamètres, tels que le nombre de couches cachées, le nombre de neurones par couche et les fonctions d'activation, ont été optimisés par validation croisée et recherche sur grille.

Multilayer Perceptron

Nous avons commencé par construire un réseau de neurones avec une architecture comprenant trois couches. La couche de sortie produit une prédiction pour le résultat du match (victoire, nul, défaite). Les couches cachées sont composées de neurones avec des fonctions d'activation non linéaires, telles que la fonction ReLU. Nous avons utilisé la rétropropagation et l'optimisation par descente de gradient stochastique pour entraîner le réseau de neurones, en minimisant une fonction de coût appropriée pour le problème de classification multiclasse.

Deep Neural Network

Ensuite, nous avons tenté d'utiliser des réseaux de neuronnes plus profonds que le MLP. Malheureusement, notre modèle de DNN produit des résultats très mauvais, probablement car nous avons essayé nous mêmes quelques valeurs d'hyperparamètres, et que nous n'avons pas utilisé des valeurs trouvées dans la litérature, pour lesquelles il a été démontré que les résultats étaient bons, comme pour les MLP.

NET

Afin d'améliorer les performances des DNN, nous avons essayé d'utiliser un réseau de neuronnes décrit dans l'article [Batista, 2023]. Dans cet article, l'auteur nous présente un réseau de neuronne qui lui permet d'obtenir une précision de 75 % de précision dans la prédiction de matchs de football.

Nous avons donc décidé d'expérimenter son modèle dans le cadre de notre projet.

3.3 Évaluation des modèles

Avant de comparer les performances de nos modèles, nous avons établi des baselines en utilisant des modèles naïfs prédisant le même résultat pour tous les matchs de notre jeu de données d'entraînement :

- Modèle prédisant que des victoires à domicile: Erreur de 57,11%
- Modèle prédisant que des victoires à l'extérieur: Erreur de 66,05%
- Modèle prédisant que des matchs nuls: Erreur de 76,84%

Le modèle prédisant uniquement des victoires à domicile présente l'erreur la plus faible, ce qui est cohérent avec le fait que les équipes à domicile gagnent environ 1,5 fois plus souvent que celles à l'extérieur. Les matchs nuls sont moins fréquents, d'où le taux d'erreur élevé du modèle prédisant uniquement des matchs nuls.

Après avoir établi ces baselines, nous avons entraîné plusieurs modèles de MLP, de DNN, et même notre propre réseau, en optimisant leurs hyperparamètres localement. Le modèle qui a donné les meilleurs résultats est MLP. Ce modèle donne une erreur d'apprentissage de 37,59%, et une erreur de validation de 39,29%. Il est bien plus performant que le modèle naïf prédisant uniquement des victoires à domicile.

Nous séléctionnons donc le réseau de neurones MLP avec les hyperparamètres suivants:

- Profondeur du réseau: 3
- Taille des couches : 3
- Fonction d'activation: Logistique
- Solveur Adam
- Itérations maximums : 500

Ce réseau de neurones prédit le résultat de chaque match et retourne une variable catégorielle à 3 modalités. L'erreur de généralisation du modèle, obtenue avec les données de la saison 2021-2022, est de 35,26%. Les points sont ensuite cumulés pour chaque équipe en fonction des résultats des matchs prédits pour obtenir le classement final de cette saison.

4 Résultats

4.1 Analyse du classement prédit

Le classement obtenu par notre modèle le plus performant, le réseau de neurones MLP, est présenté ci-dessous :

		Home Win	Away Win	Home Draw	Away Draw	Home Lose	Away Lose	Points
1	Man City	18	16	0	0	1	3	102
2	Liverpool	18	15	0	0	1	4	99
3	Tottenham	16	13	0	0	3	6	87
4	Chelsea	16	12	0	1	3	6	85
5	Man United	15	10	0	0	4	9	75
6	Arsenal	14	9	0	0	5	10	69
7	Leicester	13	9	0	0	6	10	66
8	Aston Villa	11	10	0	0	8	9	63
9	West Ham	11	9	0	0	8	10	60
10	Brighton	8	11	1	0	10	8	58
11	Crystal Palace	8	9	0	0	11	10	51
12	Newcastle	10	6	0	0	9	13	48
13	Leeds	5	10	0	0	14	9	45
14	Brentford	9	6	0	0	10	13	45
15	Wolves	8	6	1	0	10	13	42
16	Watford	7	6	0	0	12	13	39
17	Everton	6	4	1	1	12	14	32
18	Burnley	8	2	0	0	11	17	30
19	Southamton	5	4	0	1	14	14	28
20	Norwich	3	1	0	0	16	18	12

Figure 1: Prédiction du classement de la PL saison 2021-2022

Le véritable classement de la saison 2021-2022 de la PL est :

		Home Win	Away Win	Home Draw	Away Draw	Home Lose	Away Lose	Points
1	Man City	15	14	2	4	2	1	93
2	Liverpool	15	13	4	4	0	2	92
3	Chelsea	13	9	1	4	5	6	74
4	Tottenham	9	12	7	4	3	3	71
5	Arsenal	10	6	5	5	4	8	69
6	Man United	13	9	2	1	4	9	58
7	West Ham	10	4	4	6	5	9	56
8	Leicester	6	7	5	1	8	11	52
9	Brighton	9	7	5	3	5	9	51
10	Wolves	5	7	7	8	7	4	51
11	Newcastle	7	4	8	7	4	8	49
12	Crystal Palace	8	5	6	4	5	10	48
13	Brentford	4	5	6	5	9	9	46
14	Aston Villa	7	6	3	4	9	9	45
15	Southamton	7	8	3	3	9	8	40
16	Everton	2	4	2	3	15	12	39
17	Leeds	9	2	2	4	8	13	38
18	Burnley	5	2	6	8	8	9	35
19	Watford	6	3	7	6	6	10	23
20	Norwich	3	2	3	4	13	13	22

Figure 2: Classement réel de la PL saison 2021-2022

Dans l'ensemble, notre modèle semble produire un classement plutôt cohérent. Cependant, il est important de noter que notre modèle n'a prédit que 3 matchs nuls pour les 380 matchs de la saison, alors qu'il y en a eu 88 dans la réalité.

Analyse détaillée des résultats

En examinant de plus près les résultats, on constate que notre modèle ne prédit parfaitement la position que de quatre équipes (Manchester City, Liverpool, Burnley et Norwich). On pourrait se dire que les résultats ne sont vraiment pas terrible. Cependant, on voit aussi qu'il prédit avec une précision parfaite ou à une position près pour 14 équipes sur les 20. En particulier, c'est le cas pour les six premières positions du classement. Ces équipes correspondent également aux six meilleures équipes de l'histoire récente du championnat, connues sous le nom de "Big 6" (Manchester United, Liverpool, Arsenal, Chelsea, Tottenham, Manchester City). L'écart moyen de position est de 1,7 et l'écart-type est de 1,71 positions. Deux des trois équipes reléguées (finissant parmi les trois dernières) ont été correctement identifiées, Burnley et Norwich.

Concernant les points obtenus par chaque équipe, notre modèle prédit avec une précision à 5 matchs près (soit une différence inférieure à 9 points, ce qui correspond à 3 victoires) pour 14 équipes. L'écart moyen de points est de 6,4 et l'écart-type est de 4,35 points.

4.2 Limites des prédictions

Notre modèle prédit un championnat moins compétitif que la réalité, avec un écart de 90 points entre le premier et le dernier du classement, contre 71 points dans la réalité.

Notre modèle fait plus d'erreurs dans le "ventre mou" du classement (de la 7^e à la 17^e place). Pour les six premières places et les trois dernières, il ne se trompe de plus d'une place que pour une équipe (à la dix-neuvième place).

En conclusion, notre modèle de réseau de neurones a réussi à prédire un classement relativement précis pour la saison 2021-2022 de la Premier League, bien que certaines limitations demeurent, en particulier pour les équipes moins représentées dans notre échantillon d'apprentissage.

5 Discussion

Les résultats obtenus montrent que notre modèle est capable de prédire les résultats des matchs de la première ligue anglaise de football avec une certaine cohérence. Cependant, ces prédictions ne sont pas parfaites et peuvent être améliorées. Plusieurs facteurs peuvent expliquer les erreurs de prédictions et les limites du modèle actuel.

Tout d'abord, il est possible que notre modèle ne prenne pas en compte toutes les variables pertinentes pour prédire les résultats des matchs. Les données utilisées pour entraîner et tester le modèle peuvent ne pas être suffisamment représentatives de la complexité du jeu. Par exemple, la qualité des joueurs, l'état de forme, les blessures et les suspensions, ainsi que les tactiques utilisées par les entraîneurs peuvent avoir un impact significatif sur les résultats.

De plus, notre modèle peut souffrir de problèmes d'overfitting ou d'underfitting, c'est-à-dire qu'il peut être trop adapté aux données d'entraînement et ne pas généraliser correctement aux nouvelles données, ou qu'il peut être trop simple pour capturer la complexité réelle des données. Des techniques de régularisation, la sélection de modèles et l'ajustement des hyperparamètres pourraient être utilisées pour remédier à ces problèmes.

En outre, le football est un sport où l'incertitude et la chance jouent un rôle important. Même avec un modèle parfaitement entraîné et ajusté, il est probable que certaines prédictions seront incorrectes en raison de la nature imprévisible du jeu. Il est donc essentiel de considérer les prédictions du modèle comme des estimations probabilistes plutôt que des résultats garantis.

Enfin, il est possible que des modèles plus avancés, tels que les réseaux de neurones profonds et les modèles de séries temporelles, puissent offrir de meilleures performances de prédiction que le modèle utilisé dans cette étude. Ces modèles pourraient être plus à même de capturer les relations complexes entre les variables et de prendre en compte les dynamiques temporelles des données.

Dans l'ensemble, malgré les limites et les erreurs de prédiction, notre modèle fournit un point de départ intéressant pour l'analyse des résultats des matchs de la première ligue anglaise de football. Les perspectives futures pourraient inclure l'intégration de variables supplémentaires et l'exploration de modèles plus avancés pour améliorer les performances de prédiction. En outre, les prédictions du modèle pourraient être utilisées pour orienter les décisions des bookmakers, des entraîneurs et des analystes sportifs, tout en gardant à l'esprit les incertitudes inhérentes au jeu.

6 Conclusion et ouverture

6.1 Conclusion

Notre étude a mis en évidence que le modèle utilisé produit des résultats cohérents mais imparfaits pour prédire le classement de la saison 2021-2022 de la Premier League.

Cela souligne la complexité de la tâche, étant donné la multitude d'incertitudes et de variables qui affectent les résultats des matchs de football. En effet, notre modèle, basé sur l'historique des performances des équipes, ne peut pas prendre en compte tous les facteurs qui influencent le déroulement et l'issue d'un match.

6.2 Perspectives futures

Les perspectives futures pour améliorer notre modèle et obtenir des prédictions plus précises peuvent inclure l'intégration de variables additionnelles et l'exploration de nouvelles approches pour modéliser les performances des équipes. Voici quelques pistes à explorer pour renforcer la prédiction de notre modèle :

- Effectif des équipes : Intégrer la qualité des joueurs qui composent les différentes équipes, notamment le niveau international ou national des joueurs, les blessures, et les transferts. En tenant compte de ces facteurs, le modèle pourrait être plus précis dans ses prédictions.
- Dynamique des équipes : Prendre en compte la dynamique et la forme des équipes et des joueurs pour mieux refléter l'impact des performances récentes sur les matchs à venir. Par exemple, une équipe en pleine confiance après plusieurs victoires consécutives devrait être plus encline à remporter son prochain match.
- Impact de l'entraîneur : Les stratégies et tactiques déployées par les entraîneurs peuvent influencer significativement le résultat des matchs. Intégrer ces éléments pourrait renforcer la précision du modèle.
- Enchaînement des événements : Prendre en compte les événements survenus lors des matchs précédents et leur influence sur les prochains matchs pourraient aider à affiner les prédictions.
- Approches alternatives : Explorer d'autres méthodes de modélisation, telles que les modèles basés sur les réseaux de neurones ou les modèles ensemblistes, pour comparer leurs performances avec notre modèle actuel et potentiellement découvrir des approches plus performantes.

En intégrant ces éléments supplémentaires et en explorant de nouvelles approches, nous pourrions développer un modèle plus robuste et précis pour prédire le classement

des équipes dans la première ligue anglaise de football. Ces améliorations pourraient également s'étendre à d'autres compétitions et ligues pour fournir des prédictions plus fiables dans le domaine du football en général.

References

[Batista, 2023] André Luiz França Batista. Pytorch neural networks to predict matches results in soccer championships: Part 2. https://medium.com/@andreluiz_4916/pytorch-neural-networks-to-predict-matches-results-in-soccer-championships-part-ii-3d02b2ddd538, 2023.

[Football-Data, 2020] Football-Data. England football results betting odds. https://www.football-data.co.uk/englandm.php, 2020.

[Herbinet, 2018] Corentin Herbinet. Predicting football results using machine learning techniques. Available at https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf, 2018.

[Petterson and Nyquist, 2017] Daniel Petterson and Robert Nyquist. Football match prediction using deep learning. Available at http://publications.lib.chalmers.se/records/fulltext/250411/250411.pdf, 2017.

[Sushant, 2019] Deepanshu Sushant. Premier league match result prediction using machine learning. Available at http://ir.juit.ac.in:8080/jspui/handle/123456789/7597, 2019.

[Ulmer and Fernandez, 2014] Ben Ulmer and Matthew Fernandez. Predicting soccer match results in the english premier league. http://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf, 2014.