h e g

Haute école de gestion
Genève

Analyse et optimisation de l'expected goal: application au machine learning

Travail de Bachelor HES réalisé en vue de l'obtention du Bachelor par :

David Paulino

Conseillers au travail de Bachelor :

Pr Alexandros KALOUSIS

Dr Nils SCHÄTTI

Genève, le 6 mai 2023

Haute école de Gestion de Genève (HEG-GE)

Filière Informatique de gestion

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

Table des matières

1 Introduction

1.1 Introduction à la problématique

Lorsque les premiers sports sont apparus, l'information la plus importante était le score et le vainqueur de la confrontation. Au fur et à mesure, plus d'informations sur les matchs sont venues s'ajoutées. Le nombre de tirs dans un matchs par équipes, le nombre de passes, le nombre de tirs cadrés, la possession du ballon le pourcentage de passes réussies dans le football, le nombre de passes décisives et d'autres sont venus s'ajouter aux statistiques dans le football. Le pourcentage de réussite aux lancers francs, le nombre de rebonds, le nombre de passes décisives, le pourcentage de réussite aux tirs, le nombre de fautes, le nombre de minutes jouées, le pourcentage de réussite à 3 points et d'autres sont venus s'ajouter aux statistiques dans le basketball. Ces statistiques sont également devenus personnelles à chacun des joueurs. On peut également compter pour le baseball le nombre de fois qu'un joueur était au bâton, son nombre de double, de triples, son nombre de buts et bien d'autres.
C'est d'ailleurs dans le baseball que l'on peut retrouver la première utilisation de statistiques avancées pour établir des stratégies. En effet, au début des années 1970, le joueur des Baltimore Orioles a développé une analyse statistiques pour choisir le meilleur alignement possible pour son équipe de départ. Cependant, il n'a pas pu l'utiliser à ce moment-là puisque le président de sa franchise n'avait pas confiance. C'est qu'à partir de 1984 où il fut le coach des New York Mets qu'il a pu mettre en place son analyse statistiques avancées pour établir le meilleur choix pour son équipe de départ. [4] Deux saisons plus tard, il remporte la Série mondiale 1986 [1]. Les Mets étaient situés à la dernière place de leur conférence avant l'arrivée de Davey Johnson et son management orienté sur les statistiques.
Après cette réussite, les autres franchises de la MLB [2] ont également commencé à adopter l'analyse de statistiques dans le sport et cela a également été populaire dans les autres sports avec par exemple Daryl Morey qui a été le premier coach analyste statistiques recruté chez les Rockets de Houston en NBA en 2007. [1] Les franchises de la NBA [3] ont par la suite également

1. En MLB, la Série mondiale est la série finale qui permet de déterminer qui est l'équipe championne de la ligue.
2. Ligue majeure de baseball
3. Ligue nationale de basketball

adopté une approche managériale statistique. On constate alors que cette culture de la statistique dans le sport provient des États-Unis.
Il est désormais important d'amener l'arrivée des expected goals. L'une des premières études sur un modèle d'expected goals vient d'Alan Ryder qui a publié une étude sur la qualité des tirs effectuées dans des matchs de hockey. [8] Ce dernier a pu analyser les différentes circonstances lors d'un tir et développer un modèle qui prédit la probabilité d'un tir selon les circonstances de ce tir. Dans le football, l'une des premieres études sur l'expected goal vient de Richard Pollard, Jake Ensum et Samuel Taylor qui ont analysé les facteurs qui influent la chance de marquer un but. [7]
La problématique de ce travail est donc de pouvoir analyser et optimiser l'expected goal.
Il semble maintenant important de savoir ce qu'est l'expected goal. L'expected goal [4] est une métrique qui permet de déterminer la probabilité qu'un tir soit transformé en but selon les données de ce tir [2]. Un tir qui a un xG de 0.4 a une probabilité de 40% d'être transformé en but. Un tir avec un xG à 1 est la plus grande valeur possible et aurait donc 100% de chance d'être transformé en but. [5] [6]
Pour observer ce qu'est réellement un xG, nous allons l'observer avec l'emplacement de deux tirs sur un terrain. Le schéma 1 montre un terrain de football avec deux emplacements de tirs. Par exemple, le tir en rouge aurait un xG de 0.1 et le tir en vert aurait un xG de 0.5 [6].

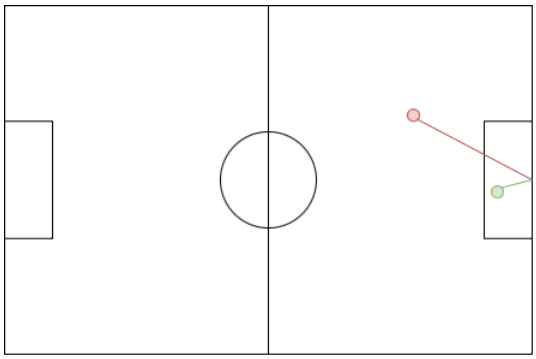

FIGURE 1 – Emplacement de deux tirs sur un terrain de foot

4. Très souvent réduit par xG
5. Il est important d'indiquer qu'il est très rare qu'un xG d'un tir soit égal à 1 mais il va généralement sans rapprocher fortement selon ses paramètres.
6. Ces métriques sont que des exemples et n'ont pas été produites par un modèle.

Maintenant que nous avons vu ce qu'est un xG, il semble pertinent de décrire l'utilisation de cette métrique dans le football actuel.

1.2 Intérêt de la problématique

Cette problématique est très intéressante puisque c'est une donnée qui est dernièrement très populaire dans le monde du football. Elle donne plus d'informations sur le match que les autres statistiques d'un matchs (possession, nombre de passes, nombre de tirs cadrés, nombre de buts, etc.). Par exemple en ce qui concerne la possession, une équipe peut avoir la possession du ballon pendant 70% du match mais ne pas marquer de buts, ni même être dangereuse avec le ballon. Le nombre de tirs cadrés, s'ils sont effectuées tous à l'extérieur de la surface de réparation, ne sont pas forcément dangereux. L'expected goal permet de donner une valeur à chaque tir et de déterminer si un tir a une chance d'être transformé en but.

Ensuite, certaines personnes utilisent cette métrique pour leurs paris sportifs. Ils observent la métrique lors des derniers matchs et la compare avec le score réel du match. Si la différence est grande, cela peut permettre de constater un manque de réalisme ou un sur-régime d'une des deux équipes. [10]

D'autres personnes utilisent cette donnée pour analyser les performances de leurs joueurs et de leurs équipes. Par exemple, la comparaison des xG et du nombre de buts marqués d'un joueur sur une période donnée peut permettre de déterminer la dangerosité et la capacité d'un buteur à terminer des actions. [6] Les xG peuvent aussi être utilisés pour analyser les performances d'une équipe dans des situations bien précises. Une équipe qui possède un haut xG sur des contre-attaques montre qu'elle est très dangereuse sur ce type de situation. [2]

1.3 Questions que l'on souhaite répondre dans ce travail

Il y a plusieurs questions qui peuvent se poser dans ce travail. Je vais donc les lister.

- Comment est calculé l'expected goal ?
- Quels sont les paramètres qui influencent le plus l'expected goal ?
- Quel est le meilleur modèle pour prédire l'expected goal ?

— Quels sont les variables qui peuvent être ajoutée pour améliorer le résultat de la prédiction ?
— Quel est le meilleur modèle pour prédire l'expected goal en utilisant les variables ajoutées ?

1.4 Intérêt de ces questions

Le but de ces questions est de comprendre comment implémenter un modèle qui permet de prédire l'expected goal. Grâce à cela, on pourra ensuite connaître quels sont les variables obligatoires pour implémenter un tel modèle et quels sont les paramètres qui influencent le plus l'expected goal. Ensuite, on pourra voir quel est le meilleur modèle pour prédire l'expected goal et quels sont les variables qui peuvent être ajoutées pour améliorer le résultat de la prédiction. Enfin, on pourra voir quel est le meilleur modèle pour prédire l'expected goal en utilisant les variables ajoutées.

2 Synthèse des travaux existants

Parmi les travaux existants, il y a le travail de David Sumpter [9] qui permet de comprendre comment implémenter un modèle d'expected goal. Ce travail va être une bonne base pour implémenter le premier modèle.
Ensuite, le travail nommé "Expected goals in soccer :explaining match results using predictive analytics" de Eggels [3] est un travail très détaillé qui permet d'avoir des pistes d'améliorations et qui explique bien les différentes étapes pour implémenter un modèle d'expected goal. Ce travail va être très utile pour permette d'apporter des améliorations au modèle.

3 Dataset

3.1 Présentation du dataset

Comme indiqué dans le section 2, le travail de David Sumpter [9] va être utilisé comme base pour implémenter le premier modèle. Il utilise un dataset qui contient les données de Wyscout. Wyscout est une entreprise qui fournit des données sur le football. Elle fournit des données sur les matchs, les joueurs, les équipes, les compétitions, etc. Il n'a pas été possible pour moi d'accéder aux données de Wyscout. J'ai donc utilisé un dataset public qui

est un sample du dataset de Wyscout. [5]
Cette source permet d'avoir l'information des événements lors d'un match de foot, comme par exemple les passes, les tirs, les fautes, etc. Cela permet d'avoir les informations nécessaires pour calculer l'expected goal. Un autre avantage de ce dataset est qu'il est déjà nettoyé et qu'il est facilement utilisable. Il est aussi très complet puisqu'il contient des données sur les matchs de 5 championnats européens (Angleterre, Espagne, Italie, Allemagne et France) de la saison 2017-2018. Il contient aussi des données sur les matchs de la coupe du monde 2018 et de l'Euro 2016.

3.2 Events

3.3 Players

Références

[1] Daryl Morey's 13-year run with the Rockets summed up in 5 incredible stats.

[2] xG Explained | FBref.com.

[3] Hph Harm Eggels. Expected goals in soccer :explaining match results using predictive analytics.

[4] Ziff Davis Inc. *PC Mag*. Ziff Davis, Inc.

[5] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. 6(1) :236.

[6] Luke Petty. What is expected goals ? Expected goals explained.

[7] Richard Pollard, Jake Ensum, and Samuel Taylor. Estimating the probability of a shot resulting in a goal : The effects of distance, angle and space. 2.

[8] Alan Ryder. Isolating Shot Quality - Hockey Analytics.

[9] David Sumpter. Fitting the xG model — Soccermatics documentation.

[10] David Tennerel. Bien utiliser les expected goals (xg) pour vos paris sportifs.