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 14 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. [5] 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. [9] 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. [8] 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] [7]

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].

Par ailleurs, les xG se sont tellement développés que des métriques dérivées ont été créées. On peut par exemple citer le xA qui est l'expected assist. C'est une métrique qui permet de déterminer la probabilité qu'une passe soit transformée en passe décisive selon les données de cette passe. [2] Également, les xGA qui est les expected goals against. Cette métrique permet de déterminer la probabilité qu'un tir soit transformé en but selon les données de ce tir mais pour l'équipe adverse. [7]
Il y en a également d'autres comme les expected points qui sont les points qui est le nombre de points qu'une équipe devrait avoir gagnés basé sur les données relatives aux xG. D'autres dérivées sont indiquées sur l'article de

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 s'en rapprocher fortement selon ses paramètres.
6. Ces métriques sont que des exemples et n'ont pas été produites par un modèle.

Pinnacle écrit par Luke Petty. [7]

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

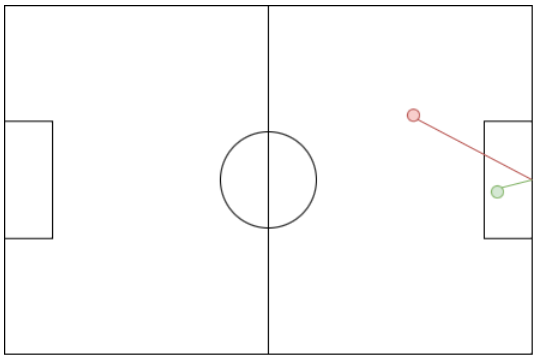

FIGURE 1 – Emplacement de deux tirs sur un terrain de foot

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.

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. [11]

Les analystes de données des équipes 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. [7] 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] L'une des choses les plus intéressante de cette métrique est qu'elle peut être utilisée pour analyser les forces et faiblesses des équipes. Par exemple, une équipe peut observer que ses xG sont très faibles lorsqu'elle tente des centres. Cela peut lui permettre de trouver son identité de jeu et de décider d'abandonner cette stratégie. Cette métrique est tout autant valable pour analyser les forces et faiblesses d'une équipe adverse. Dans la même situation, si une équipe vient à affronter l'équipe qui a dû mal à produire des xG sur des centres, elle peut décider de la forcer à jouer sur des centres en libérant de l'espace sur les côtés par exemple.

Cette métrique est également utilisée pour aider les recruteurs à juger les performances de finition d'un joueur [4]. Nous avons vu précédemment qu'il existait des dérivées des xG comme les xA [7]. Les différentes dérivées peuvent permettre de juger les performances et qualités d'un joueur qui se trouve dans un poste bien précis. Par exemple, un milieu de terrain ou un défenseur peut être jugé sur ses xA et un attaquant sur ses xG.

1.3 Questions que l'on souhaite répondre dans ce travail

Il y a deux grandes questions à répondre dans ce travail.

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

Le but de ce travail est de comprendre quelles sont les variables qui influences le plus les xG. Cela permettra de comprendre quelles sont les variables les plus importantes pour prédire les xG. Grâce à cela, il est possible pour un analyste de données d'une équipe de football de savoir quelles sont les facteurs qui influencent le plus la qualité d'un tir. Cela peut permettre de déterminer les forces et faiblesses d'une équipe et de savoir sur quels aspects travailler pour améliorer les performances de l'équipe.

7. Passes décisives attendues

Le deuxième objectif de ce travail est de trouver le meilleur modèle pour prédire les xG. En effet, le but sera d'avoir le modèle le plus performant et de comparer les différents modèles qui peuvent être utilisée pour établir cette métrique. Il faut également veiller à ce que le modèle ne fasse pas de sur-apprentissage. Il est donc important de faire attention à la complexité du modèle et de trouver le meilleur compromis entre la complexité et la performance du modèle.

2 Plan du document

Concernant le déroulement de ce travail, il y a plusieurs étapes. Tout d'abord, il y a la recherche des travaux existants sur le sujet. Parmi les travaux existants, je vais chercher à savoir quels datasets ont été utilisées ainsi que les résultats obtenus. Cela me permettra de comparer mes résultats avec les résultats obtenus dans les autres travaux.

La suite sera de trouver un dataset avec les informations nécessaires pour implémenter le modèle. Une fois ce dataset trouvé, il va falloir le documenter. En effet, il est important de comprendre ce que chaque attribut représente. L'étape suivante est également importante puisque le but sera de visualiser les données. Cela permettra de voir si les données sont exploitables et si elles sont cohérentes. Cela pourra également nous indiquer si des biais seront présents dans le modèle.

Une fois que les données sont documentées et visualisées, il va falloir les préparer. En effet, il pourrait y avoir des données manquantes mais qui sont disponibles après un traitement. Il pourrait également y avoir des données qui ne sont pas exploitables et qui doivent être supprimées. Par exemple, l'ID de la base de données d'un tir pourrait être supprimée car il n'appporte rien pour la prédiction de l'expected goal.

Ensuite, on pourra commencer à implémenter le modèle et observer les facteurs qui influencent le plus sa prédiction du xG. C'est également à ce moment-là qu'il faudra comparer les différents modèles pour voir lequel est le plus performant pour prédire les xG. Il sera également intéressant de voir quelles variables peuvent être ajoutées au modèle pour améliorer sa précision.

3 Synthèse des travaux existants

Le premier travail repertorié sur les xG et la qualité d'un tir est celui de Richard Pollard, Jake Ensum et Samuel Taylor [8]. Dans ce travail datant de 2004, la seule information indiquée concernant le dataset est que les données proviennent de la Coupe du monde 1986 et de celle de 2002. Le modèle a été implémenté en utilisant une régression logistique. Le nombre de tirs répertoriées dans ce travail est de 1096. La conclusion de ce travail est que les 3 facteurs les plus influents pour la prédiction des xG sont :

— La distance entre le tireur et le but
— L'angle du but en fonction de la position du tir
— L'espace entre le tireur et le défenseur le plus proche

Le résultat final de l'analyse de la régression logistique de ce travail ressemble à cela.

Predictor	**Coefficient**	**z**	**p**	**ratio**	**Lower**	**Upper**
Constant	0.3771	1.20	0.229			
Distance	-0.1586	-9.51	0.000	0.85	0.83	0.88
Angle	-0.0222	-3.81	0.000	0.98	0.97	0.99
Space	0.7991	3.22	0.001	2.22	1.37	3.62

TABLE 1 – Résultat de la régression logistique du travail de Pollard, Ensum et Taylor

Le deuxième travail est celui de Izzatul Umami, Deden Hardan Gutama et Heliza Rahmania Hatta [12]. Ce travail utilise les données de Wyscout des 5 championnats majeurs en Europe de la saison 2019-20. Ces derniers ont décidés de prendre comme données :

— La distance
— L'angle
— Si le tir est un tir de la tête ou pas

Dans le dataset, 32000 tirs ont été utilisées pour la création du modèle. Comme pour le travail précédent, la régression logistique a été utilisée. Il est également indiqué qu'une séparation du dataset a été faite pour avoir des données d'entraînement et de tests. Dans leurs tests du modèle, il est indiqué que le but est de faire de la classification pour de futures instances, il faut donc utiliser un seuil. Suite à l'utilisation d'un seuil pour la classification, une matrice de confusion a été faite pour ensuite calculer la spécificité et

la sensibilité du modèle. Ce principe de sensibilité et spécificité permet de choisir la meilleure performance selon le contexte d'utilisation du modèle.

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative} \tag{1}$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \tag{2}$$

La sensibilité permet de voir la capacité du modèle à prédire correctement les tirs qui sont des buts. De l'autre côté, la spécificité permet de voir la capacité du modèle à prédire correctement les tirs qui ne sont pas des buts. Comme indiqué dans le travail de Umami, Gutama et Hatta, la spécificité est plus importante que la sensibilité selon le contexte. Dans le cas où le modèle est utilisé pour prédire un cancer, nous allons chercher à avoir une meilleur sensibilité pour éviter de passer à côté d'un cancer. [12] Cela leur permet de savoir comment choisir le seuil pour la classification. En conclusion de leur travail, ils ont obtenu une sensibilité de 0.9671945701357466 et une spécificité de 0.19034406215316316 pour un seuil 0.02. Ils indiquent finalement que le modèle de xG est plus performant si l'on prend la distance et l'angle en compte plutôt que de prendre uniquement la distance. Un graphique ROC [8] a été fait pour montrer la performance du modèle. Cela permet de comparer la sensibilité et la spécificité pour chaque seuil.

Un autre travail qui fournit également du code est celui de David Sumpter [10]. Le travail de ce dernier est de créer un modèle de xG en utilisant une régression logistique. Il utilise les données de Wyscout du championnat anglais de la saison 2017-18. Ce dernier explique étape par étape ce qui est effectué pour créer et améliorer son modèle. Il y a également des pistes pour convertir les positions X et Y en distance et en angle. Il commence par créer un modèle de xG en utilisant uniquement la distance. Par la suite, il le fait uniquement avec l'angle. Ensuite, il utilise de multiples facteurs, comme la distance au carré ou encore l'angle multiplié par la position X du tir, pour créer son modèle et il produit un résumé de la régression logistique.
Il n'y a pas moyen de connaître les coefficients uniquement pour la distance et l'angle car aucun résumé n'est fait pour ces deux facteurs uniquement.

8. Receiver Operating Characteristic

Predictor	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.5103	0.887	-0.576	0.565	-2.248	1.228
Angle	-0.6338	0.319	-1.989	0.047	-1.258	-0.009
Distance	0.2798	0.118	2.381	0.017	0.049	0.510
X	-0.1243	0.124	-1.001	0.317	-0.368	0.119
C	0.0300	0.040	0.750	0.454	-0.048	0.109
X2	-0.0014	0.001	-1.422	0.155	-0.003	0.001
C2	-0.0041	0.003	-1.398	0.162	-0.010	0.002
AX	0.1251	0.118	1.063	0.288	-0.105	0.356

TABLE 2 – Résumé de la régression logistique du modèle de xG de David Sumpter

Le dernier travail est fait par H.P.H Eggels. Le but est d'expliquer le résultat d'un match en utilisant les xG. Dans ce travail, il utilise un modèle de xG pour prédire le résultat d'un match. Les données du travail proviennent de ORTEC, de Immotio et FIFA. En effet, son travail utilise trois datasets pour créer son modèle de xG. [3]. Il y a eu un travail de "merging" des données pour avoir un dataset qui contient toutes les informations nécessaires pour créer le modèle. Cependant, les datasets peuvent avoir des problèmes entre eux lors du "merging". Par exemple, les noms des joueurs qui sont différents (majuscule, accent, encodage, surnom, etc.) a été un problème mais une solution a été trouvée.
Le dataset utilisé contient donc trois sources de données différentes.

ORTEC	**FIFA**	**Immotio**
Context	Player quality	Number of attackers in line
Part of body	Goal keeper quality	Number of defenders in line
Dist to goal		Distance nearest defender in line
Angle to goal		Distance goal keeper
Originates from		
Current score		
High		

TABLE 3 – Sources de données utilisées par H.P.H Eggels

Parmi les modèles testés, ce travail utilise :
— Un modèle de régression logistique
— Random Forest
— Un arbre de décision
— Ada-boost

Pour chacun de ces modèles, il y a une liste des différents paramètres qui vont être utilisés pour trouver le meilleur modèle. Chacun des modèles a suivi une procédure avec un set d'entraînement, un set de validation et un set de test. Le set de validation permet de trouver les meilleurs paramètres pour le modèle. Le set de test permet de tester le modèle avec les paramètres trouvés et le comparer avec les autres modèles Le modèle le plus performant parmi les quatres cités est le Random Forest avec une précision de 0.771. Cependant, il n'est pas possible de connaître les meilleurs hyper paramètres pour chacun des modèles. Ce travail donne tout de même de bonnes pistes pour l'ajout de nouvelle données pour améliorer le modèle de xG.

3.1 Récapitulatif des travaux existants

Auteur(s)	**Source de données**	**Conclusion**
Richard Pollard Jake Ensum Samuel Taylor	Coupe du monde 1986 et 2002	La distance, l'angle et l'espace avec le joueur le plus proche
Izzatul Umami Deden Hardan Gautama Heliza Rahmania Hatta	Wyscout, 5 championnats majeuren Europe. Saison 2019-20	La distance et l'angle apporte plus d'informations qu'uniquement la distance
David Sumpter	Wyscout, Premier League. Saison 2017-18	La distance et l'angle sont les facteurs avec le plus d'influence
H. P. H Eggels	ORTEC, Immotio, FIFA	Le but du travail est de prédire les résultats des matchs. Rien n'indique les facteurs les plus influents

Table 4 – Récapitulatif des travaux existants

4 Dataset

4.1 Présentation du dataset

Comme indiqué dans le section 3, le travail de David Sumpter [10] 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. [6]
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.

4.2 Events

4.3 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] Fred Garratt-Stanley. What is Expected Goals (xG) ?

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

[6] 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.

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

[8] 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.

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

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

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

[12] Izzatul Umami, Deden Hardan Gautama, and Heliza Rahmania Hatta. Implementing the Expected Goal (xG) model to predict scores in soccer matches. 4(1) :38–54.