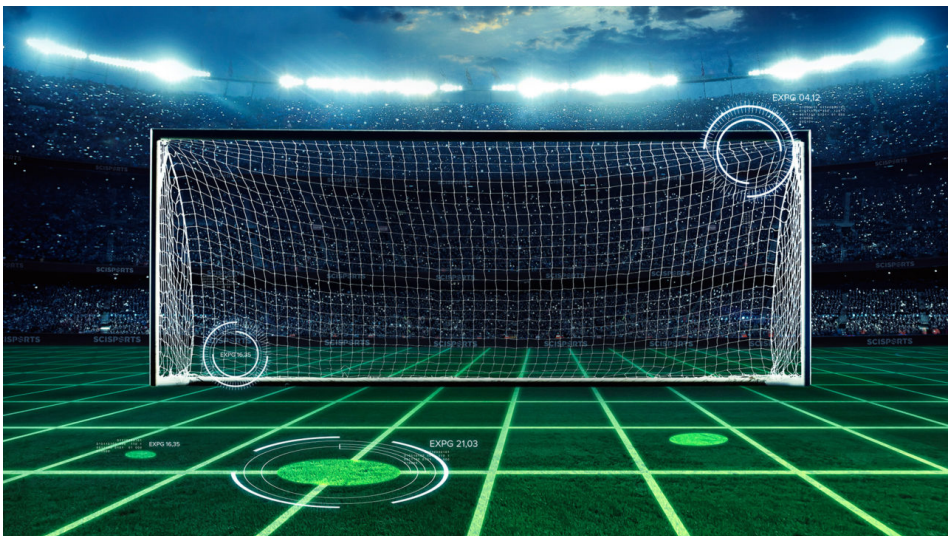
Université Côte d'Azur
M1 Ingénierie Mathématiques

L'analyse de données dans le Football : la notion d'*Expected Goals*



TERRIER Jules

Rapport de projet tutoré
encadré par Damien GARREAU

SOMMAIRE :

I. INTRODUCTION

L'utilisation des statistiques dans le sport remonte aux années 1960, quand Billy Beane directeur général de l'équipe de baseball des Oakland Athletics, franchise du Sud-Ouest des États-Unis, ne disposant pas d'un budget comparable à ceux d'autres franchises, a décidé d'utiliser l'analyse de données statistiques pour recruter des joueurs avec une faible valeur marchande mais ayant un fort potentiel, des joueurs sous-cotés. En 2002, la franchise se retrouve avec le meilleur ratio masse salariale/résultat, grâce à son recrutement, et enchaine une série de 20 victoires d'affilées.

Ces faits sont retranscrits dans le film *Le Stratège* ou *Moneyball* en anglais, et explique les débuts de l'utilisation des statistiques dans le sport aux USA. Il en a découlé, une généralisation de ce genre d'utilisation des statistiques plus précises, et au service des franchises, dans la plupart des sports américains, tel que le basketball ou encore le hockey sur glace.

Une question naturelle est la suivante : Pourquoi ne pas appliquer cela au football ?

Récemment, l'utilisation de statistiques de plus en plus précises prolifère dans le milieu du football [5]. On ne se contente plus des simples pourcentages de possession de balles par équipes, du nombre de tirs cadrés, ou encore du nombre de corners.

L'essor du Big Data et la popularité du football font que ces dernières années, les entreprises qui fournissent les données ont développé des algorithmes bien plus poussés, qui donnent une idée de la chance qu'une action a de se terminer par un but, cette valeur est communément appelée *expected goal*, et c'est le thème central de ce projet.

Cependant, le football, pourtant un des sports les plus populaires au monde, a mis du temps à se rendre compte de l'utilité des statistiques. Selon certains puristes, anciens joueurs professionnels, et autres journalistes, cela dénaturerait l'esprit du football, car les joueurs se focaliseraient plus sur les statistiques que sur l'aspect collectif.

D’un autre point de vu, le football est un sport très imprévisible et très complexe, avec très peu de buts (moyenne de 2,74 buts par match sur la saison 2018/2019 [1]), donc il est très difficile de se baser uniquement sur ceux-ci pour faire l’analyse d’un joueur, d’un match ou d’une saison.

Aujourd’hui, les équipes, les ligues, ou encore les médias sont toujours à la recherche de statistiques plus parlantes pour étayer leurs analyses. Ce qui a poussé des entreprises comme Opta Sports, StatsPerform, ProZone, et bien d’autres fournisseurs de données à se pencher encore plus sur le football, et développer leurs propres algorithmes, pour vendre leurs services.

Dans ce mémoire, nous nous penchons sur l’algorithme des *expected goals* ou xG, dont la traduction littérale en français serait *nombre de buts attendus*, qui sont le nombre de buts qu’un joueur ou une équipe aurait dû marquer — ou encaisser — sur une période donnée, qu’il s’agisse d’un match ou d’une saison.

Nous essayerons de comprendre comment en fonction des milliers de tirs des saisons précédentes, nous pouvons attribuer une certaine probabilité de marquer un but à chaque nouveau tir tenté ou concédé par une équipe. S'ils n'ont pas valeur de vérité absolue, les *expected goals* viennent s'ajouter à l'arsenal de statistiques dont disposent aujourd'hui les joueurs, entraineurs, journalistes et, par ricochet, les supporters.
S'ils affichent encore des limites, car tous les modèles ne prennent pas en compte les mêmes données, n’ont pas la même architecture pour leur algorithme, ces *expected goals* sont amenés à se développer car ils retranscrivent le match ou la saison de manière plus précise.

II. L'ESSOR DE LA DATA DANS LE FOOTBALL

1) Pourquoi un essor si tardif ?

Dans le monde du football professionnel, les décisions sont encore très souvent basées sur les émotions et les résultats récents. Nous savons aussi que le résultat d'un match de football dépend de beaucoup de choses, et plus particulièrement du facteur chance. Il est très difficile de modéliser la physique exacte des rebonds, par exemple, le fait que sur une frappe anodine, le ballon soit contré par plusieurs défenseurs et qu'il retombe dans les pieds de l'attaquant de l'équipe adverse, seul face au but, à la 89ème minute, et que cela permet à son équipe de gagner la finale d'une coupe d'Europe. C'est ce qu'on peut appeler le facteur chance.

C'est pourquoi, il est difficile d'expliquer comment une équipe ayant eu un plus grand pourcentage de possession de balle, un plus grand nombre de tir cadré, a perdu le match.

De plus, le football implique qu'il y ait peu de but dans un match, de ce fait le facteur chance à un réel impact sur le résultat. Ainsi, un seul but peut changer le résultat du match, et celui-ci les résultats de toute une saison.

On comprend donc que le résultat d'un match dépend d'énormément de facteurs, et il est donc d'autant plus difficile de l'expliquer clairement. Il est aussi bon de savoir que dans le football, les performances d'une équipe peuvent très fortement varier, d'une semaine à l'autre et en raison de pleins d'aspects, comme le temps de récupération des joueurs, le nombre de blessé, la progression des joueurs, ou encore la tactique mise en place par le coach. Si on rejoue deux fois la même action ou le même match, avec les mêmes équipes, dispositifs tactiques à la même date, nous n'aurons pas le même résultat.

Toutes ces choses font que le foot est un sport rempli d'incertitude. Il est donc difficile de pouvoir se baser uniquement sur des méthodes prédictives qui tiennent compte du passé. Il faut constamment améliorer, mettre à jour, notre approche, nos données, nos algorithmes pour qu'ils collent le plus possible au niveau actuel du jeu.

Pour comprendre le fait que l'utilisation du Big Data dans le football soit si tardif, je pense qu'il faut aussi comprendre que le football est un sport populaire.
Très souvent, les plus fervents supporters, sont ceux pour qui, la sortie du week-end, c'est l'occasion d'aller au stade encourager les joueurs du club de la ville de toutes leurs forces, après une dure semaine de travail.

Ils viennent rechercher la communion avec leur équipe favorite, les émotions que celle-ci leur procure, qu'elles soient bonnes ou mauvaises… C'est la fierté d'appartenir à un groupe, à une entité qui tire dans le même sens et pour un même but, c'est ce qui anime les spectateurs du football. Les gens recherchent des émotions, une équipe qui les faits vibrer, qui ne calcul pas justement… C'est un peu contradictoire avec l'utilisation des statistiques.

Il y a un autre aspect qui nourrit la réticence de certaines personnes à l'utilisation des datas. C'est Éric Di Meco, ancien joueur de l'Olympique de Marseille, aujourd'hui consultant, qui le dit : *« Le football n'est pas un sport de statistiques. Les statistiques tuent le jeu parce que les joueurs pensent plus aux statistiques qu'à la performance collective. C'est ce que l'on met sur le terrain qui est plus important que le nombre de passes ou la possession* » [2].
Il remet ici en cause l'aspect trop individuel et trop terre à terre des statistiques, qui occulte le côté psychologique du football, l'envie de gagner, de se surpasser, l'intelligence de jeu d'un joueur, ou encore le sens du collectif.

En somme, ce sont ces réticences des joueurs et des supporters qui ont conduit à un essor tardif de l'utilisation du Big Data dans le football. Il est à souligner que malgré cela, les aspects bénéfiques des statistiques sont conséquents et que celles-ci ont fait leurs preuves sur plusieurs points que nous évoquerons dans la sous-partie 4.

2) Qui produit et utilise les données ?

L'émergence d'entreprises spécialisées dans la statistique sportive a poussé toute la sphère footballistique à s'intéresser de plus près aux données. Ainsi, les entreprises telles que ProZone(1998), Opta Sports(1996), Statsperform(1981), Wyscout(2004), et bien d'autres, fournissent les médias et les clubs de football en toute sorte de contenus. Aussi, de nombreux clubs ont commencé à créer des cellules d'analyses statistiques.

C'est surtout le cas en Angleterre et en Allemagne, où les outils statistiques sont bien intégrés à l'optimisation de la performance dans les clubs de l'élite, mais aussi dans les divisions inférieures. On peut remarquer néanmoins, que dans d'autres pays comme la France, on commence à intégrer les données, notamment au Paris Saint-Germain, ou à l'Olympique de Marseille où des cellules spécialisées ont vu le jour ces deux dernières années [3].

On peut aussi souligner les cas de deux clubs, le Brentford FC (D2 anglaise) et le FC Midtjylland (D1 danoise). Ils ont pour particularité d'avoir le même propriétaire, Matthew Benham, un ex-trader londonien, qui a fini par s'intéresser de près au domaine des paris sportifs, et a développé son entreprise de statistiques *SmartOdds*, grâce à laquelle il a fait fortune. En tant que propriétaire des deux clubs, il s'est fixé pour objectif une gestion du club et des performances de l'équipe basé uniquement sur la Statistique.

D'ailleurs, ce type de fonctionnement porte ses fruits, puisque le FC Midtjylland est 1er du championnat Danois, et que le Brentford FC est aux portes de la Premier League anglaise, avec une 4ème place en Championship (D2 anglaise) [4] [5].

3) Comment sont-elles utilisées ?

C'est le point principal du projet. Nous allons voir que les données collectées sur un match, sur une saison entière, ou sur un joueur en particulier, vont être utilisées à travers différents types d'algorithmes, qui donneront une lecture de la performance plus précises et plus objective.

Il est bon de rappeler que les statistiques que l'on peut voir à la fin des matchs, telles que, la possession de balle ou le nombre de tirs cadrés, peuvent nous donner une idée complètement erronée du déroulement du match.

Nous pouvons très bien avoir une équipe avec un nombre de tirs cadrés et une possession de balles plus élevés que son adversaire, alors que celle-ci a perdu le match 3-0, car sa possession a été stérile, sa finition devant le but également, et l'équipe adverse a procédé en contre en mettant au fond des filets chacune de ses occasions.

C'est pourquoi, les statisticiens ont voulu développer des modèles qui retranscrivent plus précisément ce qui s'est réellement passé dans le match. Ces modèles nous donnent une valeur qui transmet une idée sur le jeu.

Nous nous intéressons ici aux *expected goals*, développés par Opta en 2012, qui traduit « *la probabilité qu'un tir se transforme en but, et qui est calculée en fonction des milliers de tirs précédents pris au même endroit »,* comme nous l'explique Kevin Jeffries, Data Editor chez Opta [5].

Dans la même lignée, on retrouve les *expected assists* qui sont les places qui auraient dû être décisives selon l'endroit où elles ont été délivrées, ou encore la *possession value* qui détermine la contribution d'un joueur pour chaque action de son équipe. Nous ne traiterons pas les cas de ces deux dernières mesures, car le sujet des *expected goals* est déjà assez vaste, et contient plusieurs pistes de questionnement que nous évoquerons tout au long de ce mémoire.

En somme, les données sont utilisées à travers ces modèles plus ou moins complexes. Ces modèles font le pont entre *le machine learning* et la performance sportive. Ils sont créés en grande partie par les entreprises de statistiques citées ci-dessus, mais il n'est pas rare aujourd'hui, de trouver des jeunes passionnés de Machine Learning et de football créer leurs propres modèles *d'expected goals.*

C'est d'ailleurs en partant du modèle d'un Data Scientist nommé Gabriel Manfredi, sur la plateforme de Machine Learning Kaggle, que nous avons commencé ce projet [7].

Il faut souligner que les premiers modèles de ce type, n'ont pas été implémenté que par les grandes entreprises de statistiques sportives tel qu'Opta ou Prozone. Mais aussi par des passionnés de football, comme Colin Trainor*,* un anglais, auditeur dans le domaine bancaire qui a trouvé l'utilisation poussées des Statistiques plus qu'intéressantes dans ses articles d'analyses pour le site internet de Statsbomb.

Néanmoins, les *expected goals* ont été inventés par Sam Green, qui décrit le concept pour la première fois en 2012. Green travaillait pour la société d'analyses de données OptaPro au sein du département qui fournit aux clubs des informations allant au-delà des statistiques régulièrement publiées dans les médias.

OptaPro avait pour mission de trouver de nouvelles mesures de performance. C'était chose faite avec les *expected goals*, qui ont rapidement été reconnus comme un outil très utile pour les clubs et par les médias, ces dernières années. Canal+ a notamment lancé une émission durant la saison 2019/2020, appelé *3-5-2*, où les experts utilisent uniquement les *statistiques avancées* comme les *expected goals* pour décrypter les performances des équipes, des joueurs, et avoir un meilleur aperçu des matchs par les chiffres. On y retrouve d'anciens footballeurs professionnels, des consultants, et des Data Scientists de l'entreprise Opta Sport.

4) Quels sont les objectifs de l'utilisation des données ?

Le but principal du développement des Statistiques avancées dans le football est de donner une lecture plus juste, plus objective du jeu, et d'aider à améliorer la performance sportive de l'équipe.
Comme l'explique Christoph Biermann dans son livre *Big Data Foot*, en tant qu'être humain, nos analyses, nos perceptions et donc notre sens de l'évaluation sont biaisés [6].

On peut remarquer cela dans la presse spécialisée. Dans un certain journal un joueur va recevoir la note de 7/10 pour son match, dans un autre ce sera un 5/10, et un supporter qui l'a vu jouer au stade lui donnera un 3/10. Les différences de perception sont monnaie courante.

C'est pourquoi il est facile de se faire une opinion erronée au sujet de la qualité d'une équipe ou d'un joueur en particulier. Il suffit de regarder, dans un premier temps le score du match, deuxièmement le nombre de buts ou de passes décisives délivrées par tel joueur, et enfin le nombre de ballons joués, de duels gagnés, ou de tirs tentés. Les évaluations à l'œil nu sont d'autant plus subjectives.

Justement, il est bon de souligner que les évaluations personnelles dépendent souvent de plusieurs facteurs extérieurs (exemple : humeur du moment, préjugés, etc.) mais aussi des *biais psychologiques* ou *biais cognitifs*, qui distordent notre analyse.

- ## Un point sur les biais cognitifs

Les biais cognitifs sont présents dans tous les domaines : au travail, dans les relations, ou encore dans l'interprétation d'évènements politiques, comme nous l'explique Christoph Biermann dans son livre cité précédemment [6].

Parmi ces biais, il existe le biais de confirmation ; par exemple : je ne retiens que les éléments qui confirment ce que je pense donc si un joueur rate une occasion, c'est que ce n'est pas un bon joueur comme je le pensais, et je ne prends pas en compte toutes les bonnes actions qu'il a effectués dans le match et qui ne se voit pas également dans les statistiques basiques.

Le biais de réciprocité, par exemple : je suis journaliste ou consultant, ce joueur a été sympa avec moi, il m'a donné une info, a accordé une interview à mon média, je vais le noter généreusement sur ce match.

Le biais du résultat, par exemple : en tant qu'être humain, on prend le résultat comme point de départ et non la prestation générale, c'est-à-dire que si notre équipe favorite a gagné son match 1-0, avec 2 tirs cadrés et 1 but, et avec 30% de possession de balle, alors que l'adversaire a eu bien plus d'occasions franches et méritait de gagner, on ne retiendra que la victoire de notre équipe, sans prendre en compte la réalité de la prestation.

Ou encore, le biais de conformité, par exemple : tout le monde autour de moi a trouvé bon tel joueur, alors il a forcément été bon.

Ainsi, c'est pour parer à cela, que les mesures de performances types *expected goals* ont été inventées, cela permet d'avoir une lecture plus précise que les statistiques *simples* de possession de balle ou de tirs cadrés, et cela permet de prendre la performance dans son ensemble de manière objective.

Ces outils permettent de prendre de meilleures décisions sur le plan sportif, comme l'évoque Rasmus Ankersen, président du FC Midtjylland. Convaincu par la méthode de gestion sportive du propriétaire du club Matthew Benham évoqué plus haut : « *cela permet de suivre le plan malgré le classement quand ça ne va pas, ou de tout changer même si en apparence tout roule ».*

En d'autres termes, si les résultats ne sont pas au rendez-vous sur une période mais que statistiquement les chiffres sont bons, c'est que c'est une période de malchance, que la chance va tourner, et qu'il ne faut rien changer. De ce fait, cela permet certainement de perdre moins d'argent en licenciant trois entraineurs sur la même saison, avec toutes les indemnités que cela représente, et ainsi de maintenir un projet sur le long terme…

- L'exemple de Smartodds

C'est Benham, grâce à sa société Smartodds (spécialisée dans l'analyse statistique), qui a établi ce type de gestion pour ces deux clubs (Brentford FC et FC Midtjylland), véritable laboratoire de la Statistique sportive dans le monde du football, qui se base sur la production et la performance d'une équipe et non pas sur les résultats purs et durs. Ainsi, dans ces deux clubs, les joueurs sont majoritairement recrutés grâce aux données statistiques.

Les données sont devenues un outil incontournable pour les recruteurs. D'ailleurs le Brentford Community Stadium, nouveau stade du club, devait être livré au printemps dernier, et fut financé par les plus-values des transferts de joueurs recrutés grâce à l'analyse de données tel que l'ancien joueur de l'OGC Nice, formé au club, Neal Maupay partit (pour 22 millions d'euros) au Brighton FC promu en Premier League. L'exemple de Saïd Benrahma (lui aussi ancien niçois), Ollie Watkins, et Bryan Mbeumo, fers de lance de la deuxième meilleure attaque de Championship (D2 Anglaise), tous recrutés dans des divisions inférieures (Ligue 2 Française pour Mbeumo et Benrahma, Quatrième divison anglaise pour Watkins) est frappant. À eux trois réunis, la valeur totale de leurs transferts dans le club londonien s'élève à tout juste 10 millions d'euros.

C’est la technique de l’entonnoir, résume un recruteur ayant travaillé pour le club. Le *scout* remet des rapports sur les joueurs supervisés, et ensuite, une équipe dédiée qui connait tout de la potentielle recrue est capable de déterminer, grâce aux outils statistiques tels que les xG et autres, si cela vaut le coup d’investir ou non sur le joueur.

Grâce à un modèle statistique développé par les équipes de SmartOdds, qui permet de comparer le niveau de toutes les équipes d’Europe par rapport à celui du Brentford FC, et ainsi d’identifier quelles équipes surperforment par rapport à leurs moyens, ils sont susceptibles de détecter des joueurs sous-cotés qui évoluent dans leur rang, puis de les recruter. Il y a donc clairement un intérêt financier pour les clubs.

On retrouve beaucoup de clubs qui cherchent ainsi à bénéficier d’un avantage concurrentiel grâce aux Statistiques avancées, comme Arsenal FC, où le manager/entraineur français Arsène Wenger, avait initié en 2012 l’achat de l’entreprise américaine StatDNA.
Il avait déclaré en 2019 lors d’une conférence : « *Les statistiques avancées nous permettent d’être en avance sur la connaissance d’un joueur inconnu. Avant d’être une star, le joueur existe déjà. Il n’est juste pas connu donc il ne coûte pas cher »* [5].

On a aussi le cas de Leicester FC et de son recruteur Steve Walsh, qui avait recruté Jamie Vardy (troisième division anglaise), Riyad Mahrez (Ligue 2) ou N’Golo Kanté pour des modiques sommes (500 000 euros pour Mahrez, 9 millions d’euros pour Kanté) et en partie grâce aux stats. Quand on connaît la plus-value qu’a réalisée Leicester City sur les ventes de Mahrez (68 millions d’euros) et Kanté (36 millions d’euros), on peut constater un certain intérêt à utiliser ce type de données de plus en plus avancées.

Enfin, dans une moindre mesure, l’utilisation des données sert à améliorer les performances de l’équipe à l’aide des outils de statistiques avancées, pour faire une analyse plus précise et donc permettre à l’entraineur de travailler en conséquence.

Il faut tout de même relativiser le fait que les statistiques ne font pas tout, et qu'elles sont là pour confirmer ou infirmer une intuition. Elles ne sont pas faites pour remplacer totalement le travail des hommes de terrain, qui conservent leur expertise.

Kevin Jeffries, évoqué précédemment, explique : « *Le football, contrairement au basketball ou au baseball, n'est pas un sport de séquences. Il y a peu de temps mort donc il y a une multitude d'événements qui se produit en continu et qui reste difficile à évaluer. Il y a aussi plein de choses qu'on ne peut pas analyser avec les stats comme le mental, les problèmes relationnels, la gestion des émotions, etc.* ».
En réalité, les chiffres ne sont pas pertinents sans une bonne interprétation, surtout que les données, les algorithmes, les modèles sont implémentés par des êtres humains qui font donc naturellement des erreurs.

III. NOS DONNÉES

Les données les plus facilement disponibles au public sont le nombre de tirs, le nombre de buts, le nombre de fautes commises, par telle ou telle équipe. Cependant, dans un modèle prédictif, ce type de données, implémentées sans contexte ne nous permet pas de tirer de conclusion précise sur un match, une équipe ou un joueur.
Par exemple, une équipe qui effectue 10 tirs de loin (c'est-à-dire à plus de 40 mètres des buts) n'a pas les mêmes chances de marquer qu'une équipe qui a tiré 10 fois à l'intérieur de la surface de réparation, soit dans les 16,5 premiers mètres devant le but. Il est plus facile de marquer lorsqu'on est proche de la cage. Ici, le contexte est donc *où a eu lieu le tir*, ce qui nous donne bien plus d'informations, et donc plus de précision pour prédire si un tir va finir au fond des filets ou non.
Un match de football donne lieu à de nombreux évènements, et il est donc important de pouvoir prendre en compte le contexte dans lequel les évènements arrivent.

1) Quels types de données avons-nous utilisées ?

Pour notre modèle, nous avons utilisé les mêmes données que Gabriel Manfredi sur la plateforme Kaggle. Il a été aidé de Alin Secareanu, un Data Scientist travaillant pour l'entreprise Avira.

Le jeu de données créer est le résultat d'un long travail de recherche internet, de l'intégration de plusieurs jeux de données différents, et de l'utilisation de toutes les informations importantes et commentaires disponibles sur les matchs, afin de remettre chaque évènement dans son contexte. Les commentaires sur les matchs ont été tirés de bbc.com, espn.com, et onefootball.com.

Ainsi la base de données nous donne une vue précise de 9074 matchs, totalisant 941 009 évènements (tirs, passes, cartons, etc.), des cinq plus grands Championnats Européens, c'est-à-dire, l'anglais, l'espagnol, l'italien, l'allemand et le français, de la saison 2011/2012 jusqu'au 25 janvier 2017 soit durant la saison 2016/2017. Ces données nous donnent le contexte de chaque tir, allant de la passe qui le précède, à la situation d'attaque, ou encore la position du tireur.

Tout ce contenu se trouve sur la Kaggle dans un fichier .csv disponible gratuitement.
https://www.kaggle.com/gabrielmanfredi/expected-goals-player-analysis/data

2) Comment avons-nous traité nos données ?

Le but de notre modèle est de pouvoir prédire, en fonction de plusieurs caractéristiques, si un tir va finir en but ou non. Ainsi, l'évènement qui nous intéresse est le *tir*, car un but provient forcément d'un tir, que cela soit du pieds, de la tête, et qu'il soit contré ou non par un adversaire. On mettra de côté le cas des buts contre son camp, qui restent très rares dans le football. Sur la saison 2019/2020, où il restait 10 journées de championnats, soit 100 matchs a joué, il y a eu 23 buts contre son camp sur 704 buts marqués en tout, soit environ 3%, ce qui est très faible [18].

Il faut souligner que nous ne prenons pas en compte l'identité du tireur, et ses qualités intrinsèques. Notre but est d'avoir un modèle qui donne une mesure pour n'importe quel joueur. Nous voulons une méthode qui standardise pour savoir si un tir va finir au fond des filets ou non, en fonction de plusieurs caractéristiques du tir, et de la situation dans laquelle il a été pris.

C'est donc sur l'évènement *tir* que nous nous focaliserons, ainsi nous aurons besoin de tous les éléments relatifs aux tirs. Toutes les caractéristiques qui nous intéressent sont disponibles dans notre jeu de données.

En quelques lignes de code Python, nous tirerons toutes les informations importantes de chaque tir, à savoir :

Le lieu du tir : moitié de terrain offensive, moitié de terrain défensive/ au milieu du terrain, aile droite, aile gauche/ angle difficile (excentré au niveau du poteau de corner), tir de loin (plus de 50 mètres), angle difficile côté gauche (proche de la cage), angle difficile côté droit/ côté gauche de la surface de réparation (à 16,5 mètres de la cage), côté droit de la surface de réparation / côté gauche de la surface des 6 mètres, côté droit de la surface des 6 mètres/ très proche de la cage (moins de 2 mètres), au niveau du point de pénalty/ hors de la surface de réparation (à 16,5 mètres de la cage), à plus de 35 mètres, à plus de 40 mètres / non enregistrés (il y a quelques tirs dans notre jeu de données où nous n'avons pu avoir de localisation par manque d'informations). Nous n'avons pas les coordonnées exactes du tireur, car les données proviennent de commentaires faits sur les matchs, ainsi nous avons seulement des positions approximatives données par nos variables.

La partie du corps, avec laquelle le tir a été effectué : pied droit/ pied gauche/ tête.

La passe qui a amené le but, le type de passe : aucune (le joueur a pris le ballon tout seul, a dribblé plusieurs joueurs adverses et a tiré)/ passe simple (dans les pieds du tireur), centre (longue passe excentré sur un côté/ au niveau du poteau de corner adverse)/ passe de la tête/ passe en profondeur (passe dans l'espace dans la course du tireur).

La situation, le contexte de jeu : jeu ouvert (le tir est intervenu après progression dans le camp adverse par un jeu de passes)/ jeu indirect après un arrêt de jeu (après une touche, un coup franc, ou un corner, joué en passes pour progresser vers le but)/ corner (centre aux poteaux de corner)/ coup franc (coup de pied arrêté après une faute adverse).

Si le tir est venu d'un contre ou non (récupération de balles et projection rapide vers l'avant).

Enfin, s'il y a eu but ou non suite au tir. C'est la variable réponse de notre modèle, celle qui nous intéresse.

Rappelons que nos données sont en anglais, c'est pourquoi dans notre code, toutes ces caractéristiques sont en anglais.

Nous avons d'abord transformé ces données catégorielles en données binaires, à l'exception des *contres* qui sont déjà sous forme binaire.
Ainsi j'ai obtenu le tableau suivant :

	is_goal	fast_break	loc_centre_box	loc_diff_angle_lr	diff_angle_left	diff_angle_right	left_side_box	left_side_6ybox	rig
0	0	0	0	0	0	0	1	0	
11	0	0	0	0	0	0	0	0	
13	1	0	0	0	0	0	1	0	
14	0	0	0	0	0	0	0	0	
17	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	
940983	0	0	0	0	0	0	0	0	
940991	0	0	0	0	0	0	0	0	
940992	0	0	0	0	0	0	1	0	
940993	0	0	0	0	0	0	0	0	
941006	0	0	0	0	0	0	1	0	

229135 rows × 29 columns

Figure 1 : Données traitées dans notre modèle.

Une ligne correspond à un évènement, dans notre cas, c'est un tir. Et chaque colonne correspond à une des caractéristiques de ce tir introduite précédemment.

Ensuite nous avons créé nos deux ensembles X et y, y contient tous les tirs soit 229 135 tirs, et la première caractéristique à savoir s'il y a but ou non, c'est ce que l'on veut prédire. X, en revanche, contient toutes les autres caractéristiques liées aux tirs, ce sont nos variables explicatives. Nous avons simplement séparé notre jeu de données de la Figure 1, en deux ensembles distincts.

On a un total de 229 135 tirs, avec 28 caractéristiques qui décrivent chaque tir. Ce sont des caractéristiques binaires, c'est-à-dire que si la ligne contient 0 cela veut dire que la caractéristique ne fait pas partie de celles qui composent le tir. Par exemple : si pour un tir j'ai 0 dans la colonne *fast_break* (contre), cela nous indique que ce tir ne provient pas d'un contre.

Ensuite, nous avons séparé X et y, en un ensemble d'entrainement et un ensemble de test. Nous avons appris dans notre cours de Data Mining avec M. Nicolas PASQUIER cette année, qu'il était bon de prendre entre 65% et 70% des données pour entrainer notre modèle, et entre 30% et 35% pour le tester [8].

Nous avons choisi la répartition 65%-35% car nous avons un jeu de données conséquent (229 135 tirs), et cela nous permet d'entrainer le modèle sur 148 937 tirs, et ainsi le tester sur 80 198 tirs.

IV. NOTRE MODÈLE

Au vu du type de données que nous avions et la manière dont nous les avons traitées pour avoir des matrices de données binaires, nous avons choisi d'utiliser la Régression Logistique comme classifieur en nous inspirant du modèle de Gabriel Manfredi sur Kaggle.

Le rôle d'un classifieur est de classer dans des groupes (des classes) les échantillons qui ont des propriétés similaires, mesurées sur des observations [9].

1) La régression logistique

La régression logistique est une approche statistique qui peut être employée pour évaluer et caractériser les relations entre une variable réponse de type binaire (dans notre cas : but=1 / pas but=0), et une, ou plusieurs, variables explicatives, qui peuvent être de type catégoriel (par exemple : contre, pénalty, tir de la tête), ou numérique continu (si nous avions les coordonnées exactes du tir par exemple). Dans notre cas, nous avons transformé nos données catégorielles en données binaires pour faciliter la compréhension.

La régression logistique nous permet d'avoir une probabilité de réalisation d'une des deux modalités cibles (par exemple, dans notre cas, probabilité que le tir finisse par un but). En effet ce n'est pas la réponse binaire qui est directement modélisée.

Cette probabilité de réalisation ne peut pas être modélisée par une droite car celle-ci conduirait à des valeurs, soit inférieure à 0, soit supérieure à 1 [10].

Ce qui est impossible, car une probabilité est forcément bornée par 0 et 1.

Cette probabilité est donc modélisée par une fonction sigmoïde, bornée par 0 et 1 :

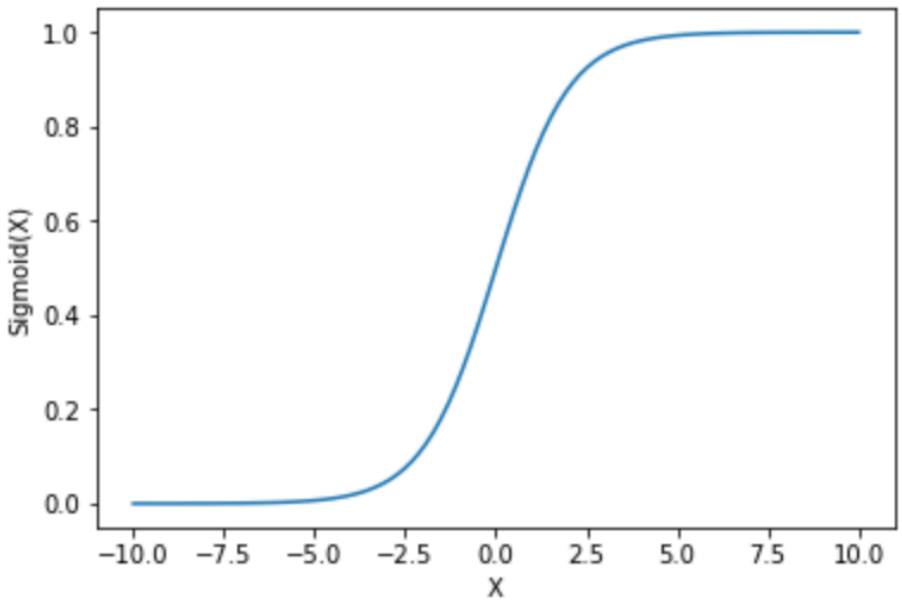


Figure 2 : Représentation de la fonction sigmoïde.

Cette courbe est définie par la fonction logistique, d'équation :

$$f(x) = \frac{\exp(x)}{1 + \exp(x)} = p.$$

Lorsque la fonction logistique est ajustée à des données observées, la forme de la courbe sigmoïde s'adapte à ces données, par l'estimation de paramètres.
Dans le cas d'une seule variable explicative X, l'équation de la courbe logistique est alors :

$$P(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

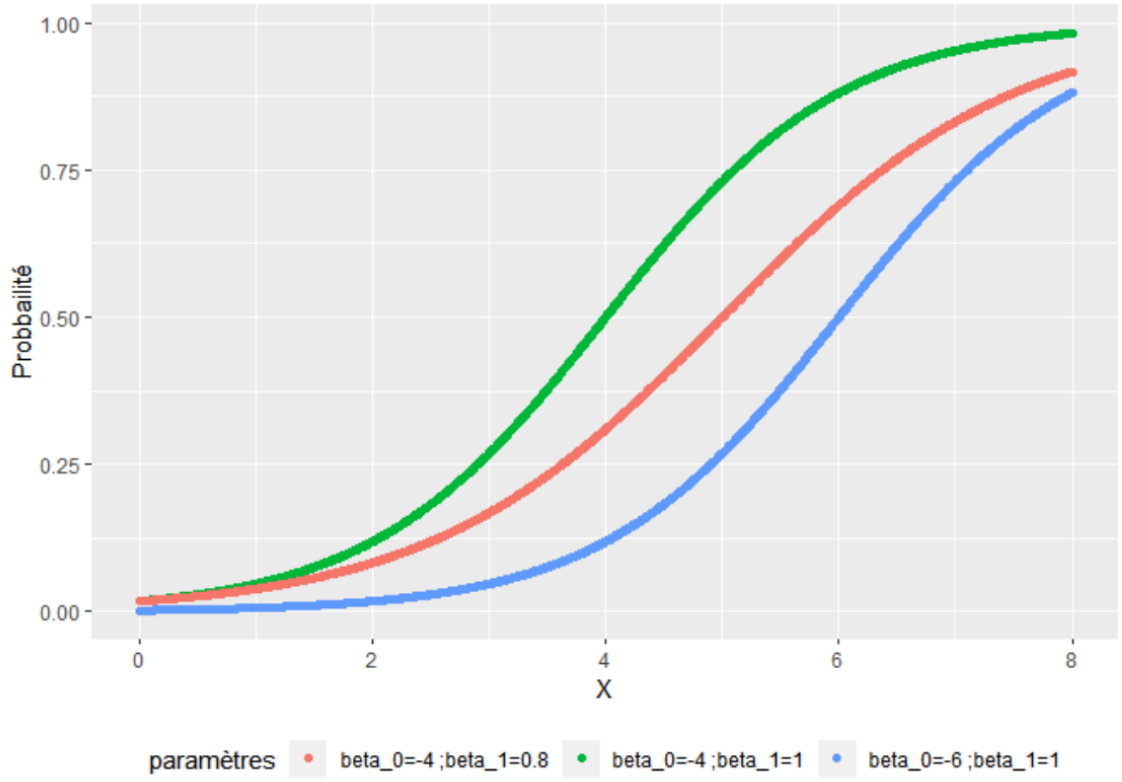


Figure 3 : Trois exemples de courbes logistiques obtenues avec des paramètres β_0 et β_1 différents.

Dans une situation de variables explicatives multiples l'équation se généralise en :

$$P(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)} = \frac{\exp(\sum \beta X)}{1 + \exp(\sum \beta X)}.$$

Un autre point positif de la régression logistique est qu'elle nous permet d'avoir des coefficients pour chacune de nos variables prédictives, et ainsi, elle permet de savoir quelles sont les variables importantes dans notre modèle pour prédire si un tir va finir au fond des filets ou non [11].

2) Exécution du modèle

Au préalable, comme nous l'avons dit précédemment, nous avons séparé nos données en un ensemble d'entrainement X_train, et un ensemble de test X_test.

Ensuite, nous avons entrainé une régression logistique sur notre ensemble d'entrainement.

Par la suite, nous avons appliqué notre classifieur à l'ensemble X_test, et obtenu des résultats assez intéressant que nous développerons dans la partie suivante.

Nous avons par la suite essayé d'améliorer ces résultats à travers différentes approches.

V. NOS RÉSULTATS

1) Les mesures de performance du modèle

Notre ensemble d'entrainement contient 148 937 tirs, dont 15 937 finissent par un but. C'est une information non négligeable que nous exploiterons par la suite pour expliquer les résultats de notre modèle.

Pour évaluer nos résultats nous avons utilisées différentes mesures utiles en Machine Learning, dont nous allons rappeler les définitions :

La matrice de confusion : c'est une matrice qui permet de mesurer rapidement la qualité de notre classifieur. Chaque ligne correspond à la classe réelle, et chaque colonne à la classe prédite. Par exemple : nous pouvons y lire, le nombre d'exemples de tests qui sont des buts et qui ont été prédit comme des buts. Cela nous donne une idée des succès et échecs de notre modèle.

La courbe Receiver Operating Characteristic (ROC) : c'est un graphe représentant les performances du classifieur pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs.

$$Taux\ de\ Vrais\ Positifs = \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Négatifs}.$$

Cela donne la proportion de résultats positifs, dans notre cas *but*, qui a été prédit correctement.

$$Taux\ de\ Vrais\ Négatifs = \frac{Faux\ Positifs}{Faux\ Positifs + Vrais\ Négatifs}.$$

Donne la proportion de résultats prédit positif donc *but*, alors qu'en réalité il n'y avait *pas but*, par rapport à tous les exemples où il n'y avait *pas but*.

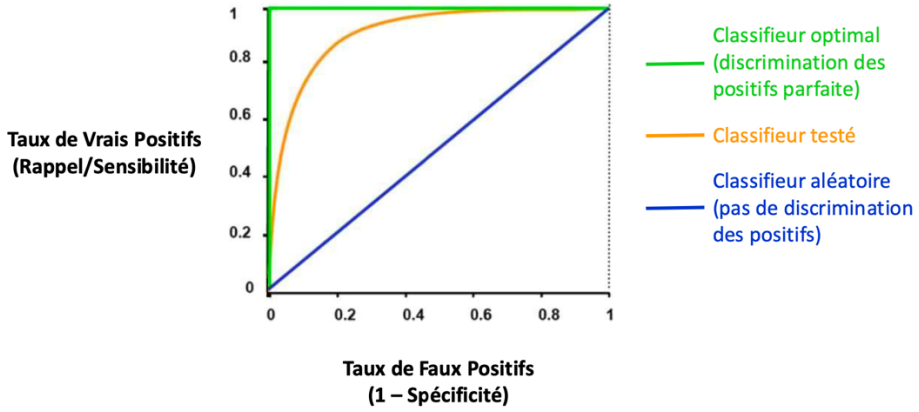


Figure 4 : Taux de vrais positifs et de faux positifs pour différents seuils de classification [8].

L'Area Under the Curve (AUC) : qui signifie *aire sous la courbe ROC*, cette valeur mesure de l'aire sous la courbe ROC (par calcul d'intégrale de (0,0) à (1,1)). C'est un indicateur numérique calculé à partir de la courbe qui permet d'en évaluer la qualité globale. Les valeurs d'AUC sont comprises entre 0 et 1. Un modèle dont 100% des prédictions sont erronées aura une AUC de 0. Si toutes ces prédictions sont correctes, son AUC sera de 1.

La Précision : elle répond à la question, *quelle proportion d'identifications positives étaient effectivement correcte ?*. Par exemple : si notre modèle a une précision de 0,7, cela veut dire que quand il prédit qu'un tir va se transformer en but, sa prédiction est juste dans 70% des cas.

2) Commentaires sur les coefficients de la régression logistique

Comme expliqué précédemment, nous obtenons les coefficients de la régression linéaire pour chaque variable explicative du modèle. Ainsi, nous pouvons identifier les variables importantes dans notre modèle. Si un coefficient est supérieur à 0, cela signifie que la variable explicative associée accroit la probabilité que le tir se finisse par un but. Si celui-ci est égale à 0, cela signifie qu'elle n'influe pas grandement sur la conversion du tir en but.

Enfin, si le coefficient est inférieur à 0, cela veut dire que la variable explicative associée va faire baisser la probabilité que le tir finisse par un but.

Nous remarquons que le coefficient de la variable *fast_break*, c'est-à-dire *contre-attaque*, est de 1,71, une valeur supérieure à 0, ce qui signifie que les tirs précédés par une contre-attaque ont une plus grande probabilité de finir en *but*. Tout comme la variable *penalty*, ou encore *close_range*, qui signifie *tir de près*, qui ont des coefficients environ égaux à 2. Ce qui parait logique, un tir proche de la cage, ou un penalty (à 11 mètres de la ligne de but), ont plus de chance de finir en but.

Aussi, la variable explicative qui concerne le type de passe avant un tir, à savoir *through_ball* (= passe en profondeur), a un coefficient positif (environ 0,66) tandis que tous les autres types de passes ont un coefficient négatif. Cela signifie que les tirs survenus après une passe en profondeur ont une plus grande probabilité de finir au fond des filets dans notre modèle.

3) Résultats

Nous avons obtenu une valeur d'AUC de 0,8177, ce qui est plutôt intéressant, en termes de qualité des prédictions du modèle.

Aussi, avec une précision de 0,91098 notre classifieur prédit correctement s'il y a but ou non, dans 91% des cas.

La courbe ROC obtenu est la suivante :

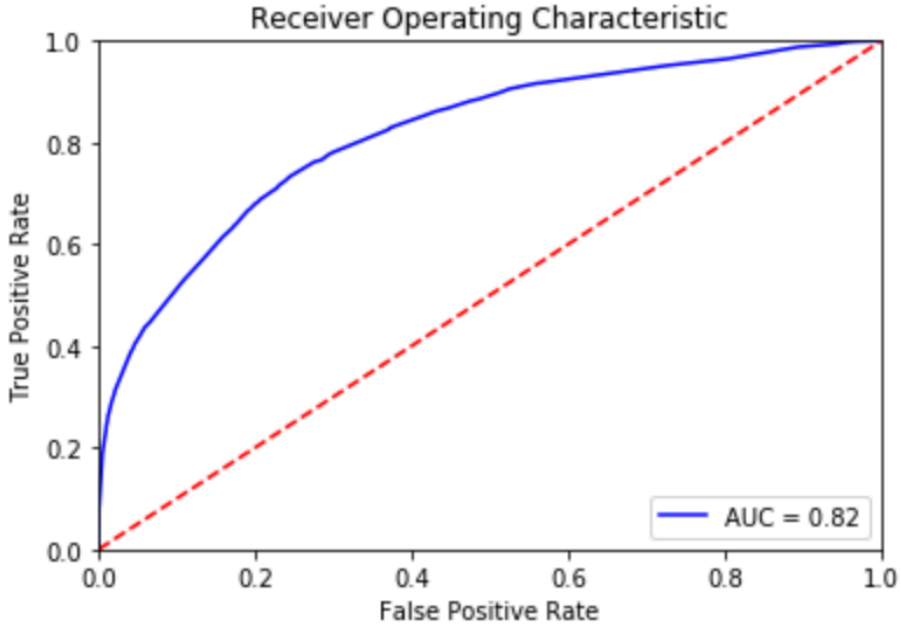


Figure 5 : Courbe ROC pour la régression logistique simple.

Ce sont de bons résultats, mais ces mesures ne tiennent pas compte d'une chose, c'est le fait que notre jeu de données est déséquilibré ; ce qui est normal, cela vient du fait que dans le football, même sur un grand nombre de tirs tentés, le taux de conversion de ceux-ci en *but* est très faible (10,7% de buts sur 229 135 tirs). C'est pourquoi notre dataset est fortement déséquilibré. On a 90% de notre ensemble d'apprentissage qui appartient à la classe 0 = *pas de but*, et 10% à la classe 1 = *but*.

Pour faire le parallèle, si nous avions un classifieur *bête* qui prédisait tout le temps *pas but*, alors dans 90% du temps il aura bien prédit.

Enfin la matrice de confusion, nous a permis de nous rendre compte de ce déséquilibre.

De tous les tirs où il n'y a *pas but*, notre modèle a correctement prédit *pas but* 70 820 fois, et s'est trompé 6265 fois en prédisant *pas but* alors qu'il y avait bien *but*.

En revanche, il a correctement prédit « but » 874 fois, et s'est trompé sur 2239 tirs, où il a prédit *pas but* alors qu'il y avait bien *but*.

On remarque donc que notre modèle sait mieux prédire la classe 0, et se trompe bien plus souvent pour prédire la classe 1, le *but*, car il a été entrainé majoritairement sur des exemples de la classe 0.

Pour remédier à ces problèmes nous avons essayé d'améliorer notre modèle en utilisant différentes techniques.

VI. AMÉLIORATION DU MODÈLE

1) Méthode d'« Oversampling »

Comme nous l'avons dit précédemment, notre jeu de données est très déséquilibré avec 90% de classe 0 (= *pas but*), et 10% de classe 1 (= *but*).

Une des méthodes pour parer ce problème est l'*oversampling*, cela peut être défini comme le fait de rajouter au jeu de données plusieurs copies d'exemple de classe minoritaire aléatoirement. Soit pour cela signifie d'ajouter plusieurs copies de la classe 1, pour arriver à un jeu de données équilibré [12].

Nous utiliserons le module de rééchantillonnage (*resample*) de Scikit-Learn pour reproduire au hasard des échantillons de la classe minoritaire.

Après le rééchantillonage, on a un ratio égal pour chaque classe de notre jeu de données, soit 50% de classe 0, et 50% de classe 1.

Nous nous retrouvons avec une valeur AUC similaire, soit 0,8177. Cependant nous avons perdu en précision, avec 0,72918. Ainsi, notre classifieur prédit correctement dans 72,918% des cas. C'est donc un gros point négatif, que nous allons essayer d'améliorer.

En revanche, en ce qui concerne la prédiction de la classe 1, le *but*, notre modèle s'est nettement amélioré. Cette fois-ci, nous avons correctement prédit le *but*, 19 687 fois, et nous nous sommes trompés 6472 fois. Enfin, le modèle a bien prédit la classe 0, *pas but*, 52 007 fois et s'est trompé 2032 fois.

On remarque tout de même encore une fois, que notre modèle se trompe moins en prédisant la classe 0.

La courbe ROC quant à elle, est assez similaire à la Figure 5 :

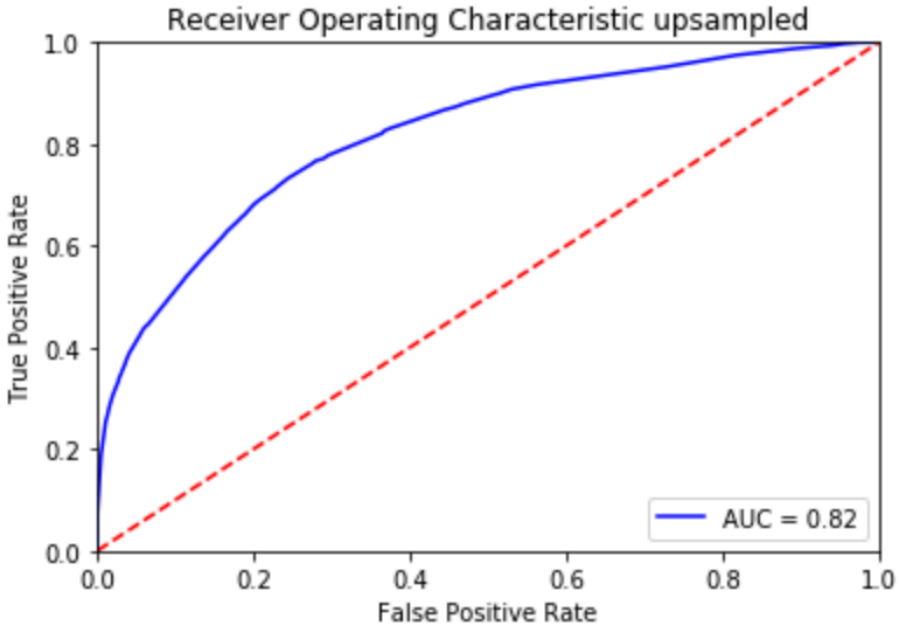


Figure 6 : Courbe ROC pour la régression logistique après *oversampling*.

Le même cheminement a été suivi en appliquant une méthode d'*undersampling*, où dans ce cas, de manière aléatoire, nous avons retiré des échantillons de la classe majoritaire de notre jeu de données, donc la classe 0, *pas but*. Dans ce cas, nous perdons des données pour entrainer notre modèle, ce qui est contraire à l'objectif de base du sujet, qui est d'avoir un maximum de données pour être le plus précis possible.

Les résultats ont été les mêmes qu'avec la méthode d'*oversampling*, et de plus celle-ci correspond mieux au projet, car plus on peut entrainer le modèle sur un jeu de données conséquents plus il aura de chance d'être précis.

Pour améliorer la précision du modèle, nous avons essayé de réitérer l'*oversampling* mais de manière moins brutale. Ainsi, au lieu de passer de 90% de classe 0, 10% de classe 1, à 50-50. Nous avons choisi, après plusieurs essais, de prendre pour répartition 70% de classe 0 et 30% de la classe 1 dans notre rééchantillonage. Grâce à cela nous avons obtenu des résultats plutôt intéressants, la valeur d'AUC reste la même environ 0,82, mais nous avons amélioré notre précision de pratiquement 20%, en passant de 0,729 à 0,899. Soit, dans pratiquement 90% des cas, notre modèle prédit correctement la classe 1 ou la classe 0.

C'est un résultat non négligeable, mais cela ne nous permet pas de battre la régression logistique appliquée simplement au début du projet. C'est pourquoi, nous avons voulu appliquer d'autres méthodes pour construire notre modèle.

2) Un autre type d'algorithme

Nous avons pensés qu'il serait bon de tester différents types d'algorithmes pour améliorer notre modèle d'*expected goals*, notamment améliorer sa précision [12].

Pour cela nous avons choisi l'algorithme de Random Forest, ou *forêts d'arbres décisionnels*. C'est un algorithme de classification qui réduit la variance des prévisions d'un arbre de décision seul. Pour cela, il effectue un apprentissage en parallèle sur de multiples arbres de décision de Breiman construits aléatoirement et entraînés sur des sous-ensembles de données différents.

L'algorithme des forêts aléatoires est connu pour être un des classifieurs les plus efficaces *out-of-the-box* (c'est-à-dire nécessitant peu de prétraitement des données). Ce qui convient très bien à notre cas de jeu de données déséquilibrés.

Nous avons appliqué ce classifieur, en modifiant les paramètres pour les ajuster un maximum au modèle, et tirer le meilleur parti de l'algorithme. C'est pourquoi, nous avons augmenté le nombre d'arbres, avec le paramètre *n_estimators*.

Ainsi les résultats obtenus sont assez proches de la régression logistique simple effectué au début du projet. Nous obtenons une valeur de l'AUC similaire 0,82, ainsi qu'une précision de 0,91. Nous avons une différence de pratiquement deux millièmes de la précision entre notre régression logistique et le random forest, en faveur de la régression logistique.

En revanche, la matrice de confusion, nous indique une légère amélioration dans le nombre de vrais positifs et de vrais négatifs, car nous obtenons 70 760 cas de classe 0, soit *pas but* bien prédits, contre 6220 cas de la classe 0 mal prédit. Enfin le modèle a bien prédit 934 fois la classe 1, le *but*, et s'est trompé 2284 fois.

On remarque que c'est une meilleure répartition que dans le cas de la régression logistique sans *Oversampling*. Cependant, si on regarde en termes de précision du modèle, la régression logistique l'emporte, même si cela se joue à deux millièmes.

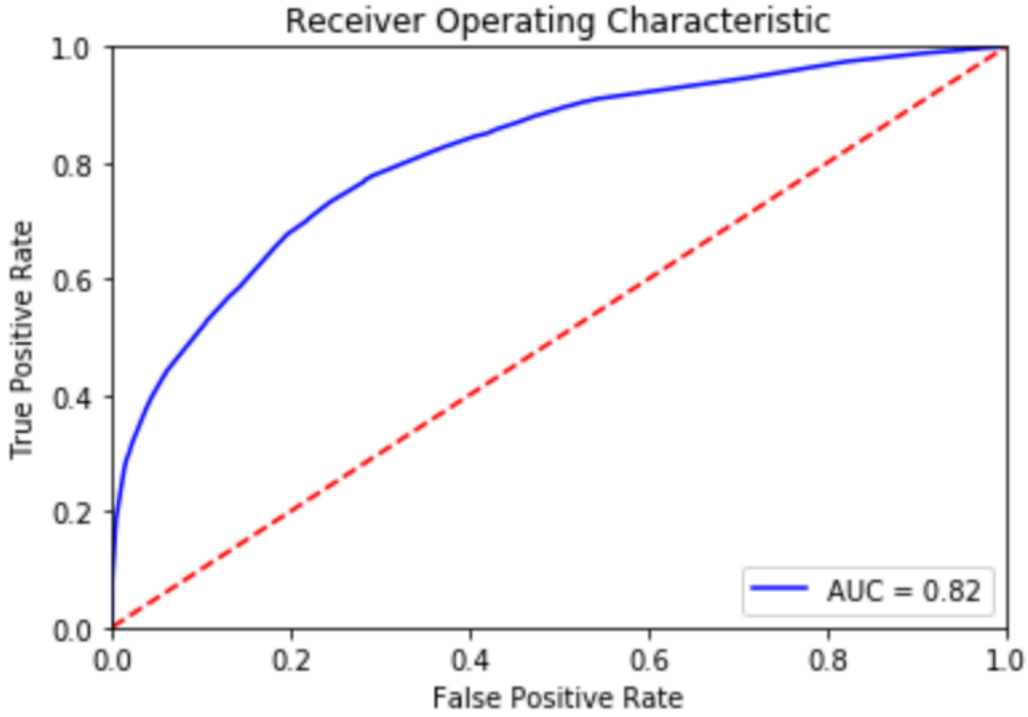


Figure 7 : Courbe ROC de notre modèle pour l'algorithme de Random Forest (n_estimators=500).

VII. CONCLUSION

À travers ce projet, j'ai voulu me concentrer sur deux sujets qui me passionnent, le football, et la Data Science. Cela m'a permis d'avoir une idée de tout le travail que demande la mise en place d'un modèle, et j'ai pu me rendre compte de l'importance du pré traitement des données avant d'appliquer nos algorithmes. Également, de l'importance du choix de l'algorithme dans l'interprétation des résultats.

Le but de ce projet était avant tout d'explorer un modèle d'*expected goals*, de comprendre son fonctionnement, et de tenter de l'améliorer. J'ai essayé d'expérimenter tous ces points en détails, et de le retranscrire dans ce rapport, en y joignant le code.

Aussi, je me suis rendu compte de l'importance de la qualité des données. Celles-ci jouent un rôle primordial dans la construction du modèle, et dans son interprétation, car selon elles, on peut faire certaines remarques sur les résultats obtenues, et les expliquer.

Dans notre cas par exemple, si nos données comportaient les coordonnées de localisation exactes du tireur au moment de la frappe, et le nombre de défenseur entre le tireur et le gardien, ainsi que l'angle exact de tir, ou encore si nous avions pris en compte les caractéristiques techniques, et les aptitudes de chaque joueur, notre modèle aurait certainement été plus précis, car capable de déchiffrer plus d'informations.

Ici, avec un jeu de données conséquent, 28 caractéristiques pour chaque tir, et tout cela en accès libre, à l'aide de mes connaissances, et de mes recherches, j'ai fait de mon mieux pour présenter un modèle cohérent, et le plus précis possible.

Je pense que le sujet des *expected goals*, et plus largement, des statistiques avancées dans le football, est un sujet qui va continuer à se développer, et où il y a encore un travail de recherche important à poursuivre pour aider à une meilleure compréhension de ce sport si imprévisible.

BIBLIOGRAPHIE

[1] Antoine Chirat. Moyenne de buts par match, la ligue 1 à la traîne, ONZE MONDIAL, 2019. https://www.onzemondial.com/ligue-1/moyenne-buts-match-ligue-1-bundesliga-premier-league-184695

[2] Didier Calcei. Big data et football : comment jongler avec les données ? The Conversation, 2018. https://theconversation.com/big-data-et-football-comment-jongler-avec-les-donnees-98477

[3] Gaétan Raoul. L'OM se transforme grâce à la donnée et au machine learning, LeMagIT, 2020. https://www.lemagit.fr/etude/LOM-se-transforme-grace-a-la-donnee-et-au-machine-learning

[4] Thymothé Pinon. Data, algorithmes, gains marginaux : Brentford, le foot 3.0, FranceFootball, 2020.
https://www.francefootball.fr/news/Data-algorithmes-gains-marginaux-brentford-le-foot-3-0/1108349

[5] Frédéric Yang. Le football est-il devenu un sport de stats ? FootMercato, 2020.
https://www.footmercato.net/a49594770965701447 0-le-football-est-il-devenu-un-sport-de-stats
[6] Christoph Biermann. Big Data Foot, edition Marabout, 2019.

[7] Gabriel Manfredi. Expected goals & player analysis, Kaggle, 2019.
https://www.kaggle.com/gabrielmanfredi/expected-goals-player-analysis/notebook

[8] Nicolas Pasquier. Cours de Data Mining et Machine Learning, Master 1 Ingénierie Mathématiques, Université Côte d'Azur, 2019.

[9] Classifieur linéaire, Wikipédia, 2020.
https://fr.wikipedia.org/wiki/Classifieur_linéaire

[10] Ayush Pant. Introduction to Logistic Regression, TowardsDataScience, 2019. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

[11] Claire Della Vedova. Introduction à la régression logistique, Statistiques et Logiciel R, 2020. https://statistique-et-logiciel-r.com/regression-logistique/

[12] Tara Boyle. Dealing with imbalanced data, TowardsDataScience, 2019.https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

[13] Raphael Cosmidis et Julien Assuncao. Les « expected goals », au cœur de la révolution statistiques, CahiersDuFootball, 2015. http://www.cahiersdufootball.net/article-les-expected-goals-au-coeur-de-la-revolution-statistique-5744

[14] Benjamin Cronin. An analysis of different expected goals model, Pinnacle, 2017. https://www.pinnacle.com/en/betting-articles/Soccer/expected-goals-model-analysis/MEP2N9VMG5CTW99D

[15] Michael Caley. Premier league projections and new expected goals, Cartillage Free Captain, 2015. https://cartilagefreecaptain.sbnation.com/2015/10/19/9295905/premier-league-projections-and-new-expected-goals#methoderology

[16] Mike Goodman. The Dual Life of Expected goals, StatsBomb, 2018. https://statsbomb.com/2018/05/the-dual-life-of-expected-goals-part-1/

[17] Peter McKeever. Building an expected goals model in python, Peter McKeever fanalytics, 2019. http://petermckeever.com/2019/01/building-an-expected-goals-model-in-python/

[18] Transfermarkt, 2020. https://www.transfermarkt.fr/ligue-1/eigentorstatistik/wettbewerb/FR1