DATA SCIENCE : PROJET FIL ROUGE

RAPPORT TECHNIQUE

Paris sportifs Football

Gaspard CANEVET gaspard.ganevet@imt-atlantique.net

Amaury TISSEAU amaury.tisseau@imt-atlantique.net

Aline BENABBOU aline.benabbou@imt-atlantique.net

Rida ENNAMIRI rida.ennamiri@imt-atlantique.net

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Décembre 2020

Introduction

Le football est de loin le sport le plus médiatisé de la planète, l'engouement qu'il provoque soulève des stades entiers mais également des sommes impressionnantes. En effet, de nombreux business fructueux tournent autour du ballon rond, et l'issue d'un seul match peut avoir des retombées financières importantes. Les jeux d'argent sont un des aspects les plus lucratifs liés à ce sport, avec le principe des paris sportifs. Lors de la coupe du monde de foot en 2018, plus de 690 millions d'euros ont été pariés par les particuliers en France (source Arjel).

Les parieurs, souvent perdants, se démènent constamment pour essayer de briser le sort et pronostiquer l'issue des matchs afin de décrocher le gros lot. Malgré tout, les sites de paris finissent toujours gagnants sur le long terme. Nous nous sommes alors intéressés à ce problème mais avec un regard de Data Scientist. Serait-il possible de prédire les résultats des matchs en utilisant les données disponibles sur ces derniers et utiliser ces prédictions pour faire pencher la balance de notre côté et devenir rentable avec des paris ?

Problématique métier

Sur quels matchs de Football et quels sites parier pour maximiser ses gains ?

Notre but est de vendre un service conseillant le parieur dans ses paris sur le Football. Cette aide statistique lui permettra de faire de meilleurs choix dans ses paris et ainsi d'augmenter sa rentabilité (ou au moins, limiter ses pertes).

Nos algorithmes auront donc pour but de prédire l'issue des matchs, et en récoltant les cotes des différents sites de paris disponibles sur le net, nous pourrons ainsi assister le parieur dans ses choix. Cet outil ne permettra certainement pas de prévoir à coup sûr l'issue des matchs qui est bien trop aléatoire, mais d'avoir une aide statistique permettant d'être en positif sur le long terme malgré la marge que prennent les bookmakers sur les paris.

***Bookmaker** : équipes de spécialistes des paris sportifs embauchés par les sites afin de fixer les cotes des matchs.*

Ce service pourrait être accessible au particulier en l'échange par exemple d'un abonnement mensuel, qui lui donnerait accès à une application web ou mobile hébergeant nos prédictions et conseils sur les matchs à venir.

Si ce business modèle B2C venait à ne pas fonctionner, nous avons également pensé à la possibilité de travailler sous la direction d'une plus grosse structure (Par exemple site web l'Équipe) donc dans un modèle plus accès B2B. Vous pouvez trouver un exemple de Business modèle en annexe (9).

Traduction en problème de Data Science

En pratique, notre projet peut être traduit comme deux tâches liées distinctes liées à la Data :

La première tâche, qui est la réalisation des **prédictions,** est un problème de **classification**, que nous allons tenter de résoudre à l'aide d'algorithmes de machine learning. En effet, les trois issues d'un match de Football sont Victoire de l'équipe à domicile, Victoire de l'équipe à l'extérieur, ou match nul. Prédire l'issue du match revient donc à classifier le match dans une de ces 3 classes. Nous aurons ainsi besoin d'un maximum de données sur les matchs et équipes des chaque saison.

La seconde tâche, est la **comparaison des cotes** sur les différents sites afin de conseiller le parieur sur le site à choisir. Ell est théoriquement assez simple, car il s'agit de trouver le site avec la cote la plus élevée sur l'issue que nous avons prédite. C'est donc plutôt une tâche de **Web Scrapping**, afin d'automatiser la récupération des cotes sur les différents sites.

Résultats attendus

Nous sommes bien conscients que nous n'arriverons pas à prédire avec une précision infaillible l'issue de tous les matchs, car ces derniers ont une part de surprise et d'imprévisible conséquente. Nous espérons tout de même obtenir des résultats de prédictions du niveau d'un être humain qui suit le Football régulièrement et connaît bien la plupart des équipes.

Nous comptons également faire des simulations d'années de paris sur les années précédentes, en regardant combien d'argent aurait-on gagné ou perdu si l'on avait parié pendant toute une saison en suivant nos pronostics. Nous espérons au moins ne pas être en négatif après une année de paris, ce qui serait déjà une victoire envers les sites de paris.

Description statistique des données

Afin de réaliser nos prédictions, nous avons besoin de données les plus complètes possibles. Par chance, le Football étant le sport le plus suivi de la planète, on peut trouver des jeux de données d'assez bonnes qualités. C'est le cas du site **football-data.uk**, un site anglais sur lequel on peut trouver des statistiques des matchs sur plus de 20 ans sur toutes les ligues européennes majeures. En tant que bons français, nous avons décidé d'effectuer nos travaux sur la **Ligue 1 française**.

Dans la suite de l'étude, nous allons analyser les données que nous avons en notre possession pour délimiter notre domaine d'étude. Nous allons donc répondre aux questions suivantes :

- Quelles saisons utiliser ?
- Quels sont les paramètres à prendre en compte dans notre étude ?
- Comment préparer les données ?

Il y a un Dataset par saison (**28 saisons**), qui contient **380 lignes**, chaque ligne correspondant à un match. Nous avons donc des informations sur **10640 matchs**. Ce Dataset contient des informations sur l'équipe qui joue à domicile (HomeTeam) et l'équipe qui joue à l'extérieur (AwayTeam). (Annexe1)

Néanmoins chaque saison ne contient pas le même nombre de caractéristiques. En effet les saisons **1993/1994 à 2006/2007** ne contiennent que **les équipes** et **le résultat match.** Aucune caractéristique sur le match n'est renseignée. Les informations sont **trop pauvres.** Nous avons donc décidé de nous passer des saisons 1993/1994 à 2006/2007.

Analysons maintenant des informations sur les saisons 2007/2008 à 2020/2021 pour souligner des différences.

Premièrement le nombre de buts moyen par saison en fonction de si l'équipe est à domicile ou à l'extérieur (Annexe 2). Le nombre de buts inscrit par les équipes à domicile est plus important que le nombre de buts inscrit à l'extérieur. Plusieurs paramètres peuvent expliquer cette information, comme **l'appui du public**. Aussi appelé le douzième joueur, il motive l'équipe à domicile et crispe l'équipe à l'extérieur. Ainsi au vu de ces informations, nous pouvons penser que l'algorithme va avantager les équipes à domiciles plutôt que les équipes à l'extérieur.

Dans un second temps, nous allons analyser les variations du nombre de cartons jaunes moyen par saison et du nombre de cartons rouges moyen par saison (Annexe3). Pour les cartons jaunes et rouges, c'est l'équipe à l'extérieur qui en prend le plus. Ceci peut, par exemple, aussi s'expliquer par la **pression mise par le public** poussant les adversaires à être plus agressifs. De plus, les données

sur les cartons sont variables, en particulier pour les cartons rouges. Ainsi, les variables sur les cartons seront source de **variabilités et d'overfitting** pour notre étude prédictive. Nous verrons par la suite si les premières observations seront confirmées.

Enfin, nous terminerons cette analyse descriptive des variables initiales par les classements des saisons 2007/2008 (Annexe4) et 2018/2019 (Annexe5) après la dernière journée de championnat.

Les équipes changent entre chaque saison, il y a la relégation des trois dernières équipes du championnat en ligue 2 et la montée des trois premières équipes de ligue 2 en ligue 1. Nous allons donc devoir, soit ne pas prendre en compte le nom des équipes et trouver une autre façon de les identifier (avec des scores FIFA par exemple), soit de n'étudier que très peu de saisons et que les matchs entre clubs qui auront réussi à se maintenir en ligue 1. La deuxième méthode réduit énormément notre champ d'étude. Nous allons donc plutôt essayer de **définir les équipes autrement que par leurs noms**.

Ensuite, nous voyons que **le résultat des équipes varie en fonction des saisons**. En effet une équipe en forme lors de la saison 2007/2008 pourrait ne plus prétendre à la victoire du championnat en 2018/2019. De nombreux paramètres extérieurs sont à prendre en compte. La **période des transferts**, où des joueurs changent d'équipe à raison d'échange d'argent, est un facteur crucial de l'évolution d'une équipe.

Pour rappel, voilà les questions que nous nous posions au début de cette analyse :

- Quelles saisons utiliser ?
- Quels sont les paramètres à prendre en compte dans notre étude ?
- Comment préparer les données ?

Pour les saisons, nous voulons prédire les résultats sur les saisons récentes, 2018/2019 par exemple. Nous allons prendre un **nombre réduit de dataset**. Nous remonterons jusqu'à trois voire quatre saisons avant la saison de prédiction.

Pour les paramètres, nous savons que les cartons n'étaient pas nécessairement intéressants à étudier. Nous allons donc devoir **choisir les paramètres en fonction des résultats du modèle**.

Enfin dans la suite nous allons modifier les variables de nos datasets pour pouvoir les rendre utilisables par nos algorithmes de prédiction.

Modifications des datasets

L'objectif sera de transformer les données de sorte qu'elles soient **utilisables par notre algorithme de prédiction**.

Dans un premier temps nous allons **modifier les variables de format "string" au format "int"**. Par exemple, le résultat du match. H -> victoire à domicile devient 2, D -> égalité devient 1 et A -> victoire à l'extérieur devient 0.

Ensuite, nous allons **calculer des moyennes** (nombre moyen de but marqué, nombre moyen de tir cadré…), **le nombre de victoires, défaites et égalités** sur les matchs précédents de chaque équipe engagée dans un match. Cela est nécéssaire car avant qu'un match ne soit joué nous n'avons aucune statistique sur ce dernier. Il faut donc regarder des statistiques moyennes de l'équipe.

Après, les premiers essais. Nous avons décidé de rajouter d'autres variables externes au dataset initial pour caractériser les matchs et les équipes, afin de voir si ces dernières peuvent améliorer nos prédictions. Le **budget annuel** de chaque équipe, en millions d'euros, *récolté sur sportune.fr.* Les **notes** des équipes caractérisant leurs capacités offensives, défensive et en milieu de terrain données par les jeux vidéos FIFA (EA sports) de la saison, r*écoltées sur FifaIndex.com.* La **dynamique** de l'équipe sur ses derniers matchs à l'extérieur et à domicile. Une équipe qui a gagné ses 5 derniers matchs à l'extérieur aura une dynamique de +5 à l'extérieur. Si elle perd ses 3 derniers matchs à domicile elle aura une dynamique de -3 à domicile, *calculée à partir du Dataset initial.* Le nombre de **supporters moyen** des équipes à domicile et à l'extérieur, *récolté sur TransferMarkt.com.* La **position de l'équipe dans le classement** après chaque journée (haut de tableau, milieu de tableau et bas de tableau), *Calculée à partir du Dataset initial*

Une fois toutes ces nouvelles features ajoutées, nous essayerons de voir lesquelles nous permettent d'obtenir de **meilleurs résultats**. Nous ne les inclurons pas forcément toutes dans nos modèles. Vous trouverez en annexe 6 un exemple ainsi qu'une explication de notre dataset final.

Nous allons maintenant passer au modèle d'étude. Nous nous sommes basés sur les dernières saisons (de 2015/2016 à 2018/2019). Nous n'avons donc pas utilisé les saisons 2019/2020 et 2020/2021. En effet, **la crise du COVID-19** a fortement impacté le monde du sport, dont le football. La saison 2019/2020 n'a pas été terminée et le début de la saison 2020/2021 est fortement chahuté par les cas de COVID-19 dans les équipes. Ainsi, de nombreux matchs sont reportés et les matchs sont joués sans public ce qui réduit considérablement l'avantage de jouer à domicile. Ces irrégularités rendent l'étude trop compliquée et trop variable pour ces deux saisons.

Modélisation

Une fois nos données à un format utilisable, nous avons essayé d'appliquer nos premiers modèles de machine learning.

Pour évaluer la qualité de nos prédictions, nous utiliserons le score subset accuracy, qui calcule simplement le ***nombre de prédictions exactes sur le nombre total de prédictions****. Nous utiliserons aussi souvent la* ***matrice de confusion****, qui est un outil visuel très efficace pour voir les forces et faiblesses des résultats.*

Nous avons d'abord dû faire le choix des modèles à appliquer. Compte tenu de notre quantité de données assez limitée (même si l'on s'entraîne sur 2 ou 3 ans on dépasse à peine le millier de matchs en Ligue 1), un réseau de neurones n'était clairement pas envisageable. Nous avons donc plutôt considéré des classifier fonctionnant bien avec des quantités de données plus faibles : **SVM**, **Régression Logistique**, **Arbre de décision** et **Random Forest.**

Pour débuter nous sommes parti au plus simple, c'est-à-dire en étudiant une seule saison. Nous entraînons notre modèle sur le début de la saison et nous testons nos prédictions sur la fin de la saison.

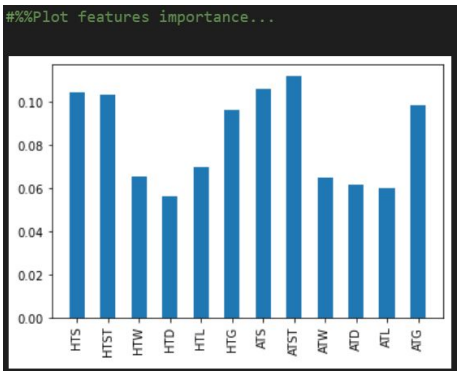
Nos résultats étaient corrects (le meilleur classifier était la **Régression logistique** avec **52%** de réussite), mais nous avons rapidement dû changer d'approche car cette dernière n'était pas compatible avec notre objectif. En effet, avec cette méthode nous ne pouvions prédire que les matchs de la **fin de la saison** (environ 70 matchs). Pour pouvoir conseiller un parieur toute l'année, ce n'est pas envisageable. Ainsi, nous avons donc commencé à entraîner nos modèles sur des saisons passées, afin de prédire les résultats d'une saison complète. L'exemple sur lequel nous avons beaucoup travaillé est la prédiction de la saison 2017-2018. Après plusieurs essais, nous avons remarqué que nous obtenions les meilleurs résultats en nous **entraînant sur 2 saisons**, afin de prédire la 3ème. Dans notre cas, nous entraînons le modèle sur les saisons 2015-2016 et 2016-2017 afin de prédire 2017-2018. En effet, si l'on prend en compte plus de saisons, il y a trop de changements dans les équipes et le classement ce qui n'est pas bon pour l'entraînement du modèle. Ces premiers tests sont faits avec uniquement le dataset initial, sans les ajouts de features. Vous pouvez retrouver les matrices de confusions des différentes méthodes en **Annexe 7**.

Les deux premiers résultats ne sont pas vraiment satisfaisant, **l'arbre de décision** donne un score de 44%, ce qui n'est pas très élevé, et le **SVM** même s'il obtient un meilleur score en théorie, obtient ce score en ne prédisant aucun match nul, ce qui n'est pas vraiment viable pour faire du conseil aux

parieurs. **La Régression logistique**, obtient un bon score, mais garde toujours une tendance à ne pas prédire beaucoup de matchs nuls, alors que ces derniers représentent quand même 25% des résultats de la saison. Celui qui est pour nous le plus utilisable ici est le modèle **Random Forest** (utilisé ici avec 300 arbres dans la forêt), avec un assez bon score mais également une meilleure répartition des prédictions, avec plus de matchs nuls. (Voir annexe 7)

```
Paris SG - Nice | algo predit : Paris SG | le vrai résultat est : Paris SG
Rennes - Reims | algo predit : Rennes | le vrai résultat est : Rennes
Toulouse - Caen | algo predit : Caen | le vrai résultat est : Toulouse
Troyes - Angers | algo predit : Angers | le vrai résultat est : Angers
Bastia - Marseille | algo predit : Bastia | le vrai résultat est : Bastia
Lorient - Lyon | algo predit : Lyon | le vrai résultat est : Lyon
Nantes - Lille | algo predit : Nantes | le vrai résultat est : Lille
Montpellier - Lyon | algo predit : Lyon | le vrai résultat est : Lyon
```

Nous avons par exemple affiché quelques prédictions, et on voit que lors de matchs à l'issue assez prévisible, il ne fait pas souvent de prédictions aberrantes. Sur les 70 matchs de fin de saison, l'algorithme prédit tout le temps la victoire du Paris Saint Germain lorsqu'il joue, club qui domine fortement la ligue 1 depuis plusieurs années.



Afin de comprendre quelles sont les features qui impactent le plus nos prédictions, nous nous sommes intéressés aux **features importances** de notre modèle Random Forest.

On observe qu'il n'y a aucune variable qui domine réellement sur les autres, avec les tirs et tirs cadrés qui sont les plus importantes. Cette répartition assez uniforme est liée au fait que les variables ne sont pas beaucoup corrélées.

Test des nouvelles Features

Afin d'obtenir de meilleurs résultats, nous allons à présent tester nos algorithmes avec les nouvelles variables que nous avons ajoutées à nos datasets. L'objectif est de voir si certaines améliorent significativement nos résultats, et quelle est la meilleure **combinaison de features** à prendre en compte. Au vu de son efficacité précédente, nous montrerons à présent uniquement les résultats du modèle Random Forest.

```
score moyen avec : dataset initial : 0.4893939393939394
score moyen avec : ['budget'] : 0.5227272727272727
score moyen avec : ['fifa'] : 0.5106060606060606
score moyen avec : ['budget', 'fifa'] : 0.5303030303030303
score moyen avec : ['streak'] : 0.5
score moyen avec : ['budget', 'streak'] : 0.5212121212121212
score moyen avec : ['fifa', 'streak'] : 0.4954545454545455
score moyen avec : ['budget', 'fifa', 'streak'] : 0.5212121212121212
score moyen avec : ['public'] : 0.5227272727272727
score moyen avec : ['budget', 'public'] : 0.5393939393939393
score moyen avec : ['fifa', 'public'] : 0.5227272727272727
score moyen avec : ['budget', 'fifa', 'public'] : 0.5318181818181819
score moyen avec : ['streak', 'public'] : 0.49393939393939396
score moyen avec : ['budget', 'streak', 'public'] : 0.5333333333333333
score moyen avec : ['fifa', 'streak', 'public'] : 0.5272727272727273
score moyen avec : ['budget', 'fifa', 'streak', 'public'] : 0.5272727272727273
```

Même si les différents scores sont assez proches en moyenne, on voit que les features **budget** et **public** sont les plus influentes. L'écart relatif entre simplement le dataset initial et la combinaison public et budget est de 10%, ce qui est une

amélioration notable, avec un score de plus de 53% qui est déjà une belle réussite pour nous. Nous utiliserons à présent cette combinaison de features et laisserons les autres de côté.

Certitude des prédictions

Afin de conseiller un parieur, donner simplement une prédiction n'est pas forcément la meilleure approche. En effet, si 2 équipes de niveau équivalent s'affrontent, les probabilités de victoires estimées par l'algorithme risquent d'être assez proches et retourner l'issue la plus probable n'est pas très représentatif. Nous allons donc présenter au parieur le **pourcentage de certitude** de chacune des issues du match.

```
Lille - Nice // prédictions: Lille= 0.36 | Draw= 0.4 | Nice= 0.24 // Le vrai résultat est Draw

Marseille - Troyes // prédictions: Marseille= 0.66 | Draw= 0.08 | Troyes= 0.26 // Le vrai résultat est Marseille

Metz - Strasbourg // prédictions: Metz= 0.25 | Draw= 0.31 | Strasbourg= 0.44 // Le vrai résultat est Metz

Monaco - Rennes // prédictions: Monaco= 0.78 | Draw= 0.2 | Rennes= 0.02 // Le vrai résultat est Monaco

Paris SG - Caen // prédictions: Paris SG= 0.72 | Draw= 0.11 | Caen= 0.17 // Le vrai résultat est Paris SG

Toulouse - Lyon // prédictions: Toulouse= 0.19 | Draw= 0.17 | Lyon= 0.64 // Le vrai résultat est Lyon
```

Ces résultats permettront au parieur d'estimer le **niveau de confiance** d'une prédiction, et de savoir quel est le risque qu'il prend avec tel ou tel pari.

Simulation d'une année de pari

En utilisant des archives des cotes de la saison 2017-2018 présentes dans nos datasets (site bet365), nous simulons une année de paris en suivant nos prédictions à la lettre. On considère que l'on mise la même somme sur tous les matchs pendant toute la saison. Voici les résultats que l'on obtient.

```
[187] def simule_annee_pari(classifier, df_cote_test, X_test, y_test, basebet=10):...

      Forest n°0 - mise de 10 € sur chaque paris - gain final : 160.4000000000019 €
      Forest n°1 - mise de 10 € sur chaque paris - gain final : 229.70000000000164 €
      Forest n°2 - mise de 10 € sur chaque paris - gain final : 255.800000000002 €
      Forest n°3 - mise de 10 € sur chaque paris - gain final : 177.50000000000182 €
      Forest n°4 - mise de 10 € sur chaque paris - gain final : -99.39999999999873 €
      Forest n°5 - mise de 10 € sur chaque paris - gain final : 83.40000000000236 €
      Forest n°6 - mise de 10 € sur chaque paris - gain final : 393.60000000000264 €
      Forest n°7 - mise de 10 € sur chaque paris - gain final : 184.10000000000127 €
      Forest n°8 - mise de 10 € sur chaque paris - gain final : 161.80000000000155 €
      Forest n°9 - mise de 10 € sur chaque paris - gain final : 110.40000000000191 €
```

On voit que sur la plupart des simulations, il y a un **gain d'argent**, allant de 2 à 10 % de la somme totale misée au cours de l'année. La variation est due au fait que les RandomForest sont liées à l'aléatoire, et que donc certaines forêts sont meilleures que d'autres. La simulation où l'on perd 100 euros montre que nos modèles restent encore trop incertains et ne sont pas encore assez **précis** et **constants**.

Récolte des cotes sur les sites de paris Français

Enfin, pour pouvoir diriger le parieur sur le bon site de pari en fonction de nos prédictions et de son choix, nous devons **récupérer** automatiquement les **nouvelles cotes** sur ces différents sites. Nous avons fait une ébauche de cette automatisation à l'aide du module python Selenium qui permet d'automatiser les navigateurs web et ainsi par exemple de faire du WebScrapping.

Ainsi, nous avons des algorithmes qui récupèrent automatiquement les cotes sur Winamax, PMU et Parionssport.

Nous n'avons malheureusement pas pu utiliser ces cotes réellement, la saison 2020-2021 étant très perturbée par la situation sanitaire liée au COVID, nos modèles entraînés sur les années passées ne sont plus du tout performants.

```
[21] print(main())

     {'Nice - Rennes': [2.95, 3.35, 2.3], 'Brest - Reims': [2.2, 3.25, 3.2], 'Lorient - Nîmes': [2.05, 3.35, 3.55], 'Nantes - Dijon': [1.63, 3.95,
     4.95], 'Strasbourg - Metz': [1.85, 3.45, 4.15], 'Lille - Bordeaux': [1.66, 3.65, 5.25], 'Paris SG - Lyon': [1.45, 5.0, 5.5], 'Angers - Strasbo
     urg': [2.1, 3.35, 3.3], 'Dijon - Lille': [6.0, 4.1, 1.5], 'Montpellier - Metz': [1.72, 3.55, 4.7], 'Nimes - Nice': [3.0, 3.35, 2.25], 'Reims -
     Nantes': [2.35, 3.15, 3.0], 'Paris SG - Lorient': [1.05, 11.0, 25.0], 'Bordeaux - Saint-Étienne': [2.1, 3.35, 3.35], 'Lyon - Brest': [1.3, 5.
     5, 8.0], 'Monaco - Lens': [1.6, 4.1, 4.75]}
```

Étude juridique de notre projet

Nous collectons nos données sur au moins cinq plateformes (cf Modification des datasets). Chacune a des **législations** différentes mais de manière générale on ne peut pas utiliser les données à des **fins commerciales**. Ainsi, pour pouvoir utiliser notre algorithme, deux possibilités s'offrent à nous. Soit nous signons des contrats avec **toutes** les plateformes, soit nous signons un contrat avec **une seule** plateforme qui dispose d'une base de **données importante** sur le football comme le journal L'Équipe. Par exemple, nous pourrions leur permettre de valoriser leur jeu de données en leur permettant de faire un article en plus sur leur **prédictions** des résultats des matchs.

Néanmoins cela n'est qu'une hypothèse pour se réorienter le cas échéant. Revenons au cas où nous allons créer une plateforme avec des clients qui vont payer un **abonnement** pour pouvoir disposer de nos prédictions. Nous traitons des données d'utilisateurs et sur les matchs de football. Nous devons donc nous assurer de respecter les normes **RGPD**.

Nous identifions deux activités principales qui nécessitent le traitement de données. La prédiction des matchs et la gestion des utilisateurs. Vous trouverez ci-dessous la démarche que nous pensions effectuer pour respecter les normes RGPD :

Données sur les matchs	Données utilisateur
Finalité : Prédire le résultat des matchs **Catégorie :** Données accessibles publiquement mais protégées par les sites qui les partagent. → Nécessité de signer un contrat avec ces sites **Accès :** Seulement l'algorithme et les data scientists qui travaillent sur les prédictions **Durée de conservation :** Dépend du nombre de saisons nécessaires à l'entraînement (pour l'instant n-3)	**Finalité :** Guider le parieur, gestion du compte et des transactions **Catégorie :** Données sensibles, à protéger. → Nécessité de respecter les normes RGPD et faire accepter les conditions d'utilisations à l'inscription **Accès** : Service client, algorithmes de suggestion **Durée de conservation :** six ans après suppression d'un compte pour pouvoir se défendre en cas d'accusation.

Nous sommes conscients que la partie juridique est cruciale dans notre projet. Ce sera l'aspect lucratif de notre business plan qui sera le plus compliqué à défendre. Notamment, l'utilisation des cotes présentes sur les sites de paris pour conseiller le parieur.

Conclusion

Même si nous avons des résultats **correct**s sur une saison avec nos modèles, ces derniers restent encore trop **aléatoires**. En effet, en faisant des tests nous pouvons évaluer si notre modèle est bon ou pas, alors qu'en situation réelle, nous ne le saurons qu'à la fin de la saison, après avoir gagné ou perdu de l'argent. Notre outil reste tout de même intéressant, surtout par sa fonctionnalité donnant les certitudes des différentes issues, qui permet de bien évaluer le risque d'un pari.

Cependant, notre modèle est encore largement **améliorable**, une approche possible serait d'ajouter plus de **variations** dans les données d'entrée. En effet, une grande partie des matchs sont des victoires de l'équipe à domicile. L'algorithme a donc moins d'informations pour prédire les victoires des équipes à l'extérieur. Il pourrait être intéressant de choisir les matchs que l'on prend entre 3 saisons pour mieux équilibrer le dataset.

Il est également possible d'ajouter d'autres features comme la **météo**, le nombre de **blessés**, **l'arbitre**... En allant plus loin il faudrait s'intéresser précisément aux joueurs même en plus des équipes. Avec la **composition** / feuille de match de chaque rencontre, il est déjà bien plus facile de faire des pronostics. Ajouter cela à nos modèles serait assez complexe, mais apporterait sans aucun doute une nette amélioration. D'autres outils statistiques assez poussés comme les **xG goals** (expected goals) pourraient encore améliorer nos prédictions, mais cela devient un travail à plein temps.

Annexes :

1) Colonnes du dataset initial

Div = Division

Date = Match Date (dd/mm/yy)

Time = Time of match kick off

HomeTeam = Home Team

AwayTeam = Away Team

FTHG and HG = Full Time Home Team Goals

FTAG and AG = Full Time Away Team Goals

FTR = Résultat du match (H=Home Win, D=Draw, A=Away Win)

HTHG = Half Time Home Team Goals

HTAG = Half Time Away Team Goals

HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

HS = Home Team Shots

AS = Away Team Shots

HST = Home Team Shots on Target

AST = Away Team Shots on Target

HF = Home Team Fouls Committed

AF = Away Team Fouls Committed

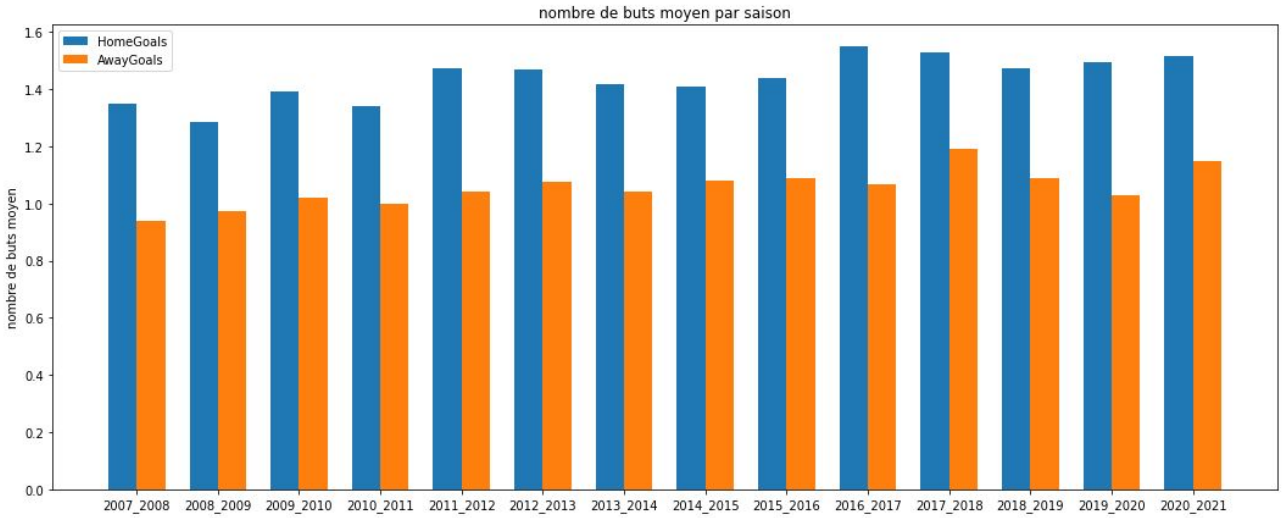
HY = Home Team Yellow Cards

AY = Away Team Yellow Cards

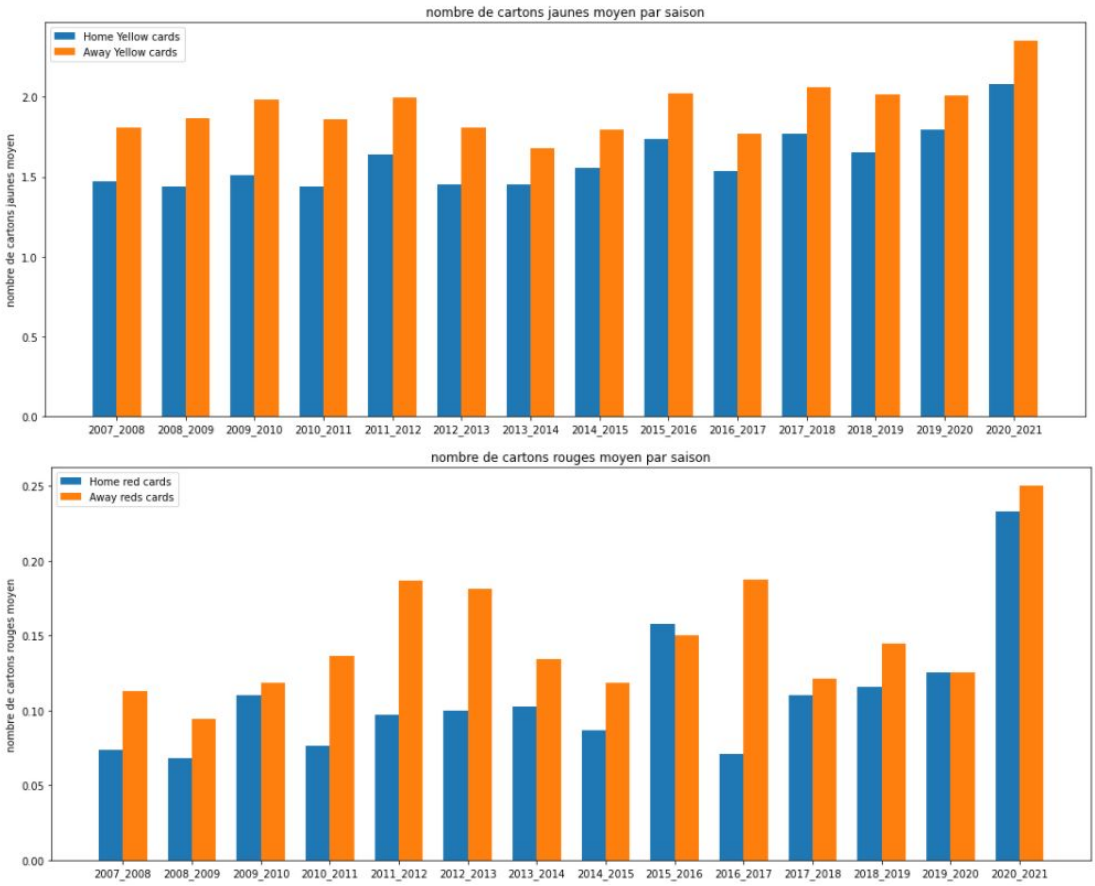
HR = Home Team Red Cards

AR = Away Team Red Card

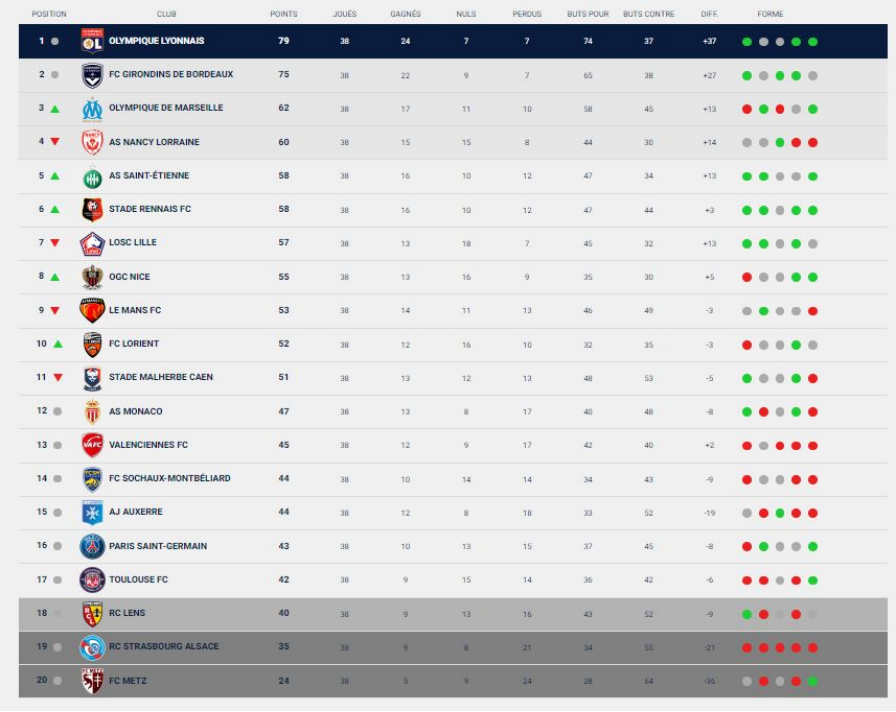
2) Nombre moyen de but par saison



3) nombre de cartons jaunes moyen par saison et du nombre de cartons rouges moyen par saison

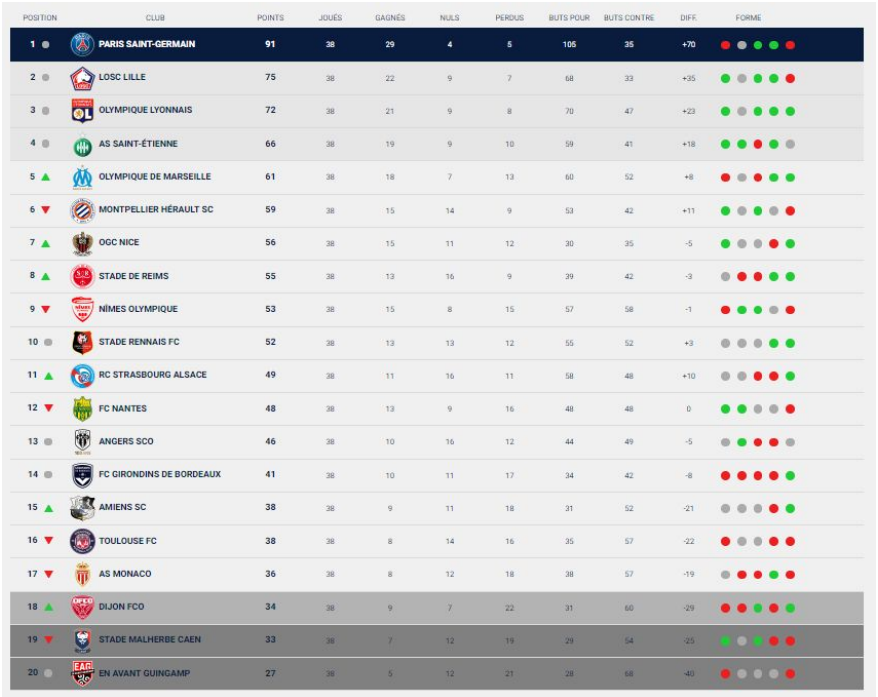


4) classement 2007/2008



POSITION	CLUB	POINTS	JOUÉS	GAGNÉS	NULS	PERDUS	BUTS POUR	BUTS CONTRE	DIFF.	FORME
1	OLYMPIQUE LYONNAIS	79	38	24	7	7	74	37	+37	
2	FC GIRONDINS DE BORDEAUX	75	38	22	9	7	65	38	+27	
3	OLYMPIQUE DE MARSEILLE	62	38	17	11	10	58	45	+13	
4	AS NANCY LORRAINE	60	38	15	15	8	44	30	+14	
5	AS SAINT-ÉTIENNE	58	38	16	10	12	47	34	+13	
6	STADE RENNAIS FC	58	38	16	10	12	47	44	+3	
7	LOSC LILLE	57	38	13	18	7	45	32	+13	
8	OGC NICE	55	38	13	16	9	35	30	+5	
9	LE MANS FC	53	38	14	11	13	46	49	-3	
10	FC LORIENT	52	38	12	16	10	32	35	-3	
11	STADE MALHERBE CAEN	51	38	13	12	13	48	53	-5	
12	AS MONACO	47	38	13	8	17	40	48	-8	
13	VALENCIENNES FC	45	38	12	9	17	42	40	+2	
14	FC SOCHAUX-MONTBÉLIARD	44	38	10	14	14	34	43	-9	
15	AJ AUXERRE	44	38	12	8	18	33	52	-19	
16	PARIS SAINT-GERMAIN	43	38	10	13	15	37	45	-8	
17	TOULOUSE FC	42	38	9	15	14	36	42	-6	
18	RC LENS	40	38	9	13	16	43	52	-9	
19	RC STRASBOURG ALSACE	35	38	9	8	21	34	55	-21	
20	FC METZ	24	38	5	9	24	28	64	-36	

5) classement 2018/2019



POSITION	CLUB	POINTS	JOUÉS	GAGNÉS	NULS	PERDUS	BUTS POUR	BUTS CONTRE	DIFF.	FORME
1	PARIS SAINT-GERMAIN	91	38	29	4	5	105	35	+70	
2	LOSC LILLE	75	38	22	9	7	68	33	+35	
3	OLYMPIQUE LYONNAIS	72	38	21	9	8	70	47	+23	
4	AS SAINT-ÉTIENNE	66	38	19	9	10	59	41	+18	
5	OLYMPIQUE DE MARSEILLE	61	38	18	7	13	60	52	+8	
6	MONTPELLIER HÉRAULT SC	59	38	15	14	9	53	42	+11	
7	OGC NICE	56	38	15	11	12	30	35	-5	
8	STADE DE REIMS	55	38	13	16	9	39	42	-3	
9	NÎMES OLYMPIQUE	53	38	15	8	15	57	58	-1	
10	STADE RENNAIS FC	52	38	13	13	12	55	52	+3	
11	RC STRASBOURG ALSACE	49	38	11	16	11	58	48	+10	
12	FC NANTES	48	38	13	9	16	48	48	0	
13	ANGERS SCO	46	38	10	16	12	44	49	-5	
14	FC GIRONDINS DE BORDEAUX	41	38	10	11	17	34	42	-8	
15	AMIENS SC	38	38	9	11	18	31	52	-21	
16	TOULOUSE FC	38	38	8	14	16	35	57	-22	
17	AS MONACO	36	38	8	12	18	38	57	-19	
18	DIJON FCO	34	38	9	7	22	31	60	-29	
19	STADE MALHERBE CAEN	33	38	7	12	19	29	54	-25	
20	EN AVANT GUINGAMP	27	38	5	12	21	28	68	-40	

6) Colonnes du dataset final

Div = Division

Date = Match Date (dd/mm/yy)

Time = Time of match kick off

HomeTeam = Home Team

AwayTeam = Away Team

FTR = Résultat du match (H=Home Win, D=Draw, A=Away Win)

HTG = Home Team Goals mean

ATG = Away Team Goals mean

HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

HTS = Home Team Shots mean

ATS = Away Team Shots mean

HTST = Home Team Shots on Target mean

ATST = Away Team Shots on Target mean

HTF = Home Team Fouls Committed mean

ATF = Away Team Fouls Committed mean

HTY = Home Team Yellow Cards mean

ATY = Away Team Yellow Cards mean

HTR = Home Team Red Cards mean

ATR = Away Team Red Cards mean

HTW = Home Team win

ATW = Away Team Win

HTD = Home Team draw

ATD = Away Team draw

HTL = Home Team lose

ATL = Away Team Lose

HATT = Home Team Fifa attack level

AATT = Home Team Fifa attack level

HMIL = Home Team Fifa midfield level

AMIL = Away Team Fifa midfield level

HDEF = Home Team Fifa defence level

ADEF = Away Team defence level

Home attendance

Away attendance

HomeTeamStreak_AtHome

AwayTeamStreak_AtHome

HomeTeamStreak_AtAway

AwayTeamStreak_AtAway

HRP = Home team ranking position (high low or middle)

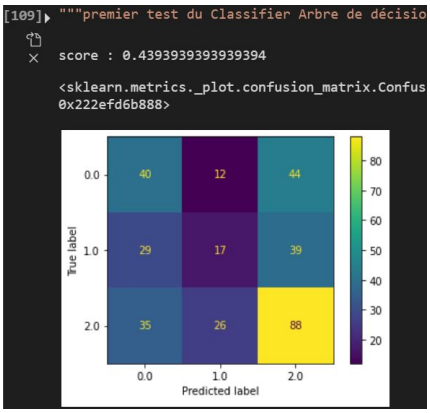
ARP = Away team ranking position (high low or middle)

7) Matrice de confusion

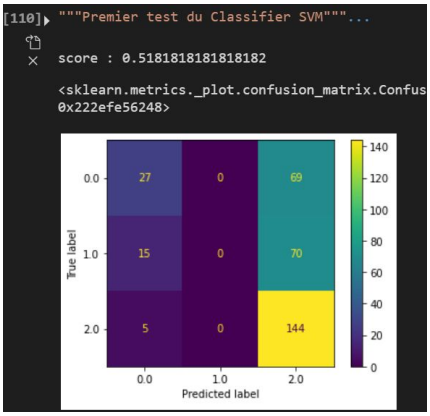
2 = Victoire du Domicile

1 = Match nul

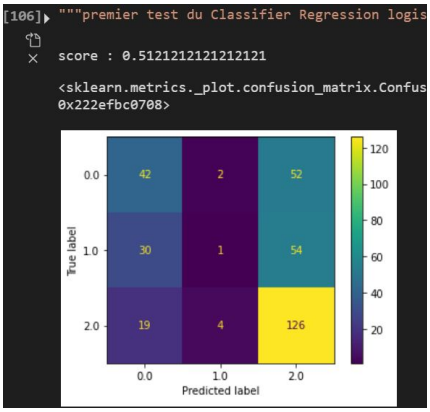
0 = Victoire de l'extérieur



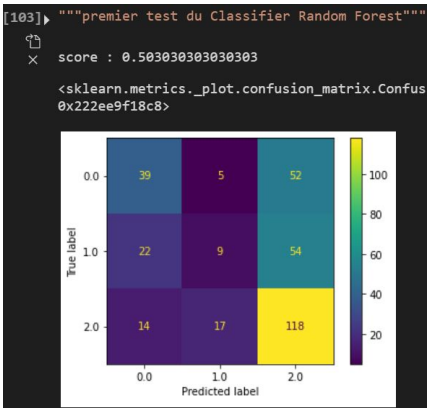
Arbre de décision



SVM

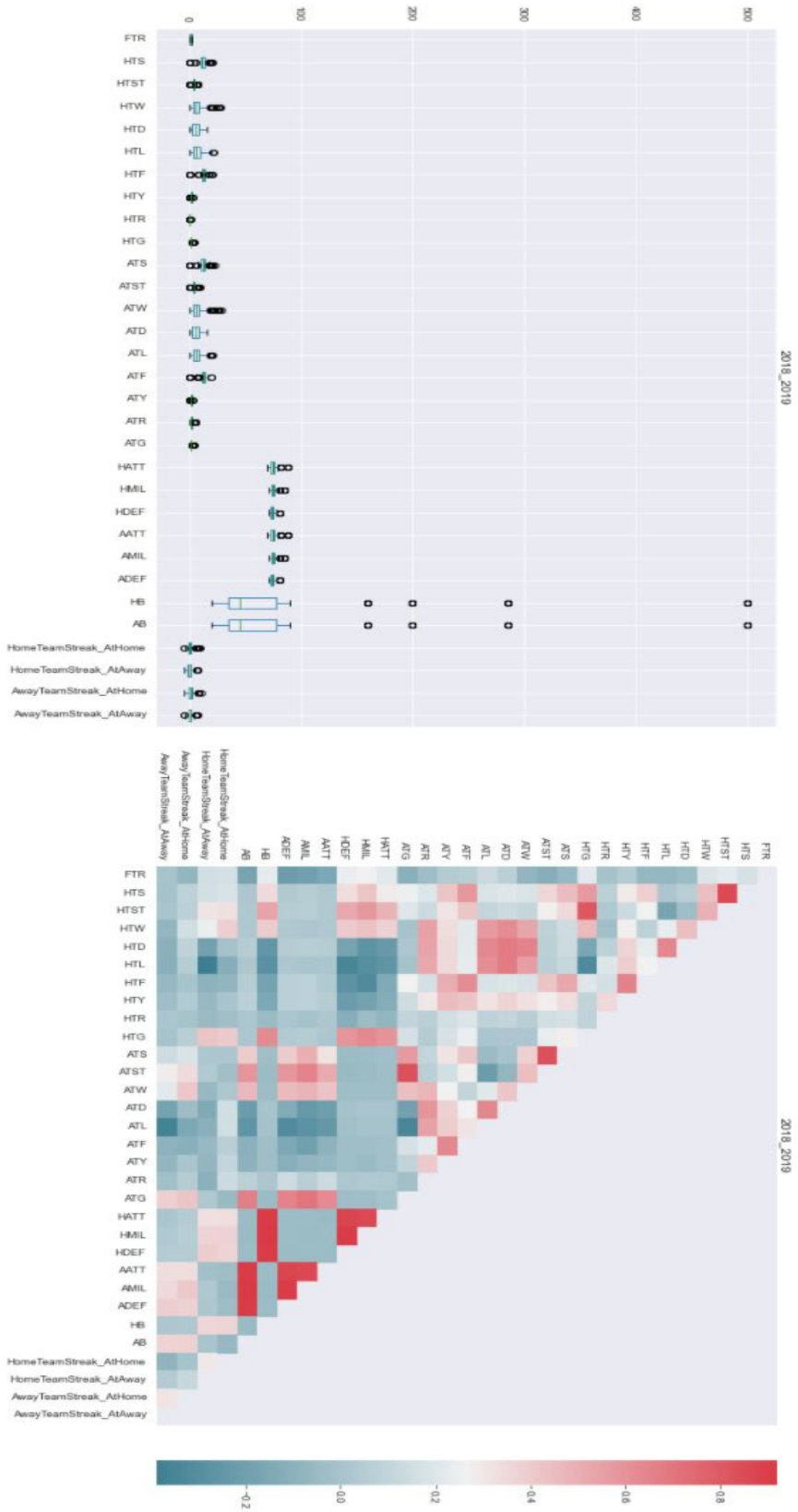


Logistic Regression



Random Forest

8) Corrélation et boxplot



9) Business model Canvas

<table>
<tr>
<td rowspan="2">Partenaires clés

-Les utilisateurs, ie les parieurs

-Les organisations qui nous partagent les données</td>
<td>Activités clés

-Développer l’algorithme de prédiction (data engineer, data scientist)
-Gérer la plateforme (dev)
-Spécialiste du droit de la donnée
-Spécialiste en cyber sécurité
-Fidéliser la clientèle (community manager)
- Gérer les comptes clients (transactions, données perso)</td>
<td rowspan="2">Proposition de valeur

-Maximiser les gains des parieurs en utilisant les caractéristiques des équipes et du Machine Learning</td>
<td>Relations clients

-Climat de confiance (on est transparent sur le fonctionnement de notre algorithme, ie mise à jour, variables décisionnelles etc, et de notre plateforme)
-Création d’une communauté</td>
<td rowspan="2">Segmentation

-Parieurs de plus de 18 ans qui cherchent à maximiser ses gains

-Parieurs fatigués de suivre des conseillers en pari qui peuvent être malhonnête.</td>
</tr>
<tr>
<td>Ressources clés

-Les données collectées sur les équipes et les matchs mais aussi sur les clients
-L’algorithme de prédiction
-La plateforme (serveur)</td>
<td>Distribution

-Plateforme internet application ou site web
-Pub par nos partenaires</td>
</tr>
<tr>
<td colspan="3">Coût de structure

-Maintenance et gestion de la plateforme et de l’algorithme de prédiction
-Contrats pour l’utilisation des données sur les équipes
-Salariés permanents
-Coûts marketing</td>
<td colspan="2">Sources de revenus

-Abonnement avec un compte pour avoir accès aux prédictions et conseils
-Plateforme gratuite mais on revend les données clients.
-Publicités sur le site web</td>
</tr>
</table>

Références bibliographiques

Datasets initiaux: https://www.football-data.co.uk
Scores FIFA: http://fifaindex.com/fr/teams.com/fr/teams
Public moyen: http://transfermarkt.com
Budget des clubs: http://sportune.fr
Chiffres clés sur le football: https://www.parieur-gagnant.com/chiffres-paris-france-infographie/
Droit de la donnée:
https://moodle.imt-atlantique.fr/pluginfile.php/28479/mod_resource/content/1/IMT%20-%20Software%20Ecology%2020171108_BJean.pdf
Machine Learning Random Forest: https://moodle.imt-atlantique.fr/mod/resource/view.php?id=27571