Machine Learning – Rapport final CHARPENTIER Nicolas – 16/03/2017

# Rappel du sujet

Le but de ce projet était d’analyser les différentes équipes composant un championnat de football, leurs victoires, défaites, buts marqués et encaissés, etc., et ensuite pouvoir prédire certaines victoires et à la finale, essayer de prédire le possible gagnant du championnat.

Pour se faire, nous savons que les grands championnats européens sont très indécis et improbables, nous pensons en particulier au club de Leicester remportant le championnat d’Angleterre l’an dernier avec un budget très inférieur aux grands clubs anglais. Ce club connu pour le rugby qui venait directement de deuxième ligue anglaise. Il est aussi possible d’énoncer Montpellier en 2012, qui fût champion de France à la surprise générale devant le grandissime favoris (le Paris Saint Germain).

C’est pour cela que nous allons nous intéresser à un championnat d’une division inférieur, car les statistiques sont beaucoup plus fiables du fait qu’il y a moins d’argent en jeux, moins de professionnalisme et donc moins de variables inconnues.

J’ai choisi la deuxième division du championnat allemand : La Bundesliga 2. J’ai choisi ce championnat car les championnats allemands ont pour réputation d’être stable et homogène, ce qui convient parfaitement pour nos analyses.

Ces analyses seront effectuées sur les 9 dernières années, ce qui permettra d'avoir un jeu de données assez denses.

Cependant, et malgré un jeu de données dense et assez loin dans le passé, il s’est tout de suite montré compliqué de prédire un vainqueur de championnat car 6 équipes différent d’une année à l’autre (les 3 premiers montent en division supérieur, et les 3 derniers descendent en division inférieur).

Sachant cela, j’ai décidé de changer d’optique, et essayer de prédire désormais le vainqueur entre 2 équipes, sur leur prochain match.

# Données

Pour réaliser ce projet, j’ai récupéré mes données sur football-data.co.uk.

J’ai récupéré les données des saisons allant de 2007 à 2016, où les saisons étaient terminées, puis la saison 2017, qui est en cours, et qui me permettra de vérifier mes analyses.

Chaque jeu de données a la même structure d’une année à l’autre, ce qui m’a permis de les analyser de la même manière, et de minimiser le nettoyage sur celles-ci.

Chaque ligne de mon dataset représente un match, avec les deux équipent qui s’affrontent, le gagnant, le nombre de buts pour chaque équipe, le nombre de tirs, etc., ainsi que les cotes de différents bookmakers.

# Algorithmes utilisés

Pour réaliser ce projet, j’ai essayé d’utiliser l’algorithme « Poisson » mais sans réel succès, puis, j’ai essayé de développer un modèle différent, se basant sur différents paramètres attribuant des scores pour chaque équipe en fonction de leurs rencontres précédentes.

Pour cela, je récupère l’ensemble des équipes présentes sur un maximum de saison, et en l’occurrence, j’obtiens 9 équipes qui ont participé à 6 saisons, mon meilleur ratio nombre d’équipe/nombre d’année.

Ensuite, je sélectionne deux équipes dans cette liste pour lesquelles je veux prédire le vainqueur, et j’analyse l’ensemble des rencontres ayant eu lieu entre celles-ci.

Comme dit précédemment, mon but est d’analyser l’ensemble des rencontres ayant eu lieu auparavant en me basant sur le nombre de but marqué, le nombre de victoire, le nombre de « large » victoire (plus de 2 buts d’écarts) et via ceci, attribué un score pour chaque paramètre et pour chaque équipe.

# Résultats

Les résultats sont encore très approximatifs car les paramètres ne sont pas encore assez nombreux, mais ils sont prometteurs car les indicateurs nous permettent de tirer de fortes tendances sur une équipe qui a potentiellement le dessus sur l’autre depuis un certain temps.

# Conclusion & améliorations possibles

Pour finir, et comme dit dans la conclusion du projet, il est nécessaire de posséder plus de paramètres permettant d’influencer le score de chaque équipe, afin d’avoir un score final le plus proche de la réalité possible, mais, il ne faut pas oublier que cela reste du sport, et qu’aucun algorithme, aussi précis et bon soit-il, ne pourra prédire un résultat d’une rencontre sure à 100%.