EXPLORATION DE GRANDS VOLUMES DE DONNEES

Golf : une analyse quantitative de l'US Open sur 4 années

Adrien ROBILIARD
MS Big Data 2020-2021

Table des matières

Notes :

1. Les graphiques et tables sont disponibles dans le fichier « Annexe » du dossier.
2. Ce rapport s'accompagne également d'un notebook présent dans le dossier remis (langage utilisé : Python)
3. Si vous n'êtes pas familier avec le golf et ses termes techniques, vous pouvez vous référer à la page suivante : http://www.lecondegolf.com/glossaire_golf.htm

I. Choix du sujet et mise en problématique

Amateur de golf, je suis de manière régulière le championnat professionnel, et notamment celui avec les golfeurs les plus en vue, ceux qui sont sur le fameux « PGA TOUR », circuit dont les tournois se déroulent principalement aux Etats Unis.

La data science s'applique à de plus en plus de secteurs/milieux, et celui du sport s'y prête très bien, notamment le golf. En effet, de nombreuses variables peuvent être mesurées et mises en relation, comme la distance de drive, le nombre de putt moyens … mais aussi les caractéristiques des golfeurs (taille, poids, âge, nationalité).
J'ai donc décidé d'analyser les performances des golfeurs au tournoi de l'US Open, joué sur le parcours d'Oakmont en Pennsylvanie, et ce sur 4 éditions : 2016, 2017, 2018 et 2020 (les données pour l'année 2019 n'étant pas disponibles).

Cette étude fait suite à des discussions que j'ai pu avoir avec mon entourage sur l'utilisation des données dans la prédiction des résultats sportifs. Le but de ce travail est donc d'explorer les statistiques de golf relatives à chaque golfeur, de les visualiser et d'utiliser des algorithmes de Machine Learning afin de faire des prédictions. Nous utiliserons notamment le clustering pour extraire des « profils de golfeurs », la régression afin de prédire les distances et le classement, mais également des algorithmes de classification pour pronostiquer la présence d'un golfeur dans le top 3 (idem pour top 10).

II. Données : récupération, nettoyage et traitement

Remarque préliminaire : les tournois professionnels de golf se jouent sur 4 jours, 1 tour par jour. Pour atteindre le tour suivant, les golfeurs doivent « passer le cut », une limite de coups qu'ils ne doivent pas dépasser pour atteindre le round suivant. La définition du « cut » dépend du tournoi : limite absolue ou en fonction des autres golfeurs ; entre chaque tour ou entre le 2ème et 3ème trou. Pour être simple, les meilleurs golfeurs « passent le cut » et peuvent joueur le tour suivant.
Pour l'US Open, le « cut » n'entre en vigueur qu'entre le 2ème et 3ème tour, et se fait sur les 60 meilleurs golfeurs (scores accumulés lors des 2 premiers tours), incluant les ex-aequo.

Les données sur le site pgatour.com ne sont disponibles que pour les golfeurs ayant « passé le cut », l'étude sera donc réalisée sur les joueurs les plus performants. Nous pouvons tout de même analyser un nombre suffisamment élevé de joueurs.

Les statistiques mentionnées dans les paragraphes ci-dessous sont répertoriées et décrites en annexe (figure 1). Nous aurons au total 26 variables pour 275 observations.

1) Statistiques de jeu

Les différentes statistiques sont récupérées sous forme de dataframes Pandas à l'aide de la méthode « read_html » pour les 4 années, pour un total de 40 dataframes. L'étape suivante consiste à concaténer les dataframes correspondant aux mêmes variables sur les différentes années. Chaque nouveau dataframe contient donc une colonne correspondant au golfeur, une colonne correspondant à une statistique et une colonne correspondant au rang du golfeur pour cette statistique (le meilleur golfeur pour cette statistique aura le rang 1). Les différents dataframes sont enfin assemblés avec un « merge » sur la clé correspond au couple (Golfeur, Année).

Les statistiques de rang doivent être modifiées : certains golfeurs peuvent être ex-aequo, un « T » s'ajoute alors à leur classement. Pour pouvoir analyser ces données par la suite, la lettre est enlevée et le type de la variable changé en entier. Les distances au driver sont initialement en yards, je les convertis en mètres.

2) Caractéristiques personnelles

J'ai également introduit des variables correspondant aux caractéristiques personnelles de chaque golfeur (taille, poids, âge, nationalité). Ces données se trouvent aussi sur le site du PGA Tour, que j'ai pu « scrapper » pour constituer un dataframe d'informations relatives à chaque golfeur. Certaines lignes n'ont que des valeurs manquantes : on les enlève.

a) Traitement de la nationalité

La colonne « Nationalité » doit être créée à partir de celle du lieu de naissance constituée pour chaque golfeur de la ville et du pays. Après avoir extrait le pays et constitué la variable « Nationalité », je me suis aperçu que pour certains golfeurs, ce n'était pas le pays mais une région/comté/ville (ou encore le pays mal orthographié) qui était renseigné.
J'ai donc créé un dictionnaire avec comme clés les bons noms de pays et les valeurs erronées en valeurs associées (exemple : "England" : ["Buckinghamshire","Staffordshire","Lancashire","London"]), puis j'ai créé une fonction de conversion, appliquée à la colonne de nationalité.

b) Traitement et conversion du poids et de la taille

En se concentrant sur l'occurrence des différentes valeurs (méthode « .value_counts() »), on se rend compte que 9 valeurs pour la taille sont mal renseignées (valeurs = « cm ») et 44 pour le poids (5 valeurs à « 0 kg », 39 à « kg »). Ces dernières sont pour la plupart disponibles en standard américain.
J'ai donc créé une fonction de conversion de taille du standard américain (feet, inches) en cm et une autre pour la conversion du poids de lbs à kg, puis effectué les conversions nécessaires.

Maintenant que les valeurs erronées et manquantes ont été traitées on transforme les variables 'HEIGHT' et 'WEIGHT' pour ne garder que les chiffres à l'aide d'une regex et convertir les données en entier (de même que pour la colonne 'AGE').

3) Mise au format parquet

Ces 2 dataframes furent mis au format parquet pour éviter de recharger à chaque fois les données, puis assemblés en effectuant un « merge » en considérant le nom des golfeurs comme clé (« left join » car le dataframe de données informationnelles regroupe plus de 600 golfeurs alors que le nombre de golfeurs différents dans le 1er dataframe est de 171).

Pour certains golfeurs, des informations sont manquantes, je préfère ne pas les effacer pour ne pas perdre trop d'observations.

III. Analyse exploratoire

Avant de développer des modèles de Machine Learning, on peut explorer et visualiser les données à travers plusieurs aspects.

1) Focus sur certaines corrélations

a) Physique et distance

On peut commencer par regarder si en moyenne la distance au driver est corrélée avec les caractéristiques personnelles des golfeurs (taille, poids, âge). On constate que la distance est négativement corrélée avec l'âge et positivement avec le poids et la taille (voir annexe, table 1 et figure 2). L'impact de la taille semble logique : plus on est grand plus l'amplitude du geste et l'inertie le seront, résultant en une distance plus importante. L'âge et le poids sont cependant plus difficiles à interpréter. On peut penser qu'il existe un âge (ou une tranche d'âge) optimal pour lequel un golfeur va être au maximum de ses capacités physiques. Concernant le poids : 2 golfeurs avec le même poids peuvent

avoir des rapports muscle/graisse bien différents, ce qui va joueur sur l'impact du club sur la balle ainsi que sur la vitesse de rotation du club. La corrélation semble donc plus complexe à justifier.

Il n'y a pour l'instant pas de lien de causalité, il faudra réaliser une régression linéaire pour cela (voir plus loin)

b) Distance vs précision

Les golfeurs perdent-ils en précision en privilégiant la puissance avec des coups de driver plus longs ? La réponse semble logiquement être oui au vu de la corrélation entre les 2 grandeurs (voir annexe, figure 3)

2) Caractérisation des vainqueurs

On peut regarder de manière individuelle quels ont été les points forts des vainqueurs de chaque édition par rapport à leurs concurrents.

Pour chaque année, on crée un dataframe en prenant les variables suivantes avec une condition sur l'année : « AVG DIST », « PRECISION (%) », « AVG AROUND », « AVG PUTTING » et « AVG APP » puis on les normalise entre 0 et 1 avec « MinMaxScaler() » de scikit-learn (pour avoir une échelle de comparaison). On calcule ensuite la moyenne pour chaque variable, ce qui va constituer notre « golfeur moyen ».

Enfin, avec le module Graph Object de plotly, on peut construire des « radar charts » où l'on peut comparer les performances du vainqueur dans chacune des variables mentionnées ci-dessus avec la moyenne des performances des autres golfeurs (voir annexe, figures 4 à 7).

Exemple : en 2020, Bryson DeChambeau a été meilleur partout par rapport à la moyenne (à égalité pour la précision au driver).

3) Résultats serrés ?

Précédemment, une modification des variables de rang a été effectuée en cas d'égalité. En effet, certains golfeurs doivent « partager » leur place si d'autres totalisent le même nombre de coups qu'eux sur les 4 tours.

J'ai décidé de m'intéresser au caractère « serré » d'un tournoi en analysant la dispersion du nombre total de coup, pour les 4 années, à l'aide d'un boxplot (voir annexe, figure 8).

On constate que les tournois des années 2017 et 2020 ont été plus serrés que les autres, la variable du nombre total de coups variant dans un range plus étroit.

4) Domination américaine

Le golf est un sport surreprésenté par les Etats Unis, le PGA Tour se jouant d'ailleurs principalement sur des parcours américains. De même, la Ryder Cup, compétition qui se joue 1 année sur 2 et oppose une équipe composée exclusivement de joueurs américains à une équipe composée de joueurs européens, met en lumière cette prépondérance américaine.

J'ai donc voulu voir si cette domination se retrouvait à l'US Open. En considérant un dataframe par année, j'ai donc calculé pour chaque édition la proportion d'Américains sur le tournoi, dans le top 10 et enfin dans le top 3, pour une représentation avec l'outil « bar » de Matplotlib (voir annexe, figure 9).
Au niveau de la participation cela ne fait pas de doute avec environ 40% de joueurs dont la nationalité est américaine sur les 4 années. Cette domination se retrouve également dans les résultats, plus précisément dans le pourcentage d'Américains faisant partie du Top 10 (entre 50% et 70% selon les années) et Top 3 (2 joueurs sur 3 étaient sur le podium en 2016, 2017 et 2020, les 3 premiers étaient américains en 2018).

On peut en tirer la conclusion suivante :

- Les Américains sont surreprésentés
- Les Américains sont plus performants que les golfeurs des autres nationalités (proportionnellement à la représentation des autres nationalités donc, en les considérant comme un tout, c'est-à-dire « USA vs Reste du monde »)

5) Importance du classement intermédiaire

Le golf est également un sport d'endurance et de psychologie où le mental joue énormément sur les performances, il faut être capable d'être constant sur 4 jours d'affilée si l'on veut gagner. On peut voir avec la figure 10 en annexe le rapport entre classement final et classement à mi-parcours (à la fin du 2ème tour). On constate qu'être mal classé à l'issue de la 2ème journée ne permet pas dans la majorité des cas d'accrocher un top 10. Cependant, on constate une concentration assez forte des points à l'extrémité sud-ouest du graphique : les joueurs les mieux classés arrivent en général à maintenir leur position, même si une dégringolade est envisageable, comme le montrent certaines observations.

IV. Machine Learning : algorithmes et interprétations

1) Clustering

a) Modèle et interprétation

On peut commencer par regarder si l'on retrouve des similarités de profils parmi les golfeurs en utilisant leurs statistiques de jeu. En implémentant l'algorithme des K-means, la « méthode du coude » nous informe que le nombre de clusters idéal se trouve entre 2 et 5, ce qui est assez imprécis. Je préfère donc compléter par la « silhouette method » qui indique un nombre de clusters optimal égal à 2.

En prenant 2 comme paramètre, on peut donc tirer une liste de labels (0 et 1) qui décrivent des styles de jeu. 71 golfeurs sont associés au style de jeu numéro 0, 204 au numéro 1.
Mais comment décrire les styles de jeu déduits de l'algorithme ? En regardant 5 observations pour chaque style on peut déjà se faire une idée (voir annexe, tables 2 et 3). En effet à première vue on peut supposer que la première catégorie (style de jeu numéro 0) a un style de jeu plus agressif avec une distance au drive plus élevée que celle des golfeurs de la seconde catégorie, au détriment d'une moins bonne précision.
On peut aller plus loin en établissant une table de statistiques descriptives des 2 groupes (voir annexe, tables 4 et 5). Cette table confirme la supposition ci-dessus avec une moyenne de la distance moyenne au drive sur un parcours de 283 m pour le groupe 0 pour une précision moyenne de 43%, la moyenne de la première variable étant de 276 m pour le groupe 1, pour une précision moyenne de 73,5%, beaucoup plus élevée (rappel : la distance moyenne au drive est une variable associée à chaque golfeur à l'US Open, sur une certaine année. La moyenne de cette variable est donc prise sur l'ensemble des golfeurs pour toutes les années). Le reste des variables semble moins définir les groupes

b) Limite

Malheureusement cette classification présente un défaut, les golfeurs peuvent avoir participé à plusieurs éditions du tournoi sur les 4 années et seront considérés comme des golfeurs différents chaque année, chaque ligne équivalant à une observation. Ainsi, un golfeur qui joue d'une certaine manière une année et d'une autre manière l'année suivante peut donc se retrouver dans les 2 catégories. Ces golfeurs sont au nombre de 36 (sur 171). De même, un golfeur dont le poids varie beaucoup d'une année à l'autre verra certainement une différence notable au niveau de la distance qu'il peut atteindre au drive (supposée comme supérieure s'il prend de la masse musculaire et vice versa). C'est le cas de Bryson De Chambeau qui a développé lors de l'année 2019-2020 sa masse musculaire d'une vingtaine de kilos. Sa distance moyenne au drive à l'US Open 2018 était de 283 m, celle de l'édition 2020 a été de 297 m.

Parallèlement, sa précision a drastiquement baissé, passant de 64,3% en 2018 à 41,1% en 2020 (voir annexe, table 6).
Une limite émanant de cet exemple est qu'ici le poids est constant, des données annuelles permettraient une meilleure précision.

2) Régressions

a) Distance au drive

Le golf est un sport de plus en plus physique où les golfeurs sont davantage à inclure des heures à la salle de musculation dans leur entraînement qu'avant. Comme nous l'avons vu lors de la section précédente la taille et l'âge sont également susceptibles de jouer un rôle.
La masse musculaire ne fait pas partie des variables disponibles, mais nous pouvons regarder l'impact des prédispositions physiques (taille et poids) et l'âge sur la distance moyenne des golfeurs au driver.
Après avoir enlevé les valeurs manquantes et considéré seulement les variables mentionnées lors de ce paragraphe, nous pouvons dresser un tableau de régression (voir annexe, table 7).
On retrouve l'interprétation des corrélations : la taille impacte positivement la distance de drive, contrairement à l'âge.
Néanmoins, il est indiqué qu'il y a un fort risque de multi-colinéarité entre nos variables. On répète donc l'expérience en enlevant le poids (voir annexe, table 8) et on obtient des résultats similaires (en termes de signe de coefficient).

b) Classement final

On peut également tenter de prédire le classement final des golfeurs avec une régression linéaire (voir annexe, table 9). Logiquement, une efficacité supérieure par rapport aux données historiques au putting, dans les coups d'approche et de parcours diminue le classement (golfeur mieux classé donc). Il s'avère que la précision au drive ainsi que le nombre moyen sur les par 3 ne sont pas significatifs.

PCA : au vu d'un risque élevé de multi-colinéarité on réduit la dimension avec l'ACP (Analyse en Composantes Principales), après avoir normalisé les données. Le R^2 obtenu est moins grand (moins de variables donc c'est logique), mais il n'y a plus de multi-colinéarité (voir annexe, table 10).

3) Classification

Entraîner un modèle de classification pour prévoir le classement des golfeurs n'est pas adapté : en cas de non ex-aequo, il y aurait autant de classements différents que de golfeurs.
Néanmoins on peut essayer de prédire si un golfeur va finir le tournoi dans le top 10 ou le top 3 en fonction des statistiques de golf (rangs non compris) ainsi que ses caractéristiques personnelles (taille, poids, âge).

a) Préparation

Je commence donc par enlever les valeurs manquantes et par ajouter 2 colonnes : « TOP 10 » et « TOP 3 » initialisées à 0 pour chaque ligne. A l'aide de conditions sur la colonne de rang final, la première colonne (respectivement la 2ème) prend 1 quand le golfeur a fini le tournoi dans les 10 premiers (respectivement dans les 3 premiers). On se trouve donc dans un cas classique de classification où la variable à prédire (en l'occurrence 2 variables ici) prend des valeurs dans {0,1}.

b) KNN et arbre de décision

On choisit par la suite de comparer le modèle des KNN avec un arbre aléatoire, les 2 étant disponibles avec scikit-learn. Pour le paramètre du nombre de voisins pour le KNN, on affiche l'erreur

engendrée dans les cas où l'on veut prédire la variable 'TOP 10' et 'TOP 3' (voir annexe, figure 11) et on choisit un nombre de voisins optimal à 5.
Après avoir séparé les données en jeu de données « train » et « test », on teste les 2 modèles pour la prédiction des 2 variables cibles et leurs performances sont évaluées (voir annexe, figures 12 et 13). Les modèles s'avèrent performants dans les 2 cas, avec un recall toutefois un peu moins élevé quand il s'agit de prédire les golfeurs sur le podium.

Les arbres de décisions sont disponibles sur le notebook mais je ne les ai pas insérés sur l'annexe par manque de visibilité.