Paris sportif - Beat the Bookmakers

Alexandre ROUET – Charles BOUDRY – David QUERIN – Marielle Odile GAHUNGU

Table des matières

[Introduction au projet 3](#_Toc143697021)

[Contexte 3](#_Toc143697022)

[Rapport d’exploration et de visualisation des données 4](#_Toc143697023)

[I - Exploration des données 4](#_Toc143697024)

[Compréhension des données 4](#_Toc143697025)

[Données – Valeurs manquantes 5](#_Toc143697026)

[Stratégies 6](#_Toc143697027)

[II – Pré-traitement des données et modèle 18](#_Toc143697028)

[Analyse multivariée 19](#_Toc143697029)

[Analyse de la distribution des valeurs 19](#_Toc143697030)

[Etape 1 – Nettoyer les données 21](#_Toc143697031)

[Etape 2 – discrétisation de la variable cible 21](#_Toc143697032)

[Etape 3 - Sélection des variables (feature selection) 21](#_Toc143697033)

[Etape 4 - Preprocessing 22](#_Toc143697034)

[Etape 5 – Identification du type de problème (discrimination, clustering…) et choisir un algorithme 22](#_Toc143697035)

[Etape 6 – Evaluer les performances de l’algorithme (Validité des structurés trouvées) 22](#_Toc143697036)

[III - Modèle 24](#_Toc143697037)

# Introduction au projet

## Contexte

Ce projet a pour objectif de proposer une rentabilité positive lors de paris sportif sur le tennis. Le jeu de donnée fourni l’intégralité des matchs recensés par l’ATP depuis 2001 est fourni, ainsi que les cotes initiales proposées par plusieurs bookmakers. Notre travail consistera à analyser les données et proposer des stratégies à rendement fiable.

Ce projet se déroulera donc en 2 temps, à savoir :

* Une étape d’exploration, de visualisation.
* Un pré-traitement des données et un modèle prédictif suggérant les cotes à jouer à une date t de telle manière que le retour sur investissement soir positif.

# Rapport d’exploration et de visualisation des données

L’objectif de cette partie est de comprendre les données de la base d’études et d’en évaluer la qualité. Pour cela nous nous sommes basés sur les techniques univariées, bivariées voir multivariées afin de garantir la qualité et l’adéquation du dataset qui servira à la modélisation de notre problématique.

En effet, la présence d’anomalies ou de valeurs manquantes peuvent influer sur le choix du modèle. Par exemple, la présence de valeurs extrêmes rend impossible l’application des méthodes d’analyse discriminante de Fisher ou de régression linéaire. Par ailleurs, si l’objet d’études est la réalisation des cartes de Kohonen, on pourra se passer du prétraitement des valeurs manquantes car elles les supportent parfaitement.

## I - Exploration des données

### Compréhension des données

Pour ce projet, nous disposons de 2 bases de données (ATP, Confidence).

**La base ATP** comporte 57420 matchs décrits selon 22 variables d’études :

* 6 portent sur les tournois dont 1 variables quantitatives et 5 variables catégorielles (ATP, Location, Tournament, Séries, Court, Surface)
* 9 portent sur les joueurs dont 3 variables quantitatives et 6 variables catégorielles (Round, Best of, Winner, Loser, WRank, LRank, Wsets, Lsets, Comment)
* 4 portent sur les bookmakers, toutes quantitatives : PSW, PWL, B365W, B365L
* 3 sur l’elo, score arbitraire attribué à chaque joueur : elo\_winner, elo\_loser, proba\_elo

On observe que les données sont extraites par script sur un site public qui est toujours actif et à jour.

Les données sont issues directement des résultats recensés par l’ATP, ce qui les rendent fiable, à l’exception des cotes des bookmakers et de l’elo.

L’elo n’est pas une information officielle mais une donnée calculée selon la base du nombre de parties gagnées et perdues.

**La base Confidence**, contient des données calculées suite à l’exécution d’un modèle prédictif sur la base ATP. Il s’agit d’une approche proposée par le détenteur du notebook sur Kaggle, et nous choisissons de l’écarter de notre analyse dans un premier temps pour nous concentrer sur les données initiales.

### Données – Valeurs manquantes

Une image contenant texte, menu, Police, document

Description générée automatiquement

Nous constatons la présence de valeurs manquantes.

Les valeurs manquantes sont réparties sur 7 variables dans les proportions suivantes :

Une image contenant texte, reçu

Description générée automatiquement

Les données manquantes sont donc principalement des matchs sans cotes. On observe également que les données manquantes sont dans les années 2001 à 2004 soit nos données les plus anciennes.

Pour l’heure, il serait prématurer de les supprimer, un match sans cote contient des informations de valeurs quant à la performance d’un joueur. La stratégie et le modèle choisi ultérieurement dicteront les opérations de prétraitement que nous effectuerons.

A noter également que notre base ne comporte pas de données dupliquées.

### Stratégies

Notre donnée cible dépend de la stratégie utilisée pour répondre à l’énoncé. Plusieurs hypothèses sont posées :

1. Mettre en concurrence les cotes de plusieurs bookmakers sur un même pari de manière à être positif quel que soit le résultat
2. Pour un bookmaker donné, mettre en place une stratégie pour ne parier que sur les matchs avec un écart de côtes très élevés
3. Evaluer les performances des joueurs pour déterminer la probabilité de victoire par joueur, et le corréler aux cotes des bookmakers pour assurer un retour sur investissement

#### Stratégie 1 : mise en concurrence des cotes entre bookmakers

Le but de cette approche est de déterminer si, pour un même pari, la cote des bookmakers s’oppose de telle manière que parier simultanément sur les deux garantiraient un bénéfice. On note que pour ce faire la cote des 2 paris doit être supérieure à 2 : En 22 ans, il n’existe que 22 matchs sur lesquels les bookmakers divergent avec une cote supérieure à 2. Le pari simultané chez les deux bookmakers garantit un gain à 100%.

On note cependant que les bookmakers prennent parfois la liberté d’annuler des paris pour « faute de cote ». Notre stratégie devient alors très risquée car elle implique de miser gros et seul le pari perdant risque de ne pas être annulé.

La faible fréquence de pari nous incite également à trouver une autre stratégie.

Outre ces 22 cas rares, nous observons une forte corrélation entre les cotes des bookmakers en notre possession B365W/PSW vs B365L/PSL (R²>90%) : on en déduit que les cotes des différents bookmakers sont très souvent similaires voire identiques.

Une image contenant texte, capture d’écran, Caractère coloré, carré

Description générée automatiquement

Une analyse temporelle, confirme cette tendance dans le temps :

Une image contenant texte, capture d’écran, nombre, Caractère coloré

Description générée automatiquement

Une image contenant texte, capture d’écran, Caractère coloré, nombre

Description générée automatiquement

De même, pour un même bookmaker, on observe moins de disparités pour les cotes des joueurs gagnants comparés à celles des perdants. En effet, les cotes des joueurs gagnants sont plus corrélées que les cotes perdantes. Le choix du bookmaker pourrait donc avoir son importance pour limiter les pertes des paris perdus.

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

**Nos conclusions sur l’analyse des données au bénéfice de l’hypothèse 1 sont donc les suivantes :**

* La stratégie 1 n’est pas viable dans le temps
* Les bookmakers ont des cotes similaires, et le choix de l’un plutôt que l’autre influencera peu un modèle prédictif
* Il reste néanmoins bénéficiable de comparer les cotes des différents bookmakers pour identifier celui qui en moyenne fournit les cotes les plus intéressantes pour le joueur. On note cependant que les données des bookmakers ne sont pas Français et que les bookmakers français ne publient pas leur cote historique.

#### Stratégie 2 : Ne parier que sur les matchs avec un écart de cotes élevés pour un bookmaker donné

Le but de cette stratégie serait de déterminer si, pour bookmaker donné, le fait de parier uniquement sur les matchs avec un écart de cotes élevés, permettrait de prédire au mieux la probabilité d’un joueur gagnant un match.

Avec cette stratégie, nous avons constaté que, sur les gains réalisés sur 100 parties en fonction de l’écart de cote entre deux joueurs, le parieur n’a que des pertes. On en conclut que cette stratégie n’est pas viable afin d’obtenir un gain convenable surtout en pariant toujours sur la cote la plus petite.

Une image contenant texte, capture d’écran, Tracé, ligne

Description générée automatiquement

#### Stratégie 3 : Approche basée sur la performance des joueurs pour déterminer le résultat des matchs

Cette stratégie est plus complexe et se divise en deux parties :

* Prédire en utilisant les données propres aux joueurs (classement mondiales, nombre de sets gagnés, nombre de points avant le tournois...) et les conditions du tournois (type de surface, de séries) le vainqueur du match.
* Vérifier que la cote associée à chaque pari réussi permet un retour sur investissement

En tenant en compte le fait que le tennis se dispute entre deux joueurs ayant des conditions de jeu identique, notre analyse se porte sur les axes suivants :

* Ratio de parties gagnées au nombre de matchs
* Influence du court (extérieur/intérieur)
* Influence de la surface (clay / hard / carpet or grass)
* Influence de la série (grand slam, masters, international)
* Influence de la localisation
* Différence de rang ATP entre les joueurs

On note aussi que cette méthodologie permet d’apporter un grand nombre d’informations supplémentaires qui ne sont pour l’heure pas présent dans le jeu de données :

* Le fait que le joueur soit droitier ou gaucher
* Influence de l’âge du joueur
* Grand écart de victoire ou non (2 set gagnés 0 perdu ?)

#### Influence des ratios de parties gagnées au nombre de matchs

Avant cela analysons les caractéristiques de notre jeu de donnée pouvant influer sur la performance de nos joueurs.

Une image contenant texte, capture d’écran, Police, diagramme

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, Tracé

Description générée automatiquement

Sur ces graphiques, nous pouvons voir les 20 meilleurs joueurs de notre jeu de données et ceux ayant le plus de défaites. Nous serons donc plus attentifs envers ces joueurs en raison de leur nombre de victoires. Les autres éléments du jeu de donnée nous donneront plus d’information si le match a lieu entre les meilleurs joueurs.

Une image contenant texte, capture d’écran, Police, Tracé

Description générée automatiquement

Sur ce graphique montrant le ratio de victoire nous pouvons voir que les 10 meilleurs joueurs détiennent un ratio supérieur à 0,6. Autrement dit leur constance au cours du temps est assez bonne. Nous remarquons aussi que plus le niveau de joueur diminue, plus la cote moyenne du joueur augmente. Cela signifie que la plupart des bookmakeurs suppose d’une victoire des meilleurs joueurs. En cas de victoire d’un joueur ayant un niveau moyen ou bas sur un meilleur joueur sera donc récompensé par un meilleur gain pour le risque pris grâce à cette cote élevée.

Une image contenant texte, capture d’écran, Police, Tracé

Description générée automatiquementUne image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

Le rendement des paris sportifs n’est pas lié au niveau des joueurs. En effet, nous pouvons voir que Novak Djokovic et Seppi A détiennent par exemple un meilleur rendement que Federer alors que nous avons un nombre de victoire de la part de Federer supérieur à Seppi. Cela rejoint donc le graphique précédent qui prouve que nos parieurs visant un gros gain préfèrent soit miser une petite somme sur des joueurs de niveau plus bas ou encore miser une grosse somme sur un joueur ayant un niveau plus élevé. Afin de maximiser leurs gains les parieurs visent donc différents types de joueurs mais surement avec un montant différent pour minimiser les pertes.

Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

Sur ce graphique le nombre de victoires par type de terrain fluctue en fonction du joueur. Nous remarquons que Nadal détient un score assez élevé de victoires sur un court à l’extérieur contrairement à Youzhny et Benneteau qui préfèrent les courts intérieurs. Cette information peut être utile lors de la confrontation de de certains joueurs afin de savoir qui pourrait gagner grâce à ces préférences.

##### Influence de la surface sur les joueurs

Sur le graphique ci-dessous nous pouvons nous rendre compte que la majorité des terrains de tennis sont des sols durs ou en argile. Nous pouvons voir que les terrains à moquette sont minoritaires et ne sont même plus utilisé depuis 2009.

Une image contenant texte, capture d’écran, diagramme, Rectangle

Description générée automatiquement

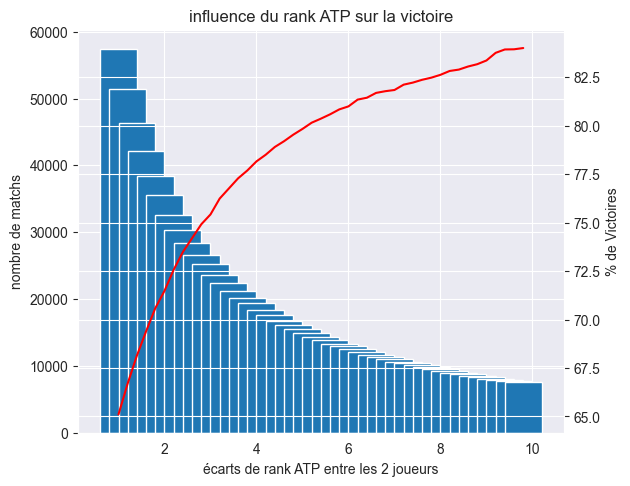
Si nous prenons en compte la surface de terrain sur le ratio de victoires des 20 meilleurs joueurs, nous pouvons voir que le type de terrain a une réelle influence sur les performances des joueurs. Les meilleurs joueurs ont de meilleures performances sur un terrain en argile comme Nadal et d’autres sur gazon comme Federer. Nous pouvons voir que les joueurs comme Federer Djokovic Nadal et Murray sont assez polyvalents et possèdent une réussite assez homogène quelque soit le type de terrain.

Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

##### Influence du Rang ATP

L’écart de rang ATP entre les joueurs augmente la probabilité de victoire. En effet plus cette différence de rang est criante et plus nous voyons une probabilité est élevée. Le rang ATP est donc un facteur important à ne pas négliger dans notre analyse de victoire d’un joueur.



##### Analyse des victoires par tournoi

L’analyse confirme que le type de tournoi influe sur la prédiction du gagnant

Une image contenant texte, capture d’écran, Police, Parallèle

Description générée automatiquement

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquement

Une image contenant texte, capture d’écran, ligne, diagramme

Description générée automatiquement

Sur ces trois graphiques nous pouvons voir les préférences des meilleurs joueurs en fonction des tournois par exemple Nadal a de nette facilité sur le tournoi français là ou Federer brille beaucoup plus à Open d’Australie et Wimbledon. De plus ces graphiques confirment les tendances découvertes lors de l’analyse de des ratios de victoires par types de terrain. Nous pouvons clairement voir la polyvalence de Djokovic qui détient des résultats assez homogènes en fonction du type de tournoi majeurs qu’il effectue. Cette variable nous sera donc utile afin de déterminer un gagnant lors d’une future confrontation. Nous notons que le nombre de victoires pour les tournois internationaux est plus faible pour la plupart des joueurs surement en raison de la rareté des tournois ou en raison de la pression lors de ces compétitions.

### Conclusion

La stratégie 3 : « Approche basée sur la performance des joueurs pour déterminer le résultat des matchs » est la plus sérieuse et la plus prometteuse.

L’analyse des données associées permettent de confirmer l’influence de plusieurs feature à notre disposition sur le résultat, ce qui est encourageant.

La suite de notre rapport présentera les modèles étudiés et leurs traitement des données associées pour optimiser au mieux cette stratégie.

## II – Pré-traitement des données et modèle

### Donnée cible

Il est clair que les variables Winner vs Loser nous permettront de définir la variable cible, et donc de prédire le gagnant d’un match.

Répartition des gagnants vs perdants

Une image contenant capture d’écran, texte, Tracé, ligne

Description générée automatiquementUne image contenant capture d’écran, texte, ligne, Tracé

Description générée automatiquement

Une analyse rapide de la réparation des gagnants vs losers permet de constater que l’on est en présence de jeu de données déséquilibrées, reparties globalement autour des joueurs les plus performants. Le choix de la méthode devra en prendre en compte

De plus, le but étant de prédire le gagnant d’un match, nous nous retrouvons dans la configuration de méthodes prédictives. Plusieurs méthodes sont éligibles selon le type des variables que l’on disposera dans le dataset d’entrainement :

|  |  |  |
| --- | --- | --- |
|  | Régression - Variable cible quantitative | Classification – Variable cible qualitative |
| Variables explicatives quantitatives | Régression,  Linéaire (lasso,) ou multiple, Régression par réseau de neurones | Analyse discriminantes (LDA, QDA, PLSDA) |
| Variables qualitatives explicatives | Analyse de la variante (ANOVA), MANOVA, SVR, K-plus proches voisins |  |
| Mélanges de variables Qualitatives et Quantitatives | Analyse de la covariance (ANCOVA),  K-plus proche voisins | Régression logistic, SVM |

Afin de choisir le type de modélisation adapté, nous allons procéder à l’analyse bivariée et multivariée, à l’aide de la visualisation des données. Cela consistera étudier les relations entre les variables à prédire (Winner vs Loser) et les variables explicatives mise à disposition. Cela permettra de réaliser en partie la feature selection et garder que les variables significatives. Pour ce faire, à l’aide de la visualisation des données, nous allons explorer les stratégies possibles afin de prédire la victoire d’un joueur lors d’un match.

## Analyse multivariée

Pour battre les algorithmes des bookmakers sur la probabilité d’un joueur gagnant un match, plusieurs axes d’études sont possibles. Nous en avons identifié 2 :

* Le premier axe d’étude consiste à prendre en compte des cotes des bookmakers lors des paris. Nous distinguons donc deux mailles d’études :
  + Comparer les cotes de chaque bookmaker sur un pari afin de s’assurer que nous obtiendrons un résultat concluant.
  + Ne parier que sur les matchs avec un écart de cotes élevés pour un bookmaker donné
* Le deuxième axe d’étude consiste à se baser sur les informations sur les performances des joueurs afin de déterminer le pourcentage de victoire possible par joueur sans prendre en compte les cotes de bookmakers.

### Analyse de la distribution des valeurs

#### Variables explicatives

En plus de la suppression des matchs n’ayant pas de côtes, après analyse de l’étendue (distribution) des variables ainsi que des statistiques descriptives, nous procéderons à une standardisation des données et éventuellement à une binarisation des variables catégorielles ayant un nombre pas très élevé de modalités :

* Standardisation des variables : les données n’ont pas la même échelle, nous procéderons dans la deuxième partie – Modélisation, à une standardisation des variables afin d’écarter l’influence des variables ayant une grande échelle ou des valeurs extrêmes

Une image contenant texte, blanc, Police, capture d’écran

Description générée automatiquement

Une image contenant capture d’écran, Tracé, diagramme, texte

Description générée automatiquement

* Binarisation de certaines variables catégorielles ayant un nombre pas très élevé sera peut-être nécessaires en fonction du modèle que l’on souhaitera entrainer notre jeu de donnée. Il s’agit de : Séries, Outdoor, surface, Round et Comment.

Une image contenant texte, diagramme, capture d’écran, cercle

Description générée automatiquementUne image contenant texte, Police, diagramme, logo

Description générée automatiquement

Une image contenant texte, diagramme, capture d’écran, cercle

Description générée automatiquementUne image contenant texte, diagramme, capture d’écran, cercle

Description générée automatiquement

Une image contenant cercle, capture d’écran, diagramme, Police

Description générée automatiquement

Et pour finir, nous constatons une saisonnalité des variables, une des pistes serait d’incorporer la notion temporelle lors de la modélisation afin de réduire le poids de la saisonnalité.

### Etape 1 – Nettoyer les données

xxxxx

### Etape 2 – discrétisation de la variable cible

Pour faciliter la prédiction, nous allons dans un premier temps transformer les variables Winner et Loser en variables numériques.

Choix du modèle ?????

### Etape 3 - Sélection des variables (feature selection)

### Etape 4 - Preprocessing

Lors du preprocessing :

* Sélectionner seulement les colonnes évoquées plus haut
* Séparer nos données en train/test
* Normaliser les variables numériques
* Encoder les variables catégorielles

### Etape 5 – Identification du type de problème (discrimination, clustering…) et choisir un algorithme

Par conséquent, étant donné, que l’on est en présence d’une variable cible numérique et un mélange de variables catégorielles et quantitatives, nous opterons pour les méthodes prédictives telles que

|  |  |
| --- | --- |
|  | Régression - Variable cible quantitative |
| Variables explicatives quantitatives | Régression,  Linéaire (lasso…) ou multiple, Régression par réseau de neurones |
| Mélanges de variables Qualitatives et Quantitatives | Analyse de la covariance (ANCOVA),  K-plus proche voisins  Modèles avancés : Arbre de décision, gradien boosting/bagging, random forest ??? |

* À quel type de problème de machine learning votre projet s’apparente-t-il ? (classification, régression, clustering, etc)
* À quelle tâche de machine learning votre projet s’apparente-t-il ? (Détection de fraude, reconnaissance faciale, analyse de sentiment, etc) ?
* Quelle est la métrique de performance principale utilisée pour comparer vos modèles ? Pourquoi celle-ci ?
* Avez-vous utilisé d’autres métriques de performances qualitative ou quantitative ? Si oui, détaillez-le.
* Entrainer les modèles aux choix
* Quels algorithmes avez-vous essayés ?
* Décrivez celui / ceux que vous avez retenu et pourquoi ?
* Avez-vous utilisé des techniques d’optimisation de paramètres de type Grid Search et Validation Croisée ?
* Avez-vous testé des modèles avancés ? Bagging, Boosting, Deep Learning…  Pourquoi ?

### Etape 6 – Evaluer les performances de l’algorithme (Validité des structurés trouvées)

* Bat-on les bookmaker
* Avez-vous analysé les erreurs de votre modèle ?
* Cela a-t-il contribué à son amélioration ? Si oui, décrivez.
* Avez-vous utilisé des techniques d’interprétabilité de type SHAP, LIME, Skater… (Grad-CAM pour le Deep Learning..)
* Qu’est ce qui a (ou non) engendré une amélioration significative de vos performances ?

## III - Modèle

Après cette analyse, des questions peuvent être soulevés. Pouvons-nous déterminer quand parier ou non.

Pouvons nous analyser un pari à partir du succès sur un nombre de matchs donnée.