DataScientest • com

Rapport Technique d'évaluation

Tennis Betting

Promotion : 02/2021

Participants :

Guillaume Rouja
Arthur David
William Romy

Objectifs

Nous sommes convaincus que les données disponibles en open source dans le milieu sportif ont un potentiel prédictif permettant de réaliser des pronostics avec un ROI positif et stable.

Nous faisons le pari qu'il est ainsi possible de mettre en œuvre une stratégie afin de déceler les meilleurs pronostics dans le milieu du tennis afin de dégager des gains constants.

La démarche de classification des résultats des parties en fonction de différentes variables et l'étude du ROI des paris sportifs en lien avec les côtes des bookmakers sont des outils facilement transposables a d'autres environnements et marchés.

Data

Cadre

Nos données sont hébergées sur GitHub :
https://raw.githubusercontent.com/GuillaumeRouja/2021-Tennis_Bet/main

Il existe des notebooks Kaggle traitant de sujets similaires :

https://www.kaggle.com/edouardthomas/atp-matches-dataset
https://www.kaggle.com/chkrdhr/tennis-player-data-analysis/data?select=Player.csv

Les données utilisées proviennent du site suivant :

http://tennis-data.co.uk/data.php

Ces données sont disponibles librement.

Organisation

3 Notebooks ont été conçus pour les 3 étapes du projet :

- Exploration : comprenant la création de certaines features apportant une valeur prédictive pour notre analyse.

- Modélisation : Tests de différents modèles et optimisation des modèles les plus prometteurs.

- ROI : Etude de différents scenarii et recherche des paramètres optimaux de la stratégie retenue.

Les notebooks sont rédigés en anglais et partagés sur la plateforme Google Colab.

Présentation du jeu de données

Le jeu de données retenu comprend environ 16,500 entrées correspondant à des statistiques de matchs masculins.

NB : Une modélisation des matchs féminins (WTA) a également été réalisé. Elle nécessitera une analyse complémentaire ultérieure pour ne pas alourdir le projet.

Ces 16,500 entrées correspondent aux rencontres principales de tennis sur les dernières années depuis le 1er janvier 2015. Nous avons jugé pertinent de conserver environ 6.5 années de données afin de réaliser le meilleur arbitrage entre :

- Le volume de données permettant d'assurer la fiabilité de notre modèle
- Des données suffisamment récentes
- La qualité et le volume des données manquantes

Nous avons extrait 18 variables du dataset initial. Nous avons ensuite créé et testé 26 variables supplémentaires, dont la majorité à partir des données existantes.
Nous n'avons ensuite conservé que les plus pertinentes à l'intérieur de nos notebooks d'exploration et de modélisation pour un total de 40 variables.
Parmi celles-ci, nous n'utilisons qu'une partie des plus prédictives dans les notebooks de modélisations et stratégie de ROI.

Pour le nettoyage des données, nous avons retiré les données nous semblant absurdes. (Exemples : Cotes < 1, classement ATP = 0)
Nous avons retenu les matchs complets, sans abandon, et nous sommes séparés de certaines variables sans intérêt prédictif telles que le numéro du tournoi.

Au regard de l'étude, le classement ATP semble être la variable a l'intérêt prédictif le plus fort. Nous avons également construit la variable 'Elo' qui permet de rendre compte de la forme des joueurs en attribuant un classement plus élevé aux joueurs ayant récemment remporté des

victoires contre des joueurs de rang supérieur. Enfin le taux de victoire, brut ou par surface a également joué un rôle important dans la construction de notre modèle.

La variable cible retenue, 'Won', est binaire. Elle vaut 1 si le joueur 1 a remporté la partie et 2 sinon.

Projet

Classification du problème

Il s'agit d'une étude prédictive (prédiction de victoire ou défaite) traitée comme un cas de classification binaire.

Dans le notebook de ROI, nous avons utilisé les probabilités de classement dans chacune des 2 catégories afin de mesurer notre modèle aux probabilités induites par les cotes des bookmakers.

Nous avons principalement retenu les métriques d'accuracy et F1 score afin de juger de la robustesse de notre modèle. Les résultats de ces 2 approches se sont avérés extrêmement proches dans la plupart des modèles.

Partie 1 : Enseignements tirés de l'exploration des données

Dans ce premier notebook, nous nettoyons et créons un certain nombre de features essentielles à nos modèles. Nous observons la distribution de nos variables ainsi que leurs corrélations au travers de différentes analyses graphiques supportées par les bibliothèques *Matplotlib* et *Seaborn*.

A ce stade, nous observons déjà des relations de corrélation importantes entre les cotes des bookmakers, les classements ATP et ELO, les % de victoires, et le nombre de matchs joués.

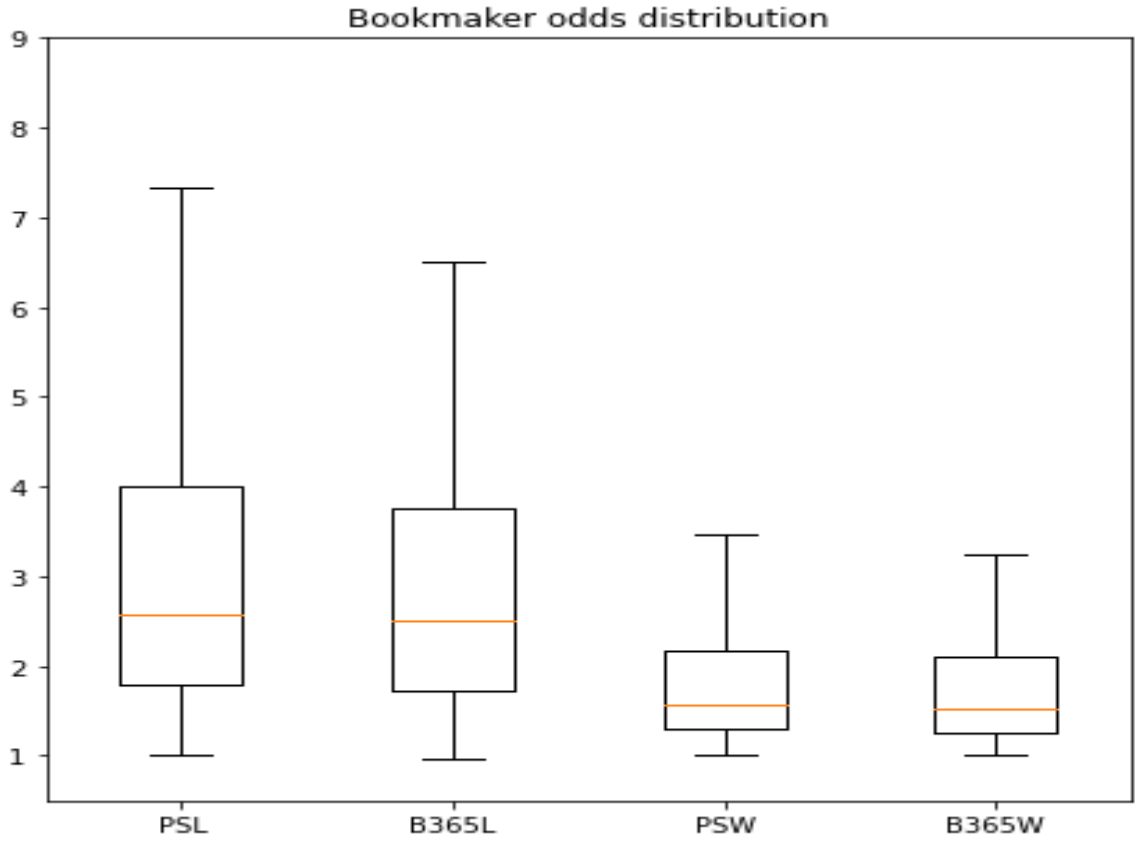
Nous observerons aussi des corrélations significatives avec la variable cible 'Won' dans le notebook de modélisation.

Que nous apprend la Data Visualisation ?

Nous pouvons tirer un certain nombre d'observations qui vont nous aider à sélectionner les features pertinentes à utiliser dans nos modèles :

- Concernant la distribution des variables, les matchs sont joués 'indoor' ou 'outdoor' et sur des surfaces dures (56%), de la terre battue (32%) et du gazon (11.9%) donnant globalement un avantage aux joueurs de surface rapide (dure + gazon).

- On note également que parmi les cotes des bookmakers dont nous disposons (Pinnacle et Bet365), les cotes de Pinnacle sont en moyenne plus généreuses de 0.1 chez les joueurs prédits vainqueurs et de 0.3 chez les joueurs prédits perdants.



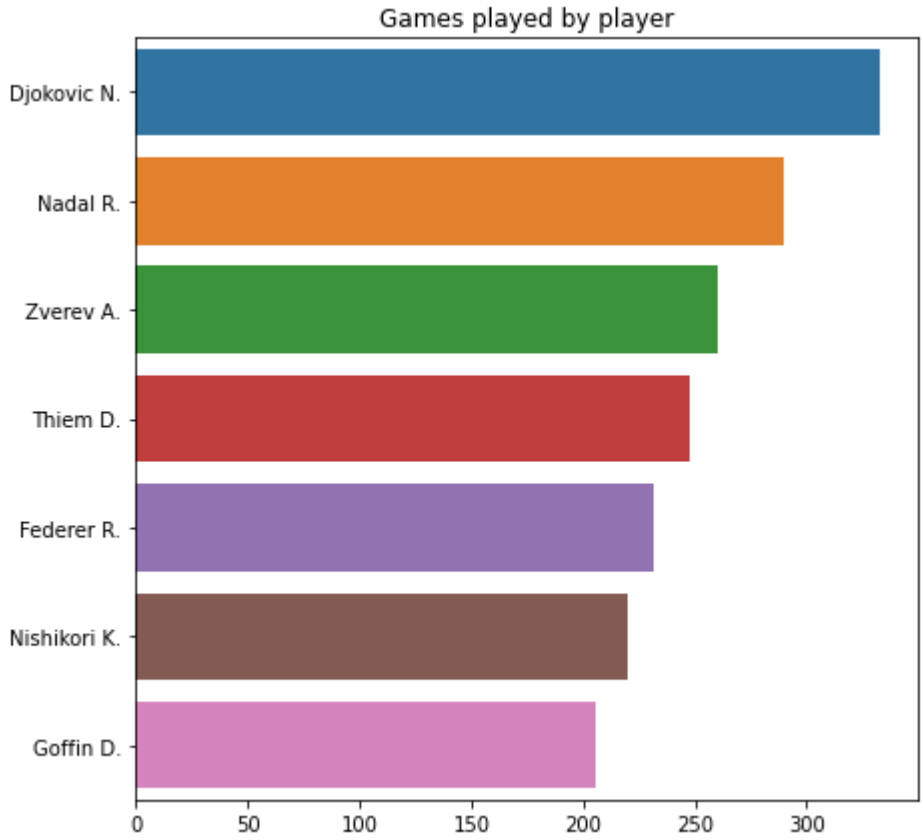
- La distribution de ces mêmes cotes est d'ailleurs variable en fonction des mois de l'année et surtout en fonction des tournois. Il est intéressant de noter que parmi les 5 tournois aux cotes moyennes les plus élevées, 4 sont des tournois du Grand Chelem !

  On peut imaginer que ces tournois populaires sont une occasion d'attirer des parieurs par le biais de cotes un peu plus intéressantes qu'à l'accoutumée. On gardera en mémoire qu'il sera possible d'affiner notre sélection de paris en privilégiant les tournois aux meilleures cotes à condition que cela ne réduise pas trop notre volume de paris.

- On notera que les cotes proposées sont négativement corrélées avec les performances des joueurs au regard de leurs classements ATP/ELO et de notre feature Winrate notamment, ce qui apparait de façon nette dans les boxplots organisés par plages de classement et Winrate.

- Nous notons enfin que la distribution des matchs joués est très inégale selon les joueurs, les 20 joueurs ayant joué le plus de matchs ont également joué presque 20% de l'ensemble des parties.

  Nous n'avons pas opté pour cette approche car cela reviendrait à nous séparer d'un nombre trop important de données : 50% des joueurs ont en effet joué moins de 10 parties et ¼ d'entre eux 1 ou 2 matchs seulement.



Partie 2 : Modélisation optimale

Ce notebook reprend les étapes d'exploration en les synthétisant de façon à avoir un notebook autonome.

Nous créons ensuite une distribution aléatoire entre joueur 1 et joueur 2 car les données initiales présentent des colonnes 'Winner' et 'Loser' et l'information du vainqueur du match ne doit pas être donnée au modèle lors de l'entrainement sous peine de créer un modèle biaisé. Il convient ensuite de rattacher les autres informations liées à chaque joueur.

L'étape suivante consiste à séparer et normaliser le jeu de données. Nous avons fait l'expérience que la normalisation via 'Quantile Transformer' présente des performances légèrement supérieures à celle par un 'Standard Scaler' par exemple.

Intuitivement, on peut considérer qu'il y a plus de différence de niveau entre les joueurs proches du top classement que ceux proches du bas.
Pour illustrer, on peut considérer que la différence de niveau entre le n.1 mondial et le n.100 mondial est plus importante que celle entre le n.300 et le n.400 mondial. A mesure qu'on se rapproche de la première place la concurrence est de plus en plus intense et la décomposition du Dataset en quantiles permet de transcrire ce phénomène de manière efficace.

Une fois cette étape réalisée, il nous faut sélectionner un ensemble de variables pertinentes. Nous avons procédé par tests et combinaisons entre les variables et étudié la matrice de

corrélation de Pearson et les performances d'accuracy et de F1 score entre les différents modèles. Il en ressort que les variables de classement Elo et ATP ainsi que les performances en termes de WinRate se révèlent les plus importantes.

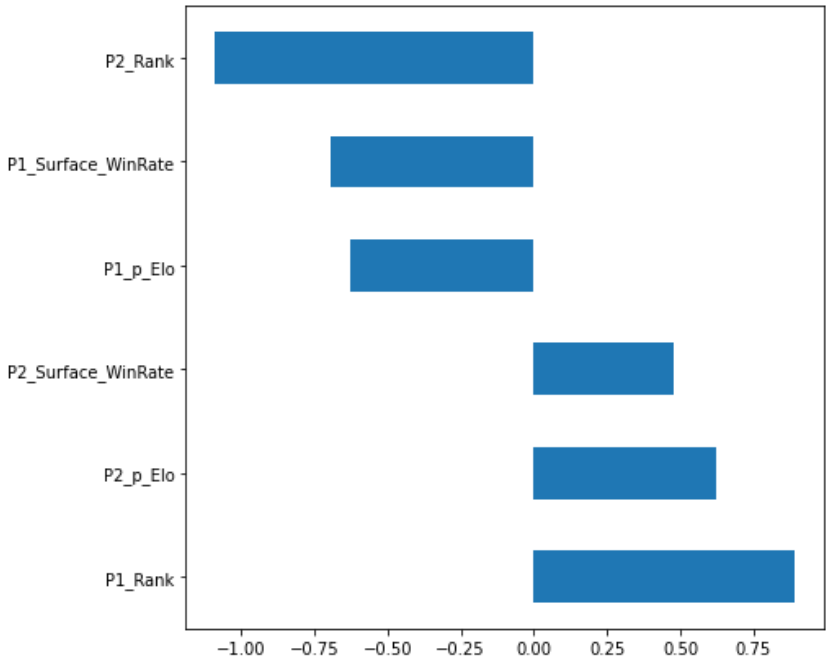
L'ajout de variables complémentaires ajoute du bruit et ne bénéficie pas au modèle de manière significative. De plus, avoir un modèle trop complexe peut conduire à du sur-apprentissage(overfitting) et rendre compliqué son alimentation en données dans la perspective de futurs paris.

Il en ressort que les modèles de Régression Logistique, SVM et Linear Discriminant Analysis se révèlent les plus performants avec des performances voisines pour les modèles simples proches de 67% d'accuracy et F1 score.

```
[28] from sklearn import linear_model

     # Model training

     clf=linear_model.LogisticRegression(random_state=10)
     clf.fit(X_train_scaled,y_train)

     # Model assessment

     y_pred = clf.predict(X_test)

     # Confusion Matrix

     cm = pd.crosstab(y_test, y_pred, rownames=['Classe réelle'], colnames=['Classe prédite'])
     print(cm)

     # Score

     print('\n''Score:',clf.score(X_test, y_test))

     #Classification report

     print('\n',classification_report(y_test, y_pred))
```

```
Classe prédite     1     2
Classe réelle
1               1019   544
2                523  1056

Score: 0.6604073838319542

              precision    recall  f1-score   support

           1       0.66      0.65      0.66      1563
           2       0.66      0.67      0.66      1579

    accuracy                           0.66      3142
   macro avg       0.66      0.66      0.66      3142
weighted avg       0.66      0.66      0.66      3142
```

```
[31] # Weights

     plt.figure(figsize=(7,7))

     pd.Series(clf.coef_[0], X_train.columns).sort_values(ascending=False).plot(kind='barh');
```



Liste des modèles testés :

- Logistic Regression
- SVM
- Decision Tree
- Random Forest
- Linear Discriminant analysis
- Quadratic Discriminant analysis
- Light Gradient Boosting Machine
- XG Boost

Une fois ces modèles sélectionnés, nous avons procédé à l'optimisation de leurs hyperparamètres et nous pouvons conclure qu'en terme de performance et de vitesse de calcul, le modèle de Régression Logistique se révèle la meilleure alternative, très proche du modèle de Linear Discriminant Analysis. Le modèle SVM se révèle quant a lui un peu plus précis mais avec des temps de processing nettement supérieurs.

Nous constatons également que le modèle n'atteint pas des prédictions basées sur les cotes des bookmakers, celles-ci se rapprochant de 70% soit environ 3 points au-dessus.

C'est pourquoi nous avons choisi de ne pas nous attarder sur une optimisation à outrance du modèle mais plutôt sur la réflexion de stratégie sur les paris à sélectionner afin de dégager un ROI consistant ce qui est l'objet de la partie suivante.

Partie 3 : Recherche d'une stratégie de ROI et de son optimisation

Ce notebook reprend les étapes d'exploration et de modélisation en les synthétisant de façon à avoir un notebook autonome.

Nous allons à présent étudier différents scenarii et leur impact sur le ROI :

SCENARIO 1 : 'On parie sur les cotes les plus élevées'.

On utilisera les cotes du Bookmaker Pinnacle qui se montre en moyenne plus généreux que BET365 (cf. étude exploratoire).
Ce scenario ne met pas en jeu de Machine Learning, il servira de base de comparaison pour les autres cas. Le ROI escompté est d'environ -5% en suivant cette stratégie.

SCENARIO 2 : 'On parie sur l'ensemble des matchs en fonction des probabilités calculées par notre modèle'.

Il s'agit d'utiliser un modèle de régression linéaire afin de prédire qui du joueur A ou du joueur B remportera la partie. On commence par sauvegarder les cotes et concaténer les résultats des matchs. On recrée ensuite un dataframe contenant nos prédictions, les matchs ou nous parions et les gains escomptés.

Les performances de ce modèle sont légèrement négatives, avec un ROI autour de - 2%. On pouvait s'attendre à un résultat de cet ordre avec un algorithme légèrement moins performant que les prédictions des bookmakers. Il s'agit à présent d'imaginer une stratégie de sélectivité des matchs.

SCENARIO 3 : 'Construction d'un seuil de sélection en fonction de nos prédictions vs les prédictions des bookmakers'.

Dans cette partie nous allons faire l'hypothèse que l'on peut dissocier les variables en 2 groupes. D'une part des variables 'statiques' qui correspondent à des classements sur une longue plage de temps par exemple (typiquement le classement ATP). D'autre part des variables 'dynamiques' qui recouvrent des facteurs de forme pouvant varier plus rapidement en fonction d'évènements récents.

Nous considérons que les modèles de prédictions élaborés par les bookmakers tiennent compte d'une combinaison de variable statiques et dynamiques.
Nous allons au contraire construire un modèle ne contenant que des variables statiques. Par itération, le choix de la variable ATP se révèle le meilleur candidat pour représenter la partie statique mais il serait possible d'étudier la combinaison avec d'autres variables. (Nous avons étudié la combinaison avec un certain nombre d'entre elles, par exemple le 'WinRate' sur une plage d'1 an mais parmi les features étudiées, le classement ATP seul se révèle plus efficient pour la stratégie qui suit).

Il s'agit ensuite de recréer un dataframe comprenant les résultats des matchs, les cotes du bookmaker Pinnacle, nos prédictions ainsi que celles de Pinnacle que nous déduisons en fonction des cotes proposées.

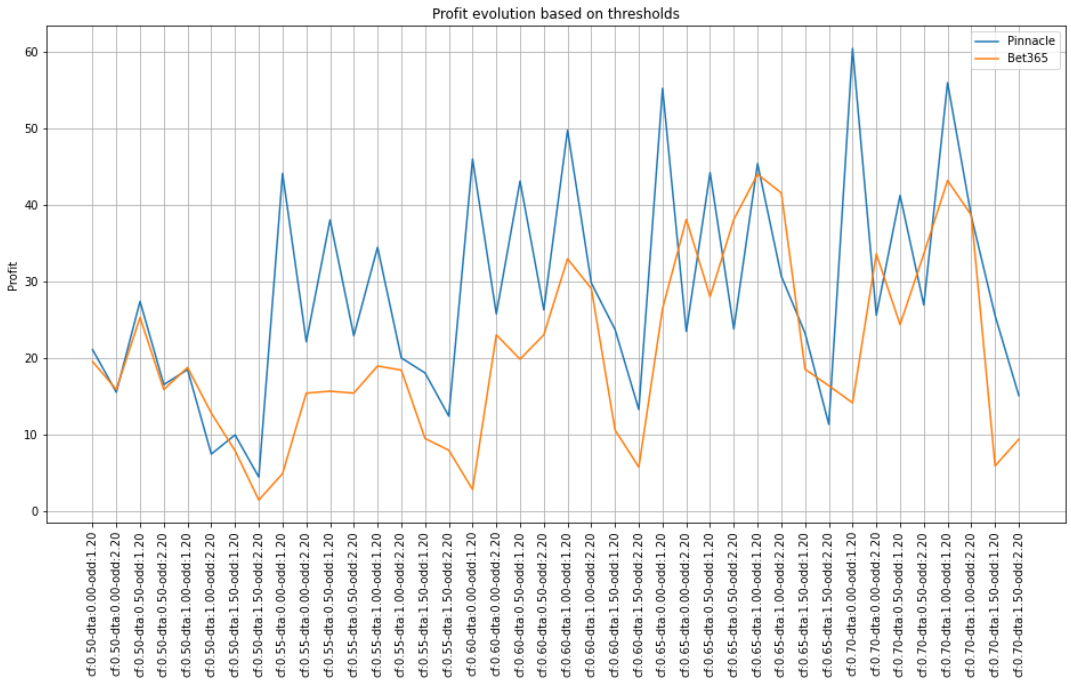
NB : A ce stade nous faisons un certain nombre d'hypothèses de travail :

- L'inverse de la cote du Bookmaker est une approximation de la probabilité de victoire du joueur. Nous admettons ce principe en dépit du fait que la cote contienne d'autres éléments telle que la marge du bookmaker ou son adaptation en fonction du volume de paris sur un joueur.
  Afin de limiter ce dernier effet, il faudra, en conditions réelles et dans la mesure du possible, parier au plus près de la disponibilité des paris sur la plateforme.

- Il existe d'autres paramètres qui, seuls, ne dégagent pas de ROI positif mais permettent de renforcer et affiner la stratégie de sélection. Dans le notebook dédié au ROI, nous faisons ainsi des hypothèses sur les cotes minimales et maximales de chaque joueur à considérer. Nous regardons également la valeur absolue de la différence des cotes des 2 joueurs avec l'idée que des cotes resserrées indiquent une incertitude quant au résultat d'un match donné.

Une fois ces principes posés, il s'agit donc de construire le seuil de décision (ratio des résultats de notre modèle / résultats du modèle du bookmaker pour chaque joueur) et ensuite d'évaluer le seul de décision le plus efficace pour réaliser un pari.

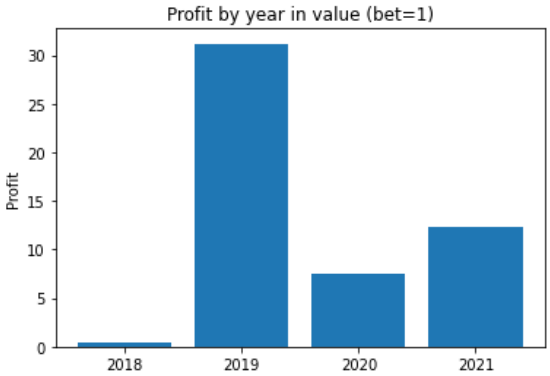
Pour cette étape, nous utiliserons une boucle afin de tester les différents niveaux de profits en fonction des seuils retenus.



Il semble qu'un seuil d'environ 65% se révèle un bon compromis entre une bonne sélectivité et un nombre de matchs suffisants pour une stabilité correcte du ROI. A ce seuil, on adjoint d'autres critères de sélectivité tels que la différence entre les cotes et les valeurs des cotes min et max. Notre boucle nous permet de tester ces valeurs également en complément du critère de sélectivité.

Le seuil de 65% s'interprète donc comme ceci : **notre modèle basé sur des variables statiques doit se dissocier suffisamment (représenter moins de 65%) du modèle du bookmaker pour que des variables dynamiques aient influé sur le taux de prédiction associé au joueur de manière significative, conduisant à une opportunité de pari intéressante.**

```
Overall Profit: 51.43

ROI: 12.0%
```



NB : à ce stade il est possible que la prédiction de victoire soit même inférieure à 50% mais cela n'est pas le plus important étant donné que l'on considère également la cote du bookmaker (ex : une cote de 3 avec 45% de chance de victoire est un bon pari).

Cette stratégie provient d'une itération manuelle mais semble confirmée par un test sur plusieurs seuils et renforcée par différents tests et calculs de ROI tous positifs et relativement stables (analyse mensuelle) sur les dernières années. Elle a l'avantage, en principe, de s'améliorer au fur et à mesure des progrès prédictifs du bookmaker (renforçant le delta entre variables statiques et dynamiques).

Toutefois une stratégie qui fonctionne à l'instant T et repose sur des données passées ne peut préjuger de l'avenir et se dérèglera probablement au fil du temps. Il faudra à minima considérer une mise à jour/maintenance de la stratégie au fil de l'eau.

Description des travaux réalisés

Répartition de l'effort sur la durée et dans l'équipe

La coordination du projet (construction des notebooks, organisation des réunions, rédaction du rapport) a été effectuée par Guillaume qui a travaillé sur tous les aspects du projet en continu sur la plage de temps impartie et élaboré la stratégie retenue.

Will a contribué à la partie liée à la Data Exploration et visualisation.

Arthur a contribué aux parties de Data exploration, data visualisation (co-construction des variables cumulées avec Guillaume, modèle et recherche de ROI (justification de la stratégie au moyen de la visualisation des intervalles de confiance).

Bibliographie

- Etudes de différents cas Kaggle et GitHub réalisés sur le sujet :

https://www.kaggle.com/edouardthomas/atp-matches-dataset
https://www.kaggle.com/chkrdhr/tennis-player-data-analysis/data?select=Player.csv

- Divers sites internet autour de l'univers des paris sportifs

Difficultés rencontrées lors du projet

- Nous avons consacré du temps à bien poser le problème et avons rencontré des difficultés dans la recherche de variables pertinentes et la construction du Dataframe joueur.

- Il fallait ensuite démontrer de façon lisible et concrète les résultats obtenus au moyen des cotes de confiance à l'aide d'une boucle qui nous a demandé des efforts de modélisation.

- Un temps considérable a été utilisé dans l'élaboration d'une stratégie combinant un modèle avec un overfitting faible et une robustesse du ROI sur un plage de temps de plusieurs années.

Bilan & Suite du projet

Ce projet nous a permis de renforcer notre pratique de la gestion de projet, définition de problématiques, modélisation Python et recherche de ROI.

Nous allons poursuivre l'étude du sujet en affinant la recherche des seuils de sélections, en renforçant l'étude sur les données de tennis féminin et en testant l'apport de mises non uniformes en s'appuyant sur des intervalles de seuils adaptés.