Master 2 Informatique STL
2017-18 - DAR

RAPPORT DE PROJET

Karim Berkane, Richard Bunel, Patrick Chen

Sommaire:

1) Problématique business

Pour ce projet, nous imaginons la problématique business suivante :

- Nous sommes une entreprise qui voulons organiser des paris sur les matchs de tennis.
- Plutôt que de simplement miser sur le vainqueur, on se propose de parier sur les statistiques des matchs, par exemple le nombre d'aces, de coups gagnants, de fautes directes …
- Pour être rentable, nous avons besoin de déterminer les côtes les plus appropriées et pour cela, d'estimer le plus précisément possible, avant un match, la valeur de ces statistiques.
- La solution va être de tenter de trouver une corrélation entre certains chiffres connus avant le match, et la valeur des statistiques que nous cherchons.

2) Dataset

Le dataset que nous utilisons provient du site Kaggle, est au format CSV et se nomme ATP.csv.
Il est disponible à l'adresse suivante : https://www.kaggle.com/sijovm/atpdata

Il contient, pour l'ensemble des matchs de tennis professionnels du circuit masculin joués entre 1968 et 2017, de nombreuses données telles que :

- Les joueurs dont leur nom, leur classement, leur nationalité, leur taille, leur âge, leur main dominante etc.
- La date du match, son emplacement, le nom du tournoi, la surface de jeu etc.
- Le score, la durée du match, le nombre d'aces, de coups gagnants, de fautes directes de chaque joueur etc.

3) Chargement des données

Notre dataset étant d'une taille relativement modeste, 31 Mo pour environ 164 000 matchs couverts, nous n'avons pas été dans la nécessité d'en réduire la taille.

Les opérations que nous effectuerons dessus s'exécutent ainsi dans un temps tout à fait raisonnable.

4) Prétraitement des données

Pour ce projet, nous avons décidé de nous focaliser sur la prédiction d'une seule variable: le nombre d'aces durant un match.

Comme nous le verrons dans la partie suivante, les variables que nous retenons pour tenter de prédire le nombre d'aces sont :

- la taille du joueur
- sa moyenne d'aces par match (ainsi que par match par surface)
- son classement

Ainsi, dans nos différentes analyses, nous avons décidé de supprimer les lignes où une des ces colonnes serait nulle.
On remarque que c'est le cas pour beaucoup de vieux matchs, ce qui n'est pas nécessairement un problème, en effet le tennis étant un sport ayant beaucoup évolué ces dernières années, on peut supposer que l'analyse de données récentes donnera des résultats plus fiables.

5) Analyse prédictive des données

La variable que nous tentons donc de prédire est le nombre d'aces.
Nous avons pour cela défini 7 modèles de prédiction qui utilisent des combinaisons de variables différentes, avec pour chacun les pourcentages d'échantillon du TrainSet et du TestSet:

A) la moyenne du nombre d'aces (80% / 20%)

B) la moyenne du nombre d'aces, la taille du joueur et celle de son adversaire (80% / 20%)

C) la moyenne du nombre d'aces, la taille du joueur et celle de son adversaire, le classement du joueur et celui de son adversaire (80% / 20%)

D) la moyenne du nombre d'aces, la taille du joueur et celle de son adversaire, le classement du joueur et celui de son adversaire (90% / 10%)

E) la moyenne du nombre d'aces, le classement du joueur et celui de son adversaire (80% / 20%)

F) la moyenne du nombre d'aces par surface, la taille du joueur et celle de son adversaire, le classement du joueur et celui de son adversaire (90% / 10%)

G) la moyenne du nombre d'aces par surface, la taille du joueur et celle de son adversaire, le classement du joueur et celui de son adversaire (80% / 20%)

Pour chacun de ces modèles, nous procédons de la manière suivante:

- Comme vu plus haut, on commence par enlever les lignes dont les données seraient manquantes.
- On divise les données en un jeu d'entraînement (TrainSet) et un jeu de test (TestSet) selon les pourcentages décrits ci-dessus.
- On entraîne le modèle sur le jeu d'entraînement.
- Puis on tente de prédire la donnée voulue sur le jeu de test.
- Enfin on compare notre analyse prédictive avec la valeur réelle de la variable pour déterminer la fiabilité du modèle (voir chap. 7 pour notre exemple).

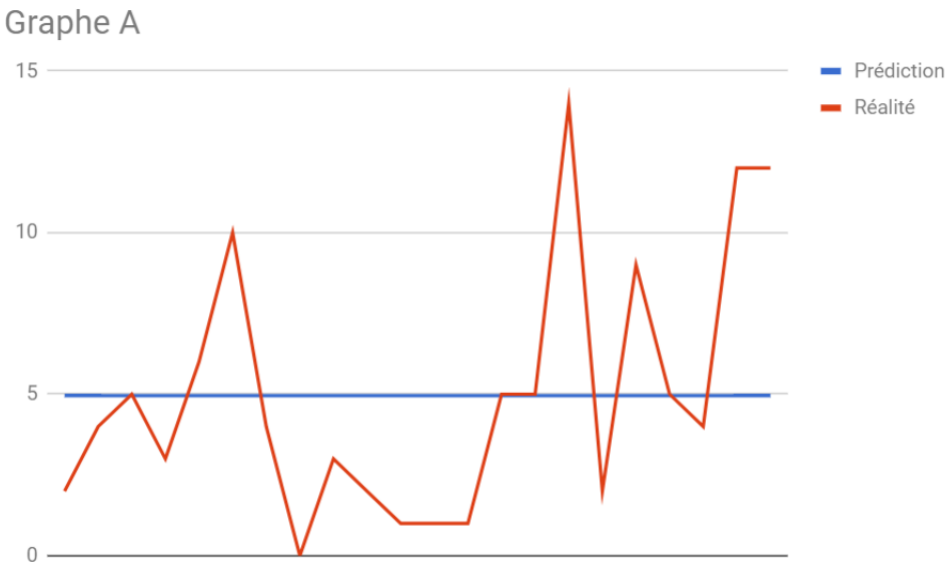
6) Exportation des résultats

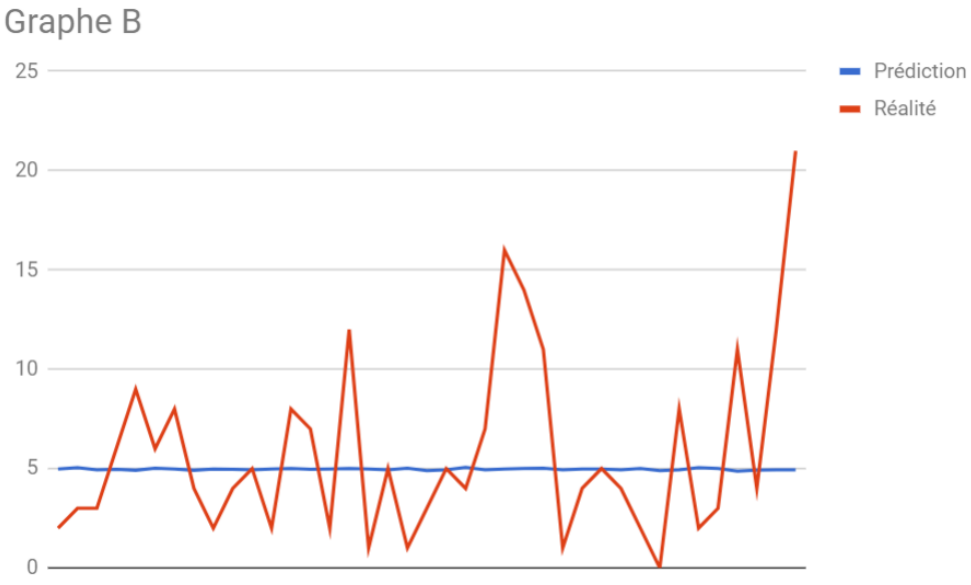
Nous exportons nos résultats dans le même format que celui d'entrée, c'est à dire le format CSV. Dans nos fichiers de sortie (un fichier par modèle), nous affichons dans la première colonne le résultat obtenu par notre analyse prédictive, dans la seconde colonne le véritable nombre d'aces marqués par le joueur, et ensuite les autres colonnes que nous avions jugées pertinentes dans le chapitre 4 (même si ces colonnes ne sont pas forcément utilisées par le modèle en question).

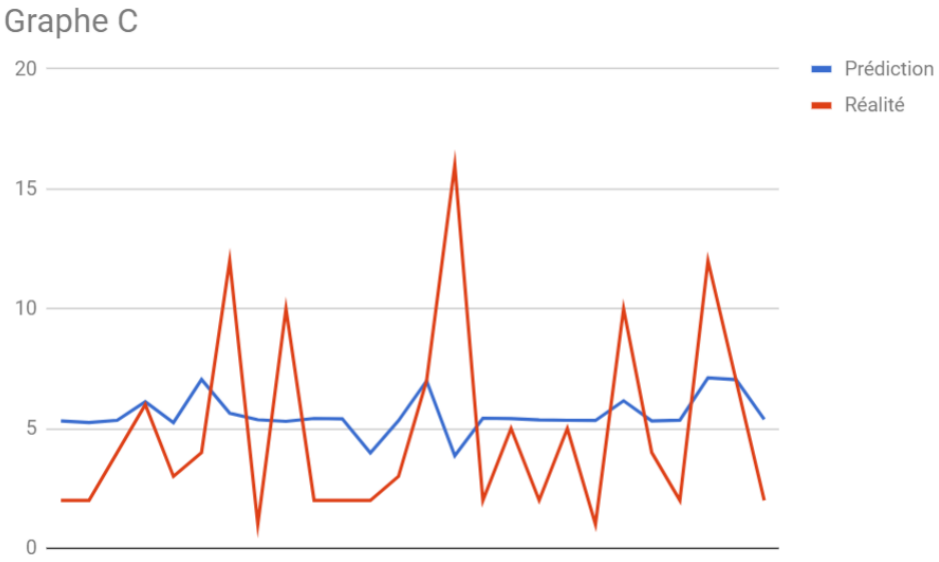
7) Visualisation des résultats

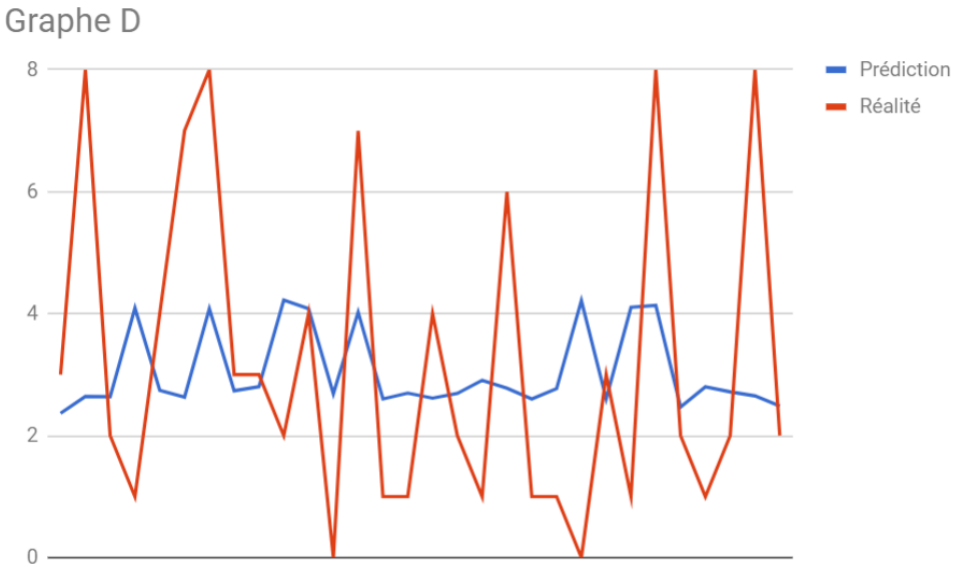
Pour chacun des 7 modèles numérotés de A à G ci-dessus, nous avons pris les résultats d'un joueur et comparé la valeur prédite du nombre d'aces avec la valeur effective.

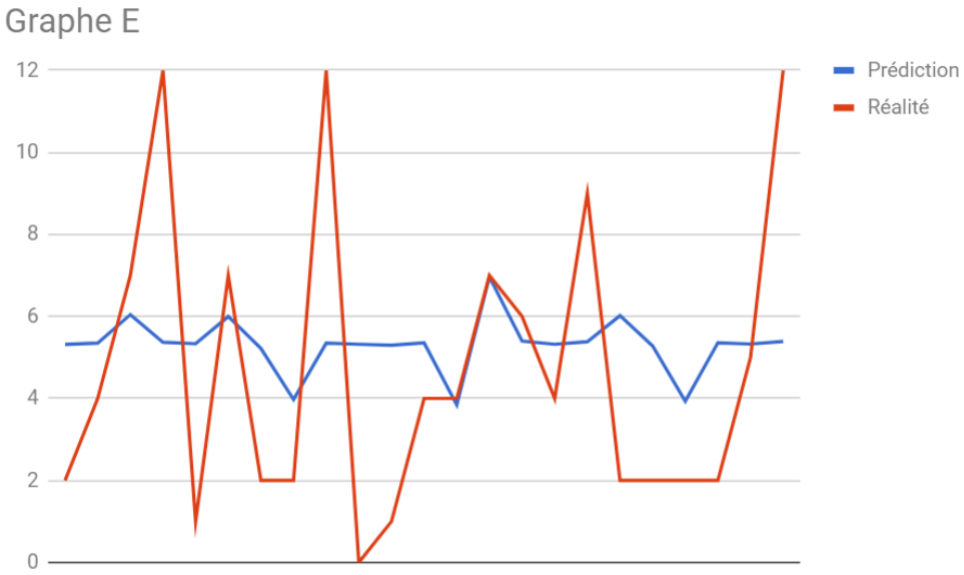
Ces résultats sont visibles dans les graphiques ci-dessous. La lettre correspond à celle affectée au modèle dans le chapitre 5. On note que plus les deux courbes sont proches l'une de l'autre, plus le modèle peut être considéré comme efficace.

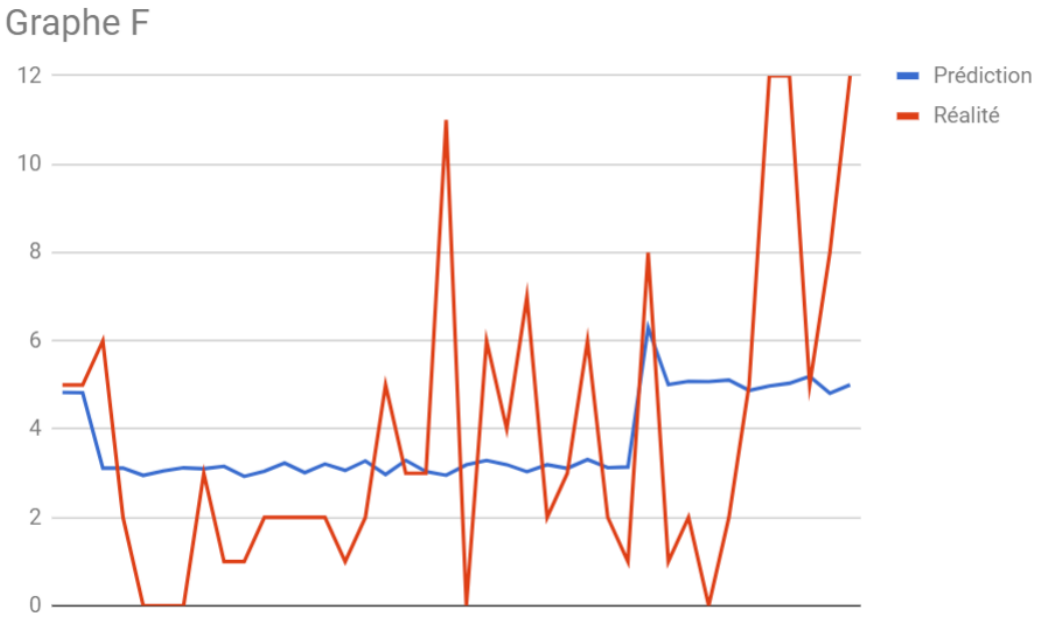


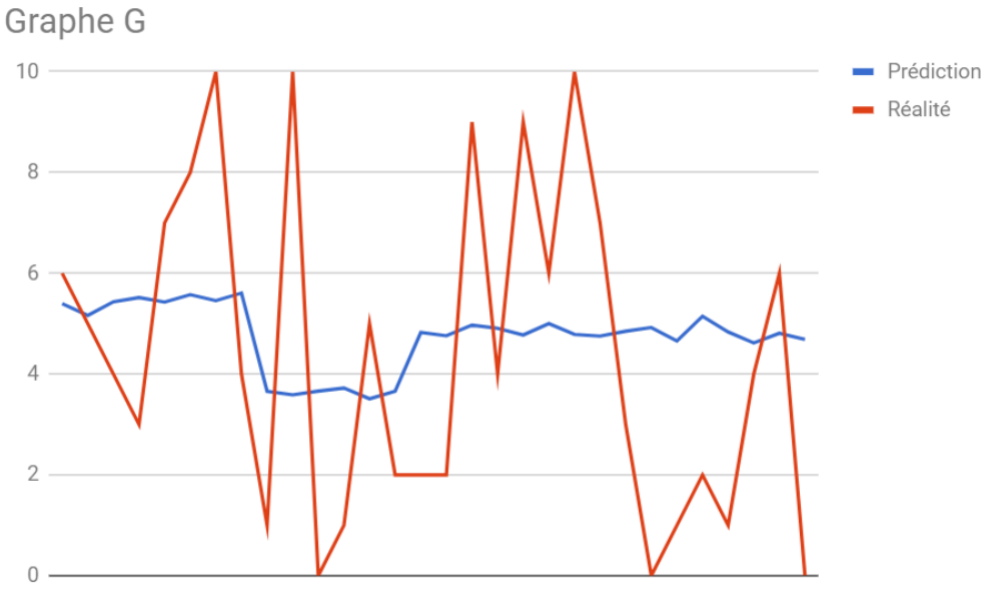

Graphe B
25
20
15
10
5
0
Prédiction
Réalité


Graphe C
20
15
10
5
0
Prédiction
Réalité


Graphe D
8
6
4
2
0
Prédiction
Réalité


Graphe E
12
10
8
6
4
2
0
Prédiction
Réalité


Graphe F
12
10
8
6
4
2
0
Prédiction
Réalité


Graphe G
10
8
6
4
2
0
Prédiction
Réalité

8) Conclusion

En conclusion de ce projet, nous pouvons dire les choses suivantes :

- Les courbes des valeurs prédites et réelles ne se confondent pas très bien. À cela plusieurs raisons possibles:
    - Les modèles choisis ne sont pas pertinents, c'est à dire que la corrélation entre les colonnes choisies et la colonne prédite n'est peut-être pas aussi importante que prévue. De plus, il faut garder à l'esprit que corrélation n'est pas synonyme de causalité.
    - Les échantillons utilisés pour l'entraînement sont trop petits. Ainsi ils ne permettent pas d'obtenir des résultats suffisamment fiables.
    - Le nombre d'aces dépend sans doute de beaucoup plus de facteurs que ceux disponibles dans le dataset, par exemple la météo, l'état de forme du joueur, le type de balles utilisées pour le match … Ces facteurs manquants pourraient expliquer le manque de précision des résultats.
- Malgré cela, la méthodologie expérimentée ici est intéressante et permet de facilement savoir si un modèle est pertinent ou non. Ainsi, dans le cadre de notre business, il faudrait continuer à expérimenter de tels modèles jusqu'à en trouver un suffisamment fiable pour établir les côtes de nos paris.
- D'autres options consisteraient à choisir une autre variable à prédire, de sélectionner un dataset beaucoup plus grand ou qui contient davantage d'informations sur chaque match.