Projet Machine Learning Python

Prédiction du MVP de la saison NBA 2021/2022

Rapport

ENSAE

IP PARIS

CANEILLES Charles – FOSSEY Oscar

MS Data Science

31 janvier 2022

1 TABLE DES MATIERES

2 TABLES DES FIGURES

3 INTRODUCTION

Dans le cadre du cours *Machine Learning with Python* nous devons réaliser un projet en lien avec les notions et méthodes abordées lors des séances en classe.

Ainsi, pour ce projet, nous nous sommes penchés sur la NBA, multinationale créée en 1947, et ligue de basket-ball américaine réputée partout dans le monde pour ses stars et son jeu spectaculaire.

Depuis 1956, et ceux lors de chaque saison, est désigné (par un système de vote, nous l'expliquerons plus tard) le joueur le plus *Valuable* (joueur ayant le plus d'impact dans le jeu, ou encore « meilleur joueur ») de la saison ; le MVP (*Most Valuable Player*).

Aujourd'hui nous sommes au beau milieu de la saison 2021-2022 et la course pour le trophée de MVP fait rage. Même si une tendance commence à se dégager et que certains prétendants se détachent du lot que constituent les 400 joueurs de la « grande ligue » ; nous allons, en nous plongeant dans le passé, essayer de prédire les potentiels prétendants et ensuite le vainqueur du trophée de MVP 2021-2022.

Figure 1 : Kobe Bryant (1978-2020) - MVP de la saison 2007-08

4 OBJECTIF DE L'ETUDE

4.1 Objectif

Notre objectif est de construire un estimateur pour le MVP de la NBA pour une saison en cours ou fini. Pour cela nous allons construire notre modèle à l'aide des statistiques de NBA passées et des résultats passés de l'élection du MVP en NBA.

4.2 Comment se déroule l'élection du MVP en NBA ?

Depuis la saison 1982-1983 : les votants sont issus d'un panel de journalistes sportifs et de commentateurs américains et canadiens. Chacun nomme cinq joueurs, classés par ordre de préférence. Une première place vaut dix points ; une deuxième en vaut sept ; une troisième en vaut cinq ; une quatrième en vaut trois et une cinquième en vaut un seul. Depuis 2010, un bulletin représentant le vote des internautes est ajouté au scrutin. Le joueur obtenant le plus grand nombre de points remporte le titre.

VOTING RESULTS: 2020-21 KIA NBA MOST VALUABLE PLAYER AWARD						
Player (Team)	**1st Place Votes (10 Points)**	**2nd Place Votes (7 Points)**	**3rd Place Votes (5 Points)**	**4th Place Votes (3 Points)**	**5th Place Votes (1 Point)**	**Total Points**
Nikola Jokić (Denver)	91	8	1	0	0	971
Joel Embiid (Philadelphia)	1	62	23	8	3	586
Stephen Curry (Golden State)	5	23	32	23	13	453
Giannis Antetokounmpo (Milwaukee)	1	5	34	41	10	348
Chris Paul (Phoenix)	2	2	8	13	26	139
Luka Dončić (Dallas)	0	1	0	7	14	42
Damian Lillard (Portland)	0	0	1	5	18	38
Julius Randle (New York)	0	0	1	2	9	20
Derrick Rose (New York)	1	0	0	0	0	10
Rudy Gobert (Utah)	0	0	0	1	5	8
Russell Westbrook (Washington)	0	0	1	0	0	5
Ben Simmons (Philadelphia)	0	0	0	1	0	3
James Harden (Brooklyn)	0	0	0	0	1	1
LeBron James (Los Angeles Lakers)	0	0	0	0	1	1
Kawhi Leonard (LA Clippers)	0	0	0	0	1	1

Figure 2 : Résultat du vote pour la saison 2020-21

Le joueur avec le plus de points est alors naturellement élu MVP. A noté que ce trophée est décerné au joueur ayant fait les meilleures performances durant la saison régulière et non pas les Playoffs qui ne sont pas pris en compte. Les Playoffs sont les phases finales de la saison sous forme de tournois à élimination directe.

5 SCRAPPING ET CONSTRUCTION DU DATASET

5.1 Sources de données

L'intégralité des données viennent de Sport Reference :

> ***"SRL believes in data democratization—open access to data and the tools that use it."***
> *https://www.sports-reference.com/termsofuse.html*

Datasets utilisés	
Contenu des datasets	*Référence*
Statistiques classiques des joueurs par match	https://www.basketball-reference.com/leagues/NBA_2021_per_game.html
Statistiques avancées des joueurs par match	https://www.basketball-reference.com/leagues/NBA_2021_advanced.html
Historique des vainqueurs du trophée de MVP	https://www.basketball-reference.com/awards/mvp.html
Historique des votes pour le trophée de MVP	https://www.basketball-reference.com/awards/awards_2021.html#mvp
Statistiques collectives (Victoires d'équipe, etc …)	https://www.basketball-reference.com/leagues/NBA_2021_standings.html

5.2 Structure du dataset

Une ligne du dataset correspond à la performance d'un joueur lors d'une saison donnée. Ainsi la « clé primaire » du dataset est le couple « Player – Season ». Avec cette clé nous avons pu coupler les différents dataset extrait (Statistiques classiques des joueurs par match, statistiques avancées des joueurs par match, vainqueur du trophée, vote pour le trophée, etc …)

Pour ajouter les statistiques collectives, nous avons couplé sur la clé « Team – Season ».

Tous les joueurs de toutes les saisons depuis 1979 (instauration de la ligne à 3 points, changement de règle trop important pour pouvoir comparer avec les saisons précédentes) représentent un peu plu que 17,000 couples « Player – Season », et donc le même nombre de lignes dans notre dataset global.

5.3 Variables explicatives

Dans ce dataset nous avons donc plus d'une cinquantaine de colonnes (listées et définies en annexes 1). Parmi ces colonnes certaines expliquent forcément plus nos labels que d'autres, il faudra donc faire une réduction de dimensions.

En plus des colonnes déjà présentes dans les sources de données il a fallu en créer d'autres qui nous semblaient pertinente :

- *Une colonne PastMVP (pour savoir si le joueur a déjà reçu le trophée), en effet, en NBA il y a un effet non négligeable sur le résultat du vote que l'on appelle la « voters fatigue » (les votants se lassent)*
- *Une colonne Game_played_proportion : en effet, la saison régulière de NBA étant dense et éprouvante (82 matchs en 5 mois), les votant ont plus tendance à récompenser les joueurs ayant joué le maximum de match possible.*

5.4 Dépendance et variables « totales »

5.4.1 Dépendance

La multitude de variables implique que certaines d'entre elles sont dépendantes les unes des autres. Voyons avec cet exemple :

- *2P% = 2P/2PA (pourcentage de réussite au tir à 2pts = nombre de tir à 2pts rentrés/nombre de tirs à 2pts rentrés), on a donc décidé de supprimer la variable 2PA. (Idem pour les 3pts)*
- *On a également FG = 3P + 2P (nb de tirs rentrés = nb de tirs à 3pts rentrés + nb de tirs à 2pts rentrés), on a également la même relation avec les tentatives donc on a décidé de supprimer les variables : FG, FGA et FG%*

Finalement pour les statistiques de tirs on a décidé de garder : 2P, 2P%, 3P et 3P%, avec ces 4 variables on arrive à expliquer les 5 autres variables supprimées.

Les autres variables supprimées directement sont expliquées dans le notebook *Predicting NBA 2021-22 MVP (première cellule).*

5.4.2 Variables « totales »

D'autres variables sont des variables totales (non moyennées), en effet, prenons l'exemple de la variable MPTot. Cette variable représente le nombre total de minutes jouées en une saison. Cette variable nous pose donc un problème car selon les saisons (nb de matchs évoluant), cette variable avantage ou désavantage trop certains joueurs.

6 METHODE DE RESOLUTION DE TRAITEMENTS DE DONNEES

6.1 Idée générale

Le problème majeur de notre projet et qu'il y'a qu'un seul MVP par saison, donc 40 étiquettes positives sur plus de 17 000 candidats : c'est donc beaucoup trop peu pour pouvoir construire un modèle sur ce label.

Nous avons créé un autre label : **ShareYN** (a reçu des votes ou non), qui détermine si le joueur est dans la course ou non. Nous créons ce labels grâce aux résultats des votes de chaque saison. Ainsi nous pouvons essayer de déterminer les joueurs qui vont possiblement recevoir un vote à l'issu de la saison.

On appellera donc candidats un joueur qui a reçu au moins un vote pour sa saison.

Ainsi pour déterminer notre MVP, nous avons construit une solution en deux étapes :

- *Identifier les candidats (via une **classification** sur le lablel **ShareYN**)*
- *Départager les candidats (via une **régression** sur les résultats actuel des votes : **MVP_share**)*

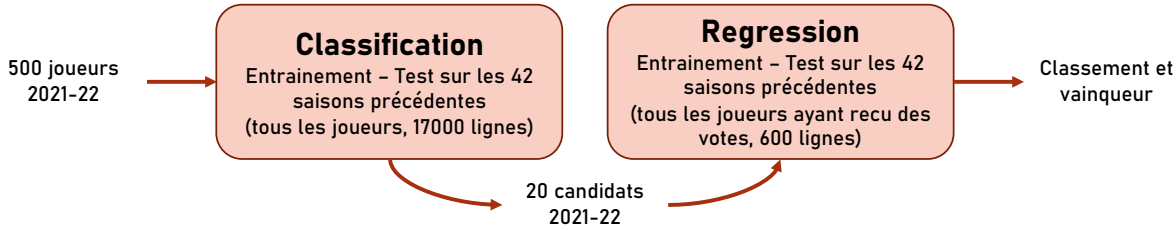


Figure 3 : Méthode de travail

6.2 Évolution du jeu et scaling par saison

6.2.1 Problème

En NBA et dans le sport en général les stratégies de jeux sont différentes en fonctions des générations. Par exemple sur les 10 dernières années les tendances montrent que le jeu en NBA devient de plus en plus offensif. L'indicateur qui nous montre c'est la pace (figure ci-dessous). La pace est le nombre moyen de points mit par match par équipe.

Étant donné les évolutions des règles et les tendances changeantes dans le jeu, le jeu dans sa globalité a beaucoup changé à travers les années. Il suffit de regarder des images du début des années 80 et les comparer avec aujourd'hui pour s'en convaincre. Ci-dessous, nous pouvons retrouver un graphique traduisant l'évolution de la *pace* (« vitesse de jeu », plus elle est haute, plus les joueurs ont des statistiques élevées) à travers les saisons.

Dans les années 2000 la pace est d'environ 87 alors qu'elle atteint une valeur de 107 en 1974. C'est une baisse d'environ 18,4% pour ce statistiques seul. Cette baisse implique que les statistiques personnels des joueurs des années 2000 sont plus faibles que celles des joueurs de 1974.

Un joueur moyen de 1974 mérite il plus d'être candidats au trophée qu'un joueur moyen des années 2000 ?

Non. Les candidats sont identifiés comme les joueurs qui se différencie par rapport à une concurrence locale (sur la saison même) puisque le trophée est annuel.

Problème : Comment comparer les chances d'être candidats de deux joueurs de deux saisons différentes ?

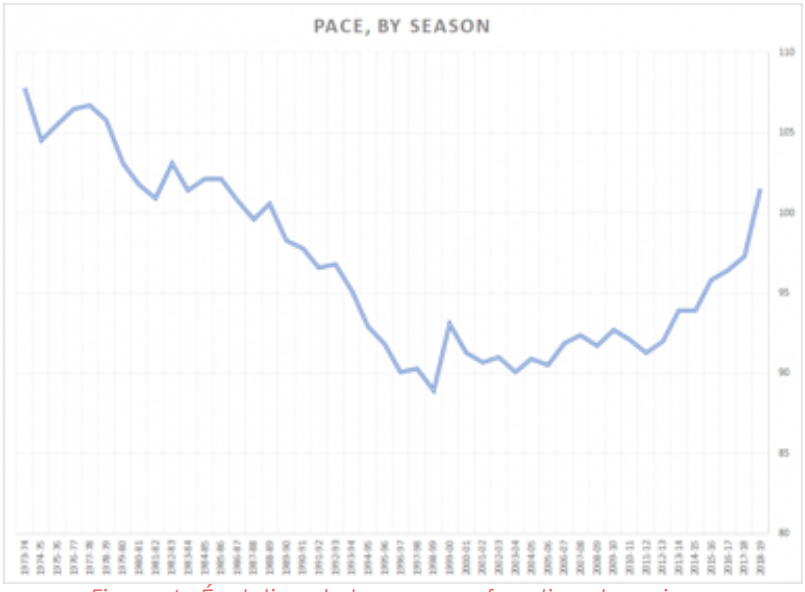


Figure 4 : Évolution de la pace en fonction des saisons

6.2.2 Solution proposée : un sclaing par saison

Pour résoudre ce problème d'évolution du jeu et de ses statistiques aux cours des saisons nous allons normaliser les variables numériques par rapport à chaque saison :

$$X_{i,scaled} = \frac{X_i - \overline{X_{season(i)}}}{\sigma(\widehat{X_{season(i)}})} \; avec \; season(i) \; la \; saison \; associée \; au \; joueur \; i$$

De cette manière les joueurs qui se seront différenciés durant leurs saisons seront identifiés comme des outliers même s'ils ne l'avaient pas été historiquement. L'idée principale est de comprendre qu'un candidat est un joueur qui se démarque de la concurrence localement (sur la saison associée) et non pas historiquement.

6.3 Répartition inégale et réduction des dimensions

6.3.1 Problème

Le nombre de candidats par saison est d'en moyenne 14. On rappelle qu'un joueur est dit candidat lorsqu'il a reçu au moins 1 vote durant l'élection du MVP. Les candidats et le MVP ensuite sont sélectionnés sur un panel de 400 joueurs en moyenne par saison. Ce qui pose le problème suivant : le nombre de joueurs labélisés « candidat » est bien inférieur aux joueurs labélisés « non-candidat ». La répartition est inégale et cela peut rendre difficile l'entrainement de notre modèle.

En effet avec de tels proportions un prédicteurs « naïf » qui labéliserais tous les joueurs comme des « non-candidat » aurait une précision moyenne de 96.7%.

6.3.2 Méthode de scoring

Une manière de supprimer les prédicteurs dit « naïf » et d'estimer l'efficacité à travers le rappel (ou alors sensibilité). Le rappel de l'estimateur « naïf » cité au-dessus est de 0% puisque qu'aucun candidat n'a été bien prédit.

Définition du rappel :

$$Rappel = \frac{Nombre \; de \; candidats \; correctement \; prédit}{Nombre \; réel \; de \; candidat}$$

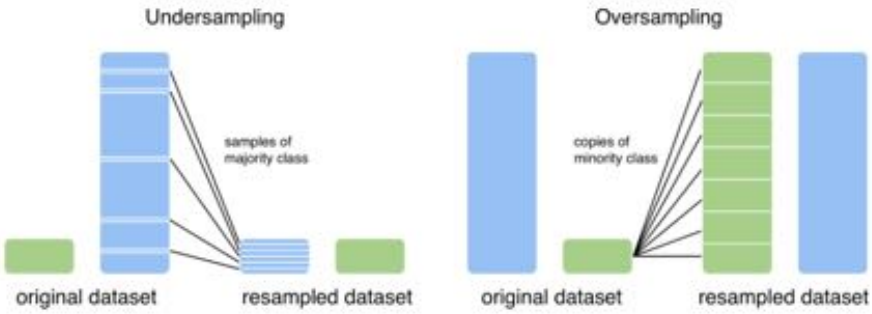
La contrepartie de l'utilisation de cette métrique pour entrainer notre classification et le taux de faux-positif. Ce phénomène de nous inquiète pas puisqu'on réalise l'hypothèse suivante. On appellera les faux positifs des « mauvais-candidats »

Hypothèse : *Les « mauvais-candidats » prédit durant la phase de classification auront un mauvais score durant la phase de régression et ne biaiserons pas la sélection du MVP (meilleur score durant la régression).*

On verra par la suite que cette hypothèse sera vérifiée.

6.3.3 Sur-échantillonnage

Pour mettre toutes les chances de notre côté et être sûr d'identifier tous les « potentiels candidats » au titre de MVP nous allons utiliser des techniques d'échantillonnage pour égaliser notre dataset. Deux grandes méthodes existent :



Le schéma est parlant. La méthode que nous allons utiliser est le sur-échantillonnage partie de droite. Dans notre cas cela reviendra à dupliquer les échantillons labéliser « candidat ». Les échantillons dupliqués ne seront pas des versions modifier (avec du bruit) comme dans la méthode d'échantillonnage SMOTE. De plus ce dataset équilibré utilisé pour l'entrainement ne sera pas utilisé pour le test. En effet les tests et les performances du modèle seront estimés avec un dataset déséquilibré, c'est-à-dire avec des conditions de prédiction réel.

7 RESULTATS ET PREDICTIONS POUR L'ANNEE 2022

7.1 Classification

7.1.1 Choix du modèle

Pour notre classification nous avons choisi un classifieur de Ridge. Auquel nous avons ajouté un recall scorer pour minimiser encore plus le taux de faux négatifs (par peur de perdre le réel MVP). Le paramètre alpha a été déterminé par Cross Validation (parmi les valeurs de 0.1, 1 et 10).

```
from sklearn.metrics import make_scorer, recall_score

scoring = True

scorer_c = make_scorer(recall_score)

if scoring :
    clf_over = GridSearchCV(clf_ridge, {'alpha' : np.array([0.01,1,10])}, scoring=scorer_c)
else :
    clf_over = GridSearchCV(clf_ridge, {'alpha' : np.array([0.01,1,10])})

clf_over.fit(X_over_c.fillna(0), y_over_c)

clf_over.cv_results_
```

```
{'mean_fit_time': array([0.14653773, 0.11637816, 0.12633367]),
 'std_fit_time': array([0.03015242, 0.00940216, 0.02090613]),
 'mean_score_time': array([0.03264809, 0.02671027, 0.02891145]),
 'std_score_time': array([0.00620699, 0.00550545, 0.00864181]),
 'param_alpha': masked_array(data=[0.01, 1.0, 10.0],
              mask=[False, False, False],
        fill_value='?',
             dtype=object),
 'params': [{'alpha': 0.01}, {'alpha': 1.0}, {'alpha': 10.0}],
 'split0_test_score': array([0.95522388, 0.95522388, 0.95522388]),
 'split1_test_score': array([0.95854063, 0.95854063, 0.95729685]),
 'split2_test_score': array([0.96600332, 0.96600332, 0.9668325 ]),
 'split3_test_score': array([0.95810867, 0.9601825 , 0.9601825 ]),
 'split4_test_score': array([0.96310116, 0.96310116, 0.96310116]),
 'mean_test_score': array([0.96019553, 0.9606103 , 0.96052738]),
 'std_test_score': array([0.00384666, 0.00370861, 0.00412459]),
 'rank_test_score': array([3, 1, 2], dtype=int32)}
```

7.1.2 Performances du modèle

Nous avons évidemment entrainé notre modèle sur le dataset oversampled et avec les dimensions les plus explicatives pour notre modèle. Nous l'avons testé sur notre échantillon de test pour évaluer ses performances.

```
classifier.fit(X_train_classifier, y_train_classifier)
```

```
RidgeClassifier(alpha=10)
```

```
classifier.score(X_test_c[features].fillna(0), y_test_c)
```

```
0.9140501110053917
```

```
from sklearn.metrics import confusion_matrix

confusion_matrix(y_test_c,classifier.predict(X_test_c[features].fillna(0)))
```

```
array([[2760,  265],
       [   6,  122]])
```

```
y_test_c.value_counts()
```

```
False    3025
True      128
Name: ShareYN, dtype: int64
```

Ainsi on retrouve un score de **91%**, avec seulement **4% de faux négatifs**. Même si la prévalence (taux de faux positifs parmi les positifs prédits) n'est pas très bonne (toujours le problème de désiquilibre de classes) ; c'est un résultat satisfaisant et nous décidons donc de l'appliquer pour définir nos candidats.

7.1.3 Prédictions pour la saison 2021-22

En faisant une prédiction sur le dataset de la saison 2021-2022, on se retrouve avec 63 potentiels candidats qu'on a choisi de trier à l'aide de la méthode `decision_function()`, on a choisis d'en sélectionner les 20 premiers.

```
y_pred_22 = classifier.predict(X_22_classif)
```

```
results_22 = Player_22.assign(Prediction = y_pred_22)

results_22['coefs'] = classifier.decision_function(X_22_classif)
```

```
results_22[results_22.Prediction].sort_values('coefs', ascending = False, ignore_index = True)
```

	Player	Prediction	coefs
0	Nikola Jokić	True	1.340512
1	Giannis Antetokounmpo	True	1.314021
2	Chris Paul	True	1.305983
3	Stephen Curry	True	1.251080
4	Joel Embiid	True	1.139622
...	...	...	...
59	Mike Conley	True	0.030248
60	Marcus Smart	True	0.028578
61	OG Anunoby	True	0.023361
62	Tyrese Haliburton	True	0.013778
63	Brandon Ingram	True	0.010641

64 rows × 3 columns

```
candidates = results_22.sort_values('coefs',ascending = False).Player.values[:20]
```

```
candidates
```

```
array(['Nikola Jokić', 'Giannis Antetokounmpo', 'Chris Paul',
       'Stephen Curry', 'Joel Embiid', 'LeBron James', 'James Harden',
       'Kevin Durant', 'Ja Morant', 'Luka Dončić', 'Jimmy Butler',
       'Devin Booker', 'DeMar DeRozan', 'Jayson Tatum',
       'Karl-Anthony Towns', 'Trae Young', 'Darius Garland',
       'Fred VanVleet', 'Rudy Gobert', 'Jarrett Allen'], dtype=object)
```

Nos candidats sélectionnés, nous pouvons passer à la partie régression.

7.2 Régression

7.2.1 Méthode

Pour la phase de régression nous avons décidé de comparer plusieurs méthodes :

- *Régression linéaire*
- *Algorithme de kNN*
- *Random Forest*
- *Multilayer Perceptron*

Les paramètres des trois derniers ont été sélectionnés par CrossValidation (voir partie III du notebook)

7.2.2 Synthèse des résultats

Modèle	R2 moyen	Prédictions pour le vainqueur du trophée		
		Classement	Équipe	Joueur
Régression linéaire	0,57	**1**	**Denver Nuggets**	**Nikola Jokić**
		2	Milwaukee Bucks	Giannis Antetokounmpo
		3	Philadelphia 76ers	Joel Embiid
Random Forest	0,63	**1**	**Phoenix Suns**	**Chris Paul**
		2	Milwaukee Bucks	Giannis Antetokounmpo
		3	Denver Nuggets	Nikola Jokić
kNN	0,64	**1**	**Denver Nuggets**	**Nikola Jokić**
		2	Milwaukee Bucks	Giannis Antetokounmpo
		3	Phoenix Suns	Chris Paul
Multilayer Perceptron	0,58	**1**	**Denver Nuggets**	**Nikola Jokić**
		2	Milwaukee Bucks	Giannis Antetokounmpo
		3	Golden State Warriors	Stephen Curry

R2 est le score de chaque régression calculée sur l'échantillon de test. Les résultats mauvais peuvent être expliqué par le fait que les votants doivent choisir les 3 meilleurs joueurs et les classer, notre modèle n'a pas l'habilité de le faire étant donné qu'il prend chaque candidat indépendamment des autres. Nos modèles ne tranchent pas alors que les votants le font.

7.3 Discussions

Parmi les résultats, un candidat clair semble se détacher du lot : le pivot des Denver Nuggets, Nikola Jokić. En regardant les résultats depuis nos yeux de fan, le résultat nous semble cohérent tant au niveau du podiums que du vainqueur.

Nous pouvons également faire une remarque sur le fait que les défenseurs (joueurs réputés pour leur habilité défensive) sont très mal représentés par notre modèle (Rudy Gobert fini 15ème en moyenne, alors qu'il est projeté dans le top 8). En effet, notre modèle prédit sur des statistiques majoritairement basées sur les qualités offensives des joueurs.

Néanmoins, les 10 prétendants retenus par le « MVP Tracker » de BasketBallReference, tous font partis des 20 candidats sélectionnés par notre modèle de classification.

8 CONCLUSION ET POINTS D'AMELIORATIONS

8.1 Conclusion

Comme expliqué après les résultats des régressions, le fait que nos modèles ne tranchant pas entre deux joueurs, il est difficile de prédire exactement les résultats exacts pour les votes (d'où les résultats de R2 médiocres).

La régression reste cependant un moyen alternatif pour réaliser un le classement et nos résultats sur la saison 2022, nous montre que c'est une méthode à laquelle on peut se fier. En effet, le but premier était de prédire le vainqueur et non les résultats exacts des votes pour chaque joueur.

8.2 Bonus : test de nos régresseurs à travers l'histoire :

Pour conclure notre projet, on s'est lancé un dernier défi. Essayer de prédire chaque MVP de chaque saison en entrainant à chaque fois nos modèle sur les autres saisons. Voici les résultats.

Season	Actual MVP	Predicted MVP Lin Reg	Predicted MVP kNN	Predicted MVP RandFor	Predicted MVP DNN
1979-1980	**Kareem Abdul-Jabbar**	Kareem Abdul-Jabbar	Kareem Abdul-Jabbar	Kareem Abdul-Jabbar	Kareem Abdul-Jabbar
1980-1981	**Julius Erving**	Julius Erving	Julius Erving	Julius Erving	Julius Erving
1981-1982	**Moses Malone**	Larry Bird	Magic Johnson	Julius Erving	Magic Johnson
1982-1983	**Moses Malone**	Moses Malone	Larry Bird	Moses Malone	Larry Bird
1983-1984	**Larry Bird**	Larry Bird	Larry Bird	Larry Bird	Larry Bird
1984-1985	**Larry Bird**	Larry Bird	Larry Bird	Larry Bird	Larry Bird
1985-1986	**Larry Bird**	Larry Bird	Larry Bird	Larry Bird	Larry Bird
1986-1987	**Magic Johnson**	Magic Johnson	Michael Jordan	Magic Johnson	Michael Jordan
1987-1988	**Michael Jordan**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1988-1989	**Magic Johnson**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1989-1990	**Magic Johnson**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1990-1991	**Michael Jordan**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1991-1992	**Michael Jordan**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1992-1993	**Charles Barkley**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1993-1994	**Hakeem Olajuwon**	David Robinson	David Robinson	David Robinson	David Robinson
1994-1995	**David Robinson**	David Robinson	David Robinson	Karl Malone	David Robinson
1995-1996	**Michael Jordan**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1996-1997	**Karl Malone**	Michael Jordan	Michael Jordan	Michael Jordan	Michael Jordan
1997-1998	**Michael Jordan**	Michael Jordan	Karl Malone	Karl Malone	Karl Malone
1998-1999	**Karl Malone**	Karl Malone	Tim Duncan	Karl Malone	Tim Duncan
1999-2000	**Shaquille O'Neal**	Shaquille O'Neal	Shaquille O'Neal	Shaquille O'Neal	Shaquille O'Neal
2000-2001	**Allen Iverson**	Shaquille O'Neal	Shaquille O'Neal	Shaquille O'Neal	Shaquille O'Neal
2001-2002	**Tim Duncan**	Tim Duncan	Tim Duncan	Tim Duncan	Tim Duncan
2002-2003	**Tim Duncan**	Tim Duncan	Tim Duncan	Tim Duncan	Tim Duncan
2003-2004	**Kevin Garnett**	Kevin Garnett	Kevin Garnett	Kevin Garnett	Kevin Garnett

2004-2005	**Steve Nash**	Dirk Nowitzki	Kevin Garnett	Amar'e Stoudemire	Kevin Garnett
2005-2006	**Steve Nash**	Dirk Nowitzki	Dirk Nowitzki	Dirk Nowitzki	Dirk Nowitzki
2006-2007	**Dirk Nowitzki**	Dirk Nowitzki	Dirk Nowitzki	Dirk Nowitzki	Dirk Nowitzki
2007-2008	**Kobe Bryant**	Chris Paul	Kevin Garnett	Chris Paul	Kevin Garnett
2008-2009	**LeBron James**	LeBron James	LeBron James	LeBron James	LeBron James
2009-2010	**LeBron James**	LeBron James	LeBron James	LeBron James	LeBron James
2010-2011	**Derrick Rose**	LeBron James	LeBron James	LeBron James	LeBron James
2011-2012	**LeBron James**	LeBron James	LeBron James	LeBron James	LeBron James
2012-2013	**LeBron James**	LeBron James	LeBron James	LeBron James	LeBron James
2013-2014	**Kevin Durant**	Kevin Durant	Kevin Durant	Kevin Durant	Kevin Durant
2014-2015	**Stephen Curry**	James Harden	Stephen Curry	Stephen Curry	Stephen Curry
2015-2016	**Stephen Curry**	Stephen Curry	Stephen Curry	Stephen Curry	Stephen Curry
2016-2017	**Russell Westbrook**	Russell Westbrook	James Harden	Kawhi Leonard	James Harden
2017-2018	**James Harden**	James Harden	James Harden	James Harden	James Harden
2018-2019	**Giannis Antetokounmpo**	James Harden	Giannis Antetokounmpo	Giannis Antetokounmpo	Giannis Antetokounmpo
2019-2020	**Giannis Antetokounmpo**	Giannis Antetokounmpo	Giannis Antetokounmpo	Giannis Antetokounmpo	Giannis Antetokounmpo
2020-2021	**Nikola Jokić**	Nikola Jokić	Nikola Jokić	Nikola Jokić	Nikola Jokić

Avec ce tableau s'accompagne les résultats d'accuracy dans la capacité de prédire les vainqueurs pour chaque algorithme (à retrouver également en fin du notebook prediction) :

- *Accuracy régression linéaire = 69 %*
- *Accuracy kNN = 62%*
- *Accuracy random forest = 66%*
- *Accuracy multilayer perceptron = 62%*

Des résultats encore une fois un peu décevant mais néanmoins explicable :

- *Les défenseurs trop pénalisés dans nos modèles (pas de prédiction pour Hakeem Olajuwon en 1993-1994)*
- *Malgré la varaible PastMVP, on n'a pas assez réussi à implémanter l'impact de la « voters fatigue » (nos modèles prédissent Micheal Jordan gagnant 7 trophées d'affillés, bien que la légende ait évoluée à un niveau impressionant sur ce range de saisons, la lassitude des votants lui ont en réalité couté quelques titres)*
- *Une part d'histoire est également importante en NBA (the « American Dream »), en effet, les votants vont souvent privilégier la « story » à la performance, variable non quantifiable et donc indétectable. Un exemple parfait, Derrick Rose 2011.*

9 ANNEXES

9.1 Détails des colonnes du dataset

Nom de la colonne	Source	Définition	Type
Player	Stats per game	Nom du joueur	Str
Pos	Stats per game	Poste du joueur	Str
Age	Stats per game	-	Int
Tm	Stats per game	Accronyme de l'équipe	Str
Team	Team stats	Nom complet de l'équipe	Str
G	Stats per game	Nb de matchs joué	Int
GS	Stats per game	Nb de match joué en tant que titulaire	Int
MPG	Stats per game	Moyenne du nb de minutes jouées par match	Float
FG	Stats per game	Moyenne de tirs rentrés par matchs	Float
FGA	Stats per game	Moyenne de tirs tentés par match	Float
FG%	Stats per game	Pourcentage de tirs rentrés	Float
3P	Stats per game	Idem avec les tirs à 3pts	Float
3PA	Stats per game		Float
3P%	Stats per game		Float
2P	Stats per game	Idem avec les tirs à 2pts	Float
2PA	Stats per game		Float
2P%	Stats per game		Float
eFG%	Stats per game	Efficacité recalculée	Float
FT	Stats per game	Idem avec les lancers francs	Float
FTA	Stats per game		Float
FT%	Stats per game		Float
ORB	Stats per game	Nb de rebonds offensifs	Float
DRB	Stats per game	Nb de rebonds défensifs	Float
TRB	Stats per game	Nb de rebonds	Float
AST	Stats per game	Nb de passe décisives	Float
STL	Stats per game	Nb d'interceptions	Float
BLK	Stats per game	Nb de contres	Float
TOV	Stats per game	Nb de pertes de balles	Float
PF	Stats per game	Nb de fautes	Float
PTS	Stats per game	Nb de points marqués	Float
Season	Created	Saison	Str
Decade	Created	Décénie	Str
Trade	Created	Joueur tradé en cours de saison	Bool
MPTot	Advanced stats	Nb total de minutes jouées	Float
PER	Advanced stats	Efficacité	Float
TS%	Advanced stats	Efficacité sur tous les types de tirs	Float
3PAr	Advanced stats	Proportion de tirs à 3 pts	Float

FTr	Advanced stats	Proportion de lancers francs	Float
ORB%	Advanced stats	% de rebons offenssifs	Float
DRB%	Advanced stats	% de rebons défensifs	Float
TRB%	Advanced stats	% de rebonds	Float
AST%	Advanced stats	% de passes décisive	Float
STL%	Advanced stats	% d'interceptions	Float
BLK%	Advanced stats	% de contres	Float
TOV%	Advanced stats	% de pertes de balles	Float
USG%	Advanced stats	% d'utilisation du joueur par son équipe	Float
OWS	Advanced stats	Estimation des victoires gagnées par le joueurs grace à ses facultés offensives	Float
DWS	Advanced stats	Estimation des victoires gagnées par le joueurs grace à ses facultés défensives	Float
WS	Advanced stats	Estimation des victoires gagnées par le joueur	Float
WS/48	Advanced stats	Estimation des victoires gagnées par le joueur ramenée sur 48 minutes	Float
OBPM	Advanced stats	Estimation du nombre de points offensifs par 100 possessions apportés par un joueur par rapport à un joueur moyen de la ligue, par rapport à une équipe moyenne.	Float
DBPM	Advanced stats	Estimation du nombre de points défensifs par 100 possessions apportés par un joueur par rapport à un joueur moyen de la ligue, par rapport à une équipe moyenne.	Float
BPM	Advanced stats	Une estimation du nombre de points par 100 possessions qu'un joueur a contribué par rapport à un joueur moyen de la ligue, dans une équipe moyenne.	Float
VORP	Advanced stats	Estimation du nombre de points par 100 possessions de l'équipe qu'un joueur a contribué au-dessus d'un joueur de niveau de remplacement (-2,0), traduite pour une équipe moyenne et calculée au prorata d'une saison de 82 matchs.	Float
MVP	Historic of trophy winners	Vainqueur du trophée ou non	Bool
PastMVP	Created	A déjà gagné le trophée ou non	Bool
W	Team stats	Nb de victoires	Int
L	Team stats	Nb de défaites	Int
W/L%	Team stats	Pourcentage de victoire	Float
GB	Team stats	Nb de victoire en moins par rapport au premier de la ligue	Int
PS/G	Team stats	Nb de points marqués par match	Float
PA/G	Team stats	Point marqués par l'adversaire	Float
SRS	Team stats	-	Float

GBC	Team stats	Nb de victoire en moins par rapport au premier de sa conférence	Int
Playoffs	Created	Est-ce que l'équipe à fait les playoffs ou non (8 premiers de chaques conférences)	Bool
Ranking_Conf	Created	Classement conférence	Int
Ranking_League	Created	Classement ligue	Int
MVP_share	Historic of MVP votes	"Pourcentage de vote" pour le trophé du mvp	Float
Game_played_prop	Created	Proportion des matchs joués	Float
ShareYN	Created	Est-ce que le joueur a recu des votes pour le trophée de MVP	Float