Baptiste Hurel

M1 IoT

Rapport de mon projet de Cloud Computing

A) Elaboration du projet

- Problématique du projet

Ce projet a pour but de créer un modèle qui utilise un algorithme et créer une API qui utilise ce modèle.

On cherche à définir les proportions de joueurs dans l'histoire de la NBA et comment se répartit les points dans une équipe / par poste.

On cherche aussi la corrélation entre point marqués et matchs joués en carrière par joueurs.

- Choix du jeu de données

Pour répondre à cette problématique, j'ai choisi le dataset suivant :

https://www.kaggle.com/drgilermo/nba-players-stats?select=Seasons_Stats.csv

Visuel du dataset :

Entrée [45]:
```
#seasons_stats.drop(['Unnamed: 0'], axis='columns', inplace = True)
seasons_stats
```

Out[45]:

	Year	Player	Pos	Age	Tm	G	GS	MP	PER	TS%	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
0	1950.0	Curly Armstrong	G-F	31.0	FTW	63.0	NaN	NaN	NaN	0.368	...	0.705	NaN	NaN	NaN	176.0	NaN	NaN	NaN	217.0	458.0
1	1950.0	Cliff Barker	SG	29.0	INO	49.0	NaN	NaN	NaN	0.435	...	0.708	NaN	NaN	NaN	109.0	NaN	NaN	NaN	99.0	279.0
2	1950.0	Leo Barnhorst	SF	25.0	CHS	67.0	NaN	NaN	NaN	0.394	...	0.698	NaN	NaN	NaN	140.0	NaN	NaN	NaN	192.0	438.0
3	1950.0	Ed Bartels	F	24.0	TOT	15.0	NaN	NaN	NaN	0.312	...	0.559	NaN	NaN	NaN	20.0	NaN	NaN	NaN	29.0	63.0
4	1950.0	Ed Bartels	F	24.0	DNN	13.0	NaN	NaN	NaN	0.308	...	0.548	NaN	NaN	NaN	20.0	NaN	NaN	NaN	27.0	59.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
24686	2017.0	Cody Zeller	PF	24.0	CHO	62.0	58.0	1725.0	16.7	0.604	...	0.679	135.0	270.0	405.0	99.0	62.0	58.0	65.0	189.0	639.0
24687	2017.0	Tyler Zeller	C	27.0	BOS	51.0	5.0	525.0	13.0	0.508	...	0.564	43.0	81.0	124.0	42.0	7.0	21.0	20.0	61.0	178.0
24688	2017.0	Stephen Zimmerman	C	20.0	ORL	19.0	0.0	108.0	7.3	0.346	...	0.600	11.0	24.0	35.0	4.0	2.0	5.0	3.0	17.0	23.0
24689	2017.0	Paul Zipser	SF	22.0	CHI	44.0	18.0	843.0	6.9	0.503	...	0.775	15.0	110.0	125.0	36.0	15.0	16.0	40.0	78.0	240.0
24690	2017.0	Ivica Zubac	C	19.0	LAL	38.0	11.0	609.0	17.0	0.547	...	0.653	41.0	118.0	159.0	30.0	14.0	33.0	30.0	66.0	284.0

- Lien du projet sur le cloud :

https://cloudcomputing1app.herokuapp.com/

B) Etapes du projet :

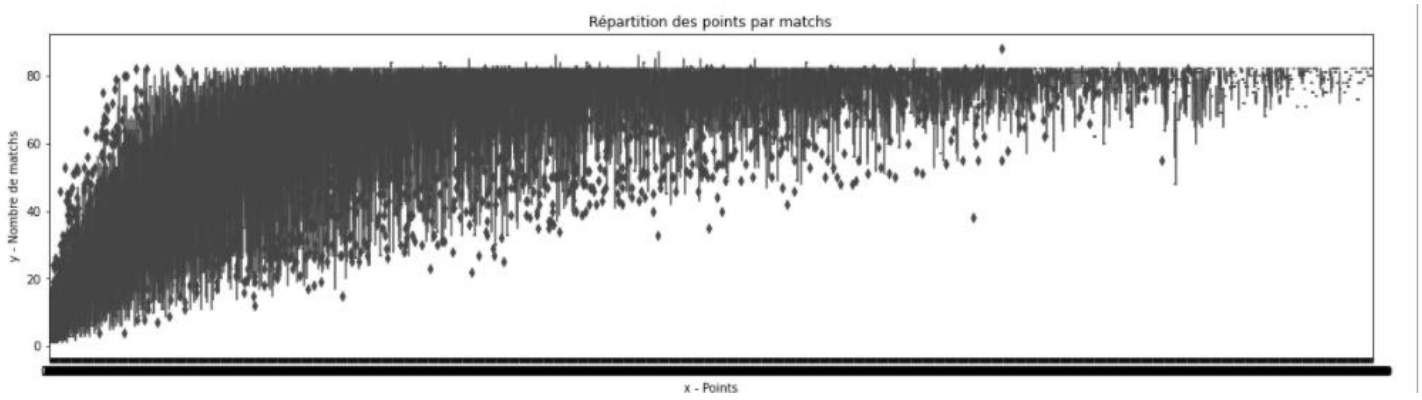
1) Analyse graphique des données (EDA)

J'ai commencé par la rédaction de mon Notebook EDA avec différentes étapes :

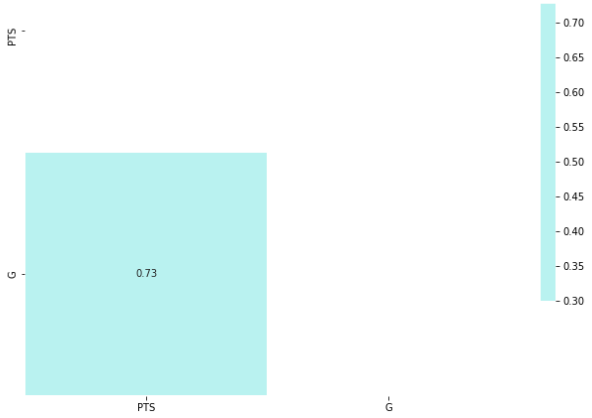
- Nettoyage rapide de mon dataset et récupération des données nécessaire à mon analyse.
- Nettoyage des données manquante avec un encodage OneHot, notamment pour les postes des joueurs. Cela facilitera la suite de mon analyse.

```
Classes : ['C' 'PF' 'PG' 'SF' 'SG']
Encodage par labels : [2 4 1 3 0]
Encodage one-hot :
[[0. 0. 1. 0. 0.]
 [0. 0. 0. 0. 1.]
 [0. 1. 0. 0. 0.]
 [0. 0. 0. 1. 0.]
 [1. 0. 0. 0. 0.]]
```

- Répartition des données avec des boites à moustaches pour visualiser les points par matchs.



- Observation des corrélations entre « Points marqués en carrières » et « Nombres de matchs joués »



2) Model Building

- Rédaction du Notebook Data Pipelines et Predict avec les 4 classes:

- Data Handler
- Feature Recipe
- Future Extactor
- Model Builder

Chaque classe comporte des méthodes qui seront utilisés plus tard.

- Dans le fichier **model.py**, création de la fonction **DataManager** avec l'instanciation de mes différentes classes (Data Handler / Feature Recipe / Future Extactor)

3) Optionnel …

4) Optionnel …

5) API, Conteneurisation et déploiement GCP

- J'ai construit mon API en utilisant la librairie Fast API.
- Conteneurisation avec Docker
- Déploiement de mon API avec la plateforme de cloud Heroku.

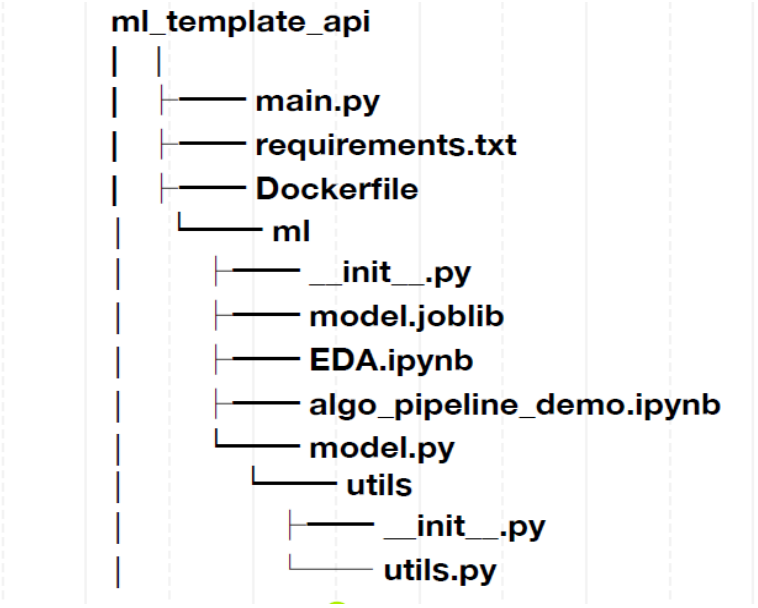
- Tout le code est disponible sur GITHUB au lien suivant :

https://github.com/BaptisteHurel/Python/tree/master/Projet%20Cloud%20Computing

6) Pacquaging, POO et refactoring

- Architecture de mon projet :

L'architecture de mon projet se présente sous la forme suivante :



Les classes créées sont présentes dans le notebook algo_pipeline_demo.ipynb.

7) Conclusion :

- Comme écrit dans la problématique on récupère bien les points marqués par les joueurs en carrière en fonction des postes et la corrélation entre point marqués et matchs joués en carrière par joueurs.

- La corrélation trouvée dans l'EDA est de 0.73.
- Il y a donc une corrélation assez importante entre « Points marqués en carrières » et « Nombres de matchs joués ».