Université de Lille

Master Informatique

Parcours Machine learning

Auto-ML

Projet FIFA Partie 3

Auteurs:
Abderrahmane Akeniouene
Fouad Boutaleb

Professeur:
Laetitia Jourdan

November 29, 2021

Contents

1 Contexte **2**

2 Description des données **2**
2.1 Attributs 2
2.1.1 Inutils 2
2.1.2 Numérique 2
2.1.3 Catégoriel 4

3 Classification **4**
3.1 DValue 4
3.1.1 Importance des Attributs 6
3.2 DWage 6
3.2.1 Importance des Attributs 8

4 Prediction **8**
4.1 Value 8
4.1.1 Importance des Attributs 9
4.1.2 Joueurs francais 9
4.2 Wage 10
4.2.1 Importances des Attributs 11
4.3 Overall 12
4.3.1 Importances des Attributs 12

5 Conclusion **13**

1 Contexte

Le but de ce rapport est de donner les étapes et l'explication du processus de la prediction de la valeur, le salaire et même le score d'un joueur, et déterminer les différents attributs (caractéristiques) qui nous permettent d'y arriver. Et cela en disposant de données relatives aux joueurs dans FIFA-19.

2 Description des données

Les données qui seront utilisées, sont celles issues de la première et le seconde partie.

2.1 Attributs

2.1.1 Inutils

Certains attributs ne nous seront d'aucune utilité dans la suite, ces attributs sont ceux de la Figure 1:

Attribut	**Description**
Name	Nom du joueur
Photo	Photo du joueur
Club Logo	Logo du Club
Flag	Drapeau du pays du joueur
Real Face	photo du joueur
Nationality	Nationalité du joueur
Joined	Date à la quelle le joueur a rejoint le club
Loaned From	le club qui a prêté le joueur
Contract Valid Until	validité du contrat

Figure 1: Attributs inutils

2.1.2 Numérique

Quelque attributs à valeurs numériques:

Attribut	**Description**
Overall	Score du joueur
Potential	Potentiel du joueur
Value	Valeur du joueur sur le marché
Wage	Salaire du joueur
Special	caracteristiques spécifiques au joueur
International Reputation	Réputation du joueur
Weak Foot	Pied faible du joueur
Skill Moves	Statistique de déplacement liée au joueur
Jersey Number	Numéro au dos du joueur
Height	Longueur du joueur (m)
Weight	Poids du joueur (kg)
Release Clause	Clause de libération
BMI	Indice Masse Corporel

Figure 2: Attributs Numériques

À travers les deux précédentes parties, nous avons vu dans notre petite étude de corrélation que certains attributs étaient fortement corrélés, ce qui signifie qu'ils ont le même comportement. On a décidé donc de n'en laisser qu'un par groupe d'attributs corrélés, c'est également ce que nous allons faire par la suite. Les figures 3,4 et 5 font office de description de la liste de positionnement, tous ont comme valeur une évaluation de la compétence du joueur sur ce poste (sur 100).

Attribut	**signification**
RWB	Right Wing Back
RB	Right back
RCB	Right center back
CB	Center back
LCB	Left center back
LB	Left back
LWB	Left Wing Back

Figure 3: Attributs Défensifs

Attribut	**signification**
CDM	Center defensive midfield
LDM	Left defensive midfield
RM	Right midfield
RCM	Right center midfield
CM	Center Midfield
LCM	Left center midfield
LM	Left midfield
RAM	Right attacking midfield
CAM	Center attacking midfield
LAM	Left attacking midfield

Figure 4: Attributs milieu de terrain

Attribut	**signification**
RF	Right forward
CF	Center forward
LF	Left forward
RS	Right Striker
ST	Striker
LS	Left Striker

Figure 5: Attributs offensifs

Il reste néanmoins quelques attributs que nous ne citons pas ici, représentant les différents scores liés aux aptitudes du joueur, comme le jeu de tête, les passes courtes, les passes longues, puissance de tir etc.Ces attributs sont tous une évaluation sur une échelle de 0 à 100.

2.1.3 Catégoriel

La Figure 6 représente les différents attributs catégoriels de notre dataset.

Attribut	**Description**
Age	Groupe d'âge: -20,20-25,25-30,330-35,30+
Preferred Foot	Droit ou Gauche
Work Rate	Niveau: Low/Low, ..., Low/Medium, ..., High/High
Body Type	Type: Normal, Lean, Stocky, Unkown
Position	GK, DEF, MID, FWD

Figure 6: Attributs Catégoriel

3 Classification

Dans cette partie, nous cherchons à prédire la classe de chaque joueur en ce qui concerne son salaire (Wage) et sa valeur (Value). Pour cela une discrétisation de nos valeurs Wage et Value (initialement continues) a été faite, on obtient alors deux nouveaux attributs DWage et DValue.

Pour la discrétisation, plusieurs choix s'offrent à nous, soit on décide de discrétiser de telle manière à respecter les écarts initialement présents entre les valeurs continues, et cela en regroupant par exemple, les valeurs faibles ensemble, les moyennes ensemble également et enfin les plus élevées.

Mais cela pourrait avoir un inconvénient, car les groupes ne seraient peut-être pas balancés, ce qui pourrait fausser nos résultats.

Nous pouvons également faire en sorte d'avoir des ensembles balancés, quitte à regrouper des valeurs plus ou moins éloignées les unes des autres.

Dans les sections 3.1 et 3.2, nous allons voir les différents scores obtenus à travers plusieurs méthodes de classification.

3.1 DValue

class	precision	recall	f1-score	support
0	0.77	0.74	0.75	1424
1	0.50	0.58	0.54	1262
2	0.64	0.64	0.64	2137
3	0.69	0.61	0.65	1186
accuracy			0.64	6009
macro avg	0.65	0.64	0.64	6009
weighted avg	0.65	0.64	0.65	6009

Figure 7: KNeighbors Classifier DValue report

class	precision	recall	f1-score	support
0	0.92	0.90	0.91	1424
1	0.85	0.87	0.86	1262
2	0.85	0.85	0.85	2137
3	0.82	0.83	0.82	1186
accuracy			0.86	6009
macro avg	0.86	0.86	0.86	6009
weighted avg	0.86	0.86	0.86	6009

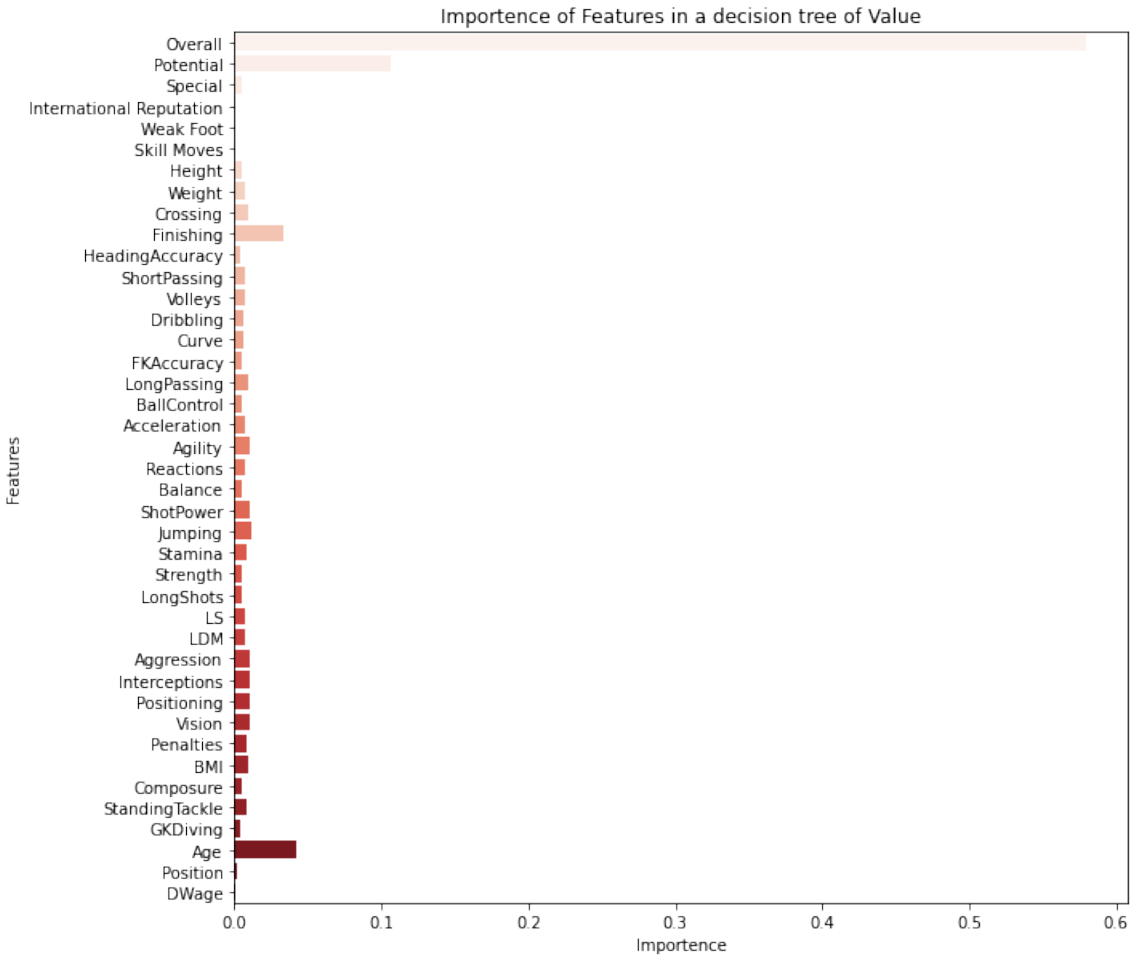
Figure 8: Decision Tree Classifier DValue report

class	precision	recall	f1-score	support
0	0.68	0.68	0.68	1424
1	0.47	0.65	0.55	1262
2	0.67	0.51	0.58	2137
3	0.68	0.69	0.68	1186
accuracy			0.62	6009
macro avg	0.63	0.63	0.62	6009
weighted avg	0.63	0.62	0.62	6009

Figure 9: Gaussian naive bayes Classifier DValue report

On peut voir que sur les trois classifieurs, Decision Tree (Figure 8) est celui qui performe le mieux, sur les 4 classes (0,1, 2 et 3) et suivant chacune des metrics.Dans la classification multilabels comme celle-ci, on aura tendance à favoriser le weighted avg, car celui-ci prend en compte le support de chaque classe.

3.1.1 Importance des Attributs



Dans la figure ci-dessus, on met en évidence les attributs qui aident le plus à la décision dans un Decision Tree pour la classification de DValue. Ces attributs sont donc Overall, Potential,Age et Finishing.

3.2 DWage

class	precision	recall	f1-score	support
0	0.92	0.96	0.94	5094
1	0.64	0.51	0.57	843
2	0.68	0.24	0.35	72
accuracy			0.89	6009
macro avg	0.75	0.57	0.62	6009
weighted avg	0.88	0.89	0.88	6009

Figure 10: KNeighbors Classifier DWage report

class	precision	recall	f1-score	support
0	0.93	0.93	0.93	5094
1	0.55	0.57	0.56	843
2	0.50	0.51	0.51	72
accuracy			0.87	6009
macro avg	0.66	0.67	0.67	6009
weighted avg	0.88	0.87	0.87	6009

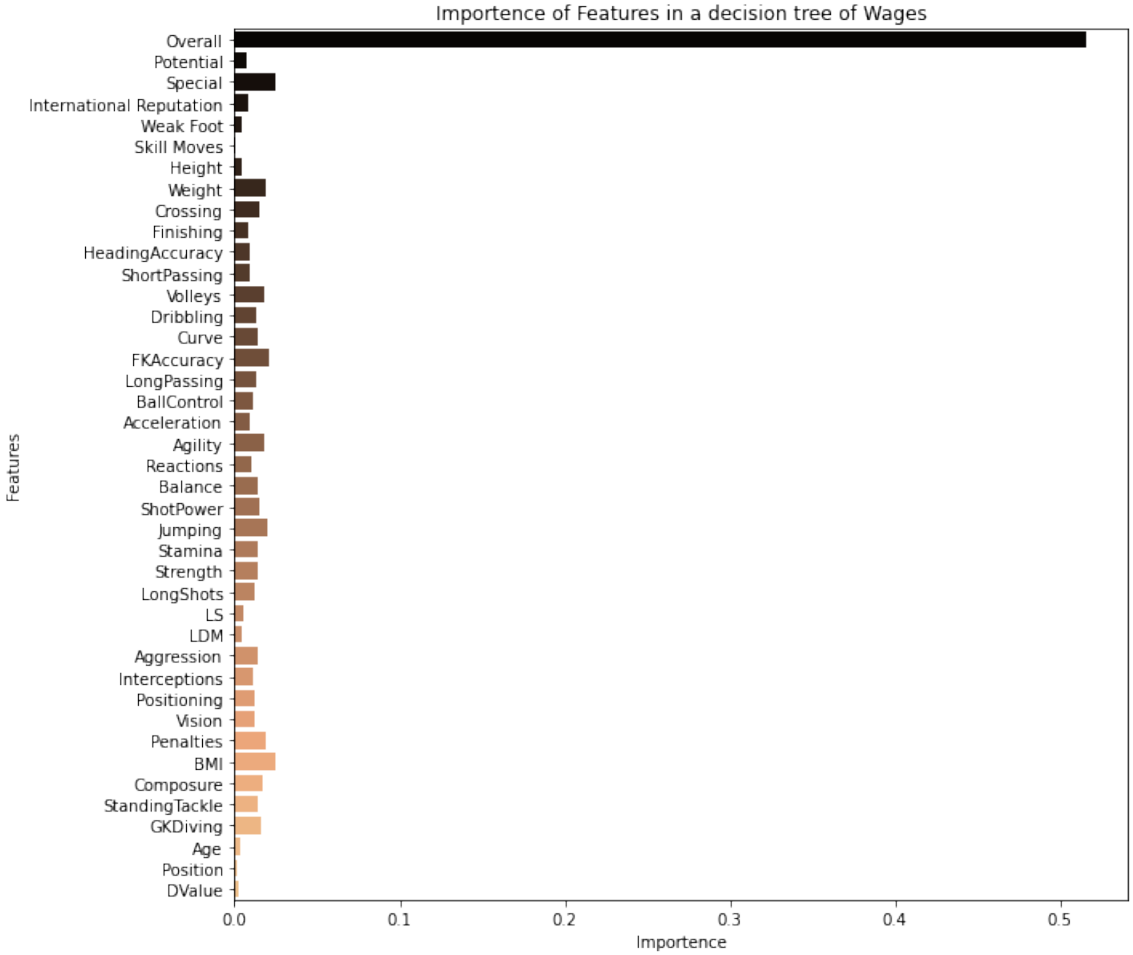
Figure 11: DecisionTree Classifier DWage report

class	precision	recall	f1-score	support
0	0.97	0.82	0.89	5094
1	0.40	0.74	0.52	843
2	0.36	0.79	0.49	72
accuracy			0.80	6009
macro avg	0.57	0.78	0.63	6009
weighted avg	0.88	0.80	0.83	6009

Figure 12: Gaussian Naive Bayes Classifier DWage report

Pour DWage deux algorithmes sortent du lot, KNeighbors Classifier et Decision Tree Classifier (Figures 10 et 11).On peut voir aussi que leurs performances sont meilleures sur la classe 0 puis moyennement bonne la classe 1 et est assez faible sur la classe 2 ayant le support le moins important des trois classes.

3.2.1 Importance des Attributs



Dans la figure ci-dessus, on met en évidence les attributs qui aident le plus à la décision dans un Decision Tree pour la classification de DWage.L'attribut dominant est Overall, puis le suivent tout un ensemble d'autres attributs ayant plus ou moins la même importance.

4 Prediction

A présent, nous aimerions prédire les valeurs originales (continues) des attributs Overall, Wage et Value. Pour cela on aura recours à de la régression.

4.1 Value

Pour cette première partie, nous allons nous focaliser sur Value.Nous allons voir comment différentes méthodes de régression se comportent en ayant comme cible cet attribut.Nous aimerions savoir également quels sont les attributs les plus importants (weights) pour cette prédiction.

Classifieur	**R^2 Score**
Lasso	0.9100516725962828
LinearRegression	0.9114466921405845
Ridge	0.9114572075302693

Figure 13: Regression Scores

À travers la figure 13, on peut voir que les trois méthodes ont à peu près la même valeur de r2 score.

4.1.1 Importance des Attributs

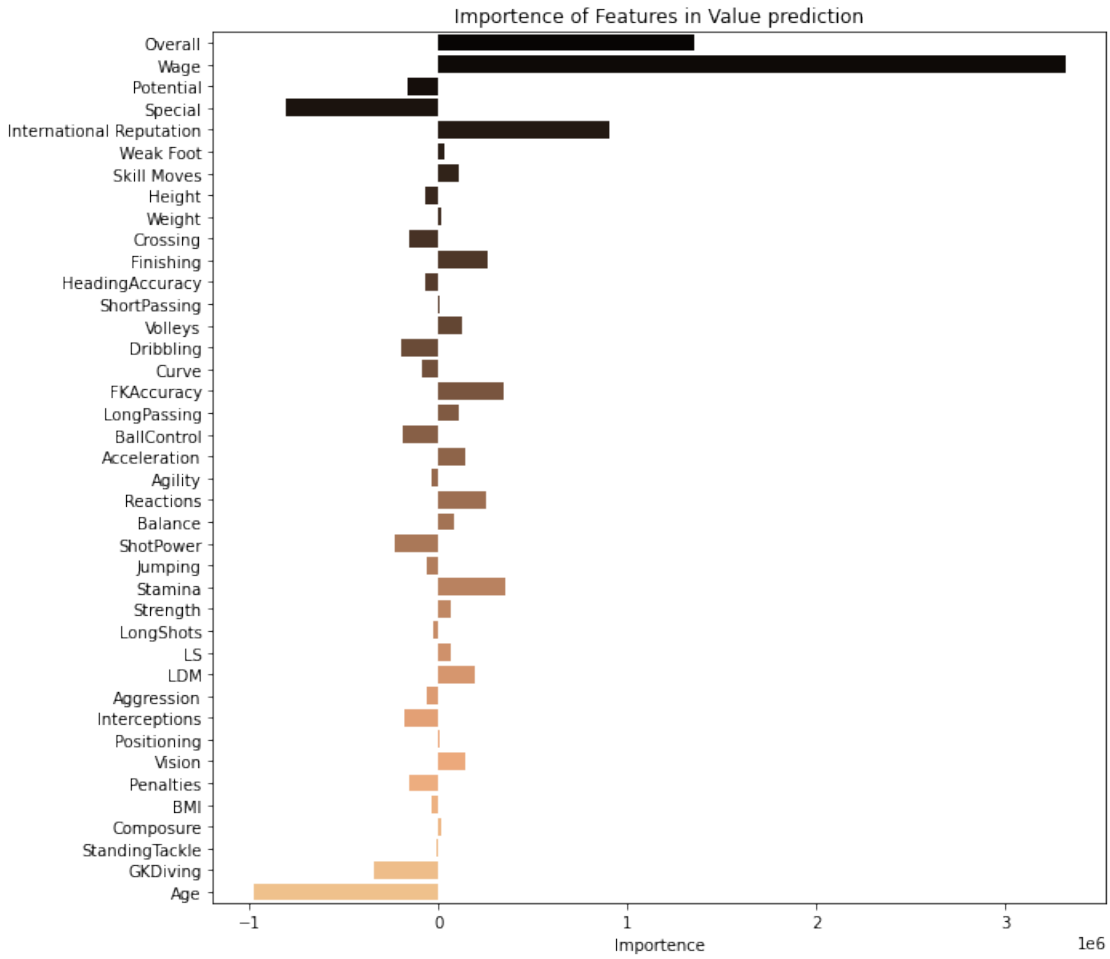


Figure 14: Importance des attributs pour Value

Nous avons 5 attributs qui se distinguent, Overall,Wage,Special,International Reputation et Age.

4.1.2 Joueurs francais

Et si on s'intéressait aux joueurs français en particulier. Aurions-nous la mêmes répartition de l'importance des attributs que dans la section 4.1.1?Les résultats obtenus avec les joueurs français sont les suivants;

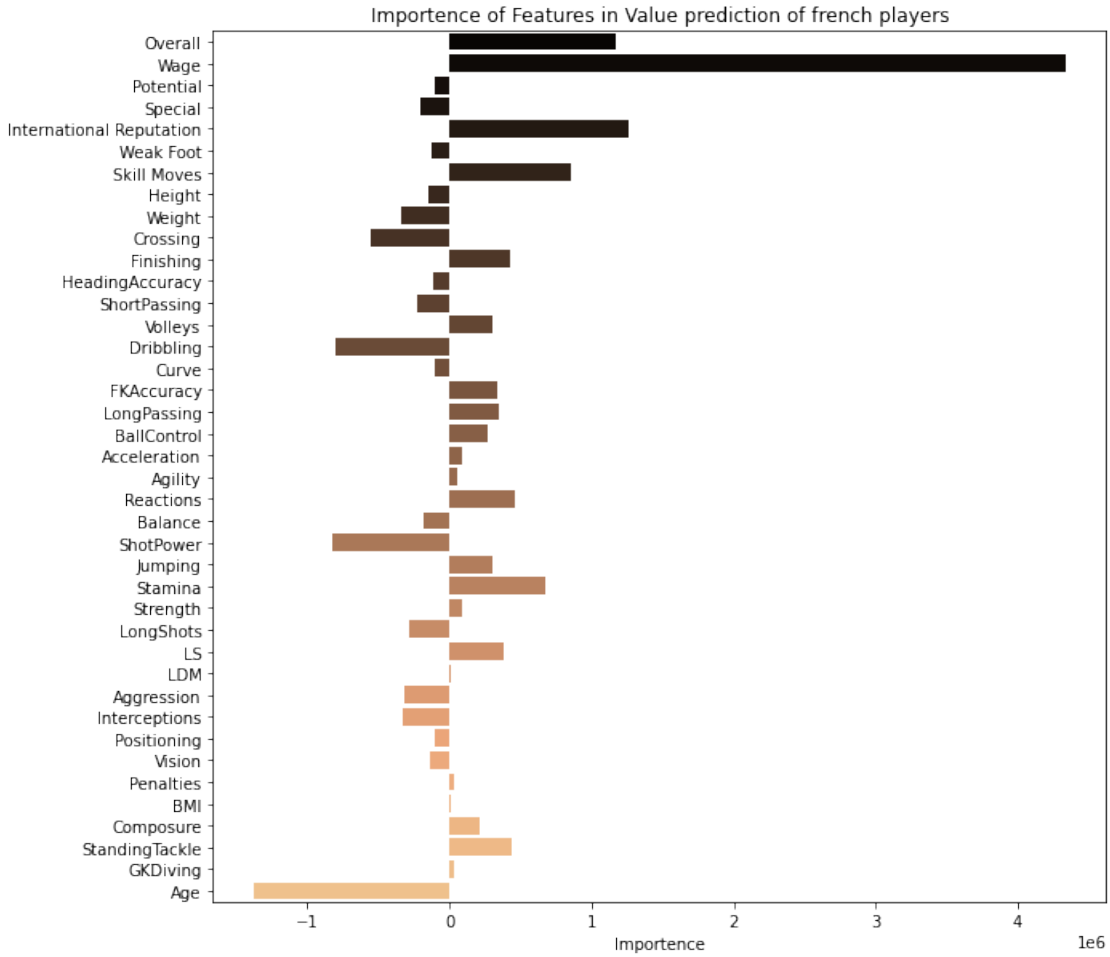


Figure 15: Importance des attributs pour les joueurs francais pour Value

On peut clairement voir que plusieurs nouveaux attributs apparaissent en plus de ceux de la section 4.1.1, par exemple SkillMoves et ShotPower.

4.2 Wage

Les différents scores obtenus avec les mêmes méthodes de régression que précédemment sont représentés dans la figure 16.

Classifieur	**R^2 Score**
Lasso	0.8135645457987263
LinearRegression	0.8136072623887647
Ridge	0.8135905767523653

Figure 16: Regression Scores

Les trois méthodes ont plus ou moins les mêmes résultats de score r2. Comme on peut le voir les résultats ne sont pas très satisfaisants vu qu'on fait en moyenne 2 erreurs de prédiction sur 10.On peut mettre en évidence cela avec la figure 17.

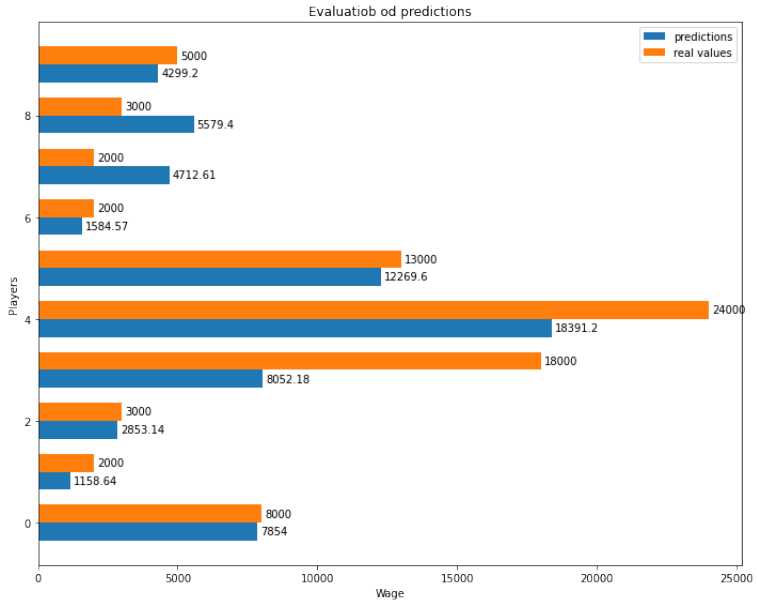


Figure 17: Comparaison entre les vraies valeurs et celles prédites

4.2.1 Importances des Attributs

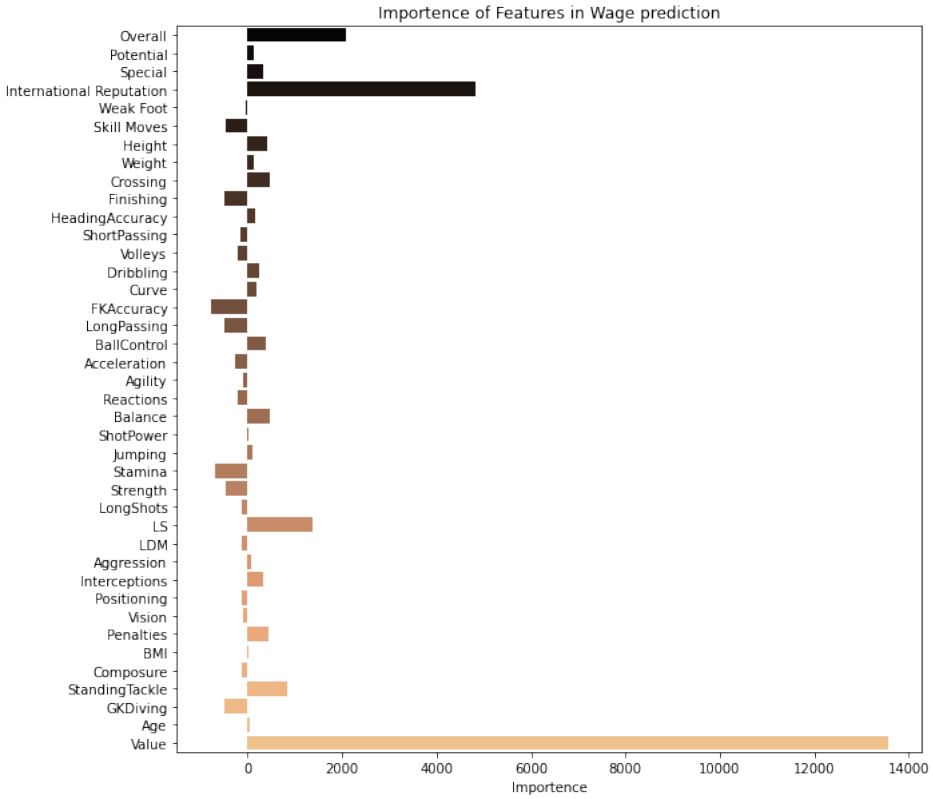


Figure 18: Importance des attributs pour Wage

Les résultats obtenus sont encore une fois cohérents, car il est normal qu'un salaire soit à la hauteur de la valeur du joueur, sa réputation internationale ainsi que ses capacités générales.

4.3 Overall

Tout comme Value et Wage nous allons prédire la valeur d'Overall grâce aux mêmes méthodes. La figure 20 représente les différents scores.

Classifieur	**R^2 Score**
Lasso	0.8358379006487198
LinearRegression	0.9174712978409312
Ridge	0.9174709111155189

Figure 19: Regression Scores

Linear Regression et Ridge ont des résultats très intéressants, car on sait que les deux attributs Potential et overall sont corrélés, c'est ce qui pousse nos modèles à des résultats aussi bons. Pour Lasso, on peut obtenir de meilleurs résultats en parametrant l'hyperparametre alpha par exemple.

4.3.1 Importances des Attributs

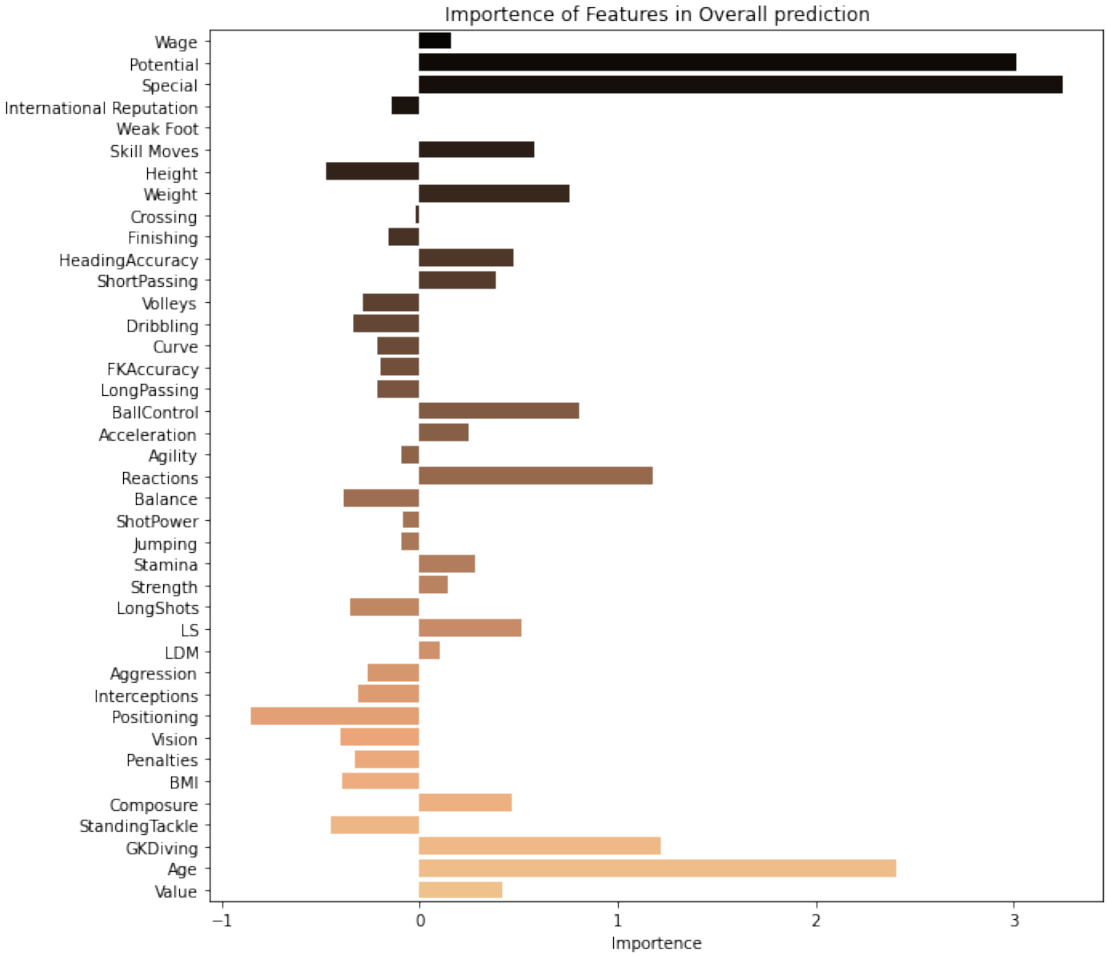


Figure 20: Importance des attributs pour Overall

5 Conclusion

Pour conclure on pourrait dire que la prédiction des valeurs, salaires et capacités des joueurs avec leurs valeurs continues est possible, néanmoins, il serait judicieux de poursuivre l'étude avec de la validation croisée, pour être sur de nos résultats et d'éliminer l'hypothèse de l'overfiting, même si nos résultats ne sont pas si élevés.Pour la classification, comme nous l'avons dit plus haut, l'étape de la discrétisation est importante quant aux résultats.Pour finir, on ne peut pas sélectionner des joueurs suivants quelques attributs particuliers uniquement. Car comme vu plus haut, les joueurs français sont sélectionnés par rapport à différents attributs par exemple. De plus, un bon milieu de terrain peut ne pas avoir un salaire élevé et une réputation internationale extraordinaire.