Université de Lille

Master Informatique

Parcours Machine learning

De la fouille de données à l'auto-ML

Projet FIFA - Part III: Prediction

Auteur:
Selim Lakhdar

Professeur:
Laetitia Jourdan

November 29, 2021

Contents

1 Context

Ce rapport a pour but d'aider à la prédiction de la valeur d'un joueur. Ainsi que de déterminer les attributs qui impactent sur sa valeur. Nous verrons aussi différents types de classification pour essayer de prédire d'autres attributs liés à la valeur du joueur.

2 Dataset

Dans cette partie nous reprendrons le dataset que nous avons obtenu après avoir "nettoyer" et discrétiser nos attributs.

2.1 Attributs

Pour résumer notre dataset nous pouvons catégoriser nos attributs ainsi;

2.1.1 Non Significatif

La Figure 1 représente les différents attributs que nous n'utiliserons pas lors de la prédiction, car elles n'apportent pas plus d'informations à notre dataset.

Attribut	**Description**
Name	Nom du joueur
Photo	Photo du joueur
Club Logo	Logo du Club
Flag	Drapeau du joueur
Real Face	Image du joueur
Name	Nom du joueur
Nationality	Nationalité du joueur
Club	Club du joueur
Joined	Date à la quelle le joueur a rejoint le club
Loaned From	Acheté depuis quel club
Contract Valid Until	durée du contrat

Figure 1: Attributs non significatif

2.1.2 Numérique

La Figure 2 représente les différents attributs numériques de notre dataset.

Attribut	**Description**
Overall	Score du joueur
Potential	Potentiel du joueur
Value	Valeur du joueur
Wage	Valeur du joueur suivant son age
Special	Statistique Spécial liée au joueur
International Reputation	Réputation du joueur
Weak Foot	Pied faible du joueur
Skill Moves	Statistique de déplacement liée au joueur
Jersey Number	Numéro du joueur
Height	Longueur du joueur
Weight	Poids du joueur
Release Clause	Clause de libération
BMI	Indice Masse Corporel

Figure 2: Attributs Numérique

La Figure 3 regroupe les différents attributs numériques liés aux statistiques des positions. Dans la partie précédente nous avons vu que ces attributs étaient fortement liés. Nous allons donc nous limiter à un attribut par position.

Attribut	**Position**	**Attribut**	**Position**	**Attribut**	**Position**
SW	DEF	RDM	MID	RF	FWD
RWB	DEF	CDM	MID	CF	FWD
RB	DEF	LDM	MID	LF	FWD
RCB	DEF	RM	MID	RW	FWD
CB	DEF	RCM	MID	RS	FWD
LCB	DEF	CM	MID	ST	FWD
LB	DEF	LCM	MID	LS	FWD
LWB	DEF	LM	MID	LW	FWD
-	-	RAM	MID	-	-
-	-	CAM	MID	-	-
-	-	LAM	MID	-	-

Figure 3: Statistique Numérique

La figure 4 représente elle, les différents scores liés aux aptitudes du joueur.

Attribut	Description	Attribut	Description
Crossing	Valeur entre 0 et 100	Finishing	Valeur entre 0 et 100
HeadingAccuracy	Valeur entre 0 et 100	ShortPassing	Valeur entre 0 et 100
Volleys	Valeur entre 0 et 100	Dribbling	Valeur entre 0 et 100
Curve	Valeur entre 0 et 100	FKAccuracy	Valeur entre 0 et 100
LongPassing	Valeur entre 0 et 100	BallControl	Valeur entre 0 et 100
Acceleration	Valeur entre 0 et 100	SprintSpeed	Valeur entre 0 et 100
Agility	Valeur entre 0 et 100	Reactions	Valeur entre 0 et 100
Balance	Valeur entre 0 et 100	ShotPower	Valeur entre 0 et 100
Jumping	Valeur entre 0 et 100	Stamina	Valeur entre 0 et 100
Strength	Valeur entre 0 et 100	LongShots	Valeur entre 0 et 100
Aggression	Valeur entre 0 et 100	Interceptions	Valeur entre 0 et 100
Positioning	Valeur entre 0 et 100	Vision	Valeur entre 0 et 100
Penalties	Valeur entre 0 et 100	Composure	Valeur entre 0 et 100
Marking	Valeur entre 0 et 100	StandingTackle	Valeur entre 0 et 100
SlidingTackle	Valeur entre 0 et 100	GKDiving	Valeur entre 0 et 100
GKHandling	Valeur entre 0 et 100	GKKicking	Valeur entre 0 et 100
GKPositioning	Valeur entre 0 et 100	GKReflexes	Valeur entre 0 et 100
Marking	Valeur entre 0 et 100	-	-

Figure 4: Statistique Numérique

2.1.3 Catégoriel

La Figure 5 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 5: Attributs Catégoriel

Par la suite nous changerons ces attributs en plusieurs attributs indicateurs (dummy), notamment pour la régression.

2.2 Discrétisation

2.2.1 Value

Nous avons vu dans la première partie qu'il y avait plusieurs façons de discrétiser la valeur Value. En effet, les valeurs ne sont pas répartis uniformément ce qui introduit une variance plus ou moins importante entre les groupes, suivant notre discrétisation. On peut observer cela grâce au traçage de la courbe qui représente Value pour chaque joueur (Figure 6). On remarque bien que la courbe est en dent de scie, et que les écarts entre ses dents de scies sont visibles (pas de continuité).

La Figure 7 représente différent découpage possible de la valeur Value. Nous observerons par la suite le score de classification sur ces groupes. Nous remarquerons qu'au plus il y a de la variance entre les groupes, plus notre classification sera moins précise (on aura des classes non balancées).

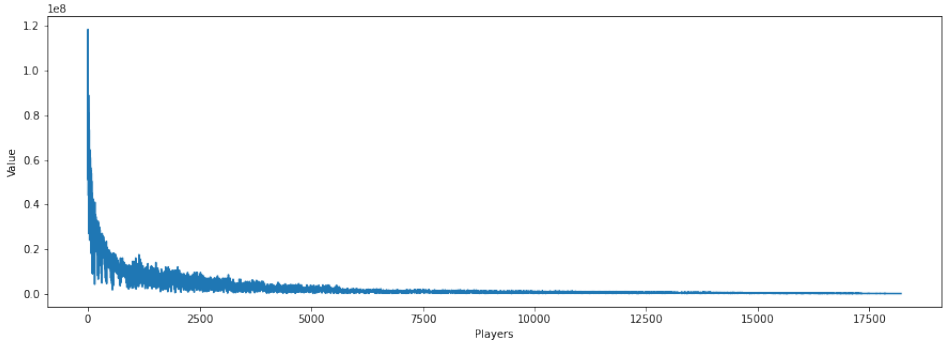

Figure 6: Value plot

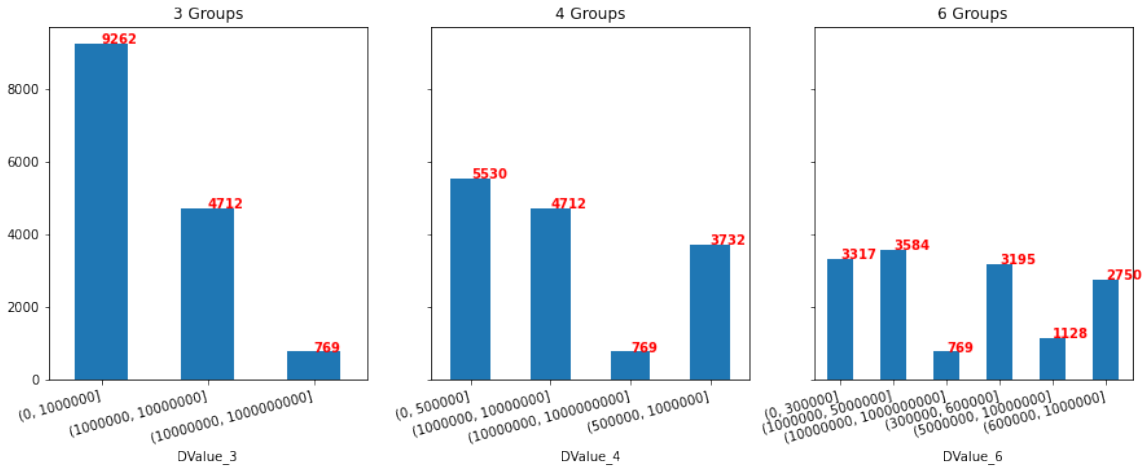

Figure 7: DValue Discretisation

2.2.2 Wage

Contrairement à Value, les valeurs de Wage sont moins dispersées. On peut l'observer grâce à la courbe (Fig 10), mais aussi à la valeur de l'écart type qui est nettement inférieur à celui de Value, **5833752** pour Value contre **22834** pour Wage.

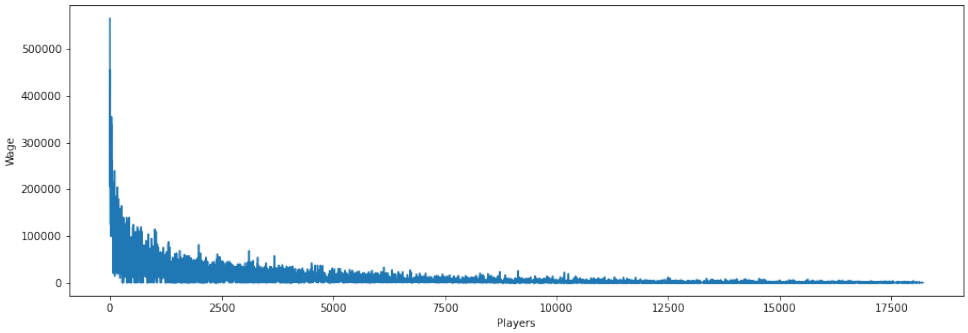

Figure 8: Wage plot

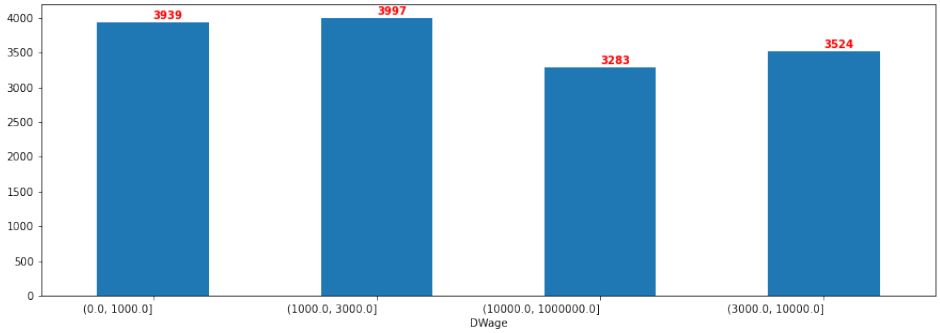

Figure 9: DWage repartition

Grâce à ce découpage nous observons une bonne répartition des éléments entre les groupes, ce qui facilitera la tâche de notre classifieur.

2.3 Correlation

Dans cette partie nous allons éliminer les différentes valeurs corrélées. En effet leur présence n'apporte pas plus d'informations à notre dataset, et peut déranger nos classifieurs. D'après la figures 10, on peut observer que la plupart des variables statistiques, que nous avons vu plus haut, sont fortement corrélées, nous allons donc les enlever. De plus, on remarque que l'attribut Release Clause est fortement corrélé avec Vale et Wage. D'autre part nous remarquons que les attributs GK* ne sont pas du tout corrélés avec le reste des attributs, cela est dû au fait d'avoir supprimer les valeurs manquantes de notre dataset pour la prédiction/classification, nous allons donc les enlever aussi ! Les attributs StandingTackle et SlidingTackle aussi.

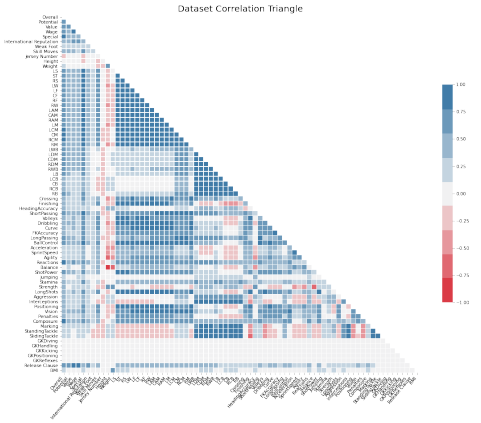

Figure 10: Correlation 1

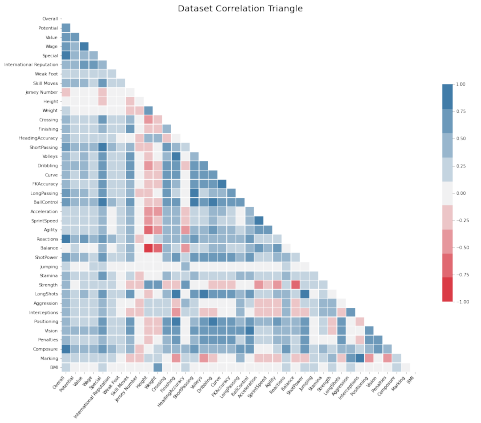

Figure 11: Correlation 2

3 Prediction

3.1 Value

Dans cette partie nous allons essayer de prédire la valeur Value pour chaque joueur. Étant donné que c'est une valeur continue nous allons prédire cette valeur grâce à de la régression. Nous allons aussi observer l'importance des attributs pris en compte pour arriver à ce résultat.
La Figure 12 résume les différents scores obtenus avec les différents classifieurs. On remarque que le score R^2 est assez élevé ce qui pourrait indiquer un overfitting.
Les colonnes R2_mean_CV10, R2_std_CV10 indiquent respectivement le score moyen de chaque classifieur avec une cross validation égale à 10, et l'écart type entre ces moyennes. On remarque un score négatif, on peut en conclure que les classifieurs linéaires ne sont pas adaptés pour prédire cette

valeur avec notre dataset.

	R2	R2_mean_CV10	R2_std_CV10
LinearRegression	0.819823	-255.783940	684.525880
LASSO	0.819823	-255.783520	684.525638
Ridge	0.819818	-255.722425	684.361062

Figure 12: Regression Scores for Value

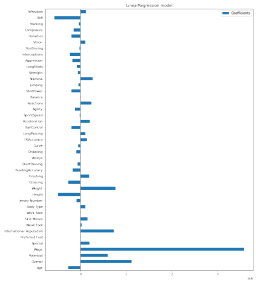

Figure 13: Value: Linear Regression Feature Importance

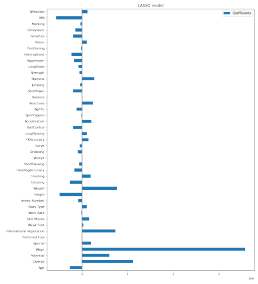

Figure 14: Value: LASSO Feature Importance

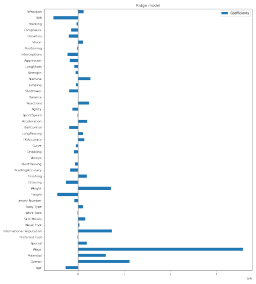

Figure 15: Value: Ridge Feature Importance

Les figures de 13 à 15 représentent les coefficients des différents attributs pris en compte pour la prédiction de Value. On remarque que les 3 classifieurs aboutissent aux mêmes coefficients à peu près. On remarque que les coefficients des attributs Wage, Age, Overall, Potential, BMI, Weight, Height et International Reputation sont plus importants que les autres. De même, les attributs Positioning, Marking, Balance, SprintSpeed, Curve, Volleys, Work Rate, Weak Foot, Preferred Foot n'impacte pas dans le choix du classifieur.

3.2 Wage

Tout come Value, nous allons prédire la valeur de Wage grâce aux mêmes classifieurs linéaires. La Figure 16 représente les différents scores. On remarque de moins bon score que pour Value.

	R2	R2_mean_CV10	R2_std_CV10
LinearRegression	0.794189	-1.285790	2.032606
LASSO	0.794198	-1.260272	1.987027
Ridge	0.794182	-1.283868	2.030639

Figure 16: Regression Scores for Wage

Les figures 17 à 19 représentent les coefficients des différents attributs pris en compte pour la prédiction de Wage. On remarque que les coefficients sont à peu près pareils entre les différents classifieurs, et presque pareille que pour Value.
On observe que les attributs Age, Value, Special et International Reputation ont des coefficients plus importants que les autres.

3.3 Overall

Tout comme Value et Wage nous allons prédire la valeur de Overall grâce aux mêmes classifieurs linéaire. La Figure 20 représente les différents scores.

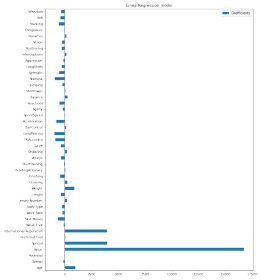

Figure 17: Wage: Linear Regression Feature Importance

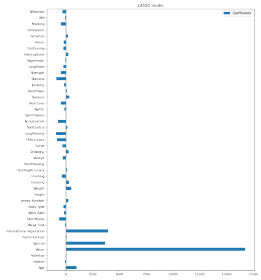

Figure 18: Wage: LASSO Feature Importance

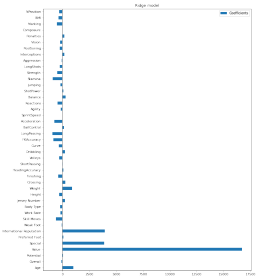

Figure 19: Wage: Ridge Feature Importance

	R2	R2_mean_CV10	R2_std_CV10
LinearRegression	0.922870	-5.797899	2.934559
LASSO	0.843138	-8.311402	2.550680
Ridge	0.922840	-5.796284	2.934056

Figure 20: Regression Scores Overall

Les figures 21 à 23 représentent les coefficients des différents attributs pris en compte pour la prédiction de Overall.

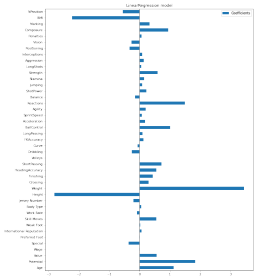

Figure 21: Overall: Linear Regression Feature Importance

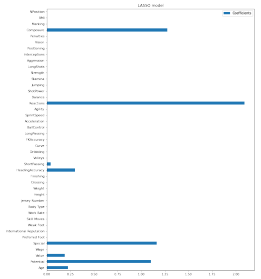

Figure 22: Overall: LASSO Feature Importance

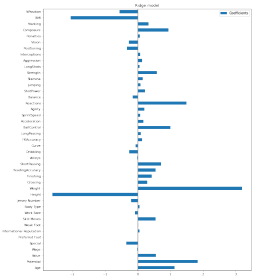

Figure 23: Overall: Ridge Feature Importance

On remarque que les coefficients ne sont pas pareils entre LASSO et les deux autres (LR et Ridge). LR et Ridge se base sur NPosition, BMI, Reaction, Weight, Height, Potential, Age. Tandis que LASSO se base sur Composure, Reaction, HeadingAccuracy, Special, Value, Age et Potential. Tout comme les anciennes prédictions, nous obtenons des scores négatifs avec de la cross validation, ce qui indiquent que ce jeu de données n'est pas adapté à la régression.

4 Classification

4.1 DValue

Si on reprend les différentes discrétisations qui ont était faite en divisant Value en 3,4 et 6 groupes (Voire Figure 7) nous obtenons différents score.

D'après les différents scores obtenus nous pouvons observer que la discrétisation en 3 groupes (Figure 24) est meilleure que celle en 4 groupes (Figure 25) qui est elle même meilleure que la discrétisation en 4 groupes (Figure 26).

4.1.1 DValue: 3 Groups

	accuracy	f1	recall	precision	f1_mean_CV10	f1_std_CV10
LogisticRegression	0.963147	0.963052	0.963147	0.963047	0.916721	0.118629
DecisionTreeClassifier	0.968799	0.968787	0.968799	0.968801	0.883725	0.151050
RandomForestClassifier	0.962921	0.962901	0.962921	0.962938	0.860834	0.186419
GaussianNB	0.840606	0.843481	0.840606	0.850207	0.810712	0.176005
KNeighborsClassifier_3	0.867737	0.866033	0.867737	0.867085	0.825663	0.089968
KNeighborsClassifier_4	0.862311	0.856590	0.862311	0.861703	0.816515	0.102246
KNeighborsClassifier_6	0.870224	0.864807	0.870224	0.869804	0.824056	0.107767

Figure 24: Classification DValue Scores

4.1.2 DValue: 4 Groups

	accuracy	f1	recall	precision	f1_mean_CV10	f1_std_CV10
LogisticRegression	0.905946	0.905821	0.905946	0.905859	0.843669	0.180698
DecisionTreeClassifier	0.931494	0.931511	0.931494	0.931529	0.816784	0.182516
RandomForestClassifier	0.926973	0.926982	0.926973	0.927013	0.773125	0.233648
GaussianNB	0.732308	0.737031	0.732308	0.751656	0.692522	0.189903
KNeighborsClassifier_3	0.730726	0.728015	0.730726	0.729527	0.692400	0.085815
KNeighborsClassifier_4	0.760118	0.750274	0.760118	0.755794	0.705030	0.093633
KNeighborsClassifier_6	0.766448	0.759450	0.766448	0.763807	0.717500	0.105438

Figure 25: Classification DValue Scores

4.1.3 DValue: 6 Groups

	accuracy	f1	recall	precision	f1_mean_CV10	f1_std_CV10
LogisticRegression	0.840380	0.840079	0.840380	0.840511	0.777335	0.207813
DecisionTreeClassifier	0.892607	0.892292	0.892607	0.892176	0.744778	0.255623
RandomForestClassifier	0.886050	0.886260	0.886050	0.887518	0.713415	0.270342
GaussianNB	0.627854	0.628603	0.627854	0.635709	0.583461	0.186409
KNeighborsClassifier_3	0.600271	0.592759	0.600271	0.596563	0.555849	0.083376
KNeighborsClassifier_4	0.633281	0.621264	0.633281	0.625496	0.577331	0.096768
KNeighborsClassifier_6	0.647072	0.638079	0.647072	0.642696	0.588475	0.105887

Figure 26: Classification DValue Scores

4.2 DWage

Tout comme DValue, nous allons essayer de prédire les valeurs discrètes de DWage que nous avions construite plus haut.

	accuracy	f1	recall	precision	f1_mean_CV10	f1_std_CV10
LogisticRegression	0.609767	0.612708	0.609767	0.618085	0.528905	0.262971
DecisionTreeClassifier	0.531540	0.531562	0.531540	0.531672	0.252740	0.066976
RandomForestClassifier	0.615193	0.615103	0.615193	0.615312	0.351820	0.193055
GaussianNB	0.556636	0.560099	0.556636	0.566438	0.500922	0.178237
KNeighborsClassifier_3	0.508026	0.507893	0.508026	0.517820	0.444934	0.109921
KNeighborsClassifier_4	0.539001	0.534621	0.539001	0.535107	0.468517	0.133113
KNeighborsClassifier_6	0.535835	0.533975	0.535835	0.536262	0.474865	0.141189

Figure 27: Classification DValue Scores

D'après la Figure 27, on observe de mauvais score de classification. Une autre discrétisation (en 3 groupes) pourrait amener à de meilleurs résultats.

5 Ce qui rend un joueur cher

Dans cette partie nous allons essayer de voir si ce qui rend un joueur français cher, rend un autre joueur cher. Plus concrètement, quel est l'importance des attributs pris en compte pour les joueurs français cher, et si c'est coefficients d'attributs sont équivalent pour d'autre joueurs.
Pour cela nous allons utiliser des arbres de régression pour prédire la valeur Value et voir l'importance des attributs pour ce choix.
Les figures suivantes représentent l'importance des attributs pour prédire la valeur Value.

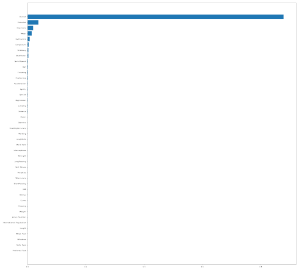

Figure 28: Value: FR Players

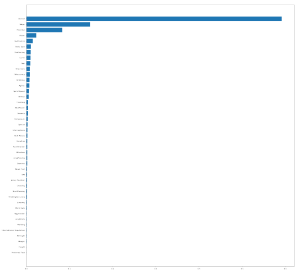

Figure 29: Value: BRZ Players

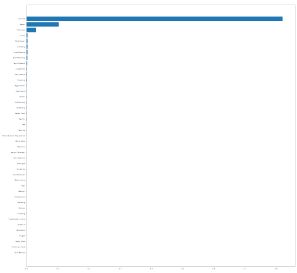

Figure 30: Value: GRM Players

La figure 28 qui représente l'importance des attribut pour prédire Value, pour les joueurs français, indique que les attributs les plus importants sont : Overall, Potential, Reaction, Wage, BallControl, Composure.
La figure 29 qui représente l'importance des attribut pour prédire Value, pour les joueurs brésiliens, indique que les attributs les plus importants sont : Overall, Wage, Potential, Vision, BallControl, BodyType.
La figure 30 qui représente l'importance des attribut pour prédire Value, pour les joueurs allemands, indique que les attributs les plus importants sont : Overall, Wage, Potential, Curve et ShotPower.

On pourrait aussi observer l'importance des attributs lorsqu'on essaie de classifier les joueurs selon le DValue que nous avons précédemment établie. Si on reprend la discrétisation en 3 groupes, on peut obtenir la figure suivante pour les joueurs français :

On peut observer d'après la figure 31 que nous obtenons à peu près les mêmes résultats que pour Value (pour les mêmes joueurs français). Ici les attributs Overall, Wage, Potential, BallControl, Reaction, Composure sont les mêmes mais avec des coefficients différents.

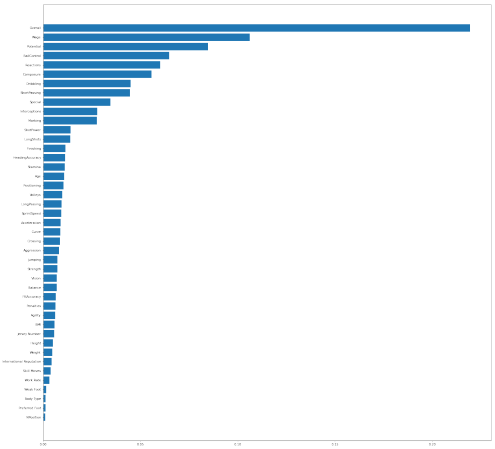

Figure 31: Value: FR Players

6 Conclusion

D'après notre étude, nous avons pu voir que prédire les valeurs continues Value, Wage et Overall était peu fiable. En effet, de très bon score indiquent un overfitting qui a été démontrer grâce au cross validation.
D'autre part, nous avons vu que les différents type de discrétisation impacte sur le score de classification.
Aussi, on ne peut pas généraliser sur un set d'attributs pour toujours sélectionner les meilleurs joueurs. En effet les différents graphiques plus haut montre une différence d'importance dans les attributs des joueurs selon leur nationalité.
Une autre analyse pourrait apporter plus d'informations. En effet la sélection des attributs en première partie (après la corrélation) peuvent nous faire aboutir à d'autres résultats.