Partie 3 - Prédiction Classification et Régression des Salaires et Valeurs Marchandes

Shaina Boutebba

Table des matières

Chapitre 1

Introduction

Ce travail vise à prédire le salaire et la valeur marchande des joueurs de football, en utilisant un ensemble de données contenant des informations techniques, physiques et contractuelles. L'objectif est double :

- **Classification** des catégories de salaire (`wage_cat`) et de valeur marchande (`value_cat`).
- **Régression** des valeurs continues du salaire (`wage_clean`) et de la valeur marchande (`value_num`).

Chapitre 2

Données et Prétraitement

2.1 Chargement et premières étapes

Nous avons utiliser les données issue du travail effectué dans la première partie du projet. Nous avons ensuite nettoyés ces données en supprimant plusieurs colonnes non pertinentes et en écartant tous les gardiens, car leurs caractéristiques ne sont pas comparables à celles des joueurs des autres postes.

2.2 Analyse des valeurs manquantes

Une analyse des valeurs manquantes montre que certaines colonnes, notamment `country_league_name`, sont inutilisables (plus de 90% de valeurs manquantes). Nous avons supprimé ces colonnes, et avons imputé certaines colonnes numériques (`shooting`, `defending`, `dribbling`, etc.) par la médiane.

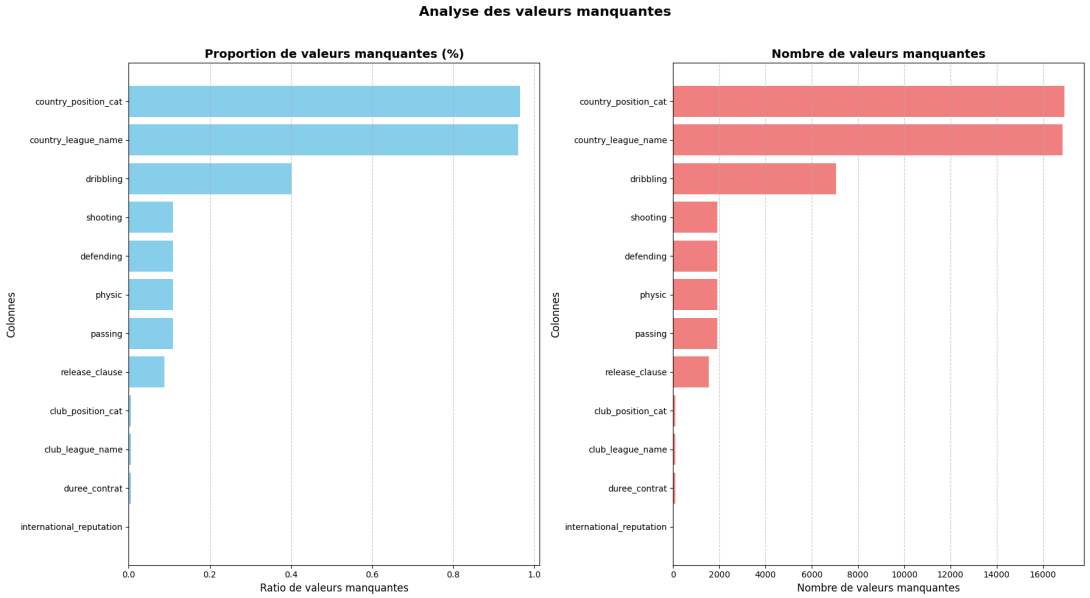

FIGURE 2.1 – Analyse des valeurs manquantes.

2.3 Encodage des variables catégorielles

Nous avons traité les variables catégorielles, notamment `preferred_foot`, `club_league_name`, `groupe_age` et `club_position_cat`, ont été encodées en **one-hot encoding**.

Un traitement particulier a été appliqué à la variable `weak_foot`, où certaines valeurs anormales (`‘‘2 -1’’`) ont été corrigées.

2.4 Colonnes utilisées

— **Colonnes numériques** : caractéristiques physiques, techniques, ratings, etc.
— **Colonnes catégorielles encodées** : poste, pied préféré, groupe d'âge, ligue.
— **Targets** :
  — Classification : `wage_cat`, `value_cat`
  — Régression : `wage_clean`, `value_num`

Chapitre 3

Analyse Exploratoire

3.1 Répartition des classes pour `wage_cat` et `value_cat`

Nous avons observé un déséquilibre de classe dans la variable `wage_cat`, ce qui nous a motivés à explorer des méthodes de rééquilibrage lors des expériences de classification.

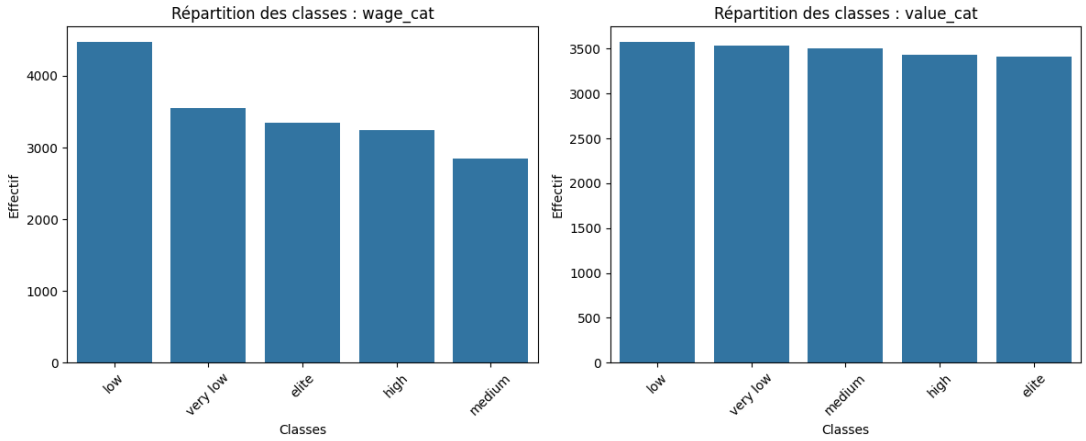

FIGURE 3.1 – Distribution des classes pour `wage_cat` et `value_cat`.

Chapitre 4

Méthodologie

4.1 Modèles utilisés

Nous avons testé les modèles ci-dessous pour la classification :
— **RandomForestClassifier**
— **SVM Classifier**
— **XGBClassifier**

Et pour la régression :
— **RandomForestRegressor**
— **GradientBoostingRegressor**
— **XGBRegressor**

4.2 Rééquilibrage des classes

Nous avons testé trois techniques :
— **Random Oversampling**
— **SMOTE**
— **Random Undersampling**

Chaque expérience combine :
— un modèle,
— une méthode de rééquilibrage,
— une validation croisée stratifiée.

Chapitre 5

Résultats : Classification

5.1 Performances globales

Modèle	**Accuracy**	**F1-macro**	**AUC**
RandomForest	0.64	0.62	0.89
SVM	0.63	0.61	-
XGBoost	0.66	0.65	0.91
Modèle	**Accuracy**	**F1-macro**	**AUC**
RandomForest	0.93	0.93	0.99
SVM	0.89	0.89	-
XGBoost	0.95	0.95	0.99

TABLE 5.1 – Performances des modèles pour la classification de wage_cat(haut) et value_cat(bas).

Les résultats montrent que les trois modèles atteignent de très bonnes performances sur la classification de la variable **value_cat**. Cependant, la variable **wage_cat** est plus difficile à prédire ; nous observons des performances moyennes, mais avec des scores **AUC** assez élevés montrant que les modèles arrivent quand même à bien séparer les classes deux à deux. Ce constat nous a menés sur la piste du rééquilibrage de classes, car comme nous l'avions évoqué dans 3, la variable **wage_cat** a un déséquilibre de classes, ce qui peut expliquer nos observations.

5.2 Analyse détaillée

Classe	XGBClassifier			RandomForest			SVM		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
very_low	0.80	0.79	0.80	0.77	0.78	0.78	0.78	0.75	0.76
low	0.55	0.58	0.56	0.52	0.54	0.53	0.51	0.53	0.52
medium	0.63	0.72	0.67	0.62	0.70	0.66	0.58	0.73	0.65
high	0.55	0.36	0.44	0.51	0.34	0.41	0.50	0.29	0.37
elite	0.77	0.81	0.79	0.76	0.79	0.78	0.76	0.77	0.76

TABLE 5.2 – Performances détaillées des modèles pour la classification de wage_cat.

Nous avons analysé les métriques par classes et avons constaté que les classes extrêmes (**elite** et**very_low**) sont mieux classées, avec généralement une précision supérieure, ce qui s'explique par le fait que ces deux classes ont des caractéristiques bien tranchées et une faible variance intra-groupe. Quant aux autres classes, elles sont généralement classées avec une précision inférieure à celle des classes extrêmes, ce qui peut s'expliquer par le fait que ces classes présentent des caractéristiques proches et une forte variance interne. Ce phénomène est connu comme **"class confusion in ordinal bands"**.

5.3 Impact du rééquilibrage

Rebalancer	Modèle	F1-macro mean
RandomOverSampler	XGBClassifier	0.657168
RandomOverSampler	RandomForest	0.641405
RandomOverSampler	SVM	0.621253
SMOTE	XGBClassifier	0.656082
SMOTE	RandomForest	0.643271
SMOTE	SVM	0.621328
RandomUnderSampler	XGBClassifier	0.653568
RandomUnderSampler	RandomForest	0.636371
RandomUnderSampler	SVM	0.617001

TABLE 5.3 – Performances des modèles avec différentes techniques de rééquilibrage pour la classification de wage_cat.

Après utilisation de techniques de rééquilibrage, nous avons observé une amélioration négligeable du score **F1**, ce qui est très insignifiant. Ce comportement est souvent observé sur des problèmes fortement ordonnés où les frontières entre classes médianes restent difficiles à séparer même après rééquilibrage.

Chapitre 6

Résultats : Régression

6.1 Performances des modèles

Modèle	MAE	RMSE	R²
XGBRegressor	7.461	34.313	0.0035
Gradient Boosting	8.306	29.584	0.2592
RandomForestRegressor	6.557	31.952	0.1359

TABLE 6.1 – Performances des modèles de régression pour la prédiction de wage.

Modèle	MAE	RMSE	R²
XGBRegressor	158622.31	946131.63	0.9885
Gradient Boosting	240622.07	740714.01	0.9930
RandomForestRegressor	156709.24	776732.15	0.9923

TABLE 6.2 – Performances des modèles de régression pour la prédiction de value.

Les performances des modèles de régression sur la variable **wage** sont très insatisfaisantes, avec des scores R^2 extrêmement faibles (voir table 6.1). Ces résultats médiocres sont attendus, reflétant les difficultés observées lors de la phase de classification.
Quant à la prédiction de la variable **value**, les modèles affichent d'excellentes performances avec un score R^2 moyen de 0.98. *XGBRegressor* présente une légère erreur moyenne absolue de 158622,30 €, tandis que les autres modèles offrent également d'excellentes performances.
Malgré les bons résultats sur **value**, nous pensons qu'une optimisation des hyperparamètres et une sélection plus rigoureuse des features pourraient améliorer davantage les performances globales, notamment pour la variable **wage**.

Chapitre 7

Réponses aux questions

7.1 Prédiction du prix (value) d'un joueur

Comme nous l'avons montré dans la section 6.1, il est tout à fait possible de prédire le prix d'un joueur. Cependant cette prédiction est sujette aux erreurs (en moyenne 156709,24 €).

7.2 Prédiction de score de joueurs

Au vu des résultats, il est possible de prédire la valeur marchande d'un joueur et sa catégorie, mais pour ce qui est du salaire, les résultats sont mauvais pour pouvoir faire une prédiction.

7.3 Analyse comparative

L'analyse comparative (fig 7.1) montre que les déterminants de la valeur marchande ne sont pas les mêmes pour l'ensemble des joueurs et pour les joueurs français uniquement. Le marché global est quasi exclusivement dominé par les variables `release_clause_num` et `overall_rating`, alors que le sous-groupe français est davantage influencé par des caractéristiques spécifiques telles que le `potential`, le `passing`, la `réputation internationale` et le `physique`. Nous notons quand même que certains sont communs notamment `release_clause_num` et `overall_rating`.

7.4 Recomandation

Nos tests de prédiction montrent que certains des critères évoqués, notamment **Overall Rating** et **International Reputation**, font effectivement partie des déterminants importants de la valeur des joueurs.

En revanche, nos analyses indiquent également que des variables telles que **Weak Foot** et **Skill Moves** ont un pouvoir prédictif très faible dans l'explication de la valeur d'un joueur. Elles sont présentes mais restent marginales par rapport aux facteurs majeurs.

Par ailleurs, nous observons que plusieurs variables non évoquées, en particulier la **release clause**, le **potential** jouent un rôle bien plus déterminant dans nos modèles.

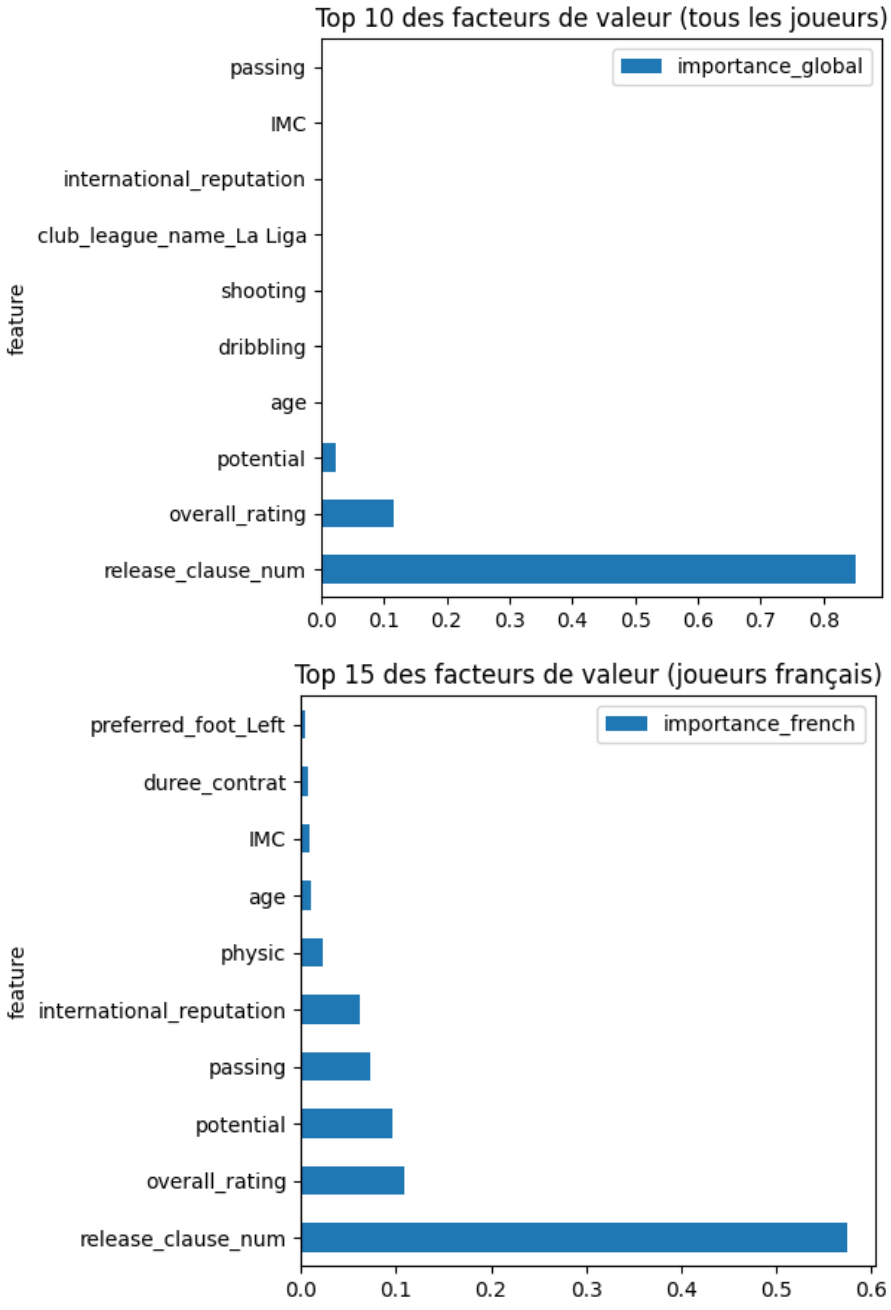

FIGURE 7.1 – Importance des caractéristiques