Compte rendu de projet

Yassine Assila et Côme Rodriguez

4 juillet 2022

Résumé

Voici le rapport du projet de SY09. L'objectif de ce projet est d'appliquer les techniques et les méthodes présentées dans SY09 à un jeu de données réel.
Dans ce rapport, nous expliquerons les méthodes que nous avons mises en place afin de prédire le salaire d'un individu à partir des données provenant du sondage *Ask a manager salary survey*. Nous considérerons deux stratégies différentes :

1. Prédire le salaire d'un individu directement à partir d'un apprentissage sur les données brutes
2. Prédire le niveau de salaire d'un individu (variable établi comme le rapport entre le salaire d'un individu et le salaire moyen du pays où exerce l'individu), puis d'utiliser cette prédiction comme descripteur pour estimer le salaire de l'individu en question

Nous comparerons et conclurons sur les résultats de ces deux méthodes, puis nous parlerons des améliorations possibles pour réduire l'erreur de prédiction et ainsi améliorer les performances de notre modèle de régression.

1 Introduction

Pour le projet de SY09 nous avons choisi le jeu de données suivant : *Ask a manager salary survey*.
Ce jeu de données provient d'un sondage réalisé par le site *Ask a Manager*. Il contient 18 variables différentes et 26232 réalisations de ces variables.
Dans ce projet, nous avons tenté de répondre au problème suivant :
Si un individu souhaite répondre au sondage mais ne souhaite pas divulguer son salaire annuel, pourrions-nous prédire son salaire annuel à partir des autres informations demandées dans le sondage ?
Nous effectuerons dans un premier temps une analyse exploratoire des données afin de mieux connaître notre jeu de données. Suite à cette exploration, nous prétraiterons certaines variables puis nous établirons un premier modèle de régression sur le salaire d'un individu qui servira de *baseline*. Une fois ce premier modèle établi, nous essayerons, via des méthodes d'analyse de données non supervisées, de définir une nouvelles variables qui nous permettraient d'améliorer les performances de notre *baseline*.

2 Analyse exploratoire des données et pré-traitement

Le jeu de données contient 18 variables différentes. Parmis ces variables, 16 sont qualitatives tandis que 2 sont quantitatives (*annual_salary* et *other_monetary_comp*). Notre analyse exploratoire nous montre que :
Les individus qui ont répondu à ce sondage sont majoritairement des femmes et ont majoritairement entre 25 ans et 44 ans. Les années d'expérience des individus varient entre moins d'un an et plus de 41 ans avec la majeur partie des individus ayant entre 2 et 20 ans d'expériences.
Certaines variables telles que *state*, *city*, *race* ou encore *job_title* ont des modalités très déséquilibrées ou alors des modalités compliquées à traiter avant de pouvoir les utiliser. Nous n'inclurons donc pas ces variables dans nos analyses futures. De même, la variable *annual_salary* présente des valeurs très extrèmes, nous n'incluerons pas ces valeurs dans nos analyses. Enfin, nous remarquons que la majorité des individus ayant répondu au sondage ont obtenu un *College degree* et un *Master's degree*.
D'autres variables sont présentes dans le jeu de données (comme la variable *country* ou encore la variable *currency*). Nous nous en sommes servies afin de créer la variable *adjusted_annual_salary* qui est une indication du niveau de salaire de l'individu :

$$\frac{harmonized\_salary\ (annual\_salary\ en\ USD)}{salaire\ moyen\ du\ pays}$$

.
Notre problème étant la prédiction du salaire d'un individu, nous concentrons notre analyse exploratoire sur la relation entre le salaire et les autres variables présentées. Le salaire d'un individu semble varier en fonction

de son genre, de son expérience, de son âge et de son diplôme.
Nous avons pré-traité la variable *industry* car celle-ci présentait trop de modalités différentes pour exprimer le même secteur. Nous avons suivi la méthode présentée dans l'article de Lily Wu, puis nous avons gardé les secteurs d'activités ayant au moins 500 représentations dans le jeu de données. Nous obtenons finalement 13 secteurs d'activité différents (au lieu de 1068 au départ). Ce traitement peut contenir certaines erreurs, que nous considérerons négligeables dans nos analyses. Comme le montre la figure 1, cette variable semble avoir un effet sur le salaire d'un individu.

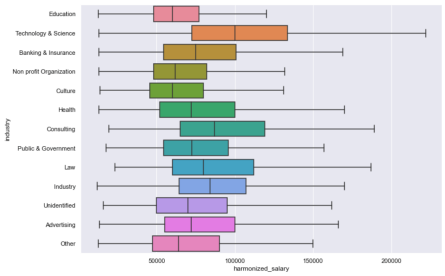

FIGURE 1 – Répartition des salaires en fonction du secteur d'activité

3 Premier modèle de régression sur le salaire

Le premier modèle de prédiction vise à établir une *baseline* pour pouvoir mesurer l'utilité de nos modifications futures.
Le critère à minimiser est la racine carré de l'erreur quadratique moyenne (RMSE) dont voici la formule :

$$\sqrt{\frac{\sum_{n=1}^{N}(\hat{y}_n - y_n)^2}{N}}$$

Nous cherchons donc :

$$Y = f(X) \text{ avec } \begin{cases} Y = harmonized\_salary \\ X = \text{Vecteurs de descripteurs} \end{cases}$$

Suite à notre analyse exploratoire, considérons comme descripteurs les variables *gender*, *diploma*, *age*, *industry* et *experience*. Les variables *age*, *experience* et *diploma* sont transformées en variables de niveau (car il existe un ordre entre les modalités) et les variables *gender* et *industry* sont transformées en *dummy variables*.

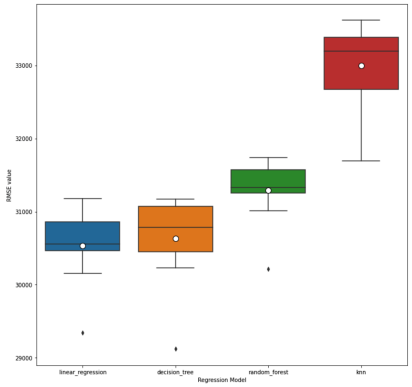

FIGURE 2 – Résultats d'une validation croisée de 10 plis pour sélectionner le modèle minimisant la valeur du RMSE

Si nous nous réfèrons à la figure 2, la régression linéaire semble être le modèle minimisant la valeur du RMSE, et donc le modèle qui prédit avec le plus de précision le salaire d'un individu, parmis les modèles présentés.
La régression linéaire n'est pas un modèle adapté pour l'apprentissage avec notre jeu de données. Les variables étant catégorielles, nous nous attendions à ce que les arbres de décision et les forêts aléatoires soient plus précis, car plus adaptés au type de données que nous avons. Néanmoins, la régression linéaire semble donner de bons résultats de par le fait que nos données contiennent une relation monotone entre nos descripteurs et le salaire à prédire : plus un individu est agé, plus il a des années d'expérience et plus son diplôme est élevé, plus son salaire sera élevé. Nous pouvons donc choisir la régression linéaire comme premier modèle d'estimation de salaire d'un individu.

3.1 Sélection de variables

Sélectionnons les variables qui, selon notre *baseline*, sont les plus importantes pour la régression sur le salaire des individus. Cette sélection de variables est justifiée par le fait que suite à nos transformations, nous avons un grand nombre de variables explicatives, dont certaines ne sont peut être pas pertinantes et vont affecter les capacités de généralisation de notre modèle. De plus, certaines variables (comme l'âge et l'expérience dans un domaine) peuvent être corrélées, ce qui rajoute de l'information redondante et augmente le risque d'*overfitting*. Nous utilisons la classe *SelectFromModel* de *sklearn* et obtenons le sous-ensemble de variables suivant : *experience*, *diploma* qui sont des variables de niveau et *non_ binary* (genre) *Consulting*, *Culture*, *Education*, *industry_ Industry*, *Non profit organization*, *Tech-*

nology & Sciences (secteurs) qui sont des variables binaires.

3.2 Régression

La valeur de RMSE obtenue après apprentissage sur les données de validation est de 30 465. Cette première valeur nous servira de valeur de comparaison pour les analyses futures.

4 Analyse non supervisée

Une fois notre *baseline* établie, analysons notre jeu de données avec des méthodes d'analyse non supervisée. Nous cherchons spécialement à déterminer la relation entre le niveau de salaire d'un individu et les autres variables le décrivant. Nous discrétisons la variable *adjusted_annual_salary* de la manière suivante :

$$\begin{cases} \textit{bad payed} \text{ si } \textit{adjusted\_annual\_salary} < 1 \\ \textit{well payed} \text{ sinon} \end{cases}$$

Autrement dit, si un individu a un salaire inférieur au salaire moyen de son pays, il a un niveau de salaire considéré comme bas, sinon il est considéré comme ayant un bon salaire.
La principale difficulté de l'analyse non supervisée sur ce jeu de données est la forme du jeu de données : les variables sont très généralement catégorielles.
Premièrement, nous chercherons à analyser les relations entre nos différentes variables. Pour ceci, nous appliquerons l'analyse des correspondances multiples. Enfin, nous essayerons de regrouper nos individus dans différents clusters. Comme notre jeu de données contient des variables catégorielles, nous ne pouvons appliquer l'algorithme des *KMeans* (qui se base sur les distances entre individus). Nous appliquerons à la place l'algorithme des *KModes*, qui se base lui sur les dissimilarités entre individus.

4.1 Analyse des correspondances multiples

4.1.1 Synthèse de la méthode

L'analyse des correspondances multiples est une méthode d'analyse factorielle adaptée aux données catégorielles. L'objectif est d'étudier le lien entre les différentes variables considérées.
Cette méthode s'appuye sur l'utilisation d'un tableau disjonctif complet. Ce tableau est de la forme $I * J$ où I est le nombre d'individus et J est le nombre total de modalités présentes dans les données (si nous avons 3 variables catégorielles pouvant prendre chacune 5 modalités, nous avons alors J=15). Les cases (i, j) ne peuvent prendre comme valeur que 1 ou 0 :

$$\begin{cases} 1 \text{ si l'individu i possède la modalité j} \\ 0 \text{ sinon} \end{cases}$$

L'analyse des correspondances multiples permet d'observer les relations entre les individus et les relations entre les modalités : plus deux individus sont similaires, plus ils sont proches dans la projection sur les axes factoriels, à l'inverse, plus deux individus sont différents, plus ils sont éloignés dans la projection sur les axes factoriels.
Pour les modalités, nous distinguons deux interprétations différentes :

1. Si les deux modalités considérées proviennent de deux variables différentes : plus les modalités ont tendance à apparaître en même temps, plus ces modalités sont proches dans la représentation des modalités sur les axes factoriels
2. Si les deux modalités considérées proviennent de la même variable : plus les individus comportant ces deux modalités sont semblables (en excluant la variable comportant les deux modalités considérées), plus ces modalités seront proches dans la représentation des modalités sur les axes factoriels

Nous pouvons également noter que plus une modalité est rare, plus celle-ci sera éloignée du centre de gravité du nuage des modalités.

4.1.2 Application de l'analyse des correspondances multiples au jeu de données

Nous appliquons l'analyse des correspondances multiples grâce à la classe *mca* de la librairie python *prince*. En considérant les variables *age*, *experience*, *diploma*, *gender*, *industry* et le niveau de salaire, nous obtenons une représentation de notre jeu de données dans un espace continu. Les inerties expliquées par les axes factoriels sont montrées sur la figure 3 :

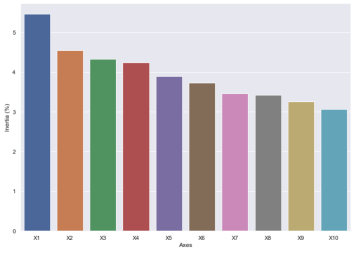

FIGURE 3 – Inerties expliquées par les 10 premiers axes factoriels (40% de l'inertie totale)

La projection du nuage de points sur les deux premiers axes factoriels est représentée sur la figure 4.

FIGURE 4 – Représentation du nuage de points sur les deux premiers axes factoriels

La distribution des modalités sur le premier plan factoriel est montrée dans la figure 5 :

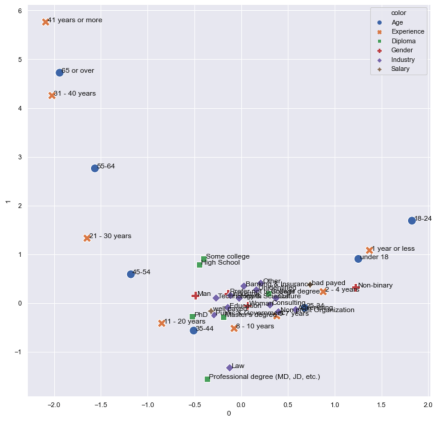

FIGURE 5 – Distribution des modalités selon le premier plan factoriel

Le nuage de points a une allure générale de fer à cheval. Cette allure témoigne d'un effet Guttman : l'axe principal oppose les extrêmes (dans notre cas les individus jeunes des individus âgés, les individus ayant peu d'expérience à ceux ayant beaucoup d'expérience). Cet effet s'explique du fait que nos modalités contiennent en réalité une notion d'ordre. Ce sont des variables quantitatives qui semblent avoir été discrétisées dans le sondage.
Le niveau de salaire des individus semblent être réparti selon la courbure du nuage : les individus situés au plus haut du nuage de points semblent avoir un niveau de salaire généralement bas, tandis que les individus situés en dessous de cette courbure semblent généralement avoir un bon niveau de salaire.
Si nous analysons la représentation des modalités dans le premier plan factoriel, nous nous apercevons premièrement que les tranches d'âges élevées et faibles (*under 18, 18-24* et *55-64, 65 or over* sont rares, car éloignées du centre de gravité), l'analyse est la même pour les tranches d'années d'expérience.
De manière générale, nous remarquons avec la représentation des modalités dans le premier plan factoriel que les différents niveaux de salaires sont relativement proches, ce qui témoigneraient d'une certaine difficulté à les différencier. Cette analyse se confirme avec la représentation des individus : des individus de niveau de salaire différents sont plus proches que des individus de même niveau de salaire. Essayons de voir si cette interprétation est vérifiée lorsqu'on essaye de regrouper nos individus dans des clusters, en appliquant l'algorithme des *KModes*.

4.2 KModes

4.2.1 synthèse de l'algorithme

A l'instar des *KMeans*, cet algorithme vise à répartir les individus dans un nombre de clusters préalablement défini. Il s'applique aux données catégorielles.
L'objectif est de minimiser les dissimilarités entre les individus d'un même cluster, et donc par conséquent, de maximiser les dissimilarités entre les individus de clusters différents.
L'initialisation de l'algorithme est la même que celle de l'algorithme des *KMeans* : nous choisissons aléatoirement k centres (k étant le nombre de clusters que nous désirons former). Une fois ces centres initialisés, nous comparons chaque individu au centre de chaque cluster (chaque variable est comparée une à une entre l'individu et le centre) et affectons l'individu au cluster présentant le centre avec le moins de dissimilarités.
Une fois tous les individus du jeu de données affectés à un cluster, les centres sont recalculés de la manière suivante : chaque variable prend la modalité la plus présente dans le cluster considéré.
Une fois le calcul des centre effectué, nous refaisons une itération d'affectation des individus au cluster minimisant les dissimilarités entre l'individu et le centre.
L'algorithme s'arrête lorsque aucun individu est affecté à un autre cluster après le calcul des nouveaux centres.
Nous pouvons alors calculer le coût total, qui est le nombre total de dissimilarités entre les individus et les centres des clusters où ils sont affectés.
Comme les *KMeans*, le résultat de cet algorithme dépend de l'initialisation. C'est donc pour cela que nous

veillons en général à effectuer plusieurs fois l'algorithme avec des initialisations différentes afin de converger vers k clusters optimaux.

4.2.2 Application des KModes au jeu de données

Nous utilisons la classe *KModes* de la librairie python *kmodes*. Pour déterminer le nombre de clusters que nous souhaitons considérer, nous appliquons la méthode du coude en conservant les variables *gender*, *industry*, *diploma*, *age* et *experience*. Pour chaque nombre de clusters (de 2 à 15), nous faisons 15 initialisations différentes et conservons l'initialisation présentant le coût le plus faible (voir figure 6).

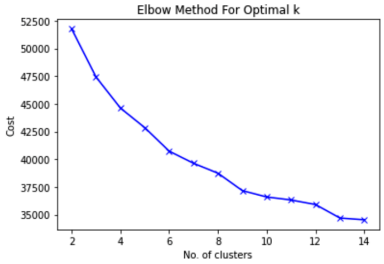

FIGURE 6 – Evolution du coût des *KModes* en fonction du nombre de clusters considéré

Suite à cette analyse, nous décidons de considérer 6 clusters différents. Nous obtenons les résultats suivants : L'algorithme semble avoir affecté une majorité des individus aux clusters 1 et 2 (comme montré sur la figure 7) :

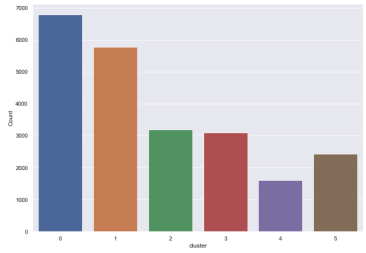

FIGURE 7 – Répartition des individus dans les clusters

La répartition qui nous intéresse le plus est la répartition des niveaux de salaires dans les différents clusters. Si l'on se réfère à la figure 8, nous remarquons que les niveaux de salaire bas sont majoritairement présents dans les trois premiers clusters, et que les bons niveaux de salaires sont majoritairement présents dans les clusters 1, 2 et 4. Cette répartition confirme les résultats de l'analyse des correspondances multiples, à savoir qu'il est difficile de trouver une distinction des individus en fonction de leur niveau de salaire si nous tentons de les regrouper en utilisant leurs autres informations données dans le sondage.

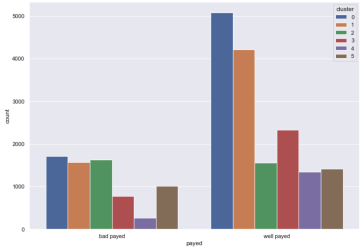

FIGURE 8 – Répartition des niveaux de salaire dans les clusters

Comme nous pouvons le remarquer, ce clustering n'est que peu utile au regard de notre objectif. Nous n'utiliserons donc pas ce résultat dans nos analyses futures. Considérons maintenant 2 cas différents :
Dans un premier temps, supposons que les individus pour lesquels on souhaite prédire le salaire aient déclaré leur niveau de salaire. Nous ajouterons un nouveau descripteur à notre modèle de régression, le niveau de salaire des individus, afin d'estimer leur salaire.
Dans un second temps, considérons le cas où les individus ne souhaitent pas communiquer d'informations à propos de leur salaire. Nous essayerons de prédire le niveau de salaire d'un individu, puis nous appliquerons notre modèle de régression sur le salaire de l'individu en ajoutant comme descripteur le niveau de salaire prédit par notre classifieur. Nous comparerons les résultats à ceux de notre *baseline* pour conclure.

5 Régression sur le salaire d'un individu en utilisant le cluster et le niveau de salaire

Pour prédire le salaire d'un individu, nous utilisons le même modèle que nous avons utilisé section 3 en ajoutant comme descripteur le niveau de salaire de l'individu. Nous observons que la valeur du RMSE baisse (30 465 en moyenne à 25 573 en moyenne). L'introduction de la variable du niveau de salaire nous permet de mieux estimer le salaire d'un individu, ce qui semble logique car le niveau de salaire est une variable issue du salaire de l'individu. Si un individu souhaite communiquer son niveau de salaire, cette variable semble indispensable pour la prédiction de son salaire. Intéressons

nous maintenant à notre capacité à estimer le niveau de salaire d'un individu.

6 Classification du niveau de salaire d'un individu

Nous cherchons ici à estimer le niveau de salaire, nous cherchons donc :

$$Y = f(X) \text{ avec } \begin{cases} Y = salary\_level \\ X = \text{Vecteurs de descripteurs} \end{cases}$$

Les descripteurs que nous avons choisis sont les suivants : *age*, *experience*, *diploma*, *industry*, *gender* et *cluster*.

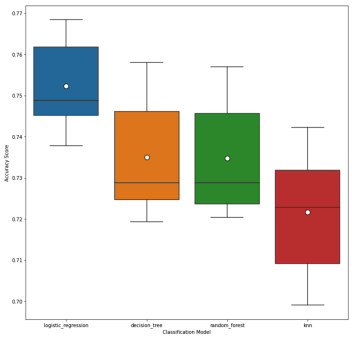

FIGURE 9 – Résultats des performances des modèles suite à une validation croisée de 10 plis pour sélectionner le modèle maximisant *l'accuracy score*

Les résultats de la validation croisée, présentés figure 9, montrent que la régression logistique est le modèle qui maximise l'*accuracy score* ($accuracy = \frac{bonnes\ predictions}{predictions\ totales}$). Nous utilisons donc la régression logistique pour prédire le niveau de salaire d'un individu.

6.1 Sélection de variables

Nous utilisons la même logique présentée dans la section 3.1 et obtenons le sous-ensemble de variables suivant : *experience*, *diploma* (variables de niveau), *non_binary* (genre, binaire) et *Consulting*, *Culture*, *Education*, *Industry*, *Non profit organization*, *Other*, *Technology & Sciences* (secteurs d'activité, binaires).

6.2 Classification

La matrice de confusion résumant les performances de ce modèle est présentée figure 10 :

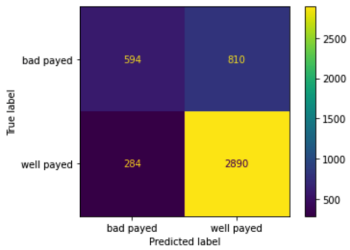

FIGURE 10 – Matrice de confusion du modèle de régression logistique

Nous remarquons que l'*accuracy score* de notre modèle est plutôt bon (75%), mais en nous intéressant à la capacité de celui-ci à prédire la modalité *bad payed*, nous remarquons qu'il ne semble pas bien prédire cette modalité. Vérifions cette hypothèse en nous intéressant au *F1 score*.
Le *F1 score* est donné par la formule suivante :

$$F1\ score = 2 * \frac{recall * precision}{recall + precision}$$

Le *recall* est une mesure donnant une indication sur la faculté du modèle de prédiction à bien prédire les positifs (par exemple, avec notre matrice de confusion, nous pouvons établir que le *recall* de la modalité *well payed* est de 0,91 : $\frac{2890}{2890+284}$. Notre modèle est donc bon dans la prédiction des bons niveaux de salaire).
La *précision* permet quand à elle d'avoir une information sur la capacité du modèle à minimiser les faux positifs. Par exemple, la précision de notre modèle sur la prédiction de la modalite *bad payed* est de 0,68 : $\frac{594}{594+284}$, ce qui signifie que si notre modèle prédit *bad payed*, il y a 68% de chance que la véritable modalité soit *bad payed*.
Nous nous intéressons au *F1 score* car cette mesure nous donne une réelle indication sur les performances de notre modèle, à savoir bien prédire les modalités et ne pas se tromper dans la prédiction.
Le *F1 score* de la modalité *well payed* est de 0,84. Notre modèle semble donc performant dans la prédiction de cette modalité.
En revanche, le *F1 score* de la modalité *bad payed* est de 0,52, ce qui signifie que notre modèle aura plus de mal à bien prédire cette modalité.
Ces observations peuvent être expliquées par le fait que la modalité *well payed* est sur-représentée par rapport à la modalité *bad payed*. De plus, les variables que nous utilisons pour prédire le niveau de salaire ne sont pas les seules variables permettant de différencier le niveau de salaire de deux individus différents. Par exemple, deux femmes ayant entre 25 et 34 ans, ayant obtenu le même diplôme, travaillant dans le même secteur d'ac-

tivité avec entre 5 et 7 ans d'expérience dans leur domaine peuvent avoir un salaire différent en fonction de leur poste.
Dans la section 5, nous avons vu qu'utiliser le descripteur *salary_ level* nous permettait d'améliorer la qualité de notre régression. Si nous tentons de réaliser une régression sur le salaire d'un individu en ajoutant comme descripteur le niveau de salaire prédit par notre modèle de classification, nous obtenons un RMSE de 32 830, ce qui est moins bon que notre *baseline*.
Se tromper dans la prédiction du niveau de salaire augmente l'erreur d'estimation du salaire d'un individu. Si nous souhaitons prédire avec plus de précision le salaire, il faut que nous améliorerions les performances de notre modèle de classification pour prédire le niveau de salaire.

7 Conclusion et pistes d'améliorations

Dans ce projet, nous avons essayé de prédire le salaire d'un individu à partir des données du sondage *Ask a manager salary survey*. Nous avons établi deux stratégies différentes : prédire le salaire d'un individu à partir des données qui nous sont communiquées, ou essayer de prédire le niveau de salaire d'un individu, puis prédire son salaire en incluant le niveau de salaire prédit.
Si le niveau de salaire est correctement prédit, ceci nous permet d'améliorer les performances de notre régression sur du salaire d'un individu. En revanche, les performances actuelles du modèle de classification ne sont pas assez bonnes pour nous permettre de réduire l'erreur de prédiction de salaire dans notre deuxième stratégie. Ce manque de performance peut venir du fait que :

1. Certaines variables qui pourraient expliquer les différences de salaire entre individus semblables ont été écartées de l'analyse (comme le pays de l'individu par exemple, ou bien son poste)
2. Des erreurs dans le pré-processing de la variable *industry* peuvent être présentes et peuvent baisser les performances du modèle. Par exemple, le secteur *Technology & Sciences* regroupe des sous secteurs comme la réparation d'ordinateurs et la recherche dans les biotechnologies, secteur n'ayant pas forcément les mêmes niveaux de salaires
3. Certaines modalités sont sur-représentées par rapport à d'autres dans le jeu de données

Pour améliorer les performances globales du modèle de prédiction de salaire, il pourrait être intéressant :

- d'inclure les variables ayant été écartées de l'analyse (en veillant à bien les traiter en amont).
- d'utiliser des techniques d'over (ou d'under) sampling (comme *SMOTE* par exemple) pour pallier le déséquilibre des modalités dans le jeu de données.

Références

[1] Hervé Abdi, Dominique Valentin (2007). Multiple Correspondance Analysis. *Disponible en ligne à cette URL.*

[2] Lily Wu (2021). Clustering Product Names with Python — Part 2. *Disponible en ligne à cette URL.*

[3] Audhi Aprilliant (2021). The k-modes as Clustering Algorithm for Categorical Data Type. *Disponible en ligne à cette URL.*

[4] Zhexue Huang (2008).. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Disponible en ligne à cette URL.*