Traitement des données manquantes dans le milieu bancaire

Ahmad Charaf,
Nexialog Consulting, Paris, France

26 octobre 2022

Les bases de données relatives au domaine bancaire et de l'assurance sont alimentées par des informations de nature contractuelle, des informations sur l'activité des clients, des enquêtes ou des questionnaires, ou encore des données macro-économiques. Ces données peuvent comporter un ensemble "d'anomalies" pouvant empêcher une prédiction ou une analyse réalisée sur la base d'un modèle statistique. Il peut s'agir de la présence de données manquantes, de valeurs extrêmes, de données censurées ou encore de variables catégorielles qu'il faut encoder.

Dans notre cas, on s'intéressera uniquement au traitement des données manquantes pour des variables statiques (ou pour lesquelles, une série temporelle n'est pas disponible). S'il est justifié, le traitement des données manquantes, peut apporter un véritable gain de performance sur la prédiction. En règle générale, la présence de valeurs manquantes dans un jeu de données est assez fréquente. Cela s'explique par diverses raisons comme des erreurs de saisies, un refus de partage d'information, l'oubli d'un opérateur, etc... Plus globalement les données manquantes se caractérisent dans le jeu de données par une valeur nulle ou la valeur NaN ou encore par une valeur extrême préalablement définie.
La première étape consiste à déceler la présence des données manquantes. Ensuite, il s'agit d'identifier la bonne typologie ou le mécanisme à l'origine de la donnée manquante, afin d'imputer cette dernière ou non.

La Section 1 présente les trois grandes typologies de données manquantes et discute dans ces trois cas de figure le biais potentiellement induit par l'omission des entrées correspondant aux données manquantes. La section 2 présente différentes approches d'imputation des valeurs manquantes, correspondant à différentes manières de compléter la donnée manquante à partir d'une méthode d'estimation. La section 3.1 présente l'application de ces différentes méthodes sur une base de données emprunteur, en présence de différentes proportions de données manquantes.

1 Typologie des données manquantes

En partant du principe que l'on a réussi à correctement identifier la présence des données manquantes, une importante question se pose. A partir de quelle proportion de données manquantes l'imputation induirait un biais trop important ? Même s'il n'y a pas de réponse tranchée et que la réponse dépend fortement du contexte de génération des données, plusieurs études ont toutefois proposé un seuil à partir duquel imputer empiéterait trop sur la validité de l'inférence statistique. On peut notamment citer Bennett (2001) qui montre de manière empirique qu'un niveau maximal de 10% de données manquantes peut être considéré. Ce seuil ne constitue en rien une vérité absolue surtout quand on sait que certains statisticiens penchent sur une possible imputation dépendant uniquement du mécanisme expliquant le manquement de la donnée. Dans une utilisation complémentaire, les règles propres à la proportion de données manquantes ainsi qu'à la typologie des données manquantes peuvent apporter plus de précision à notre analyse.

Comprendre l'origine de la donnée manquante est donc primordial pour identifier le traitement adéquat. Dans ce sens, Little and Rubin (1987) identifient 3 mécanismes qui sous-tendent la présence de données manquantes :

- MCAR (missing completely at random). Dans ce cas, l'absence d'information est entièrement liée à un effet aléatoire et cet effet est identique pour toutes les observations. Par ailleurs, l'absence ou non de la donnée pour une observation est totalement indépendante des autres variables. L'exemple le plus simple pour illustrer cette situation consiste à supposer qu'une valeur est renseignée à la suite d'un jeu de pile ou face, le manquement est ici totalement dû à un effet exogène. Cette situation est généralement considérée comme peu fréquente.
- MAR (missing at random). Moins contraignant que le cas MCAR, le MAR signifie que la probabilité d'avoir une donnée manquante pour une variable dépendant des autres variables observées dans la base. Un exemple assez intuitif, est celui d'un client possédant un revenu assez élevé, il aura alors le droit à une option particulière sur sa carte bancaire. Dans ce cas la présence ou non de l'option dépend entièrement du revenu du client.
- MNAR (missing not at random). Dans ce cas, la probabilité de données manquantes dépend de variables qui n'ont pas été observées dans le jeu de données. Prenons l'exemple d'un cadre gagnant un haut revenu et ne souhaitant pas le dévoiler, si la

catégorie socio-professionnelle n'est pas présente dans notre jeu de données, les données manquantes dépendront d'un effet extérieur à la base de donnée, ici une variable non-observée.

Les travaux de Little and Rubin (1987) sont très importants pour définir la meilleure démarche possible dans le traitement des données manquantes. Rappelons qu'une connaissance accrue de la base de données, de sa construction et de son alimentation est requise pour classer nos variables selon les typologies identifiées précédemment. Enfin, les traitements à appliquer selon la typologie sont les suivants :

- Pour MCAR, les données manquantes réduisent la taille de la population qui peut être analysée et donc le pouvoir de prédiction statistique qui en découle, mais l'omission des entrées présentant une incomplétude de données n'induit pas de biais. En effet, la sous-base complète peut être considérée comme un échantillonnage aléatoire de la base initiale et par conséquent, elle peut être considérée comme ayant la même distribution. Si la proportion des données manquantes est faible, il suffit dans ce cas de retirer les lignes présentant un manque d'information. L'hypothèse de données manquantes de type MCAR est généralement considérée comme forte et irréaliste.
- Pour MAR, omettre les données manquantes revient à supprimer de l'information car la probabilité d'absence d'une donnée manquante dépend d'une autre variable. Dès lors, cette valeur manquante concentre une information pertinente. Un biais est donc créé par la suppression de ces données-là. L'imputation dans ce cas-là est fortement recommandée.
- Pour MNAR, la présence de données manquantes est la conséquence d'un effet extérieur à la base ou de variables non-observées. Dans le cas où cet effet est indépendant des variables de la base, la suppression des données manquantes n'induira pas de biais sur l'estimation. La suppression de la variable peut être considérée en particulier si la proportion de données manquantes est élevée. Dans le cas contraire, une méthode d'imputation peut être envisagée à partir des autres variables, car ces dernières portent une information sur le mécanisme ayant causé la perte de données.

2 Traitement des données manquantes

Une fois que l'origine de la donnée manquante a été identifiée et qu'une imputation est requise, plusieurs approches de complétion de la donnée peuvent être considérées. En effet, la méthode à retenir dépend des caractéristiques de la base de données et des relations intrinsèques entre les variables. De plus, certaines méthodes performent mieux selon le type de la variable à imputer, qu'elle soit quantitative ou qualitative comme on le verra.

Tout d'abord pour les méthodes les plus simples on peut avoir recours à une simple suppression des données manquantes, cette méthode s'applique vraiment quand la proportion de données manquantes est très faible et qu'il y aura peu d'incidence à cette décision. La méthode consiste à supprimer les lignes comportant les données manquantes. Cependant, cette méthode n'est clairement pas à privilégier, mieux vaut utiliser une imputation par valeur unique pour une action rapide.

De plus, pour toute imputation statistique que l'on réalisera, il sera intéressant d'ajouter une variable binaire à notre base de données encodée en 1 si l'observation s'est vue être imputée ou 0 sinon, cela, pour ne pas perdre l'information sur le caractère manquant de la donnée.

Notons que les méthodes présentées ci-dessous s'utilisent dans le cas de variable continu comme dans le cas de variable discrète même si l'utilisation reste souvent plus évidente pour une utilisation dans le cadre continu.

2.1 Imputation simple

La première forme d'imputation des données manquantes est l'imputation par une valeur unique. Ce cas de figure consiste à remplacer les données manquantes par la moyenne, la médiane ou encore dans un cadre discret la valeur la plus fréquente de la variable concernée. La Figure 1 ci-dessous illustre une imputation par la moyenne

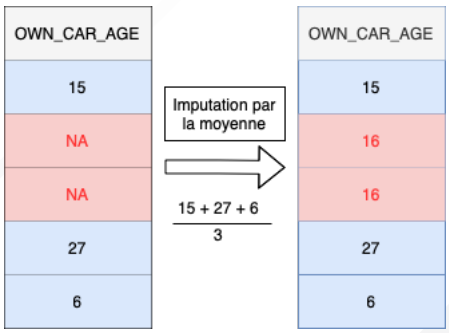

FIGURE 1 – Imputation par valeur unique

Cette méthode est relativement simple d'utilisation et rapide à mettre en place. Elle peut convenir pour un nombre très faible de données manquantes. Cependant, elle cause l'inclusion d'un grand nombre de valeurs uniques pour une même variable, ce qui peut, pour une grande proportion de données manquantes, modifier la corrélation entre les variables et biaiser nos modèles d'estimation. Par ailleurs, la valeur incluse n'est pas nécessairement représentative de la distribution de la variable concernée. Par exemple, la moyenne peut être un mauvais choix contrairement à la médiane ou le mode pour représenter l'échantillon, notamment, en présence de valeurs extrêmes.

D'autres méthodes d'imputation statistique existent et permettent une plus grande précision des valeurs im-

putées. Ces méthodes sont plus complexes que l'imputation par valeur unique et sont basées sur l'apprentissage automatique.

2.2 Les K plus proches voisins (KNN)

Ici, on part de l'hypothèse qu'il existe des clusters d'individus dans notre échantillon (entendre ici, des sous-populations d'individus ayant des caractéristiques proches). Si on prend par exemple la variable montant du crédit (AMT_CREDIT) qui a une donnée manquante pour l'individu 1, alors cette méthode nous permettra de trouver les K autres individus à données non-manquantes qui ressembleront le plus en termes de caractéristique globale à notre individu 1. De manière plus précise, une moyenne pondérée des valeurs de la variable montant du crédit de nos K observations les plus proches sera faite et imputée à la donnée manquante de la variable montant du crédit dans l'observation 1. Ainsi, plus la ressemblance des plus proches voisins est forte, plus leur contribution dans la moyenne pondérée est importante. Enfin la similarité ou la ressemblance se mesure généralement par la distance euclidienne. La donnée remplacée est la même pour l'ensemble des individus d'une même classe k.

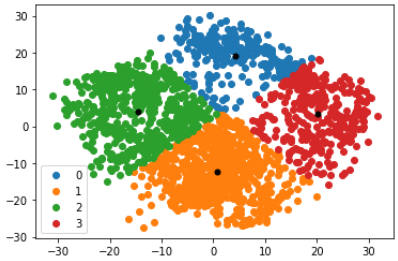

FIGURE 2 – Imputation par K plus proches voisins

Cette méthode est bien plus élaborée que l'imputation simple car elle tient compte des autres variables et des caractéristiques communes des observations de notre base de données. De plus les KNN obtiennent de bons résultats pour des variables qualitatives comme quantitatives. Cependant, le nombre de k clusters est à définir au préalable. Or, comme le choix d'un hyperparamètre, ce choix influera sur le clustering final et donc sur notre imputation. On pourrait proposer une méthode appliquée à nos données où l'on minimisera une fonction de perte pour chaque k nombre de cluster, mais cela prendra énormément de temps à concevoir spécialement dans le cas des KNN. Notons aussi que des pré-traitements seront à prendre en compte avant l'utilisation des KNN. En effet, si le critère de similarité est défini par une distance euclidienne, une normalisation ou une standardisation des variables continues (une mise à l'échelle donc), un encodage pour les données catégorielles, ou encore une transformation logarithmique pourront être appliqués afin d'assurer les capacités prédictives des KNN (pour plus de détails, voir par exemple Troyanskaya et al. (2001)). De plus, l'efficacité de l'algorithme peut être fortement affectée en présence de multicolinéarité dans nos données, c'est à dire en présence de forte corrélation entre variables.

2.3 Miss Forest

Introduite par Stekhoven and Bühlmann (2012), la méthode des MissForest vise à estimer les données manquantes à l'aide de l'algorithme ensembliste Random Forest. La méthode se construit de manière itérative et accroît sa précision à chaque itération. Une étape préliminaire consiste à fixer une première imputation basée sur une valeur unique (type moyenne/médiane/mode).

L'algorithme commence par séparer les données en deux parties, un échantillon d'entraînement composé des données non manquantes et un échantillon de prédiction composé des données manquantes du jeu de données. Un modèle Random Forest est ensuite lancé sur ces échantillons et les prédictions faites sur les valeurs manquantes sont considérées pour une éventuelle imputation.

Ensuite, l'algorithme décide ce qu'il doit encore prédire, c'est-à-dire si l'imputation faite sur une donnée manquante lors de la première itération est satisfaisante ou non. Si c'est le cas, la donnée complétera l'échantillon d'entraînement, sinon, elle reste dans l'échantillon de prédiction. Le processus d'estimation est itéré jusqu'à ce que l'écart relatif entre sorties d'imputations à l'itération actuelle et sorties d'imputation à l'itération précédente augmente pour la première fois.

L'imputation finalement générée par la méthode sera celle précédemment calculée. Alors la donnée se verra être ajoutée à l'échantillon d'entraînement composé des autres données initialement non manquantes pour accroître la qualité de la prédiction pour les données qui n'ont pas obtenu des résultats suffisamment bons [1]. Enfin, l'opération prend fin quand le nombre d'itérations atteint un niveau seuil ou selon un critère d'arrêt. La Figure 3 illustre l'algorithme d'imputation pour 3 itérations.

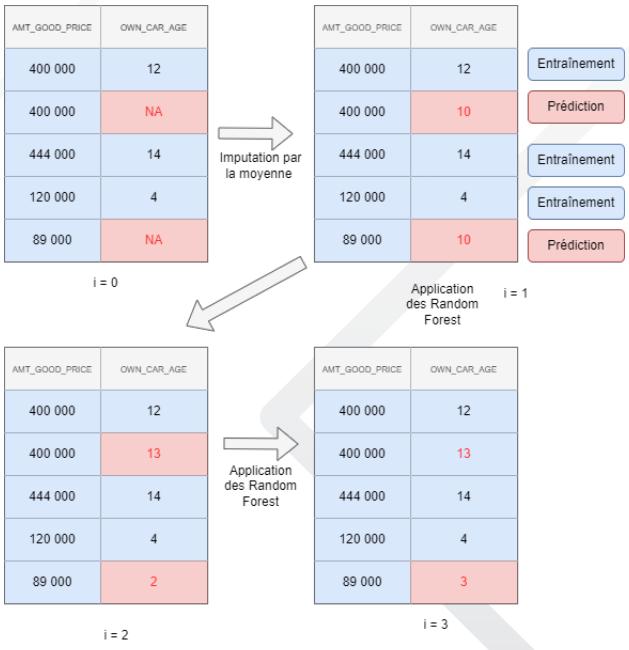

FIGURE 3 – Imputation par Miss Forest

1. Source : Numpyninja- Miss Forest

Contrairement aux KNN, les algorithmes MissForest ne nécessitent aucun pré-traitement sur les données, comme une mise à l'échelle ou un encodage des variables catégorielles par exemple. L'algorithme reste aussi performant en cas de multicolinéarité et de bruit intrinsèque. Par ailleurs, l'algorithme des MissForest qui repose sur les Random Forest est intrinsèquement non-paramétrique. L'imputation par MissForest reste moins coûteuse en temps de calcul que les KNN sur de larges jeux de données par exemple. Cependant, sur une volumétrie de données plus réduite, la complexité des MissForest peut être relativement élevée par rapport aux KNN. D'autres méthodes peuvent être plus intéressantes en terme de coût de calcul.

2.4 Imputation multiple avec MICE

MICE est une méthode d'imputation multiple, elle signifie *Multivariate Imputation via Chained Equations*. Cette méthode permet de chercher la distribution conditionnelle de chaque variable à données manquantes plutôt que de chercher à avoir la distribution conditionnelle globale (voir par exemple Stavseth et al. (2019)). Ceci permet à MICE d'être plus précis et plus flexible dans le sens ou l'imputation se fait variable par variable (voir Van Buuren (2007)). Cette méthode se base sur l'utilisation d'un algorithme de régression pour les variables continues et d'un algorithme de classification pour les variables discrètes [2].

La méthode se construit de manière itérative. Dans un premier temps les valeurs manquantes issues des variables à données manquantes sont remplacées par une valeur unique simple (type moyenne/médiane/mode). Supposons qu'à ce stade 3 variables sont concernées par des données manquantes dans notre jeu de données, soit v_1, v_2, v_3 ne contenant aucune donnée manquante car remplacée par une valeur unique. La seconde étape de MICE vise à faire revenir à l'état initial avec données manquantes une de nos 3 variables concernées par les données manquantes, soit v_1 uniquement contenant des données manquantes. Une régression est alors appliquée sur notre jeu de données avec comme seule variable cible v_1, en omettant cette fois les observations comportant des données manquantes (pour rappel, on ne peut lancer une régression en présences données manquantes). A la suite de cela, un modèle de régression sera créé puis utilisé pour estimer les données manquantes de la variable v_1. L'action devra être répétée pour chaque variable à données manquantes.

Il s'agit d'une méthode assez modulable dans le sens où différents modèles de régression peuvent être utilisés selon la nature des données. En effet on peut utiliser une régression logistique ou encore des Random Forest dans l'imputation grâce à cet algorithme. Cependant, l'ordre selon lequel les variables à données manquantes sont retenus dans l'algorithme peut avoir un impact sur la distribution jointe des données et sur la performance de prediction. Certaines études comme Drechsler and Rässler (2008) montrent qu'on peut s'attendre à ce que les distributions conditionnelles reconstruites par l'algortihme convergent vers la distribution jointe. De plus, un encodage au préalable des variables qualitatives est requis pour appliquer l'imputation par MICE.

2. Source : Numpyninja - MICE

3 Etude de cas

Désormais, il serait intéressant de comparer la performance ainsi que le temps de calcul de nos algorithmes sur un échantillon de données. Pour cela, nous avons considéré une base de données composée de 307 511 observations et de 122 variables, la target représente quant à elle 1 si le client a des difficultés de remboursement de son emprunt bancaire et 0 sinon. La base de données a été récupérée sur Kaggle. Elle provient de la banque tchèque Home Credit qui a mis à disposition cet ensemble de données anonymisées. L'information fournit par la table peut être utilisée pour évaluer la capacité de remboursement d'un emprunteur, en fonction de ses caractéristiques et des données contractuelles.

3.1 Comparaison de nos méthodes

3.1.1 Évolution de l'erreur

Pour apprécier et comparer la performance de nos méthodes d'imputation, nous avons généré aléatoirement un sous-échantillon de notre base initiale, constituée d'aucune donnée manquante. En effet, imputer l'ensemble de nos données directement serait beaucoup trop coûteux en temps de calcul. Pour se faire un sous-échantillon composé de 50 000 observations sans aucune donnée manquante sera utilisé avec un total de 8 variables. Ces 8 variables sont définies dans le tableau ci-dessous.

Variable	Signification
AMT_GOODS_PRICE	Le montant du prêt à la consommation
AMT_ANNUITY	Annuité du prêt
AMT_INCOME_TOTAL	Revenu total
AMT_CREDIT	Montant du credit
CNT_FAM_MEMBERS	Le nombre de membres de la famille
CNT_CHILDREN	Le nombre d'enfant
NAME_EDUCATION_TYPE	Les études les plus longues du client
DAYS_EMPLOYED	Nombre de jours employes

FIGURE 4 – Description des variables utilisées

Nous avons ensuite retiré une certaine proportion de données pour les quatre variables suivantes : AMT_GOODS_PRICE, AMT_CREDIT,

CNT_CHILDREN, NAME_EDUCATION_TYPE. Ces variables étant corrélées aux autres, l'information manquante peut être estimée à partir du reste de l'échantillon : il s'agit donc d'une perte de données de type MAR (voir Section 1). Ceci nous permettra notamment de comparer notre ensemble connu et l'ensemble imputé à l'aide de nos algorithmes. Pour l'étude des performances il serait intéressant de regarder la performance de nos algorithmes suivant la proportion de données manquantes.

En guise d'analyse de performance on utilise la RMSE (Root Mean Square Error) défini par :

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{x}_i - x_i)^2}{n}}, \qquad (1)$$

où n désigne le nombre de données manquantes et $\hat{x}_i - x_i$ l'écart entre la donnée estimée et la vraie donnée.

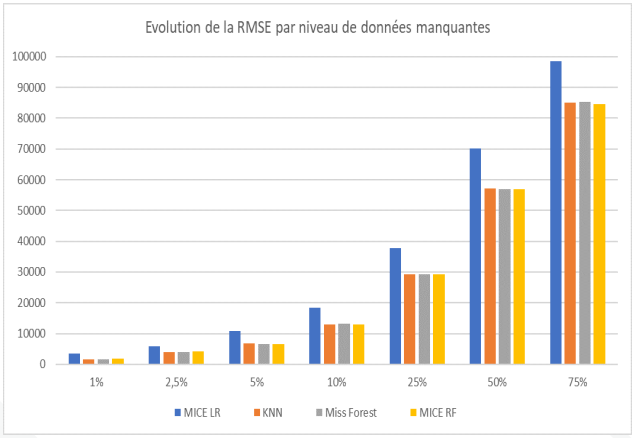

FIGURE 5 – RMSE de nos algorithmes par proportion de données manquantes

Comme nous pouvons le constater en Figure 5, les méthodes KNN (K = 4), MissForest et MICE Random Forest se distinguent par une erreur bien plus faible que la méthode de MICE basée sur la régression logistique. A titre d'information la RMSE de l'imputation simple a été retirée car elle étirée trop l'ordonnée du graphique étant donné que sa RMSE est bien plus forte (voir annexe pour le graphique avec l'imputation simple). Les RMSE de nos algorithmes MICE RF, KNN (K = 4) et MissForest évoluent de manière homogène pour les différentes proportions de données manquantes considérées. En effet, la RMSE de nos méthodes des KNN (K = 4), Miss Forest et de MICE par Random Forest sont relativement proches, comme nous pouvons le voir en Figures 6 et 7.

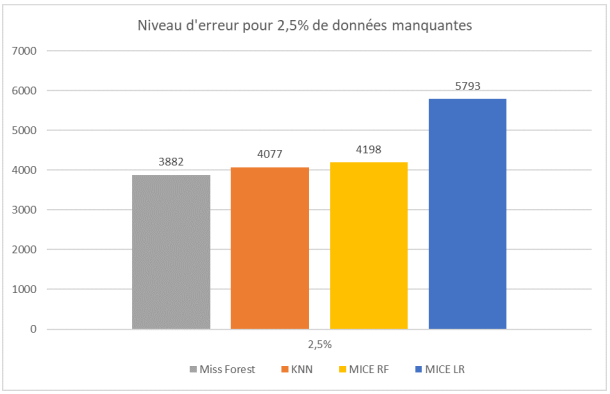

FIGURE 6 – RMSE de nos méthodes pour 2.5% de données manquantes

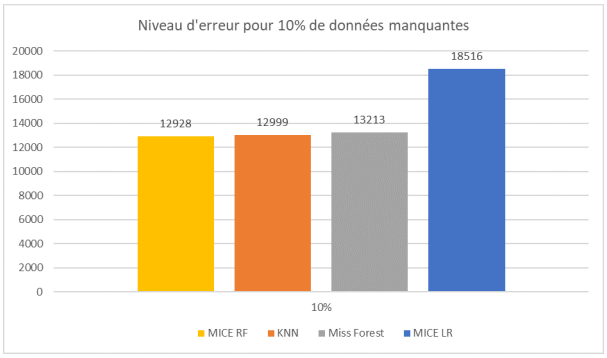

FIGURE 7 – RMSE de nos méthodes pour 10% de données manquantes

Cependant, nous constatons que la meilleure méthode d'imputation peut varier selon la proportion de données manquantes. Pour 2.5% de données manquantes, c'est l'imputation Miss Forest qui donne les meilleurs résultats d'ajustement, alors que, pour 10% de données manquante, c'est la méthode MICE par Random Forest.

3.1.2 Point de rupture

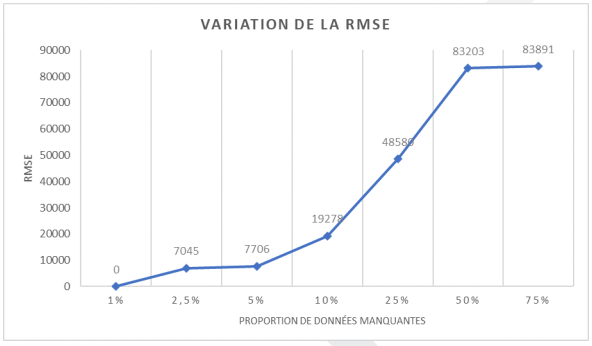

FIGURE 8 – Variation relative de la RMSE par proportion de données manquantes

Nous constatons en Figure 8 que le point de rupture se fait au delà de 10% de données manquantes, c'est à dire à partir 10% de données manquantes la hausse relative de l'erreur augmente fortement.

3.2 Temps de calcul

Reprenons le même échantillon composé de nos 8 mêmes variables et de nos 50 000 observations et retirons 10% de données sur le même ensemble de variables : AMT_GOODS_PRICE, AMT_CREDIT, CNT_CHILDREN, NAME_EDUCATION_TYPE.

Enregistrons le temps pris par l'algorithme pour effectuer l'imputation et répétons l'opération 10 fois pour chaque algorithme pour ensuite calculer la moyenne des résultats. Les temps de cal cul moyens pour chaque méthode sont comparés en Figure 9.

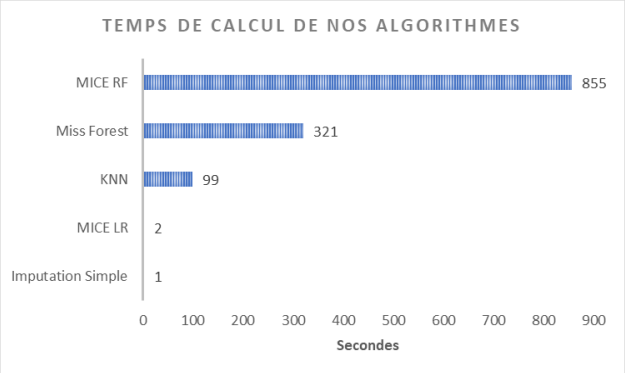

FIGURE 9 – Temps de calcul en secondes par algorithme

Nous constatons que l'imputation par KNN (K = 4) est l'approche la plus rapide parmi nos 3 meilleures méthodes. L'algorithme MICE en régression logistique est bien moins coûteux en temps de calcul. Fait important, notons que pour toute augmentation du nombre d'observations et de variables, le temps de calcul des KNN réagit très fortement. Cette méthode n'est pas à privilégier pour une plus importante volumétrie de données.

Proposition d'une mesure de pénalité

Nous proposons ici une mesure de performance qui tient compte à la fois de la performance de nos méthodes et du temps de calcul. Cette mesure est construite en appliquant une pénalité au RMSE liée au temps de calcul. Un peu à l'image d'un critère d'information qui viendrait pénaliser l'ajout de paramètre afin de garantir la parcimonie, ici l'objectif est d'identifier l'algorithme qui le plus performant en terme d'ajustement dans un temps de calcul raisonnable.

Nous considérons alors le temps de calcul en seconde par variable pour chacun de nos algorithmes, en faisant l'hypothèse que c'est essentiellement le nombre de variables ou la dimension de la base qui va accroître le temps de calcul de nos méthodes d'imputation. Ainsi, nous introduisons le critère de pénalité suivant :

$$\text{Pénalité} = P_i := \log(X_i) \tag{2}$$

où X_i représente le temps de calcul par variable pour la méthode i d'imputation testée. Pour une proportion p de données manquantes, le nouveau critère de performance pénalisé s'écrit alors :

$$y_i^p := \text{RMSE}_i^p * P_i \tag{3}$$

où RMSE_i^p est l'erreur associé à la méthode i pour la proportion p de données manquantes. Les Figures 10 et 11 donnent la RMSE pénalisée pour les différentes méthodes, pour une proportion $p = 2.5\%$ et $p = 10\%$ de données manquantes.

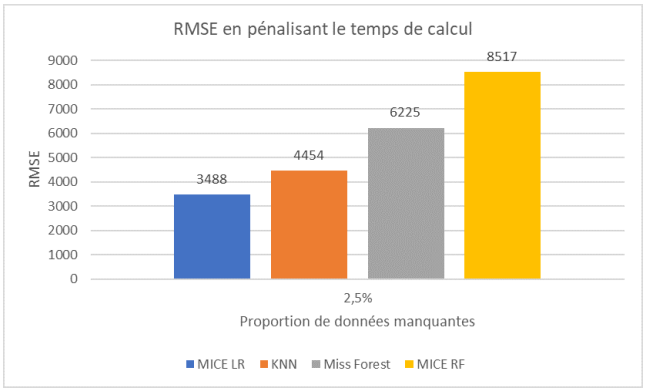

FIGURE 10 – RMSE pénalisé pour 2.5% de données manquantes

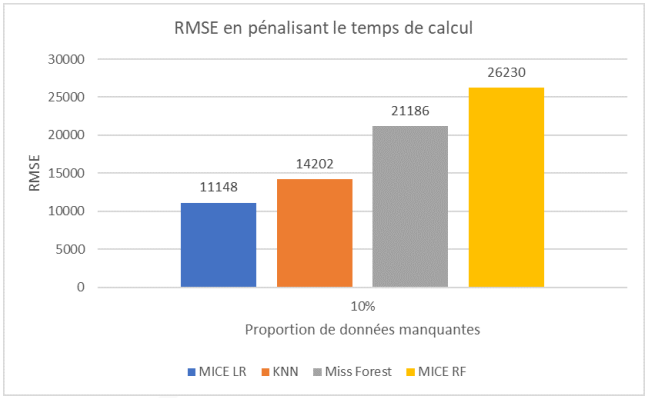

FIGURE 11 – RMSE pénalisé pour 10% de données manquantes

Nous constatons en appliquant la pénalité sur notre jeu de données que l'algorithme des MICE Regression Logistique ressort comme étant le meilleur arbitrage entre performance et temps de calcul. Les KNN qui étaient une des meilleures méthodes d'imputation affichent de bons résultats même si on peut s'attendre à une augmentation du temps de calcul si la dimension du problème augmente. Enfin, le compromis Miss Forest semble être assez intéressant si l'on cherche un algorithme performant avec des temps de calcul modéré et qui n'augmente pas autant que les KNN avec la dimension des données.

4 Conclusion

Répondre à la présence de données manquantes relève avant tout de bonnes connaissances du contexte dans lequel les données ont été construites, afin d'identifier si une imputation est envisageable pour pallier à ces manquements dans nos données. En effet, dans beaucoup de circonstances, les phénomènes ayant causé

l'incomplétude des données constituent une information à part entière, qu'il est important de prendre en compte. L'imputation de nos données peut être intéressante et peut accroître la performance des modèles qui seront ajustés, même s'il faut avoir en tête que cela ajoute un biais d'estimation et que les relations entre nos variables peuvent être légèrement modifiées. Enfin, notons que certains algorithmes notamment ensemblistes peuvent directement fonctionner sur des données manquantes, dans la phase de modélisation, ou que des étapes de discrétisation peuvent évincer de manières directe ces dernières.

5 Annexes

La Figure 12 décrit l'évolution de la RMSE par niveau de données manquantes en tenant de l'imputation simple par la médiane.

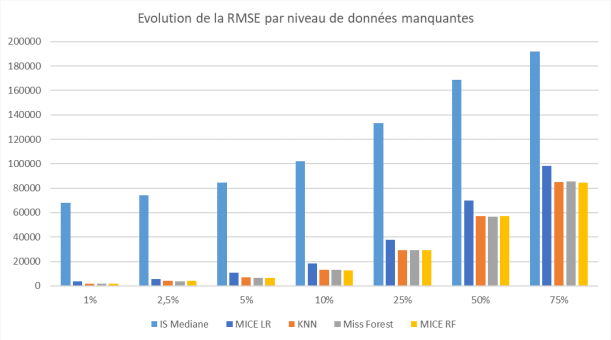

FIGURE 12 – RMSE de nos algorithmes par proportion de données manquantes avec l'imputation Simple

Références

Bennett, D. A. (2001). How can I deal with missing data in my study ? *Australian and New Zealand journal of public health 25*(5), 464–469.

Drechsler, J. and S. Rässler (2008). Does convergence really matter ? In *Recent advances in linear models and related areas*, pp. 341–355. Springer.

Little, R. and D. B. Rubin (1987). Statistical analysis with missing data. *New York*.

Stavseth, M. R., T. Clausen, and J. Røislien (2019). How handling missing data may impact conclusions : A comparison of six different imputation methods for categorical questionnaire data. *SAGE open medicine 7*, 2050312118822912.

Stekhoven, D. J. and P. Bühlmann (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics 28*(1), 112–118.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics 17*(6), 520–525.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research 16*(3), 219–242.

Nexialog Consulting est un cabinet de conseil spécialisé en Banque et en Assurance. Organisés autour de 3 domaines d'activité - Risques Bancaires, Financiers & Assurantiels - nous intervenons au sein des équipes métiers afin de les accompagner depuis le cadrage jusqu'à la mise en œuvre de leurs projets. Associant innovation et expertise, le savoir-faire de notre cabinet a permis de consolider notre positionnement sur ce segment et de bénéficier d'une croissance forte et régulière.

Les besoins de nos clients étant en constante évolution, nous nous adaptons continuellement pour proposer le meilleur accompagnement. Le département R&D de Nexialog Consulting se donne pour objectif de proposer des solutions innovantes à des problématiques métier ou d'actualité. Pour cela, nous nous appuyons sur des librairies internes et sur le travail de nos consultants. Le pôle R&D Nexialog a également pour mission de former les collaborateurs sur l'évolution des techniques et la réglementation en lien avec leur activité.

Site web du cabinet : https://www.nexialog.com
Publications : https://www.nexialog.com/publications/

Contacts

Ali BEHBAHANI
Associé, Fondateur
Tél : + 33 (0) 1 44 73 86 78
Email : abehbahani@nexialog.com

Christelle BONDOUX
Associée, Directrice commerciale
Tél : + 33 (0) 1 44 73 75 67
Email : cbondoux@nexialog.com

Areski COUSIN
Directeur scientifique
Email : acousin@nexialog.com