Master économiste D'entreprise

Mémoire de Recherche

Construction d'un modèle prédictif Analyse autour des données manquantes

SONDEJI Karl

Promotion 2023

Université de Tours

Mme. SCHOLLER Julie

21 Août 2023

Table des matières :

1 Introduction

Dans l'ère numérique actuelle, les données sont au cœur de chaque aspect de notre société. Qu'il s'agisse de la recherche scientifique, de l'analyse économique, de la prise de décision politique ou même des applications quotidiennes de la vie, les données jouent un rôle fondamental. Cependant, dans le monde réel, il est rare d'avoir des ensembles de données complets et parfaits. Des données manquantes peuvent apparaître pour diverses raisons, allant de problèmes techniques à des raisons liées à la collecte, à la saisie ou au stockage des données. On note trois principale causes à la survenue de données manquantes:

Dans un premier temps, suite à une fusion d'ensembles de données sources par un identifiant déchantillon (ID), si cet identifiant est présent dans la première base de données mais pas dans la deuxième alors on aura des données manquantes on parlera ici de lacunes structurelles.

Ensuite, de manière simplement aléatoire. Dans ce cas précis, un évènement aléatoire est venue empêcher la collecte de données comme par exemple un disfonctionnement d'un appareil de collecte, empêchant ainsi la collecte de données.

Enfin, des données manquantes peuvent se produire en cas d'échec de mesure. Par exemple, un patient qui arrête de suivre un traitement, on dira alors que ces données manquantes ont pour origine, une cause spécifique. Dans un tel contexte, comprendre la nature et l'importance des données manquantes devient essentiel afin de mettre en application des méthodes adaptées et efficaces pour réaliser des analyses fiables.

Ce mémoire est destiné à un public non-familier au traitement des données manquantes et a pour objet de présenter les principales méthodes permettant de gérer au mieux les données manquantes dans nos bases de données. Ainsi nous commencerons par définir les différents types de données manquantes, puis nous verons ensembles l'importances de celles-ci dans le traitement des données, enfin nous présenterons quelques méthodes de gestion des données manquantes et nous testerons celles-ci sur deux bases de données différentes et concluerons quant à l'éfficacité de ces méthodes. Pour cette dernière partie nous commencerons par prendre une base de données avec très peu de valeurs manquantes, ensuite nous injecterons de manières aléatoires des valeurs manquantes à l'intérieur de notre base de données afin de tester nos différentes méthodes.

La seconde base de données, plus volumineuse, sera sur la détection de fraude dans les transactions financières téléphoniques. La base de données contient 50000 observations.

2 Nature des Données Manquantes

Précdement nous avons énoncé trois types de données manquantes, à présent voyons plus en détail quels différents problèmes posent ces données manquantes en fonction de leur nature.

2.1 Lacunes structurelles

Les valeurs manquantes ayant pour origine une déficience structurelle, sont définies comme des données où une composante d'un prédicteur a été omise des données. imaginons qu'on ait une base données mesurant les différentes caractéristiques impactant le prix d'un t-shirt, et que l'on ait une variable *position_logo* prenant les modalités **gauche**, **droite** ou **manquant** et que 90% de t-shirt aient une valeur manquantes pour la variable *logo*, si on supprimait les valeurs manquantes cela voudrait dire que ces t-shirt non pas de logo, il serait donc préférable de remplacer la modalité **manquant** par **pas_dinformation_position** par exempe. On voit donc que pour ce type de valeurs manquantes, le problème est assez simple à résoudre une fois le composant nécessaire identifié.

2.2 Occurence aléatoire

- Données manquantes complètement au hasard (MCAR/ missing completly at random) : Ces données manquantes ne montrent aucune relation entre les valeurs manquantes et les autres variables présentes dans l'ensemble de données, la probabilité d'un résultat manquant est égale pour tous les points de données. Elles sont simplement dû au hasard et n'ont pas d'impact significatif sur les résultats finaux.
- Données manquantes au hasard (MAR/ missing at random) : La probabilité d'un manquant n'est pas égale pour tous les points de données, les données manquantes sont liées à d'autres variables, mais elles ne sont pas directement liées à la valeur manquante elle-même. Par exemple, les réponses manquantes à un sondage peuvent être liées à la période de l'année où le sondage a été réalisé. Dans ce scénario la probabilité d'une valeure manquante dépend des données observées.

En réalité il est souvent très difficile de déterminer si les valeurs manquantes ont la même probabilité d'occurrence pour toutes les données, ou ont des probabilités différentes pour les données observées ou non observées. Les méthodes qu nous verrons pourront s'appliquer à l'un ou l'autre cas.

2.3 Valeurs manquantes dû à une cause spécifique

Données manquantes non au hasard (NMAR/ not missing at random) : Ces données manquantes sont liées à des facteurs spécifiques qui ne sont pas aléatoires et peuvent introduire des biais importants dans les analyses. Par exemple, si des personnes ayant un certain revenu sont moins susceptibles de divulguer leurs informations financières dans un sondage, cela crée une NMAR. Il est important de distinguer les valeurs manquantes de causes spécifiques de celles de causses aléatoire afin de pouvoir utiliser les bonnes méthodes de traitement. Pour nous aider à les distinguer

il convient de visualiser nos données manquantes, de les quantifiés et de regarder les co-occurences (combinaisons de prédicteurs manquants les plus fréquents). Si on remarque que certaines combinaisons de prédicteurs on souvent ensemble de valeurs manquantes cela n'est surement pas dû au hasard.

3 Importance des Manquants

La présence de données manquantes peut entraîner divers problèmes lors du traitement des données et de la réalisation d'analyses :

- Biais : Les données manquantes, en particulier lorsque les valeurs manquantes ne sont pas au hasard, peuvent introduire un biais significatif dans les résultats. Les conclusions basées sur des données biaisées peuvent conduire à des décisions erronées et potentiellement coûteuses. Par exemple, si les répondants d'un sondage évitent de répondre à certaines questions gênantes, cela peut affecter les résultats de l'enquête, car les personnes ayant des caractéristiques spécifiques sont surreprésentées ou sous-représentées.
- Perte de Puissance Statistique : Lorsque des observations avec des valeurs manquantes sont exclues de l'analyse, cela peut réduire la puissance statistique de l'étude, rendant difficile la détection de relations significatives entre les variables.
- Erreur dans les Prédictions : Si les valeurs manquantes ne sont pas gérées correctement lors de la construction de modèles prédictifs, cela peut entraîner des erreurs importantes dans les prédictions futures.
- complexité dans les analyses: Les données manquantes peuvent rendre les analyses plus complexes. Les méthodes traditionnelles peuvent ne pas être applicables sans adaptation aux données manquantes. Cela nécessite l'utilisation de méthodes plus avancées, telles que l'imputation ou les méthodes de modélisation spécifiques

A noter que certains modèles sont plus résistant que d'autres aux données manquantes, il est donc préférable de les utiliser si on est face à un nombre important de données manquantes. C'est le cas du modèle d'arbre de décision CART, dont la méthodologie utilise l'idée de répartitions de substitution, soit si la valeur d'un prédicteur est manquant, l'algorithme de partitionnement utilisera un prédicteur alternatif dont la logique de fractionnement se rapproche du prédicteur d'origine. Les modèles tels que le Support Vector Machine ou le Logit sont à éviter en cas de valeurs manquantes.

4 Présentation des méthodes

Une fois que l'origine des données manquantes a été identifiée, plusieurs approches sont possibles, nous allons voir ensembles une liste de méthodes, non-exhaustive, permettant de traiter les données manquantes et ainsi de minimiser leur impact sur les analyses.

4.1 Supprimer les données manquantes

Lorsqu'il est nécessaire d'utiliser des modèles qui ne tolèrent pas de données manquantes, les valeurs manquantes doivent être extraites des données. L'approche la plus simple pour traiter les valeurs manquantes consiste à supprimer l'intégralité des prédicteurs contenant des valeurs manquantes. Cependant, il faut examiner attentivement un certain nombre d'aspects des données avant d'adopter cette approche. Il existe deux cas possibles, on élimine les valeurs manquantes en supprimant tous les prédicteurs contenant au moins une valeur manquante, ou alors on supprime les prédicteurs contenant plus d'un certains seuil de valeurs manquantes (par exemple 30%). Pour certains ensembles de données, il peut être vrai que des prédicteurs particuliers sont beaucoup plus problématiques que d'autres ; en supprimant ces prédicteurs, le problème des données manquantes est résolu, cependant si ces prédicteurs contiennent beaucoup d'information, ou si les valeurs manquantes en question ont un sens particulier il sera préférable de garder les variables en question, et de procéder autrement.

Une autre considération importante est la valeur intrinsèque des échantillons par rapport aux prédicteurs. Lorsqu'il est difficile d'obtenir des échantillons ou lorsque la base de données est assez petite, il n'est pas souhaitable de supprimer une partie de nos données, en effet on donne généralement priorité au plus grands nombres d'observations c'est pourquoi lorsque l'on a peu de données on ne supprimera que les prédicteurs dépassant un seuil d'absence.

4.2 Encoder les manquants

La présence de valeurs manquantes peut être dû à un problème d'encodage ou à un mauvais encodage, comme vu précedemment dans la catégorie "Lacune structurelles", recoder les valeurs manquantes peut changer le sens de l'information ainsi que son interpétation,mais aussi permettre à nos algorithmes de mieux traiter celles-ci. Toutefois il faut faire très attention car ce type de recodage sur les données peut avoir un impact non-négligeable sur notre études. Par exemple, Kuhn et Johnson (dans le livre "Applied Predictive Modeling",2013) utilisent un ensemble de données dont le but est de prédire l'acceptation ou le rejet des propositions de subvention. L'un des prédicteurs catégoriels était le sponsor de la subvention, qui prenait des valeurs telles que «subventions compétitives australiennes», «centre de recherche coopérative», «industrie», etc. Au total, il y avait plus de 245 valeurs possibles pour ce prédicteur avec environ 10% de les demandes de subvention ayant une valeur de parrain vide. Pour permettre aux applications qui avaient un parrain vide d'être utilisées dans la modélisation, les valeurs de parrain vides ont été encodées comme "inconnues". Pour bon nombre des modèles qui ont été étudiés, l'indicateur d'un parrain inconnu était l'un des prédicteurs les plus importants du succès de la subvention. En fait, l'odds-ratio qui opposait sponsor connu versus inconnu était supérieur à 6. Cela signifie qu'il était beaucoup plus probable qu'une subvention soit financée avec succès si le prédicteur du promoteur

était inconnu. En fait, dans l'ensemble de formation, le taux de réussite de la subvention associé à un parrain inconnu était de 82,2 % contre 42,1 % pour un parrain connu.

Dans ce cas précis on ne peut pas rééllement dire que l'encodage des données était une stratégie russie, et il est impossible de savoir pourquoi le mécanisme qui a conduit à l'identification de l'étiquette de sponsor "manquante"à une forte probabilité d'acceptation de la subvention, était vraiment important. Ce qui nous a permis de détecter le problème d'encodage, est le fait que les conséquences ont eu un impact important.

4.3 Imputer les données manquantes

L'imputation consiste à estimer les valeurs manquantes en utilisant des techniques telles que la moyenne, la médiane (on parle alors d'interpolation), la régression ou les méthodes de machine learning. Cependant, l'imputation doit être réalisée avec précaution, car elle peut introduire un biais supplémentaire si elle n'est pas effectuée de manière appropriée. L'objectif de ces techniques est de s'assurer que les distributions statistiques sont traitables et d'une qualité suffisante pour prendre en charge les tests d'hypothèses ultérieurs. La principale approche dans ce scénario consiste à utiliser plusieurs imputations ; plusieurs variantes de l'ensemble de données sont créées avec différentes estimations des valeurs manquantes. Les variations des ensembles de données sont ensuite utilisées comme données d'entrée dans les modèles et les répétitions des statistiques de test sont calculées pour chaque ensemble de données imputées. À partir de ces statistiques répétées, des tests d'hypothèse appropriés peuvent être construits et utilisés pour la prise de décision.

4.3.1 Approche prédictive ou inférentielle

Il existe plusieurs différences entre les modèles inférentiels et prédictifs qui ont un impact sur ce processus :

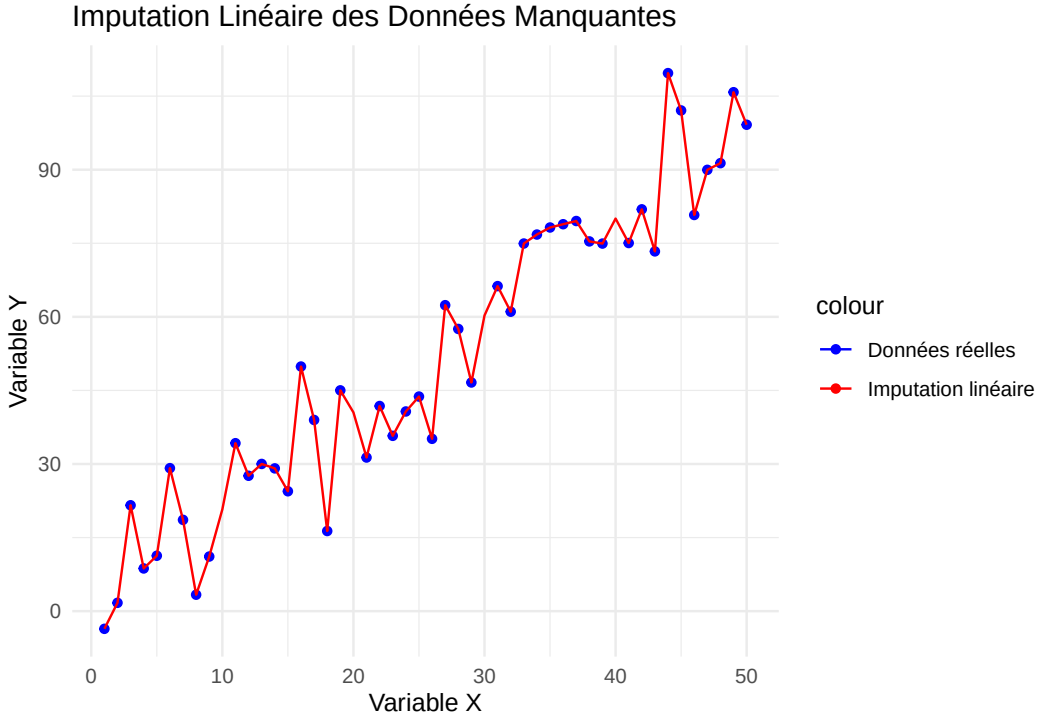
- Dans de nombreux modèles prédictifs, il n'y a pas de notion d'hypothèses distributionnelles (ou elles sont souvent insolubles). Par exemple, lors de la construction de la plupart des modèles basés sur des arbres, l'algorithme ne nécessite aucune spécification d'une distribution de probabilité pour les prédicteurs. Ainsi, de nombreux modèles prédictifs sont incapables de produire des résultats inférentiels, même si cela était un objectif principal 73 . Compte tenu de cela, les méthodes traditionnelles d'imputation multiple peuvent ne pas être pertinentes pour ces modèles.
- De nombreux modèles prédictifs sont coûteux en calcul. Une imputation répétée augmenterait considérablement le temps de calcul et les frais généraux. Il existe cependant un intérêt à saisir le bénéfice (ou le préjudice) causé par une procédure d'imputation. Pour garantir que la variation conférée par l'imputation soit prise en compte dans le processus de formation, nous recommandons que l'imputation soit effectuée dans le cadre du processus de rééchantillonnage.
- Étant donné que les modèles prédictifs sont jugés sur leur capacité à prédire avec précision des échantillons non encore observés (y compris l'ensemble de tests et les nouveaux échantillons inconnus), par opposition à leur pertinence statistique, il est essentiel que les valeurs imputées soient aussi proches que possible de leurs valeurs. valeurs vraies (non observées).

- L'objectif général des modèles inférentiels est de bien comprendre les relations entre le prédicteur et la réponse pour les données disponibles. À l'inverse, l'objectif des modèles prédictifs est de comprendre les relations entre les prédicteurs et la réponse qui sont généralisables à des échantillons encore à voir. Les méthodes d'imputation multiples ne conservent pas le générateur d'imputation après l'estimation des données manquantes, ce qui complique l'application de ces techniques à de nouveaux échantillons.

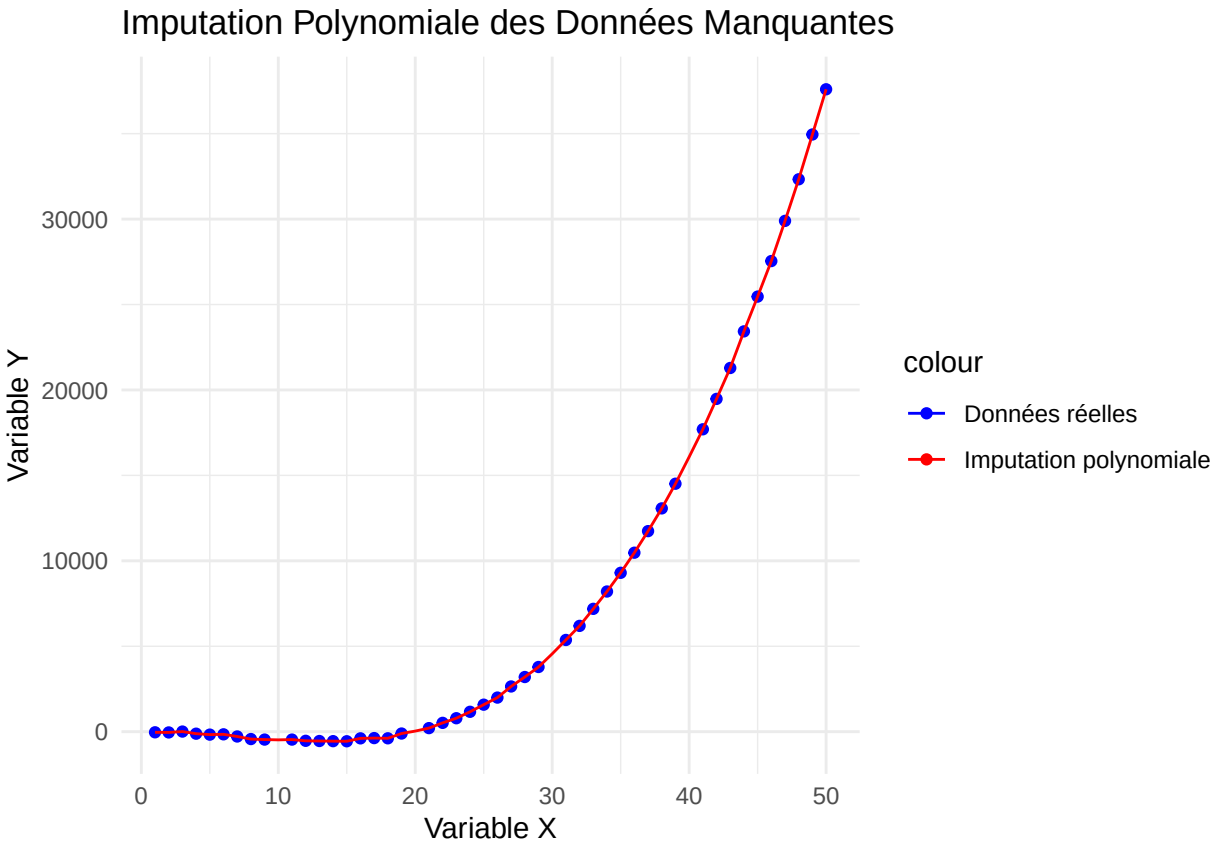
L'imputation soulève la question de savoir quelle quantité de données manquantes est trop importante pour être imputée, on considère généralement que l'on peut imputer au delà de 20% de valeurs manquantes dans une colonne. L'imputation est la première étape de toute séquence de prétraitement, elle doit avoir lieu avant les autres étapes qui impliquent l'estimation des paramètres. Par exemple, si le centrage et la mise à l'échelle sont effectués sur les données avant l'imputation, les moyennes et les écarts-types résultants hériteront des biais et des problèmes potentiels liés à la suppression des données.

4.3.2 Méthodes polynomiales

- Au degré 1 il s'agit simplemenent d'une regréssion linéaire, appellé aussi imputation par valeur unique ou interpolation. 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. Cette méthode est relativement simple d'utilisation et rapide à mettre en place. Elle cpnvient 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.

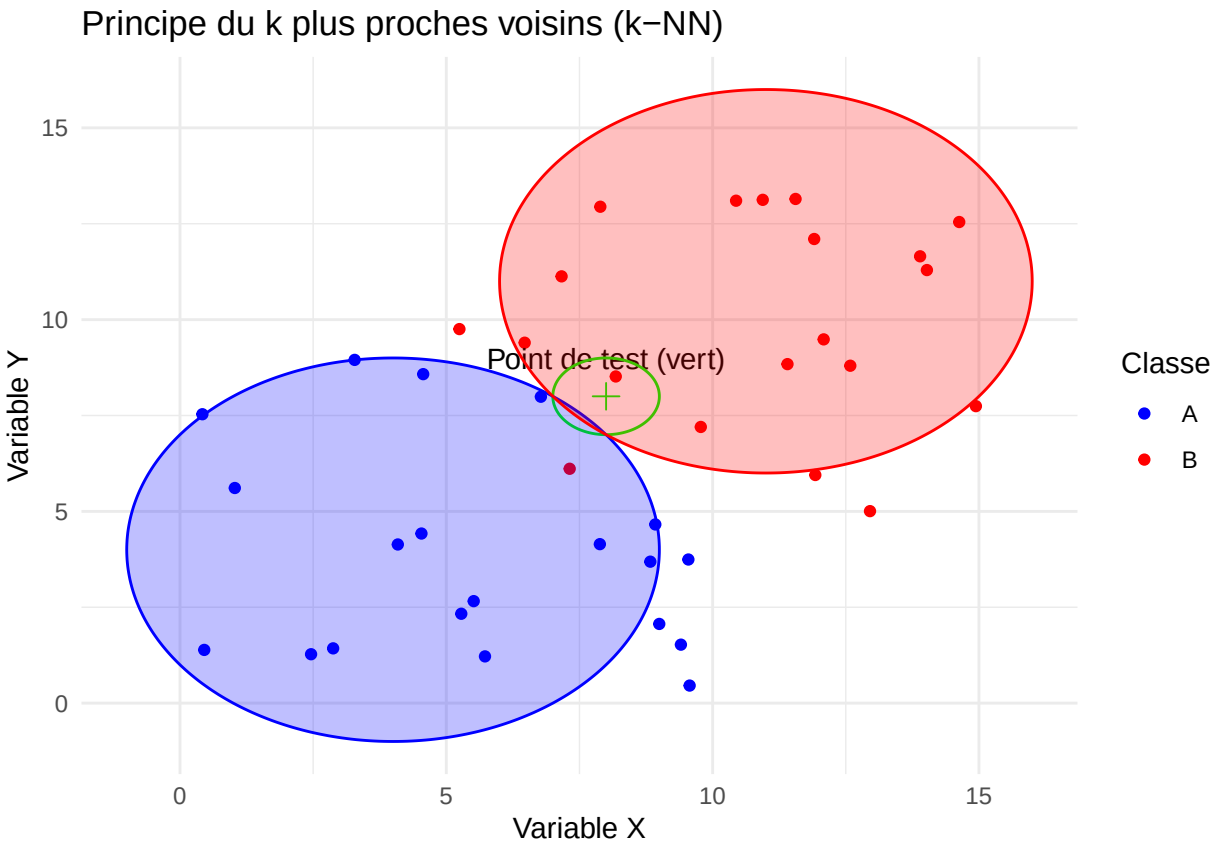


- Au delà du degré un on parle de regression quadratique (pour le degré 2) ou de regression poynomiale. L'inconvénients de ces approches est que vous pouvez interpoler uniquement les variables qui présentent des valeurs manquantes bien réparties. Ces méthodes d'apprentissage automatiques sont simplistes mais largement utilisées dans les séries temporelles.



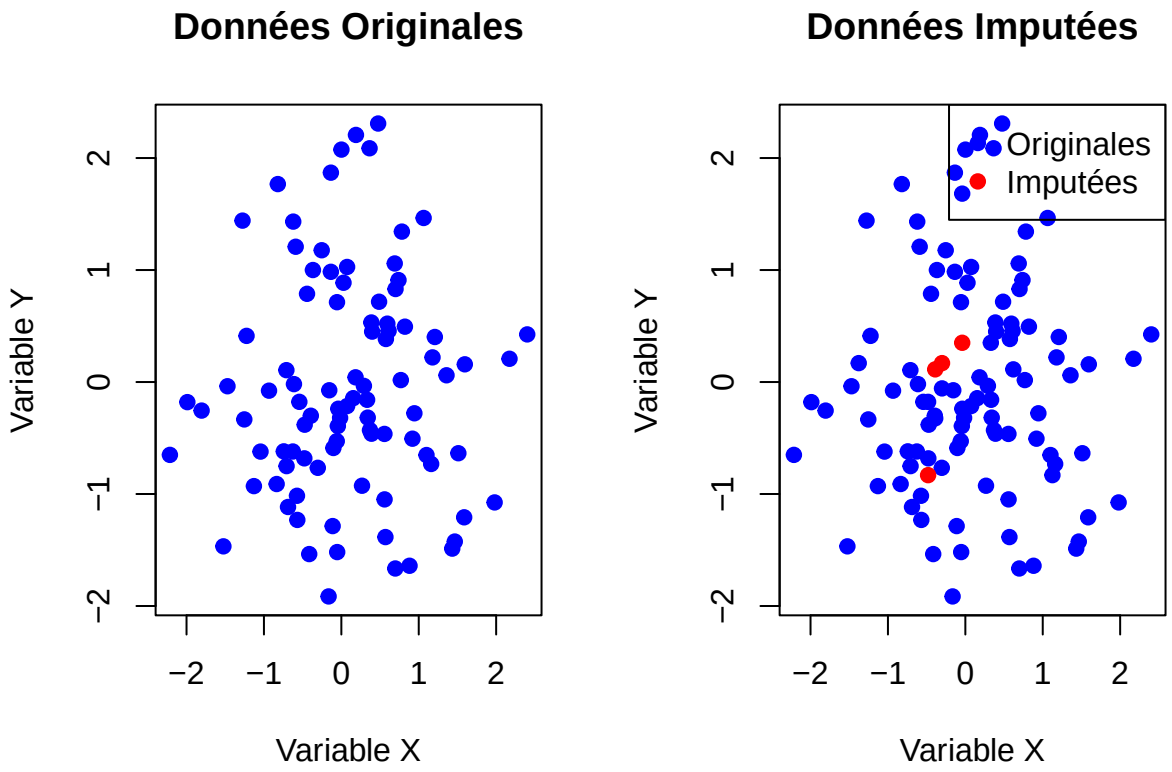
4.3.3 K-voisins les plus proches

Lorsque l'ensemble d'entraînement est de taille petite ou moyenne, les K-voisins les plus proches peuvent être une méthode rapide et efficace pour imputer les valeurs manquantes. Cette procédure identifie un échantillon avec une ou plusieurs valeurs manquantes. Ensuite, elle identifie les K-échantillons les plus similaires dans les données complètes (qui n'ont aucune valeur manquante dans certaines colonnes). La similarité des échantillons pour cette méthode est définie par une métrique de distance. Lorsque tous les prédicteurs sont numériques, la distance euclidienne standard est couramment utilisée comme métrique de similarité. Après avoir calculé les distances, les K-échantillons les plus proches de l'échantillon présentant la valeur manquante sont identifiés et la valeur moyenne du prédicteur d'intérêt est calculée. Une fois les K-voisins trouvés, leurs valeurs sont utilisées pour imputer les données manquantes. Le mode est utilisé pour imputer des prédicteurs qualitatifs et la moyenne ou la médiane est utilisée pour imputer des prédicteurs quantitatifs. K peut être un paramètre réglable, mais les valeurs autour de 5 à 10 sont une valeur par défaut raisonnable.



4.3.4 Arbres/Miss Forest

Les modèles arborescents sont un choix raisonnable pour une technique d'imputation puisqu'un arbre peut être construit en présence d'autres données manquantes. De plus, les arbres ont généralement une bonne précision et n'extrapolent pas les valeurs au-delà des limites des données d'apprentissage. Bien qu'un arbre unique puisse être utilisé comme technique d'imputation, il est connu qu'il produit des résultats à faible biais mais à variance élevée. Les ensembles d'arbres, cependant, fournissent une alternative à faible variance. Les forêts aléatoires sont une de ces techniques, cependant, l'utilisation de cette technique dans un contexte de modélisation prédictive présente quelques inconvénients notables. Avant tout, la sélection aléatoire des prédicteurs à chaque scission nécessite un grand nombre d'arbres (500 à 2 000) pour obtenir un modèle stable et fiable. Chacun de ces arbres n'est pas élagué et le modèle résultant a généralement une empreinte carbonne importante. Cela peut présenter un défi à mesure que le nombre de prédicteurs avec des données manquantes augmente puisqu'un modèle distinct doit être construit et conservé pour chaque prédicteur. Une bonne alternative qui a une empreinte informatique plus petite est un arbre en sac. Un arbre en sac est construit de la même manière qu'une forêt aléatoire. La principale différence est que dans un modèle groupé, tous les prédicteurs sont évalués à chaque division de chaque arbre. Les performances d'un arbre en sac utilisant 25 à 50 arbres se situent généralement dans la fourchette des performances d'un modèle de forêt aléatoire. Et le plus petit nombre d'arbres constitue un net avantage lorsque l'objectif est de trouver des valeurs imputées raisonnables pour les données manquantes.



4.3.5 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. Ceci lui permet d'être plus précis et plus flexible dans le sens ou l'imputation se fait variable par variable. 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.

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 v1, v2, v3 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 v1 uniquement contenant des données manquantes. Une régression est alors appliquée sur notre jeu de données avec comme seule variable cible v1, en omettant cette fois les observations comportant des données manquantes. A la suite de cela, un modèle de régression sera créé puis utilisé pour estimer les données manquantes de la variable v1. 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. Certaines études comme "Drechsler andRässler (2008)" montrent qu'on peut

s'attendre à ce que les distributions conditionnelles reconstruites par l'algortihme convergent vers la distribution jointe.

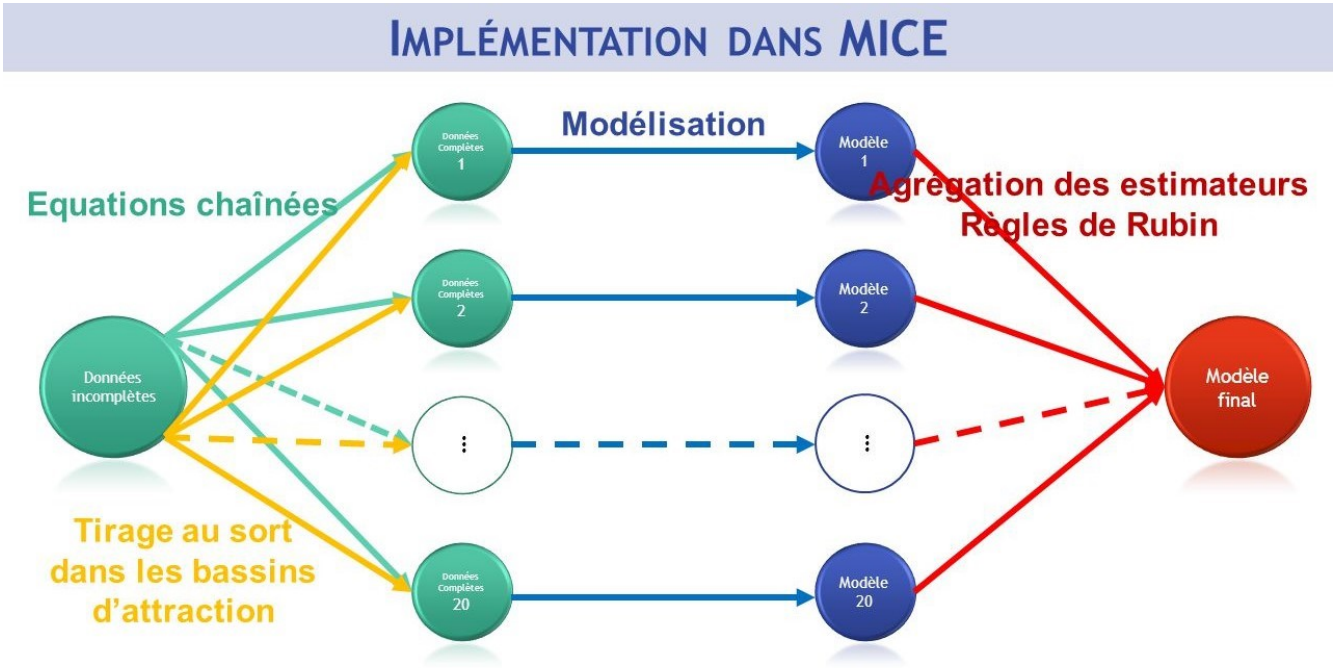


Figure 1: MICE

4.4 Cas particuliers

Il existe des situations où un point de données n'est pas manquant mais n'est pas non plus complet . Par exemple, lors de la mesure de la durée jusqu'à un événement, on peut savoir que la durée est d'au moins un certain temps T (puisque l'événement ne s'est pas produit). Ces types de valeurs sont dites censurées. Diverses méthodes statistiques ont été développées pour analyser ce type de données. Les durées sont souvent censurées à droite car la valeur de fin n'est pas connue. Dans d'autres cas, une censure à gauche peut se produire. Par exemple, les mesures de laboratoire peuvent avoir une limite de détection inférieure , ce qui signifie que l'instrument de mesure ne peut pas quantifier de manière fiable les valeurs inférieures à un seuil "X". Lorsqu'un prédicteur a des valeurs inférieures à la limite inférieure de détection, ces valeurs sont généralement signalées comme "<X", cela n'affecte pas négativement certains modèles de partitionnement, tels que les arbres ou les règles, cela peut avoir un impact négatif sur d'autres modèles car il suppose que ce sont les vraies valeurs. Pour atténuer le problème de variabilité, les valeurs censurées à gauche peuvent être imputées en utilisant des valeurs uniformes aléatoires entre zéro et X.

Dans les cas où les données ne sont pas définies en dehors des limites inférieures ou supérieures, les données seraient considérées comme tronquées

5 Mise en application

Dans cette partie nous nous concentrerons sur les diférentes méthodes d'imputations, il sera intéressant de comparer la performance ainsi que le temps de calcul de nos algorithmes sur un échantillon de données.

5.1 Visualisation de nos données

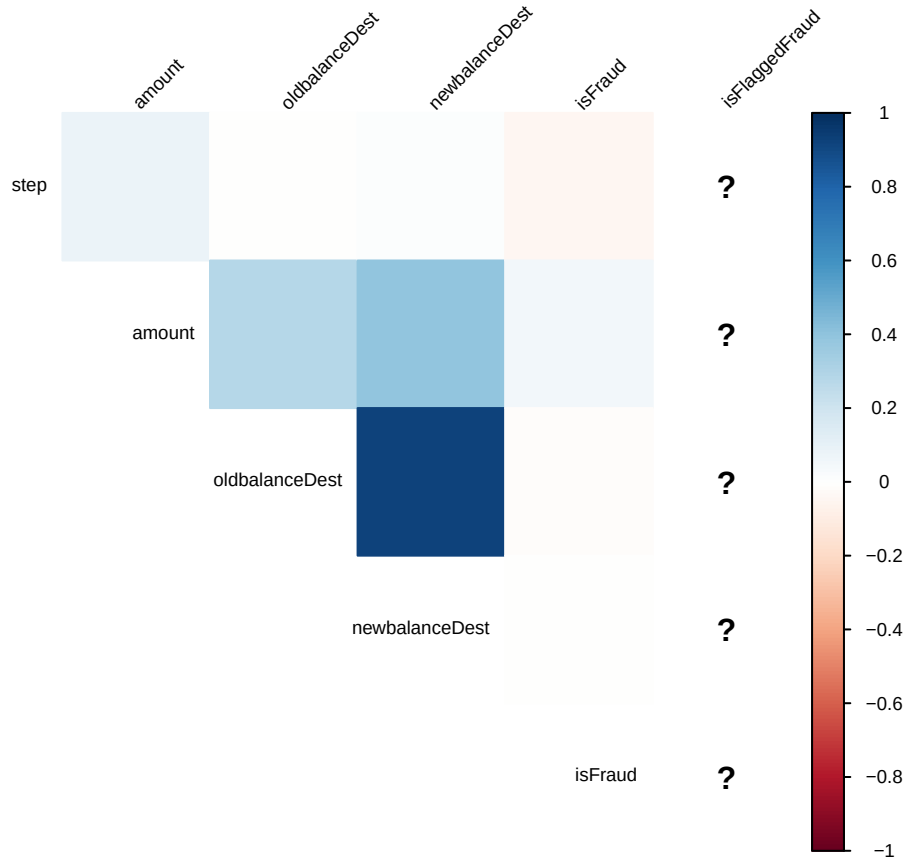
Commençons par présenter nos données.

La base de données concerne la détection de fraude dans les transactions financières, elle contient 11 variablees pour 50000 observations, nous ne conserverons que 6 d'entre elles pour notre études car celles relatives à l'identifiant de l'auteur de la transaction ne nous intéresse pas, par ailleurs nous conserverons certaines variables qui seraient inutiles pour déterminer la probabilité de fraude d'une transaction car ici nous voulons simplement étudier le comportement de nos données lors d'une imputation.

Les variables retenues sont les suivantes :

- **step**, cartographie une unité de temps dans le monde réel. Dans ce cas, 1 "step" équivaut à 1 heure. Total d'étapes 744 (simulation de 30 jours), de type integer.
- **type**, de type catégorielle : CASH-IN, CASH-OUT, DÉBIT, PAIEMENT et TRANSFERT.
- **amount**,montant de la transaction en devise locale, de type numérique.
- **oldbalanceOrg**,solde initial avant la transaction, numérique.
- **newbalanceOrig**, nouveau solde après la transaction, numérique.
- **isFraud**, Il s'agit des transactions effectuées par les agents frauduleux dans la simulation. Dans cet ensemble de données spécifique, le comportement frauduleux des agents vise à tirer profit en prenant le contrôle des comptes des clients et en essayant de vider les fonds en les transférant vers un autre compte, puis en les encaissant du système, binaire.
- **isFlaggedFraud**, Le modèle économique vise à contrôler les transferts massifs d'un compte à un autre et signale les tentatives illégales. Une tentative illégale dans cet ensemble de données est une tentative de transférer plus de 200 000 en une seule transaction, binaire.

Nous nous assurons qu'il n'y a pas de trop forte corrélations entre nos variables numériques qui pourraient biaiser nos résultats.



5.2 Imputation de nos données manquantes

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

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

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

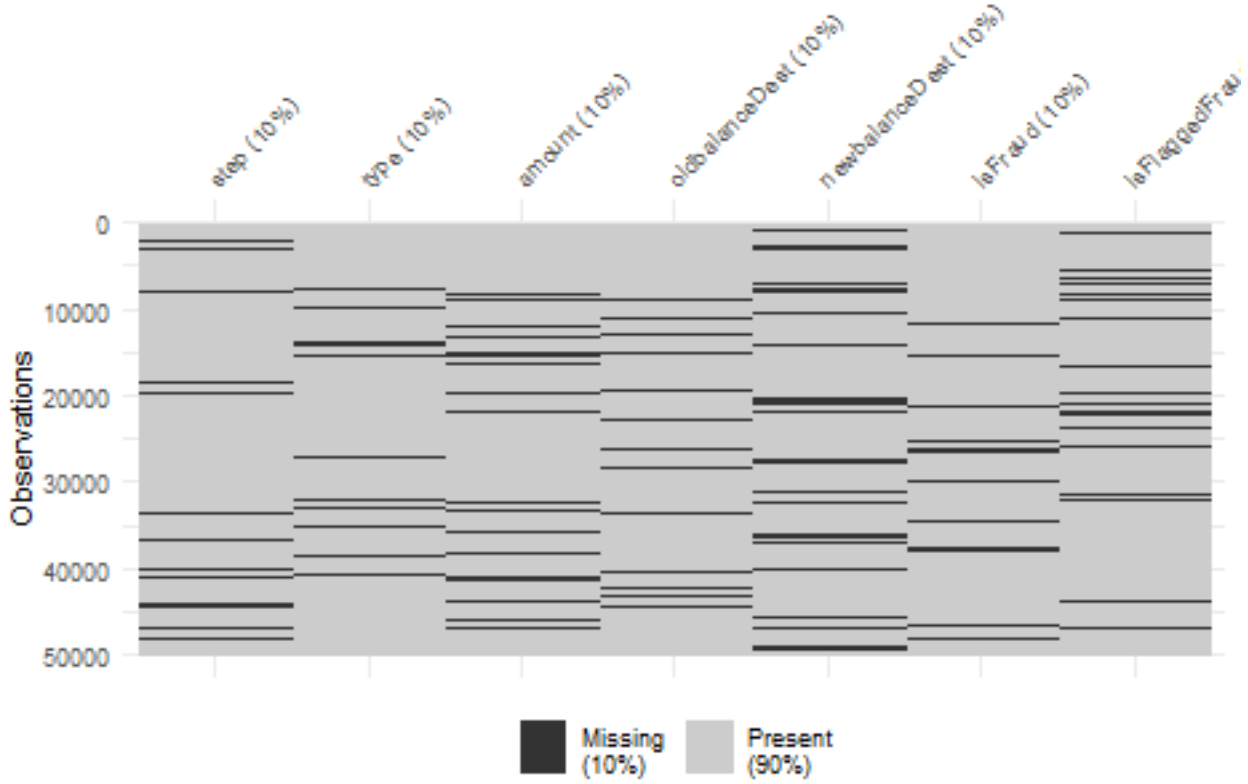
Pour injecter les données manquantes nous procéderons de la manière suivante:

```
library(dplyr)

inject_missing <- function(data, columns, prop_missing) {
  for (col in columns) {
    data <- data %>%
      mutate({{col}} := ifelse(runif(nrow(data)) < prop_missing, NA, {{col}})
    )
  }
  return(data)
}
```

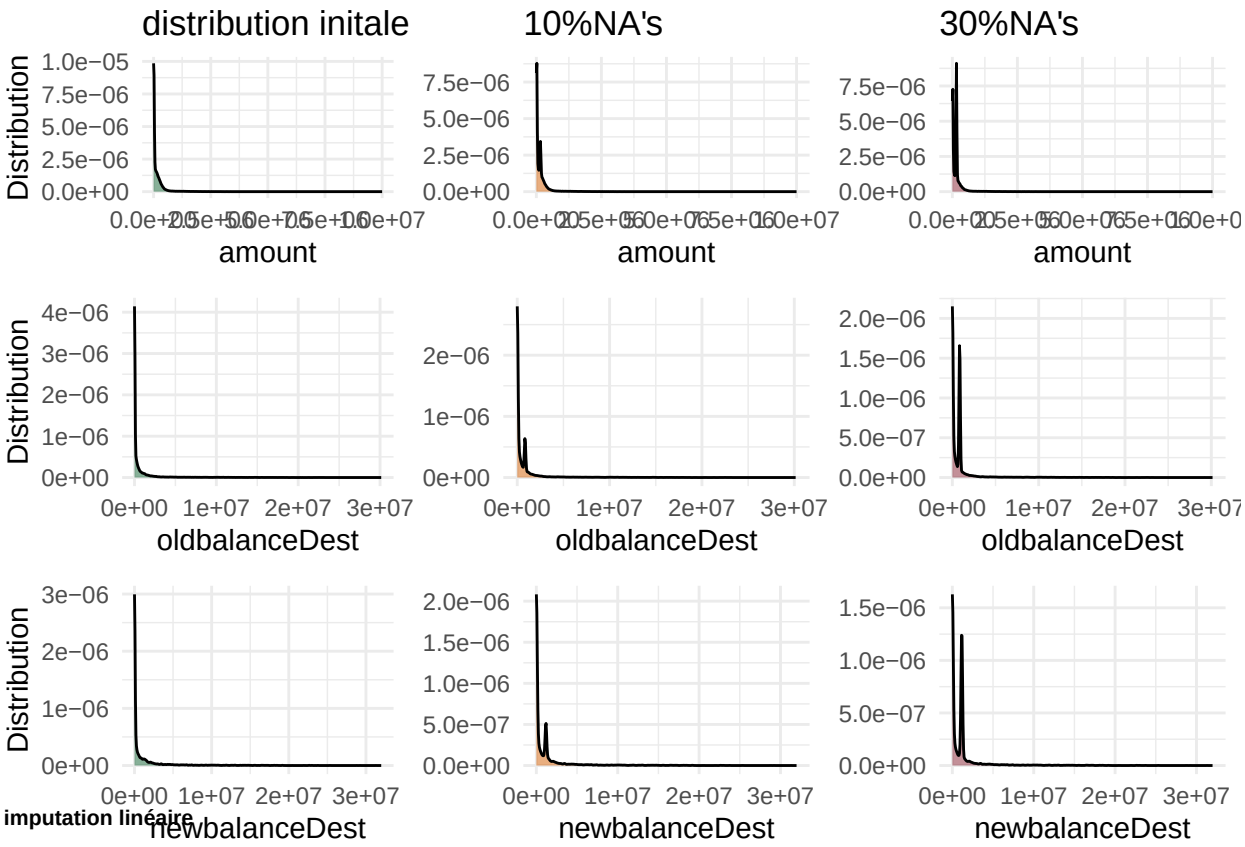
Dans cet exemple, nous utilisons la fonction mutate() de la bibliothèque dplyr pour modifier les valeurs des colonnes spécifiées en utilisant la fonction ifelse(). La fonction inject_missing prend en compte la proportion de valeurs manquantes que l'on souhaite ajouter.

Observons notre base de données avec 10% de données manquants injectées dans les variables numériques.



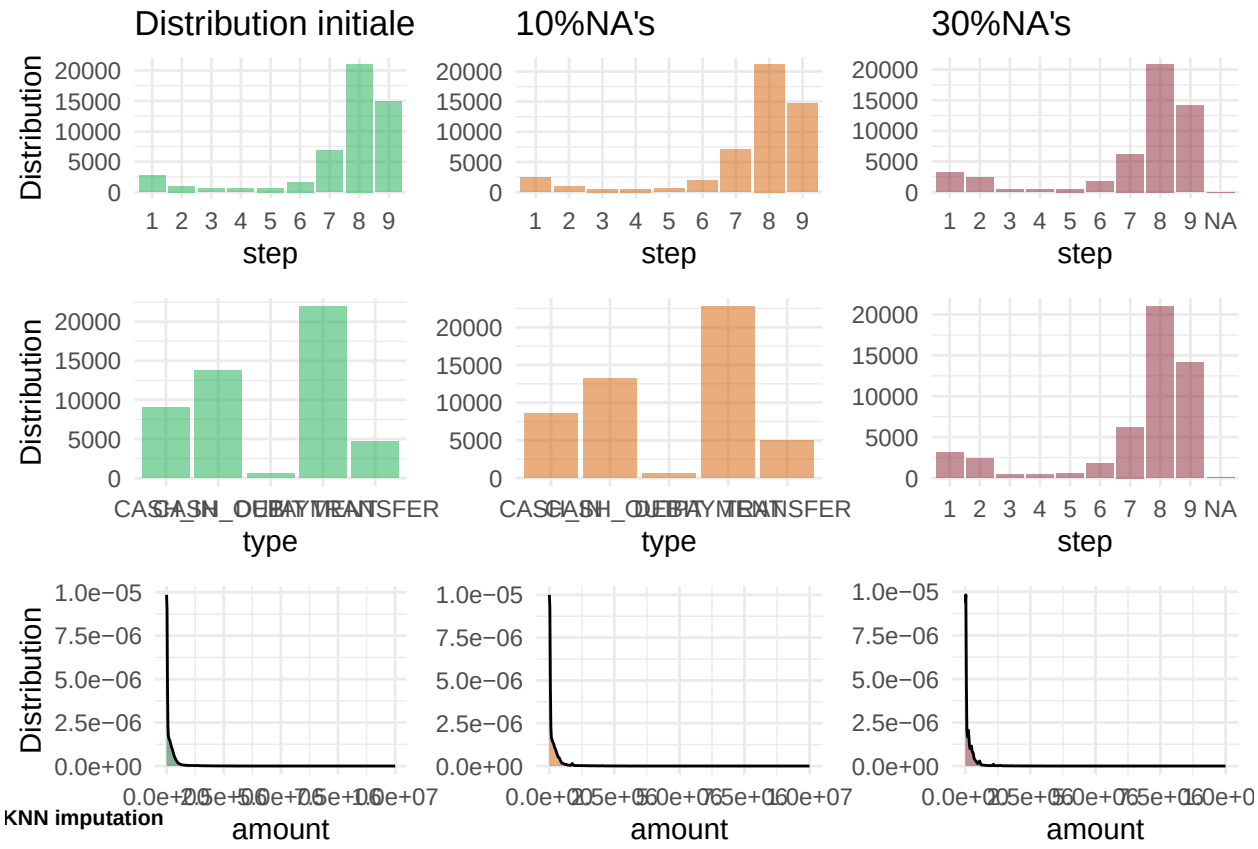
5.3 Imputation linéaire

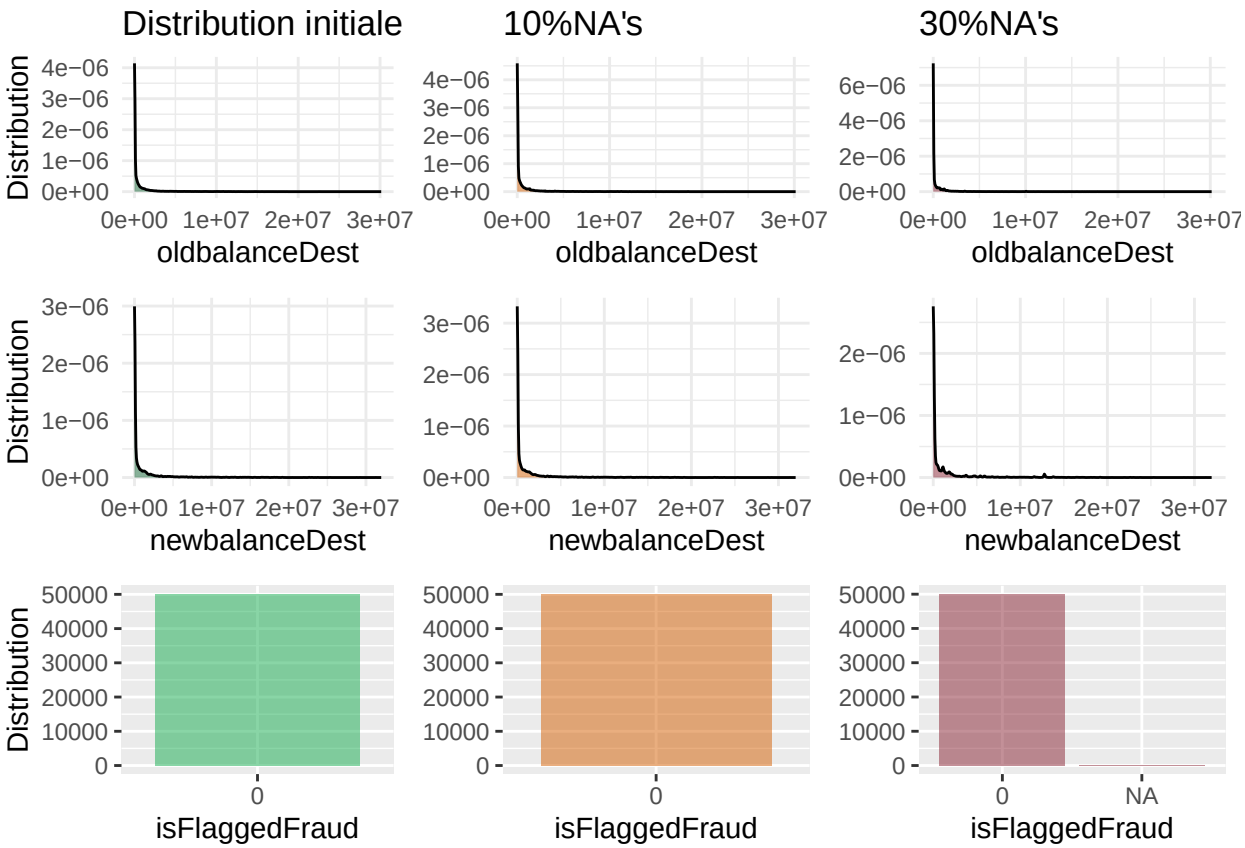
Pour réaliser l'imputation linéaire je n'ai utilisé que les variables numériques. Une fois l'imputation effectuée, on véridfie que la distribution de nos données n'a pas trop évolué.



Sans suprise plus on a de données manquantes, plus on s'éloigne de la distribution initiale.

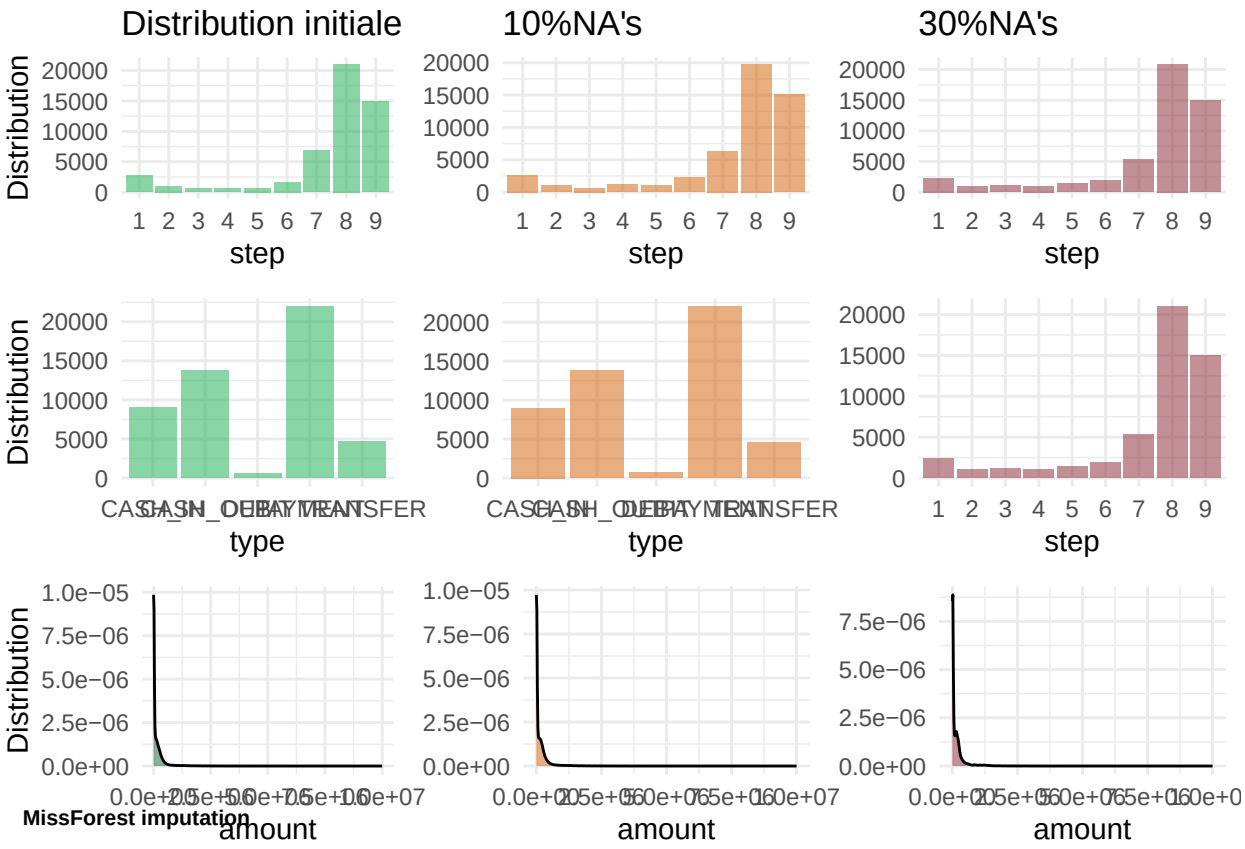
5.4 Imputation par Knn

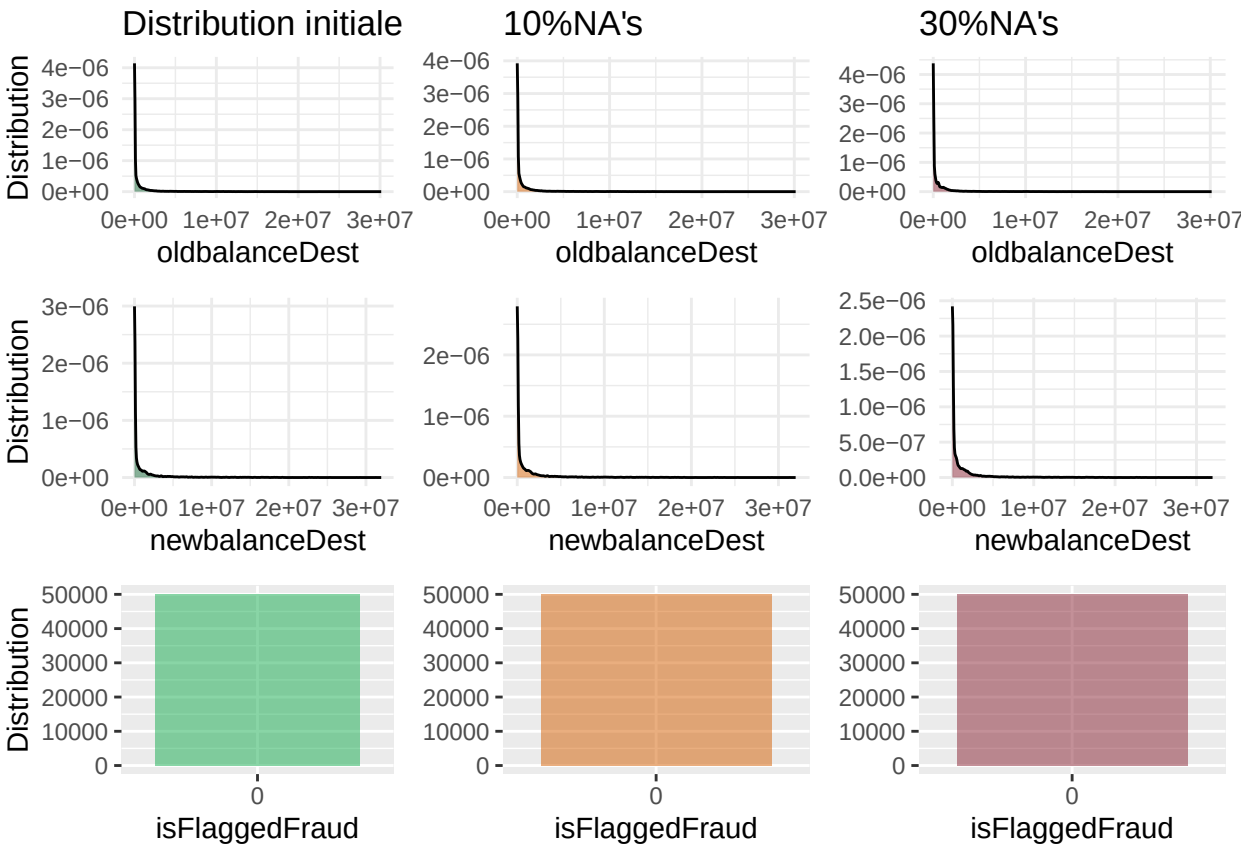




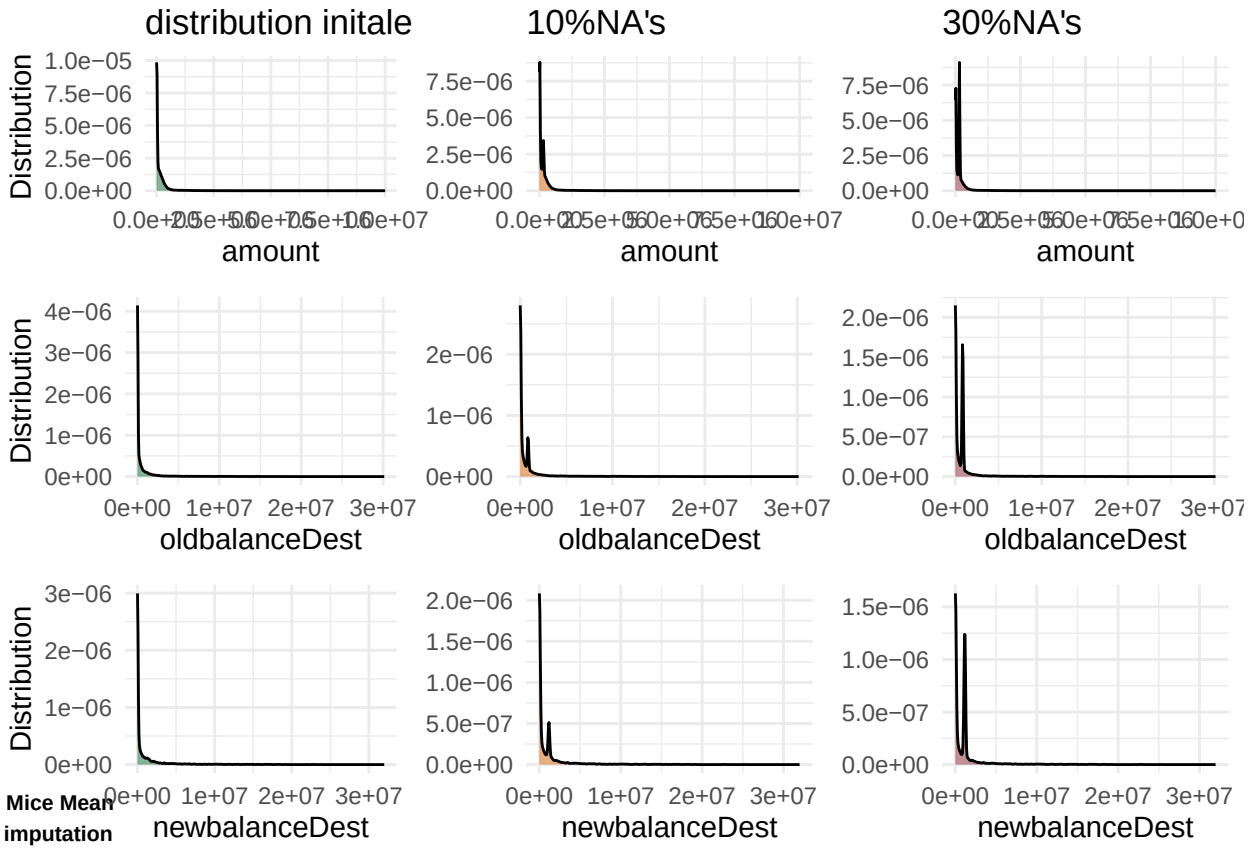
Lorsque l'on voit ces données on constate que l'imputation est plus précise piur les variables numériques que pour les variables catégorielles.

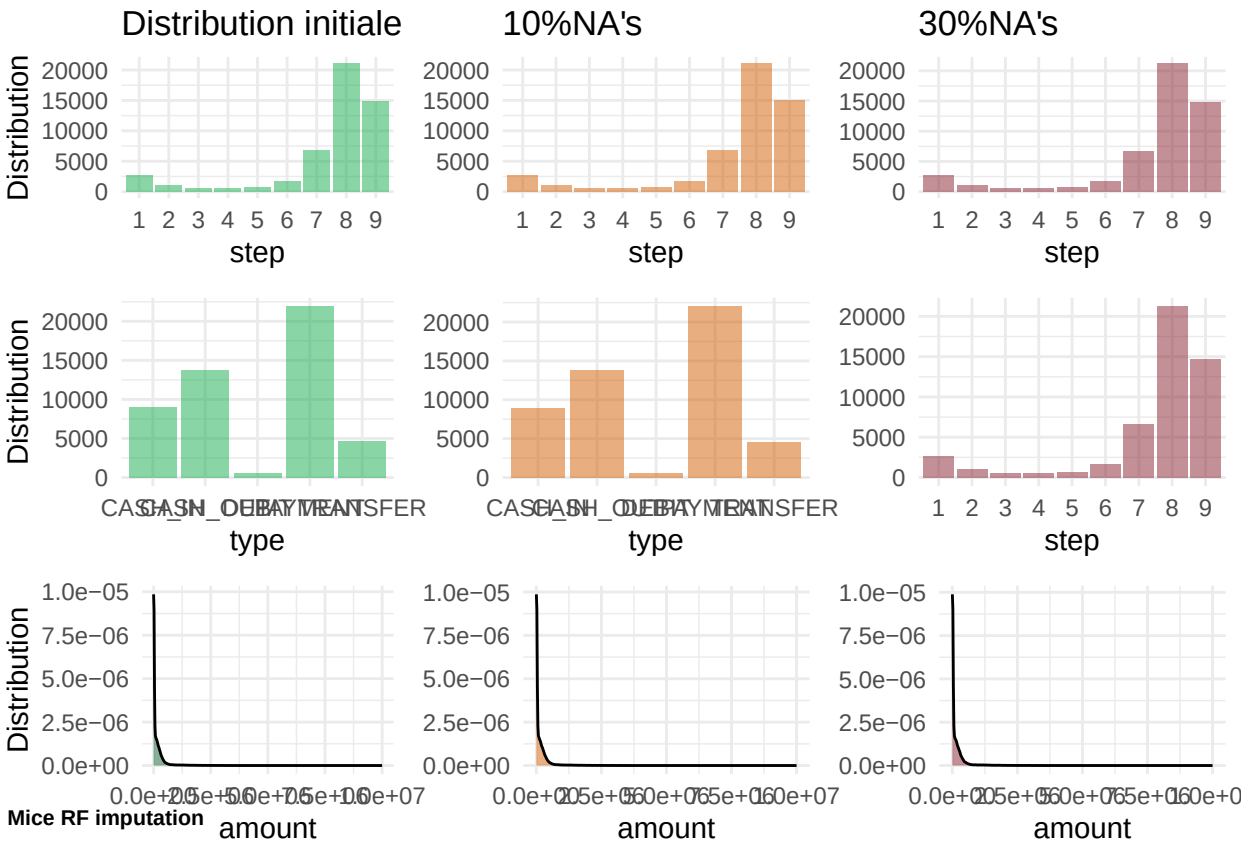
5.5 Imputation par MissForest

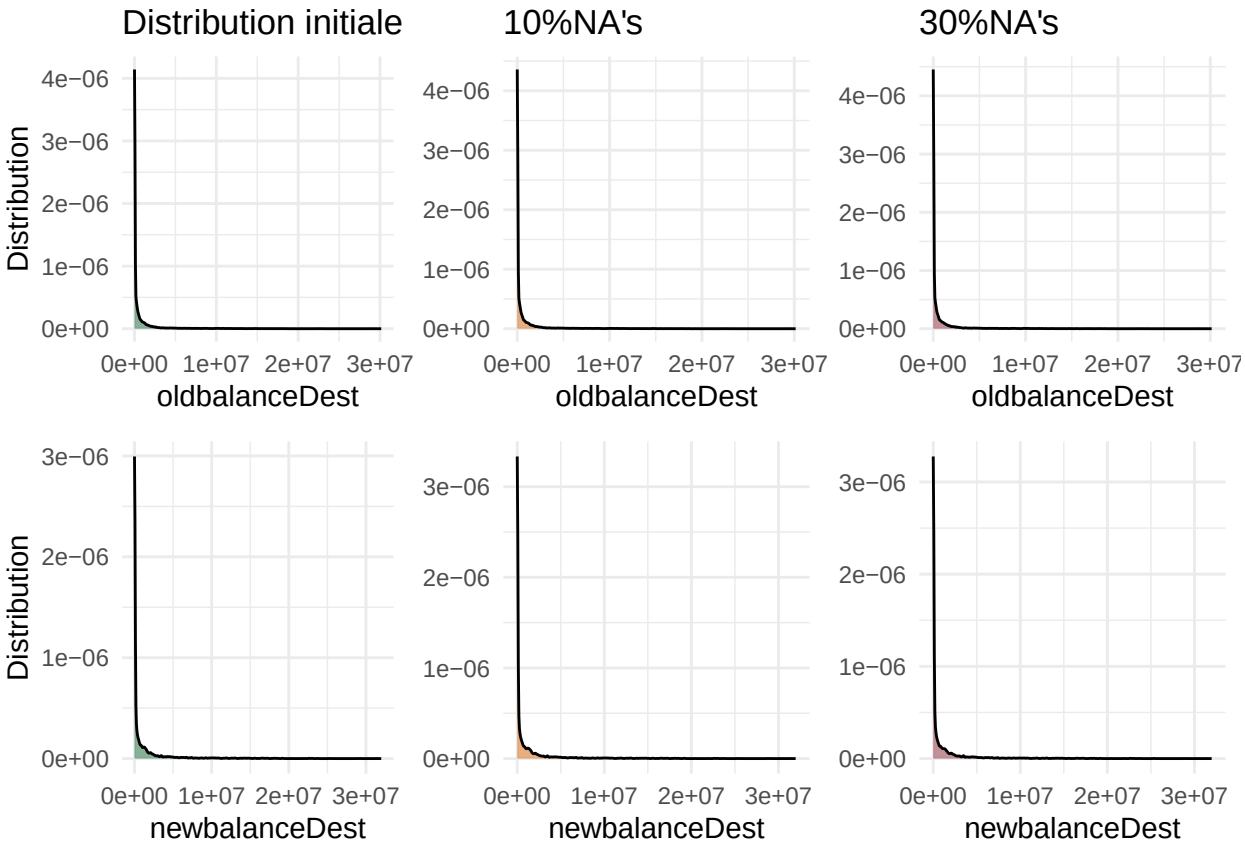



Distribution initiale
10%NA's
30%NA's
Distribution
oldbalanceDest
newbalanceDest
isFlaggedFraud

5.6 MICE


distribution initale
10%NA's
30%NA's
Distribution
amount
oldbalanceDest
newbalanceDest
Mice Mean imputation


Distribution initiale
10%NA's
30%NA's
Distribution
step
type
amount
Mice RF imputation


Distribution initiale
10%NA's
30%NA's
Distribution
oldbalanceDest
newbalanceDest

5.7 Temps de calcul

5.8 Performances

5.9 Résultats

Tableau de la mesure RMSE :		
Modèle	$RMSE_{10}$	$RMSE_{30}$
Linéaire	0.254	0.459
KNN	0.125	0.244
RF	0.116	0.202
$MICE$	0.091	0.174

On peut voir que le modèle qui fournit la meilleure précision est la MICE avec un taux d'erreur de 9% pour une proportion de 10% de données manquantes dans un échantillon de 50000 observtions.

6 Conclusion

En conclusion, les données manquantes constituent un défi important dans le traitement des données, et leur gestion adéquate est essentielle pour obtenir des résultats précis et fiables. En comprenant la nature des données manquantes et en utilisant des techniques appropriées pour les gérer, on peut garantir que nos analyses et nos conclusions sont robustes et représentatives de la réalité. Une approche réfléchie pour traiter les données manquantes est donc une étape cruciale dans l'exploitation efficace et responsable des données à notre disposition, c'est pourquoi il est impératif de passer par la modélisation afin de comprendre la nature des données manquantes, et opter pour la meilleure façon de les supprimer ou de les imputer.

Une fois connue la gravité des valeurs manquantes, une décision doit être prise sur la manière de traiter ces valeurs. Pour cela plusieurs facteurs sont à prendre en compte si on veut choisir la méthode la plus éffisciente, les deux facteurs principaux, étant la perfomance de notre méthode avec la RMSE, et le temps de calcul de nos algorithmes. La méthode la plus chrnonophage étant la MICE, nous allons donc préférer la KNN qui offre qui a un meilleur rapport temps/performance.

Bibliographie

- https://bookdown.org/max/FES/encoding-missingness.html
- https://hal.science/hal-03526292/
- https://thinkr.fr/donnees-manquantes-causes-identification-et-imputation/
- Traitement des données manquantes dans le milieu bancaire, Ahmad Charaf, Nexialog Consulting
- https://www.kaggle.com/datasets/ealaxi/paysim1
- https://www.kaggle.com/datasets/antfarol/car-sale-advertisements