Groupe SANCHEZ Laura, DOUSSIET Lou, VIDOR Fabien, CORIN Gaëtan

Projet Maths-Datascience

1. Introduction

1.1. Contexte du projet

Ce projet intervient dans le cadre du cours de mathématiques appliquées à la data science du Master 1 Data Engineer. Il est au sujet du traitement des données dans un projet d'apprentissage automatique et plus précisément sur le traitement des données manquantes dans un jeu de données.

1.2. Importance de la qualité des données dans un projet d'apprentissage automatique

On appelle données de qualité, un ensemble d'informations qui répondent aux exigences suivantes : pertinence, exactitude, actualité, intelligibilité, cohérence et accessibilité. Dans un contexte de projet d'apprentissage automatique, la qualité des données est cruciale.
Voici quelques raisons pour lesquelles la qualité des données est essentielle :
La préparation des données sera plus simple et plus rapide. Le modèle d'apprentissage sera plus précis. Si les données avec lesquelles il est entraîné ont un biais, des valeurs manquantes ou des données aberrantes, alors le modèle sera moins efficace.

1.3. Nécessité de traiter les valeurs nulles dans un jeu de données

Il est nécessaire de traiter les valeurs manquantes dans un jeu de données pour perdre le moins d'informations possible. Ensuite, la plupart des modèles d'apprentissage automatique ne peuvent pas gérer les valeurs nulles : cela entraîne des erreurs. Si les données manquantes ne sont pas traitées correctement, cela peut entraîner un biais dans les résultats.
De plus, il est essentiel de se questionner quant au pourquoi il y a ces valeurs manquantes : Quelles informations se cachent derrière cette absence de valeur? D'où proviennent-elles? Est-ce que ces données suivent un schéma spécifique dans le jeu de données ? Est-ce que la colonne où il manque des données est corrélée à une autre colonne?

1.4. Problématique et objectifs de l'étude comparative

Quelle est la stratégie de gestion de valeur manquantes à adopter pour limiter les biais d'analyse dans un jeu de données?

L’objectif de l’étude comparative consiste à étudier plusieurs types de valeurs manquantes en réalisant sur chacune plusieurs méthodologies permettant de réaliser des actions sur les valeurs nulles afin d’en tirer le meilleur résultat possible.

1.5. Annonce du plan

Voici le plan en détail de l’étude comparative:

- Simuler différents types de valeurs manquantes
- Catégoriser les types de données manquantes afin de les structurer (MCAR, MAR et NMAR)
- Appliquer différentes méthodologies de gestion de valeur manquantes sur chacun des types de données
- Étudier les réactions de ces types de valeurs suivant les méthodologies appliquées

2. Revue de la littérature

2.1. Taxonomie de type de valeurs nulles dans les jeux de données

Dans un jeu de données, nous distinguons deux catégories dites “non-réponse”:

- La “**non-réponse totale**”, lorsque aucune information n’est recueillie
- La “**non-réponse partielle**”, lorsque le manque d’information est limité à certaines variables.

Il existe trois mécanismes distincts de non-réponse:

- **MCAR** (Missing Completely At Random): réponse manquante complètement aléatoire
- **MAR** (Missing At Random): réponse manquante aléatoire
- **NMAR** (Non missing At Random): réponse manquante non aléatoire

2.2. Stratégies existantes pour gérer les valeurs nulles

Il existe différentes stratégies de gestion des valeurs manquantes dans les données:

- **La suppression des données manquantes : (méthode utilisée au sein du projet)**

Cette méthode consiste à supprimer les lignes ou les colonnes contenant des valeurs manquantes.

Cette dernière est recommandée si le nombre de valeurs manquantes est négligeable par rapport à la taille totale du jeu de données ou si les valeurs manquantes sont réparties aléatoirement.

Avantage	Inconvénient
● Simple et facile à mettre en œuvre ● Peut être approprié lorsque les valeurs manquantes sont peu nombreuses et aléatoires	● Peut entraîner une perte d'informations et de précision

- **Imputation par statistiques descriptives : (méthode utilisée au sein du projet)**

Cette méthode consiste à remplacer les valeurs manquantes par des statistiques descriptives telles que la moyenne, la médiane ou le mode de la variable

Cette méthode est simple mais peut introduire des biais si les données ne sont pas distribuées de manière normale.

Avantage	Inconvénient
● Simple et facile à mettre en œuvre ● Préserve la taille du jeu de données	● Ne prend pas en compte les relations entre les variables ● Peut introduire un biais

- **Imputation par modèle : (méthode utilisée au sein du projet)**

Cette méthode consiste à utiliser des modèles statistiques ou des algorithmes d'apprentissage automatique pour prédire les valeurs manquantes en fonction des valeurs observées et d'autres variables du jeu de données.

Nous allons utiliser l'**Algorithme d'espérance-maximisation (EM).**

C'est un méthode itérative utilisée pour estimer les paramètres de modèles statistiques lorsque certaines données sont manquantes.

Avantage	Inconvénient
● Obtention d'une estimation robuste des paramètres du modèle ● Utilisation de toutes les données du jeu de données ● Utilise différents types de modèles statistiques ● Convergence optimal de la fonction de vraisemblance	● Plus complexe à mettre en œuvre. ● Sensible à l'initiation des paramètres du modèle ● Nécessite de sélectionner un modèle approprié et de traiter les éventuelles erreurs de prédiction.

- **Imputation multiple :**

Cette méthode génère plusieurs ensembles de données complets en imputant les valeurs manquantes de manière stochastique.

Les ensembles de données complets sont ensuite analysés séparément, et les résultats sont combinés pour obtenir des estimations finales et des intervalles de confiance.

- **Imputation par interpolation :**

Cette méthode est utilisée pour les données séquentielles ou temporelles. Elle peut être utilisée pour estimer les valeurs manquantes en se basant sur les valeurs observées adjacentes dans le temps ou dans l'espace.

- **Imputation basée sur des règles métier :**

Cette méthode applique des règles spécifiques au domaine pour remplir les valeurs manquantes.

- **Imputation par clusterisation :**

Cette méthode divise les données en clusters basés sur les valeurs observées, puis remplace les valeurs manquantes par la moyenne ou le mode du cluster auquel elles appartiennent.

- **Imputation par propagation de la dernière observation valide (LOCF) :**

Cette méthode est utilisée principalement pour les séries temporelles, cette méthode consiste à remplacer les valeurs manquantes par la dernière observation valide connue.

3. Méthodologie de comparaison

3.1. Types de valeurs nulles à traiter

Pour ce projet nous avons décidé de générer nous-même différents type valeur nulles tels que :

- **MCAR** (Missing Completely At Random): réponse manquante complètement aléatoire
- **MAR** (Missing At Random): réponse manquante aléatoire
- **NMAR** (Non missing At Random): réponse manquante non aléatoire

Pour ce faire, nous avons créé trois copies du data frame original, un qui va contenir des MCAR, un second qui va contenir des MAR, un troisième qui va contenir des MNAR.

Dans le df "df_mcar", La colonne "climbRate" représente le "Missing Completely at Random" (MCAR) où 10% des valeurs sont effacées.

Dans le df "df_mar", La colonne "climbRate" représente le "Missing at Random" (MAR) en lien avec la colonne "Sgz" où 10% des valeurs sont effacées.

Dans le df "df_mnar", La colonne "climbRate" représente le "Missing Not at Random" (MNAR) où 10% des valeurs les plus hautes sont effacées.

3.2. Caractéristiques à examiner

Les caractéristiques à examiner dans cette étude sont les réactions de ces différents types de valeurs nulles suivant les stratégies de gestions appliquées.
Chaque type de données de valeurs nulles va réagir différemment à ces stratégies, et le scoring de différentes méthodes permettra d'établir la méthode la plus approprié suivant le type.

3.3. Les différentes stratégies de gestion des valeurs nulles à comparer

Les stratégies de gestion des valeurs nulle que nous allons comparer sont:

- la suppression des valeurs nulles
- la méthode d'imputation par le remplacement d'une valeur fixe
- la méthode d'espérance maximisation

3.4. Les jeux de données à utiliser et/ou protocole de génération de données

Les données qui ont été utilisées pour ce projet proviennent du site OpenML.
Le fichier "dataset_2202_elevators.arff" contient des données liées au contrôle d'un avion F16, en se concentrant sur les actions effectuées sur les élévateurs de l'avion. Les caractéristiques du fichier sont les suivantes : il comporte 8752 cas et 19 attributs continus. Les attributs incluent des variables telles que le taux de montée, la vitesse verticale, la pression, etc. Les données sont fournies sous forme de valeurs numériques pour chaque attribut.

Les attributs dans le fichier "dataset_2202_elevators.arff" ont les significations suivantes :

'climbRate' : Taux de montée
'Sgz' : Vitesse verticale
'p' : Pression
'q' : Variable non spécifiée
'curRoll' : Roulis actuel
'absRoll' : Roulis absolu
'diffClb' : Différence de taux de montée
'diffRollRate' : Différence de taux de roulis
'diffDiffClb' : Différence de différence de taux de montée
'SaTime1', 'SaTime2', 'SaTime3', 'SaTime4' : Temps d'action spécifique 1, 2, 3 et 4

'diffSaTime1', 'diffSaTime2', 'diffSaTime3', 'diffSaTime4' : Différence de temps d'action spécifique 1, 2, 3 et 4

'Sa' : Action spécifique

'Goal' : Objectif ou but lié à une action prise sur les élévateurs de l'avion

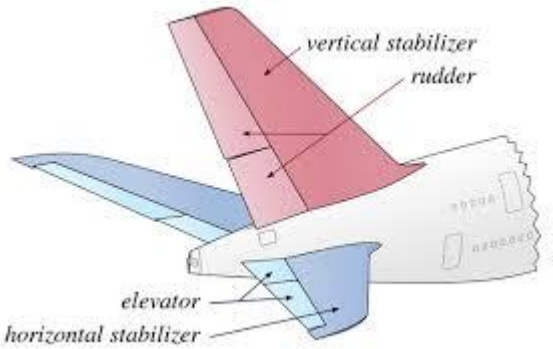
Ces attributs représentent différentes variables mesurées ou calculées dans le contexte du contrôle des élévateurs d'un avion F16, et ils sont utilisés pour analyser les actions effectuées sur ces éléments

Elevators ou gouverne est une surface mobile, agissant dans l'air et servant à piloter un avion, un dirigeable ou une fusée, par l'intermédiaire des commandes de vol, selon un de ses trois axes:

- tangage : rotation dans le plan vertical pour cabrer (monter) ou piquer (descendre) ;
- roulis : inclinaison latérale en virage ;
- lacet : rotation dans le plan horizontal pour « tourner » à gauche ou à droite.

Il s'agit généralement d'une surface articulée dont le changement d'orientation génère une force aérodynamique, de même que le gouvernail d'un bateau utilisé une force hydrodynamique.

Le jeu de données concerne les gouvernes de l’axe de lacet.



3.5. Langage et outils informatiques utilisés

Le langage de programmation utilisé est Python, pour le traitement des données données nous avons utilisé Jupyter Notebook ainsi que les librairies Scikit-Learn et Pandas…

3.6. Métriques d’évaluation des stratégies

Pour évaluer les différentes stratégies d’imputation des données, nous avons utilisé le modèle Random Forest Regressor, son score R^2, RMSE et MAE.

4. Analyse de stratégies

- **La suppression des données manquantes : (méthode utilisée au sein du projet)**

Cette méthode consiste à supprimer les lignes ou les colonnes contenant des valeurs manquantes.

Cette dernière est recommandée si le nombre de valeurs manquantes est négligeable par rapport à la taille totale du jeu de données ou si les valeurs manquantes sont réparties aléatoirement.

Avantage	Inconvénient
● Simple et facile à mettre en œuvre ● Peut être approprié lorsque les valeurs manquantes sont peu nombreuses et aléatoires	● Peut entraîner une perte d'informations et de précision

- **Imputation par statistiques descriptives : (méthode utilisée au sein du projet)**

Cette méthode consiste à remplacer les valeurs manquantes par des statistiques descriptives telles que la moyenne, la médiane ou le mode de la variable

Cette méthode est simple mais peut introduire des biais si les données ne sont pas distribuées de manière normale.

Avantage	Inconvénient

● Simple et facile à mettre en œuvre ● Préserve la taille du jeu de données	● Ne prend pas en compte les relations entre les variables ● Peut introduire un biais

- **Imputation par modèle : (méthode utilisée au sein du projet)**

Cette méthode consiste à utiliser des modèles statistiques ou des algorithmes d'apprentissage automatique pour prédire les valeurs manquantes en fonction des valeurs observées et d'autres variables du jeu de données.

Nous allons utiliser l'**Algorithme d'espérance-maximisation (EM).**

C'est un méthode itérative utilisée pour estimer les paramètres de modèles statistiques lorsque certaines données sont manquantes.

Avantage	Inconvénient
● Obtention d'une estimation robuste des paramètres du modèle ● Utilisation de toutes les données du jeu de données ● Utilise différents types de modèles statistiques ● Convergence optimal de la fonction de vraisemblance	● Plus complexe à mettre en œuvre. ● Sensible à l'initiation des paramètres du modèle ● Nécessite de sélectionner un modèle approprié et de traiter les éventuelles erreurs de prédiction.

5. Comparaison des stratégies et recommandations

Comparaison des différentes stratégies de gestion des valeurs manquantes en comparant le ***score du Random Forest Regressor.***

Type de valeurs manquantes	Suppression valeurs manquantes	Imputation par la moyenne	Espérance Maximisation
MCAR	0,8324	0,8405	0.84344
MAR	0,8429	0,8442	0,8470
MNAR	0,8429	0,8453	0,8458

Conclusion du tableau :

Les résultats de la suppression des valeurs manquantes sont inférieurs aux résultats des autres stratégies.

Résultats du RMSE et MAE entre le data frame original et le data frame où les valeurs manquantes ont été remplacées par la moyenne (stratégie de d'imputation statistique) et remplacées grâce à l'utilisation de l'algorithme d'espérance maximisation.

	DF / DF MCAR AVG	DF / DF MAR AVG	DF / MNAR AVG	DF / DF EM MAR	DF / DF EM MNAR	DF / DF EM MCAR
RMSE	89,56	379,667	357,3177	379,72	356,93	379.6669
MAE	22,91	302,1352	284,6281	302,31	284,45	302.1356

6. Conclusion

6.1. Récapitulation des résultats obtenus

On a testé deux types de méthodes de remplacement de valeurs manquantes. La moyenne et l'espérance maximisation. Les scores semblent peu varier sauf pour le remplacement par la moyenne dans le cas des MCAR.

6.2. Importance de choisir une stratégie adaptée en fonction du contexte d'un projet donné d'apprentissage automatique et perspectives

Dans un cas de MAR les deux solutions de remplacements testés sont très similaires. Pour les MNAR, l'espérance de maximisation semble meilleure mais les scores sont tellement proches que des conclusion définitives ne peuvent être tirées. Par contre, il y a une nette différence dans le cas des MCAR où la moyenne obtient les meilleurs résultats de nos tests.

6.3. Perspectives

Dans le cas des MAR et MNAR, les scores de nos méthodes sont trop faibles, il faudrait étudier d'autres méthodes comme celle des proches voisins qui semble plus cohérente à ces problématiques.

7. Références

7.1. Travaux et ressources utilisées pour la revue de la littérature et justification des choix méthodologiques.

valeur nulles
https://www.sciencedirect.com/science/article/abs/pii/S1551741121001157?via%3Dihub

https://www.frontiersin.org/articles/10.3389/fdata.2021.693674/full
https://dl.acm.org/doi/10.1145/3533381

data quality
https://medium.com/nerd-for-tech/its-time-to-focus-more-on-data-quality-and-ethics-than-algorithms-656890da247a
https://medium.com/@evertongomede/an-essay-on-the-multifaceted-approach-to-assessing-data-quality-fb5f45dd7f50
https://www.youtube.com/watch?v=U5AVaI86Jws&ab_channel=StatistiqueCanada
https://medium.com/investbegin/the-secret-ingredient-of-ai-success-navigating-the-world-of-data-quality-e533bcdb0188