Test Cas Spéciaux - Niveau 2 (Avancé)

UniBank Haiti - Data Analyst

Durée: 50 minutes
Questions: 30
Niveau: Avancé
Sujets: Valeurs manquantes avancé, Outliers complexes, ACP interprétation, ANOVA multi-facteurs

---

Section A: Valeurs Manquantes Avancé (8 questions)

Q1. Dans votre dataset de scoring crédit, la variable "revenus" a 25% de valeurs manquantes. Les clients sans revenu renseigné ont un taux de défaut 3× plus élevé. Quel type de mécanisme de manquant cela suggère-t-il?

A) MCAR (Missing Completely At Random)

B) MAR (Missing At Random)

C) MNAR (Missing Not At Random)

D) Erreur de saisie aléatoire

Réponse: C) MNAR (Missing Not At Random)

Le fait que les manquants soient corrélés avec la variable cible (défaut) suggère que le revenu est manquant PARCE QU'il est bas (personnes en difficulté financière ne déclarent pas). C'est MNAR - le plus problématique.

Q2. Vous utilisez IterativeImputer (MICE) pour imputer les valeurs manquantes. Quelle est son avantage principal par rapport à l'imputation simple par médiane?

A) Plus rapide

B) Capture les relations entre variables pour une imputation plus réaliste

C) Ne nécessite pas de données

D) Toujours plus précis

Réponse: B) Capture les relations entre variables pour une imputation plus réaliste

MICE (Multiple Imputation by Chained Equations) modélise chaque variable manquante en fonction des autres, préservant les corrélations. L'imputation médiane ignore ces relations.

Q3. Avant d'imputer les valeurs manquantes dans un modèle de ML, pourquoi faut-il splitter train/test AVANT?

A) Pour gagner du temps

B) Pour éviter le data leakage - les statistiques de test ne doivent pas influencer l'imputation

C) Ce n'est pas nécessaire

D) Pour avoir plus de données de test

Réponse: B) Pour éviter le data leakage - les statistiques de test ne doivent pas influencer l'imputation

Si on impute sur l'ensemble des données, les statistiques du test "fuient" dans l'entraînement via les valeurs imputées. Il faut fit l'imputer sur train, puis transform sur test.

Q4. La variable "nb_enfants" a 5% de valeurs manquantes. Après analyse, ces manquants semblent corrélés avec l'âge (plus de manquants chez les jeunes). Quelle stratégie d'imputation?

A) Suppression des lignes

B) Imputation par la médiane globale

C) Imputation par groupe d'âge (MAR)

D) Remplacer par 0

Réponse: C) Imputation par groupe d'âge (MAR)

Si le mécanisme est MAR (dépend de l'âge), l'imputation conditionnelle par groupe d'âge est appropriée. Elle préserve la relation âge-nb_enfants.

Q5. Vous créez un indicateur binaire "revenu_manquant" en plus de l'imputation. Pourquoi cette approche peut-elle être utile?

A) Pour augmenter le nombre de colonnes

B) Le pattern de manquant peut lui-même être prédictif (surtout si MNAR)

C) Pour compliquer le modèle

D) C'est inutile

Réponse: B) Le pattern de manquant peut lui-même être prédictif (surtout si MNAR)

Si le fait d'avoir une valeur manquante est informatif (ex: clients qui cachent leur revenu car faible), l'indicateur capture cette information que l'imputation seule perd.

Q6. KNNImputer avec k=5 est utilisé sur votre dataset. Si un client a 3 voisins avec revenus [50K, 60K, 70K] et 2 voisins avec revenus manquants, quelle valeur sera imputée?

A) 60K (médiane des 3)

B) Moyenne des 3 disponibles = 60K

C) Impossible sans les 2 manquants

D) 0

Réponse: B) Moyenne des 3 disponibles = 60K

KNNImputer utilise les k plus proches voisins mais ignore leurs valeurs manquantes pour le calcul. Moyenne de [50K, 60K, 70K] = 60K.

Q7. Après imputation, vous remarquez que la variance de la variable imputée a diminué de 30%. Est-ce normal?

A) Non, c'est une erreur d'imputation

B) Oui, l'imputation par valeur centrale (moyenne/médiane) réduit naturellement la variance

C) La variance devrait augmenter

D) La variance ne change jamais

Réponse: B) Oui, l'imputation par valeur centrale (moyenne/médiane) réduit naturellement la variance

L'imputation simple par moyenne/médiane "tire" les valeurs vers le centre, réduisant la dispersion. C'est un effet connu. MICE ou imputation multiple préservent mieux la variance.

Q8. Vous avez un dataset avec 3 variables ayant respectivement 5%, 15%, et 40% de valeurs manquantes. Comment les traiter?

A) Supprimer toutes les lignes avec au moins un manquant

B) Imputer les 3 variables de la même façon

C) Stratégie différenciée: imputer 5% et 15%, considérer suppression de la variable à 40%

D) Ignorer les manquants

Réponse: C) Stratégie différenciée: imputer 5% et 15%, considérer suppression de la variable à 40%

< 5%: suppression ou imputation simple OK. 5-20%: imputation recommandée. > 40%: questionner la valeur de la variable, potentiellement supprimer ou utiliser indicateur uniquement.

Section B: Outliers Complexes (7 questions)

Q9. Dans la détection de fraude, vous identifiez une transaction de 500,000 HTG pour un client dont la moyenne est 5,000 HTG. Le Z-score est 15. Comment traiter cette observation?

A) Supprimer automatiquement (Z > 3)

B) NE PAS supprimer - investiguer car potentiellement une vraie fraude (le signal recherché)

C) Remplacer par la moyenne

D) Winsoriser au percentile 99

Réponse: B) NE PAS supprimer - investiguer car potentiellement une vraie fraude (le signal recherché)

En détection de fraude, les outliers SONT souvent les cas intéressants. Supprimer automatiquement reviendrait à éliminer les fraudes potentielles. Investigation contextuelle requise.

Q10. Vous utilisez Isolation Forest avec contamination=0.05 sur vos données. Que signifie ce paramètre?

A) Le modèle supprime 5% des données

B) Le modèle s'attend à ce qu'environ 5% des observations soient des anomalies

C) L'erreur attendue est de 5%

D) 5% des arbres sont utilisés

Réponse: B) Le modèle s'attend à ce qu'environ 5% des observations soient des anomalies

contamination indique la proportion estimée d'anomalies dans le dataset. C'est utilisé pour calibrer le seuil de décision.

Q11. La méthode MAD (Median Absolute Deviation) est préférée au Z-score quand:

A) Les données sont normalement distribuées

B) Les données sont fortement asymétriques ou contiennent déjà des outliers extrêmes

C) L'échantillon est très grand

D) Les données sont binaires

Réponse: B) Les données sont fortement asymétriques ou contiennent déjà des outliers extrêmes

La MAD utilise la médiane (robuste) au lieu de la moyenne/écart-type (sensibles aux outliers). Elle est donc plus appropriée quand les données sont déjà contaminées.

Q12. Vous appliquez le winsorizing au percentile 1-99 sur les montants de transaction. Quelle est la conséquence sur la distribution?

A) La distribution devient normale

B) Les queues extrêmes sont “coupées” et remplacées par les valeurs seuils

C) La moyenne ne change pas

D) Toutes les valeurs deviennent identiques

Réponse: B) Les queues extrêmes sont “coupées” et remplacées par les valeurs seuils

Le winsorizing remplace les valeurs au-delà des percentiles par les valeurs seuils (ex: tout ce qui > P99 devient = P99). Cela limite l’impact des extrêmes sans supprimer d’observations.

Q13. Un outlier est identifié dans les données de revenus: 2,000,000 HTG/mois pour un employé déclaré. Après vérification, c’est le PDG d’une grande entreprise. Que faire?

A) Supprimer car c’est un outlier statistique

B) Conserver car c’est une valeur légitime bien que extrême

C) Remplacer par la médiane

D) L’exclure de tous les modèles

Réponse: B) Conserver car c’est une valeur légitime bien que extrême

Un outlier vérifié et légitime ne doit pas être supprimé. On peut le traiter spécifiquement (segment VIP) ou utiliser des méthodes robustes, mais pas le supprimer arbitrairement.

Q14. La transformation Box-Cox avec λ=0 équivaut à quelle transformation?

A) Identité (pas de transformation)

B) Transformation logarithmique

C) Transformation racine carrée

D) Transformation inverse

Réponse: B) Transformation logarithmique

Box-Cox avec λ=0 est un cas spécial qui correspond à log(Y). λ=0.5 correspond à √Y, λ=1 à l’identité, λ=-1 à 1/Y.

Q15. Dans un contexte de scoring, vous détectez que 2% des observations ont des valeurs extrêmes sur 3+ variables simultanément. Comment les traiter?

A) Supprimer les 2%

B) Créer un segment “profil atypique” séparé pour analyse spécifique

C) Imputer par les moyennes

D) Ignorer

Réponse: B) Créer un segment “profil atypique” séparé pour analyse spécifique

Des observations extrêmes sur plusieurs dimensions peuvent représenter un profil client spécifique (ex: très riches, fraudeurs). Les isoler pour analyse séparée est plus informatif que la suppression.

Section C: ACP Avancée (8 questions)

Q16. Après ACP, les 3 premières composantes expliquent 45%, 25%, et 15% de la variance. Combien de composantes retenir selon le critère de Kaiser?

A) 1 composante

B) 2 composantes

C) 3 composantes (celles avec eigenvalue > 1, si c’est le cas)

D) Impossible à dire sans les eigenvalues

Réponse: D) Impossible à dire sans les eigenvalues

Le critère de Kaiser retient les composantes avec eigenvalue > 1. La variance expliquée ne suffit pas - il faut connaître les eigenvalues réelles. (Indice: 85% cumulé avec 3 CP sur beaucoup de variables → probablement toutes > 1)

Q17. Le test de Bartlett donne p < 0.001 et le KMO = 0.82 pour votre matrice de corrélation. Que concluez-vous?

A) L’ACP n’est pas appropriée

B) L’ACP est appropriée - les corrélations sont significatives et l’échantillonnage est adéquat

C) Il faut plus de données

D) Les variables sont indépendantes

Réponse: B) L’ACP est appropriée - les corrélations sont significatives et l’échantillonnage est adéquat

Bartlett p < 0.05: la matrice n’est pas identité (corrélations existent). KMO > 0.8: très bon (0.6-0.8 acceptable, > 0.8 excellent). L’ACP est justifiée.

Q18. Sur le graphique des loadings de PC1, vous observez: revenus (+0.7), patrimoine (+0.6), épargne (+0.5), dettes (-0.4). Comment interpréter PC1?

A) PC1 représente le risque de crédit

B) PC1 représente la “richesse globale” ou capacité financière

C) PC1 n’a pas d’interprétation

D) PC1 représente l’âge du client

Réponse: B) PC1 représente la “richesse globale” ou capacité financière

Les loadings positifs sur revenus, patrimoine, épargne et négatif sur dettes suggèrent que PC1 capture une dimension de “situation financière globale” ou richesse.

Q19. Après rotation Varimax, les loadings deviennent plus “tranchés” (proches de 0 ou ±1). Quel est l’objectif de cette rotation?

A) Augmenter la variance expliquée

B) Simplifier l’interprétation en maximisant les loadings forts et minimisant les faibles

C) Réduire le nombre de composantes

D) Normaliser les scores

Réponse: B) Simplifier l’interprétation en maximisant les loadings forts et minimisant les faibles

Varimax est une rotation orthogonale qui rend les facteurs plus facilement interprétables en “poussant” les loadings vers les extrêmes (0 ou ±1).

Q20. Vous utilisez les scores PC1 et PC2 comme features dans un modèle de clustering. Pourquoi standardiser les données AVANT l’ACP?

A) Pour accélérer le calcul

B) Pour éviter que les variables à grande échelle dominent les composantes

C) La standardisation n’est pas nécessaire pour l’ACP

D) Pour augmenter la variance

Réponse: B) Pour éviter que les variables à grande échelle dominent les composantes

L’ACP est sensible à l’échelle. Sans standardisation, une variable en millions (revenus) dominera une variable en unités (nb_enfants). La standardisation met toutes les variables sur un pied d’égalité.

Q21. Le scree plot montre un “coude” net après la 4ème composante. Que suggère le critère du coude?

A) Retenir 1 composante

B) Retenir 4 composantes (avant le coude)

C) Retenir toutes les composantes

D) Le scree plot n’est pas informatif

Réponse: B) Retenir 4 composantes (avant le coude)

Le critère du coude (elbow rule) suggère de retenir les composantes avant le point où la pente de décroissance devient faible (“coude”). Ici, 4 composantes.

Q22. En ACP, la communalité d’une variable est de 0.35. Que cela signifie-t-il?

A) La variable est très bien représentée par les composantes retenues

B) Seulement 35% de la variance de cette variable est expliquée par les composantes retenues

C) La variable a un loading de 0.35

D) 35% des observations ont cette variable

Réponse: B) Seulement 35% de la variance de cette variable est expliquée par les composantes retenues

La communalité = somme des loadings² sur les composantes retenues. 0.35 signifie que 65% de la variance de cette variable n’est pas capturée → potentiellement mal représentée.

Q23. Vous effectuez une ACP pour réduire 50 variables à 10 composantes pour un modèle de scoring. Quelle précaution prendre pour le déploiement?

A) Recalculer l’ACP à chaque prédiction

B) Sauvegarder les paramètres de l’ACP (mean, std, composantes) pour transformer les nouvelles données de la même façon

C) L’ACP n’a pas besoin d’être sauvegardée

D) Utiliser des composantes différentes pour chaque client

Réponse: B) Sauvegarder les paramètres de l’ACP (mean, std, composantes) pour transformer les nouvelles données de la même façon

Pour le scoring en production, les nouveaux clients doivent être transformés avec les MÊMES paramètres que l’entraînement. Il faut sauvegarder le PCA.fit() (composantes, moyennes).

Section D: ANOVA Avancée (7 questions)

Q24. Une ANOVA à deux facteurs (Agence × Segment) donne un effet d'interaction significatif (p < 0.01). Comment interpréter?

A) Les deux facteurs sont indépendants

B) L'effet d'un facteur dépend du niveau de l'autre facteur

C) Aucun effet n'est significatif

D) Il faut refaire l'analyse

Réponse: B) L'effet d'un facteur dépend du niveau de l'autre facteur

Une interaction significative signifie que l'effet de l'Agence sur la variable dépendante varie selon le Segment (et vice versa). Il faut analyser les effets simples.

Q25. Le test de Levene donne p = 0.03 avant votre ANOVA. Quelle action est appropriée?

A) Procéder normalement avec l'ANOVA classique

B) Utiliser une alternative robuste (Welch's ANOVA) ou transformation

C) Abandonner l'analyse

D) Augmenter la taille d'échantillon

Réponse: B) Utiliser une alternative robuste (Welch's ANOVA) ou transformation

Levene p < 0.05 rejette l'homogénéité des variances. L'ANOVA classique est sensible à cette violation. Welch's ANOVA ou transformation peuvent être utilisés.

Q26. Après ANOVA one-way significative (p < 0.001), le test post-hoc Tukey HSD montre que seul le groupe "VIP" diffère significativement des autres. Que conclure?

A) Tous les groupes sont différents

B) Le segment VIP a une moyenne significativement différente, les autres segments sont similaires entre eux

C) L'ANOVA était un faux positif

D) Il faut un autre test

Réponse: B) Le segment VIP a une moyenne significativement différente, les autres segments sont similaires entre eux

Tukey HSD compare toutes les paires. Si seules les comparaisons impliquant VIP sont significatives, c'est VIP qui "tire" la différence globale.

Q27. L'Eta-carré (η^2) de votre ANOVA est 0.15. Comment l'interpréter?

A) L'effet est négligeable

B) 15% de la variance totale est expliquée par le facteur - effet modéré à grand

C) Le facteur explique 85% de la variance

D) η^2 n'a pas d'interprétation

Réponse: B) 15% de la variance totale est expliquée par le facteur - effet modéré à grand

η^2 est la taille d'effet en ANOVA. Conventions: 0.01 = petit, 0.06 = moyen, 0.14 = grand. Donc 0.15 est un effet de grande taille.

Q28. Dans une ANOVA à mesures répétées (même clients avant/après campagne marketing), quelle hypothèse supplémentaire doit être vérifiée?

A) Indépendance des observations

B) Sphéricité (égalité des variances des différences entre niveaux)

C) Les données doivent être ordinales

D) Absence de valeurs manquantes

Réponse: B) Sphéricité (égalité des variances des différences entre niveaux)

Pour l'ANOVA à mesures répétées, l'hypothèse de sphéricité (testée par Mauchly) doit être vérifiée. Si violée, utiliser corrections (Greenhouse-Geisser, Huynh-Feldt).

Q29. Vous effectuez 10 tests post-hoc après ANOVA. Sans correction, quel est le risque approximatif de commettre au moins une erreur de Type I (si α=0.05)?

A) 5%

B) Environ 40% ($1 - 0.95^{10}$)

C) 50%

D) 0%

Réponse: B) Environ 40% ($1 - 0.95^{10}$)

Sans correction, le risque familywise d'erreur Type I est 1 - (1-α)^k ≈ 1 - 0.95^10 ≈ 0.40 ou 40%. C'est pourquoi Bonferroni ou Tukey corrigent.

Q30. La correction de Bonferroni avec 10 comparaisons ajuste α à:

A) 0.05

B) 0.005 (0.05/10)

C) 0.5

D) 0.10

Réponse: B) 0.005 (0.05/10)

Bonferroni: α_ajusté = α/k = 0.05/10 = 0.005. Chaque test individuel doit avoir p < 0.005 pour être considéré significatif.

Résumé des Concepts Clés

Valeurs Manquantes

- **MCAR:** Test de Little, imputation simple OK
- **MAR:** Imputation conditionnelle, MICE
- **MNAR:** Le plus problématique, indicateur de manquant utile
- **Data leakage:** Toujours split AVANT imputation

Outliers

- **Détection:** IQR, Z-score, MAD, Isolation Forest
- **Contexte fraude:** Ne PAS supprimer automatiquement
- **Traitement:** Winsorizing, transformation, segmentation

ACP

- **Kaiser:** Eigenvalue > 1
- **Coude:** Scree plot
- **KMO > 0.6:** Échantillonnage adéquat
- **Varimax:** Simplifie interprétation

ANOVA

- **Levene:** Homogénéité variances
- **Post-hoc:** Tukey HSD
- **$\eta^2 > 0.14$:** Grand effet
- **Bonferroni:** α/k

Score: ___/30