Ce document a été mis en ligne par l'organisme FormaV®.

Toute reproduction, représentation ou diffusion, même partielle, sans autorisation préalable, est strictement interdite.

Pour en savoir plus sur nos formations disponibles, veuillez visiter :
formav.co/explorer

BACHELOR UNIVERSITAIRE DE TECHNOLOGIE

SCIENCE DES DONNÉES

ÉVALUATION DE LA COMPÉTENCE : MODÉLISER LES DONNÉES DANS UN CADRE STATISTIQUE, NIVEAU 2, PARCOURS EXPLORATION ET MODÉLISATION STATISTIQUE

SESSION 2026

Durée : 3 heures

SPÉCIALITÉ / UNITÉ	COEFF
SCIENCE DES DONNÉES	3,0 INDICATIF, SELON LES DONNÉES FOURNIES

Moyens de calculs et matériel autorisés :

L'usage de calculatrice avec mode examen actif est autorisé. L'usage de calculatrice sans mémoire, « type collège », est autorisé.

Aucun document ou matériel autorisé en dehors de ceux fournis avec le sujet.

Dès que le sujet vous est remis, assurez-vous qu'il est complet.

Ce sujet comporte 9 pages numérotées de 1/9 à 9/9.

Les documents réponses en fin de sujet sont à rendre avec la copie.

DOSSIER SUJET

MISE EN SITUATION PROFESSIONNELLE ET CAHIER DES CHARGES : RÉSEAU VÉLOCITÉ, SERVICE D'ABONNEMENT DE VÉLOS EN LIBRE-SERVICE, LYON

Le réseau VéloCité souhaite repérer les abonnés susceptibles de ne pas renouveler leur abonnement au cours du mois suivant. Le service relation client demande une démarche de modélisation documentée afin de prioriser les actions de fidélisation.

Vous disposez des documents techniques DT 1 à DT 3. La variable à prédire est **resiliation**, égale à 1 si l'abonné résilie au cours du mois suivant et à 0 sinon. L'unité statistique est l'abonné. Les données couvrent 2 400 abonnés observés sur une période de référence définie avant la résiliation.

Votre travail doit traiter la qualité des données, la préparation, le choix et l'interprétation d'un modèle, son évaluation sur des données distinctes de celles utilisées pour l'ajustement, puis les conditions d'un usage opérationnel. Les modalités de durée et de matériel indiquées dans ce document s'appliquent à cette évaluation.

PARTIE 1 : Analyse de la problématique et fiabilisation des données (4 points)

Définition de la cible, qualité des données et prévention des biais

À partir du cahier des charges et du DT 1, caractérisez le problème statistique avant toute modélisation. Justifiez les opérations proposées en fonction de l'usage métier visé.

Extrait du dictionnaire de données et des contrôles qualité

Variable	Nature	Description	Valeurs manquantes
resiliation	Binaire	1 si résiliation au mois suivant, 0 sinon	0 %
anciennete_mois	Quantitative	Ancienneté de l'abonnement en mois	0 %
trajets_moyens_semaine	Quantitative	Nombre moyen de trajets par semaine sur la période de référence	4 %
contacts_support_90j	Quantitative discrète	Nombre de contacts au support pendant les 90 jours précédents	0 %
remise_active	Binaire	1 si une remise est active pendant la période de référence, 0 sinon	2 %
date_resiliation_effective	Date	Date de prise d'effet de la résiliation	Disponible uniquement après résiliation

Les taux de valeurs manquantes sont calculés sur les 2 400 lignes.

1.1 (1,5 point) Formulez l'unité statistique, la variable cible et l'horizon de prédiction. Indiquez s'il s'agit d'un problème de régression ou de classification et précisez pourquoi.

1.2 (1,5 point) Calculez le nombre de valeurs manquantes attendu pour *trajets_moyens_semaine* et pour *remise_active*. Proposez, pour chacune, une règle de traitement applicable sans utiliser les informations du jeu de test.

1.3 (1 point) Expliquez pourquoi *date_resiliation_effective* ne doit pas être utilisée comme variable explicative au moment du calcul du score. Proposez une règle de séparation des données cohérente avec la finalité de prédiction.

PARTIE 2 : Préparation et construction du modèle (7 points)

Choix d'une méthode, préparation des variables et interprétation des paramètres

Les données ont été réparties aléatoirement de façon stratifiée sur la cible : 1 680 abonnés pour l'apprentissage et 720 pour le test. Les paramètres de préparation ont été calculés uniquement sur l'échantillon d'apprentissage. Pour les valeurs manquantes de *trajets_moyens_semaine*, la médiane d'apprentissage est de 3,1 trajets par semaine. Pour *remise_active*, la modalité la plus fréquente dans l'apprentissage est 0.

Un modèle de régression logistique a été ajusté sur l'échantillon d'apprentissage. Les coefficients sont présentés dans le DT 2. Pour un abonné donné, la probabilité estimée est définie par $p = \frac{1}{1+\exp(-\eta)}$, où η est la combinaison linéaire des variables et des coefficients.

Coefficients du modèle logistique ajusté

Terme	Coefficient estimé
Constante	1,10
anciennete_mois	-0,08
trajets_moyens_semaine	-0,55
contacts_support_90j	0,42
remise_active	0,30

Les variables explicatives sont codées comme indiqué dans le DT 1.

2.1 (2 points) Décrivez une chaîne de préparation adaptée aux quatre variables explicatives du modèle, depuis le contrôle des valeurs manquantes jusqu'à l'ajustement. Précisez quelles opérations doivent être estimées sur l'apprentissage puis appliquées sans réestimation au test.

2.2 (2 points) Calculez η, puis la probabilité estimée de résiliation pour un abonné ayant 3 mois d'ancienneté, 2,4 trajets moyens par semaine, 2 contacts au support et aucune remise active. Détaillez le calcul et arrondissez la probabilité à trois décimales.

2.3 (1,5 point) Interprétez le signe du coefficient associé à *trajets_moyens_semaine*. Calculez l'odds ratio correspondant à une augmentation d'un trajet moyen par semaine, à autres variables constantes, et interprétez-le sans le confondre avec une variation de probabilité.

2.4 (1,5 point) Le DT 2 donne les AIC du modèle avec variables explicatives et du modèle avec constante seule, ajustés sur le même échantillon. Choisissez le modèle que l'AIC privilégie et justifiez votre réponse. Indiquez une limite de cette comparaison.

PARTIE 3 : Validation, comparaison des seuils et diagnostic (6 points)

Évaluation hors apprentissage et prise en compte du coût métier des erreurs

Le modèle a été évalué sur les 720 abonnés du jeu de test, non utilisés pour l'ajustement. Le seuil de décision transforme la probabilité estimée en prédiction de résiliation : prédire 1 si la probabilité est supérieure ou égale au seuil. Les résultats aux seuils 0,40 et 0,50 sont fournis dans le DT 3.

Pour l'exercice, le coût d'un faux négatif, abonné qui résilie mais n'est pas repéré, est fixé à 15 unités monétaires. Le coût d'un faux positif, abonné repéré à tort, est fixé à 4 unités monétaires. Le coût total considéré est donc $C = 15 \times FN + 4 \times FP$.

Matrices de confusion sur le jeu de test

Seuil	Réel résiliation, prédit résiliation	Réel résiliation, prédit non-résiliation	Réel non-résiliation, prédit résiliation	Réel non-résiliation, prédit non-résiliation
0,40	72	36	54	558
0,50	60	48	30	582

Les lignes correspondent à la classe réelle et les colonnes à la classe prédite.

3.1 (3 points) Pour le seuil 0,40, calculez la précision, le rappel, la spécificité et l'exactitude globale. Donnez les formules utilisées et arrondissez les résultats à trois décimales. Identifiez le nombre de vrais positifs, faux négatifs, faux positifs et vrais négatifs.

3.2 (2 points) Calculez le coût total des erreurs pour chacun des deux seuils. Au regard des coûts fournis, retenez un seuil et justifiez ce choix. Indiquez une conséquence possible de ce choix pour les équipes de fidélisation.

3.3 (1 point) Indiquez une limite de l'évaluation présentée et proposez un contrôle complémentaire pertinent avant un déploiement réel.

PARTIE 4 : Restitution et suivi en conditions d'usage (3 points)

Prise en compte du domaine d'application et adaptation à l'évolution des données

Le modèle pourrait être utilisé chaque semaine par le service relation client. Les équipes doivent comprendre les scores, pouvoir vérifier la qualité des données entrantes et suivre la performance après mise en service.

4.1 (1,5 point) Rédigez une restitution de trois à cinq phrases à destination du responsable relation client. Elle doit présenter le résultat principal de la validation, le seuil retenu et une précaution d'interprétation du score.

4.2 (1,5 point) Proposez deux indicateurs ou contrôles à suivre après déploiement, en précisant pour chacun ce qu'il permettrait de détecter et une action possible en cas de dégradation.

DOSSIER TECHNIQUE ET RESSOURCES

Documents ressources nécessaires à la résolution du sujet

DT 1 : DICTIONNAIRE DES DONNÉES ET PROTOCOLE DE CONSTITUTION DE LA CIBLE

Population étudiée : abonnés actifs au début du mois de référence. La cible *resiliation* vaut 1 lorsque l'abonnement prend fin pendant le mois qui suit la période de référence, et 0 sinon. L'effectif total est de 2 400 abonnés, dont 720 résiliations et 1 680 non-résiliations.

Les quatre variables explicatives candidates sont calculées uniquement à partir d'informations disponibles avant la date de prédiction. *anciennete_mois* est un entier positif ou nul ; *trajets_moyens_semaine* est une moyenne calculée sur les huit semaines précédentes ; *contacts_support_90j* est un compte entier ; *remise_active* vaut 1 si une remise est active, 0 sinon. Les taux de valeurs manquantes sont ceux du tableau ci-dessous.

La variable *date_resiliation_effective* est postérieure à la date de prédiction pour les abonnés résiliés. Elle est fournie uniquement pour établir la cible lors de la constitution historique du jeu de données.

Dictionnaire et contrôles à réaliser

Variable	Type	Domaine attendu ou définition	Valeurs manquantes
resiliation	Binaire	0 ou 1, cible observée après la période de référence	0 %
anciennete_mois	Quantitative entière	Valeur supérieure ou égale à 0	0 %
trajets_moyens_semaine	Quantitative	Valeur supérieure ou égale à 0	4 %
contacts_support_90j	Quantitative entière	Valeur supérieure ou égale à 0	0 %
remise_active	Binaire	0 ou 1	2 %
date_resiliation_effective	Date	Date d'effet, connue après la période de prédiction	Dépend de la classe

Contrôles réalisés avant toute imputation ou modélisation.

DT 2 : AJUSTEMENT DE LA RÉGRESSION LOGISTIQUE ET COMPARAISON PAR AIC

La régression logistique modélise la probabilité de résiliation conditionnellement aux variables explicatives. L'écriture du modèle est $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, où p est la probabilité estimée de résiliation, x_1 l'ancienneté en mois, x_2 le nombre moyen de trajets par semaine, x_3 le nombre de contacts au support et x_4 l'indicateur de remise active.

Coefficients estimés sur l'échantillon d'apprentissage : constante 1,10 ; ancienneté -0,08 ; trajets moyens par semaine -0,55 ; contacts au support 0,42 ; remise active 0,30. AIC du modèle avec les quatre variables explicatives : 194. AIC du modèle avec constante seule : 201. Les deux AIC sont calculés sur le même

échantillon d'apprentissage et la même variable cible. Un AIC plus faible indique un compromis plus favorable entre ajustement et complexité parmi les modèles comparés.

La fonction logistique est $p = \frac{1}{1+\exp(-\eta)}$, où η est le membre de droite de l'équation de régression logistique.

Coefficients et critères d'information

Élément	Valeur
Constante	1,10
anciennete_mois	-0,08
trajets_moyens_semaine	-0,55
contacts_support_90j	0,42
remise_active	0,30
AIC, modèle avec quatre variables	194
AIC, modèle avec constante seule	201

Estimations obtenues sur les seules données d'apprentissage.

DT 3 : ÉVALUATION SUR LE JEU DE TEST ET COÛTS MÉTIER DES ERREURS

Le jeu de test contient 720 abonnés : 108 ont résilié et 612 n'ont pas résilié. Il est indépendant de l'échantillon d'apprentissage utilisé pour estimer les coefficients. Les matrices ci-dessous sont calculées à partir des probabilités produites par le modèle sur le jeu de test.

Pour une matrice de confusion, VP désigne les résiliations correctement repérées, FN les résiliations manquées, FP les abonnés signalés à tort et VN les non-résiliations correctement reconnues. Les indicateurs sont : précision $= \frac{VP}{VP+FP}$, rappel $= \frac{VP}{VP+FN}$, spécificité $= \frac{VN}{VN+FP}$, exactitude $= \frac{VP+VN}{N}$. Les coûts unitaires convenus pour cette étude sont de 15 unités monétaires par faux négatif et de 4 unités monétaires par faux positif.

Comptages par seuil de décision

Seuil	VP	FN	FP	VN
0,40	72	36	54	558
0,50	60	48	30	582

Seuil appliqué à la probabilité estimée : prédiction positive si le score est supérieur ou égal au seuil.

NE RIEN ÉCRIRE DANS CETTE PARTIE

DR 1 : Document réponse à rendre avec la copie : Questions 1.2 et 2.1 **À RENDRE AVEC LA COPIE**

Compléter directement le tableau. Pour les effectifs de valeurs manquantes, donner un nombre entier d'observations. Pour les règles de préparation, préciser la règle retenue et l'échantillon sur lequel ses paramètres sont calculés.

Contrôle et préparation des variables

Variable	Effectif manquant attendu	Traitement proposé	Paramètre estimé sur l'apprentissage
trajets_moyens_semaine	.....	.....	.....
remise_active	.....	.....	.....
date_resiliation_effective	.....	.....	.....

Compléter les cases signalées par des points de suspension.

Cadre réservé aux justifications, calculs complémentaires ou schémas du candidat :

.....

.....

.....

.....

.....

.....

.....

NE RIEN ÉCRIRE DANS CETTE PARTIE

DR 2 : Document réponse à rendre avec la copie : Questions 3.1 et 3.2 **À RENDRE AVEC LA COPIE**

Compléter les résultats arrondis à trois décimales lorsque cela s'applique. Présenter les coûts sous forme d'entiers d'unités monétaires.

Indicateurs et comparaison des seuils

Indicateur ou calcul	Résultat
Précision au seuil 0,40	.....
Rappel au seuil 0,40	.....
Spécificité au seuil 0,40	.....
Exactitude au seuil 0,40	.....
Coût total au seuil 0,40	.....
Coût total au seuil 0,50	.....
Seuil retenu et justification	.....

Utiliser les définitions et coûts fournis dans le sujet.

Cadre réservé aux justifications, calculs complémentaires ou schémas du candidat :

.....

.....

.....

.....

.....

.....

.....

Copyright © 2027 FormaV®. Tous droits réservés.

FormaV® est une marque déposée auprès de l'INPI sous le numéro 5227897.

Ce document a été élaboré par FormaV® avec le plus grand soin afin d'accompagner chaque apprenant vers la réussite de ses examens. Son contenu (textes, graphiques, méthodologies, tableaux, exercices, concepts, mises en forme) constitue une œuvre protégée par le droit d'auteur.

Toute copie, partage, reproduction, diffusion ou mise à disposition, même partielle, gratuite ou payante, est strictement interdite sans accord préalable et écrit de FormaV®, conformément aux articles L.111-1 et suivants du Code de la propriété intellectuelle. Dans une logique anti-plagiat, FormaV® se réserve le droit de vérifier toute utilisation illicite, y compris sur les plateformes en ligne ou sites tiers.

En utilisant ce document, vous vous engagez à respecter ces règles et à préserver l'intégrité du travail fourni. La consultation de ce document est strictement personnelle.

Merci de respecter le travail accompli afin de permettre la création continue de ressources pédagogiques fiables et accessibles.