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

BUT SCIENCE DES DONNÉES • CORRIGÉ DÉTAILLÉ

DÉMARCHE DE MODÉLISATION

Session 2026 • Code sujet : **26-BUT-SD-EP1** • Durée : **3 heures** • Coefficient : **3,0**
indicatif, selon les données fournies

Sommaire

Cliquez sur une partie pour y accéder directement

1. **Analyse de la problématique et fiabilisation des données** (3 questions) — **Page 3**
2. **Préparation des variables et construction du modèle logistique** (4 questions) — **Page 5**
3. **Validation du modèle, comparaison des seuils et diagnostic** (3 questions) — **Page 8**
4. **Restitution et suivi du modèle en conditions d'usage** (2 questions) — **Page 10**
5. **Ce qu'il faut retenir & Questions fréquentes (FAQ)** (Synthèse & Q&A Examen) — **Page 11**

Sujet officiel (PDF) : https://butsd.fr/assets/annales/but-sd/C8 - Réaliser l'ensemble de la démarche de modélisation/2026/BUT Science des Données - C8 - Réaliser l'ensemble de la démarche de modélisation - Session 2026 - 1-Sujet.pdf

1. Analyse de la problématique et fiabilisation des données

Le réseau VéloCité veut repérer les abonnés susceptibles de résilier au cours du mois suivant afin de prioriser les actions de fidélisation. L'étude porte sur 2 400 abonnés et mobilise un dictionnaire de variables, des taux de données manquantes et des informations sur la disponibilité temporelle des données.

Définition du problème statistique

Question 1.1. 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.

L'unité statistique est un abonné actif au début du mois de référence. La cible **resiliation** vaut 1 si l'abonnement prend fin pendant le mois suivant la période de référence, et 0 sinon : l'horizon est donc ce mois suivant. Il s'agit d'une **classification binaire**, car le résultat attendu est l'une de deux classes, et non une valeur numérique continue.

> ✓ RÉPONSE À RETENIR
>
> Unité : abonné. Cible : resiliation (0 ou 1). Horizon : mois suivant la période de référence. Problème : classification binaire.

Valeurs manquantes

Question 1.2. 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. Complétez également le document réponse DR 1.

Les effectifs manquants attendus se calculent sur les 2 400 abonnés : 4 % pour les trajets et 2 % pour les remises. On impute les variables explicatives à partir de paramètres estimés sur l'apprentissage uniquement, puis on applique ces mêmes paramètres au test.

Variable	Manquants attendus	Traitement proposé	Paramètre estimé sur l'apprentissage
trajets_moyens_semaine	96	Imputation par la médiane	3,1
remise_active	48	Imputation par le mode	0
date_resiliation_effective	Dépend de la classe	Exclure des prédicteurs; établir la cible	Aucun

La date de résiliation sert uniquement à construire la cible historique; elle ne fait pas partie des variables à imputer pour le score.

DR 1 complété

Variable	Effectif manquant attendu	Traitement proposé	Paramètre estimé sur l'apprentissage
trajets_moyens_semaine	96	Médiane	3,1
remise_active	48	Mode	0
date_resiliation_effective	Selon classe	Exclure des X	Aucun

✓ RÉPONSE À RETENIR

Trajets : 96 valeurs manquantes, imputées par 3,1. Remises : 48, imputées par 0. Paramètres estimés sur l'apprentissage; date de résiliation exclue des prédicteurs.

Prévention des biais

Question 1.3. 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.

date_resiliation_effective n'est connue qu'après la période de prédiction et révèle, pour les abonnés concernés, l'événement que le modèle doit anticiper. L'utiliser comme prédicteur créerait une fuite de cible; cette information n'est donc utilisée que pour constituer la cible historique. Il faut évaluer le modèle sur des abonnés distincts de ceux de l'apprentissage, idéalement issus d'une période ultérieure pour reproduire l'usage futur. Les paramètres de préparation sont estimés sur l'apprentissage, puis appliqués sans réestimation au test.

✓ RÉPONSE À RETENIR

Exclure date_resiliation_effective des prédicteurs. Tester sur des données distinctes, de préférence postérieures à l'apprentissage, sans réestimer la préparation sur le test.

2. Préparation des variables et construction du modèle logistique

Les données sont réparties entre un échantillon d'apprentissage de 1 680 abonnés et un jeu de test de 720 abonnés, selon un tirage stratifié sur la cible. Une régression logistique est ajustée à partir de quatre variables explicatives, avec les coefficients, les règles d'imputation et les AIC fournis dans les documents techniques.

Chaîne de préparation

Question 2.1. 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. Complétez également le document réponse DR 1.

Contrôler d'abord les domaines : valeurs quantitatives non négatives, indicateurs binaires codés 0 ou 1, cible binaire. La date de résiliation, connue après la prédiction, sert uniquement à construire la cible et doit être exclue des prédicteurs.

Variable	Manquants attendus	Traitement	Paramètre estimé sur l'apprentissage
trajets_moyens_semaine	96	Imputation par la médiane	3,1
remise_active	48	Imputation par le mode	0
date_resiliation_effective	1 680 structurels	Exclusion des prédicteurs	Aucun

Vérifier aussi les deux variables explicatives complètes. Estimer les imputations sur les 1 680 abonnés d'apprentissage, puis appliquer ces mêmes valeurs aux 720 abonnés de test, sans réestimation. Ajuster ensuite la régression logistique sur l'apprentissage.

DR 1 complété

Variable	Effectif manquant attendu	Traitement proposé	Paramètre estimé sur l'apprentissage
trajets_moyens_semaine	96	Médiane	3,1 apprentissage
remise_active	48	Mode	0 apprentissage
date_resiliation_effective	1 680 structurels	Exclure des prédicteurs	Aucun

✓ RÉPONSE À RETENIR

Imputer les trajets par 3,1 et les remises par 0, paramètres estimés sur l'apprentissage. Exclure la date de résiliation des prédicteurs.

Calcul d'une probabilité estimée

Question 2.2. 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.

On remplace chaque variable par sa valeur et son coefficient dans la combinaison linéaire. La remise est codée 0 :

$$\eta = 1,10 - 0,08 \times 3 - 0,55 \times 2,4 + 0,42 \times 2 + 0,30 \times 0 = 0,38.$$

La fonction logistique donne alors $p = 1/(1+\exp(-0,38)) \approx 0,594$. La probabilité est arrondie à trois décimales. Il s'agit du score estimé par le modèle pour cet abonné, et non d'une certitude de résiliation ou d'une décision sans seuil.

> **✓ RÉPONSE À RETENIR**
>
> $\eta = 0,38$ et $p = 0,594$, soit une probabilité estimée de résiliation de 59,4 %.

Interprétation d'un coefficient

Question 2.3. 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.

Le coefficient de *trajets_moyens_semaine* est négatif : à ancienneté, contacts et remise identiques, une hausse des trajets est associée à une baisse des log-odds de résiliation. Pour un trajet moyen supplémentaire, l'odds ratio est $\exp(-0,55) \approx 0,577$. Les odds de résiliation sont donc multipliées par 0,577, soit une baisse d'environ 42,3 %. Cette valeur décrit les odds, pas la probabilité : la variation de probabilité dépend de son niveau initial et des autres variables.

> **✓ RÉPONSE À RETENIR**
>
> $OR = \exp(-0,55) \approx 0,577$: les odds diminuent d'environ 42,3 % par trajet moyen supplémentaire, toutes choses égales par ailleurs.

Comparaison par AIC

Question 2.4. À partir des AIC fournis dans le DT 2 pour le modèle avec variables explicatives et le modèle avec constante seule, choisissez le modèle privilégié et justifiez votre réponse. Indiquez une limite de cette comparaison.

Le modèle à quatre variables est privilégié : son AIC vaut 194, contre 201 pour le modèle à constante seule. L'AIC étant plus faible, il indique ici un compromis ajustement-complexité plus favorable; l'écart est de 7 points. Cette comparaison est relative aux deux modèles considérés, ajustés sur le même échantillon et la même cible. Elle ne prouve pas que le modèle prédira bien sur de nouvelles données. Il faut donc compléter ce choix par une évaluation hors apprentissage.

✓ **RÉPONSE À RETENIR**

L’AIC privilégie le modèle à quatre variables : 194 contre 201. Cela ne garantit pas ses performances sur de nouvelles données.

3. Validation du modèle, comparaison des seuils et diagnostic

Le modèle est évalué sur 720 abonnés du jeu de test, indépendant de l'échantillon d'apprentissage, avec deux seuils de décision et leurs matrices de confusion. Les faux négatifs et faux positifs ont des coûts métier différents, à prendre en compte pour comparer les seuils.

Indicateurs de classification

Question 3.1. Pour le seuil 0,40, calculez la précision, le rappel, la spécificité et l'exactitude globale. Donnez les formules utilisées, arrondissez les résultats à trois décimales et identifiez les nombres de vrais positifs, faux négatifs, faux positifs et vrais négatifs. Complétez le document réponse DR 2.

Au seuil 0,40, la matrice donne VP = 72, FN = 36, FP = 54 et VN = 558, sur 720 abonnés. La précision vaut VP/(VP + FP) = 72/126 ; le rappel, VP/(VP + FN) = 72/108 ; la spécificité, VN/(VN + FP) = 558/612 ; l'exactitude, (VP + VN)/720 = 630/720. Les indicateurs sont arrondis à trois décimales.

DR 2 complété, indicateurs au seuil 0,40

Indicateur ou calcul	Résultat
Précision au seuil 0,40	0,571
Rappel au seuil 0,40	0,667
Spécificité au seuil 0,40	0,912
Exactitude au seuil 0,40	0,875
Coût total au seuil 0,40	
Coût total au seuil 0,50	
Seuil retenu et justification	

✓ RÉPONSE À RETENIR

VP = 72, FN = 36, FP = 54, VN = 558. Précision : 0,571 ; rappel : 0,667 ; spécificité : 0,912 ; exactitude : 0,875.

Coût métier et choix du seuil

Question 3.2. 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. Complétez le document réponse DR 2.

Le coût total suit la formule C = 15 × FN + 4 × FP. On compare les deux seuils avec les coûts unitaires indiqués :

1. Seuil 0,40 : 15 × 36 + 4 × 54 = 756.
2. Seuil 0,50 : 15 × 48 + 4 × 30 = 840.

Le seuil 0,40 minimise le coût parmi les deux options évaluées. Il permet de repérer davantage de résiliations, mais conduit les équipes à contacter davantage d'abonnés qui n'auraient pas résilié.

DR 2 complété, comparaison des coûts

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	756
Coût total au seuil 0,50	840
Seuil retenu et justification	0,40

✓ RÉPONSE À RETENIR

Retenir le seuil 0,40 : coût total de 756 unités monétaires, contre 840 au seuil 0,50.

Limites de l’évaluation

Question 3.3. Indiquez une limite de l’évaluation présentée et proposez un contrôle complémentaire pertinent avant un déploiement réel.

Le résultat du test unique ne garantit ni une performance stable dans le temps ni des résultats équivalents pour tous les profils d’abonnés. Avant le déploiement, compléter l’évaluation par une validation temporelle sur une période ultérieure.

- Conserver le seuil retenu et calculer les indicateurs ainsi que les coûts métier sur cette période.
- En cas de dégradation, examiner l’évolution des données et réévaluer le seuil ou le modèle avant généralisation.

✓ RÉPONSE À RETENIR

Limite : un seul jeu de test. Contrôle : validation temporelle complémentaire et suivi des indicateurs et coûts métier.

4. Restitution et suivi du modèle en conditions d’usage

Le service relation client envisage d’utiliser le modèle chaque semaine pour orienter ses actions de fidélisation. Les équipes doivent comprendre les scores, contrôler la qualité des données entrantes et surveiller la performance après le déploiement.

Restitution au responsable métier

Question 4.1. 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.

Sur le jeu de test indépendant, le modèle atteint une exactitude de 0,875 et repère 72 des 108 résiliations, soit un rappel de 0,667 au seuil de 0,40. Parmi les deux seuils évalués, 0,40 minimise le coût estimé des erreurs : 756 unités, contre 840 au seuil de 0,50. Ce seuil est retenu pour prioriser les actions de fidélisation. Le score est une estimation, pas une certitude individuelle : il guide les prises de contact sans garantir qu’un abonné résiliera.

> **✓ RÉPONSE À RETENIR**
>
> Retenir le seuil de 0,40 pour prioriser les contacts ; interpréter les scores comme des estimations, non comme des certitudes.

Suivi après déploiement

Question 4.2. Proposez deux indicateurs ou contrôles à suivre après déploiement. Pour chacun, précisez ce qu’il permettrait de détecter et une action possible en cas de dégradation.

Après le déploiement, suivre régulièrement les deux éléments suivants :

- **Performance sur les résiliations observées :** recalculer le rappel et le coût des erreurs lorsque les résultats réels sont connus. Une baisse du rappel ou une hausse du coût invite à réexaminer le seuil et, si nécessaire, à réajuster le modèle.
- **Qualité et évolution des données entrantes :** contrôler les valeurs manquantes, les domaines attendus et la distribution des variables. Une dégradation peut révéler un problème de collecte ou une évolution des abonnés ; corriger la collecte et réentraîner le modèle si le décalage persiste.

> **✓ RÉPONSE À RETENIR**
>
> Surveiller la performance sur les résultats réels et la qualité des données ; corriger la collecte ou réajuster le modèle si nécessaire.

Ce qu'il faut retenir de ce sujet (Session 2026)

Bilan de l'épreuve : Cette épreuve de démarche de modélisation, session 2026, code 26-BUT-SD-EP1, dure 3 heures et porte un coefficient indicatif de 3,0 selon les données fournies. À partir du cas VéloCité, elle étudie la prédiction des résiliations par régression logistique. Elle évalue la qualité et la préparation des données, l'interprétation du modèle, sa validation, le choix d'un seuil selon les coûts métier et le suivi opérationnel.

Notions-clés, formules & règles à maîtriser

- La cible binaire resiliation indique si l'abonné résilie pendant le mois suivant la période de référence.
- Pour une variable quantitative, l'imputation par la médiane doit être estimée sur l'apprentissage puis appliquée au test sans réestimation.
- En régression logistique, $p = 1 / (1 + \exp(-\eta))$ et une variation d'un prédicteur multiplie les odds par exp(coefficient).
- La précision vaut VP/(VP+FP), le rappel VP/(VP+FN), la spécificité VN/(VN+FP) et l'exactitude (VP+VN)/N.

Points de vigilance & attentes du jury

- Écarter les variables connues seulement après la date de prédiction pour éviter une fuite d'information.
- Estimer les paramètres de préparation sur l'apprentissage seulement, puis les appliquer tels quels au jeu de test.
- Distinguer l'effet d'un coefficient ou d'un odds ratio d'une variation directe de probabilité.
- Choisir le seuil selon les coûts des erreurs et vérifier aussi les conséquences opérationnelles du nombre d'abonnés signalés.

Questions fréquentes sur cette épreuve (Q&A Examen)

Q : Comment choisir une règle d'imputation sans contaminer le jeu de test ?

Séparez d'abord les données en apprentissage et test. Estimez la médiane ou la modalité la plus fréquente uniquement sur l'apprentissage, puis utilisez ce paramètre inchangé pour remplacer les valeurs manquantes dans les deux ensembles.

Q : Comment interpréter l'odds ratio d'un coefficient logistique ?

Pour une augmentation d'une unité du prédicteur, à autres variables constantes, les odds sont multipliées par exp(coefficient). Cet effet porte sur les odds et non sur la probabilité, dont la variation dépend aussi de la situation de départ.

Q : Pourquoi ne suffit-il pas de choisir le seuil qui donne la meilleure exactitude ?

L'exactitude ne distingue pas le coût des erreurs. Si manquer un abonné qui résilie coûte davantage qu'un signalement injustifié, comparez les seuils avec la fonction de coût fournie et considérez la charge créée pour les équipes.

Q : Que faut-il surveiller après le déploiement du modèle ?

Contrôlez la qualité et la distribution des données entrantes, puis mesurez les performances lorsque les résiliations réelles sont connues. Une dégradation peut conduire à vérifier les données, réévaluer le seuil ou réentraîner le modèle.

Sujet officiel associé (PDF) : https://butsd.fr/assets/annales/but-sd/C8 - Réaliser l'ensemble de la démarche de modélisation/2026/BUT Science des Données - C8 - Réaliser l'ensemble de la démarche de modélisation - Session 2026 - 1-Sujet.pdf
Ressources BUT Science des Données : Toutes les annales & corrigés • Fiches de révision complètes • Simulateur de notes d'examen

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.