Comparer sans se raconter d’histoire

Arbres, forêts aléatoires et validation temporelle

Aurélien Nicosia

À la fin de la matinée

Vous saurez choisir entre une régression logistique, un arbre et une forêt sans utiliser le futur pour fabriquer le gagnant.

Une matinée, trois décisions

1. Protéger le futur
2. Choisir avant le test
3. Ouvrir le futur et décider

> Chaque mission ajoute une règle, une preuve et une limite à la fiche finale.

Plus complexe signifie-t-il plus intelligent?

Votez: oui, non ou cela dépend.

Puis nommez la preuve qui pourrait vous faire changer d’avis.

Notre terrain: les demandes 311

Le fichier pédagogique contient 18 000 demandes créées en 2024.

Élément	Définition
Unité	une demande de service
Cible	non terminée dans les sept jours
Moment de prédiction	à la création de la demande
Usage pédagogique	comparer des probabilités futures

Une cible construite avec le futur

Pour savoir si une demande est terminée dans les sept jours, il faut observer ce qui arrive après sa création.

Cette information définit la réponse, mais elle n’est pas un prédicteur disponible au moment de la décision.

> Construire la cible avec le futur est permis. Prédire avec ce futur serait une fuite.

Train, validation, test: trois rôles

Ensemble	Rôle	Règle
entraînement, janvier à septembre	apprendre le prétraitement et les paramètres	réutilisable pendant le développement
validation temporelle, dans janvier à septembre	comparer les candidats sur des mois suivants	guide le choix avant le test
test futur, octobre à décembre	estimer la performance après le choix	ouvert une seule fois

> Si le test sert à choisir ou à régler le modèle, il devient une validation et ne confirme plus la performance finale.

Le futur ne ressemble pas exactement au passé

```
# A tibble: 4 × 4
  ensemble     verite           n proportion
  <chr>        <chr>        <int>      <dbl>
1 Entraînement Non terminée  6326      0.441
2 Entraînement Terminée      8025      0.559
3 Test futur   Non terminée  1446      0.396
4 Test futur   Terminée      2203      0.604
```

La part non terminée passe de 44.1% à 39.6%.

Six prédicteurs disponibles à la création

Temps	Demande	Contexte
jour de la semaine	activité	type de lieu
plage horaire	arrondissement	provenance

Nous n’utilisons ni le statut futur, ni sa date, ni le délai final.

Trois candidats, trois compromis

Modèle	Force	Fragilité
Logistique	structure globale inspectable	formes à spécifier
Arbre	règles et interactions lisibles	instabilité
Forêt	flexibilité et agrégation	interprétation indirecte

Choisir un modèle est une décision

La performance ne suffit pas. Il faut aussi considérer:

1. la stabilité dans le temps;
2. l’interprétabilité requise;
3. le coût de calcul et de surveillance;
4. les conséquences des erreurs;
5. la calibration des probabilités.

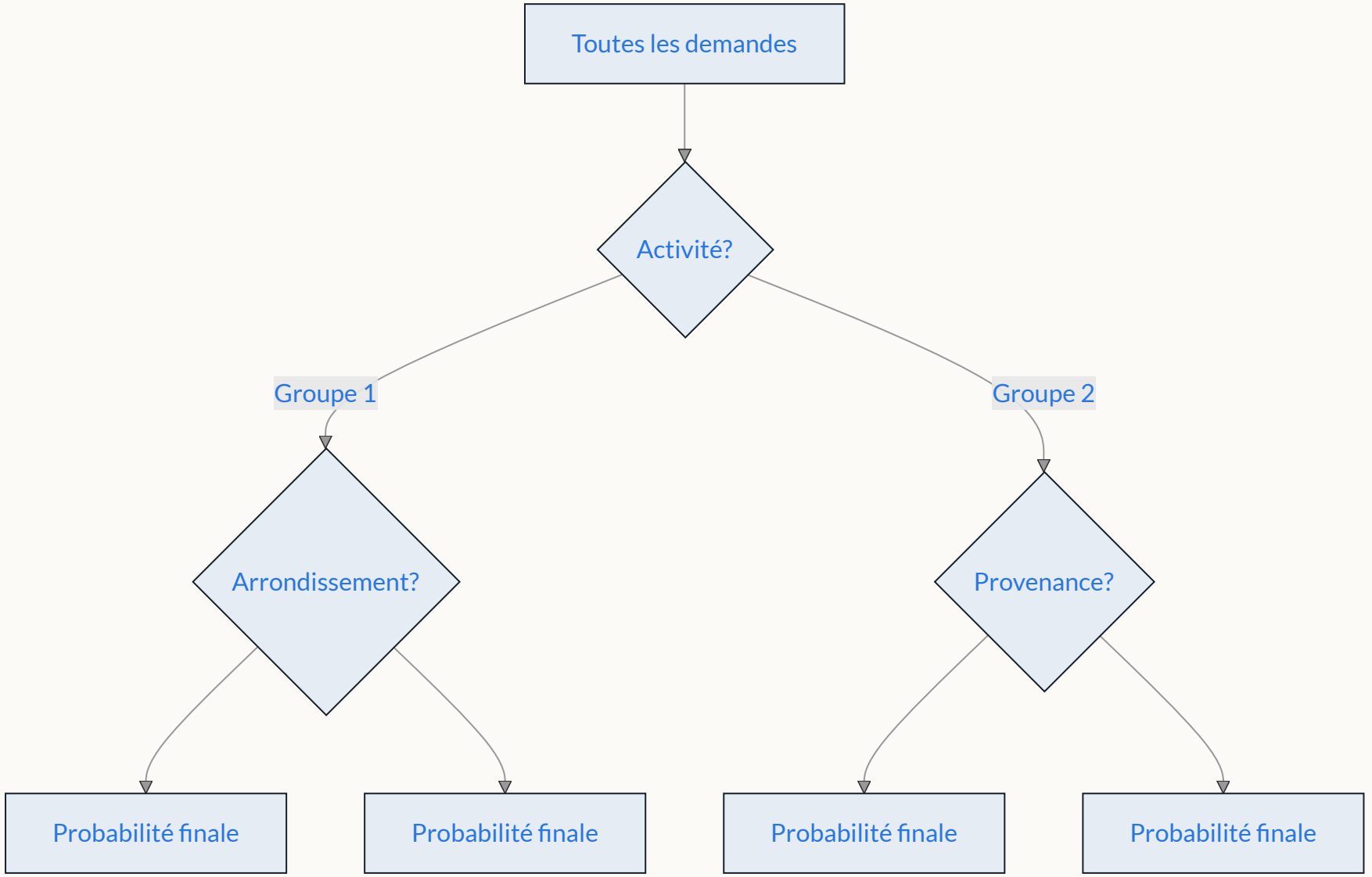
La logistique impose une structure globale

Elle additionne des effets sur l'échelle de la log-cote, puis transforme le résultat en probabilité.

Elle peut être très performante si la structure choisie est suffisante.

> Simple ne signifie ni naïf ni automatiquement stable dans le futur.

L’arbre pose des questions successives



Une coupure recherche de l’homogénéité

À chaque noeud, l’algorithme compare des coupures candidates et retient celle qui sépare le mieux les réponses.

Le processus recommence dans chaque groupe obtenu.

Il s’arrête lorsque les règles de complexité l’imposent.

Une feuille produit une probabilité

Si une feuille contient 30 demandes, dont 18 non terminées en sept jours, la probabilité estimée est:

$$18/30 = 0{,}60.$$

Le seuil de classement est une décision supplémentaire.

Un arbre profond peut mémoriser

Arbre peu profond	Arbre très profond
règles grossières	règles très spécifiques
biais plus élevé	biais plus faible
variance plus faible	variance plus élevée

> L'erreur d'entraînement récompense souvent une complexité qui échoue dans le futur.

Trois freins à la complexité

Paramètre	Question pratique
profondeur maximale	combien de questions successives?
taille minimale d'un noeud	combien de lignes avant de recouper?
pénalité de complexité	le gain justifie-t-il une branche?

La forêt perturbe et agrège

Pour chaque arbre:

1. rééchantillonner les observations;
2. proposer seulement quelques prédicteurs à chaque coupure;
3. construire l’arbre;
4. agréger les probabilités de 200 arbres.

Trois réglages de la forêt

Réglage	Valeur	Rôle
`trees`	200	taille de l'ensemble
`mtry`	4	prédicteurs candidats par coupure
`min_n`	30	taille minimale avant une coupure

Plus d'arbres stabilisent l'agrégation, mais ne corrigent pas un protocole biaisé.

L’importance prédictive répond à une question limitée

On permute les valeurs d’un prédicteur et on mesure la perte de performance.

Une grande perte signifie que le modèle utilisait cette information.

Elle ne donne ni le sens de l’association, ni un effet causal, ni une priorité de politique publique.

Flexibilité et variance forment un compromis

La place sur cet axe ne détermine pas le gagnant. Le futur tranche.

Le test futur reste fermé

Pourquoi répéter la question « mois suivant »?

Une seule séparation peut être chanceuse ou défavorable.

Plusieurs fenêtres montrent:

1. la performance moyenne;
2. la variabilité entre périodes;
3. les mois difficiles;
4. la sensibilité du classement des modèles.

Cinq fenêtres temporelles

```
# A tibble: 5 × 3
  pli    apprentissage             validation
  <chr>  <chr>                     <chr>
1 Slice1 2024-01-01 au 2024-04-30 2024-05-01 au 2024-05-31
2 Slice2 2024-02-01 au 2024-05-31 2024-06-01 au 2024-06-30
3 Slice3 2024-03-01 au 2024-06-30 2024-07-01 au 2024-07-31
4 Slice4 2024-04-01 au 2024-07-31 2024-08-01 au 2024-08-31
5 Slice5 2024-05-01 au 2024-08-31 2024-09-01 au 2024-09-30
```

Chaque modèle apprend sur quatre mois, puis prédit le mois suivant.

Pourquoi pas cinq plis aléatoires?

Plis aléatoires	Fenêtres temporelles
lignes contemporaines mélangées	passé vers mois suivant
question d’interpolation	question de déploiement futur
ignore l’ordre temporel	respecte l’ordre temporel

Les deux protocoles peuvent être valides, mais ils ne répondent pas à la même question.

Une recette commune évite un avantage caché

Les trois candidats reçoivent:

1. la même cible;
2. les mêmes prédicteurs;
3. les mêmes regroupements de catégories rares;
4. les mêmes fenêtres;
5. les mêmes mesures.

Le prétraitement doit apprendre dans chaque fenêtre

Problème	Traitement
catégorie jamais vue	niveau « Nouveau »
catégorie manquante	niveau « Non précisé »
catégorie très rare	regroupement « Autres »
prédicteur constant	retrait

> La recette est réestimée sur chaque sous-ensemble d’apprentissage.

Trois mesures, trois questions

Mesure	Question	Sens favorable
aire ROC	classe-t-il bien deux demandes?	élevée
aire précision-rappel	repère-t-il l'événement positif?	élevée
Brier	les probabilités sont-elles proches des résultats?	faible

Le score de Brier regarde la probabilité

Pour une observation:

$$(y - \hat{p})^2,$$

où $y = 1$ pour « non terminée » et $\hat{p}$ est la probabilité correspondante.

Une prédiction confiante et fausse est fortement pénalisée.

Mission 1: protéger le futur

Ouvrez la mission 1.

En 25 minutes, produisez:

1. la définition de la prédiction;
2. la liste des variables permises et interdites;
3. le schéma des cinq fenêtres;
4. les règles de comparaison équitable.

Débrief: notre protocole est maintenant verrouillé

Élément	Décision commune
événement	non terminée en sept jours
apprentissage	janvier à septembre
validation	cinq mois futurs successifs
test fermé	octobre à décembre
information	disponible à la création

Pause

10 h 20 à 10 h 30

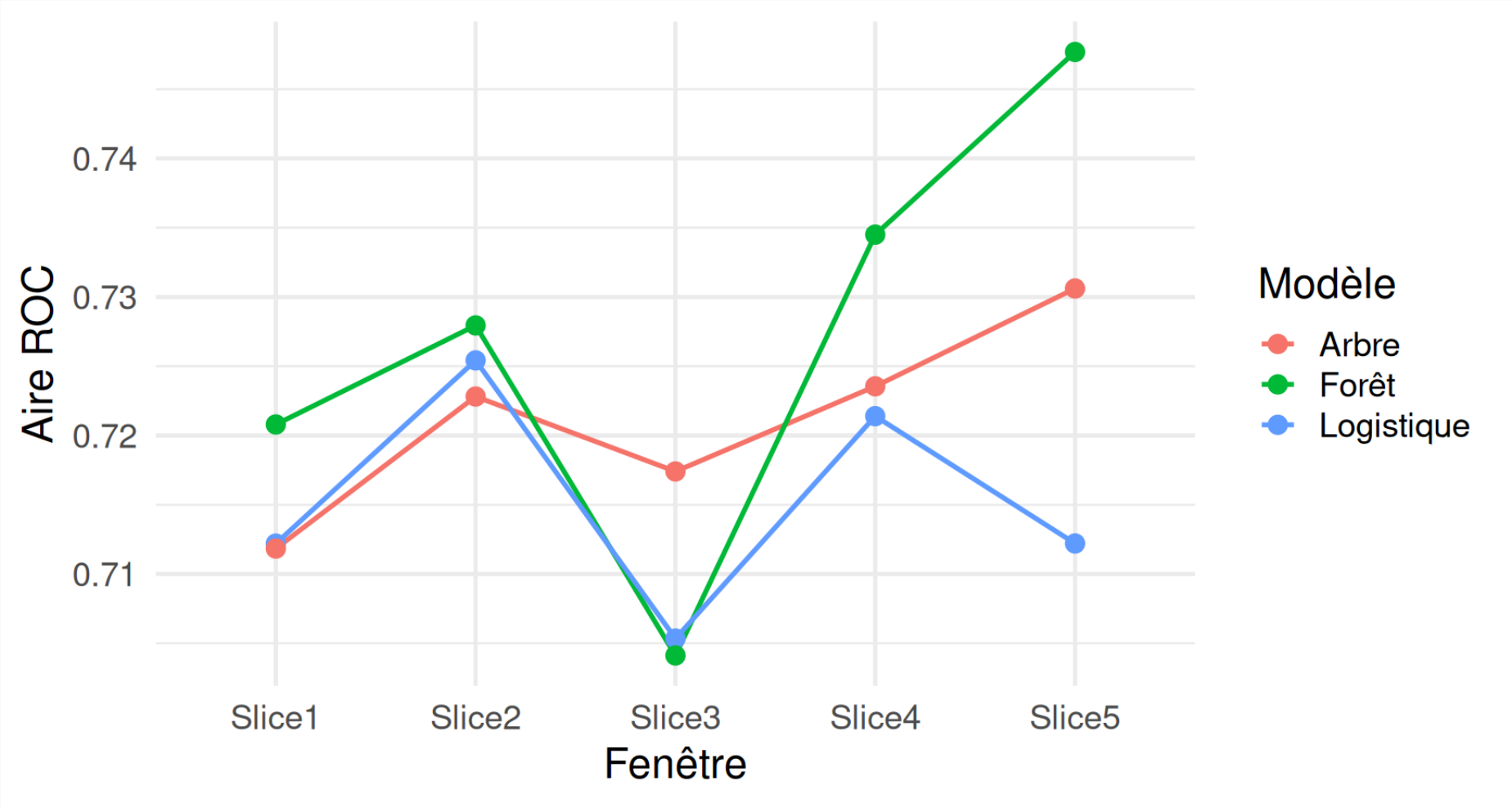
Reprise avec les résultats de validation, sans ouvrir le test futur.

Résultats moyens avant le test

```
# A tibble: 3 × 4
  modele     Brier `Aire précision-rappel` `Aire ROC`
  <chr>      <dbl>                   <dbl>      <dbl>
1 Logistique 0.214                   0.663      0.715
2 Arbre      0.213                   0.674      0.721
3 Forêt      0.210                   0.680      0.727
```

La forêt est meilleure en moyenne sur les trois mesures, mais les écarts sont modestes.

La moyenne ne suffit pas



La stabilité change la lecture

```
# A tibble: 3 × 4
  modele     minimum maximum ecart_type
  <chr>        <dbl>   <dbl>      <dbl>
1 Arbre        0.712   0.731    0.00705
2 Forêt        0.704   0.748    0.0162 
3 Logistique   0.705   0.725    0.00803
```

La forêt varie de 0.704 à 0.748. L'arbre varie de 0.712 à 0.731.

> Le meilleur score moyen n'est pas la meilleure réponse à tous les mandats.

Quatre mandats, quatre classements possibles

1. communication scientifique claire;
2. meilleur tri prédictif expérimental;
3. règle inspectable ligne par ligne;
4. infrastructure et surveillance minimales.

Le mandat doit être attribué avant l'ouverture du test.

Mission 2: choisir avant le test

Ouvrez la mission 2.

Mission 1 incomplète? Exécutez d'abord le bloc de reprise autonome au début de l'énoncé.

En 25 minutes:

1. comparez moyenne et stabilité;
2. classez les trois modèles pour votre mandat;
3. écrivez le choix et son compromis;
4. signez la décision avant de voir le test.

Débrief: une recommandation avant le test

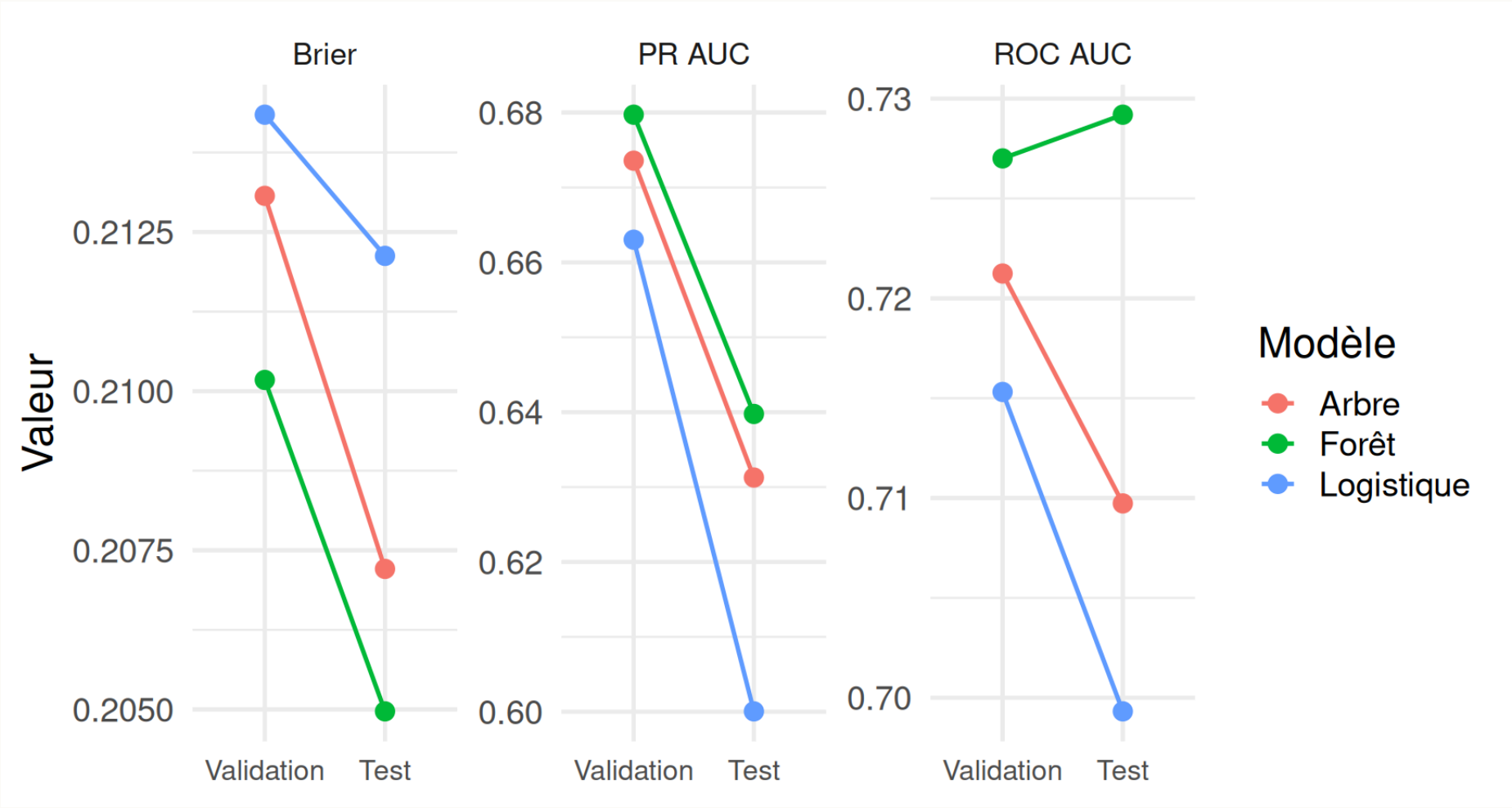
Une recommandation défendable contient:

1. le mandat;
2. le modèle retenu;
3. deux preuves chiffrées;
4. une preuve de stabilité ou d’instabilité;
5. un compromis accepté;
6. une condition de révision.

Le test futur n’est ouvert qu’une fois

```
# A tibble: 3 × 4
  modele     `Aire ROC` `Aire précision-rappel` Brier
  <chr>           <dbl>                   <dbl> <dbl>
1 Arbre           0.710                   0.631 0.207
2 Forêt           0.729                   0.640 0.205
3 Logistique      0.699                   0.600 0.212
```

Validation et test racontent-ils la même histoire?



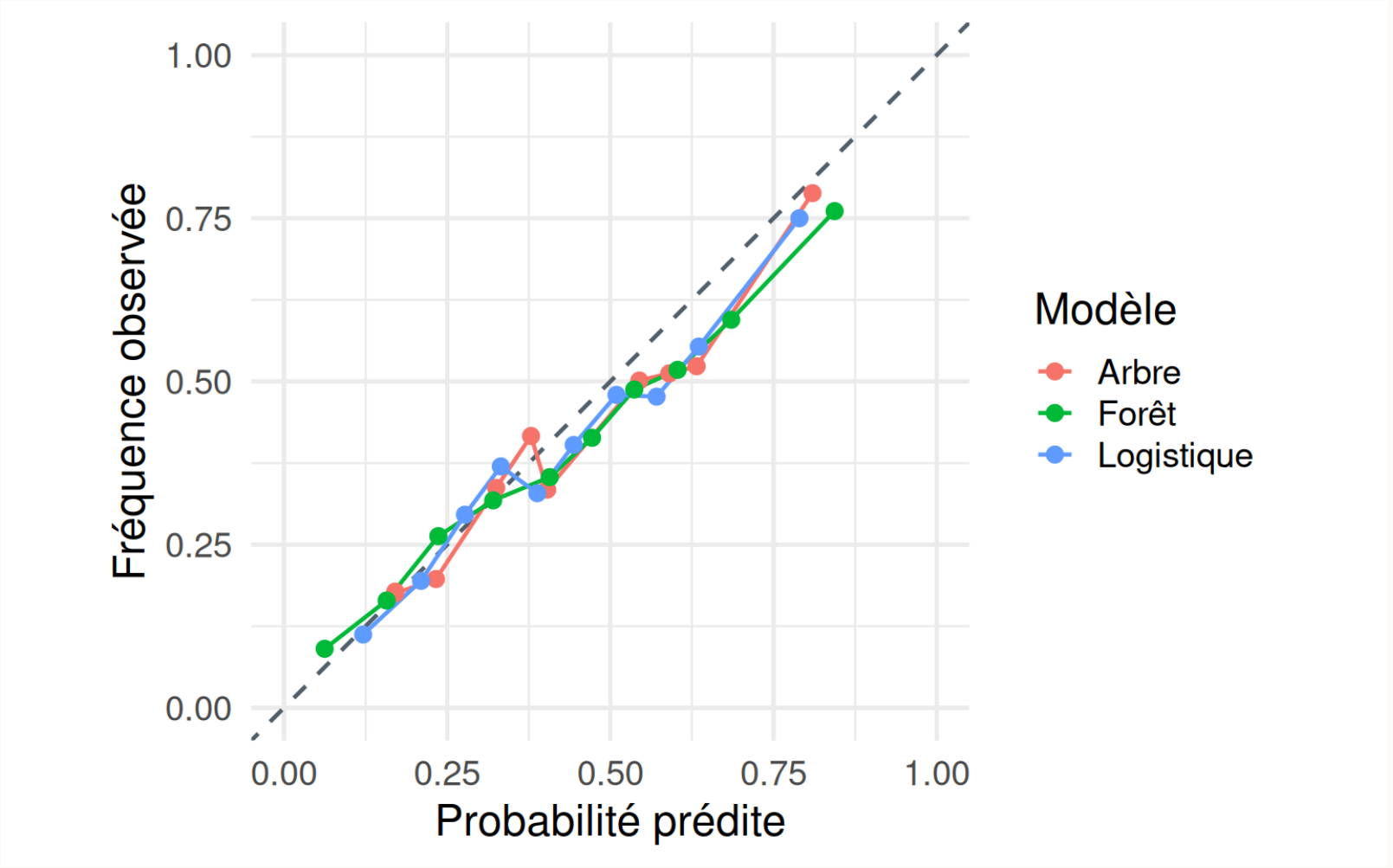
Si le test déçoit

Vous pouvez:

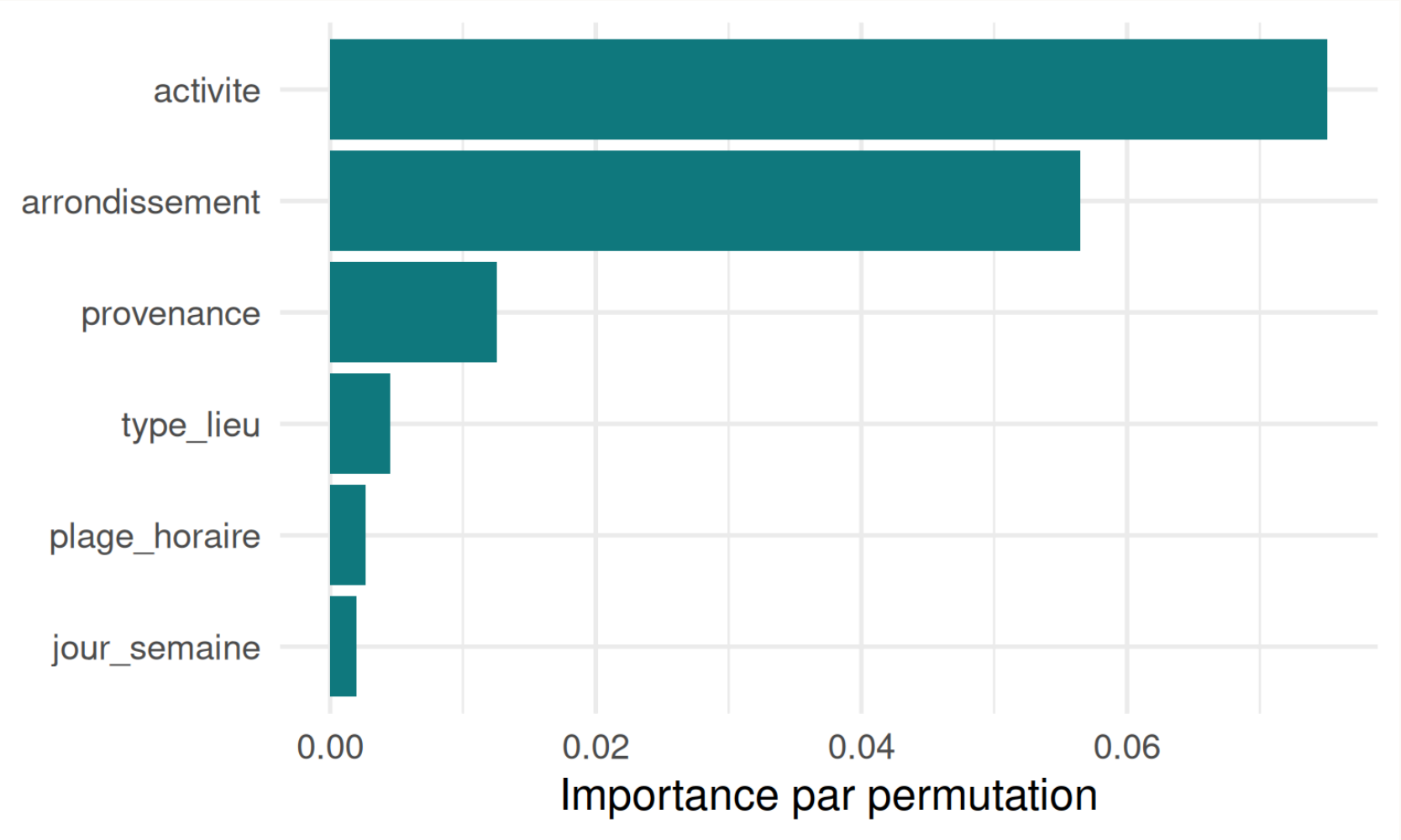
1. documenter l'écart;
2. formuler une nouvelle hypothèse;
3. modifier le pipeline;
4. attendre une nouvelle période indépendante.

Vous ne pouvez pas recycler silencieusement le test comme validation et confirmation.

La calibration vérifie les probabilités



L’importance n’est pas une cause



Une performance globale peut cacher des écarts

Avant un usage réel, il faudrait vérifier les résultats par:

1. période;
2. arrondissement;
3. type de demande;
4. canal de provenance;
5. volume et taux d'événement.

> Un modèle acceptable en moyenne peut échouer pour un sous-groupe ou une période.

Surveiller, réviser, retirer

Niveau	Question
données	les volumes et catégories ont-ils changé?
prédictions	les probabilités se déplacent-elles?
résultats	discrimination et calibration tiennent-elles?
décision	le modèle sert-il encore le mandat?

Une condition de retrait doit être écrite avant l’incident.

Mission 3: ouvrir le futur et décider

Ouvrez la mission 3.

En 25 minutes:

1. comparez validation et test;
2. examinez calibration et importance;
3. maintenez ou révisez votre choix explicitement;
4. rédigez la fiche et la règle de retrait.

La fiche modèle en sept rubriques

1. question et population cible;
2. moment de prédiction et variables interdites;
3. périodes d’apprentissage, de validation et de test;
4. modèles et mesures comparés;
5. décision et compromis;
6. limites et sous-groupes à examiner;
7. surveillance et condition de retrait.

Sept questions avant de conclure

1. Que représente une ligne?
2. Quel futur cherchons-nous à prédire?
3. Quelle information était disponible à ce moment?
4. Le protocole imite-t-il l’usage réel?
5. Le gain est-il stable et utile?
6. Les probabilités sont-elles calibrées?
7. Quand cesserons-nous d’utiliser le modèle?

Quel résultat vous ferait retirer un modèle déjà utilisé?

Écrivez une règle qui contient:

1. une mesure;
2. une période;
3. un seuil;
4. une action.