Analyse de Survie et Prédiction Conforme

Projet SigBERT

Auteurs : Mathis Derenne, Djida Boukari, Ines Nakhli
Date : 05 Février 2026

1. Contexte : Projet SigBERT

- **Objectif** : Modélisation du risque patient à partir de données cliniques complexes.
- **Données** : Signatures mathématiques issues d'embeddings de textes cliniques (OncoBERT).
- **Problématique** : Dépasser la simple estimation de risque (ranking) pour **quantifier l'incertitude** via des intervalles de prédiction fiables.

2. Données et Prétraitement

- **Jeu de données** : `df_study_L18_w6` (3555 patients).
- **Haute Dimension** : 756 variables (signatures).
- **Censure** : 33% (données censurées à droite).
- **Prétraitement** :
  - Standardisation.
  - **PCA** (Conservation de 90% de la variance).
  - Split Train/Test/Calibration.

3. Focus Baseline : Modèle de Cox

Configuration :

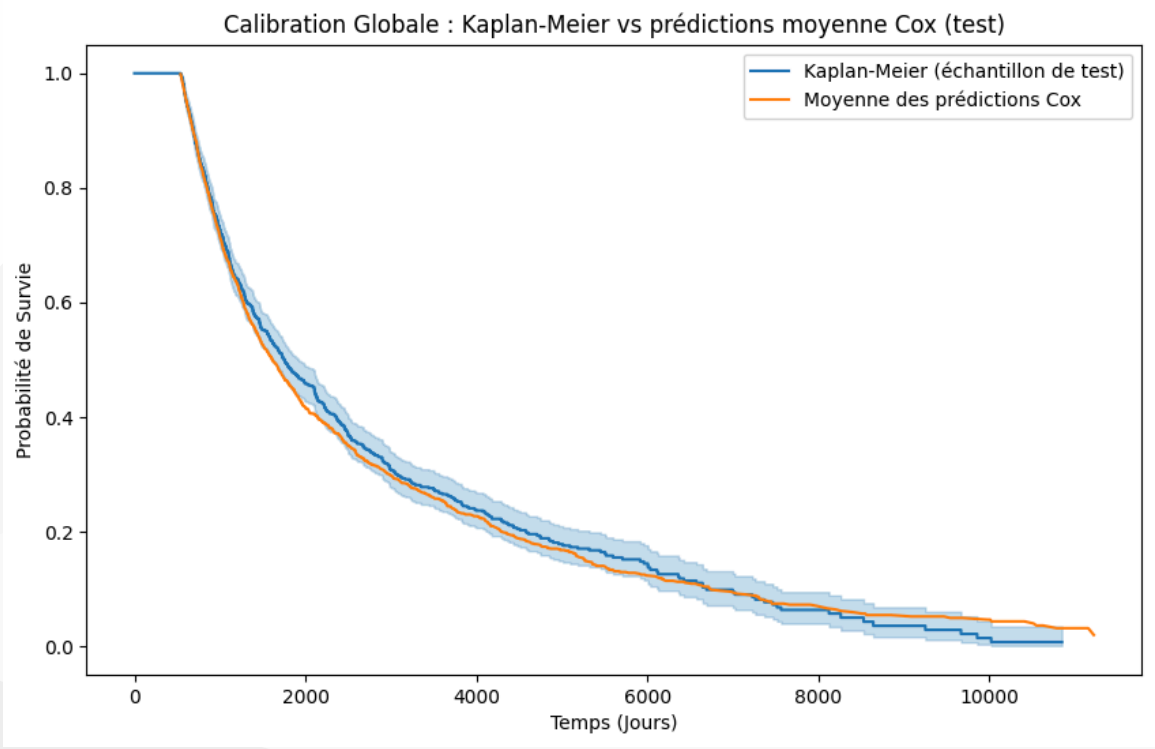
- **Optimisation** : Grid Search + Cross-Validation.
- **Meilleurs Hyperparamètres** : Pénalité 0.1 + ElasticNet ($L1/L2 = 0.5$).

penalizer	l1_ratio	mean_c_index	std_err
0.100	1.000	0.698	0.011
0.100	0.500	0.698	0.012
1.000	1.000	0.673	0.004
1.000	0.000	0.657	0.002
1.000	0.500	0.650	0.000
0.100	0.000	0.641	0.009

4. Focus Baseline : Modèle de Cox

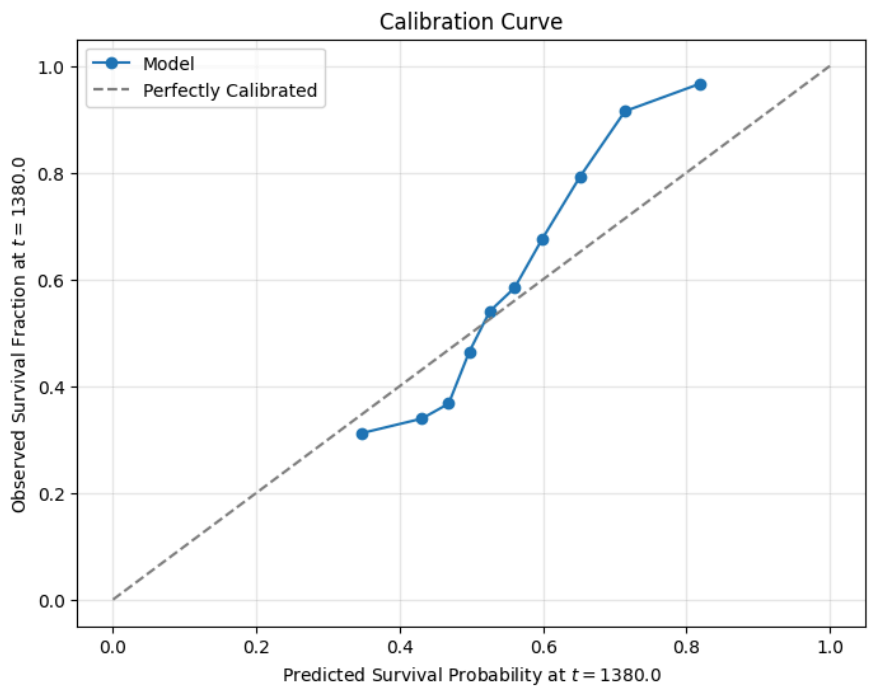
Analyse de Performance :

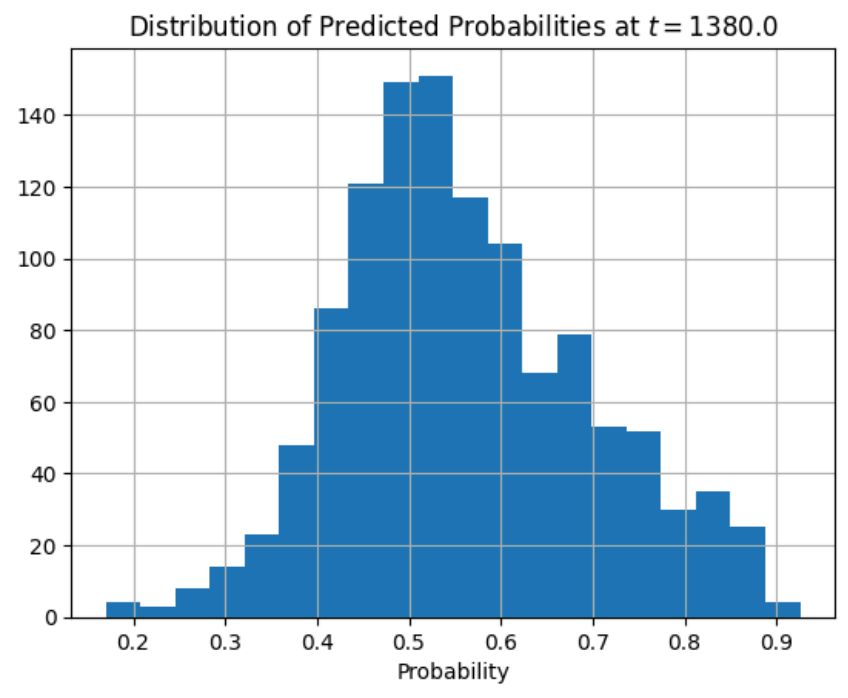
- **Calibration Globale** : ≈ Parfaite (courbe orange vs bleue).
- **Fiabilité** : Très bonne calibration locale.
- **Discrimination** : C-index 0.70.



4. Détails Performance Cox

Calibration Locale vs **Discrimination**





- **Gauche** : Bonne fiabilité (points sur la diagonale).
- **Droite** : Bonne capacité à distinguer les risques (étalement).

5. Comparaison des Modèles de Survie

Comparaison sur le jeu de test :

Modèle	C-index	IBS	Conclusion
Survival SVM	**0.73**	-	Meilleur Ranking
XGBoost AFT	0.72	-	**Choisi pour le Conforme**
Gradient Boosting	0.72	**0.18**	Meilleure Calibration
Random Forest	0.72	0.17	Robuste
Cox ElasticNet	0.70	0.20	Moins performant (non-linéarités)

6. Défi de la Prédiction Conforme en Survie

Objectif : Construire un intervalle $C(X)$ tel que $P(T \in C(X)) \geq 1 - \alpha$.

- **Obstacle** : La **censure**.
  - Pour un patient censuré, le temps réel T est inconnu.
  - Impossible de calculer le résidu $|T - \hat{T}|$ classique.
- **Risque** : Ignorer les censurés biaise l'échantillon vers les décès précoces (sous-estimation du risque).

7. Méthodologie : Weighted Split Conformal Prediction

Utilisation de l'**IPCW** (Inverse Probability of Censoring Weighting).

1. **Modèle** : XGBoost AFT (prédit $\log(T)$).
2. **Calibration** :
   - Calcul des scores $s_i = |\log(T_i) - \log(\hat{T}_i)|$ pour les non-censurés.
   - **Pondération** : $w_i = \frac{1}{\hat{G}(T_i)}$ où $\hat{G}$ est la survie de la censure (Kaplan-Meier).
   - Les patients ayant survécu longtemps (là où la censure est forte) comptent plus.
3. **Quantile** : Calcul du quantile pondéré $1 - \alpha$ des scores.

8. Résultats (Cible $\alpha = 0.1$)

Métrique	Valeur	Interprétation
Couverture Observée	**91.9%**	Objectif (>90%) atteint. Méthode valide.
Quantile $\hat{q}$	1.21	Facteur d'incertitude $\approx 3.35\times$.
Largeur Médiane	6504 j	Incertitude élevée malgré la validité.

> La méthode garantit la couverture statistique, mais révèle une incertitude intrinsèque importante dans la prédiction temporelle exacte.

9. Conclusion

- **Validation** : La méthode pondérée (WSCP) corrige efficacement le biais de censure.
- **Performance** : XGBoost AFT offre un bon compromis performance/flexibilité.
- **Impact** :
  - Pour le **classement** (triage), utiliser **Survival SVM**.
  - Pour la **planification** (estimation de durée), utiliser les intervalles conformes, en tenant compte de leur largeur.

Merci de votre attention.