Evaluation des Performances de Modèles de Survie par Cross-Validation : Une Approche Statistique en Science des Données

Auteurs : **BA Adama Abdoul** , **SAHAKIAN Marc Antonio** et **AIT SADI Charlotte**

M2 Mathématiques Appliquées et Statistiques parcours Data Science

2023-2024

Table des matières

Abstract

Ce projet explore l'utilisation de la cross-validation dans l'analyse de survie avec des modèles d'accélération du temps de défaillance (AFT) ou des modèles de Cox. L'objectif est d'évaluer la performance prédictive des modèles et leur généralisation aux nouvelles données, en se basant sur diverses mesures telles que le C-statistique, le test du log-rank, le score Brier, l'Integrated Brier Score (IBS), le Time-dependent Area Under the Curve (AUC), et d'autres critères classiques tels que l'AIC et le BIC. Le projet commence par la génération de données synthétiques, incluant la simulation de covariables corrélées, du temps jusqu'à l'événement, et de la censure. Ensuite, plusieurs modèles sont ajustés aux données synthétiques, et la cross-validation est utilisée pour évaluer et comparer leurs performances. La conclusion du projet inclut une analyse détaillée des résultats obtenus

1 Introdution

Dans le domaine de la Science des données, les techniques statistiques jouent un rôle central dans le développement de modèles robustes pour l'analyse prédictive. Un domaine particulièrement important est l'analyse de survie, qui explore le temps jusqu'à ce qu'un événement d'intérêt se produise. L'application des modèles de temps de défaillance accélérée (AFT) ou des modèles de Cox dans l'analyse de survie est répandue, et évaluer leur performance prédictive est essentiel pour une sélection de modèles efficace et une généralisation à de nouvelles données.
La validation croisée émerge comme une technique précieuse pour évaluer le pouvoir prédictif de ces modèles. Cependant, les critères traditionnels de comparaison entre les valeurs prédites et observées rencontrent un défi unique dans l'analyse de survie en raison de la censure. La censure survient lorsque le temps réel de l'événement est inconnu, rendant une comparaison directe impossible.
Pour surmonter cet obstacle, diverses mesures sont utilisées pour évaluer la performance des modèles AFT ou Cox. Ces mesures comprennent les statistiques C (C-index), le test du log-rank, le score Brier, le score Brier intégré (IBS), l'aire sous la courbe (AUC) dépendante du temps, ainsi que des critères d'information tels que l'AIC ou le BIC.
L'objectif principal de ce projet est de plonger dans les subtilités de ces scores de qualité et de se familiariser avec leur application dans le contexte de l'analyse de survie. Le projet se déroule en deux grande étapes : la préparation des données, l'ajustement des modèles.

2 Préparation des données

2.1 Création du Dataframe

Dans le cadre de ce rapport, un DataFrame a été élaboré, comprenant initialement 12 variables générées selon une distribution normale. Dans le processus de manipulation des variables, des transformations significatives ont été opérées. Plus précisément, la distribution de quatre variables a été modifiée en les rendant uniformes, en utilisant la fonction de répartition cumulative (CDF) de la distribution normale standard noté $\Phi(x)$. Elle représente la probabilité qu'une variable aléatoire normale standard soit inférieure ou égale à x. Mathématiquement, cela s'exprime comme suit :

$$\Phi(x) = \frac{1}{2}(1 + erf(\frac{x}{\sqrt{2}}))$$

où erf est la fonction d'erreur. La fonction de répartition cumulative inverse (ou fonction quantile) de la distribution normale standard, notée $\Phi^{-1}(y)$, donne la valeur x correspondant à une probabilité cumulée y. Cette fonction permet de convertir une variable normale en une variable continue. La formule de conversion est la suivante :

$$X_{uniforme} = \Phi^{-1}(X_{normale})$$

Par ailleurs, afin d'introduire une composante catégorique dans l'ensemble de données, deux variables ont été transformées en les catégorisant, enrichissant ainsi la complexité de l'information contenue dans le DataFrame. Ces modifications stratégiques visent à créer un ensemble de données varié et représentatif, propice à des analyses statistiques et exploratoires approfondies. Comme nous pouvons Visualiser maintenant les distributions de nos variables générées à l'aide d'histogrammes, offrant un aperçu graphique des tendances et des patterns au sein de notre jeu de données

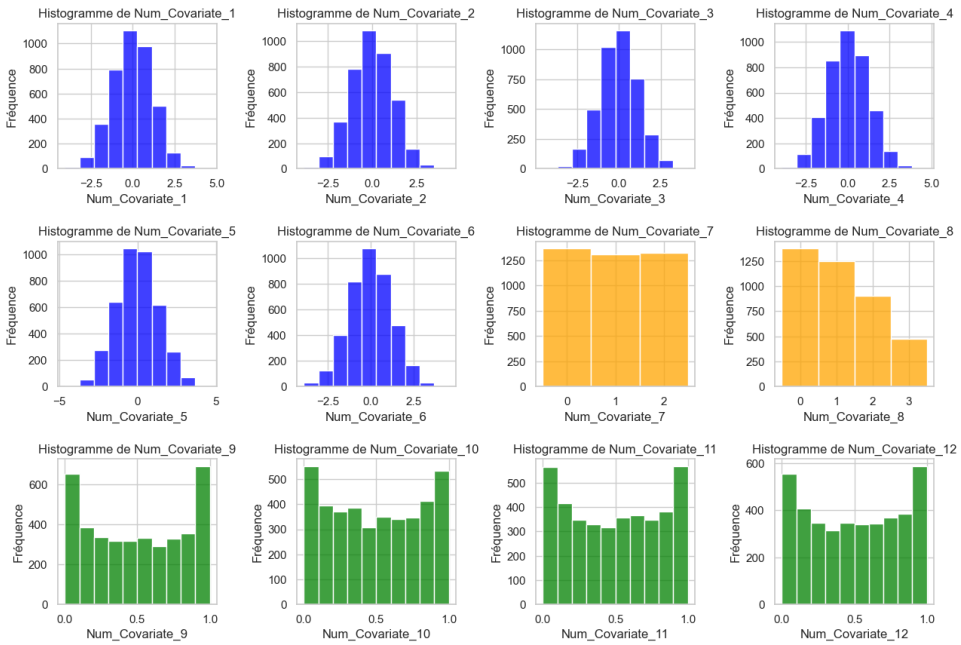

FIGURE 1 – Histogramme des Covariables

2.2 Creation des variables temps et evenement

2.2.1 Temps

Dans le cadre de cette étude, dix covariables ont été sélectionnées et utilisées pour générer le true-time-to-event en se basant sur une distribution de Weibull formulée comme suit :

$$T_{ref} = cx^{c-1}exp(-x^c)$$

où c est le paramétre de forme. Ensuite les temps sont générés par la formule suivante :

$$T = T_{ref} \times e^{\beta_0 + \sum_{i=1}^{10} \beta_i X_i}$$

où les X_i sont les covariables et les β_i sont des coefficients fixés.
On répète la même procédure, mais cette fois-ci en utilisant uniquement quatre covariables :

$$Z = Z_{ref} \times e^{G_0 + \sum_{i=1}^{4} G_i X_i}$$

2.2.2 Censure

La censure désigne la situation où la fin du suivi d'une observation survient avant qu'elle n'ait subi l'événement d'intérêt. Cela se produit lorsque l'on n'observe pas l'événement pour toutes les unités expérimentales pendant la période d'étude, et ces données censurées doivent être traitées spécifiquement dans l'analyse statistique. La censure est calculé par :

$$C = 1_{Z>T}$$

2.2.3 Evenement

Les evenements Y, sont calculé avec la formule suivante :

$$Y = \begin{cases} \text{T}, & \text{if } Z > T \\ \text{Z}, & \text{Sinon} \end{cases}$$

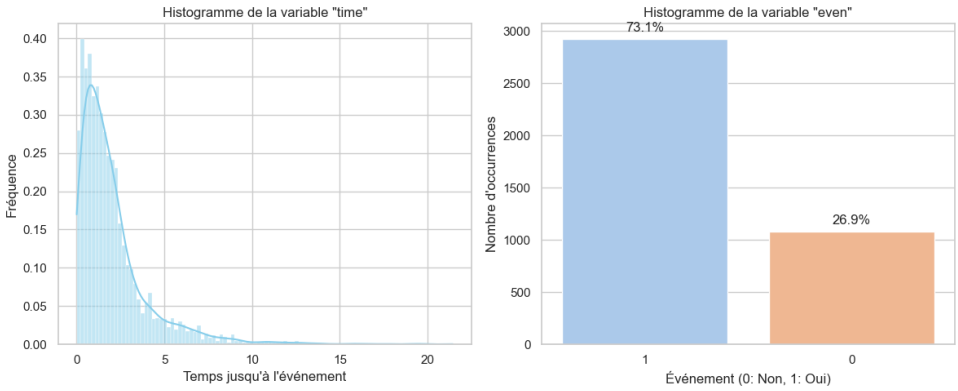

FIGURE 2 – Histogramme des Variables Temps et evenement

3 Les différentes modéles et transformation de la statistique de survie

3.1 Prétraitement des Données

Pour améliorer la qualité de nos données et les rendre compatibles avec notre modèle de survie, nous avons effectué une étape de prétraitement appelée "One-Hot Encoding". Cette opération consiste à convertir nos variables catégorielles en un format adapté à l'analyse de survie. Pour les variables $Num\_Covariate\_7$ et $Num\_Covariate\_8$, nous avons créé de nouvelles colonnes binaires pour chaque catégorie unique, éliminant ainsi la nécessité de les représenter sous forme de nombres.

3.2 Division des Données

Afin d'évaluer la performance de notre modèle de survie, nous avons divisé nos données en ensembles d'entraînement, de test et de validation. Cela nous permet de construire notre modèle sur l'ensemble d'entraînement, de le valider sur l'ensemble de validation, et enfin de le tester sur l'ensemble de test.

3.2.1 Modéle de Kaplan-Meier

La méthode de Kaplan-Meier est une technique non paramétrique utilisée en statistique de survie pour estimer la fonction de survie empirique à partir de données de survie censurées. Cette méthode est souvent employée lorsque l'on souhaite estimer la probabilité qu'un événement survienne avant un certain temps. La fonction de survie empirique est calculée à partir des temps d'événements observés et des temps de censure. La formule pour la fonction de survie empirique à un temps donné tt est la suivante :

$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i})$$

où :

- t_i sont les temps d'événements observés,
- d_i est le nombre d'événements à t_i,
- n_i est le nombre d'individus à risque à t_i

La fonction de survie est estimée de manière itérative à chaque instant où un événement survient.

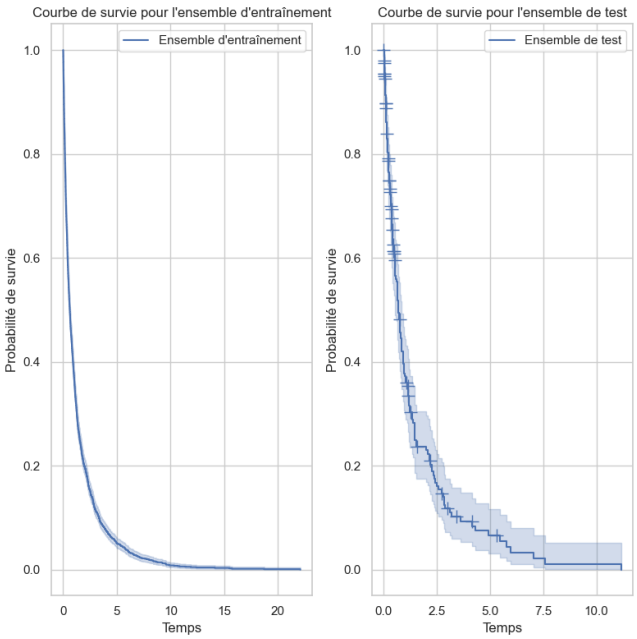

FIGURE 3 – Courbe de Kaplan Maier des données d'entrainement et test

3.2.2 Modéle à risque proportionel de Cox

La régression de Cox, est un modèle semi-parametrique utilisé en statistique de survie pour analyser le temps jusqu'à un événement. Ce modèle repose sur la fonction de risque $h(t)$, qui représente le risque instantané d'événement à un moment donné. La régression de Cox exprime le log de cette fonction de risque comme une combinaison linéaire de covariables, ce qui permet de modéliser l'effet des variables explicatives sur le risque de manière exponentielle. La fonction de risque (h(t)h(t)) dans le modèle de Cox est définie comme suit :

$$h(t) = h_0(t) \exp(\sum_{i=1}^{k} \beta_i X_i)$$

où :

- $h(t)$ est le risque instantané à un moment t.
- $h_0(t)$ est la fonction de risque de base, représentant le risque de référence non modifié par les covariables,
- X_i sont les covariables,
- β_i sont les coefficients de régression, mesurant l'effet de chaque covariable sur le risque.

La log-vraisemblance partielle du modèle de Cox est maximisée pour estimer les coefficients β_i. L'hypothèse sous-jacente est que le rapport des risques pour deux individus reste constant au fil du temps. Ainsi, l'interprétation des coefficients β_i se fait en termes de rapport de risque instantané.

3.2.3 Modéle Accélération du Temps de Défaillance(AFT)

Le modèle AFT (Accélération du Temps de Défaillance), est une approche en statistique de survie qui modélise directement le temps de survie d'un événement. Contrairement au modèle de Cox, le modèle AFT exprime le temps de survie comme une fonction linéaire des covariables, affectée par un paramètre d'accélération. Mathématiquement, la fonction de survie dans le modèle AFT est définie comme suit :

$$S(t) = S_0(\exp(-\sum_{i=1}^{k} \beta_i X_i))^{\lambda}$$

où :

- $S(t)$ représente la fonction de survie à un momentt.
- $S_0(t)$ est la fonction de survie de base non modifiée par les covariables,
- X_i sont les covariables,
- β_i sont les coefficients de régression, mesurant l'effet de chaque covariable sur le risque,
- λ est le paramétre d'accélération.

Le modèle AFT offre une interprétation directe des coefficients, indiquant comment chaque covariable modifie le temps de survie. Par exemple, un coefficient de 0.5 signifie que la covariable double le temps de survie, tandis qu'un coefficient de 2 le divise par deux.

3.3 Transformation des variables

La transformation des covariables dans le contexte de l'analyse statistique joue un rôle crucial pour plusieurs raisons. En modifiant la forme des variables indépendantes, on peut mieux adapter les modèles statistiques aux caractéristiques sous-jacentes des données. Les transformations, telles que l'utilisation de splines ou de polynômes, sont particulièrement utiles pour capturer des relations non linéaires entre les variables, offrant ainsi une flexibilité accrue par rapport aux modèles linéaires traditionnels. Ces techniques permettent de modéliser des comportements plus complexes, d'améliorer la précision des prédictions et de mieux rendre compte de la diversité des relations présentes dans les données. De plus, elles peuvent faciliter l'interprétation des résultats en rendant les relations fonctionnelles entre les variables plus compréhensibles. En somme, la transformation des covariables constitue une étape importante dans la création de modèles statistiques robustes et adaptés aux spécificités des données étudiées.

3.3.1 Transformation splines

a forme générale de la spline cubique naturelle peut être écrite comme une combinaison linéaire de polynômes cubiques et d'une fonction de base de spline. La fonction spline est définie entre des points de référence appelés nœuds (ϵ_1, ϵ_2,...,ϵ_k).

$$S(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3 + \sum_{j=1}^{k} \gamma_j \dot{(X - \epsilon_j)}_+^3$$

où :

- $S(X)$ est la fonction spline résultante,
- $\beta_0\beta_1\beta_2\beta_3$ sont les coefficients associés aux termes polynomiaux,
- γ_j sont les coefficients associés aux termes de la fonction de spline,
- $(X - \epsilon_j)_+^3$ est une fonction qui vaut $X - \epsilon_j$ si $X > \epsilon_j$ et zéro sinon.

Les coefficients $\beta_0\beta_1\beta_2\beta_3$ sont estimés à partir des données, tout comme les coefficients γ_j Le choix des nœuds ϵ_j peut avoir un impact significatif sur la flexibilité de la spline. Des nœuds bien placés permettent à la spline de s'ajuster aux caractéristiques importantes des données, tandis qu'un nombre excessif de nœuds peut conduire à un surajustement.

3.3.2 Transformation polynomiale

La transformation polynomiale est une technique fondamentale en statistique qui permet d'adapter les modèles aux relations non linéaires entre les variables. Cette approche consiste à élever une variable X à différentes puissances, créant ainsi de nouveaux termes polynomiaux. Mathématiquement, la transformation polynomiale d'ordre dd pour une variable X est définie comme suit :

$$X_{poly} = X, X^2, X^3 ..., X^d$$

Chaque terme X^i représente XX élevé à la puissance i. En incorporant ces termes polynomiaux dans un modèle statistique, on peut mieux capturer la complexité des relations entre les variables. Par exemple, un terme quadratique (X^2) permet de modéliser une relation quadratique, tandis qu'un terme cubique (X^3) peut rendre compte de non-linéarités plus complexes. Cette flexibilité permet d'ajuster le modèle de manière plus précise aux caractéristiques des données, offrant ainsi une meilleure compréhension des relations fonctionnelles. Cependant, il est essentiel de trouver le bon équilibre pour éviter le surajustement du modèle.

4 Implementation et resultats

4.1 Application

4.1.1 Applicaton1 : Modèle de Cox

Dans cette partie nous appliquons le modèle a risque proportionel de Cox sur un sous ensemble de nos donnés.

```
from lifelines import CoxPHFitter
data1 = train_data.iloc[:,:14]

modele1 = CoxPHFitter()
modele1.fit(data1, duration_col = 'time', event_col = 'event')
modele1.print_summary()
```

FIGURE 4 – Application du modèle sur les 14 premiers covariables

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Num_Covariate_1	0.81	2.24	0.03	0.75	0.86	2.12	2.36	0.00	29.86	<0.005	648.39
Num_Covariate_2	-0.42	0.66	0.03	-0.47	-0.37	0.62	0.69	0.00	-16.18	<0.005	193.17
Num_Covariate_3	0.25	1.28	0.03	0.20	0.30	1.22	1.35	0.00	9.48	<0.005	68.38
Num_Covariate_4	-0.53	0.59	0.02	-0.58	-0.49	0.56	0.61	0.00	-24.61	<0.005	441.90
Num_Covariate_5	0.30	1.34	0.02	0.26	0.34	1.29	1.40	0.00	14.57	<0.005	157.33
Num_Covariate_6	0.68	1.97	0.02	0.63	0.73	1.88	2.07	0.00	28.02	<0.005	571.67
Num_Covariate_9	-0.15	0.86	0.08	-0.31	0.01	0.73	1.01	0.00	-1.89	0.06	4.10
Num_Covariate_10	-0.12	0.88	0.09	-0.30	0.05	0.74	1.05	0.00	-1.39	0.17	2.60
Num_Covariate_11	-0.63	0.54	0.08	-0.78	-0.47	0.46	0.62	0.00	-7.89	<0.005	48.24
Num_Covariate_12	0.40	1.49	0.09	0.22	0.58	1.25	1.78	0.00	4.45	<0.005	16.81
Num_Covariate_7_1	0.31	1.36	0.06	0.18	0.43	1.20	1.53	0.00	4.87	<0.005	19.80
Num_Covariate_7_2	0.60	1.83	0.07	0.47	0.74	1.59	2.10	0.00	8.59	<0.005	56.67

Concordance	0.79

FIGURE 5 – Résultats

Concordance (C-index) : Le coefficient de concordance, égal à 0.79, mesure la capacité du modèle à classer correctement les paires d'individus en termes de temps de survie. Une valeur plus proche de 1 indique une meilleure performance.
Coefficients des covariables : Les coefficients estimés pour chaque covariable fournissent des informations sur l'impact relatif de ces covariables sur le risque de l'événement. Par exemple, une unité d'augmentation dans la Num_Covariate_3 entraîne une augmentation de 1.28 du risque d'événement, tandis qu'une unité d'augmentation dans la Num_Covariate_11 réduit le risque de 0.08.
Intervalle de confiance : Les intervalles de confiance associés à chaque coefficient fournissent une plage plausible dans laquelle le véritable effet de la covariable pourrait se situer.
Tests de signification : Les tests de signification (z et p) évaluent si les coefficients sont significativement différents de zéro. Les p-values indiquent si une covariable a un impact significatif sur le risque d'événement.
AIC : Le critère d'information d'Akaike (AIC) est utilisé pour comparer la qualité du modèle par rapport à d'autres modèles possibles. Un AIC plus bas est préférable.

4.1.2 Applicaton2 : Accélération du Temps de Défaillance(AFT)

Dans cette partie nous appliquons le modèle d'accélération du temps de défaillance sur un sous ensemble de nos donnés.

		coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
lambda_	**Num_Covariate_1**	-0.562	0.570	0.016	-0.592	-0.531	0.553	0.588	0.000	-36.040	<0.0005	942.427
	Num_Covariate_10	0.038	1.039	0.058	-0.077	0.153	0.926	1.165	0.000	0.650	0.516	0.956
	Num_Covariate_11	0.353	1.423	0.053	0.249	0.457	1.283	1.579	0.000	6.678	<0.0005	35.260
	Num_Covariate_12	-0.253	0.777	0.056	-0.362	-0.144	0.696	0.866	0.000	-4.533	<0.0005	17.391
	Num_Covariate_2	0.279	1.322	0.016	0.248	0.311	1.282	1.365	0.000	17.477	<0.0005	224.789
	Num_Covariate_3	-0.160	0.852	0.017	-0.193	-0.126	0.824	0.881	0.000	-9.372	<0.0005	66.931
	Num_Covariate_4	0.387	1.473	0.013	0.361	0.413	1.435	1.511	0.000	29.346	<0.0005	626.401
	Num_Covariate_5	-0.187	0.830	0.012	-0.211	-0.163	0.810	0.850	0.000	-15.123	<0.0005	169.231
	Num_Covariate_6	-0.486	0.615	0.014	-0.514	-0.458	0.598	0.633	0.000	-34.253	<0.0005	851.756
	Num_Covariate_7_1	-0.143	0.867	0.039	-0.220	-0.066	0.803	0.936	0.000	-3.630	<0.0005	11.786
	Num_Covariate_7_2	-0.352	0.703	0.044	-0.439	-0.265	0.644	0.767	0.000	-7.941	<0.0005	48.826
	Num_Covariate_9	0.113	1.120	0.049	0.016	0.210	1.017	1.234	0.000	2.293	0.022	5.515
	Intercept	0.949	2.582	0.051	0.850	1.048	2.339	2.851	0.000	18.770	<0.0005	258.691
rho_	**Intercept**	0.373	1.451	0.015	0.343	0.402	1.409	1.495	0.000	24.795	<0.0005	448.442

FIGURE 6 – Resultats

Concordance	0.787
AIC	7778.552
log-likelihood ratio test	2922.445 on 12 df
-log2(p) of ll-ratio test	inf

Concordance (C-index) : Le coefficient de concordance, égal à 0.787, mesure la capacité du modèle à classer correctement les paires d'individus en termes de temps de survie. Une valeur plus proche de 1 indique une meilleure performance.

4.2 Concordance Statistique

Dans le domaine des statistiques de survie, la concordance statistique est une mesure couramment utilisée pour évaluer la performance d'un modèle de survie, tel que le modèle de régression de Cox. La concordance dans ce contexte mesure la capacité du modèle à correctement ordonner les paires d'événements de survie. En d'autres termes, elle évalue la probabilité que le modèle classe une paire d'individus dans le bon ordre de survie. La concordance c-statistique (ou concordance de Harrell) est une mesure spécifique de la concordance dans le contexte des modèles de survie. Elle varie de 0 à 1, où 0.5 indique une performance équivalente à un classement aléatoire, et 1 indique une classification parfaite. Plus la concordance est élevée, meilleure est la capacité du modèle à prédire la survie. La formule de calculer la concordance s'écrit comme suit :

$$C = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\mu_j>\mu_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}$$

où :

- μ_j est l'indice du risque,
- $1_{T_j<T_i} = 1$ si $T_j < T_i$, 0 sinon,
- $1_{T_j<T_i} = 1$ si $T_j < T_i$, 0 sinon.

```
def compute_cindex(predictions, durations, events):
    n = len(predictions)
    concordant = 0
    permissible = 0

    for i in range(n):
        for j in range(i + 1, n):
            if events[i] == 1 and durations[i] < durations[j]:
                permissible += 1
                if predictions[i] > predictions[j]:
                    concordant += 1
            elif events[i] == 1 and durations[i] > durations[j]:
                permissible += 1
                if predictions[i] < predictions[j]:
                    concordant += 1
            elif events[i] == 0 and durations[i] < durations[j]:
                permissible += 1
            elif events[i] == 0 and durations[i] > durations[j]:
                permissible += 1
                concordant += 1

    if permissible > 0:
        cindex = concordant / permissible
    else:
        cindex = 0.0

    return cindex
```

FIGURE 7 – Fonction C-index

La concordance avec la formule manuellement implementé donne une valeur de 0.71 environ comparé au concordance dans le modèle Cox dont la concordance est égale à : 0.79

4.3 Cross-validation

La validation croisée est une méthode de validation des modèles d'apprentissage automatique. Elle consiste à diviser le jeu de données d'entraînement en plusieurs sous-ensembles, puis à entraîner le modèle sur différents sous-ensembles et à le tester sur les autres. Cela permet d'obtenir une estimation plus précise des performances du modèle sur des données inconnues.

Afin d'approfondir l'analyse de survie des ensembles de données, les donnee. Pour chaque sous-ensemble de données ainsi constitué, les deux modèles, regression Cox et AFT sont appliquées.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Num_Covariate_1	0.74	2.10	0.02	0.69	0.79	2.00	2.20	0.00	30.94	<0.005	695.59
Num_Covariate_2	-0.37	0.69	0.02	-0.41	-0.32	0.66	0.72	0.00	-16.37	<0.005	197.70
Num_Covariate_3	0.27	1.31	0.02	0.23	0.31	1.26	1.37	0.00	12.27	<0.005	112.56
Num_Covariate_4	-0.51	0.60	0.02	-0.55	-0.47	0.58	0.62	0.00	-25.60	<0.005	477.75
Num_Covariate_5	0.31	1.36	0.02	0.27	0.34	1.31	1.41	0.00	17.20	<0.005	217.87
Num_Covariate_6	0.65	1.91	0.02	0.60	0.69	1.83	1.99	0.00	29.88	<0.005	649.07
Num_Covariate_9	0.06	1.06	0.07	-0.07	0.19	0.93	1.21	0.00	0.91	0.36	1.47
Num_Covariate_10	-0.06	0.94	0.08	-0.22	0.09	0.80	1.10	0.00	-0.80	0.42	1.24

Concordance	0.78
Partial AIC	36131.14
log-likelihood ratio test	2665.59 on 8 df
-log2(p) of ll-ratio test	inf

FIGURE 8 – Cox

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Num_Covariate_5	0.26	1.30	0.02	0.23	0.30	1.26	1.35	0.00	15.19	<0.005	170.79
Num_Covariate_6	0.39	1.48	0.02	0.35	0.43	1.42	1.53	0.00	19.35	<0.005	274.79
Num_Covariate_9	0.08	1.09	0.07	-0.05	0.22	0.95	1.24	0.00	1.21	0.23	2.14
Num_Covariate_10	0.52	1.68	0.08	0.36	0.68	1.44	1.97	0.00	6.46	<0.005	33.11
Num_Covariate_11	-0.11	0.90	0.07	-0.25	0.03	0.78	1.03	0.00	-1.51	0.13	2.94
Num_Covariate_12	0.51	1.67	0.07	0.37	0.66	1.45	1.93	0.00	7.12	<0.005	39.80
Num_Covariate_7_2	-0.01	0.99	0.04	-0.10	0.07	0.90	1.07	0.00	-0.32	0.75	0.42
Num_Covariate_8_1	0.09	1.09	0.04	0.01	0.17	1.01	1.19	0.00	2.13	0.03	4.91

Concordance	0.72
Partial AIC	37433.83
log-likelihood ratio test	1362.90 on 8 df
-log2(p) of ll-ratio test	957.46

FIGURE 9 – Cox

		coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
alpha_	Num_Covariate_1	-0.502	0.606	0.017	-0.535	-0.468	0.586	0.626	0.000	-29.631	<0.0005	638.569
	Num_Covariate_10	0.031	1.031	0.064	-0.095	0.156	0.910	1.169	0.000	0.478	0.633	0.660
	Num_Covariate_2	0.246	1.279	0.017	0.212	0.280	1.236	1.323	0.000	14.241	<0.0005	150.464
	Num_Covariate_3	-0.191	0.826	0.018	-0.226	-0.157	0.798	0.855	0.000	-10.890	<0.0005	89.334
	Num_Covariate_4	0.364	1.440	0.015	0.335	0.394	1.398	1.482	0.000	24.435	<0.0005	435.645
	Num_Covariate_5	-0.218	0.804	0.014	-0.245	-0.191	0.783	0.826	0.000	-15.839	<0.0005	185.290
	Num_Covariate_6	-0.447	0.639	0.016	-0.478	-0.417	0.620	0.659	0.000	-28.699	<0.0005	599.300
	Num_Covariate_9	-0.021	0.979	0.053	-0.125	0.083	0.882	1.087	0.000	-0.395	0.693	0.529
	Intercept	0.571	1.771	0.040	0.492	0.651	1.636	1.917	0.000	14.139	<0.0005	148.356
beta_	Intercept	0.690	1.993	0.016	0.658	0.721	1.931	2.057	0.000	42.817	<0.0005	inf

Concordance	0.783
AIC	8085.883
log-likelihood ratio test	2473.372 on 8 df
-log2(p) of ll-ratio test	inf

FIGURE 10 – AFT

		coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
alpha_	Num_Covariate_10	-0.443	0.642	0.077	-0.595	-0.292	0.552	0.747	0.000	-5.742	<0.0005	26.673
	Num_Covariate_11	0.136	1.146	0.068	0.002	0.270	1.002	1.310	0.000	1.985	0.047	4.407
	Num_Covariate_12	-0.505	0.603	0.070	-0.643	-0.368	0.526	0.692	0.000	-7.225	<0.0005	40.857
	Num_Covariate_5	-0.236	0.790	0.016	-0.267	-0.204	0.765	0.816	0.000	-14.523	<0.0005	156.344
	Num_Covariate_6	-0.358	0.699	0.019	-0.395	-0.322	0.674	0.725	0.000	-19.214	<0.0005	270.906
	Num_Covariate_7_2	0.004	1.004	0.043	-0.080	0.088	0.923	1.092	0.000	0.083	0.934	0.099
	Num_Covariate_8_1	-0.030	0.970	0.041	-0.110	0.049	0.896	1.051	0.000	-0.742	0.458	1.126
	Num_Covariate_9	-0.015	0.985	0.066	-0.145	0.114	0.865	1.121	0.000	-0.231	0.817	0.291
	Intercept	0.976	2.654	0.056	0.867	1.085	2.380	2.959	0.000	17.574	<0.0005	227.258
beta_	Intercept	0.497	1.644	0.016	0.465	0.529	1.593	1.697	0.000	30.805	<0.0005	689.799

Concordance	0.717
AIC	9195.420
log-likelihood ratio test	1363.835 on 8 df
-log2(p) of ll-ratio test	958.137

FIGURE 11 – AFT

Lors de l'évaluation du modèle de survie sur l'ensemble de validation, une constatation notable émerge quant à la proximité entre la concordance prédite et les performances du modèle. Les résultats obtenus révèlent une similitude remarquable entre les prédictions de concordance et les mesures réelles de survie. Cette cohérence souligne la robustesse du modèle, démontrant sa capacité à généraliser efficacement à de nouvelles données. La concordance étant une mesure cruciale dans l'évaluation des modèles de survie, cette concordance similaire entre les prédictions et les résultats observés sur l'ensemble de validation suggère une validité et une fiabilité étendues du modèle. Cette concordance étroite indique que les relations entre les variables du modèle et le temps de survie, telles qu'apprises à partir de l'ensemble d'apprentissage, sont bien généralisables et restent cohérentes lors de l'application du modèle à des données indépendantes.

En résumé, les résultats de l'analyse comparative entre les modèles AFT (Accelerated Failure Time) et CoxPH (Cox Proportional Hazards) sur l'ensemble de données spécifique suggèrent les conclusions suivantes :

Performances Similaires : Les performances des deux modèles, évaluées par le biais du C-index, sont très proches. Les valeurs moyennes du C-index pour l'AFT et le CoxPH dans le cadre de la validation croisée indiquent une similitude dans leur capacité à prédire les événements.

Bonne Capacité de Prévision Les valeurs du C-index autour de 0.79 pour les deux modèles signalent une bonne capacité à prévoir l'ordre relatif des événements. Cette mesure suggère que les modèles sont efficaces dans la discrimination des risques.

Consistance et Stabilité : La cohérence des résultats sur les différents plis de la validation croisée témoigne de la stabilité des modèles AFT et CoxPH sur divers sous-ensembles de données. Cette constance renforce la fiabilité des performances observées.

Aucune Supériorité Évidente d'un Modèle : Les résultats ne révèlent pas de nette supériorité d'un modèle par rapport à l'autre. La différence minime entre les moyennes du C-index suggère une performance comparable des deux modèles.

4.3.1 Moyenne des indices de concordance des modéles AFT et Cox sur différents jeux de données

	Model	Average_Concordance_Index
0	AFT	0.788171
1	CoxPH	0.788170

FIGURE 12 – Cox

4.4 Conclusion

ce projet a permis d'explorer et de comparer deux modèles de survie majeurs, à savoir l'Accéléré Failure Time (AFT) et le Cox Proportional Hazards (CoxPH), sur un ensemble de données spécifique. Les résultats obtenus indiquent des performances similaires entre les deux modèles, avec des valeurs moyennes de l'indice de concordance (C-index) proches et une capacité robuste à prévoir l'ordre relatif des événements.