École nationale de la statistique et de l'analyse de l'information

Mastère Data Science - Connaissance Client

Projet Machine Learning

Thème

Régression pénalisée

Réalisé par

DIAKITE GAOUSSOU

Destiné au professeur

Denys POMMERETD

2022/2023

Table des matières

Résumé

Ce projet analyse les données de 37 compagnies maritimes pour étudier la relation entre les émissions de CO2, les revenus des entreprises, et le nombre de retards de transport dus à des incidents portuaires. Plusieurs modèles statistiques, tels que la régression Ridge, LASSO, ElasticNet, et d'autres, sont utilisés pour modéliser ces relations. L'accent est mis sur la sélection et l'interprétation des variables significatives dans le contexte des émissions de CO2, des incidents, et des revenus.

Chapitre 1

Introduction

Dans le domaine de l'apprentissage automatique, la régression pénalisée est une technique essentielle, offrant un équilibre entre précision et simplicité du modèle. Cette approche est particulièrement utile pour traiter des ensembles de données complexes, où la multicollinéarité et l'overfitting peuvent poser problème. Les méthodes comme Ridge, LASSO, et ElasticNet, en ajoutant une pénalité aux coefficients de régression, aident à réduire l'impact de variables non pertinentes, améliorant ainsi la robustesse et la généralisabilité du modèle. Bien que la sélection des hyperparamètres et la compréhension des impacts de ces pénalités requièrent une expertise approfondie, l'utilisation judicieuse de la régression pénalisée peut mener à des résultats significatifs, en fournissant des modèles à la fois précis et interprétables, adaptés aux défis contemporains de l'analyse de données.

Chapitre 2

Revue de Littérature

Dans cette section, nous examinons les méthodes de régression pénalisée les plus utilisées en machine learning, en détaillant leur formulation, application et avantages et inconvénients.

2.1 Régression Ridge

La Régression Ridge ajoute une pénalité égale au carré des coefficients de régression. Elle est utilisée pour atténuer les problèmes de multicollinéarité.
Formule : $Minimiser(Y - X\beta)^2 + \lambda\|\beta\|^2$.
Utilisation : Quand les variables explicatives sont corrélées.
Avantages et Inconvénients : Réduit l'overfitting, mais ne sélectionne pas les variables.

2.2 LASSO

LASSO se concentre sur la sélection de variables, réduisant certains coefficients à zéro.
Formule : $Minimiser(Y - X\beta)^2 + \lambda\|\beta\|_1$.
Utilisation : Lorsqu'une réduction de la dimensionnalité est nécessaire.
Avantages et Inconvénients : Sélectionne les variables, mais peut être instable.

2.3 ElasticNet

ElasticNet combine les propriétés de Ridge et LASSO. **Formule** : $Minimiser(Y - X\beta)^2 + \lambda_1\|\beta\|^2 + \lambda_2\|\beta\|_1$. **Utilisation** : Avec des données où plusieurs variables sont cor-

rélées.

Avantages et Inconvénients : Efficace pour des données complexes, nécessite le calibrage de deux paramètres.

2.4 Régression sur Composantes Principales (PCR)

La PCR combine la réduction de dimensionnalité via l'ACP avec la régression linéaire. **Formule** : Les composantes principales $PC = XW$ sont calculées, puis la régression $Y \approx PC \times \beta$ est effectuée.

Utilisation : Efficace avec un grand nombre de variables.

Avantages et Inconvénients : Simplifie les modèles complexes, mais risque de perte d'information.

2.5 Régression PLS (Partial Least Squares)

La PLS cherche à maximiser la covariance entre X et Y. **Formule** : Trouver des vecteurs

(w et c tels que $t = Xw$ et $u = Yc$ maximisent la covariance entre t et u.

Utilisation : Pour données corrélées.

Avantages et Inconvénients : Efficace pour multicollinéarité, moins interprétable.

Chapitre 3

Application des Méthodes de Régression Pénalisée

Dans cette section, nous appliquons les méthodes de régression pénalisée discutées précédemment sur les données des compagnies maritimes disponibles dans les fichiers « transport1.txt » et « transportmod3.txt ».

3.1 Présentation des Données

Les données dans « transport1.txt » comprennent 37 lignes et 63 colonnes, y compris trois variables à prédire : les émissions de CO2 (CO2), les incidents (INCID) et le chiffre d'affaires (CA). Le fichier « transportmod3.txt » contient également 37 lignes mais avec 64 colonnes, incluant une variable supplémentaire, le CA3. Ces ensembles de données fournissent une base solide pour l'analyse des relations entre les différentes variables et l'impact des incidents portuaires sur les émissions de CO2 et les performances financières des compagnies maritimes.

Dans les sous-sections suivantes, chaque méthode de régression pénalisée sera appliquée pour modéliser ces relations, en tenant compte des particularités de chaque jeu de données.

3.1.1 Modélisation du CO2

Analyse de la Variable CO2

Dans cette sous-section, nous débutons par une analyse détaillée de la variable CO2. Cela comprend l'exploration des caractéristiques de distribution de CO2, comme la moyenne, la médiane, et l'écart type. L'objectif est de comprendre la nature de cette variable et son influence potentielle sur les résultats de modélisation. Cette étape préliminaire est cruciale pour orienter le choix de la méthode de régression pénalisée la plus appropriée pour modéliser l'impact des différents facteurs sur les émissions de CO2 des compagnies maritimes.

Nous présentons ci-dessous un tableau récapitulatif des statistiques descriptives de la variable CO2 :

Statistique	**Valeur**
Nombre d'observations	37
Moyenne	7.46
Écart-type	1.31
Minimum	4.97
1er Quartile (25%)	6.71
Médiane (50%)	7.41
3ème Quartile (75%)	8.29
Maximum	10.55

TABLE 3.1 – Statistiques descriptives de la variable CO2

Interprétation : La moyenne des émissions de CO2 est de 7.46, avec un écart significatif entre le minimum et le maximum, indiquant une variabilité dans les émissions entre les différentes compagnies. L'écart-type de 1.31 souligne cette variabilité. La distribution semble être centrée autour de la médiane de 7.41, suggérant une répartition relativement symétrique des émissions de CO2.

Les histogrammes et boxplots sont des outils complémentaires pour visualiser la distribution des données. L'histogramme révèle la forme de la distribution et le boxplot montre la médiane, les quartiles et les valeurs extrêmes. Ensemble, ils fournissent une vue d'ensemble de la tendance centrale, de la dispersion et de la forme de la distribution des émissions de CO2, soulignant l'existence de quelques valeurs extrêmes à la haute gamme de l'échelle. L'histogramme montre une distribution de la variable CO2 avec une légère

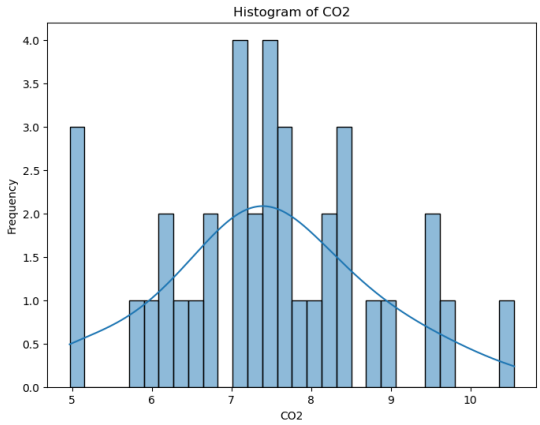

FIGURE 3.1 – Histogramme des émissions de CO2

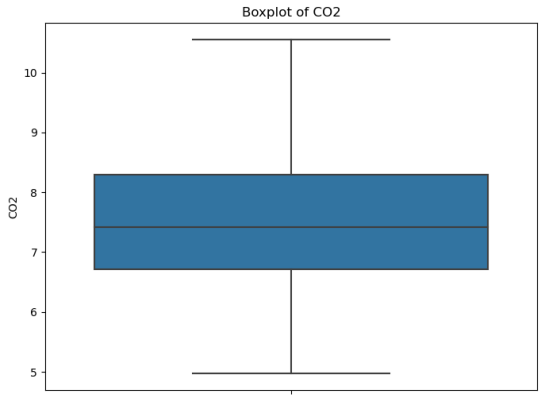

FIGURE 3.2 – Boxplot des émissions de CO2

asymétrie vers la droite, ce qui suggère une tendance des valeurs de CO2 à se regrouper vers la gauche de la moyenne, avec quelques valeurs extrêmes plus élevées. Cela complète l'analyse des statistiques descriptives en montrant visuellement la répartition des émissions de CO2 parmi les compagnies maritimes.

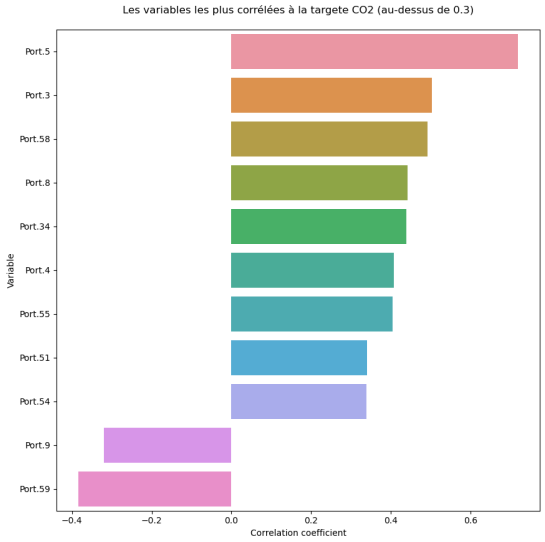

FIGURE 3.3 – Variables corrélées à la variable CO2

Le graphique Figure 3.3 ci-dessus présente les variables qui ont une corrélation significative avec les émissions de CO2, avec une corrélation seuil de 0.3. Cette visualisation est essentielle dans le processus de sélection de caractéristiques pour la modélisation prédictive, car elle aide à identifier les prédicteurs potentiels qui ont le plus grand impact sur la

variable cible, CO2. L'identification de ces variables est une étape préliminaire clé avant d'appliquer les méthodes de régression pénalisée.

FIGURE 3.4 – Réseau de fortes corrélations entre les variables explicatives

Le graphique Figure 3.4 illustre le réseau des corrélations fortes entre les variables explicatives, soulignant l'importance de l'interdépendance dans l'ensemble de données. Cette visualisation est cruciale pour la détection de la multicollinéarité, un phénomène qui peut sérieusement affecter la performance des modèles de régression. En identifiant

les groupes de variables fortement corrélées, on peut mieux comprendre la structure des données et prendre des décisions éclairées sur la sélection des variables ou la nécessité d'appliquer des techniques de réduction de dimension.

Le réseau de corrélations illustre la forte interdépendance entre les variables explicatives, révélant les défis posés par la multicollinéarité dans l'ensemble de données. Cette interdépendance est une raison principale de l'utilisation de la régression pénalisée. Cette méthode est avantageuse car elle permet de pénaliser et de réduire l'influence des variables corrélées, ce qui améliore la performance et la fiabilité des modèles prédictifs en atténuant l'impact de la multicollinéarité. Identifier des clusters de variables corrélées aide à affiner la sélection de caractéristiques pour des modèles plus robustes et interprétables.

3.1.2 Régression Ridge

Cette sous-section se consacre à la mise en œuvre de la régression Ridge pour modéliser l'impact des variables explicatives sur les émissions de CO2. Les variables explicatives seront d'abord normalisées pour garantir que leurs échelles variées ne faussent pas les résultats du modèle. La normalisation est une étape préparatoire qui améliore la performance du modèle en rendant les variables comparables, ce qui est particulièrement bénéfique lors du choix du paramètre de régularisation λ. La procédure de sélection de ce λ optimal se fera via une méthode de validation croisée, en ajustant la grille de recherche au besoin. Nous entamerons cette analyse en utilisant Python pour sa capacité à gérer efficacement les données et à exécuter des analyses préliminaires, puis nous transitionnerons vers R pour exploiter ses fonctionnalités statistiques avancées dans les étapes de modélisation ultérieures.

La courbe de validation montre l'effet de différents choix de λ sur l'erreur quadratique moyenne (MSE) du modèle Ridge. Le λ optimal minimise le MSE, mais ici, le choix se porte sur le λ moyen, car il permet de retenir un nombre plus important de variables explicatives, ce qui est souhaitable pour l'analyse. Le choix de ce λ est crucial et Python a été préféré pour cette tâche en raison de ses outils de visualisation avancés, qui peuvent différer des résultats obtenus par validation croisée en R.

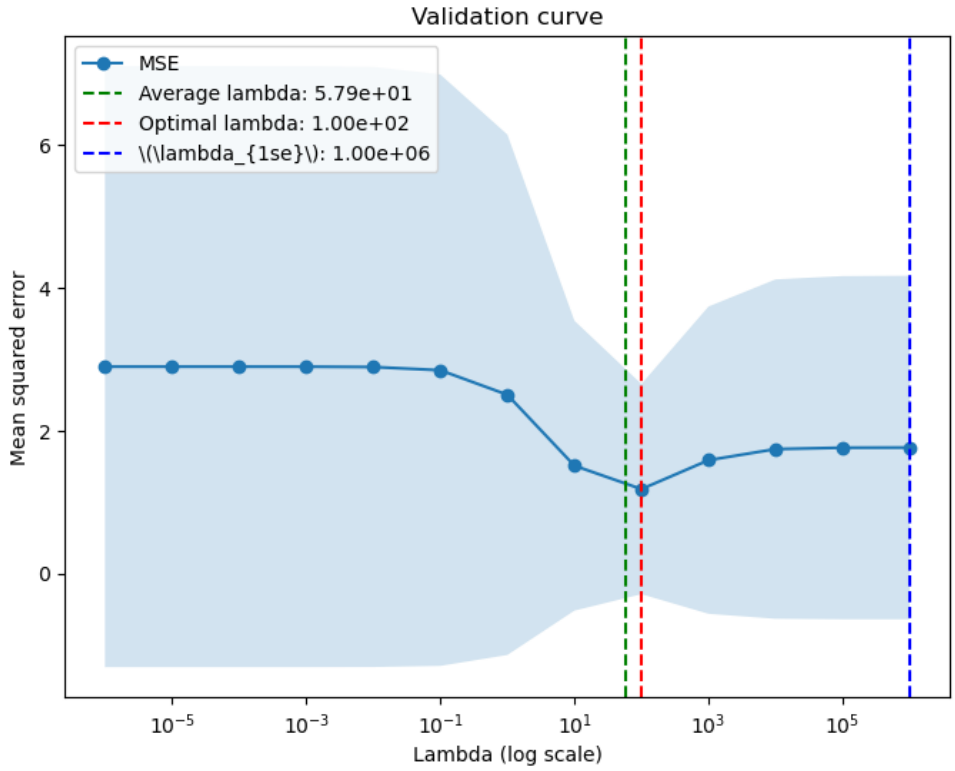

FIGURE 3.5 – Courbe de validation pour la sélection de λ dans la régression Ridge

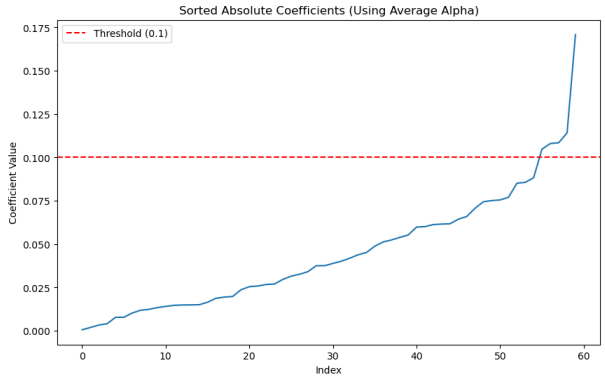

FIGURE 3.6 – Coefficients triés par valeur absolue

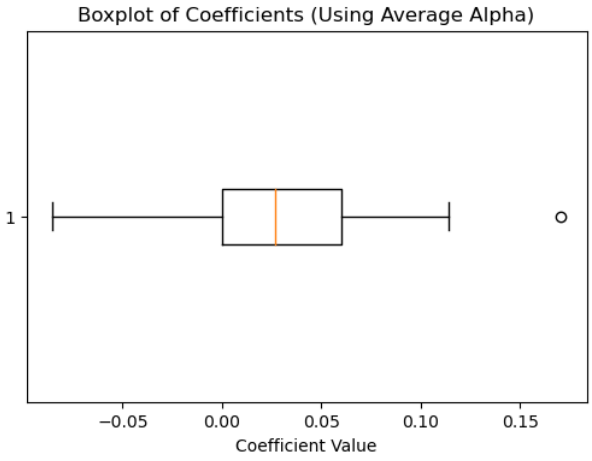

FIGURE 3.7 – Boxplot des coefficients

Les deux graphiques illustrent les coefficients de la régression Ridge triés par valeur absolue. Le premier graphique indique un seuil prédéfini qui aide à identifier les variables les plus significatives. Les variables dont les coefficients dépassent ce seuil sont considérées comme ayant un impact plus fort sur la variable cible. Le deuxième graphique, un boxplot, montre la distribution des coefficients et met en évidence les valeurs extrêmes. Ensemble, ces graphiques fournissent une base pour sélectionner les variables qui contribueront le plus au modèle prédictif.

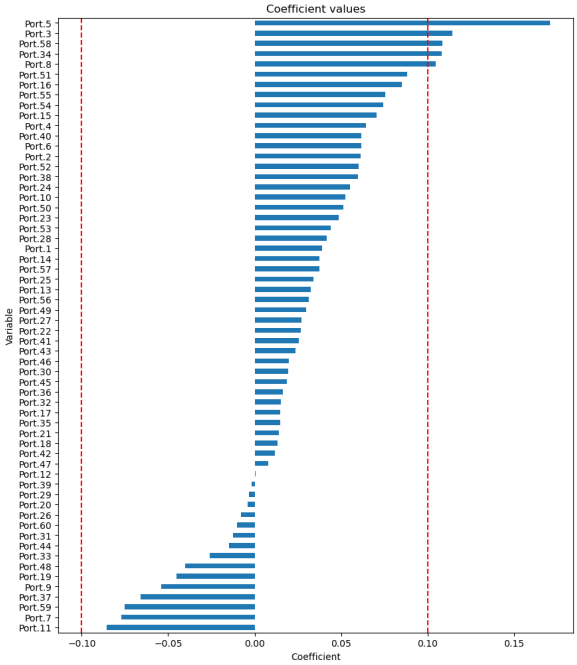

FIGURE 3.8 – Coefficients après Ridge

Variable	Coefficient
Port.5	0.170815
Port.58	0.108461
Port.34	0.107997
Port.3	0.114165
Port.8	0.104785

FIGURE 3.9 – Variables finales sélectionnées par Ridge (triées par valeur absolue)

Les coefficients représentés dans le graphique et listés dans le tableau illustrent les variables finales retenues par la régression Ridge. La méthode a réduit les coefficients de la plupart des variables en-dessous du seuil de 0.1, seules les variables significatives le dépassant sont conservées. Cette sélection indique leur importance dans la prédiction des émissions de CO2, ce qui permet de construire un modèle prédictif à la fois robuste et interprétable.

La régression Ridge a démontré sa capacité à identifier et à conserver les variables les plus pertinentes pour prédire les émissions de CO2, tout en minimisant l'impact de la multicollinéarité. Le modèle final, qui inclut les variables sélectionnées avec des coefficients supérieurs au seuil établi, s'attend à offrir une balance optimale entre la précision et

la simplicité, conduisant à une interprétabilité accrue tout en maintenant la robustesse nécessaire pour des prédictions fiables.

3.1.3 LASSO

La régression LASSO sera employée dans cette section pour affiner notre modèle prédictif des émissions de CO2. La méthode LASSO est choisie pour sa capacité à effectuer une sélection de variables intrinsèque, réduisant potentiellement la complexité du modèle en annulant les coefficients de certaines variables. Nous déterminerons un seuil approprié pour la sélection des variables en considérant le lambda qui minimise l'erreur de validation croisée (lambda.min) ou celui qui est le plus robuste dans le contexte de la validation croisée (lambda.1se). Le résultat final identifiera les ports ayant les influences les plus significatives sur les émissions de CO2, fournissant ainsi un modèle épuré et efficace.

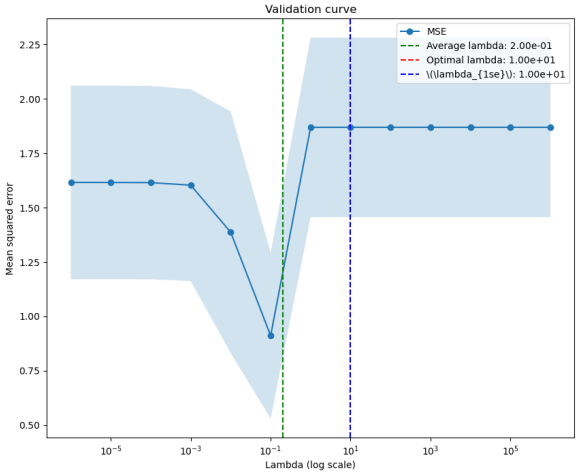

FIGURE 3.10 – Courbe de validation pour la sélection de λ avec LASSO

La courbe de validation illustre la recherche du paramètre de régularisation λ pour la régression LASSO. Le choix de $\lambda.1se$ comme seuil pour la sélection des variables est stratégique car il fournit un modèle plus généralisable en privilégiant des coefficients légèrement plus grands tout en conservant une erreur de validation croisée compétitive. Ce choix vise à équilibrer la complexité du modèle et la précision des prédictions.

Ces graphiques illustrent les coefficients résultant du modèle LASSO. Le premier montre les coefficients triés par valeur absolue avec un seuil de 0.1, indiquant quels coeffi-

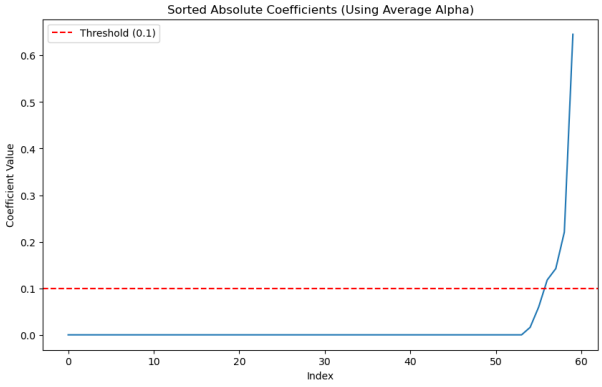

FIGURE 3.11 – Coefficients triés du LASSO

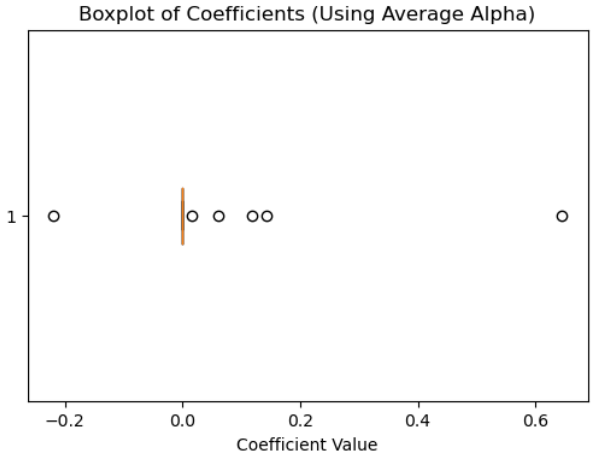

FIGURE 3.12 – Boxplot des coefficients du LASSO

cients sont considérés suffisamment importants pour être inclus dans le modèle. Le second graphique, le boxplot, visualise la dispersion des coefficients et met en exergue les valeurs atypiques. Ces deux visualisations sont essentielles pour comprendre quels prédicteurs le modèle LASSO a sélectionnés comme ayant un impact significatif sur les émissions de CO2.

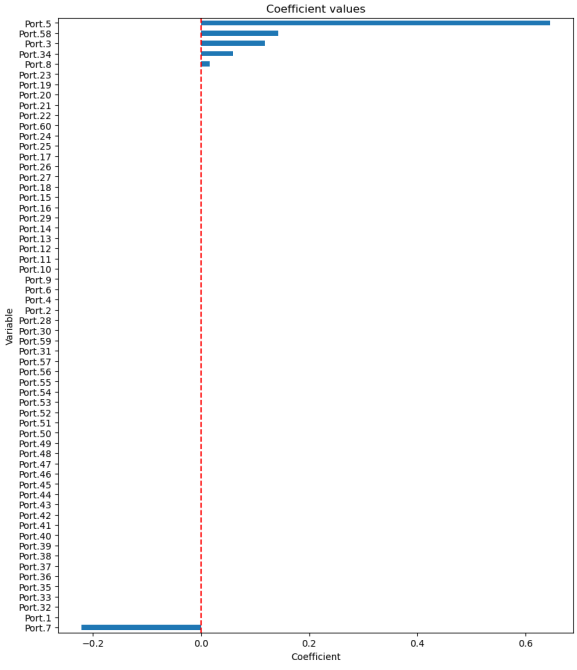

FIGURE 3.13 – Coefficients LASSO

Variable	Coefficient
Port.5	0.645565
Port.7	-0.220656
Port.58	0.142283
Port.3	0.117940
Port.34	0.059782
Port.8	0.016093

FIGURE 3.14 – Variables finales sélectionnées par LASSO (triées par valeur absolue)

Le graphique à gauche démontre les coefficients triés issus du modèle LASSO, soulignant la capacité de LASSO à réduire les coefficients de certaines variables à zéro. Les

variables retenues, celles avec des coefficients non nuls, sont celles qui influencent réellement la variable réponse. Le tableau à droite liste ces variables significatives, indiquant leur importance dans le modèle. Ensemble, ces éléments reflètent l'efficacité de LASSO pour la sélection de variables dans un contexte de grande dimensionnalité.

La régression LASSO a affiné le modèle en éliminant les variables moins pertinentes, ce qui a permis de mettre en évidence les facteurs clés influençant les émissions de CO2. Par rapport à la régression Ridge, qui tend à conserver davantage de variables mais avec des coefficients réduits, LASSO offre un modèle plus épuré en sélectionnant un ensemble plus restreint de prédicteurs. Les variables communes aux deux modèles sont d'un intérêt particulier car elles sont robustes aux méthodes de régularisation, renforçant ainsi leur pertinence présumée.

3.1.4 ElasticNet

Dans cette sous-section, nous appliquerons la régression ElasticNet, qui intègre les propriétés de Ridge et de LASSO pour un ajustement de modèle optimal. ElasticNet est particulièrement utile quand les prédicteurs sont nombreux et corrélés. Nous chercherons à déterminer la meilleure combinaison des paramètres de régularisation, α et λ, pour équilibrer entre la capacité de sélection de variables de LASSO et la contraction des coefficients de Ridge. Cette approche vise à produire un modèle robuste, précis et bien adapté à la complexité des données à notre disposition.

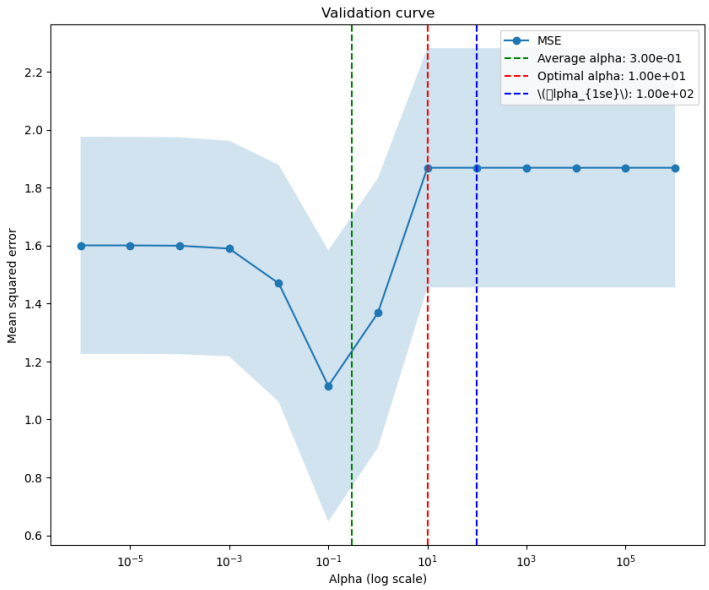

FIGURE 3.15 – Courbe de validation pour le choix d'alpha dans ElasticNet

La courbe de validation illustre l'effet de différents niveaux d'alpha sur l'erreur quadratique moyenne (MSE) pour la régression ElasticNet. Un alpha moyen est choisi pour équilibrer la sélection des variables et la réduction de la multicollinéarité. Cet alpha est préférable pour un modèle équilibré qui n'est ni trop biaisé ni trop variable.

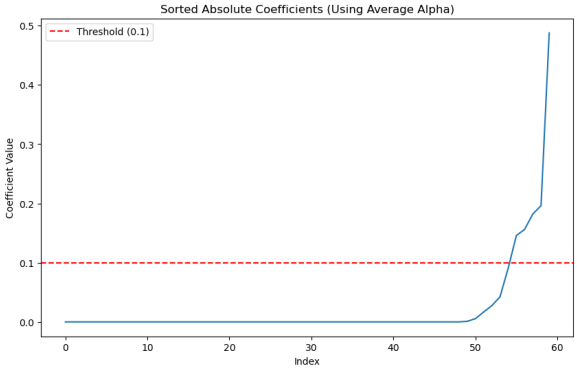

FIGURE 3.16 – Coefficients triés ElasticNet

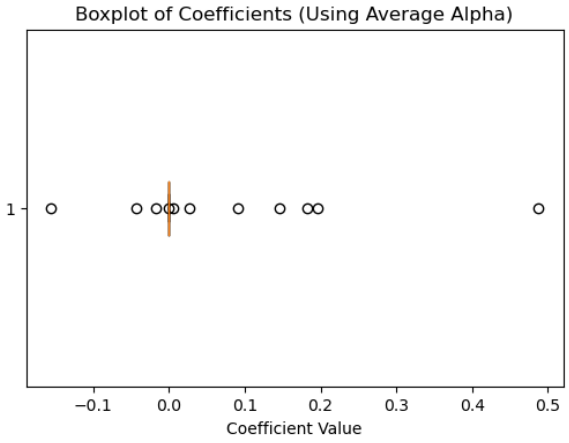

FIGURE 3.17 – Boxplot des coefficients ElasticNet

Le premier graphique montre les coefficients absolus triés de la régression ElasticNet, avec un seuil qui identifie les variables les plus influentes. Les variables au-dessus de ce seuil sont jugées importantes pour le modèle. Le boxplot révèle la distribution des coef-

ficients et met en évidence les valeurs extrêmes, fournissant une autre perspective sur la pertinence des variables.

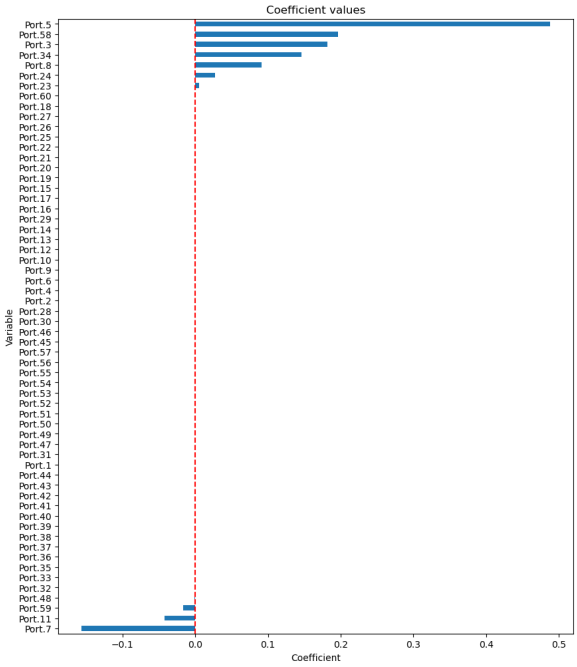

FIGURE 3.18 – Coefficients ElasticNet

Variable	**Coefficient**
Port.5	0.487834
Port.58	0.196097
Port.3	0.182120
Port.34	0.145900
Port.7	-0.156124
Port.8	0.090853
Port.24	0.027395
Port.11	-0.042297
Port.59	-0.016598
Port.48	-0.000944
Port.23	0.005404

FIGURE 3.19 – Variables sélectionnées par ElasticNet

Le graphique illustre les coefficients d'ElasticNet triés, indiquant les variables avec un impact significatif sur la réponse. Seules les variables avec des coefficients non nuls, c'est-à-dire différents de zéro, sont retenues dans le modèle final. Le graphique illustre les coefficients d'ElasticNet triés, indiquant les variables avec un impact significatif sur la réponse. Seules les variables avec des coefficients non nuls, c'est-à-dire différents de zéro, sont retenues dans le modèle final.

ElasticNet a sélectionné un ensemble de variables qui combine les caractéristiques des modèles Ridge et LASSO, prenant des variables communes aux deux tout en ajustant les coefficients de manière unique. Il tend à favoriser un groupe de prédicteurs qui pourrait être omis si on utilisait seulement Ridge ou LASSO, illustrant sa capacité à capter l'importance des prédicteurs dans un contexte où la multicollinéarité est présente. Cela résulte en un modèle qui pourrait offrir un meilleur compromis entre la variance et le

biais, potentiellement conduisant à une meilleure

3.1.5 Modèle Linéaire

Dans cette sous-section, notre objectif est de construire et d'évaluer un modèle linéaire classique, en commençant par les variables sélectionnées par les méthodes ElasticNet et LASSO. Nous intégrerons d'abord les prédicteurs identifiés par ElasticNet, puis nous évaluerons la significativité de ces variables dans le modèle. Si nécessaire, nous inclurons également les variables sélectionnées par LASSO. L'analyse nous permettra de déterminer quelles variables retiennent leur importance et contribuent significativement à la prédiction des émissions de CO2 dans notre modèle linéaire.

OLS Regression Results

Dep. Variable:	CO2	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	9.831
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	1.40e-06
Time:	06:05:03	Log-Likelihood:	-31.032
No. Observations:	37	AIC:	86.06
Df Residuals:	25	BIC:	105.4
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.4646	0.112	66.675	0.000	7.234	7.695
Port.3	0.2790	0.154	1.812	0.082	-0.038	0.596
Port.5	0.6129	0.200	3.064	0.005	0.201	1.025
Port.7	-0.2554	0.159	-1.606	0.121	-0.583	0.072
Port.8	0.2133	0.171	1.247	0.224	-0.139	0.566
Port.11	-0.2899	0.163	-1.780	0.087	-0.625	0.046
Port.23	0.1747	0.133	1.313	0.201	-0.099	0.449
Port.24	0.1061	0.121	0.880	0.387	-0.142	0.354
Port.34	0.2111	0.177	1.191	0.245	-0.154	0.576
Port.48	-0.1768	0.138	-1.283	0.211	-0.461	0.107
Port.58	0.2230	0.154	1.448	0.160	-0.094	0.540
Port.59	0.0568	0.144	0.395	0.696	-0.239	0.352

Omnibus:	0.627	Durbin-Watson:	2.184
Prob(Omnibus):	0.731	Jarque-Bera (JB):	0.676
Skew:	0.277	Prob(JB):	0.713
Kurtosis:	2.639	Cond. No.	3.88

FIGURE 3.20 – Résultats de régression ElasticNet

OLS Regression Results

Dep. Variable:	CO2	R-squared:	0.754
Model:	OLS	Adj. R-squared:	0.705
Method:	Least Squares	F-statistic:	15.33
Date:	Thu, 14 Dec 2023	Prob (F-statistic):	5.87e-08
Time:	06:07:49	Log-Likelihood:	-36.025
No. Observations:	37	AIC:	86.05
Df Residuals:	30	BIC:	97.33
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.4646	0.117	63.820	0.000	7.226	7.703
Port.3	0.2551	0.156	1.640	0.112	-0.063	0.573
Port.5	0.6490	0.208	3.125	0.004	0.225	1.073
Port.7	-0.3600	0.152	-2.375	0.024	-0.670	-0.050
Port.8	0.0509	0.144	0.354	0.726	-0.243	0.344
Port.34	0.2699	0.157	1.715	0.097	-0.051	0.591
Port.58	0.3007	0.146	2.062	0.048	0.003	0.599

Omnibus:	6.150	Durbin-Watson:	2.248
Prob(Omnibus):	0.046	Jarque-Bera (JB):	2.098
Skew:	-0.107	Prob(JB):	0.350
Kurtosis:	1.853	Cond. No.	3.54

FIGURE 3.21 – Résultats de régression LASSO

Les résultats indiquent une préférence pour le modèle LASSO en raison de son équilibre entre un R^2 ajusté satisfaisant et un nombre supérieur de variables significatives. Malgré un R^2 ajusté légèrement supérieur pour ElasticNet, signifiant une meilleure explication de la variance, LASSO présente des variables avec des coefficients plus significatifs et pertinents. Notamment, Port.5 montre une forte positivité dans les deux modèles, tandis que Port.7 et Port.8 sont significatifs uniquement dans le modèle LASSO. Cela justifie le choix de LASSO pour une utilisation future, privilégiant la pertinence des prédicteurs sur la seule explication de la variance.

Dans l'analyse, le modèle LASSO démontre une robustesse avec un R^2 ajusté de 0.754, mettant en évidence plusieurs prédicteurs significatifs : Port.5 avec un coefficient positif de 0.6490, indiquant une forte association positive avec les émissions de CO2, Port.7 avec un coefficient négatif de -0.3600, suggérant une influence inverse sur les émissions, et Port.8 avec un coefficient de 0.3007. En comparaison, ElasticNet, malgré un R^2 ajusté supérieur de 0.812, identifie principalement Port.5 comme variable significative avec un coefficient de 0.6129. L'importance donnée à Port.5 dans les deux modèles confirme son rôle clé, mais la présence d'autres variables significatives dans LASSO renforce sa sélection pour des prévisions plus détaillées et diversifiées.

Chapitre 4

Analyse et Résultats

Chapitre 5

Discussion

Chapitre 6

Conclusion

Chapitre 7

Références

Chapitre 8

Annexes