UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TRAVAIL PRATIQUE

PRÉSENTÉ À

PHILIPPE GOULET COULOMBE

DANS LE CADRE DU COURS

DONNÉE MASSIVE & APPRENTISSAGE AUTOMATIQUE AVEC APPLICATION EN ÉCONOMIE

ECO930J, Groupe 40

PAR

VAUDESCAL GUILLAUME (VAUG30119904)

SAMUEL LALONDE-TREMBLAY (LALS12039506)

NAILA AGHANIM (AGHN21599906)

DATE DE REMISE

15 AVRIL 2022

SOMMAIRE :

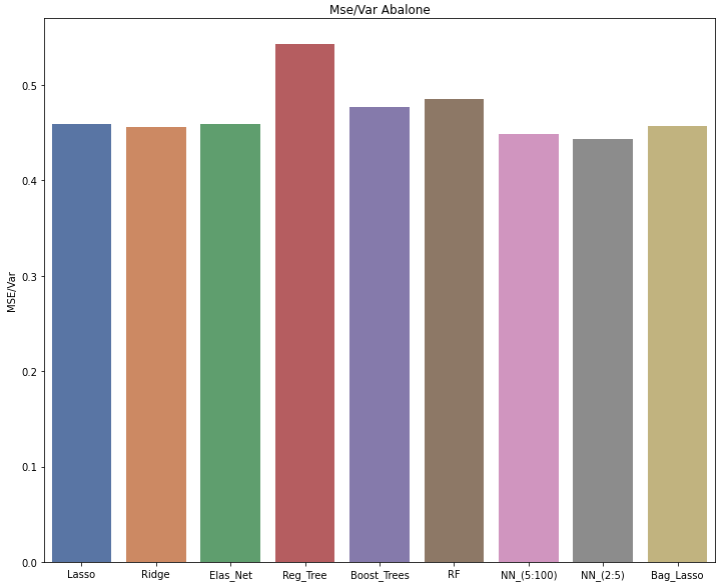
Introduction :

Le présent travail rend compte de l'utilisation de plusieurs algorithmes de machine learning tel que random forest ou neural network sur des bases de données relativement "classiques". Afin d'entraîner et tester nos modèles, on sépare chaque base de données en deux échantillons distincts ; en un training et test set, qui respectivement correspondent à 80% et 20% (à la fin de l'échantillon pour les bases de données macro) de l'échantillon pour chaque base de données. Puis, on réalise un test de performance des différents algorithmes en fonction des différentes bases de données, que nous illustrons en bar plot. Enfin, après avoir préciser nos choix d'hyper-paramètres pour chaque base de données, nous réalisons un histogramme des variables les plus importantes selon deux algorithmes ; Lasso et Random Forest pour les datasets *California Housing*, *Unemployment (h=1)*, et *Inflation (h=1)*. Les données de forme numérique ont été standardisées.

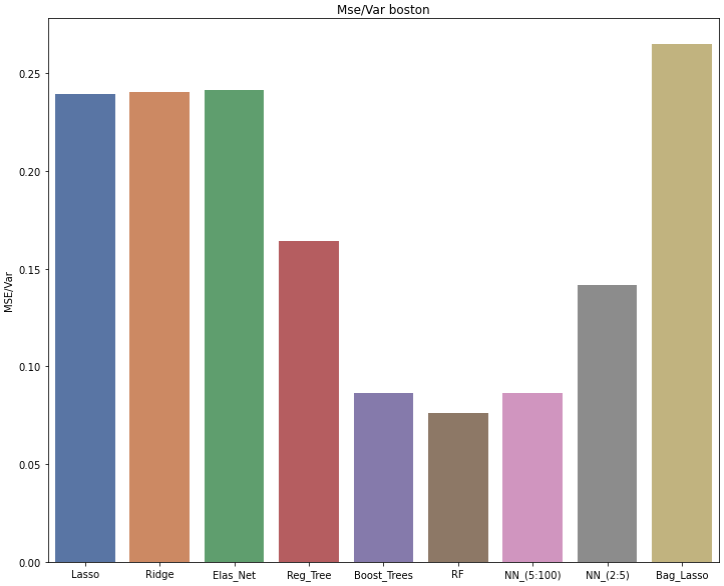
I - Rapportez les test MSEs (dans un barplot pour chaque data set) divisés par la variance de Y_test. Discutez :

Dans cette partie nous rapportons les tests MSEs divisés par la variance Y_test et nous décrivons les principaux résultats. Mais avant cela, il est intéressant de rappeler la définition de l'erreur quadratique moyenne et pourquoi nous divisons cette dernière par la variance.

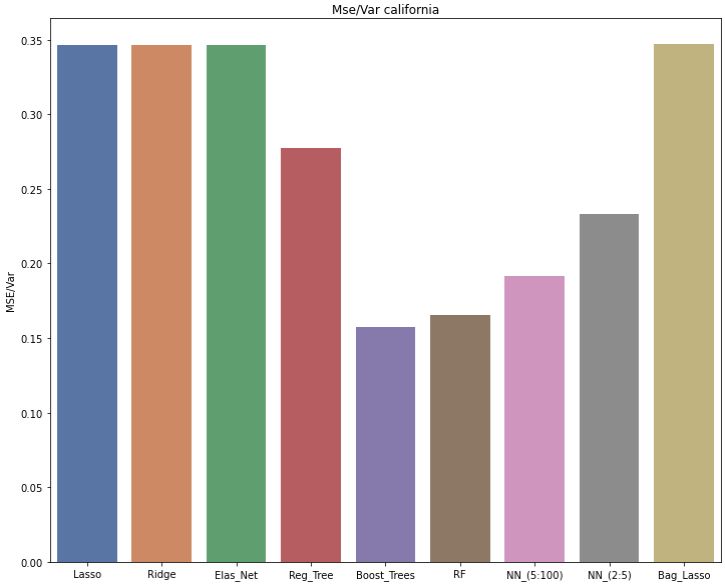
En statistique, l'erreur quadratique moyenne (MSE) d'un estimateur mesure la moyenne des carrés des erreurs, c'est-à-dire la différence quadratique moyenne entre les valeurs estimées et la valeur réelle. On la divise ici par la variance comme métrique de mesure de performance afin de standardiser les résultats et pouvoir effectuer des comparaisons.



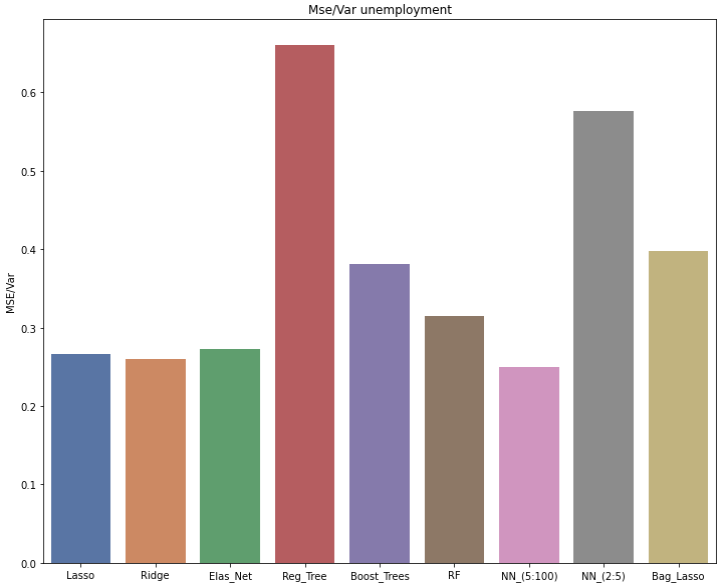
a) *Abalone*



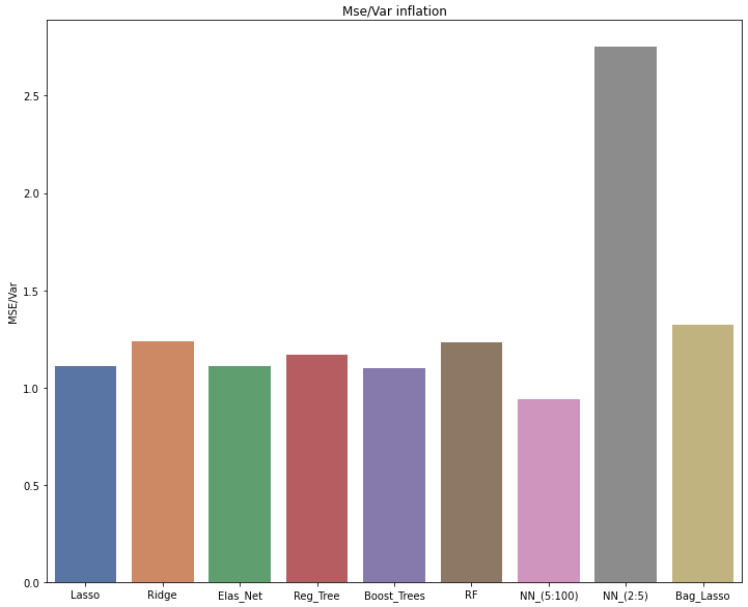
b) *Boston Housing*



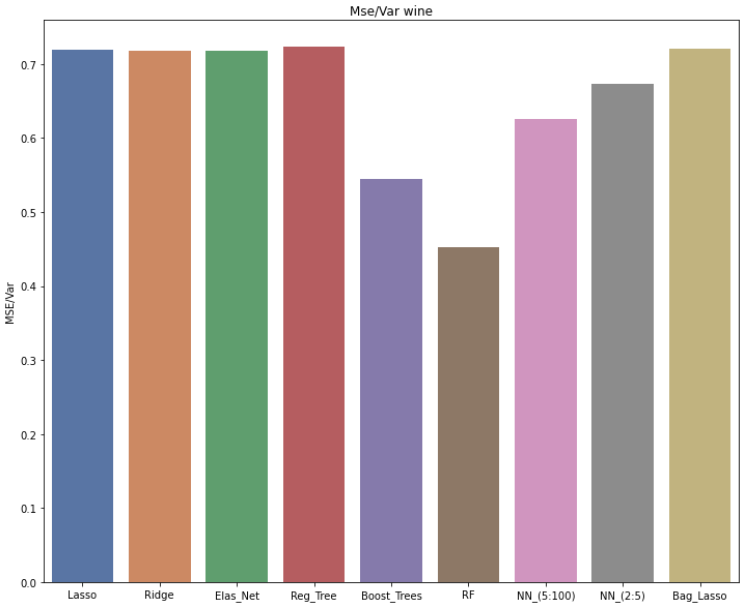
c) *California Housing*



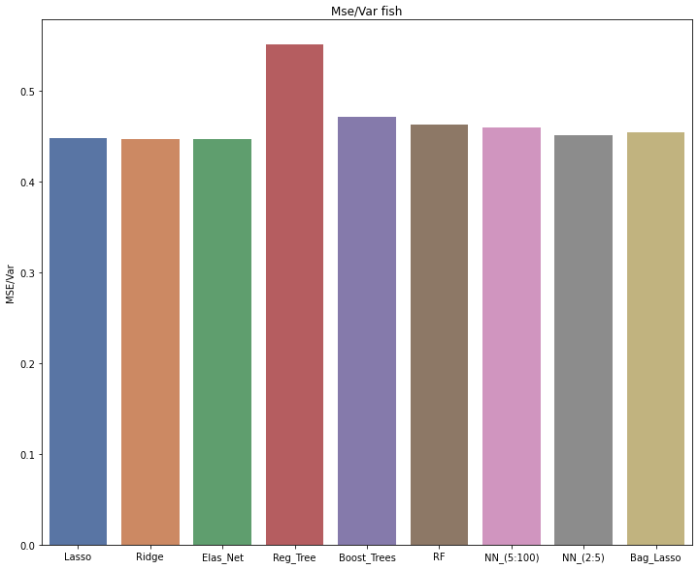
d) *US Unemployment Rate (h=1)*



e) *US Inflation (h = 1)*



f) *White Wine*



g) *Fish Toxicity*

Figure 1 : Résultats des prédictions empiriques.

Notes : Outil de mesure de la performance : MSE/Variance. Les modèles comparés pour chaque base de données sont : Lasso / Ridge / Elastic Net / Regression Trees / Boosted Trees / Random Forest / Neural Network (5;100) / Neural Network (2;5) / Bagging Lasso.

Les résultats des prédictions empiriques dans un premier temps semblent nous illustrer que les modèles non paramétriques tels que Random Forest ou Neural Network, qui captent plus efficacement les relations non-linéaires entre les variables prédictives, semblent mieux performer ou aussi bien que les modèles paramétriques comme Lasso ou Ridge. Pour illustrer cela, on peut prendre l'exemple du dataset *Boston Housing* et *white wine*. En effet, on peut noter dans le cas de *Boston Housing* que Random Forest est le modèle le plus performant (MSE/Variance = 0.07) suivit ex aequo de Boosted Trees et Neural Network 5;100 (MSE/Variance = 0.08) comparativement à Lasso et Ridge (MSE/Variance = 0.24) qui performent relativement moins bien. Également dans la même optique, avec le dataset *white wine*, Random Forest est le modèle le plus performant (MSE/Variance = 0.45) en contraste avec Lasso et Ridge (MSE/Variance = 0.71).

D'autre part, on peut noter pour les datasets *Fish Toxicity* et *Abalone*, que les résultats entre tous les modèles sont relativement similaires. En ce sens, pour le dataset *Fish Toxicity*, les résultats oscillent autour 0.45 (MSE/Variance) où les modèles paramétriques comme Lasso et Ridge, performent légèrement mieux. De surcroît, pour le dataset *Abalone*, les résultats oscillent également autour de 0.45 (MSE/Variance), où pour ce dernier ce sont les modèles non paramétriques comme les Neural Network qui performent légèrement mieux.

Enfin, pour le dataset *US Inflation (h = 1)*, le Neural Network 2;5 semble sous-apprendre (trade-off biais-variance) avec une performance de prédiction de 2.7 (MSE/Variance). Cela peut être expliqué par le faible nombre de couches et de neurones du Neural Network, qui semble peiner à estimer efficacement les relations "complexes" et non-linéarités entre les variables prédictives (plus de 621) de cette base de données. En ce sens, un Neural Network avec un plus grand nombre de couches et de neurones, tiendrait alors mieux compte des non-linéarités entre les variables prédictives, comme l'indique d'ailleurs nos résultats où le Neural Network 5;100 est le modèle le plus efficace, avec une performance de prédiction de 0.92 (MSE/Variance).

II - Rapportez les valeurs d'hyper-paramètres choisis par la validation croisée. Pour le modèle (i), rapportez un histogramme des 200 lambdas. Inclure 3 lignes verticales : la moyenne, la médiane, et la valeur lambda choisie pour le modèle (a). Discutez en faisant des liens avec les résultats en 4.

TABLE 1 – Valeurs des hyper-paramètres

	L lambda	R lambda	EN lambda	EN alpha	RT deph	BT learn rate	BT deph
Abalone	0,002	0,8	0,002	1	5	0,01	4
Boston	0,0009	8	0,03	0,001	9	0,01	3
California	0,00000001	0,2	0,00001	0,01	9	0,1	4
Unemployment	0,06	911	6	0,001	4	0,05	3
Inflation	0,05	162	0,05	1	3	0,05	1
Wine	0,006	49	0,01	0,001	4	0,1	4
Fish Tox	0,00000001	20	0,03	0,001	4	0,005	4

TABLE 2 – Valeurs des hyper-paramètres (suite)

	BT tree	RF mtry	NN 5 :100 learn rate	NN 5 :100 early stop
Abalone	750	1/3	0,01	50
Boston	750	1/3	0,001	20
California	750	0,5	0,001	20
Unemployment	750	sqrt	0,01	10
Inflation	500	sqrt	0,001	50
Wine	750	1/3	0,001	10
Fish Tox	750	1/3	0,01	10

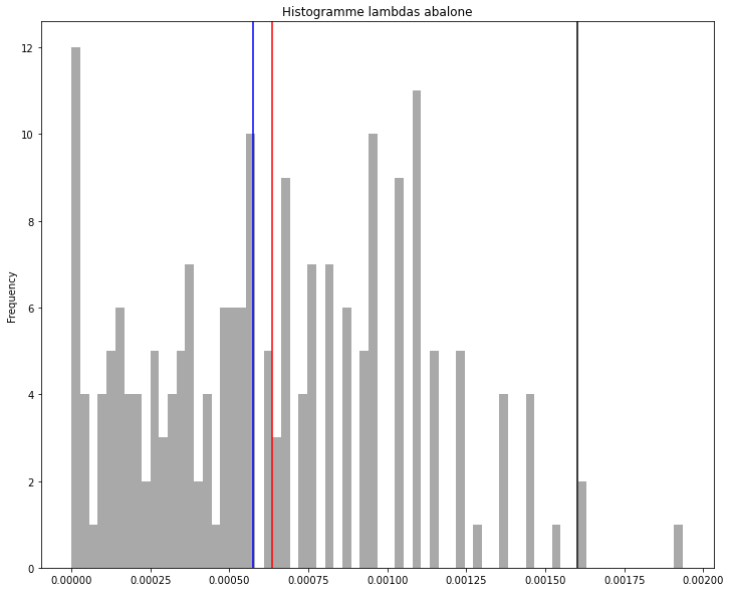
TABLE 3 – Valeurs des hyper-paramètres (suite)

	NN 2 :5 learn rate	NN 2 :5 learn rate
Abalone	0,1	100
Boston	0,01	50
California	0,01	100
Unemployment	0,001	20
Inflation	0,01	10
Wine	0,001	50
Fish Tox	0,05	10

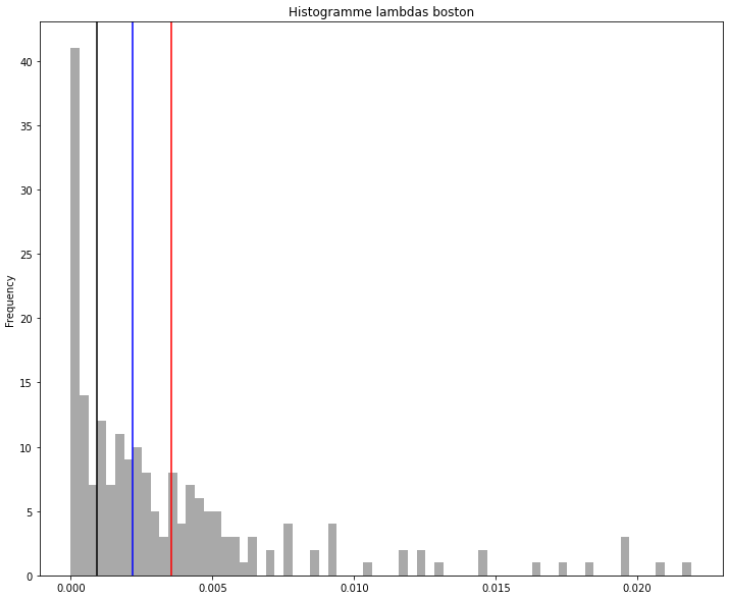
Les tableaux précédents portant sur les choix des hyperparamètres montrent que pour le Boosting Tree, le nombre d'arbres choisit par la validation croisée est dans la majorité des cas 750 (pour 6 dataset / 7), ce qui correspond au nombre maximum que l'on a testé. Cela laisse penser qu'on pourrait tester avec un nombre plus grand d'arbres afin d'optimiser un maximum l'algorithme.

De même pour la profondeur du Boosting Tree, la valeur maximale est de 4 et il serait intéressant de l'augmenter sachant que la validation croisée a retenu cette valeur pour plus de la moitié des modèles.

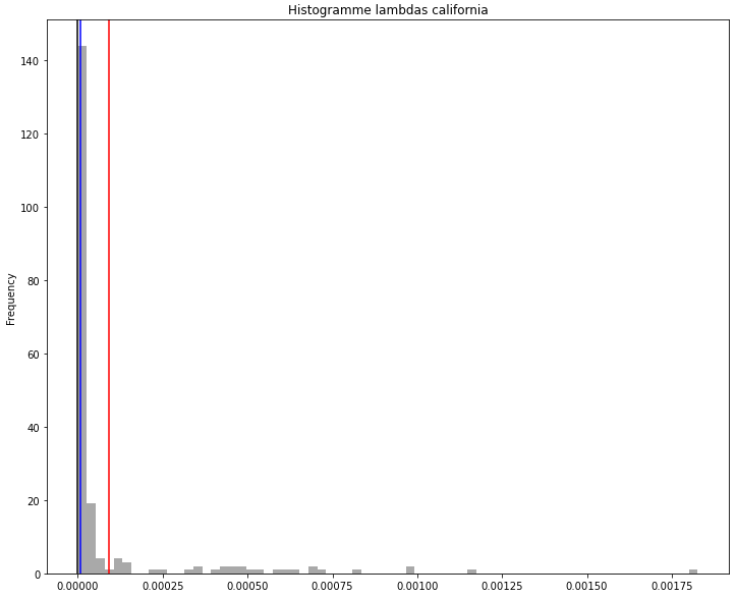
Concernant les alphas de lasso pour le dataset California Housing, on observe sur les graphiques ci-dessous que la majorité des observations sont à gauche, ce qui montre que l'on pourrait tester des valeurs plus faibles pour les alphas.



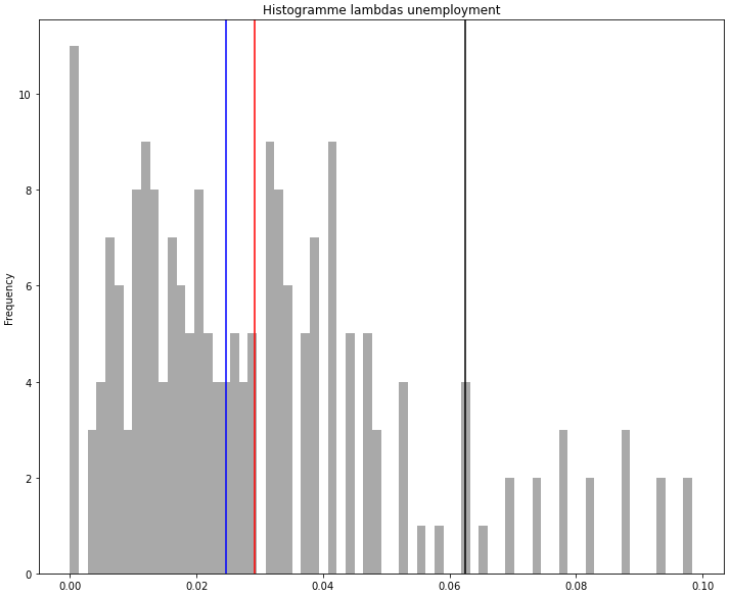
a) *Abalone*



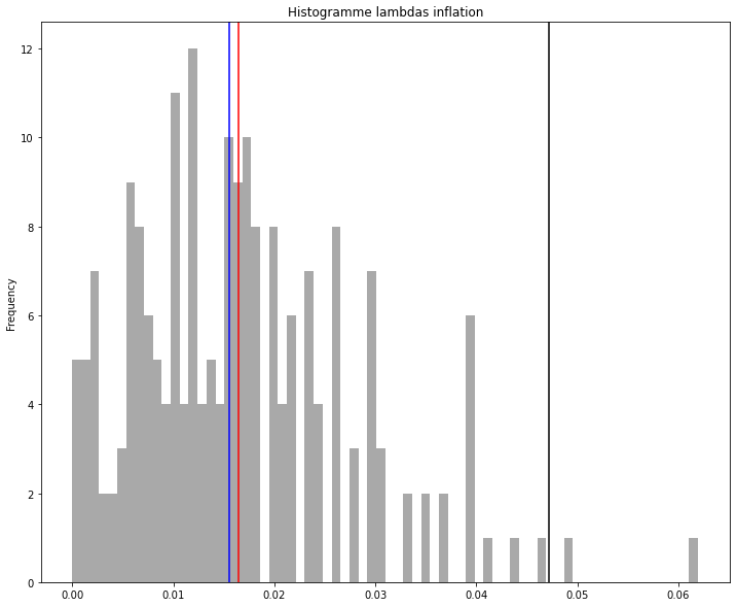
b) *Boston Housing*



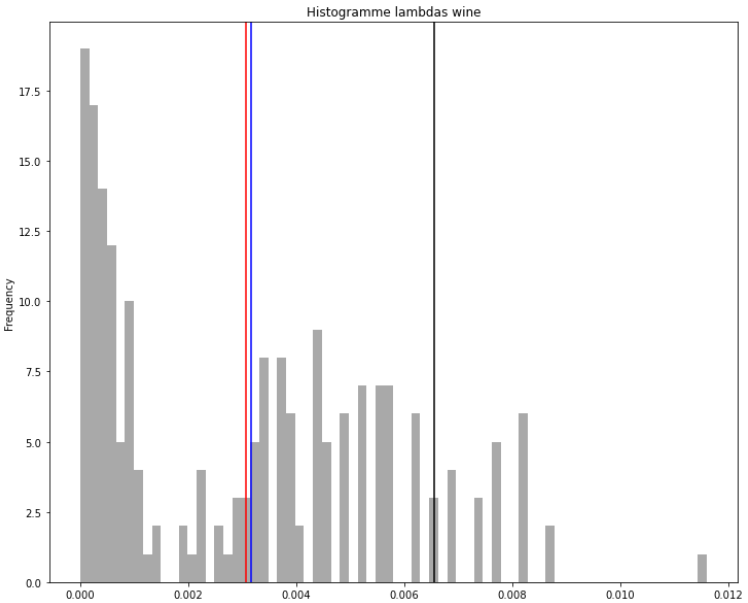
c) *California Housing*



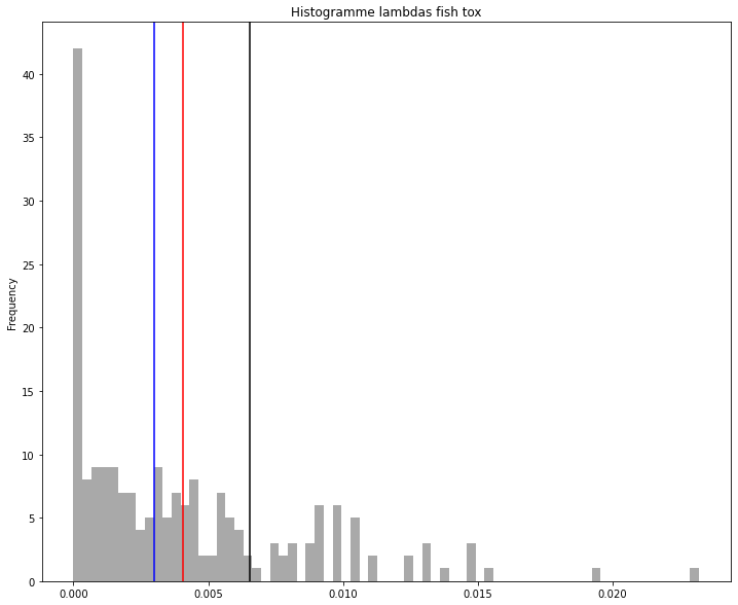
d) *US Unemployment Rate (h=1)*



e) *US Inflation (h = 1)*



f) *White Wine*



g) *Fish Toxicity*

Figure 2 : Histogrammes des 200 lambdas pour chaque datasets.

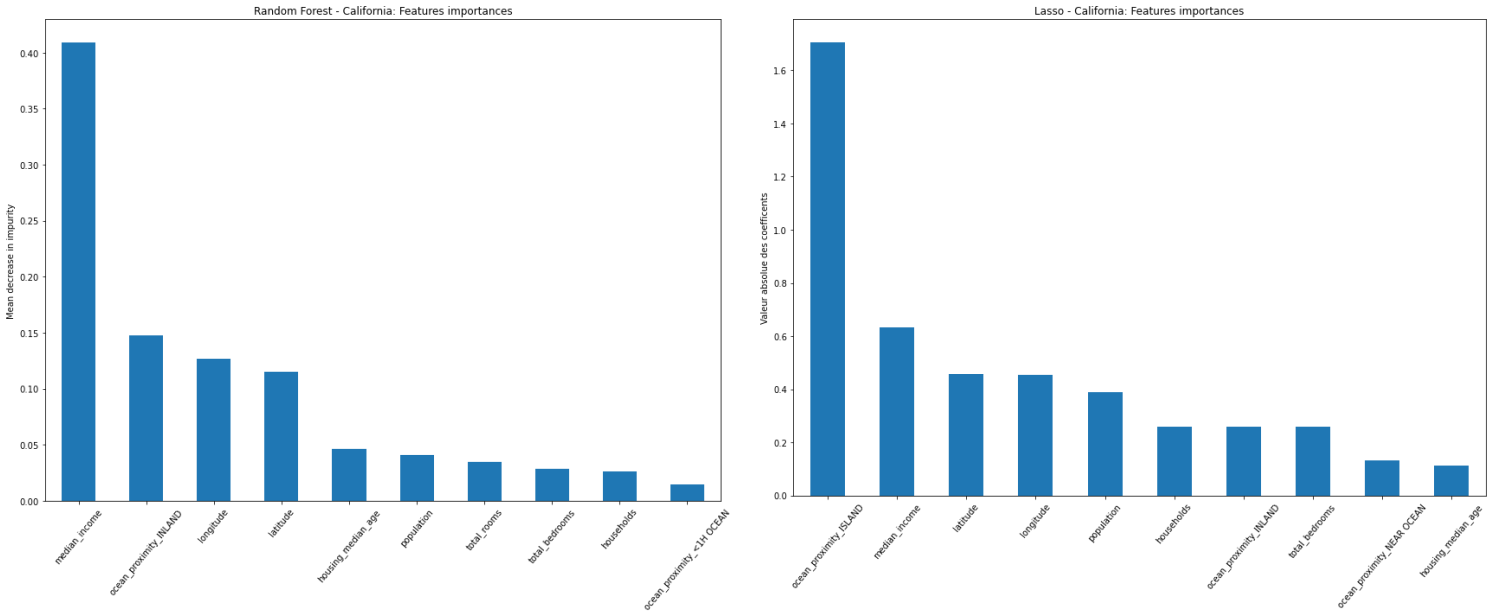
Notes : Moyenne ; ligne rouge / Médiane ; ligne bleu / Valeur lambda ; ligne noire. Nombre de lambdas pour chaque dataset : 300.

Nous allons désormais mettre en lien les valeurs des hyperparamètres choisit avec les erreurs de prédiction représentées en I) pour le lasso. Premièrement, pour le dataset *Abalone*, on observe sur le graphique en I) que la prédiction n'est pas bonne, donc le choix de l'hyperparamètre n'est pas optimal. Tester d'autres valeurs pour les lambdas permettra sûrement une meilleure prédiction. Nous observons la même chose pour les datasets *Boston Housing* ainsi que *California Housing*. On remarque que ces 3 dataset ont une valeur assez faible pour le lambda retenu (0.002, 0.0009 et 0.00000001 respectivement), ce qui laisse penser qu'une valeur faible de l'hyperparamètre pour le lasso mène à moins bonne prédiction.

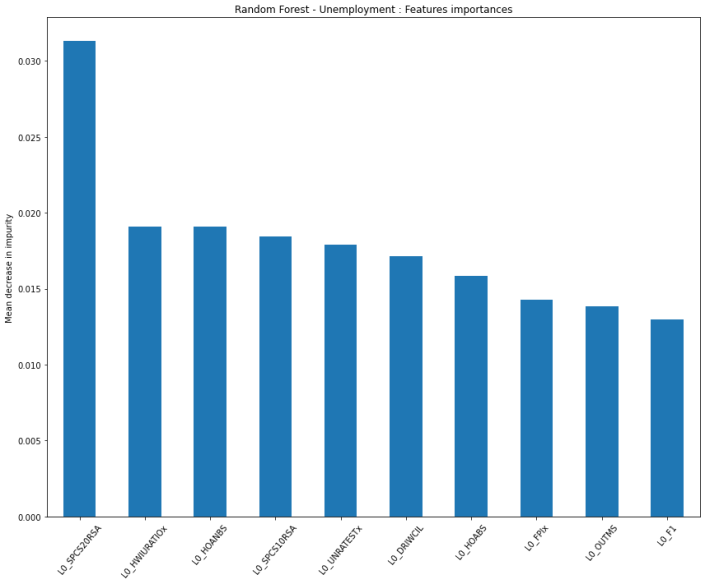
Cela se confirme car on observe pour les datasets *US Unemployment* et *US Inflation* une meilleure prédiction du lasso, accompagnée d'une valeur du lambda plus élevée (0.06 et 0.05 respectivement). Pour finir, *White Wine* et *Fish Toxicity* ont des prédiction moins bonnes avec lasso et ont une valeur du lambda faible (0,006 et 0.00000001 respectivement). On peut donc conclure qu'avec une valeur très faible pour le lambda, la prédiction du lasso est moins bonne.

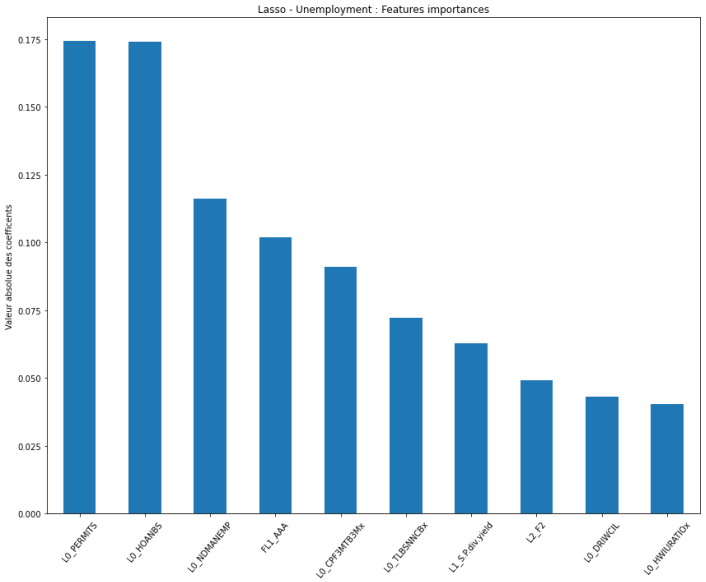
III - Dans le cas de California Housing, Unemployment (h=1), et Inflation (h=1). Rapportez les variables qui semblent les plus importantes selon Lasso. Rapportez les variables qui semblent les plus importantes selon Random Forest en utilisant le «variable importance ». Discutez.

Dans cette partie nous rapportons les variables qui semblent les plus importantes selon Lasso et Random Forest pour les datasets de *California Housing*, *Unemployment (h=1)*, et *Inflation (h=1)*.

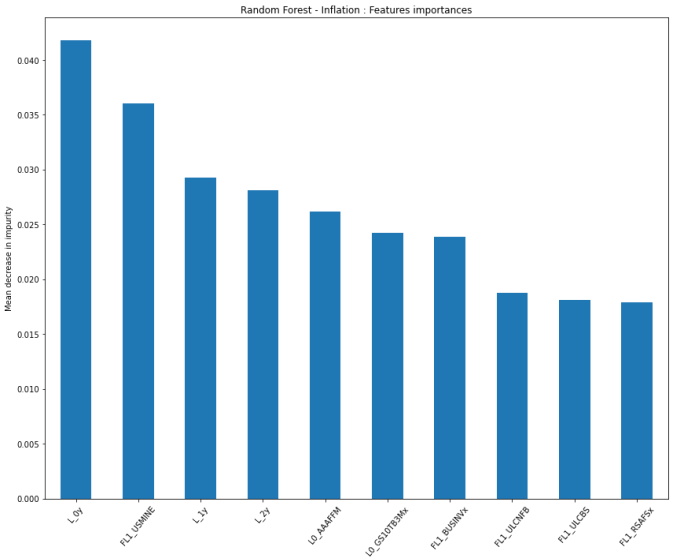


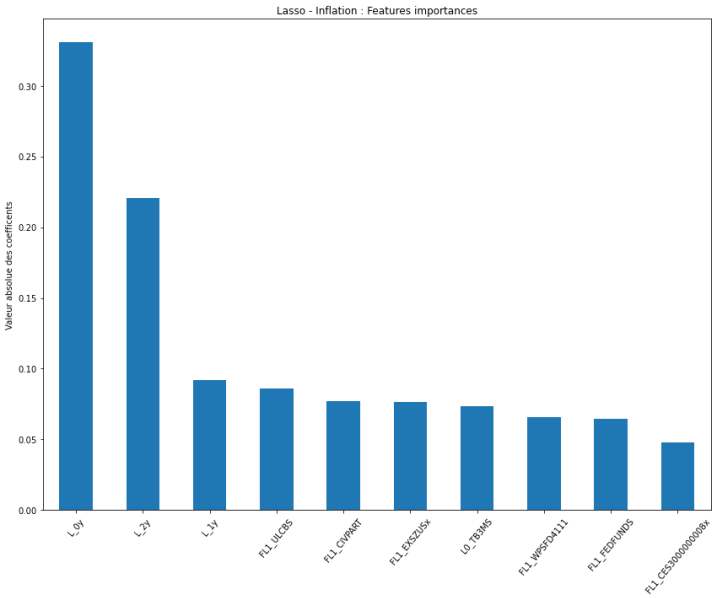
a) *California Housing features importances RF/Lasso*





b) *US Unemployment Rate (h=1) features importances RF/Lasso*





c) *US Inflation (h = 1) features importances RF/Lasso*

Figure 3 : Features importances selon *RF*/*Lasso* :

Notes : Les datasets utilisées sont *California Housing*, *Unemployment (h=1)*, et *Inflation (h=1)*. Voir **annexe** pour le détail des variables.

La figure 3 nous indique dans le cas du dataset de *California Housing*, que les variables prédictives les plus importantes sont relativement différentes selon Random Forest et Lasso. En effet, on peut noter que la variable la plus importante selon Random Forest est le *median_Income* (revenu médian des ménages d'un bloc de maisons) comparativement à Lasso qui est *oceanProximity* (emplacement de la maison par rapport à l'océan/la mer). Bien que dans chacun des modèles, ces variables se classent deuxième respectivement en termes d'importance. En guise d'interprétation de ces deux facteurs qui influent sur le prix des maisons en Californie, on sait que statistiquement[1] les individus ont une préférence pour habiter proche d'un océan ou littoral, augmentant ainsi la demande et de ce fait le prix de l'immobilier dans ces zones convoitées. Également, le revenu médian des ménages d'un bloc de maisons est un facteur important car on peut s'attendre que plus cette variable augmente, plus les prix de l'immobilier dans une zone sont élevés (corrélation positive). Il serait alors intéressant de tester statistiquement cette hypothèse en guise d'extension de ce travail.

Pour le dataset de *US Unemployment Rate (h=1)*, on peut noter que les variables prédictives les plus importantes sont très différentes selon Random Forest et Lasso. En effet, selon Random Forest c'est *L0_SPCS20RSA* qui correspond à l'indice composé des prix des maisons S&P/Case-Shiller pour 20 villes, alors que pour Lasso c'est *L0_PERMITS* qui correspond aux nouveaux logements privés autorisés par permis de construire dans la région de recensement du Sud. De surcroît, on peut remarquer sur la figure 3, que les variables importantes de *US Unemployment Rate (h=1)* sont légèrement asymétriques, ce qui indique que ce n'est pas un faible nombre de variables qui permet de prédire efficacement notre variable dépendante mais au contraire un ensemble de variables (surtout dans le cas de Random Forest).

Enfin, pour le dataset *US Inflation (h = 1)*, on observe que les variables prédictives selon Random Forest et Lasso sont relativement similaires. En effet, les variables les plus importantes, qu'importe le modèle, sont surtout les lags de l'inflation. En ce sens, cela indique que l'inflation est essentiellement corrélée avec elle-même et on peut donc utiliser ses valeurs passées (avec un certain nombre de lags) afin de prédire ses valeurs futures. Ici, pour Random Forest et Lasso la variable prédictive la plus importante est *L_0y* qui correspond au premier lag de l'inflation et soutient par conséquent notre point précédent.

[1] United States Census Bureau California 2020 Census, URL : https://www.census.gov/library/stories/state-by-state/california-population-change-between-census-decade.html

Références :

Cortez, P, A. Cerdeira, F. Almeida, T. Matos and J. Reis. (2009). Viticulture Commission of the Vinho Verde Region (CVRVV). *University of Minho, Guimarães, Portugal. Retrieved*, *15*, 2020.

Federal Reserve Economic Data, Inflation rate
Federal Reserve Economic Data, Unemployment rate

Géron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow: Concepts. *Tools, and Techniques to build intelligent systems*.

Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81-102.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., & Ford, W. B. (1994). The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, *48*, p 411.

United States Census Bureau, Administrative and Customer Services Division, Electronic Products Development Branch.

United States Census Bureau, California 2020 Census.

Annexe :

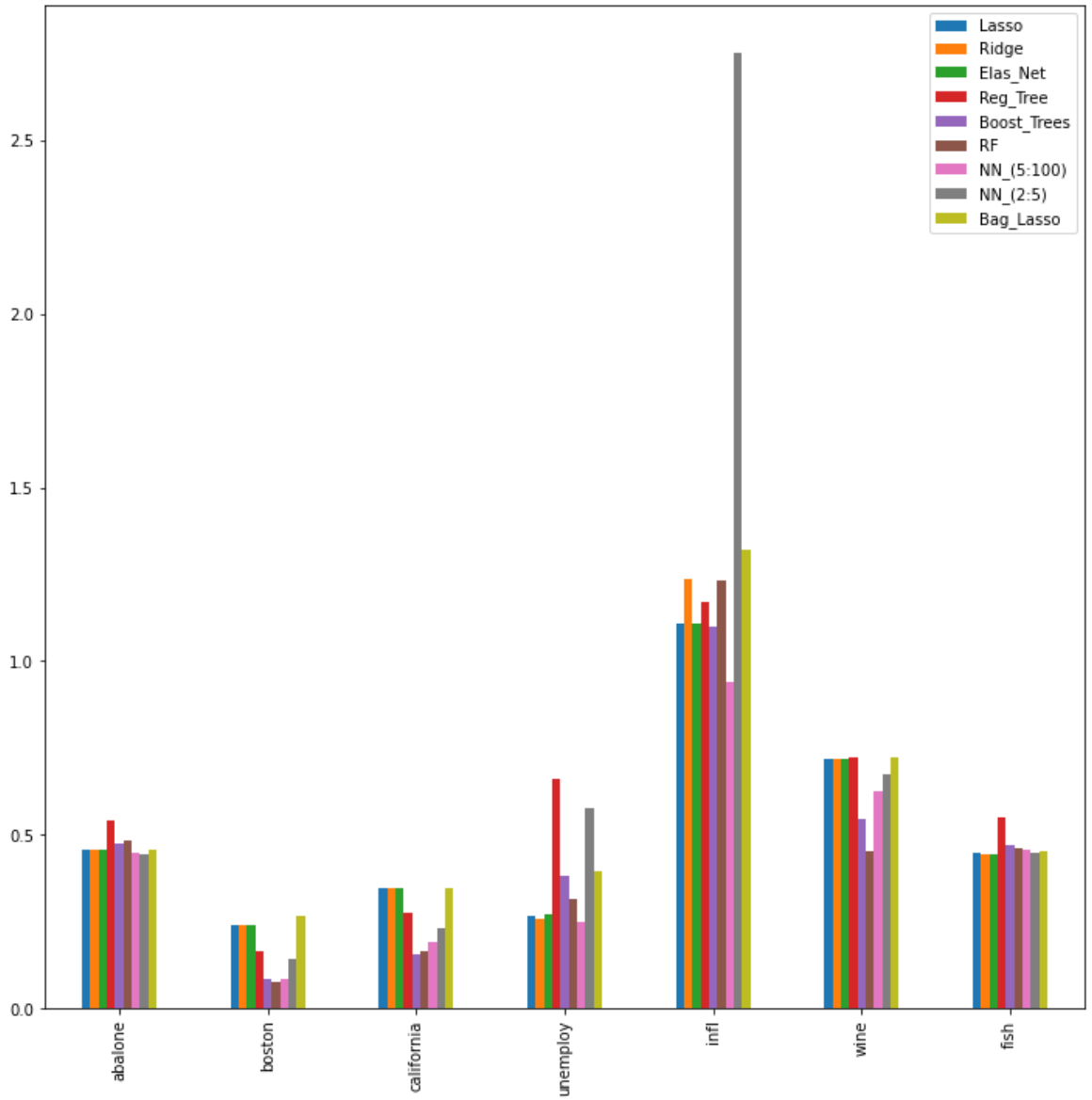


Figure 1.1 : Comparaison des résultats des prédictions empiriques entre les datasets.

Notes : Outil de mesure de la performance : MSE/Variance. Les modèles comparés pour chaque base de données sont : Lasso / Ridge / Elastic Net / Regression Trees / Boosted Trees / Random Forest / Neural Network (5;100) / Neural Network (2;5) / Bagging Lasso.

Détail des variables les plus importantes de la figure 3 :

California Housing RF :

1. *median_Income* : Revenu médian des ménages d'un bloc de maisons (mesuré en dizaines de milliers de dollars US).
2. *oceanProximity* : Emplacement de la maison par rapport à l'océan/la mer.
3. *longitude* : Mesure de la distance à l'ouest d'une maison ; plus la valeur est élevée, plus la maison est à l'ouest.

California Housing Lasso :

1. *oceanProximity* : Emplacement de la maison par rapport à l'océan/la mer.
2. *median_Income* : Revenu médian des ménages d'un bloc de maisons (mesuré en dizaines de milliers de dollars US).
3. *latitude* : Mesure de la distance qui sépare une maison du nord ; plus la valeur est élevée, plus la maison est au nord.

US Unemployment Rate (h=1) RF :

1. *L0_SPCS20RSA* : Indice composé des prix des maisons S&P/Case-Shiller pour 20 villes.
2. *L0_HWIURATIOx* : Rapport entre le nombre de personnes recherchées et le nombre de chômeurs.
3. L0_HOANBS : Indice des heures travaillées dans le secteur des entreprises non agricoles.

US Unemployment Rate (h=1) Lasso :

1. *L0_PERMITS* : Nouveaux logements privés autorisés par permis de construire dans la région de recensement du Sud (milliers, SAAR).
2. *L0_HOANBS* : Indice des heures travaillées dans le secteur des entreprises non agricoles.

3. *L0_NDMANEMP* : Tous les employés, biens non durables

US Inflation (h = 1) RF :

1. *L_0y* : Premier lag
2. *FL1_USMINE* : Tous les employés, Mines et exploitation forestière
3. *L_1y* : Deuxième lag

US Inflation (h = 1) Lasso :

1. *L_0y* : Premier lag
2. *L_2y* : Troisième lag
3. *L_1y* : Deuxième lag