3) Method

Ce mémoire à pour objectif de tester les performances des algorithmes de sélection des variables présenté dans la partie précédente. Pour ce faire, nous devons générer des ensembles de données avec une structure de corrélation (paramétré manuellement) connue et précise, chaque DGP a une corrélation unique et approprié. Nous avons créés 1000 bases de données par DGP (processus de génération de donnée), soit 4000 bases de données au total. Chaque base de données contient un ensemble de 51 variables (1 variable expliquée Y et 50 variables explicatives x1-x50) et 100 observations. Enfin, nous testons chaque algorithme sur les 4000 ensembles de données et concluons sur leurs résultats.

3.1 Simulation de variables corrélées avec une distribution multinormale

Pour chaque DGP, 1000 échantillons de taille N = 100 ont été générés à partir d'une distribution normale multivariée. C'est-à-dire que chaque vecteur d'un échantillon suit une distribution normale univariée et la relation entre deux vecteurs d'un échantillon est déterminée par une matrice de covariance de variance spécifiée.

3.2 Correlation matrix

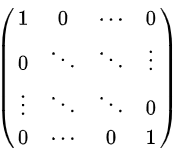
Simuler une structure de corrélation entre les variables implique de créer une matrice de corrélation. On sait que les matrices de corrélation sont semi-définies positives symétriques. Afin de remplir ces conditions, nous avons choisi d’utiliser des matrices Toeplitz. La matrice Toeplitz est une matrice à constante diagonale, ce qui signifie que les entrées de la matrice ne dépendent que des différences des indices, si elle est symétrique, elle possède toutes les propriétés d'une matrice de corrélation. On obtient donc une matrice de corrélation avec une diagonale qui vaut 1 et des covariances entre les variables de valeur différente en fonction du DGP. En valeur absolue, on estime que deux variables fortement corrélées ont un coefficient de corrélation compris entre 1 et 0,6, deux variables moyennement corrélées ont un coefficient de corrélation compris entre 0,6 et 0,3, Deux variables faiblement corrélées ont un coefficient de corrélation compris entre 0,3 et 0.

3.4 Data Generating Processes (DGP)

Chaque DGP est composé de 1000 bases de données, eux-mêmes composées de 100 lignes (taille de l’échantillon) et de 51 colonnes(nombres de variables), dans lequel la 1ere colonne est la colonne Y et se suit par les colonnes X1 à X50 (Y||X1-X50).

3.4.1 First Data Generating Process (DGP1)

Tout d’abord, nous avons généré un ensemble de donnée classique représentant le modèle parfait, c’est-à-dire sans présence de corrélation entre les variables, et sans variables extrêmes. Pour la création d’une base de données, nous avons simulé un ensemble de données (matrice X) de 50 variables en utilisant la distribution normale multivariée et utilisé la matrice d'identité comme matrice de covariance de variance.



Ensuite, on génère Y en fonction de l'équation suivante :

Y = 1.5 x X1 + 0.9 x X2 + 1 x X3 + 0.1 x X4 - 0.5 x X5

Enfin, nous avons joint la colonne Y avec la colonne X1 à X50, afin de créer une base de données avec 51 colonnes. La valeur de Y est dépendante des valeurs obtenues aléatoirement des variables X1 à X5, on pourrait interpréter la relation entre Y et les variables X1 à X5 comme des corrélations positives et négative, et la relation entre les variables X6 à X50 et Y comme des corrélations nulles.

3.4.2 Deuxième processus de génération de données (DGP2)

Pour le 2e DGP, nous avons généré un ensemble de données dans laquelle seulement les variables X1 à X5 sont corrélées de manières positives entre-elle et les variables X6 à X50 ne sont pas corrélées entre elles. Pour créer cette base de données, nous avons simulé un 1er ensemble de données (matrice Xa) de 5 variables (X1 à X5) suivant une loi de distribution normale multivariée (moyenne=0), et de corrélation positive et symétrique entre les 5 premières variables. La matrice de corrélation pour les variables X1 à X5 s’est fait à l’aide de la matrice Toeplitz. On obtient une variable X1 qui est faiblement corrélée à X2, moyennement corrélée à X3 et X4, et fortement corrélée à X5 (Cov(X1,X5)=0.714 ).

Une image contenant table

Description générée automatiquement

Figure 1 : Corrélation entre les 5 variables (X1 à X5)

Ensuite, nous avons simulé un 2e ensemble de données (matrice Xb) de 45 variables (X6 à X50) suivant une distribution normale multivariée (moyenne =0) et de corrélation nulle entre ces variables. Pour la matrice de corrélation, nous avons utilisé la matrice identité comme dans le DGP1. Après avoir simulé les 2 ensembles de données, nous avons créé une matrice X qui joint les 2 ensemble de données, afin d’avoir qu’une seule table de donnée. Puis, on génère Y par l'équation suivante :

Y = 1.5 x X1 + 0.9 x X2 + 1 x X3 + 0.1 x X4 - 0.5 x X5

Enfin, nous avons joint la colonne Y à la matrice X, afin de créer une base de données avec 51 colonnes dans lequel seulement Y est corrélée aux 5 premières variables, qui sont elles-mêmes corrélées entre-elles. Y n’est pas corrélée aux variables X6 à X50, et elles ne sont elles-mêmes pas corrélées entre-elles.

3.4.3 Troisième processus de génération de données (DGP3)

Pour le 3e DGP, nous avons généré un ensemble de données dans laquelle les variables X1 à X50 sont corrélées de manières positives entre-elle. Pour réaliser une base de données, nous avons simulé un ensemble de données (X) de 50 variables (X1 à X50) suivant une loi de distribution normale multivariée de moyenne 0, et de corrélation positive et symétrique entre les 50 variables. La matrice de corrélation pour les variables X1 à X50 s’est fait à l’aide de la matrice Toeplitz. Chaque variable est fortement corrélée avec les variables les plus proches dans la matrice. Plus une variable est éloignée d'une autre, moins elle est corrélée. Par exemple, on obtient une variable X1 qui est fortement corrélée aux variables X2 à X16, moyennement corrélée aux variables X17 à X36 , et faiblement corrélée aux variables X37 à X50.

Une image contenant table

Description générée automatiquement

Figure 2 : Corrélation entre les 16 premières variables (X1 à X16)

Une image contenant table

Description générée automatiquement Une image contenant table

Description générée automatiquement

Figure 3 : Corrélation de X1 par rapport aux variables X17 à X50

Ensuite, on génère Y selon l'équation suivante :

Y = 1.5 x X1 + 0.9 x X2 + 1 x X3 + 0.1 x X4 - 0.5 x X5.

Enfin, nous avons joint la colonne Y à la matrice X, afin de créer une base de données avec 51 colonnes dans lequel Y est le résultat de la génération aléatoire de 5 premières variables. Les variables X1 à X50 sont corrélées entre-elles de manière positive et symétrique.

3.4.3 Quatrième processus de génération de données (DGP4)

Pour le DGP4, nous avons généré un ensemble de données dans laquelle les variables X1 à X50 ne sont pas corrélées mais qui présentent des valeurs extrême. Pour la création d’une base de données, nous avons simulé un ensemble de données (matrice X) de 50 variables en utilisant la distribution normale multivariée avec des valeurs extrêmes (« ExtremeValue ») et utilisé la matrice d'identité comme matrice de covariance de variance. Ensuite, on génère Y via l'équation suivante :

Y = 1.5 x X1 + 0.9 x X2 + 1 x X3 + 0.1 x X4 - 0.5 x X5

Enfin, nous avons joint la colonne Y à la matrice X, afin de créer une base de données avec 51 colonnes dans lequel Y est le résultat de la génération aléatoire de 5 premières variables. Les variables X1 à X50 présentes des valeurs extrêmes sans corrélation.

Une image contenant table

Description générée automatiquement

Figure 4 : Valeur maximal et minimal par variable aléatoire

3.5 Extraction des résultats

Pour évaluer la performance de chaque méthode, on calcul la probabilité que la procédures de sélection utilisé sélectionne seulement les 5 variables d’intérêts (c’est-à-dire les variables X1 à X5). Lorsqu’on lance la procédure de sélection pour une « table de donnée générée », les variables sélectionnées se voient attribuer un coefficient de corrélation. Par extension, les variables sans coefficient assigné sont les variables "non sélectionnées", elles apparaissent comme une valeur manquante dans les résultats de coefficient, ces variables sont automatiquement retirées. On retrouve donc une nouvelle table qu’on nommera « table de sélection » avec seulement les colonnes des variables sélectionnées. Par exemple, si dans une table de données générée, la méthode utilisée sélectionne seulement les variables X1 à X10, alors la table de sélection sera composée de colonne X1 à X10.

Afin de comparer les différentes méthodes de sélection, on compare premièrement les fréquences de sélection individuelles de chaque variable selon les différents ensembles de données. Et deuxièmement on compare les fréquences de sélections des 5 variables d’intérêts uniquement selon les différents ensembles de données.

Pour la fréquence individuelle, nous avons créer une tables qu’on nommera la « table de sélection en 1 », permettant de rassembler toutes les tables de sélection. Puis, nous avons remplacer chaque coefficient de la « table de sélection en 1 » par 1 et chaque valeur manquante par un 0. La matrice ainsi obtenue sera utilisée pour établir 1 diagrammes à barres qui donne la fréquence de chaque variable qui a été choisie individuellement. La fréquence de sélection de chaque variable obtenue est divisé par 1000 afin de pouvoir donner une fréquence de sélection en pourcentage (car maintenant la fréquence maximal de sélection est de 100), on peut donc les utiliser pour exprimer les probabilité de sélection des variables sur une échelle entre 0 et 100.

Pour la fréquence de sélection des 5 variables d’intérêt uniquement, nous construisons une nouvelle base de données composée de 0 et 1, qui attribue 1 lorsque ces 5 variables sont sélectionnées dans le même ensemble de données ; 0 si au moins une variable n'est pas sélectionnée avec les autres. Par exemple, nous définissons stepwise12345, comme la sélection conjointe de X1, X2,X3,X4, et X5 avec la méthode Stepwise : si cette condition est remplie, la valeur 1 est inscrite dans la colonne, sinon 0. Une fois ceci fait, il suffit de compter le nombre de 1 parmi les 1000 bases et de diviser la fréquence par 1000 pour avoir des résultats en pourcentages. Si la fréquence jointe est de 0/100, cela veut dire que X1, X2,X2,X3,X4 et X5 ne sont jamais sélectionnés conjointement, tandis que si la fréquence jointe est des 100/100, alors ces variables d’intérêts sont jointes à chaque fois.

Nous aurons pour chaque procédé énoncé ( Stepwise, Backward, Forward, LARS, LASSO, ElasticNet) deux histogrammes, dont l’une représentant la probabilité qu’une variable est choisie parmi les 1000 régressions, et l’autre représentant la probabilité que les 5 variables d’intérêt soient uniquement sélectionnées parmi les 1000 régressions.

4 Résultats

Maintenant que nous avons les graphiques associés à la probabilité qu'une variable soit sélectionnée par chacun des modèles ; ainsi que ceux des probabilités conjointes que plusieurs variables soient sélectionnées en même temps, décrivons-les afin de comparer et d'évaluer la précision de chaque modèle.

Pour évaluer et comparer les résultats de chaque modèle, nous avons choisi de structurer notre raisonnement en 4 grandes parties, représentant les 4 DGP. Puis nous avons crée 2 sous-partie dont l’une représente les méthodes de Statistical learning et l’autre les méthodes de Machin learning. Au sain de ces 2 sous-parties nous avons comparer les méthodes de statistical learning entre-elles et les méthodes de machin learning entre-elles. Enfin nous avons comparer le modèle le plus pertinent de statistical learning contre celui de machin learning afin de déterminer la méthode de sélection la plus efficace en fonction de chaque situation (c’est-à-dire en fonction de chaque DGP).

4.1 DGP1

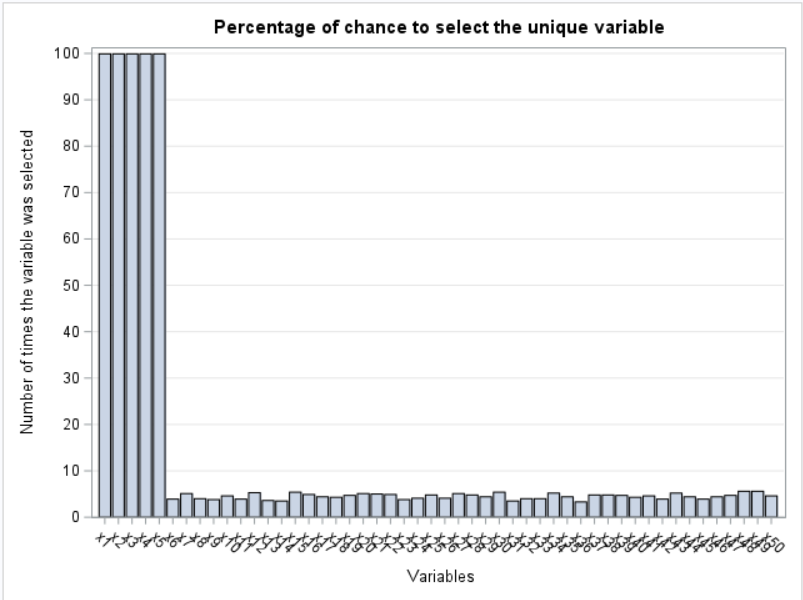
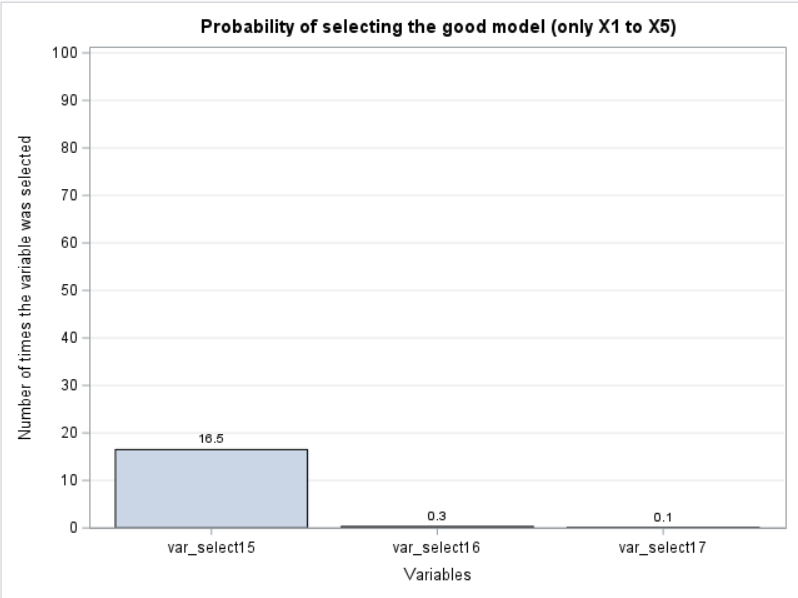
4.1.1 Statistical learning

Nous allons tout d’abord comparer les différentes méthodes de statistical learning en fonction des critère de sélection imposés. Dans cette partie, les méthode de sélection testées sont : Stepwise Forward et Backward.

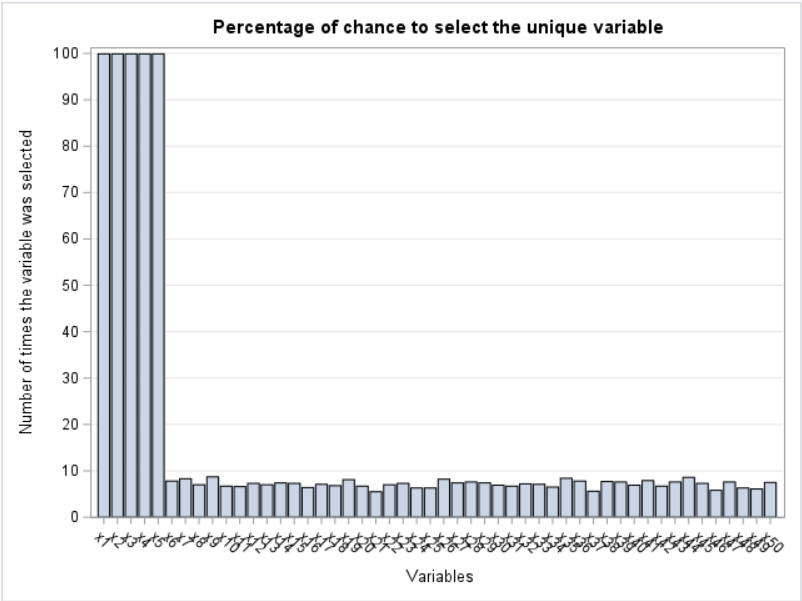
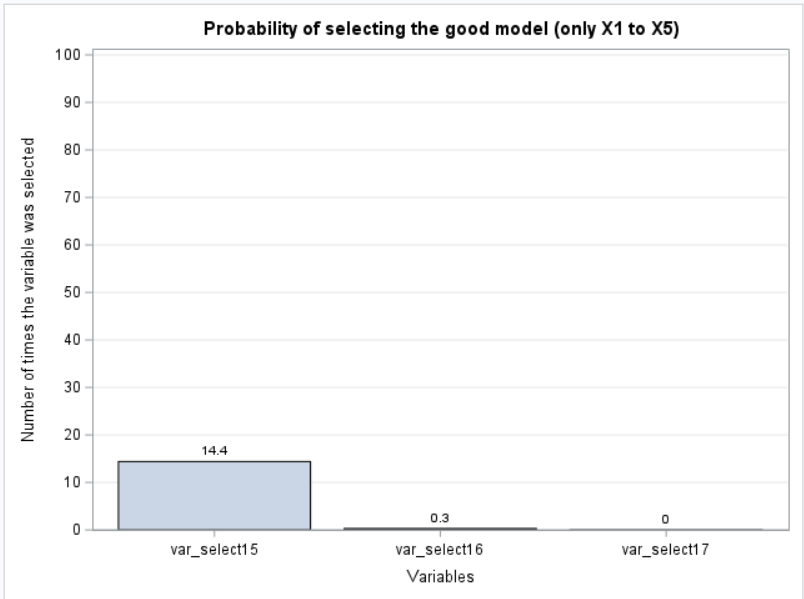
Chaque modèle testé présente 2 graphiques. A gauche, on retrouve les probabilités de sélection individuelle des variables. A droite, on retrouve les probabilités de sélections Jointes (càd la sélection de certaine variable en même temps) dans lequel nous allons principalement nous appuyer sur l’analyse de la sélection jointes des variables d’intérêts (X1 à X5) car c’est à partir de ce résultats que l’on va pouvoir juger l’efficacité de la méthode.

**SBC**

Forward :

Backward :

Avec le critère de sélection SBC, on peut observer que toutes les variables d’intérêts ont été sélectionnées individuellement dans 100% des cas. Cependant on remarque que les méthodes de sélection ne retrouve les variables d’intérêts jointes que dans 15% des cas en moyenne, ce qui est faible. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Forward qui présente une probabilité de sélection parfaite de 16,5%.

**BIC**

Forward :

Une image contenant table

Description générée automatiquement

Backward :

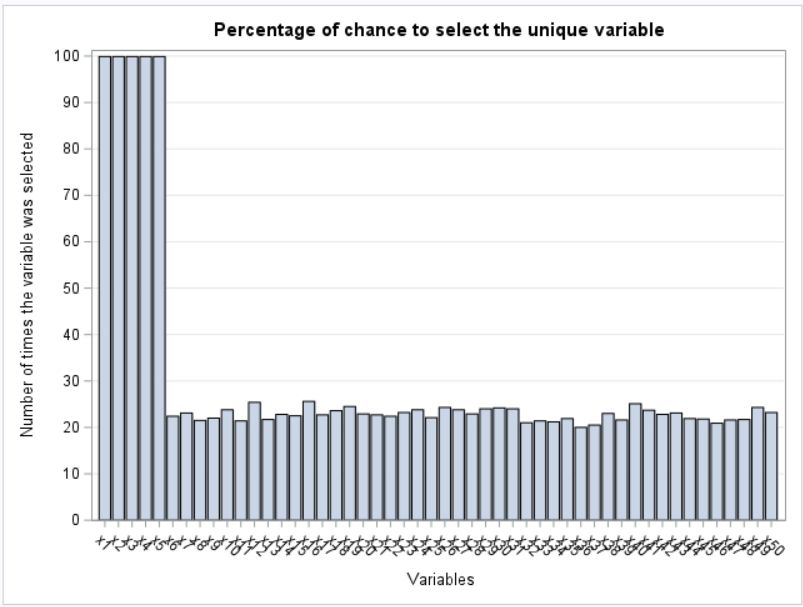
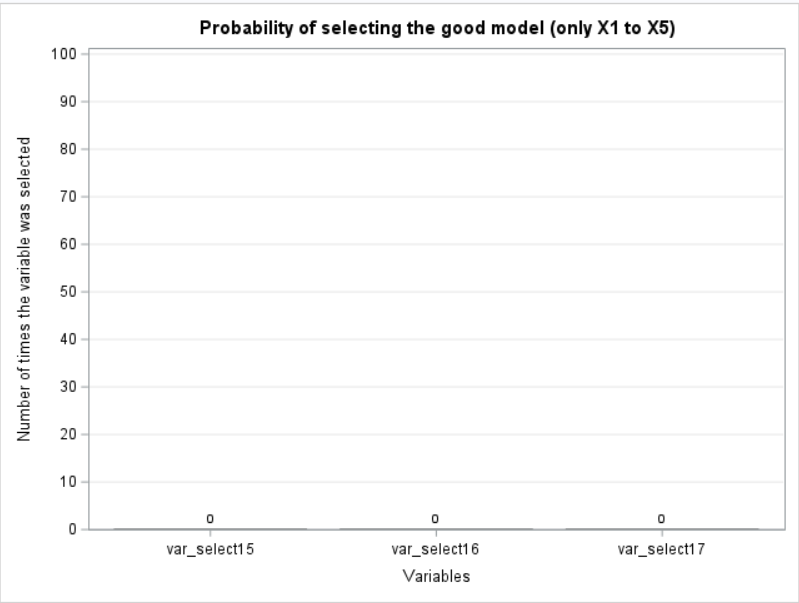
Une image contenant table

Description générée automatiquement

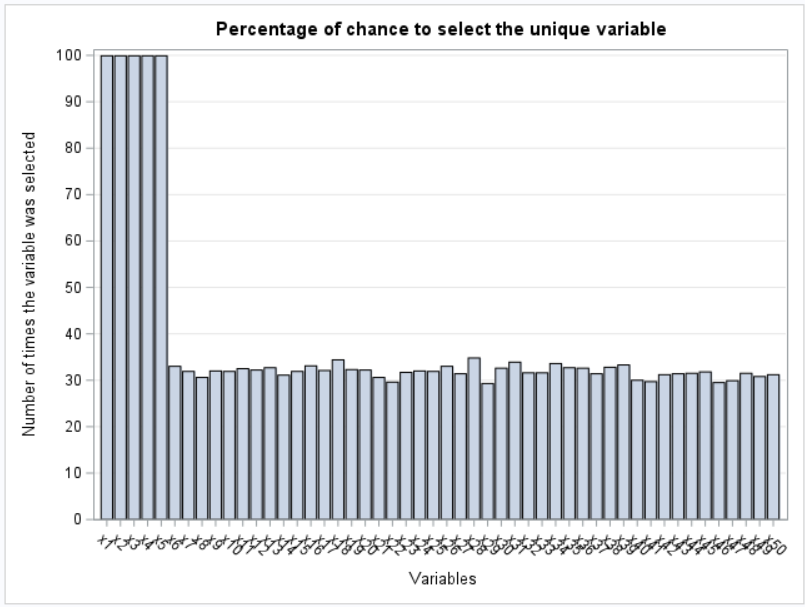
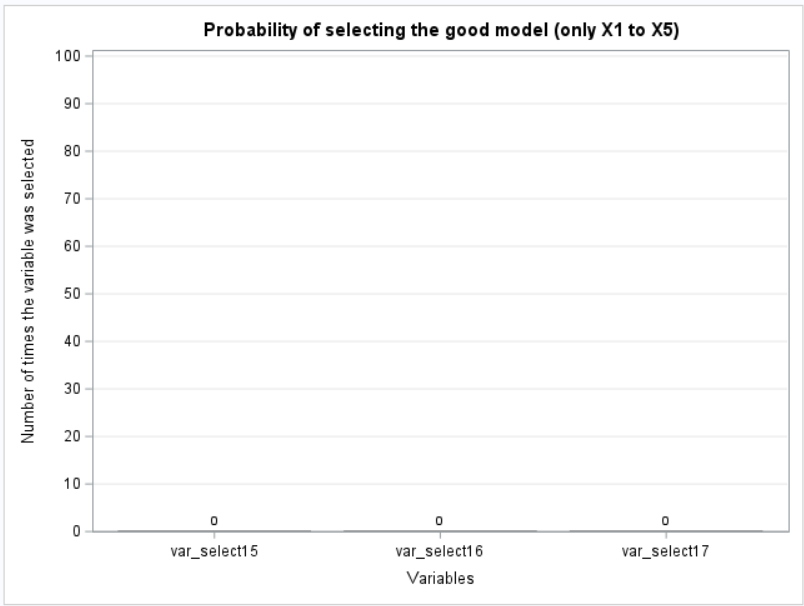
Avec le critère de sélection BIC, nous avons choisi de présenter un seul graphique de sélection individuel car tous ces graphiques mènent au même résultat. Quelque soit la méthode, on peut observer que toutes les variables d’intérêts ont été sélectionnées individuellement avec une probabilité de 100%, mais on peut aussi voir que les autres variables (X6 à X50) ont été plus fréquemment sélectionner par rapport au critère précédent. Du côté de la sélection jointe, remarque que les méthodes de sélection ne retrouve les variables d’intérêts jointes que dans 1,8% des cas en moyenne, ce qui est très faible. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Forward qui présente une probabilité de sélection parfaite de 1,9%, cependant la méthode Forward avec le critère BIC est loin d’être la méthode parfaite.

**AIC**

Forward :

Backward :

Avec le critère de sélection AIC, on peut observer qu’avec toute les méthodes que toutes les variables d’intérêts ont été sélectionnées individuellement, mais on peut aussi voir que les autres variables (X6 à X50) ont fortement été sélectionnées par rapport au critère précédent, on appelle cela de l’overfitting. Ce dernier se défini par le fait qu’une méthode sélectionne beaucoup de variable lorsqu’il doit retrouver le modèles justifiant la valeur de Y, ce qui à tendance à mener à de mauvais résultat. Du côté de la sélection jointe, on remarque que les méthodes de sélection retrouvent les variables d’intérêts jointes dans 0% des cas. Ce résultat s’explique par un taux d’over fitting élevé, ce qui fait que ces modèle de sélection auront tendance à sélectionner plus de variables que prévu, il est donc peu probable que ces procédures de sélection retrouve seulement les bonnes variables avec ce critère de sélection. Alors, aucune procédure ne peut fonctionner avec ce critère de sélection. Avec le critère de sélection AICC, on obtient la même conclusion qu’avec le critère AIC.

Conclu : le meilleur de Stat Learning :

Pour des bases de données du type DGP1, la méthode de procédure de sélection la plus efficace en Statistical learning est la méthode Forward avec le critère de sélection SBC, proposant une probabilité de sélection des variables d’intérêts joints de 16,1%. Backward n’est pas considéré comme efficace car cette méthode a tendance à faire de l’over fitting quelque soit le critère de sélection, donnant ainsi des résultats peu convaincant.

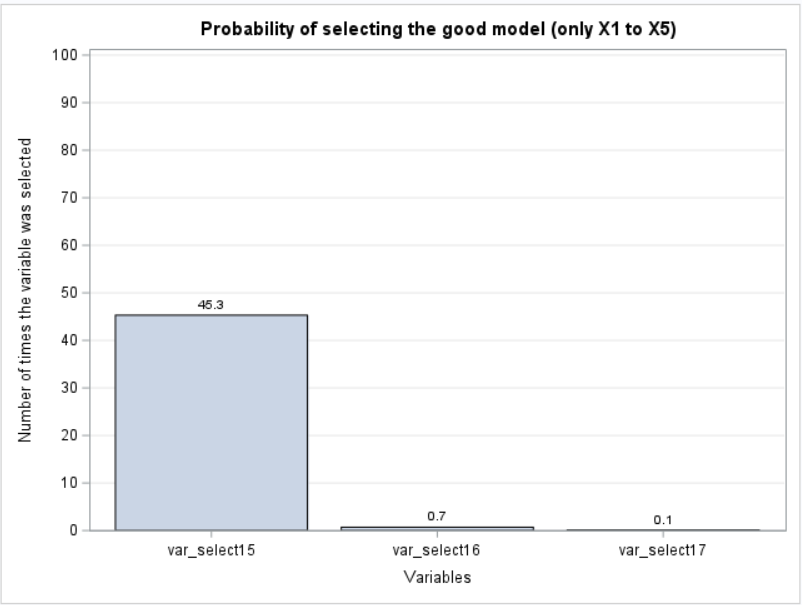
4.1.2) Machin Learning :

Dans cette partie, nous allons comparer les différentes méthodes de Machin learning en fonction des critères de sélection imposés. Dans cette partie, les méthode de sélection testées sont : Lars, Lasso et Elasticnet.

**SBC**

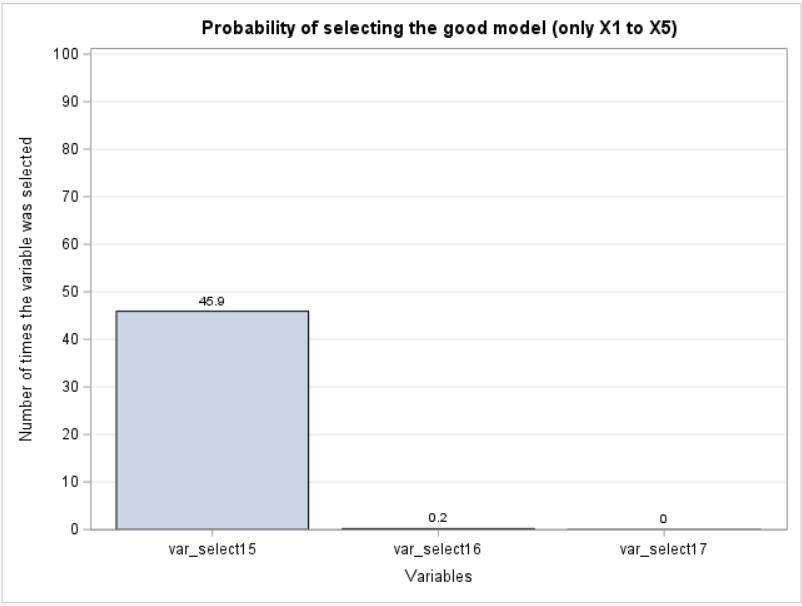
Lars :

Une image contenant table

Description générée automatiquement 

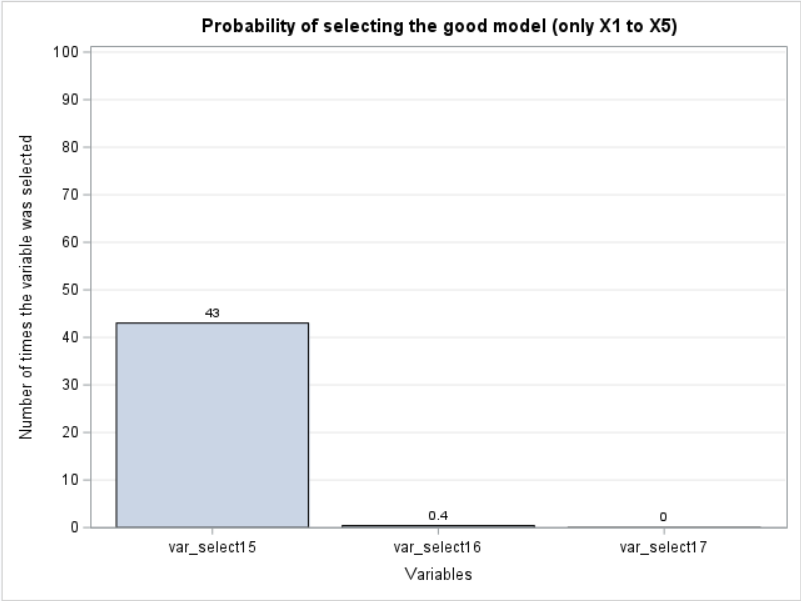
Lasso :

Une image contenant table

Description générée automatiquement 

Elasticnet :

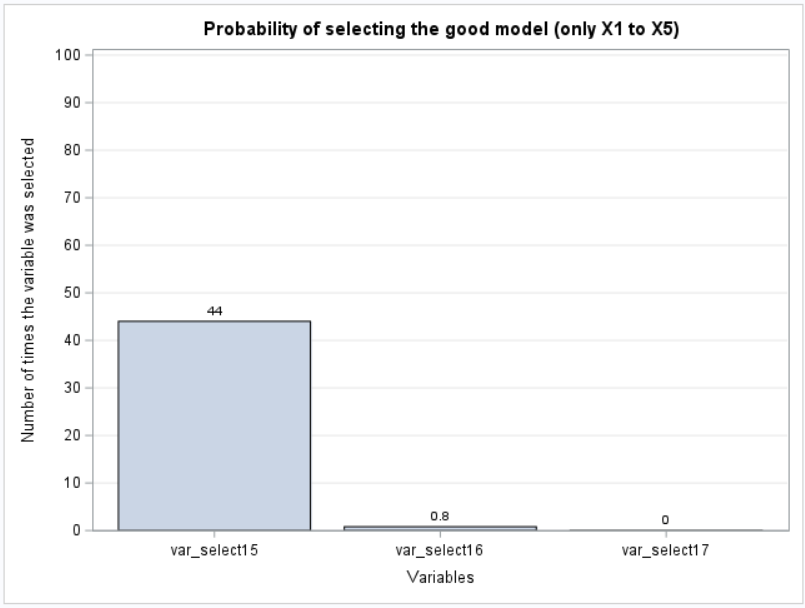
Une image contenant table

Description générée automatiquement 

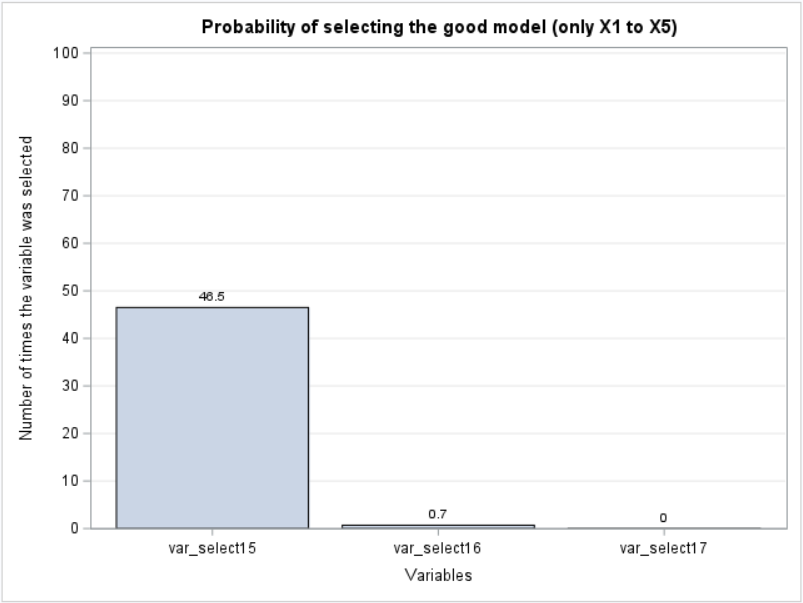
Avec le critère de sélection SBC, on peut observer que toutes les variables d’intérêts ont été sélectionnées individuellement à tous les coups avec les méthodes de LARS et LASSO, ce qui n’est pas le cas pour la méthode Elasticnet qui présente tout de même une probabilité minimum de 90% de sélectionner variables d’intérêts individuelles. Ensuite, on peut aussi voir que les autres variables (X6 à X50) ont une faible probabilité d’être sélectionnées individuellement. Du côté de la sélection jointe, on remarque que les méthodes de sélection retrouve les variables d’intérêts jointes dans 45% des cas en moyenne, ce qui est plus élevé que l’ensemble des procédures de sélection en Statistical learning. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode LASSO qui présente une probabilité de sélection parfaite de 45,9%.

**BIC**

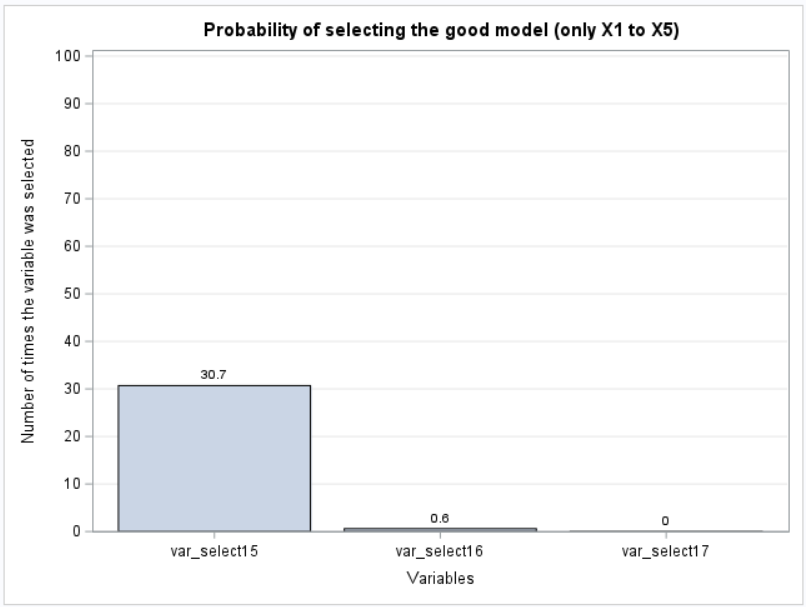
Lars :



Lasso :



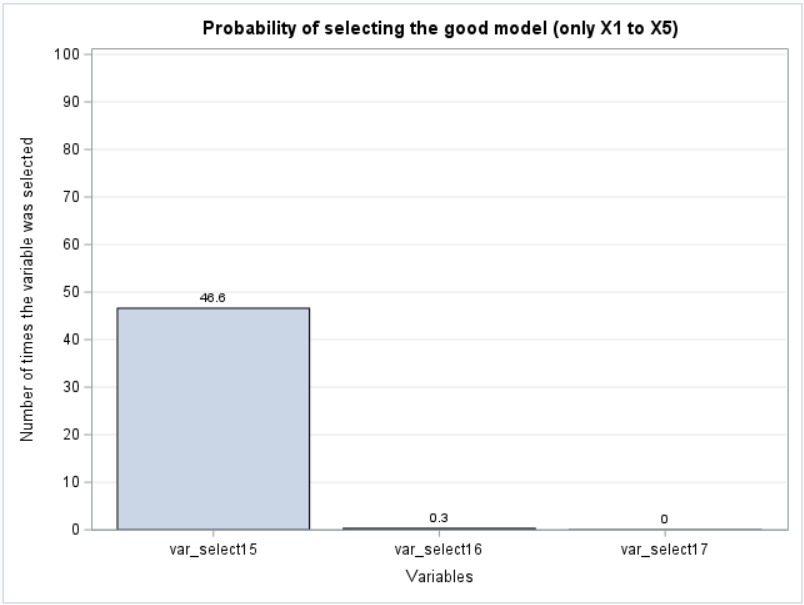
Elasticnet :



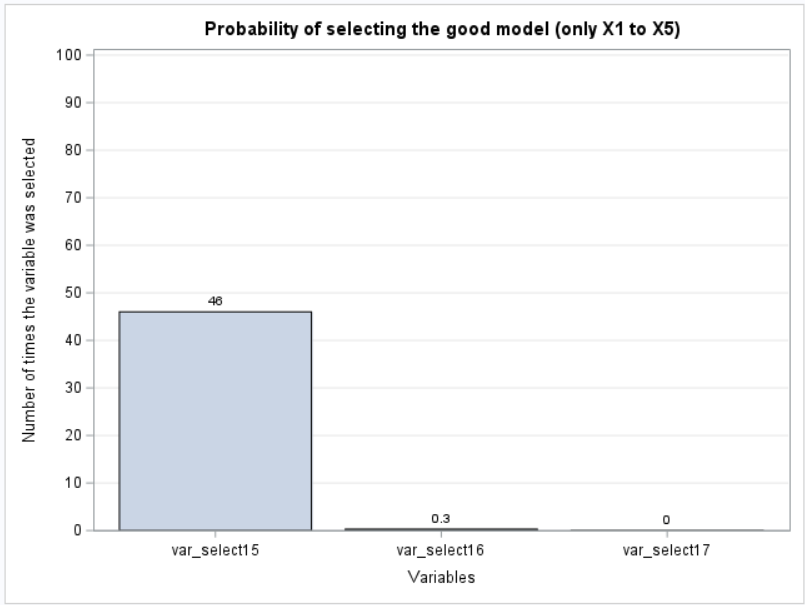
Pour les parties sur les critères BIC AIC et AICC, nous avons choisi de ne présenter aucun graphique de sélection individuel car les graphiques de sélection individuelles de Lars, Lasso et Elasticnet mènent approximativement au même résultat que celle présenté avec le critère SBC. Du côté de la sélection jointe, on remarque que les méthodes de sélection retrouve les variables d’intérêts jointes dans 40% des cas en moyenne, ce qui est un peu plus faible qu’avec le critère de sélection SBC. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode LASSO qui présente une probabilité de sélection parfaite de 46,5%.

**AIC**

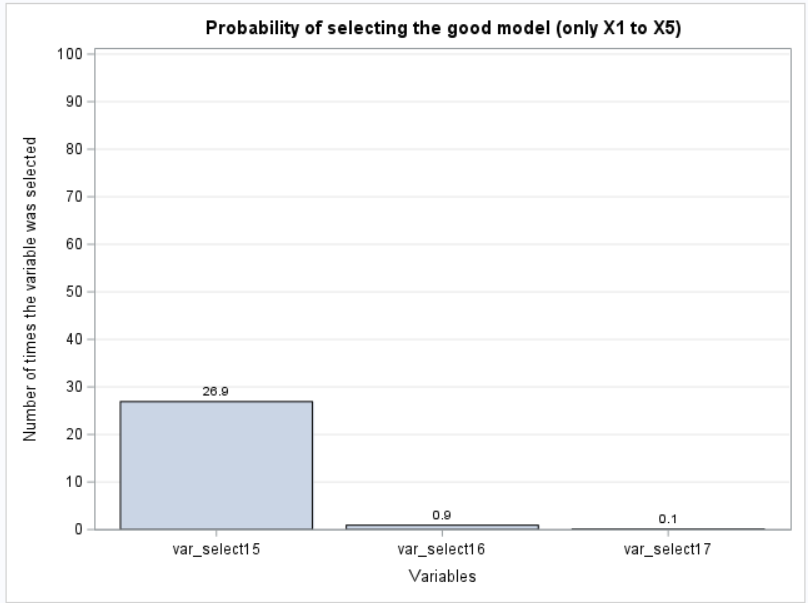
Lars :



Lasso :



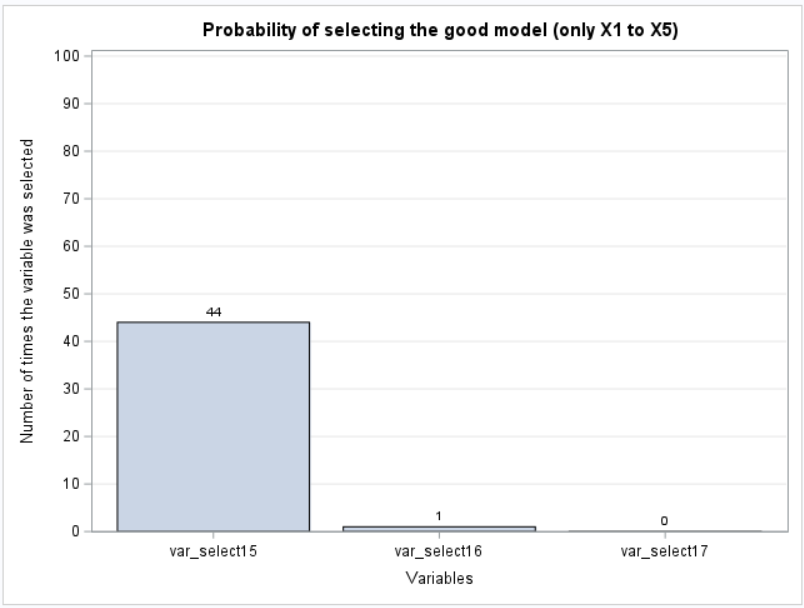
Elasticnet :



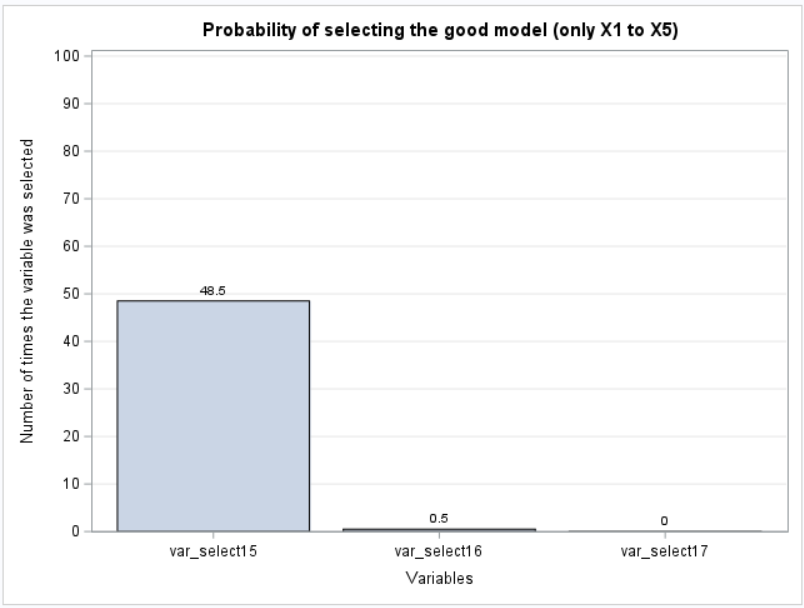
Avec le critère de sélection AIC, on remarque que les méthodes de sélection retrouvent les variables d’intérêts jointes dans 36% des cas en moyenne, ce qui est plus faible que les deux critères de sélections précédent. Cependant on remarque que l’Elasticnet avec AIC a une probabilité de retrouver les variables d’intérêts jointes de 27%, ce qui est plus faible par rapport à l’Elasticnet avec BIC et SBC. Tandis que les procédures Lars et Lasso ont la même probabilités de retrouver les variables d’intérêts avec les critères SBC, AIC et AIC (qui tourne généralement autour 44% à 46,6% pour Lars et autour des 45%-46% pour Lasso). La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode LARS qui présente une probabilité de sélection parfaite de 46,5%.

**AICC**

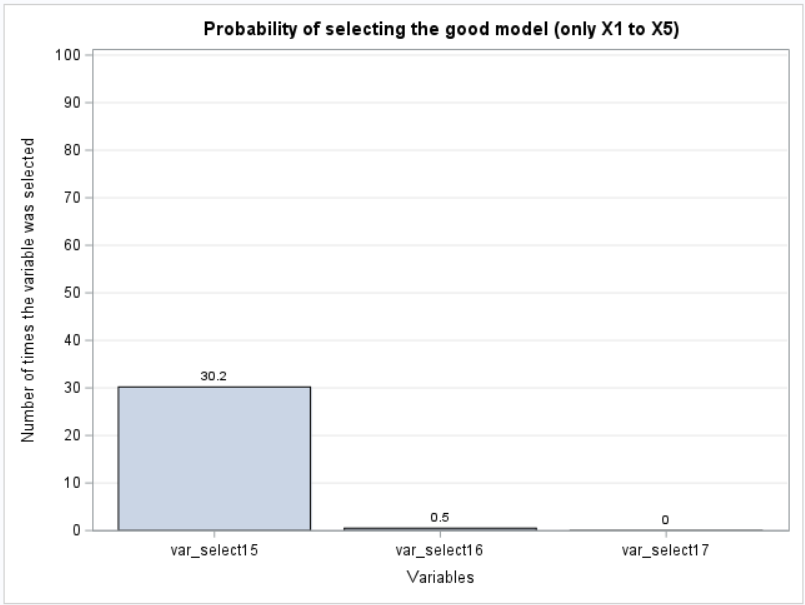
Lars :



Lasso :



Elasticnet :



Avec le critère de sélection AICC, on remarque que les méthodes de sélection retrouvent les variables d’intérêts jointes dans 42% des cas en moyenne, ce qui correspond à peu près aux mêmes probabilité qu’avec critères de sélections SBC et BIC. L’Elasticnet avec AICC a une probabilité de retrouver les variables d’intérêts jointes de 30%. La procédure Lasso a une probabilités de 48,5% et Lars une probabilité 44%. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode LASSO qui présente une probabilité de sélection parfaite de 48,5%.

Conclu : le meilleur de Machin Learning :

La procédure de sélection la plus efficace en Machin learning dans ce cadre de générateur de donnée(DGP1) est la méthode Lasso avec le critère de sélection AICC, proposant une probabilité de sélection des variables d’intérêts joints de 48,5%. On peut aussi dire que Lars peut être un bon outil de sélection au vu de ses probabilités de sélection très proche de celle de Lasso. Elasticnet présente des probabilités de sélection supérieur aux procédures de Statistical learning, mais il n’est pas le meilleur outils de sélection dans ce contexte là.

Comparer meilleur de Stat et meilleur de Machin, et dire quel est la meilleure méthode pour DGP1 :

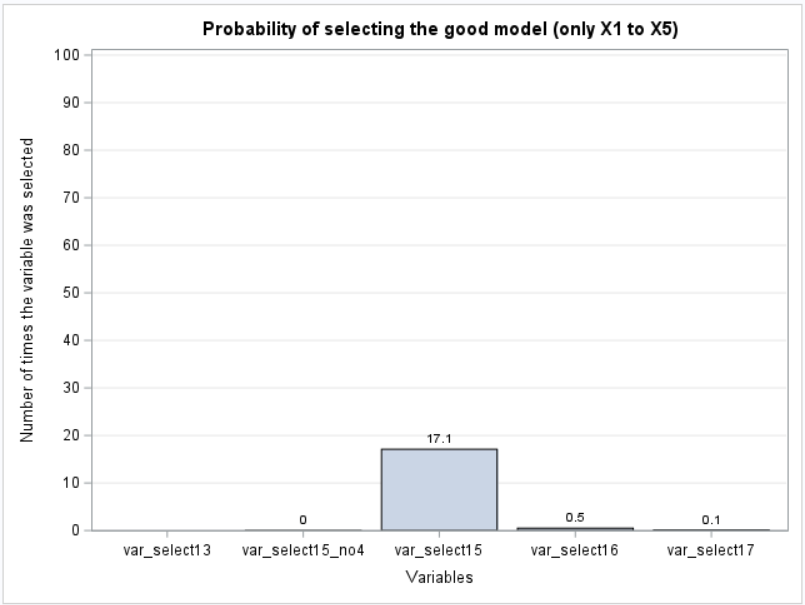
Si on compare la meilleure procédure en Statistical learning (Forward) et la meilleure en Machin learning (Lasso) , on s’aperçoit de manière évidente que la procédure Lasso avec le critère de sélection AICC (48,5% de réussite) est plus performant que la procédure Forward avec le critère de sélection SBC (16,5% de réussite). Lasso avec le critère de sélection AICC est donc la méthode la plus performante pour retrouver les variables d’intérêt dans une base de donnée du type DGP1.

4.2) DGP2

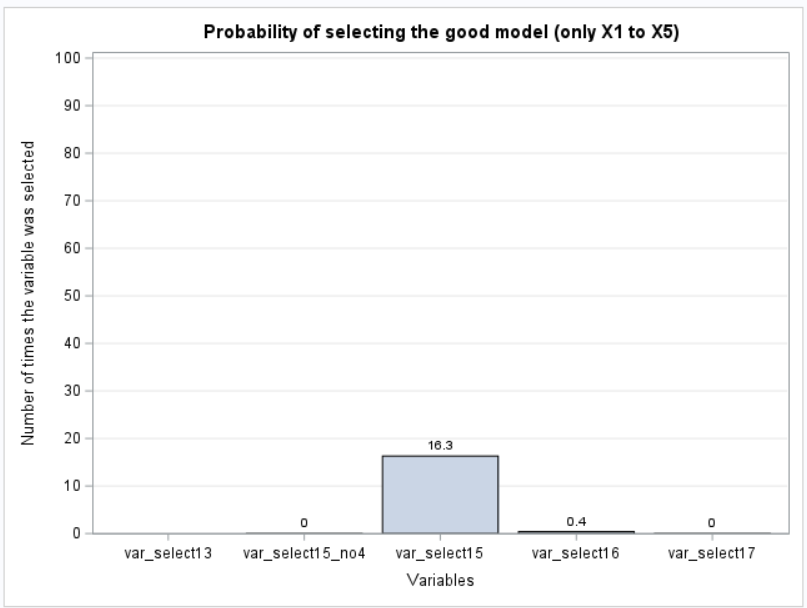
4.2.1 Statistical learning

**SBC**

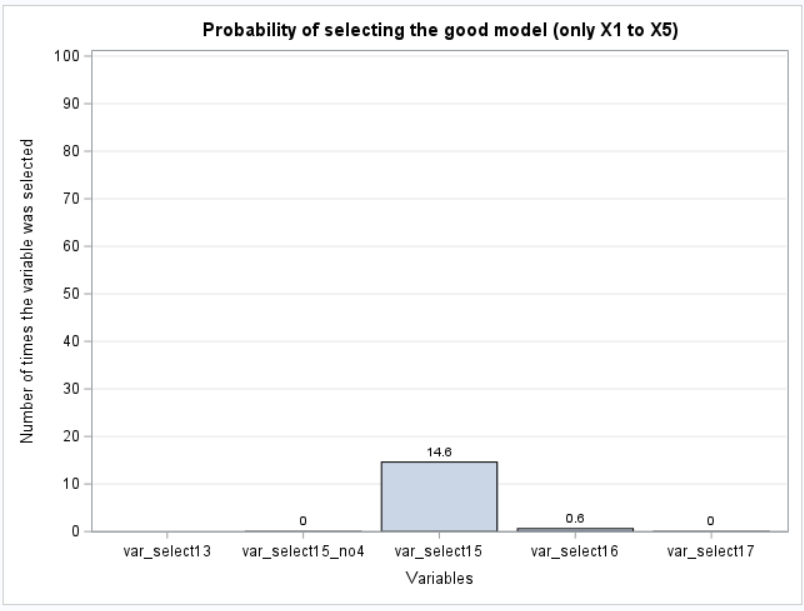
stepwise :



forward :



backward:



Avec le critère de sélection SBC, on peut observer que toutes les variables d’intérêts ont été sélectionnées individuellement à tous les coups. Cependant, les méthodes de sélection ne retrouve les variables d’intérêts jointes que dans 16% des cas en moyenne, il y a peu de différence par rapport au probabilité de sélection moyen effectué avec le DGP1 (=15%) . La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Stepwise qui présente une probabilité de sélection parfaite de 17,1%. Sachant que Stepwise et Forward ont des méthodes de sélection proche, on peut aussi déduire que Forward peut aussi bien fonctionner avec ce critère de sélection.

**BIC**

forward :

Une image contenant table

Description générée automatiquement

backward:

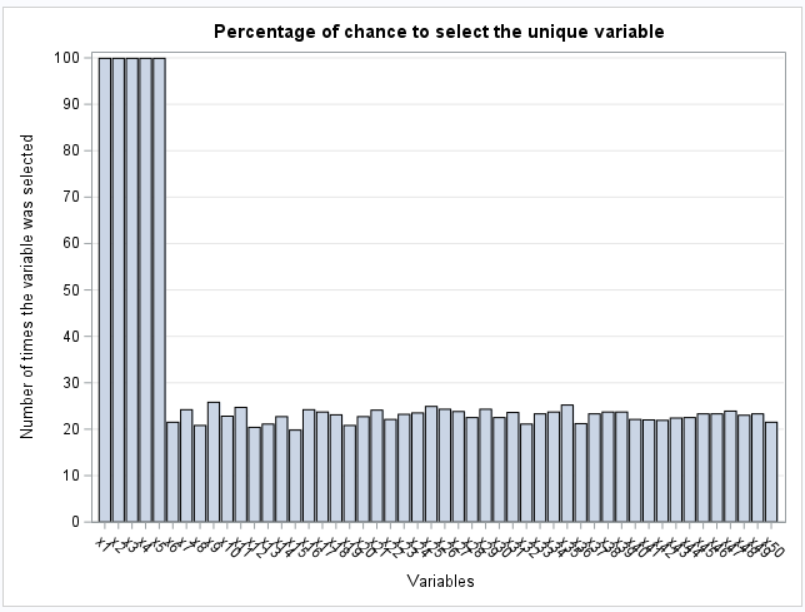
Une image contenant table

Description générée automatiquement

Avec le critère de sélection BIC, nous avons choisi de présenter un seul graphique de sélection individuel car tous ces graphiques mènent au même résultat. Quelque soit la méthode, on peut observer que toutes les variables d’intérêts ont été sélectionnées individuellement avec une probabilité de 100%. Les autres variables (X6 à X50) ont été plus fréquemment sélectionner par rapport au critère précédent. Du côté de la sélection jointe, remarque que les méthodes de sélection ne retrouve les variables d’intérêts jointes que dans 1,35% des cas en moyenne, ce qui est un peu plus faible par rapport au probabilité de sélection moyen effectué avec le DGP1 (=1,5%), la différence est peu marquante. On peut tout de même proposer la procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Forward, présentant une probabilité de sélection parfaite de 1,5%. Au vu de la faible probabilité de sélection parfaite, on peut conclure qu’ aucune procédure ne peut fonctionner avec ce critère de sélection.

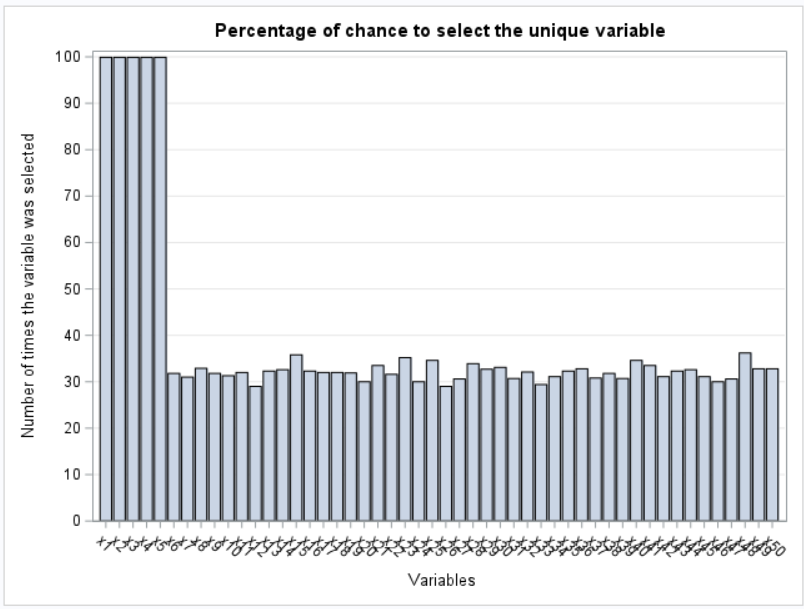
**AIC**

stepwise :

Une image contenant table

Description générée automatiquement

backward:

 Une image contenant table

Description générée automatiquement

Avec le critère de sélection AIC, nous avons choisi de ne pas présenter le graphique de sélection individuel de la méthode Forward car ce graphique mène au même résultat que celle de Stepwise. Quelque soit la méthode, on peut observer que toutes les variables d’intérêts ont été sélectionnées individuellement dans 100% des cas, mais on peut aussi voir que les autres variables (X6 à X50) sont fortement sélectionnées par rapport aux analyses précédent, on fait face à un taux d’overfitting élevé. Du côté de la sélection jointe, on remarque que les méthodes de sélection retrouvent les variables d’intérêts jointes dans 0% des cas. Ce résultat s’explique par un taux d’over fitting élevé, ce qui fait que ces modèle de sélection auront tendance à sélectionner plus de variables que prévu, il est donc peu probable que ces procédures de sélection retrouve seulement les bonnes variables avec ce critère de sélection. Aucune procédure ne peut fonctionner avec ce critère de sélection.

Avec le critère de sélection AICC, on obtient la même conclusion qu’avec le critère de sélection précédents. Il est donc peu probable que ces procédures de sélection retrouve seulement les bonnes variables avec ce critère de sélection. Aucune procédure ne peut fonctionner avec ce critère de sélection.

Conclu : le meilleur de Stat Learning -> Forward

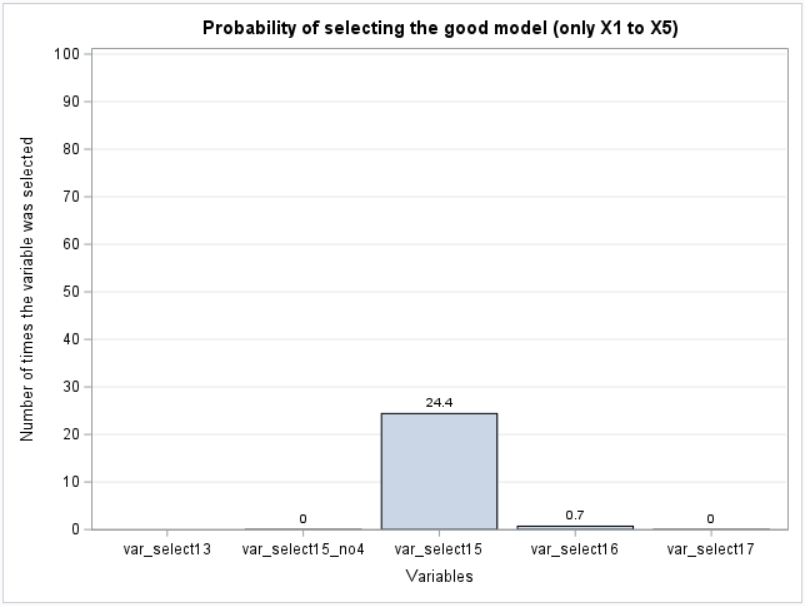
A travers l’analyse des résultats obtenus, on sait que SBC est le seul critère de sélection pertinent pour les procédures de Statistical learning. Cependant il a été remarquer que les résultats obtenus sont très proche voire identique à celle obtenue dans l’analyse des résultats des méthodes de sélections sur le DGP1. Cela nous montre donc que les corrélation positives entre les variables X1 à X5 effectuées dans le DGP2 ont peu d’impact sur la probabilité de sélection des variables d’intérêts des procédures de Statistical learning. Ainsi, pour des bases de données du type DGP2, la méthode de procédure de sélection la plus efficace en Statistical learning est les méthodes Stepwise et Forward avec le critère de sélection SBC, proposant une probabilité de sélection des variables d’intérêts joints de 17,1% et 16,3%. Backward n’est pas considéré comme efficace car cette méthode a tendance à faire de l’over fitting quelque soit le critère de sélection, donnant ainsi des résultats peu convaincant.

4.2.2) Machin Learning :

**SBC**

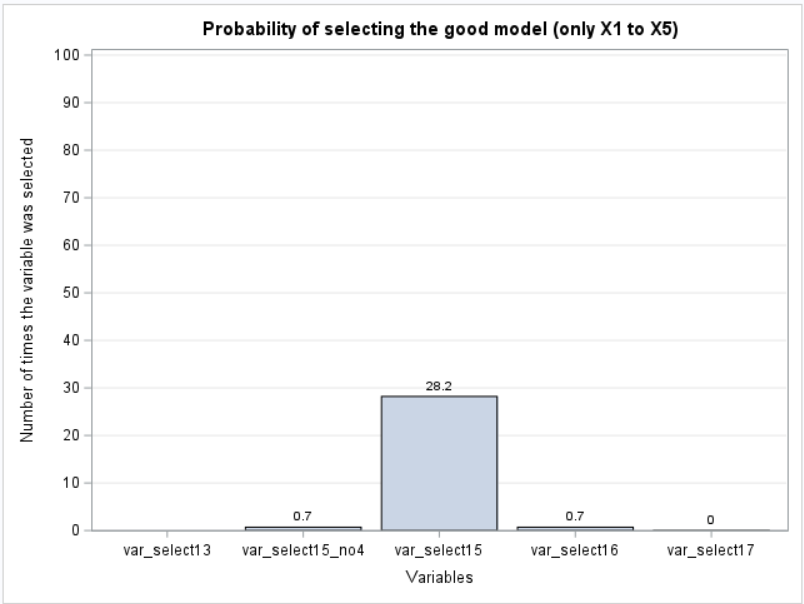
Lars :

Une image contenant table

Description générée automatiquement

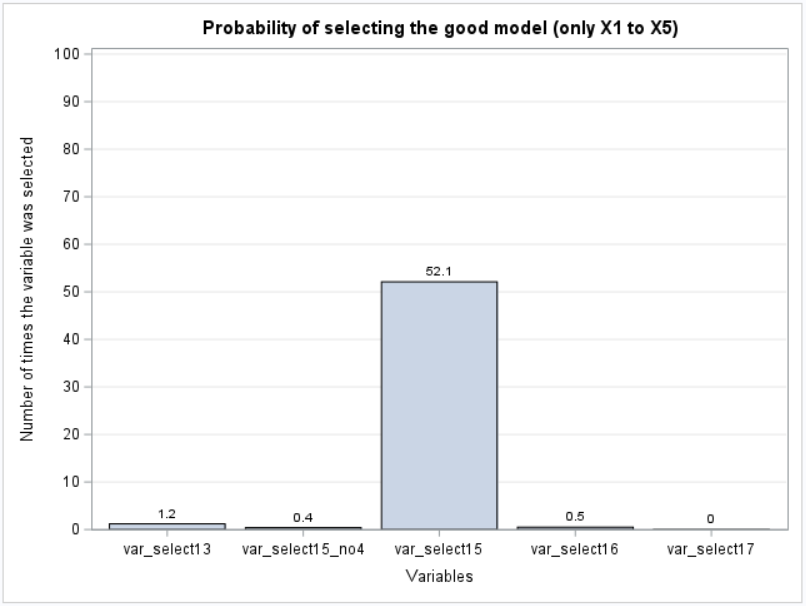
Lasso :

Une image contenant table

Description générée automatiquement

Elasticnet :

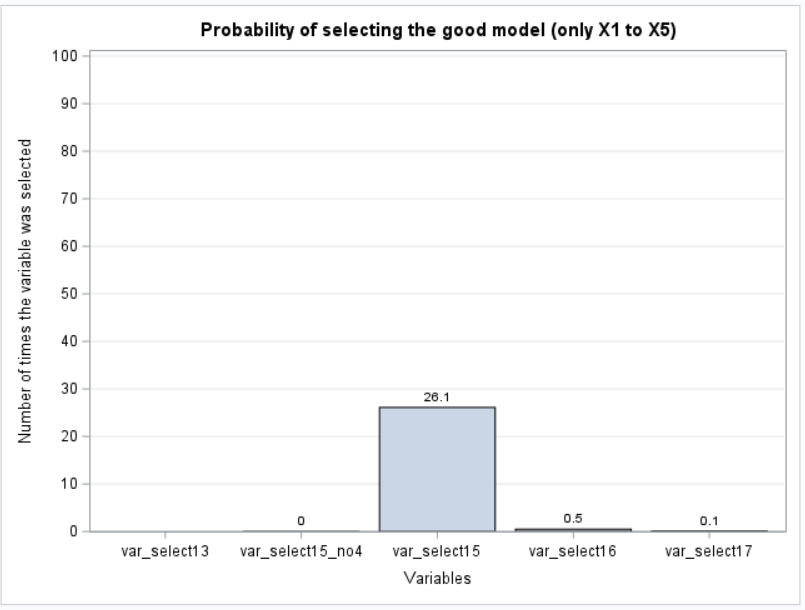
Une image contenant table

Description générée automatiquement

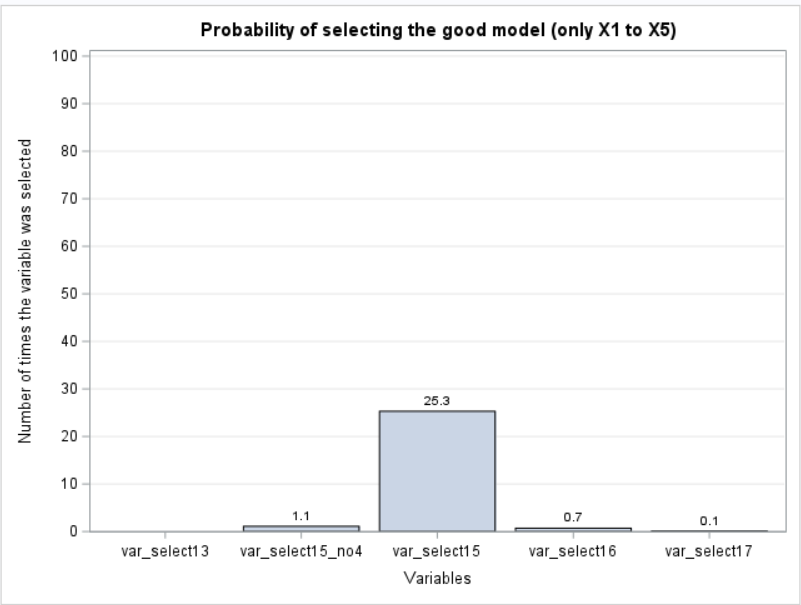
Avec le critère de sélection SBC, on peut observer que seulement les 4 premières variables d’intérêts ont été sélectionnées individuellement à tous les coups avec les méthodes de LARS et LASSO. On remarque que la 5e variables d’intérêts (X5) à été sélectionné seulement dans 72% des cas avec Lars et Lasso. Premièrement, cela s’explique tout simplement par la forte corrélation entre X5 et X1 (programmé dans le DGP2). Deuxièmement lorsque Lars ou Lasso doivent sélectionner des variables, elles vont choisir les variables qui sont les moins corrélées entre-elles, mais aussi les plus significatives. Ainsi, lorsqu’elles tombent sur 2 variables fortement corrélées entre-elles, elles doivent faire le choix entre une des deux variables car selon elles une des deux variables n’est pas la bonne variable (variable d’intérêt). Dans ce cadre là, elles devaient faire le choix entre X1 et X5, ce qui explique donc pourquoi X5 n’est pas toujours sélectionnée. En ce qui concerne la méthode Elasticnet les 5 variables d’intérêts sont sélectionnées individuellement dans 90% des cas. Quelque soit la procédure, on peut voir que les autres variables (X6 à X50) ont une faible probabilité d’être sélectionnées individuellement. En observant les graphiques de droite, on remarque que la procédure Elasticnet est la plus performante (avec une probabilité de sélection jointe égal à 52,1%) parmi les 3 procédures testées. Cela peut se justifier par le fait que Elasticnet avec le critère SBC est moins sensible aux corrélations, il aura donc plus tendance à sélectionner à la fois les variables X1 et X5, contrairement à Lars et Lasso. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Elasticnet qui présente une probabilité de sélection parfaite de 52,1%.

**BIC**

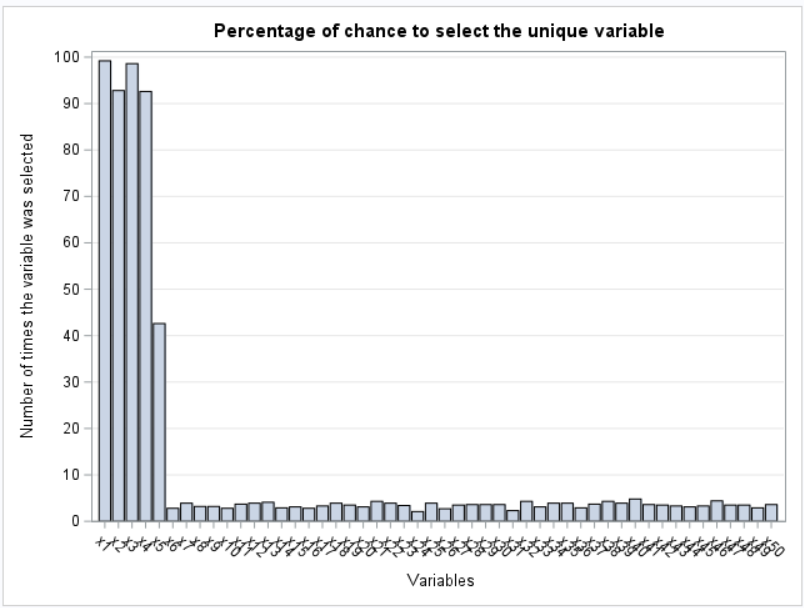
Lars :



Lasso :



Elasticnet :

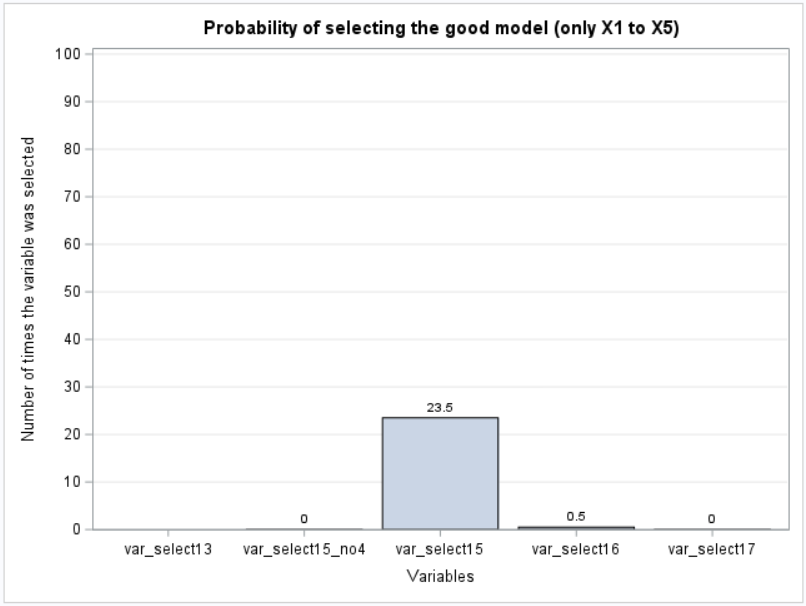
Une image contenant table

Description générée automatiquement

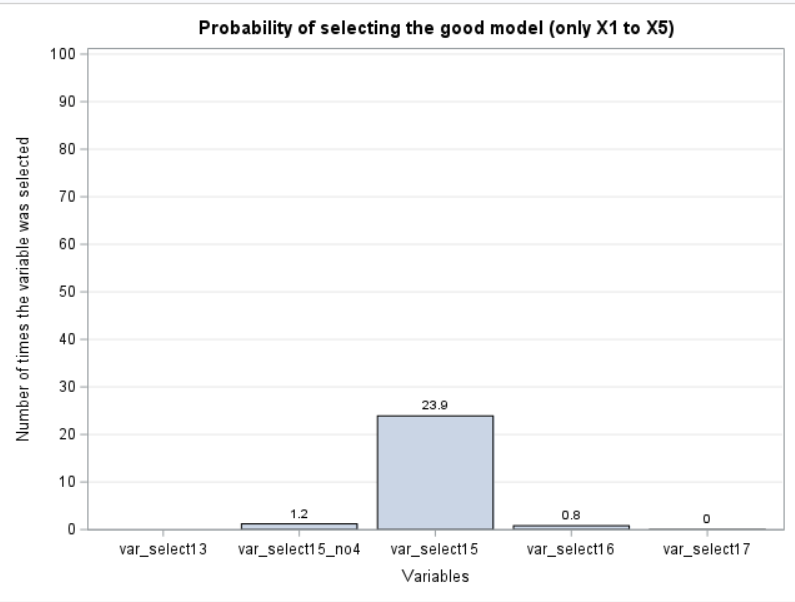
Pour les parties sur les critères BIC AIC et AICC, nous avons choisi de présenter seulement le graphique de sélection individuel de l’Elasticnet car les graphiques de sélection individuelles de Lars, Lasso mènent approximativement au même résultat que celle présenté avec le critère SBC. Avec le critère de sélection BIC, la méthode de l’Elasticnet (avec le critère BIC) sélectionne les 4 premières variables d’intérêts avec une probabilité minimum de 90%, mais on remarque que la 5e variables d’intérêts (X5) à été sélectionné seulement dans 42% des cas. Ensuite, on remarque que la procédure Lars est la plus performante (avec une probabilité de sélection jointe égal à 26,1%) parmi les 3 procédures testées. Contrairement à l’analyse précédente, l’Elasticnet (avec BIC) est de loin la moins performante des 3 procédures. Cela peut se justifier par le fait que Elasticnet est beaucoup plus sensible aux corrélations avec le critère BIC, il aura donc plus tendance à sélectionner X1 ou X5, mais rarement les 2 à la fois. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Lars qui présente une probabilité de sélection unique des 5 variables d’intérêt de 26,1%.

**AIC**

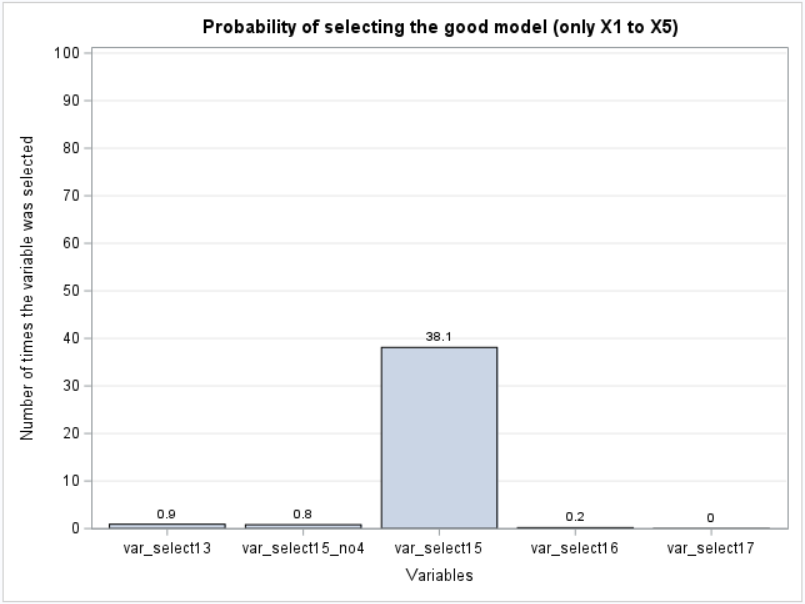
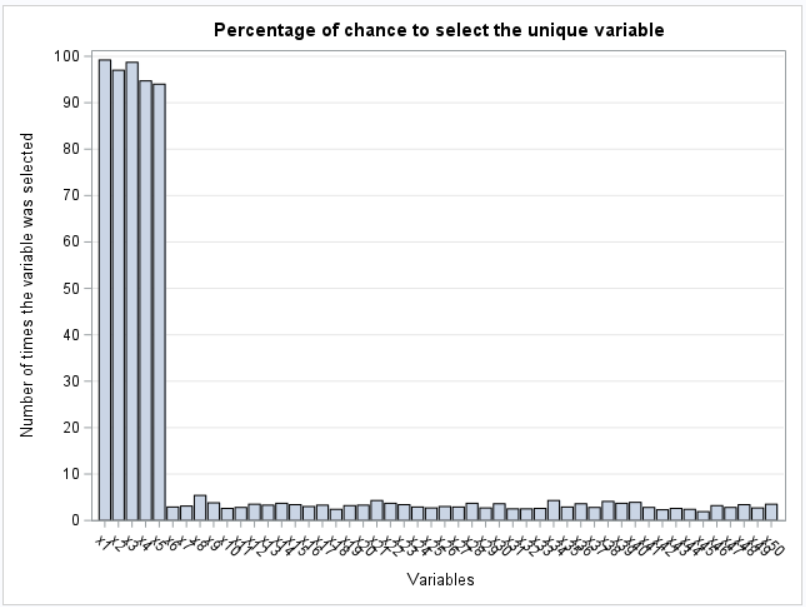
Lars :



Lasso :



Elasticnet :



Avec le critère de sélection AIC ou AICC, on peut observer que toutes les variables d’intérêts sont sélectionnées individuellement dans plus de 90% des cas avec la méthode de l’Elasticnet. En observant les graphiques de sélection jointes de l’Elasticnet avec AIC et de l’Elasticnet avec AICC, on remarque que la procédure Elasticnet est la plus performante (avec une probabilité de sélection jointe d’environ 40%) avec les 2 critères. Cela peut se justifier par le fait que Elasticnet avec le critère AIC (ou AICC) est moins sensible aux corrélations, il aura donc plus tendance à sélectionner à la fois les variables X1 et X5, contrairement à Lars et Lasso. La procédure qui fonctionnerait le mieux avec ces 2 critères de sélection est la méthode Elasticnet qui présente une probabilité de sélection parfaite d’environ 40%.

Conclu : le meilleur de Machin Learning

La procédure de sélection la plus efficace en Machin learning dans ce cadre de générateur de donnée(DGP2) est la méthode Elasticnet avec le critère de sélection SBC, proposant une probabilité de sélection des variables d’intérêts joints de 52,1%. Lars et Lasso présentent des probabilités de sélection supérieur aux procédures de Statistical learning, mais ce ne sont pas les meilleurs outils de sélection dans ce contexte là.

Comparer meilleur de Stat et meilleur de Machin, et dire quel est la meilleure méthode pour DGP2 :

Si on compare la meilleure procédure en Statistical learning (Stepwise) et la meilleure en Machin learning (Elasticnet) , on s’aperçoit de manière évidente que la procédure Elasticnet avec le critère de sélection SBC(52,1% de réussite) est plus performant que la procédure Stepwise avec le critère de sélection SBC (17,1% de réussite). Elasticnet avec le critère de sélection SBC est donc la méthode la plus performante pour retrouver les variables d’intérêt dans une base de donnée du type DGP2.

4.3) DGP3 :

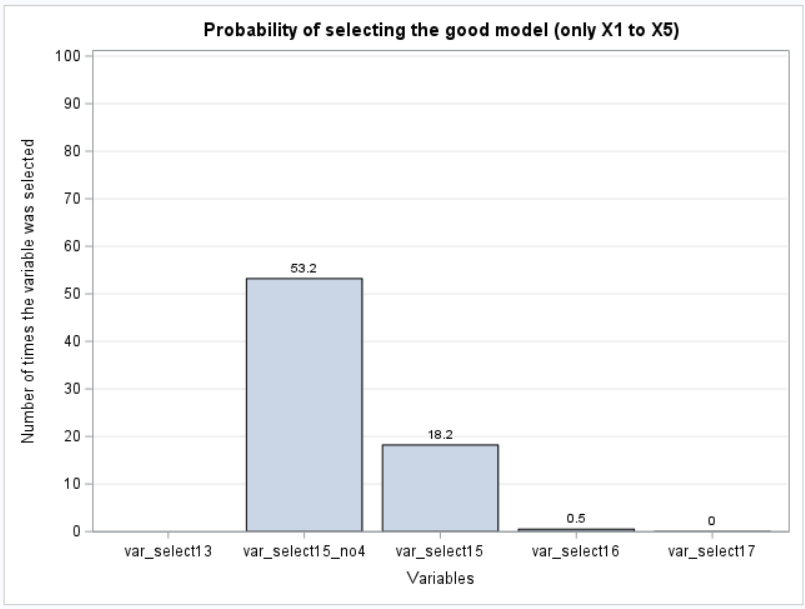
Quelque soit la méthode, on peut remarquer que toutes les variables d’intérêts X1 X2 X3 X5 ont été sélectionnées individuellement à tous les coups. Le fait que la variable d’intérêt X4 n’est pas choisi dans tout les cas s’explique par la présence de corrélation entre les variables, mais aussi par le fait que son coefficient multiplicateur (=0,1) est une valeur proche de 0. Ce qui fait que les méthodes de sélections vont avoir moins tendance à sélectionner cette variable car ils considèrent le coefficient multiplicateur de X4 comme peu significative pour justifier la valeur de Y. Le fait que la variable x4 soit faiblement choisi par les modèles a des conséquences sur l'évaluation des compétences des méthodes de sélection qui seront étudié dès à présent.

Stat Learning :

**SBC**

stepwise :

Une image contenant table

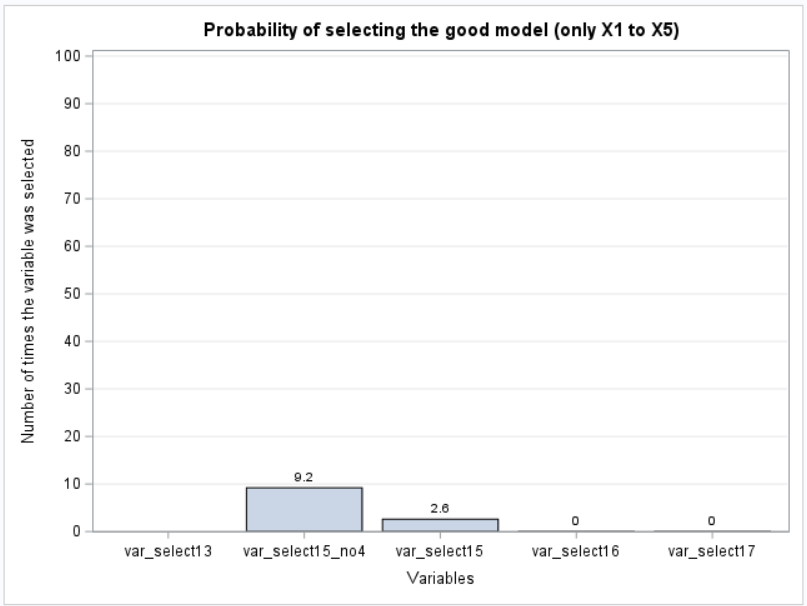
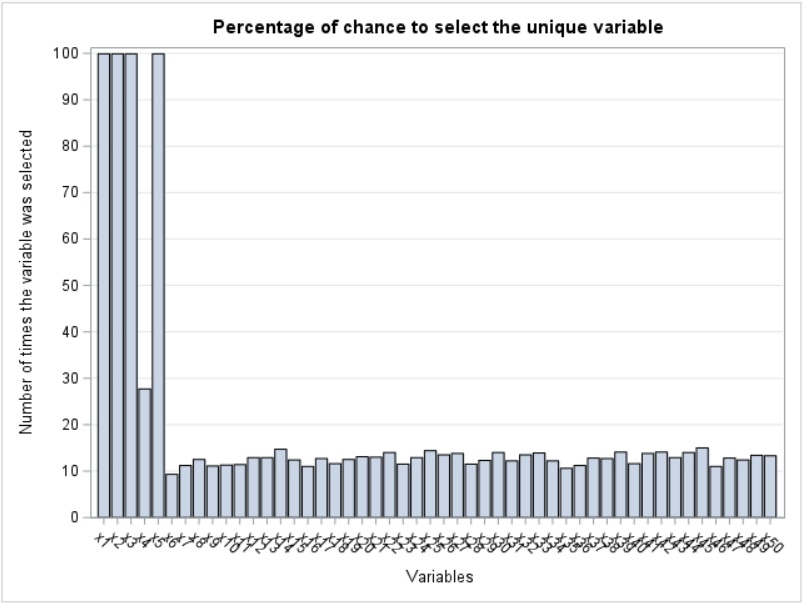
Description générée automatiquement 

forward :

Une image contenant table

Description générée automatiquement 

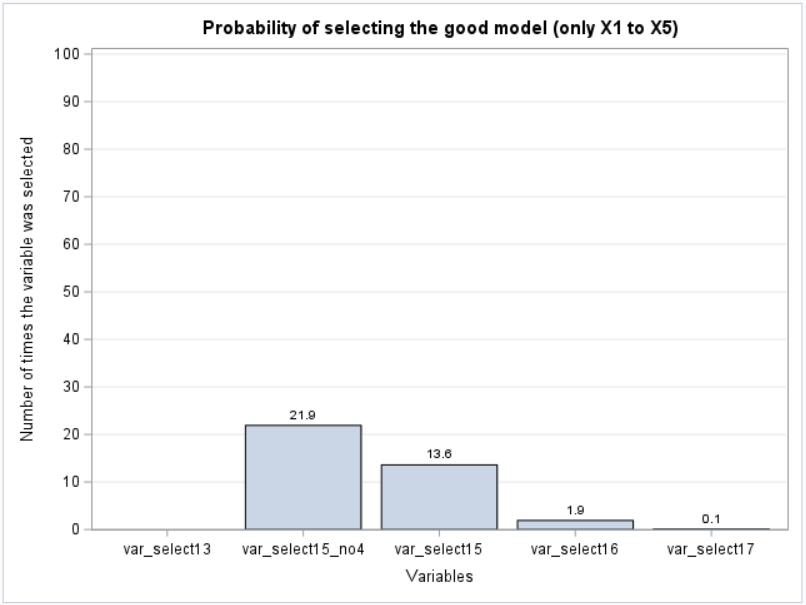
backward:



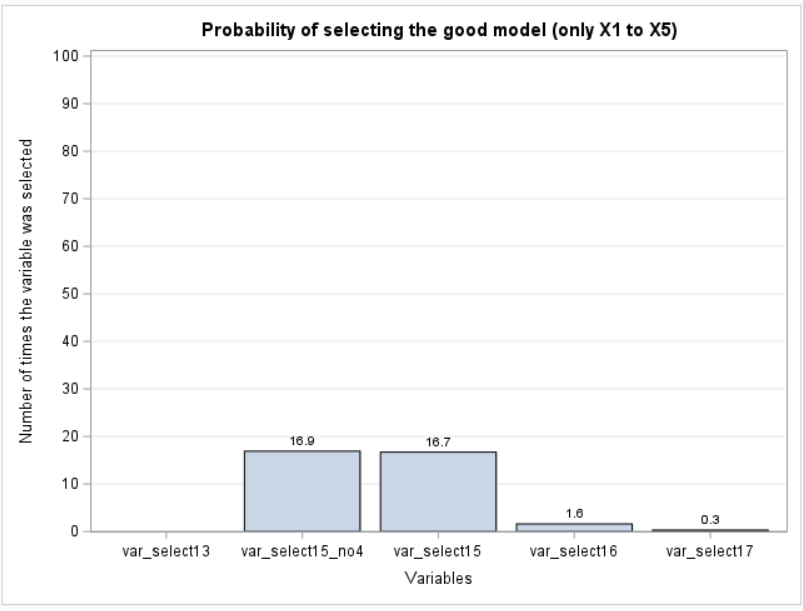
Avec le critère de sélection SBC, la méthodes Stepwise et Forward retrouve les variables d’intérêts jointes avec une probabilité allant de 16,6% à 18,2%, ce qui est environ 5 fois supérieurs par rapport à la méthode Backward. On peut remarquer que les méthodes Stepwise et Forward sélectionnent les variables d’intérêts sans le X4 dans plus de 50% des cas, ce qui nous permet de conclure que ces 2 méthodes rejette directement la variables si sont coefficient multiplicateur (Béta de X4) n’est pas assez significatif pour justifier Y (c’est à dire proche de 0). La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Stepwise qui présente une probabilité de sélection parfaite de 18,2%. Sachant que Stepwise et Forward sont des méthodes assez proche, on peut aussi déduire que Forward peut aussi bien fonctionner avec ce critère de sélection.

**BIC**

stepwise :



forward :



backward :

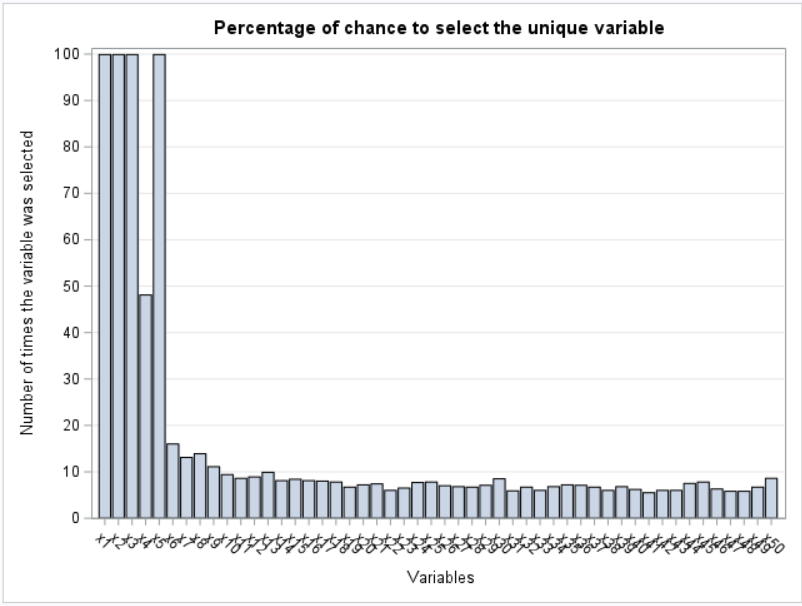
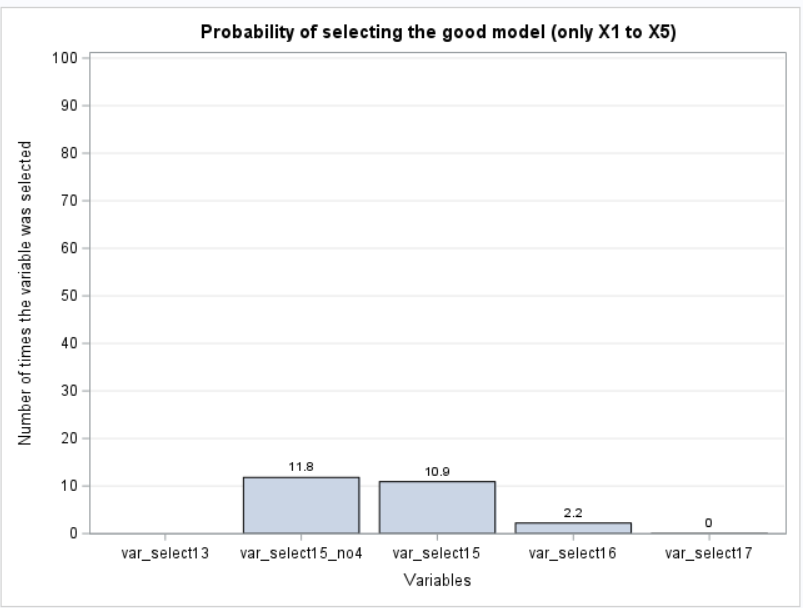
Une image contenant table

Description générée automatiquement

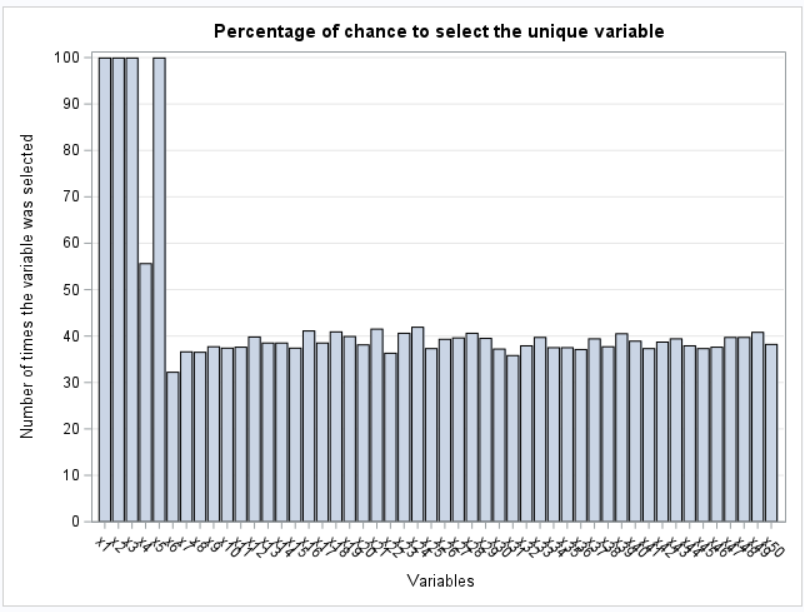
Avec le critère de sélection BIC, la méthodes Stepwise et Forward retrouve les variables d’intérêts jointes avec une probabilité entre 13% et 16,7%, tandis que Backward ne les retrouve dans seulement 0,4% des cas dû à l’over fitting. Le critère de sélection BIC semble être moins efficace que le critère SBC pour les modèle Stepwise et Backward. Cependant, la procédure Forward affiche 2 probabilités de sélection identiques avec les critères SBC et BIC. La procédure qui fonctionnerait le mieux avec ce critère de sélection est la méthode Forward qui présente une probabilité de sélection parfaite de 16,7%.

**AIC**

stepwise :

backward:

 Une image contenant table

Description générée automatiquement

Avec le critère de sélection AIC et AICC, les méthodes Stepwise et Forward retrouvent les variables d’intérêts jointes avec une probabilité allant de 10,4% à 10,9% avec le critère AIC et une probabilité allant de 12,7% et 14,4%. Quelque soit le critère de sélection(AIC et AICC), Backward ne retrouve aucun groupe de variables d’intérêts jointes dû à l’overfitting. Les critères de sélection AIC et AICC semblent donc être moins efficace que le critère SBC pour les méthodes de Statisitcal learning. La procédure qui fonctionnerait le mieux avec ces critères de sélections (AIC et AICC) est la méthode Stepwise qui présente une probabilité de sélection parfaite de 10,9% avec AIC et 14,4% avec AICC. Sachant que Stepwise et Forward affichent des résultats assez proche, on peut aussi déduire que Forward peut aussi bien fonctionner avec ce critère de sélection.

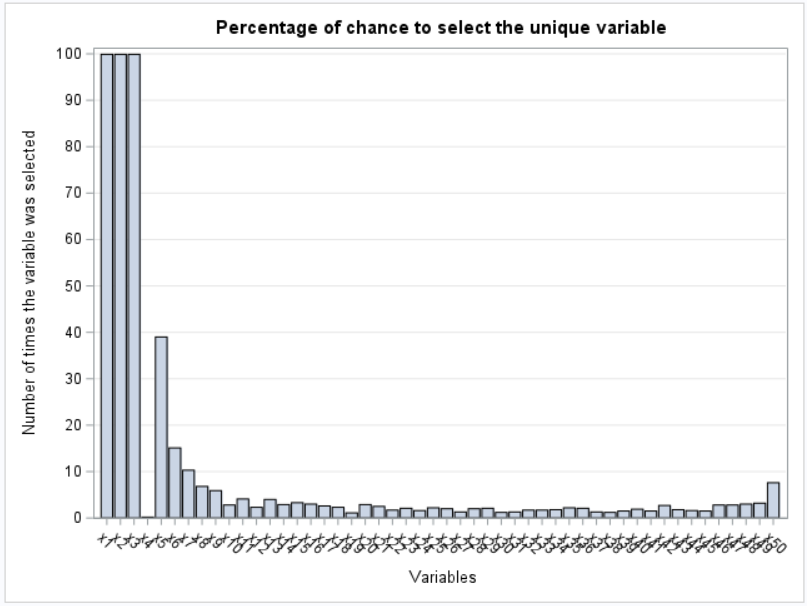
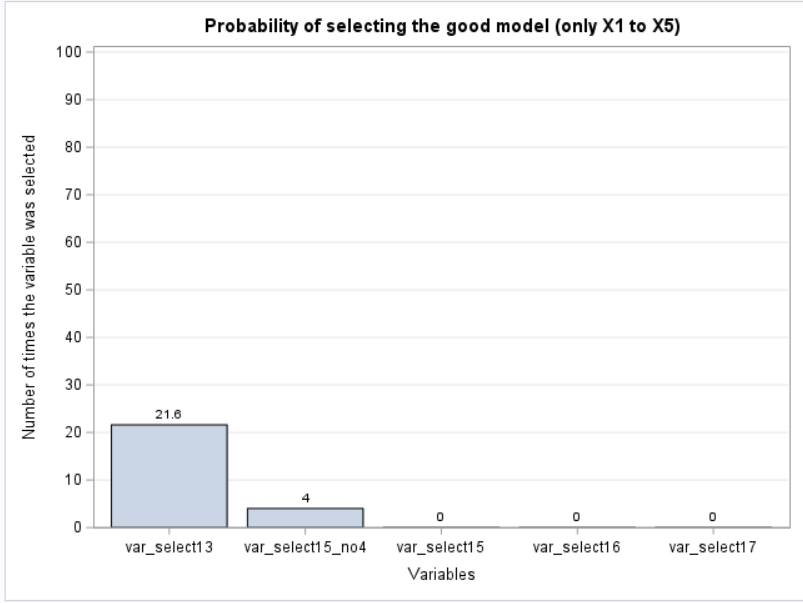
Conclu : le meilleur de Stat Learning -> Stepwise (sbc)

Pour des bases de données du type DGP3, la méthode de procédure de sélection la plus efficace en Statistical learning est la méthode Stepwise avec le critère de sélection SBC, proposant une probabilité de sélection des variables d’intérêts joints de 18,2%. Cependant, Stepwise et Forward affichent des résultats assez proche, on peut donc déduire que Forward peut aussi être une méthode efficace avec le critère SBC. Backward n’est pas considéré comme efficace car cette méthode a tendance à faire de l’over fitting quelque soit le critère de sélection, donnant ainsi des résultats peu convaincant.

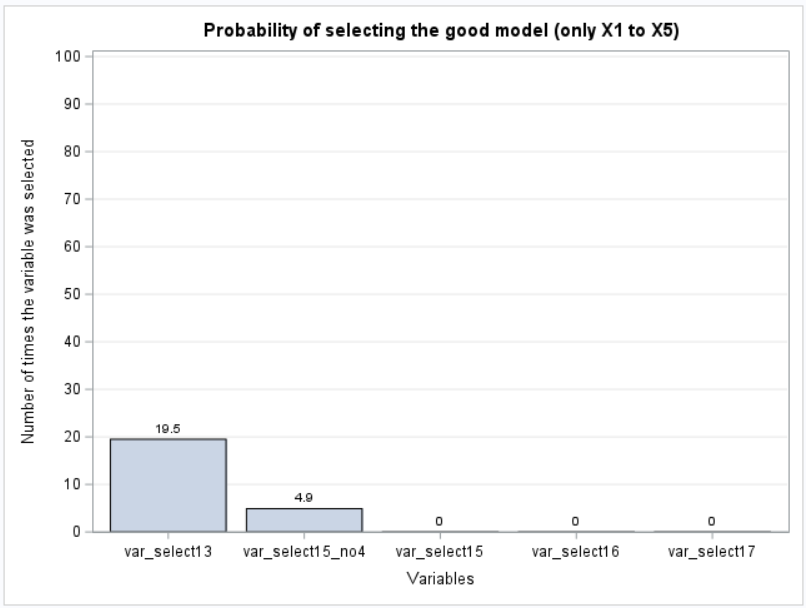
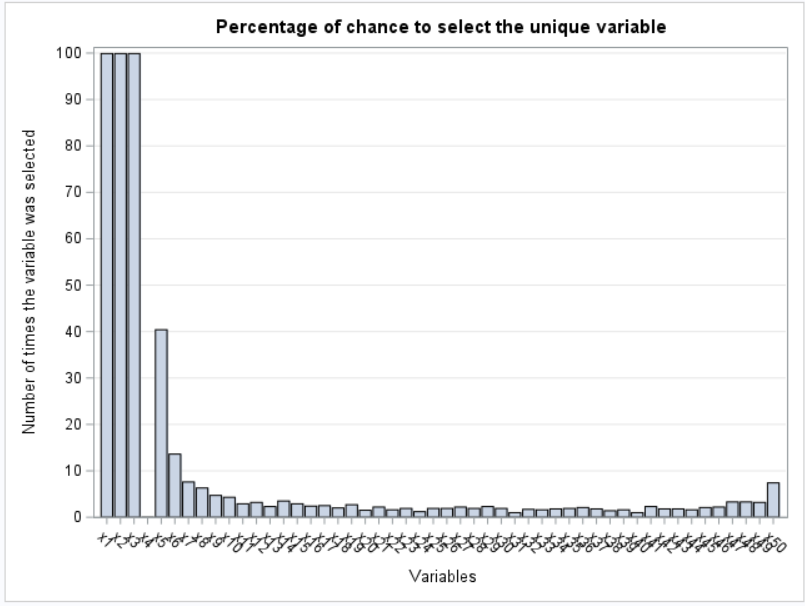
Machin Learning :

**SBC**

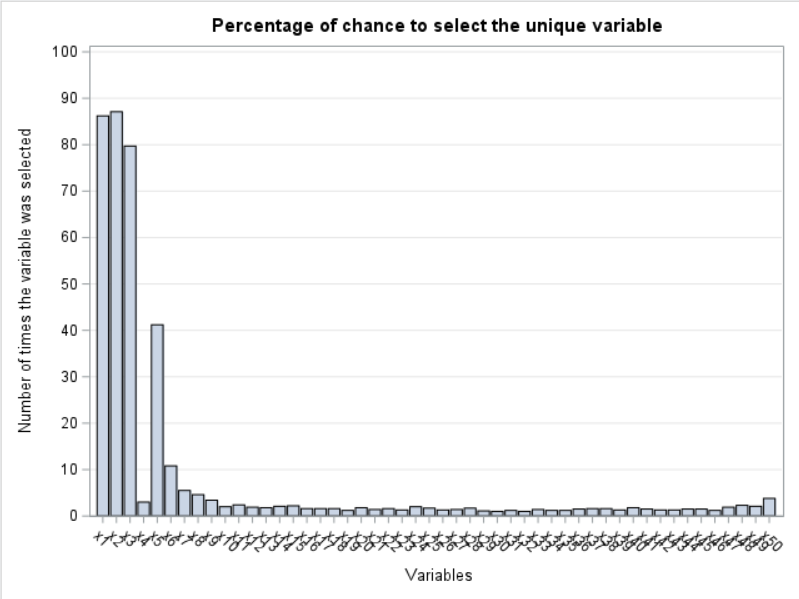
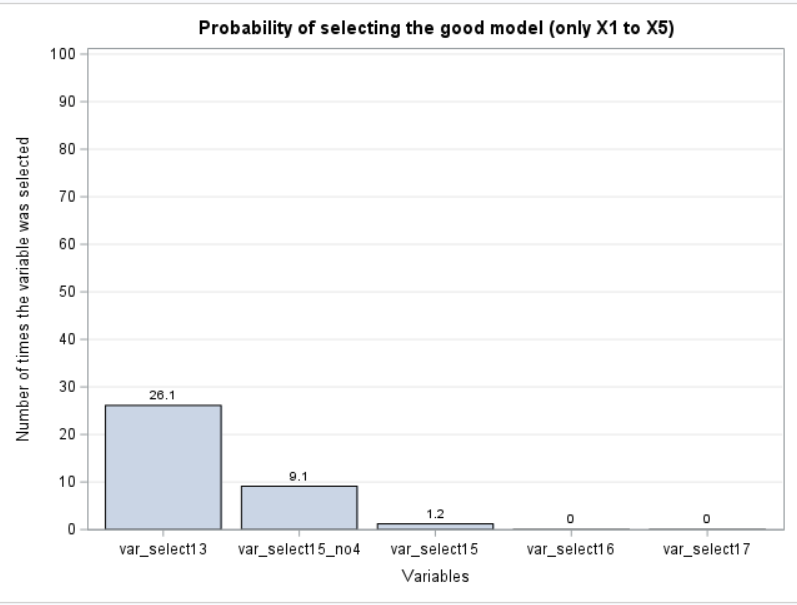
Lars :

Lasso :



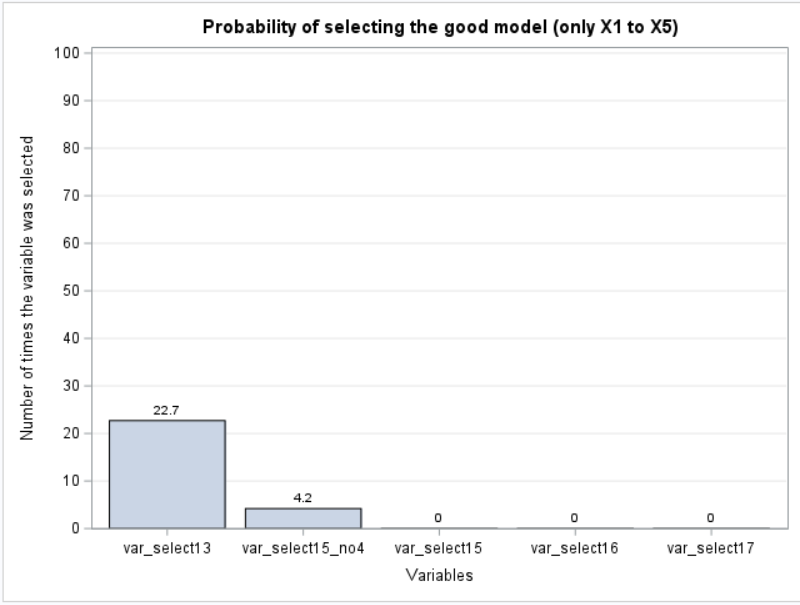
Elasticnet :

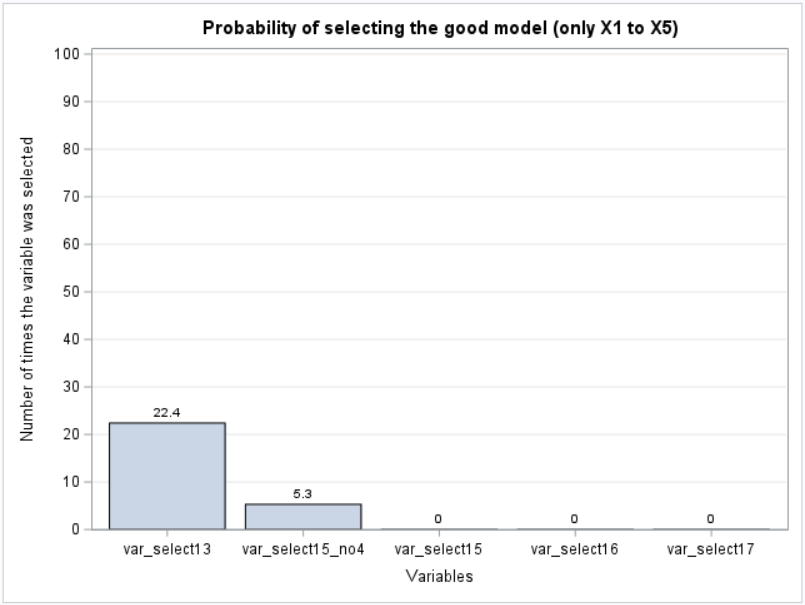
Avec le critère de sélection SBC, la méthode Elasticnet est la seule méthode qui retrouve les variables d’intérêts jointes avec une probabilité de 1,2%. Tandis que Lars et Lasso ne retrouvent aucune variables d’intérêts jointes dû à l’under fitting. Ce dernier se défini par le fait qu’une méthode sélectionne un minimum de variable lorsqu’il doit retrouver le modèles justifiant la valeur de Y. Les méthodes Lars et Lasso couplées au critère de sélection SBC ne sont pas du tout efficace. La procédure de machin learning qui fonctionnerait le mieux en avec ce critère est la méthode Elasticnet qui présente une probabilité de sélection parfaite de 1,2%. Ce résultat est faible, on peut dire que la méthode Elasticnet avec le critère SBC n’est pas pertinent.

**BIC**

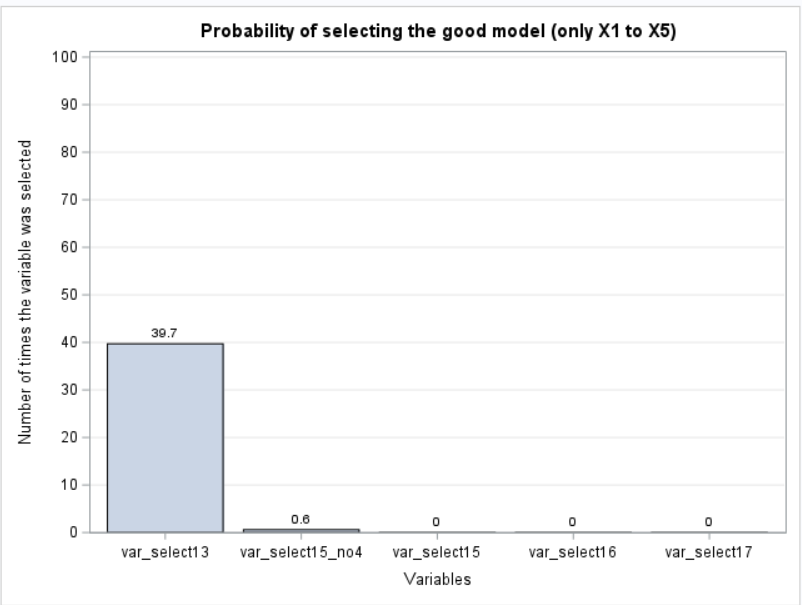
Lars :



Lasso :



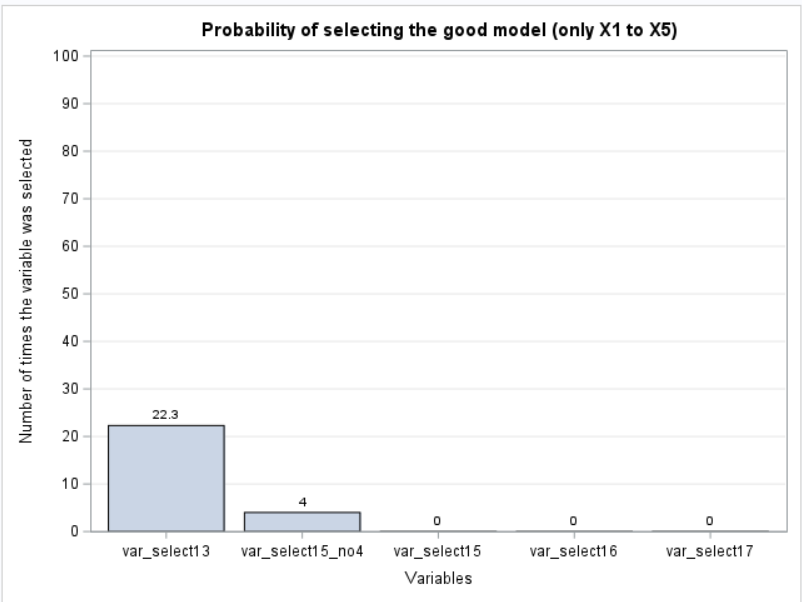
Elasticnet :



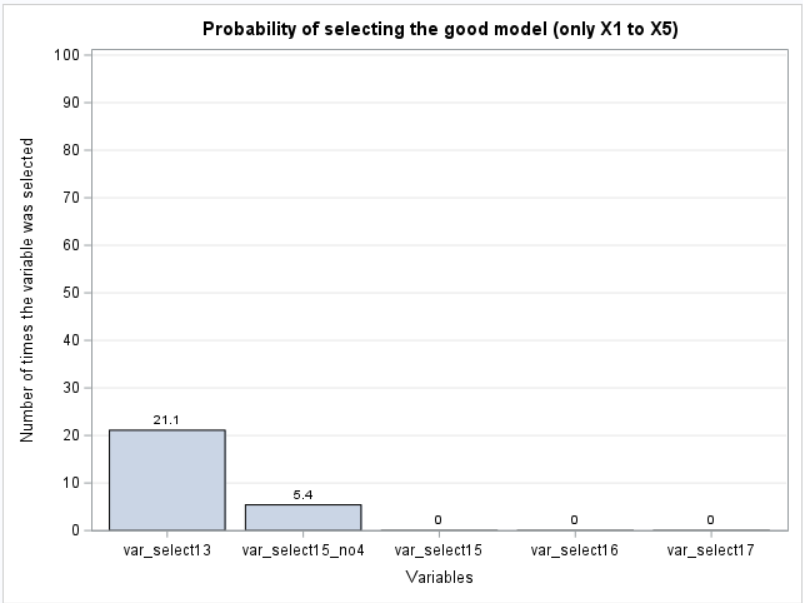
Avec le critère de sélection BIC, les méthodes de machin learning ne retrouvent aucune variables d’intérêts jointes dû à l’under fitting. Ces méthodes ne sont donc pas efficace avec le critère BIC.

**AIC**

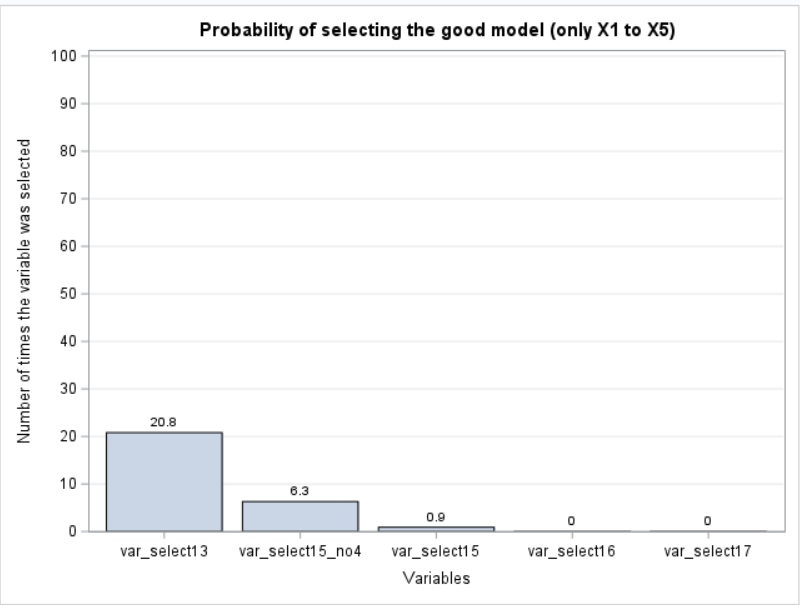
Lars :



Lasso :



Elasticnet :



Avec les critères de sélection AIC et AICC, la méthode Elasticnet est la seule méthode qui retrouve les variables d’intérêts jointes avec une probabilité de 1%. Tandis que Lars et Lasso ne retrouvent aucune variables d’intérêts jointes. Les méthodes Lars et Lasso couplées aux critères de sélection AICC ne sont pas du tout efficace. La procédure de machin learning qui fonctionnerait le mieux en avec ce critère est la méthode Elasticnet qui présente une probabilité de sélection parfaite de 1%. Ce résultat est faible, on peut dire que la méthode Elasticnet avec le critère AIC et AICC n’est pas pertinent.

Conclu : le meilleur de Machin Learning

Pour des bases de données du type DGP3, la méthode de procédure de sélection la plus efficace en Machin learning est la méthode Elasticnet avec le critère de sélection SBC, proposant une probabilité de sélection des variables d’intérêts joints de 1,2%. La méthode Elasticnet est plus performant que Lasso car

(anglais)« elastic Net is proved to better it combines the regularization of both lasso and Ridge. The advantage of that it does not easily eliminate the high collinearity coefficient. »

Ainsi, aucune méthode de Machin learning n’est considéré comme réellement efficace car cette méthode a tendance à faire de l’underfitting quelque soit le critère de sélection, donnant ainsi des résultats peu convaincant.

Comparer meilleur de Stat et meilleur de Machin, et dire quel est la meilleure méthode pour DGP3.

Si on compare la meilleure procédure en Statistical learning (Stepwise) et la meilleure en Machin learning (Elasticnet) , on s’aperçoit de manière évidente que la procédure Stepwise avec le critère de sélection SBC(18,2% de réussite) est plus performant que la procédure Elasticnet avec le critère de sélection SBC (1,2% de réussite) car lorsqu’il y a de corrélation entre les variables, les méthodes de Machin learning ont tendance à faire de l’underfitting quelque soit le critère de sélection, ces méthodes peuvent passer à côté de la sélection de certaines variables significatives. Stepwise avec le critère de sélection SBC est donc la méthode la plus performante pour retrouver les variables d’intérêt dans une base de donnée du type DGP3.

In DGP3 case, on a remarqué que les méthodes de statistical learning ont plus de chances de retrouver les variables d'intérêts, mais ils ont aussi plus de chance de retrouver des variables supplémentaires et donc plus de chance de proposer des modèles avec des erreurs de sélection. Ainsi il y aura plus de chances que la méthode propose un modèle composé de toutes les variables d'intérêts mais avec des variables d'erreurs.

Tandis que les méthodes de machine learning ont moins de chances de retrouver les variables d'intérêt dans le DGP 3, mais ont aussi moins de chances de retrouver des variables supplémentaires et donc moins de chance de proposer des modèles avec des erreurs de sélection. Ainsi il y aura plus de chances que la méthode propose un modèle seulement composé d’une partie des variables d'intérêts mais sans variables d'erreurs.

We notice that statistical learning methods are more likely to propose a model composed of all the variables of interest but with error variables due to under fitting. While machine learning methods are more likely to propose a model composed of only a part of the variables of interest but without error variables due to over fitting.

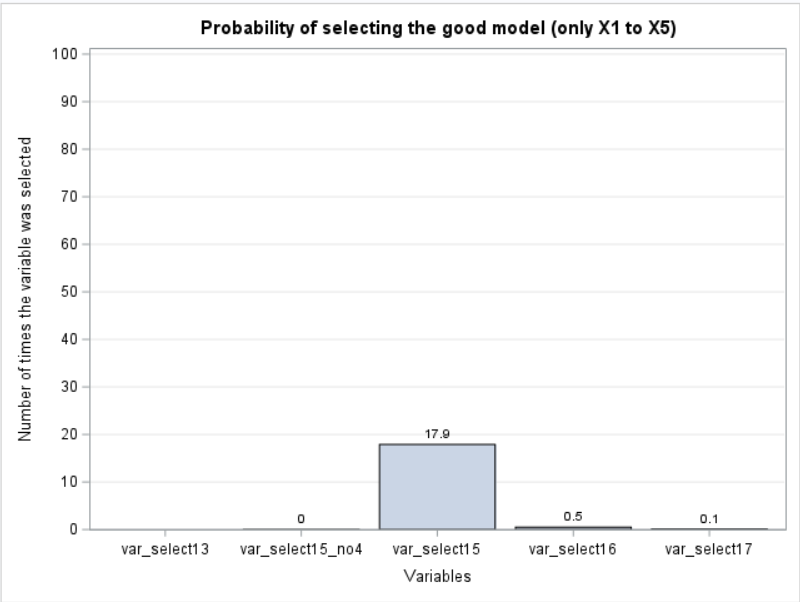
4.4)DGP4 :

Stat Learning :

**SBC**

forward :

Une image contenant table

Description générée automatiquement 

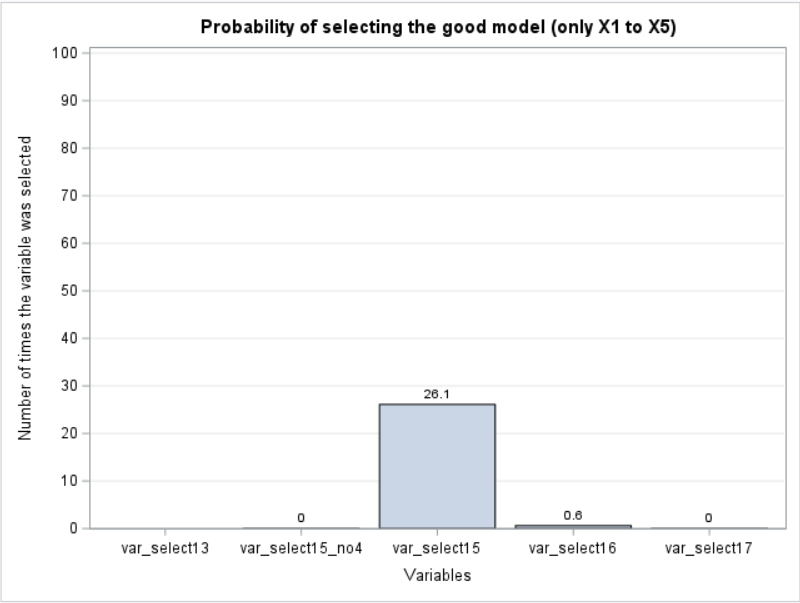
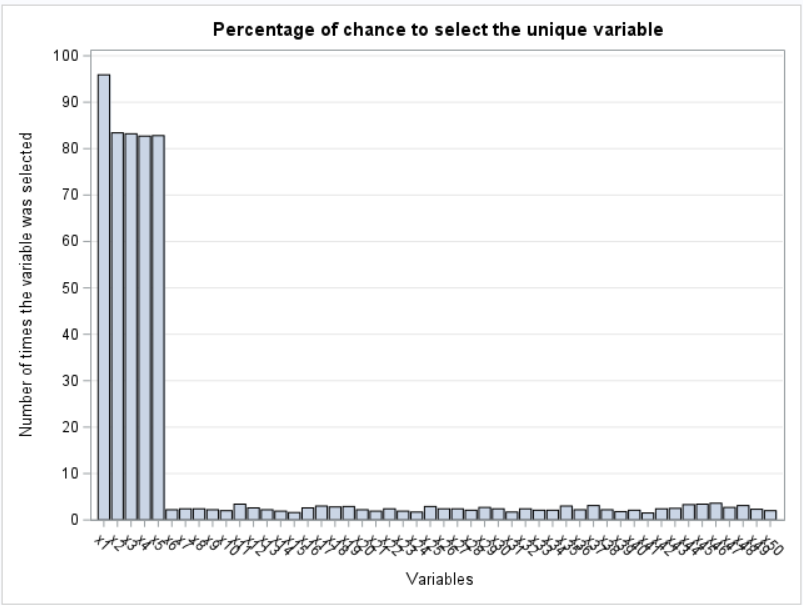
Conclu : le meilleur de Stat Learning

Les méthodes de Statistical learning affichent approximativment les mêmes résultats de sélection parfaite avec les bases de données du type DGP1 et celle du type DGP4. Ainsi, les méthodes de Statitiscal learning ne sembles pas êtres impacté par les valeurs extrêmes. La méthode de procédure de sélection la plus efficace en Statistical learning est la méthode Forward avec le critère de sélection SBC, proposant une probabilité de sélection des variables d’intérêts joints de 17,9%. Backward n’est pas considéré comme efficace car cette méthode a tendance à faire de l’over fitting quelque soit le critère de sélection, donnant ainsi des résultats peu convaincant.

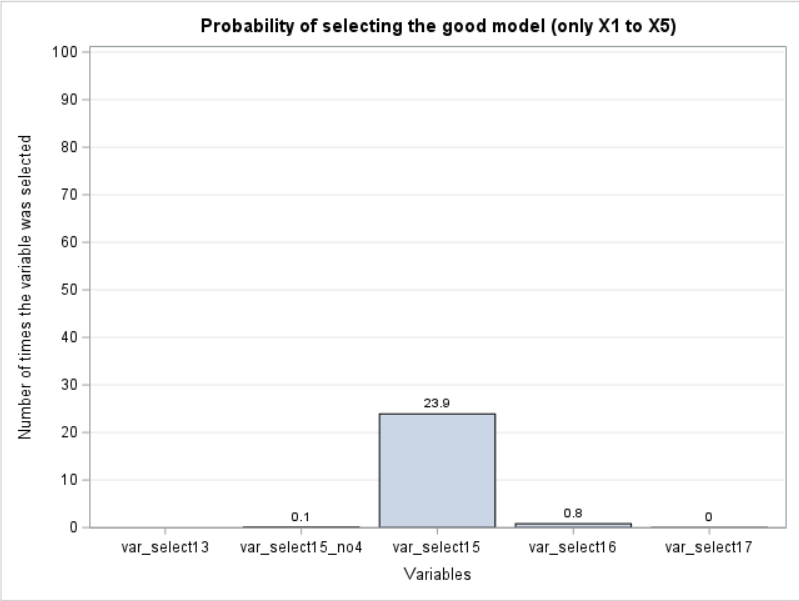
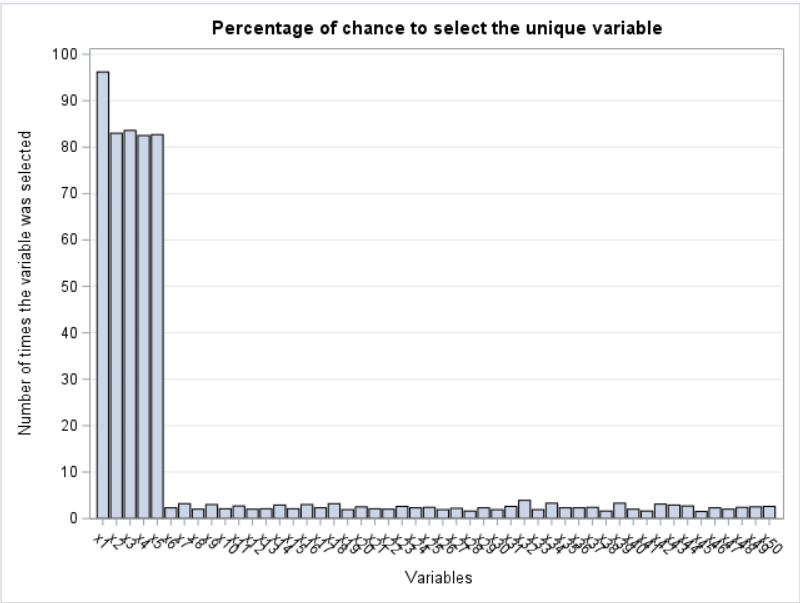
Machin Learning :

**SBC**

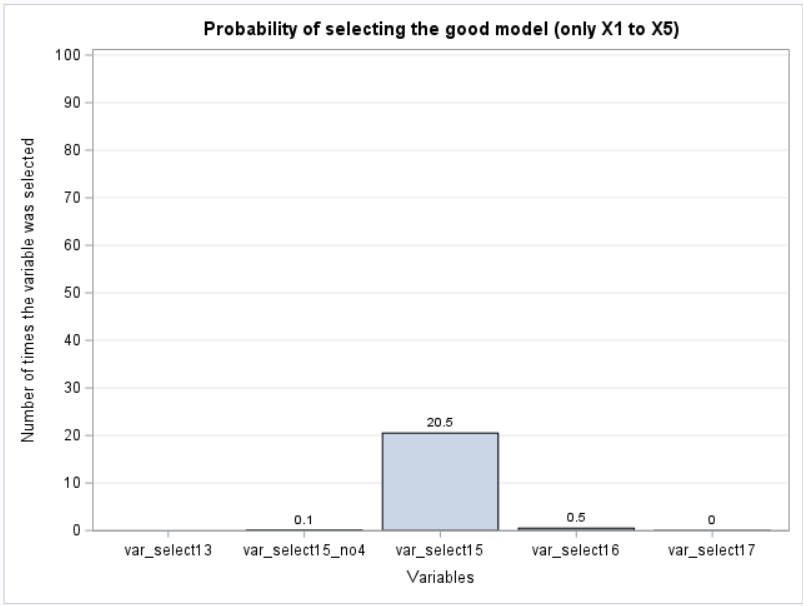
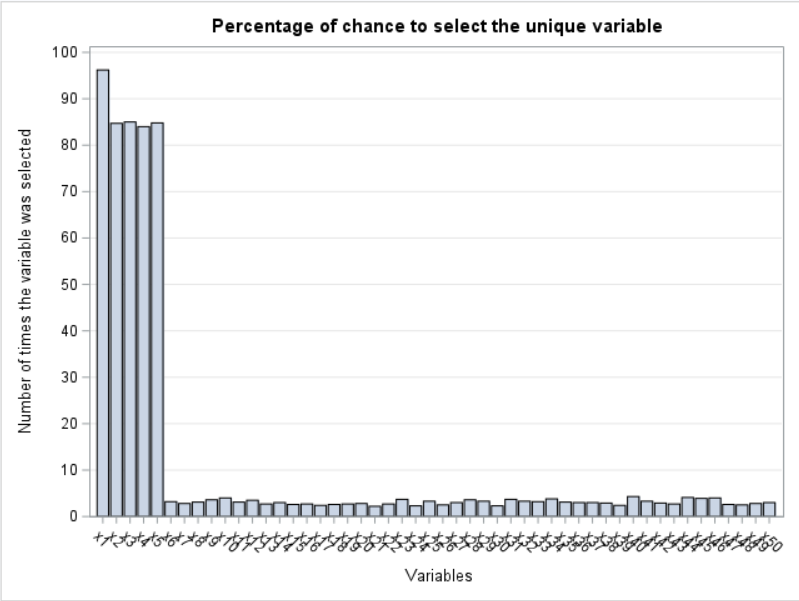
Lars :



Lasso :

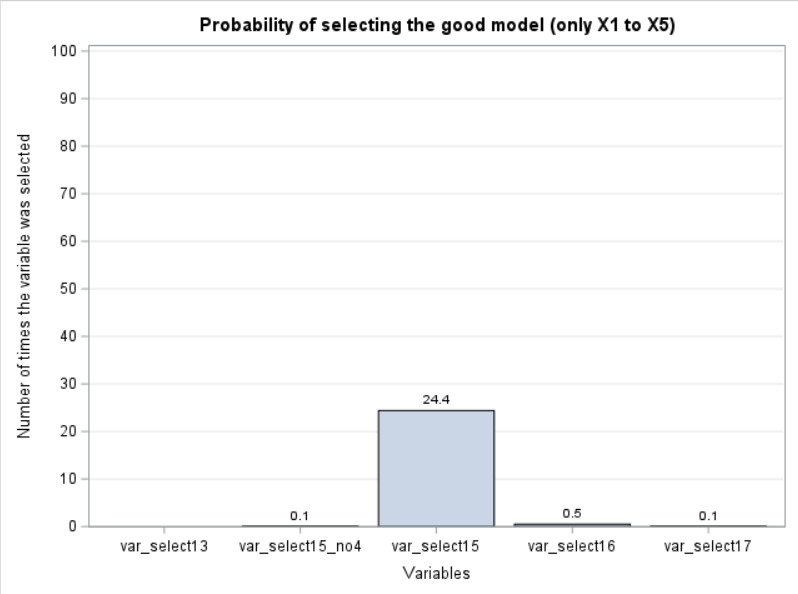


Elasticnet :

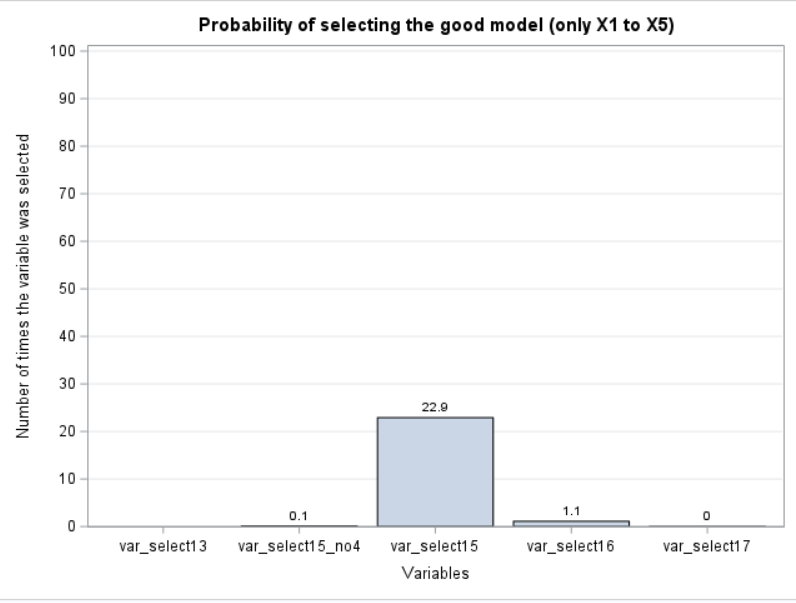


**BIC**

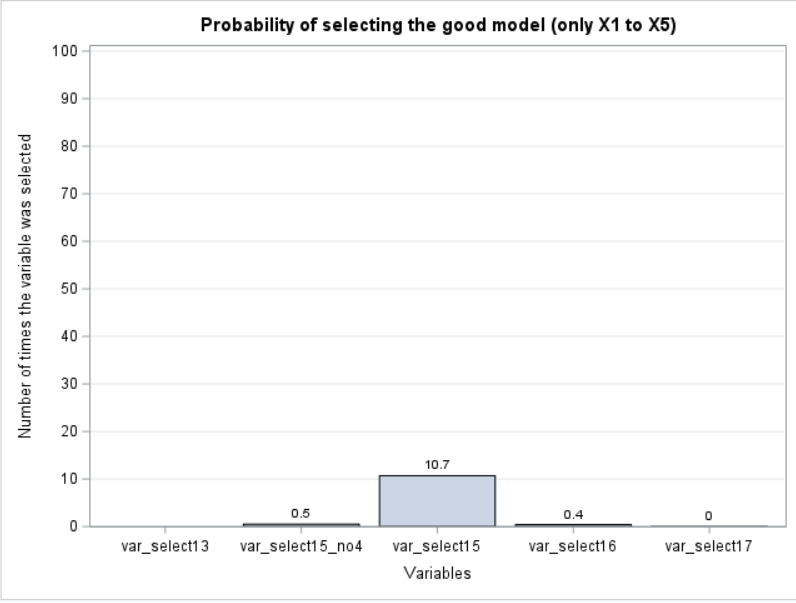
Lars :



Lasso :



Elasticnet :

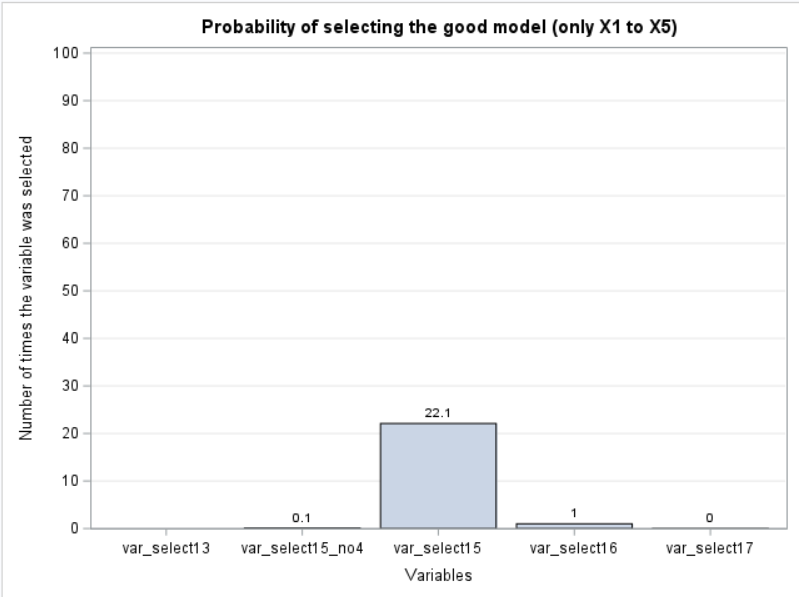


**AIC**

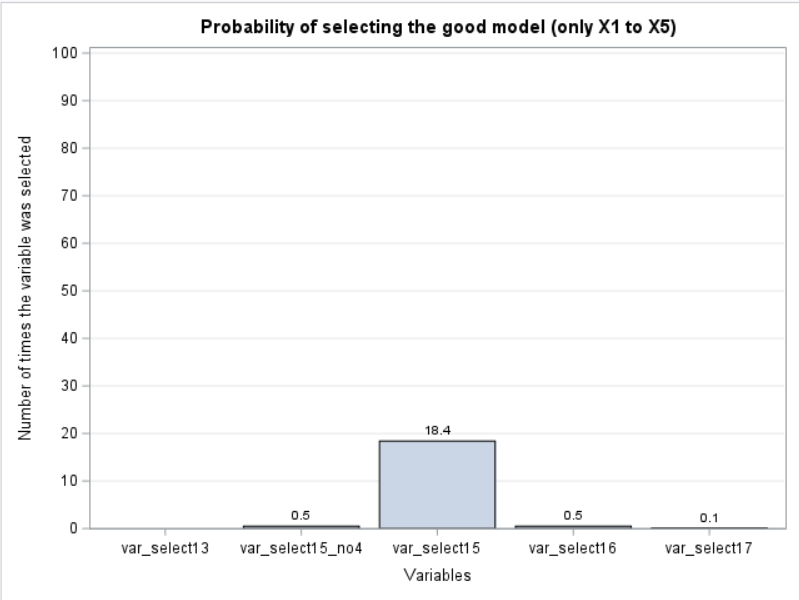
Lars :



Lasso :



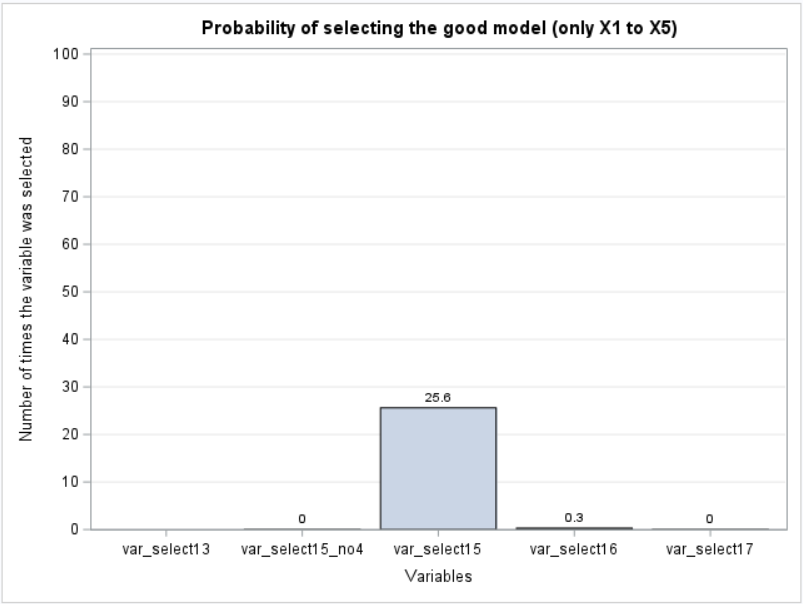
Elasticnet :



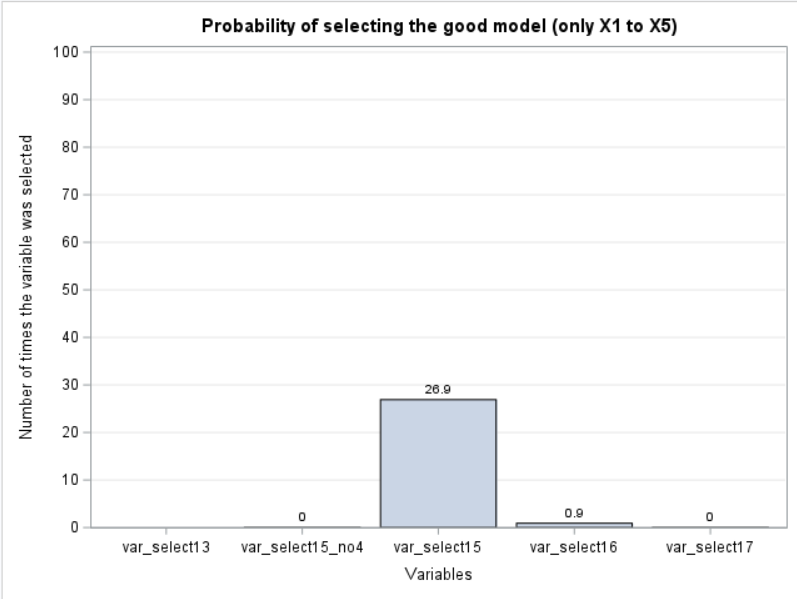
**AICC**

Lars :

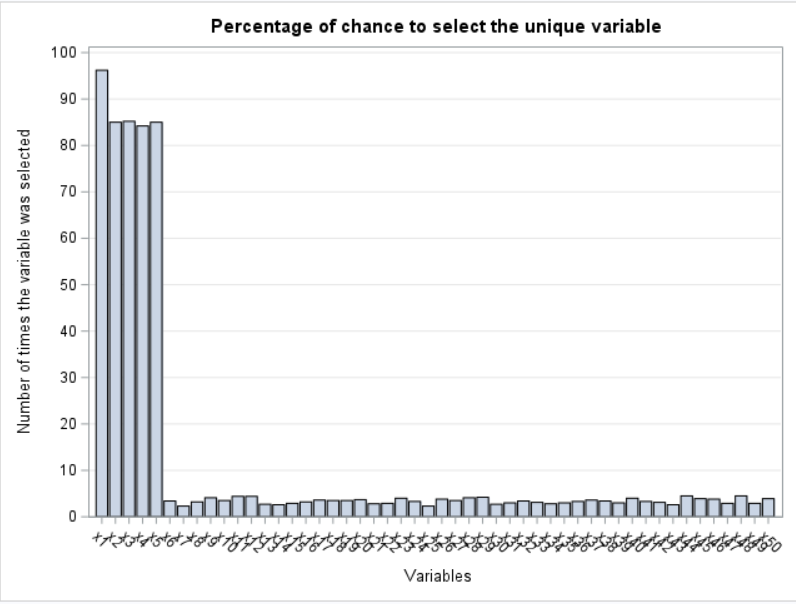
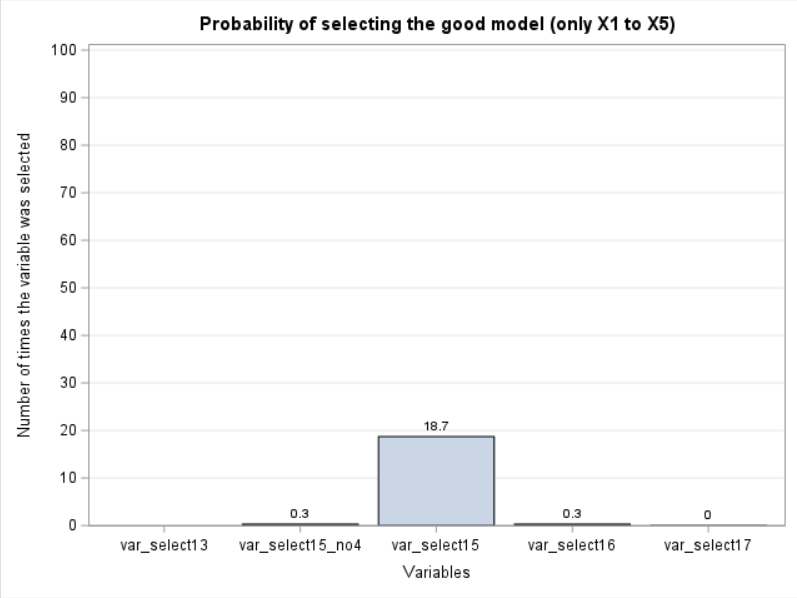
Une image contenant table

Description générée automatiquement 

Lasso :

Elasticnet :

Conclu : le meilleur de Machin Learning

Les méthodes de Machin learning affichent des résultats de sélection parfaite plus faibles avec les bases de données du type DGP4 qu’avec celle du type DGP1. Contrairement aux méthodes de Statistical learning, les méthodes de Machin learning sont impactées par l’existence de valeurs extrêmes dans les bases de données. Lars et Lasso présente à peu près les mêmes performances. Si nous devons choisir la procédure de sélection la plus efficace en Machin learning dans ce cadre de générateur de donnée(DGP4), nous choisirons la méthode Lasso avec le critère de sélection AICC. Ce dernier propose une probabilité de sélection des variables d’intérêts joints de 26,9%. On peut aussi dire que Lars peut être un bon outil de sélection au vu de ses probabilités de sélection très proche de celle de Lasso. Elasticnet présente des probabilités de sélection supérieur aux procédures de Statistical learning, mais il n’est pas le meilleur outils de sélection dans ce contexte là.

Comparer meilleur de Stat et meilleur de Machin, et dire quel est la meilleure méthode pour DGP4 :

Si on compare la meilleure procédure en Statistical learning (Forward SBC) et la meilleure en Machin learning (Lasso AICC) , on s’aperçoit de manière évidente que la procédure Lasso avec le critère de sélection AICC (26,9% de réussite) est plus performant que la procédure Forward avec le critère de sélection SBC (17,9% de réussite). Malgré la présence de valeurs extrêmes et la baisse de performance des méthodes de machin learning, Lasso avec le critère de sélection AICC reste la méthode la plus performante pour retrouver les variables d’intérêt dans une base de donnée du type DGP4.

Conclusion des Résultats :

Après avoir analysé toutes les méthodes de sélection dans différentes situations, on peut conclure que les méthodes de Machin Learning sont plus efficaces par rapport aux méthodes de Statistical Learning dans les situations du type DGP1, DGP2, et DGP4. Tandis que dans une situation du type DGP3, les méthodes de Statistical Learning sont plus efficaces par rapport aux méthodes de Machin learning. Si nous devons comparer l’ensemble des résultats, on conclura que les meilleurs méthodes sont Lasso et Elasticnet.

After examining all of the selection methods in various conditions, it can be concluded that in DGP1, DGP2, and DGP4 type situations, Machine Learning technics are more successful than Statistical Learning methods. Statistical Learning techniques find the variables of interest more easily than Machine Learning methods only in the scenario of DGP3, but Statistical Learning is not necessary better.

If we compare all of the outcomes, we can conclude that LASSO and ElasticNet are the best approaches because c’est mieux d’avoir un modèle avec peu de varirables mais dans lequel on sait qu’elles sont toutes significatives, plutôt qu’un modèles composé de toutes les variables d’intérêts avec des variables supplémentaires mais dans lequel on ne sait connait pas réellement les variables d’intérêts.

If we compare all of the outcomes, we can conclude that LASSO and ElasticNet are the best approaches because It is more interesting to have a model with few variables of interest but in which we know that they are all variables of interest, rather than a model composed of all variables of interest with additional variables but in which we don’t really know the variables of interest.

A travers ces résultats, on remarque que les méthodes Lasso et Elasticnet sont limitées dans certaine situation. Tout d’abord, on sait que Lasso a tendance à sélectionner un sous-ensemble de variable due à la nature de la contrainte exercée sur les coefficients, cependant il aura donc tendance à faire de l’under fitting. Lasso aura plus de chance de ne pas proposer des variables sans intérêts, mais elle a aussi des chance d’ignorer certaine variables d’intérêts. De plus, si des variables sont fortement corrélées entre elles et qu'elles sont importantes pour la prédiction, le lasso en privilégiera une au détriment des autres comme on a pu l’observer dans le cas du DGP2. Un autre cas, où les corrélations posent problème, est quand les variables d'intérêts sont corrélées avec d'autres variables. Dans ce cas, la consistance de la sélection du lasso n'est plus assurée.

L’Elasticnet utilise une combinaison pondérée de régularisations L1 et L2. Comme vous pouvez probablement le voir, la même fonction est utilisée pour la régression LASSO et Ridge, seul le L1\_wt argument change. Cet argument détermine combien de poids va à la norme L1 des pentes partielles. Si la régularisation est L2 pure (Ridge) et si L1\_wt = 1.0 la régularisation est L1 pure (LASSO).

L'Elasticnet permet d’ajouter au Lasso une pénalité Ridge, afin de réduire les biais de sélections du Lasso, Elasticnet retrouve donc de meilleur résultat par rapport au Lasso dans des situations du type DGP2 et DGP3.

Discussion :

A travers ce mémoire nous avons découvert de nombreuses limites, ce qui pourrait nous permettre de nous remettre en question le principe du projet qui est de comparer les différentes de sélection, mais aussi sur le fait de savoir si les DGP réalisés correspondent réellement à la réalité ?

Tout d'abord on peut remettre en question le principe du projet car le fait de comparer les différentes méthodes de sélection ne mène pas forcément à chercher une amélioration de ces méthodes. Pour une entreprise, il serait donc plus intéressant de mener ce type de mémoire vers un objectif d'amélioration plutôt que vers un objectif de comparaison.

Cependant nous ne pouvions proposer un moyen d'améliorer les méthodes de machine learning en paramétrant certains critères de sélection, de choix, ou d'arrêt. Par exemple avec l'Elasticnet, on pourrait proposer un critère de choix C(p), donnant une probabilité de sélection de variables d'intérêt de plus de 50% en DGP1 et en DGP2.

Ensuite on pourrait se poser la question de la cohérence des DGP générées pour ce mémoire avec la réalité. Pour approfondir notre mémoire, il serait donc plus intéressant de créer un DGP6 qui prendrait en compte à la fois l'existence de corrélations entre les variables et la présence des valeurs extrêmes. On aura donc plus de chance de rencontrer des bases de données du type DGP6 dans la réalité plutôt que des bases de données du type DGP3 ou DGP4 car dans la réalité il est rare de retrouver des bases de données où il existe seulement des corrélations sans valeurs extrêmes et inversement.