ÉTUDE STATISTIQUE SUR LE PRIX DES VOITURES

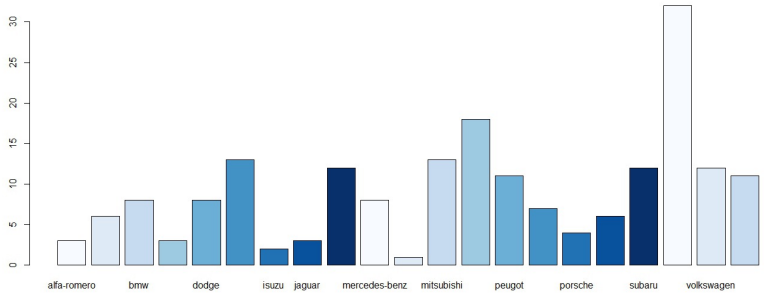
Par Buisson Marianne et Grégoire Lecallier

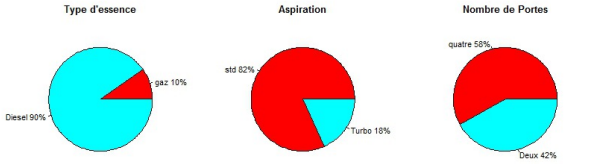
I. Description des données

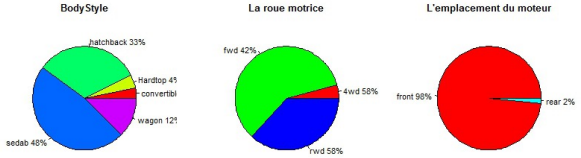
1. Statistique univariée

Nous disposons de 24 variables dont 8 qualitatives. Ces variables correspondent à des caractéristiques de voitures. Grâce à ces données nous allons pouvoir étudier les caractéristiques qui influencent le prix d'une voiture.

La variable make correspond à la marque de la voiture. Il y a plus de 21 marques. Le graphique ci dessous nous permets d'étudier le nombre de voiture par marque.

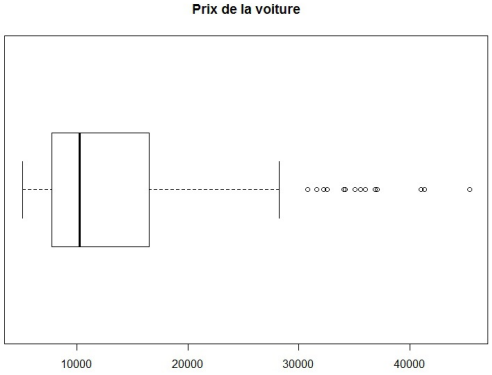






Ces 6 diagrammes circulaires permettent de rendre compte de la répartition des types de caractéristique dans l'ensemble des voitures étudiées.

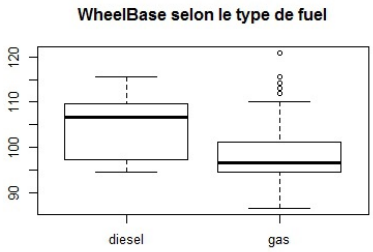
La variable réponse que nous étudions est le prix de la voiture. Ci dessous un boxplot du prix.

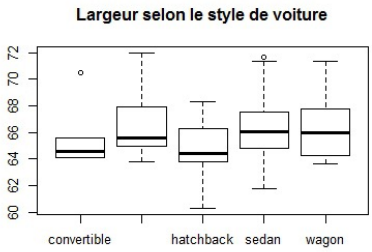


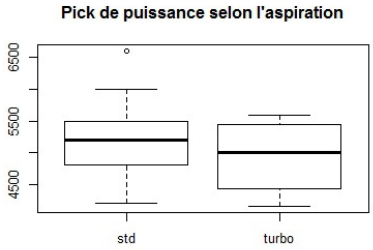
Le prix d'une voiture varie entre 5 118€ et 45 400€. Le prix médian est 13 290€.. On peut voir que les prix au delà de 30 000€ sont plus rares.

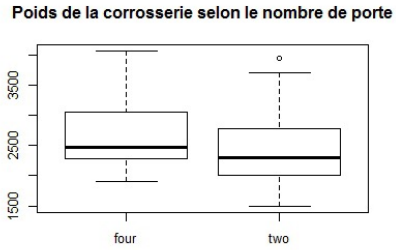
Nous montrerons à travers ce tableau des indicateurs simples, pour décrire les caractéristiques numériques.

2. Statistique bivariée





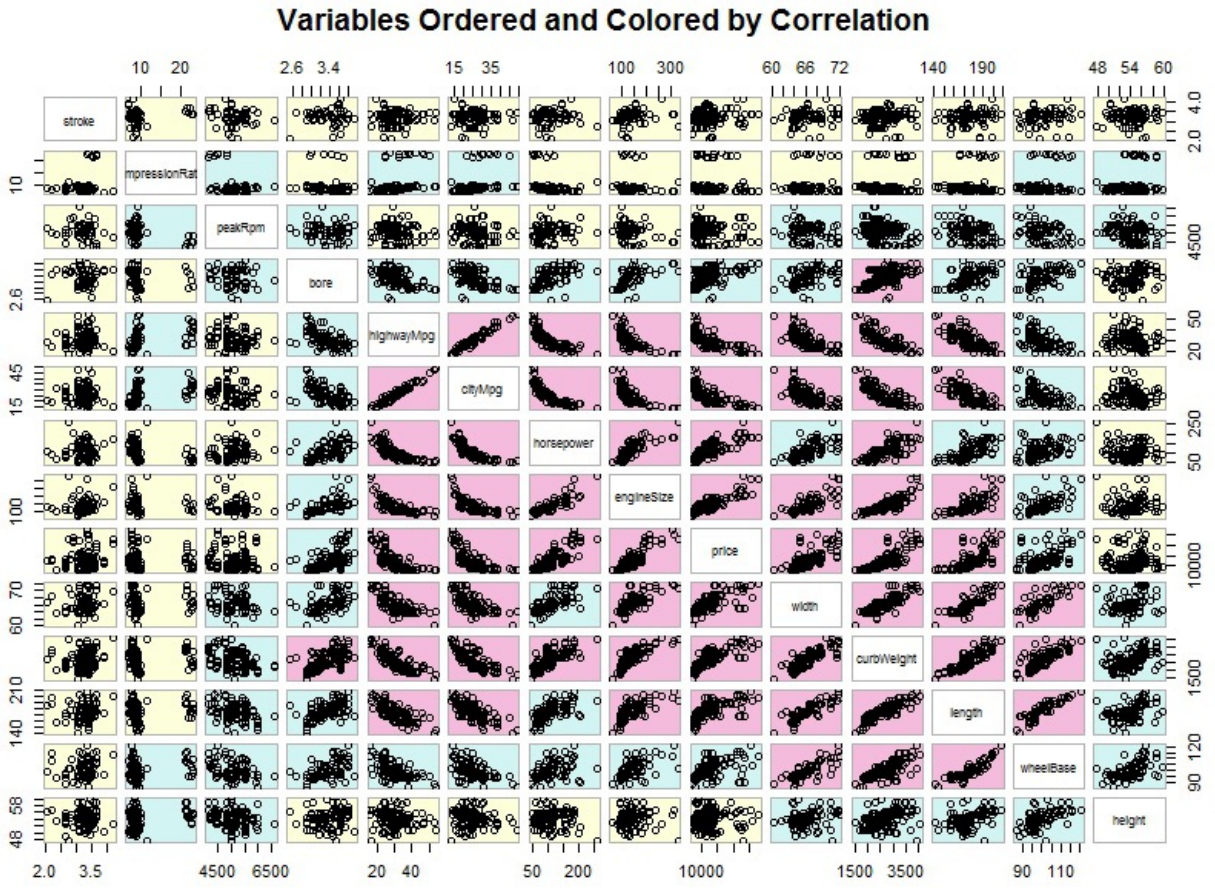




Caractéristiques	Moyenne	Maximum	Minimum
L'empattement	97	121	86
La longueur	174	208	141
Largeur	65	72	60
Hauteur	54	60	48
Poids de la carrosserie	2 562	4 066	1 488
Taille du moteur	120	326	61
L'alesage	3,33	3,9	2,5
La puissance du piston	3,3	4,1	2,07
Ratio de compression	10,14	23	7
Puissance	103,5	262	48
Pic de puissance	5 100	6 600	4 150
CityMPG	25,33	13	49
HightwayMpg	30,79	54	16

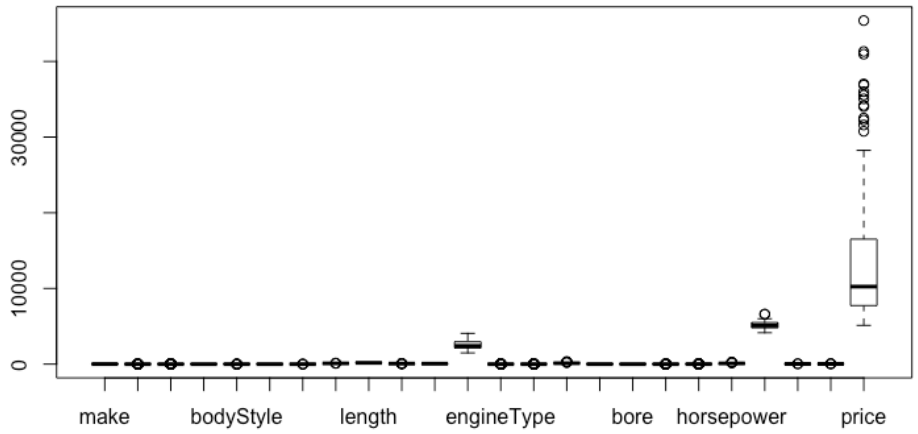
Dans cette partie, nous montrerons rapidement les quelques variable qualitative qui influent sur une variable quantitative à partir d'un box plot.

Quant aux variables quantitatives, le graphique ci dessous permet d'avoir une vision des corrélations entre chaque variable.



On remarque une corrélation entre higway et City/mapg, le prix selon le poids de la carrosserie et la largeur. Le rose montre une corrélation nette quand le bleu et le jaune n'exprime pas de corrélation significative.

Ci-dessous la représentation des boites à moustache nous indique que les variables qualitatives n'ont pas le même ordre de grandeur ainsi que la présence de fortes variabilités. Nous avons donc décidé pour la suite de l'exercice d'effectuer une normalisation des variables numériques (on centre et on réduit les variables quantitatives).
Les prédictions de prix se feront donc sur des prix centrés et réduits.



II. Étude sur les données quantitatives

Désormais nous allons étudier les variables quantitatives qui ont un impact significatif sur le prix. Pour choisir le meilleur modèle on séparera les données en deux : une partie d'essaie et une partie de test.

1. Régression simple

On note Y la variable aléatoire réelle à expliquer (variable endogène, dépendante ou réponse) et Xi les variables explicatives. L'écriture du modèle suppose implicitement une notion préalable de causalité dans le sens où Y dépend de X car le modèle n'est pas symétrique.

Nous choisissons de transformer la variable prix en log afin de limiter les évolutions qui biaiseraient notre modèle.

Variable sélectionnées : La largeur, le poids de la carrosserie, la taille du moteur, la puissance du piston, le ratio de compression, la puissance, les pics de puissance, cityMPG et highwayMPG.

Erreur de prédiction : 0.104
Remarque : le modèle n'est pas validé en raison du non respect de l'hypothèse de normalité des résidus (cf qq plot sur R).

2. Régression rigde

L'objectif de la régression ridage est d'imposer une contrainte sur la taille des coefficients.

Ce programme revient à :

$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \sum \beta_k x_i^k) + \lambda \sum_{k=1} \beta_k$$

λ est le paramètre de complexité :
- Si λ est grand alors on contraint les β à être petit
- Si λ est petit alors on contraint les β à être plus grand

- Variables les plus influentes : La longeur, la largeur et la hauteur.
- Erreur de prédiction sur echantillon test : 0.109

3. Régression lasso

Rappelons que l'idée du Lasso est non pas de faire un régression linéaire classique mais une régression linéaire sous contrainte que la norme L1 des coefficients soit inférieur à un certain seuil.
Contrairement à Ridge (norme L2), la norme L1 sous lasso peut atteindre la borne 0.

- Variables les plus influentes sur le prix : La longueur, la largeur et la hauteur.

On produit une validation croisée ci dessous :
Le graphique montre les lambda en abscisse : le lambda minimum est -6,5 pour une erreur de prédiction de 0,0035.
Notre meilleur modèle est celui ci.

4. Régression par CART

La méthode CART consiste à construire le découpage pas à pas de la variable Y à expliquer.
L'intérêt est de minimiser :

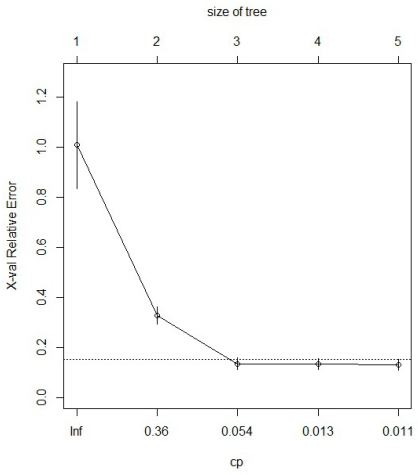
$$min[\sum_{i=1} - \hat{c}_1] \text{ si } x_i \in \tilde{R}_1^j(s)$$

$$min[\sum_{i=1} - \hat{c}_2] \text{ si } x_i \in \tilde{R}_2^j(s)$$

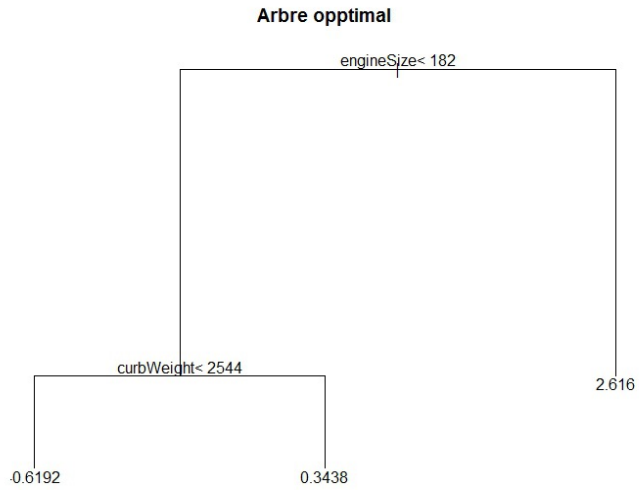
Où c_1 est la moyenne de y_i et x_i appartenant à $\tilde{R}_i$.

On cherche à ajuster un arbre de taille M (m noeuds terminaux). On choisi M

- Variables selctionnees : la taille du moteur (ceux dont la taille est inférieur à 182 contre le reste) et le poids de la carrosserie. $M = 3$



Une première façon d'estimer l'erreur par validation croisée consiste à tracer la décroissance de l'estimation de l'erreur relative (erreur divisée par la variance de la variable à modéliser) en fonction du coefficient de complexité, c'est-à-dire plus ou moins aussi aussi en fonction de la taille de l'arbre ou nombre de feuilles. Le choix optimal suggéré est la valeur du cp la plus à gauche en dessous de la ligne , soit cp =0,054
Ci dessous notre arbre optimal :



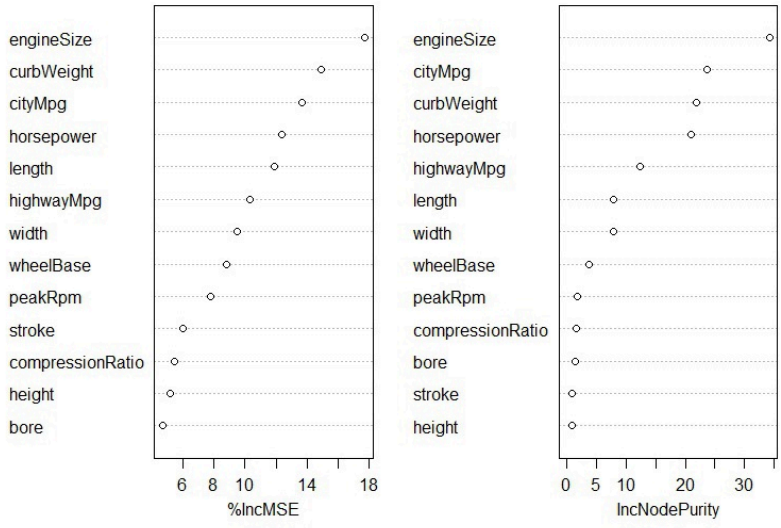
Erreur de prediction sur l'echantillon test : 0.03

5. Régression par forêt aléatoire

I

Cette méthode consiste à créer une multitude d'arbres de décision où chaque arbre est construit à partir d'échantillons aléatoires (appelés échantillons bootstrap) et où chaque noeud de décision est constitué de une ou plusieurs variables choisies au hasard (paramètre mtry). Ici , nous avons construit 500 arbres et à chaque noeud l'algorithme fait un essai sur p/3 variables par défaut, avec p = 23. Chaque arbre est donc entrainé sur une partie des données et nous pouvons estimer son taux d'erreur sur les données qu'il n'a pas utilisé pour le bootstrap ("out of bag error") et plus ce taux est faible plus le taux est juste.

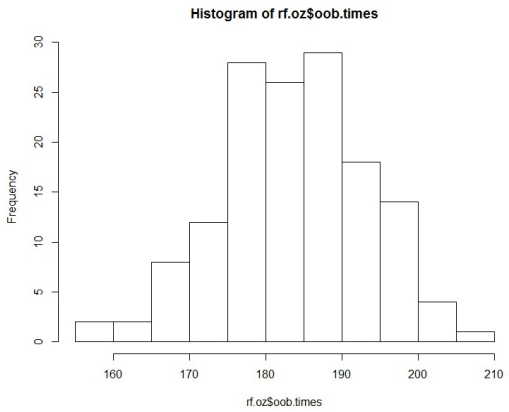
- Variables les plus importantes : taille du moteur, poids de la carrosserie,CityMPG, ce sont celles qui contribuent le plus à faire baisser le MSE.



- Validation du modèle :

Par l'indice de Gini : la diminution moyenne de l'impureté apportée par chaque variable, puis une moyenne sur l'ensemble des arbres est effectuée. (Cf graphe ci-dessus)

Par un histogramme des erreurs de prédictions :



Une fonction permet d'estimer le taux d'erreur sur les données des arbres « out-of-bag » : plus ce taux est faible, plus le modèle est juste. Ce chiffre à lui seul pourrait servir d'indicateur de performance du modèle.

Erreur de prediction sur l'echantillon test : 0.015

6. Régression ar la méthode des réseaux de neurone

Considérons le cas le plus simple de la reégression avec un réseau constitueé d'un neurone de sortie linéaire (le prix des voitures lié aux autres variables) et d'une couche à q neurones dont les paramètres (β) sont optimisés par moindres carrés. L'apprentissage est l'estimation des paramètres α_j=0,p ; k=1,q et β_k=0, par minimisation de la fonction perte quadratique :

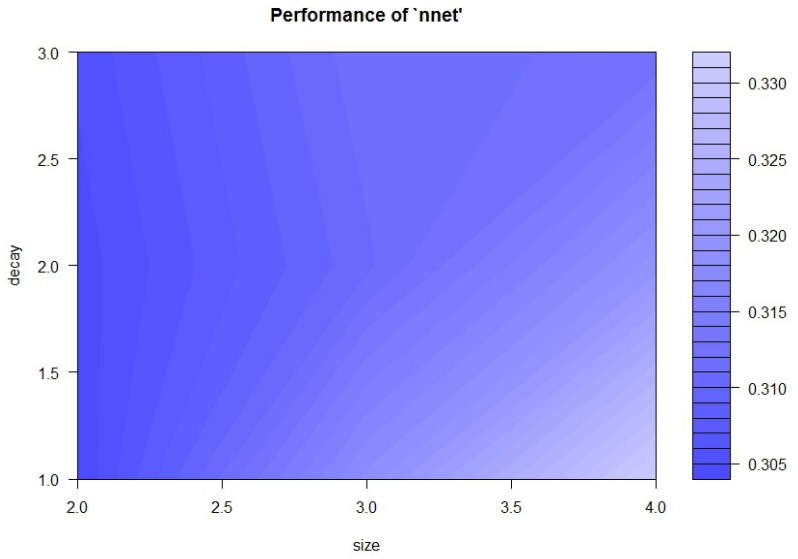
$$Q(\alpha,\beta) = \sum_{i=1}^{n} Q_i = \sum_{i=1}^{n} [y_i - f(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta})]^2 .$$

- Variables ssélectionnées:

B->h1	i1->h1	i2->h1	i3->h1	i4->h1	i5->h1	i6->h1	i7->h1	i8->h1	i9->h1	i10->h1	i11->h1	i12->h1	i13->h1
-0,01	0	0,02	-0,33	0,34	0	0,01	0,03	0,25	-0,22	-0,04	0	-0,41	0
B->h2	i1->h2	i2->h2	i3->h2	i4->h2	i5->h2	i6->h2	i7->h2	i8->h2	i9->h2	i10->h2	i11->h2	i12->h3	i13->h3
0	0,04	0,12	0,01	0	-0,02	0,07	0,01	0,03	-0,07	-0,14	0,01	-0,02	-0,04
B->h3	i1->h3	i2->h3	i3->h3	i4->h3	i5->h3	i6->h3	i7->h3	i8->h3	i9->h3	i10->h3	i11->h3	i12->h3	i13->h3
-0,01	0	0,01	0,11	-0,32	0	0,07	0,05	-0,19	0,12	-0,1	0,01	-0,41	-0,35
B->h4	i1->h4	i2->h4	i3->h4	i4->h4	i5->h4	i6->h4	i7->h4	i8->h4	i9->h4	i10->h4	i11->h4	i12->h4	i13->h4
0	0,04	-0,02	0	-0,04	0	-0,21	-0,01	0	-0,08	-0,37	0,02	0	-0,02
B->h5	i1->h5	i2->h5	i3->h5	i4->h5	i5->h5	i6->h5	i7->h5	i8->h5	i9->h5	i10->h5	i11->h5	i12->h5	i13->h5
-0,01	0,08	0,1	0,31	-0,51	0	-0,02	0,02	-0,1	-0,24	0,09	0	0,18	0,2
B->0	H1->0	H2->0	H3->0	H4->0	H5->0								
0,32	1,01	-0,47	0,87	-1,43	0,75								

Ici les nombre de la ligne grisée représentent α à estimer. Si il vaut 0 alors il n'y a pas de lien entre deux variables (pas de connexions). Nous avons 6 couches pour 13 variables explicatives.
On retrouve les variables qui ont le plus d'impact sur le prix : largeur, longueur, hauteur, taille du moteur, bore, compression ratio horse power et City MPG.

- Validation du modèle :



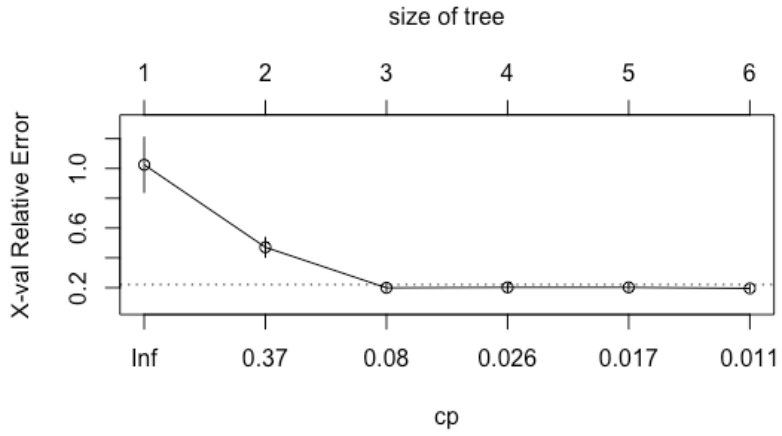
Performance of net indique que plus le bleu est foncé plus les erreurs sont faibles. Ce graph ede performance est fait sur les variables sélectionnées précédemment. On remarque que le graph eest presque entièrement bleu foncé. Il s’agit simplement de ne pas prendre une taille trop grande et une pénalité trop faible entre 1 et 1,5 afin d'ajuster au mieux le compromis biais/variance.

Erreur de prediciton : 0.03

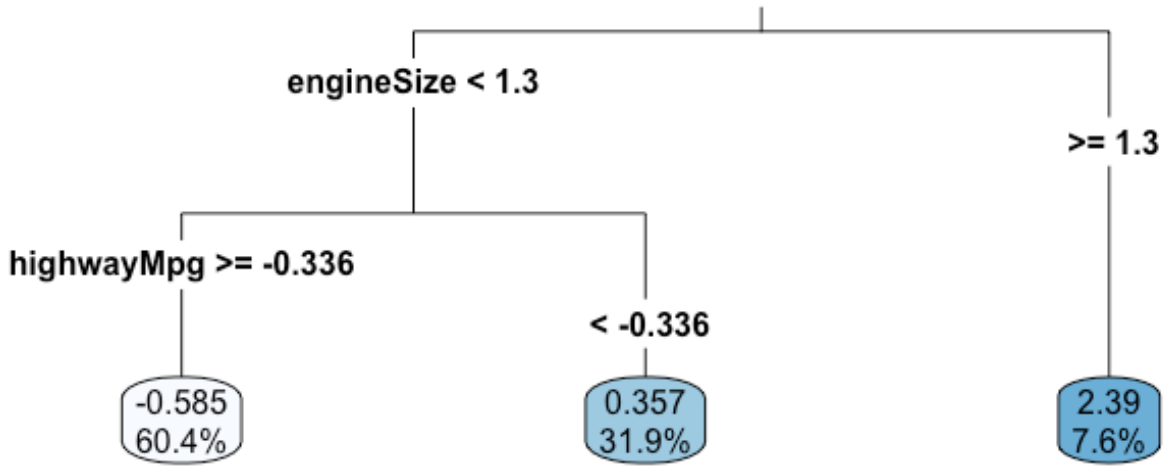
III. **Étude sur les données qualitatives**

On compare désormais réspéctivement les modèles CART, random forest et réseau de neurones précèdent avec les modèles construits sur le jeu de données comprenant les variables qualitatives.

1. Régression par CART



Ici le meilleur cp semble être égale à 0.08, ce qui donne l'arbre de décision de taille 3 suivant :

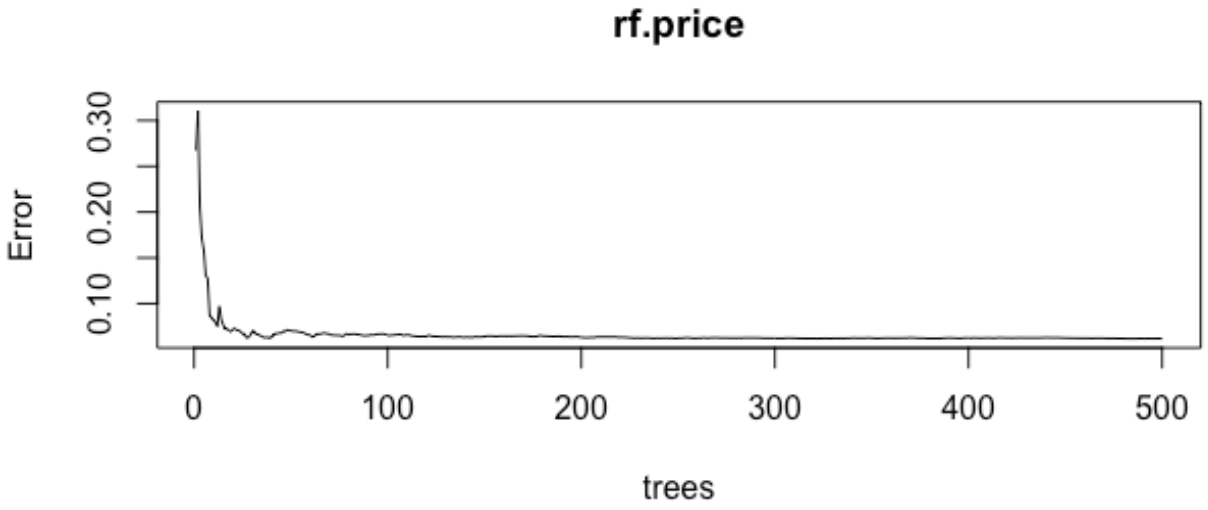


Le premier noeud de décision est la même variable que dans l'arbre précèdent. Le 2e noeud est cependant different et l'arbre est constitué de 3 feuilles comme dans l'arbre précédent.
Le résultat affiché sur chaque feuille représente la moyenne des prix du sous-groupe constitué lors du découpage des données à chaque noeud de décision. Notons que les prix on été normalisés, c'est pour cela que les résultats sont proches de 0.
Les pourcentages représentent la part des données que contient chaque feuille.
Cet arbre est aussi "complexe "que le précèdent adans le sens où il y aseulement 2 noeuds de décisions.

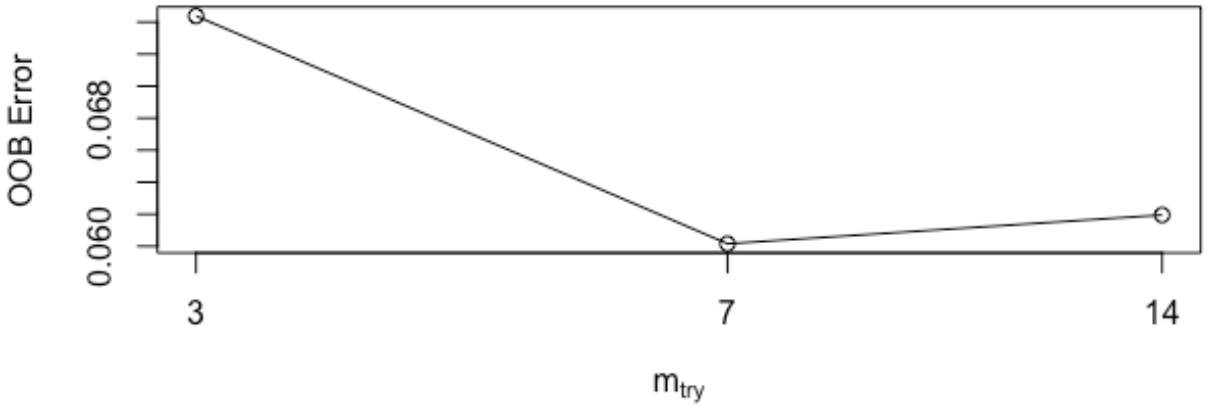
Erreur de prediction: 0.0695

2. Random Forest

Ci-dessous on représente l'erreur en fonction du nombre d'arbres construit :

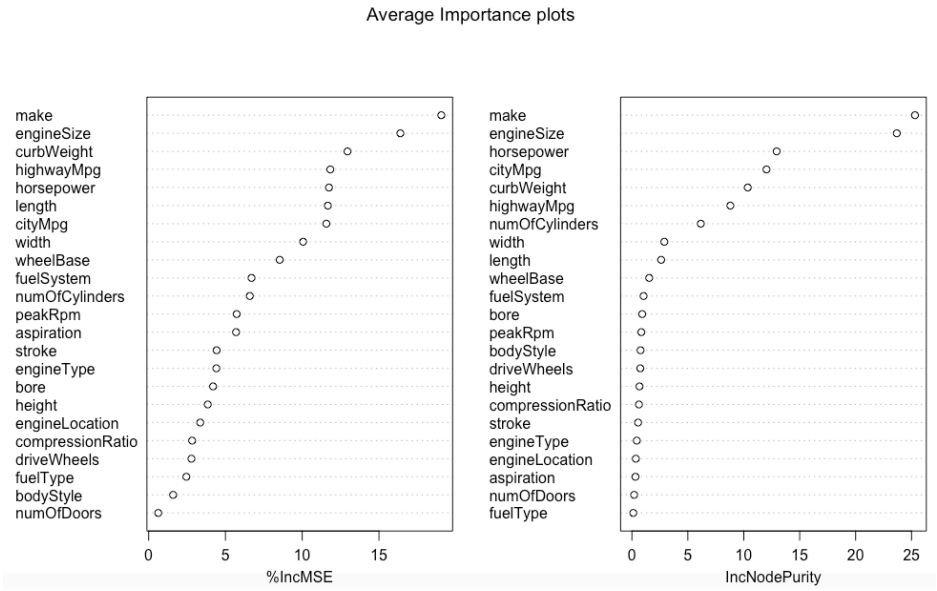


On observe que l'erreur diminue peu à partir de 100 arbres, augmenter le nombre d'arbre dans l'algorithme random forest au delà de ce seuil ne ferait que'ugmenter la complexité du modèle pour peu de gain en performance.
Nous choisissons donc pour améliorer notre random forest, d'optimiser le paramètre mtry en limitant le nombre d'arbres construits à 100 afin de limiter slaccomplexité du modèle :



Le graphe ci-dessus nous montre que l'erreur out of bag la plus petite est celle pour un mtry égale à 7.

On peut mesurer l'importance des variables en représentant la contribution de chaque variable à la décroissance moyenne du MSE :



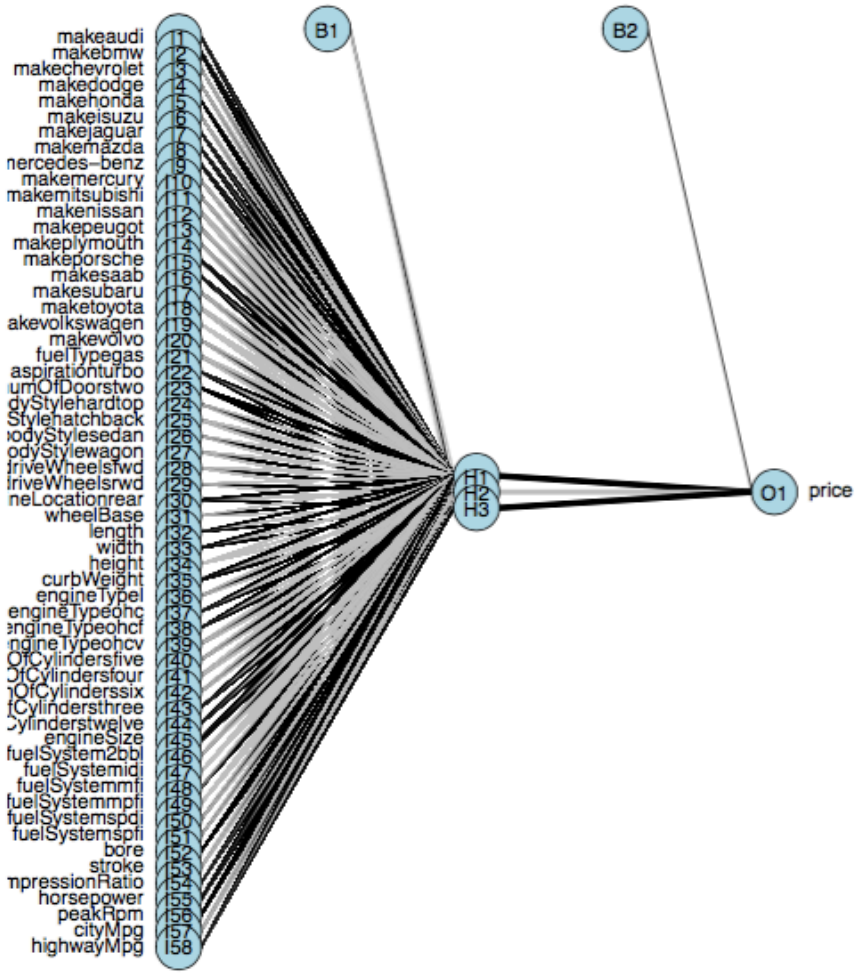
On remarque alors que la variable qualitative "make", soit la marque de la voiture, est la variable la plus importante dans le critère de décision du prix.
L'information sur cette variables qualitative est donc importante ici.
La 2e et la 3e meilleure variable correspond respectivement à la meilleure et à la 2e meilleure variable du modèle random forest construit à partir du jeux de données "purement qualitatif".
La variable CityMpg qui était en 3e position est redescendue en 7e position dans ce nouveau modèle.

Erreur de prédiction : .031

3. Réseaux de neurone

L'optimisation des paramètres (pénalisation et nombre de neurones) se fait par validation croisée.
Notant que "Plus la valeur du paramètre γ (decay) est importante et moins les poids des entrées des neurones peuvent prendre des valeurs chaotiques contribuant ainsi à limiter les risques de sur-apprentissage" (source : https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-rn.pdf), les paramètres choisis pour notre modèle de réseau de neurone sont decay = 1 et size = 3.

Ce qui nous donne le réseau ci dessous :



Erreur de prédiction 0.036

IV **Conclusion**

Pour conclure nous synthétisons les erreurs de prédictions sur les échantillons test pour chaque modèle étudié dans les parties II & III sur le tableau ci-dessous:

Sans variables qualitatives					Avec variables qualitatives		
Régression linéaire	**Ridge**	**CART**	**RF**	**NNET**	**CART**	**RF**	**NNET**
10.4 %	10.9%	3.14%	**1.54%**	3.20%	6.52%	3.10%	3.64%

Commentaires :

On remarque dans un premier temps que les modèles de régression linéaire, Lasso et Ridge obtiennent une erreur de prédiction élevée comparée aux autres modèles, de l'ordre de 10%.
La simplicité de ces modèles ont en effet pour contrepartie un biais élevé, de plus nous ne validons pas l'hypothèse de normalité des résidus. Nous ne retenons donc pas ces modèles

Parmis les modèles restant, le modèle CART avec les variables quantitatives obtient la plus forte erreur de prédiction.
En effet l'arbre de décision découpe les variables réponses en sous groupes, pour au final obtenir 3 groupes de prix. Or la variable réponse est supposée continue, ce qui crée d'important "gap" dans les prédictions, ce sont ces gaps qui contribuent à un taux d'erreur élevé.

Dans tous les cas ,l'ajout de variables quantitatives n'améliore pas la qualité de la prédiction, ce qui semble contre-intuitif.
Le modèle ayant l'erreur de prédiction la plus petite est donc le modèle random forest construit à partir du jeu de données sans les variables qualitatives avec une erreur de 1.54%.
Le modèle random forest a en effet pour avantage le fait qu'il ne requiert pas d'hypothèses sur la loi des résidus.