Cars Partie 2

Jean Clark, Lilian Boissé et Théo De Sa Morais

Contents

1 Bon ou mauvais achat?

Au premier semestre nous avions étudié cette base de données, cars, et avions observé que la première variable de celle-ci donnait l'information sur la qualité de l'achat; si c'est un bon ou mauvais achat. Il serait intéressant dans une approche du point de vue d'un acheteur d'avoir une idée précise des caractéristiques d'un bon achat et de produire un modèle prédictif qui permettrait de correctement prédire les bons achats.

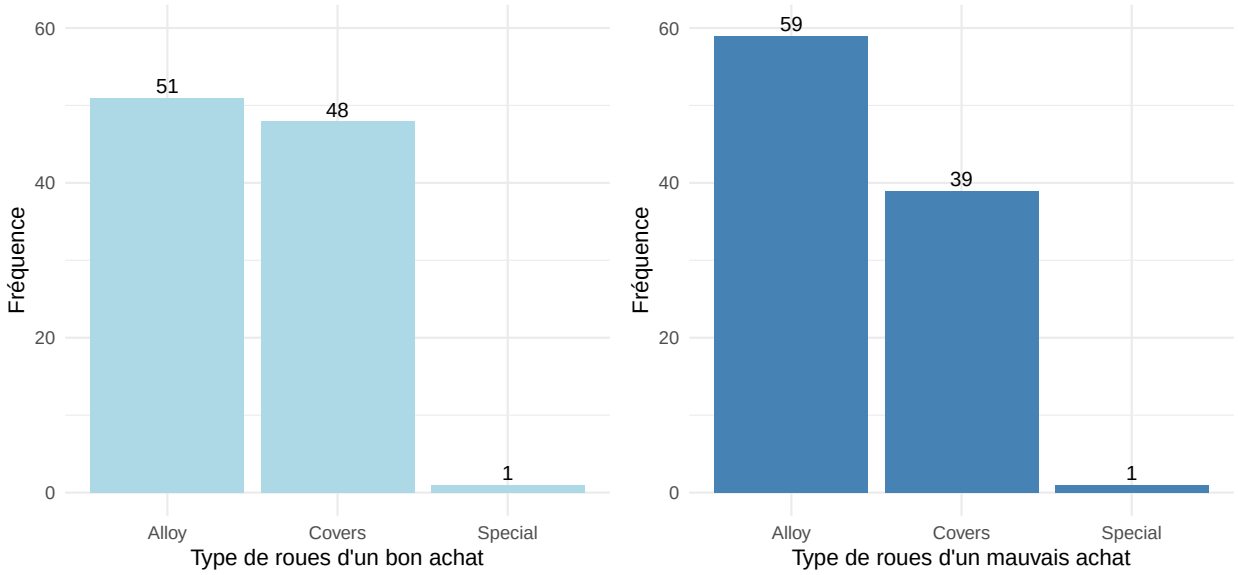
1.1 Les données

Afin d'avoir une idée des facteurs qui différencient les véhicules qui sont des mauvais achats ou non, il convient d'avoir une idée des différences entre les deux groupes(mauvais achats/bons achats) et les potentielles variables qui les différencient le plus.

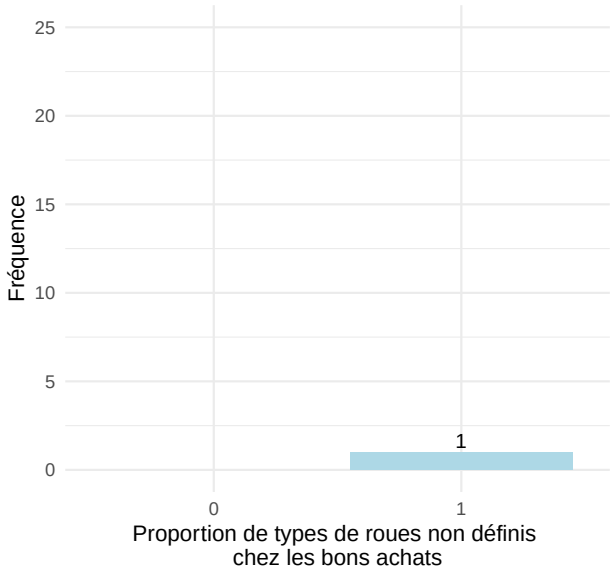
1.1.1 Les qualitatives

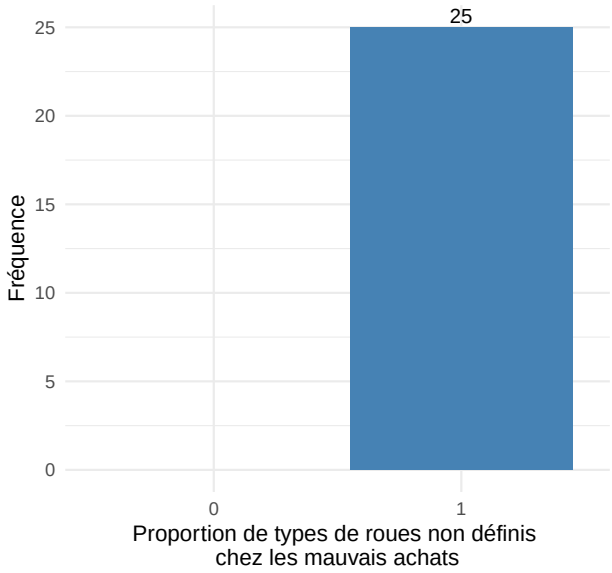
Facteurs de différenciations

Tout d'abord l'âge, il semblerait que ce soit l'un des facteurs qui différencie le plus les deux groupes. En effet, la proportion des voitures agées chez les mauvais achats est plus élevée que chez les bons achats. Il semblerait aussi que la proportion de véhicules équipés de roues alloy soit plus élevée chez les mauvais achats que chez les bons achats.



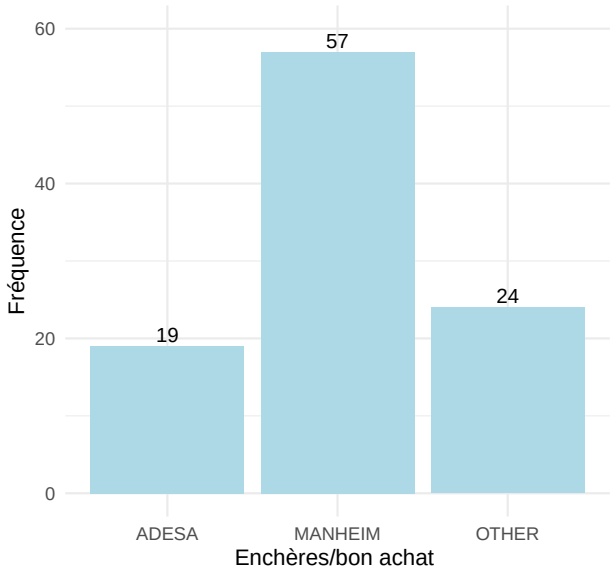
En étudiant les types de roues de plus près, on observe que certaines lignes de notre base de données contiennent des données manquantes sur le type de roues. On pourrait imaginer que lorsque le type de roues n'est pas rentré dans la base de données, c'est peut être parce qu'un remplacement de roues ou une réparation est nécessaire. Observons la répartition de ces données manquantes sur les deux groupes.

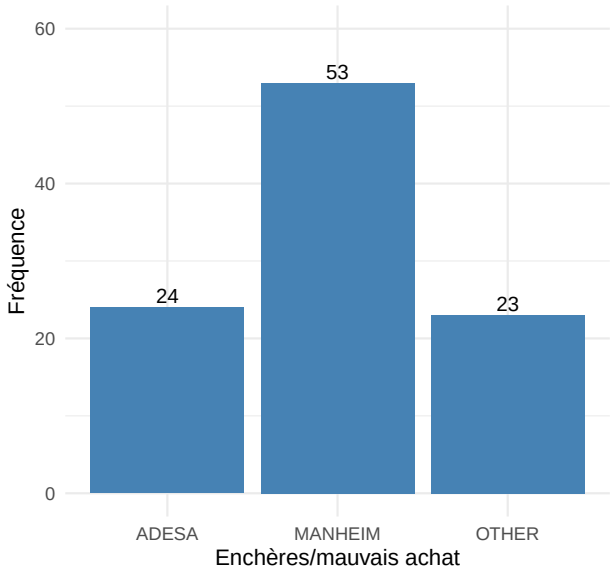


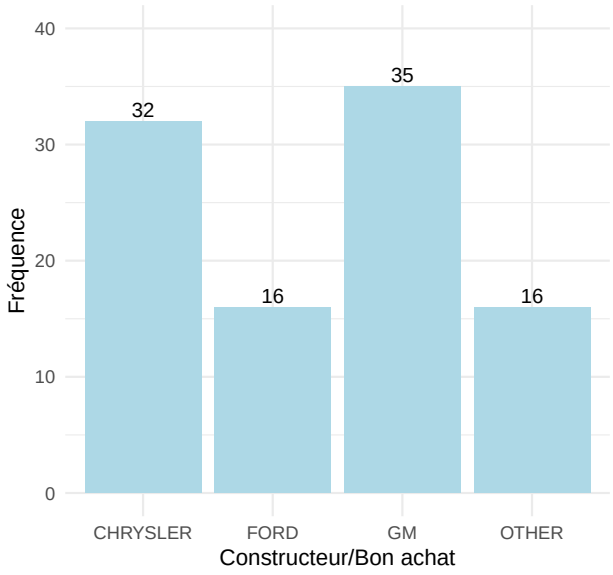


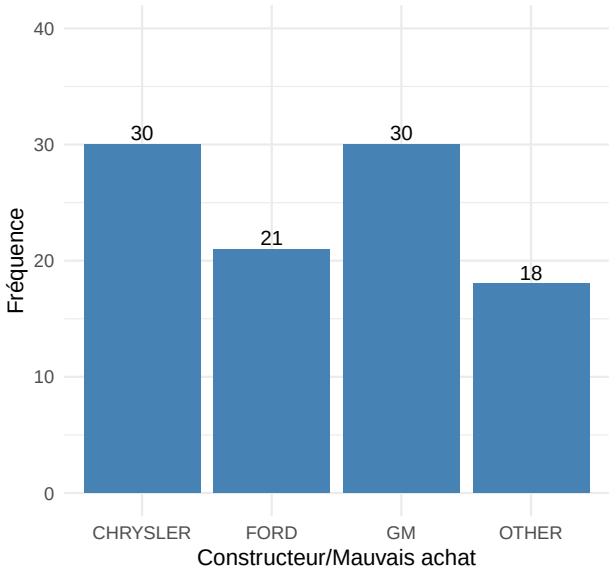
Il y a 25 fois plus de types de roues non définis chez les mauvais achats que chez les bons achats.

Il semblerait aussi y avoir relativement plus de mauvais achats aux enchères Manheim. Les mauvais achats sont aussi relativement plus représentés chez Ford que les bons achats.





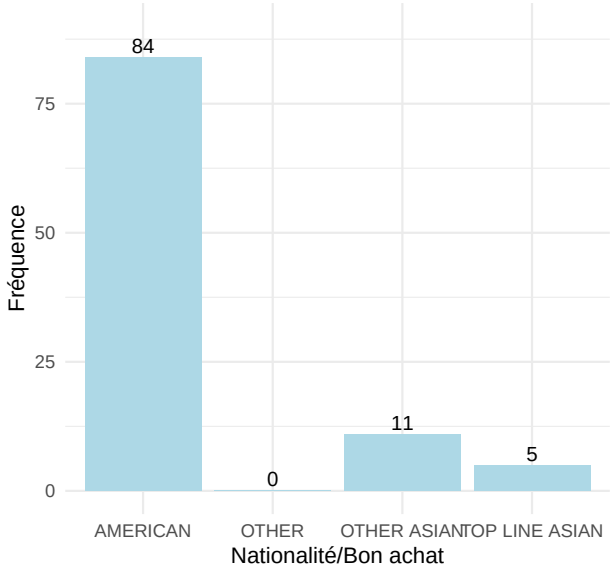


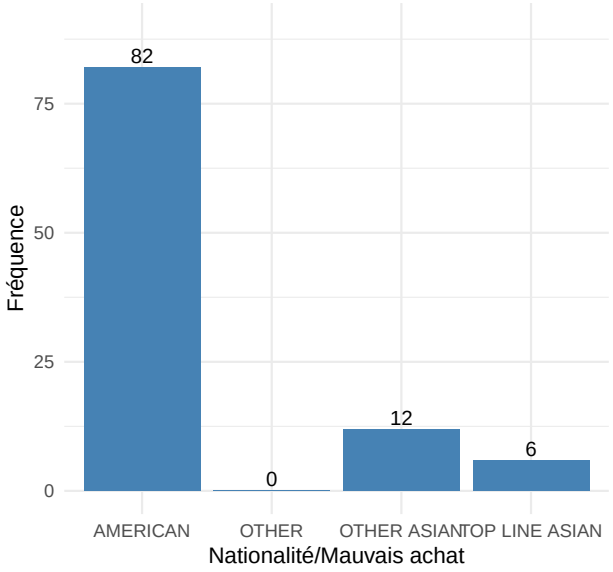


Facteurs de différenciation absents

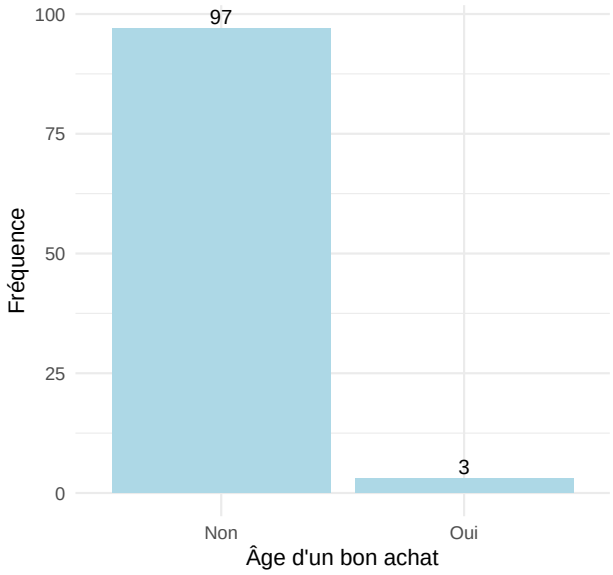
Les variables suivantes ne différencient pas les deux groupes, ils sont clairement répartis de la même façon. Ce ne sont pas des variables pertinentes pour notre analyse.

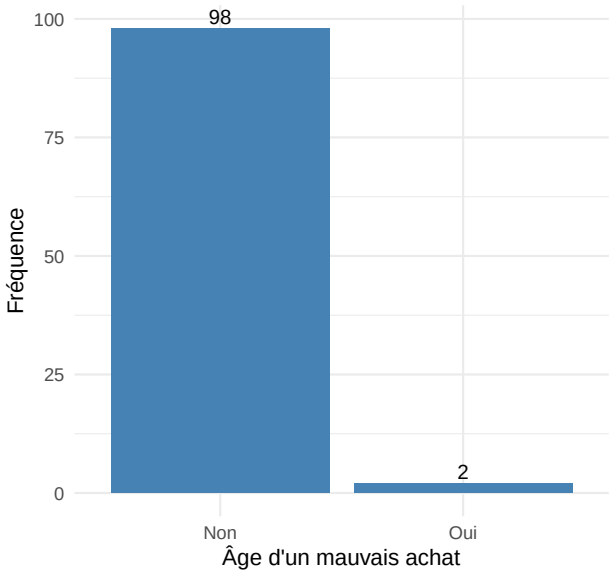
En premier la nationalité du constructeur du véhicule n'a pas vraiment de différence dans la répartition pour les deux groupes, on peut mettre en lien avec cela le fait que les différences de répartition se font uniquement entre les fabricants américains.

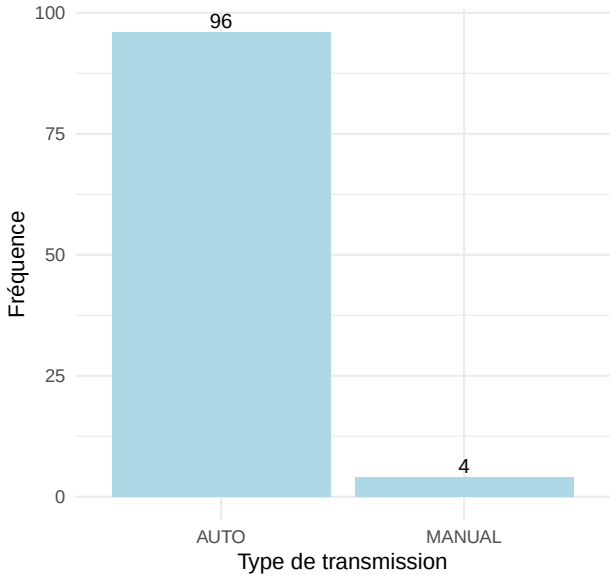


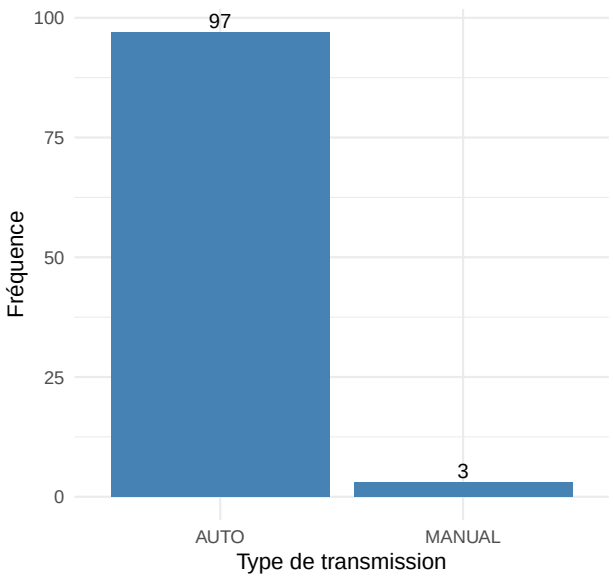


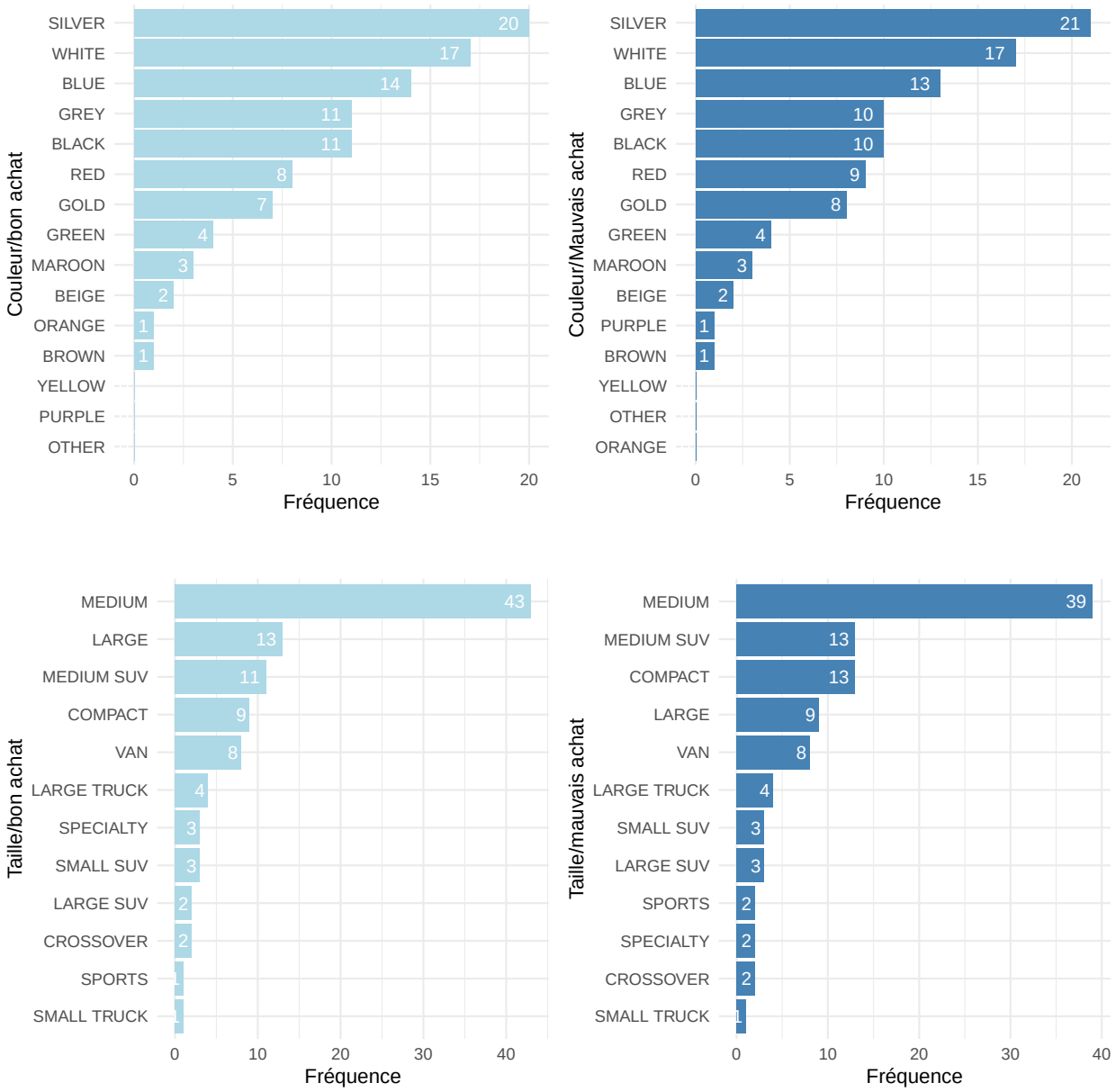
Il semblerait que les proportions de ventes en ligne chez les deux groupes soient égales. Ainsi que le type de transmission.


100
75
50
25
0
Fréquence
97
3
Non
Oui
Âge d'un bon achat


100
75
50
25
0
Fréquence
98
2
Non
Oui
Âge d'un mauvais achat


100
75
50
25
0
Fréquence
96
4
AUTO
MANUAL
Type de transmission


100
75
50
25
0
Fréquence
97
3
AUTO
MANUAL
Type de transmission



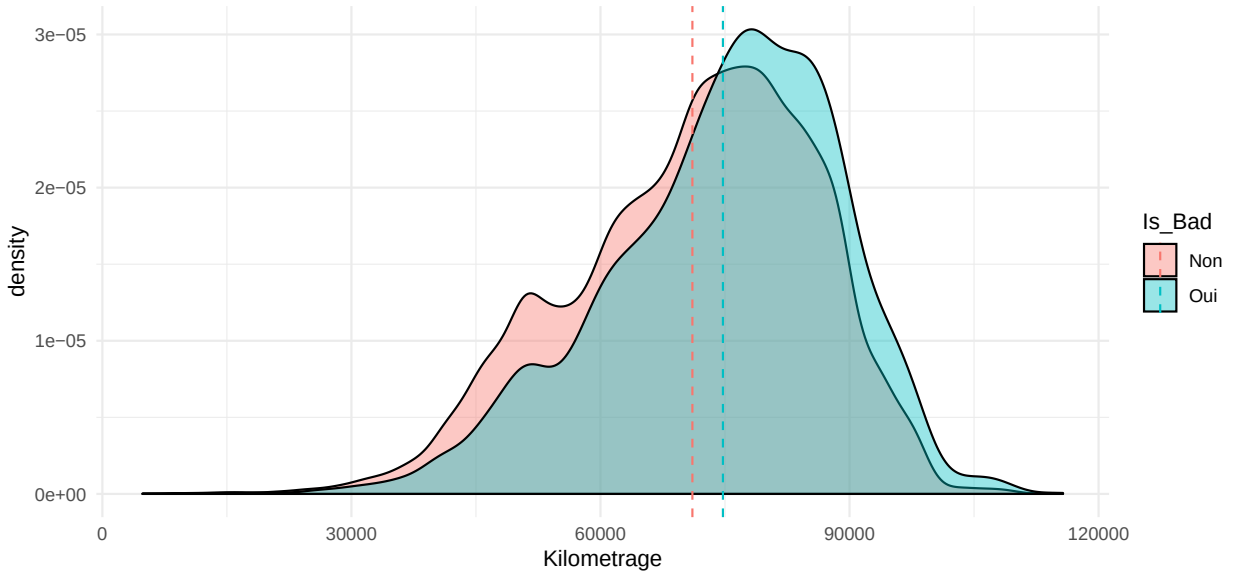
1.1.2 Quantitatives

Pour construire nos modèles prédictifs nous ne pourrons pas utiliser le prix de vente du véhicule aux enchères car c'est une donnée que nous ne possédons pas normalement à l'avance. Pour palier à ce manque, nous avons décidé de créer des mesures qui pourraient potentiellement aider notre modèle à être plus précis.

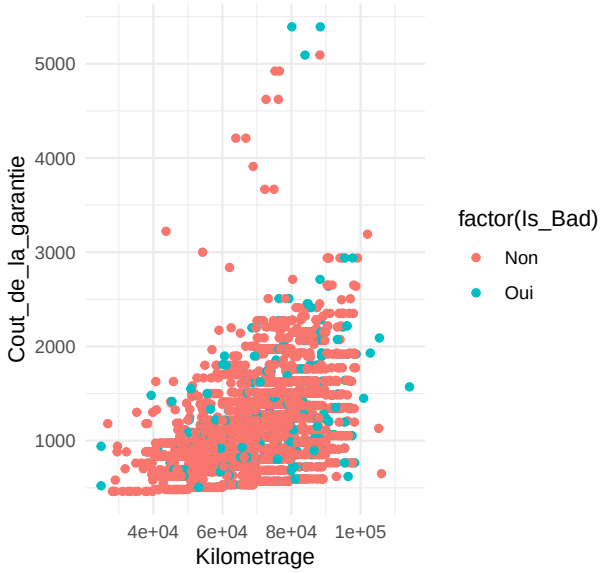
- Comparaison des variations entre les cotes aux enchères et les cotes aux détaillants pour tous les états des véhicules
- Calcul des mesures : $Mesures = \frac{CoteDetaillant_{EtatVeh} - CoteEncheres_{EtatVeh}}{CoteEncheres} * 100$
- Plus le pourcentage est élevé plus l'acheteur peut espérer gagner de l'argent sur la revente du véhicule par rapport à la valeur estimée aux enchères

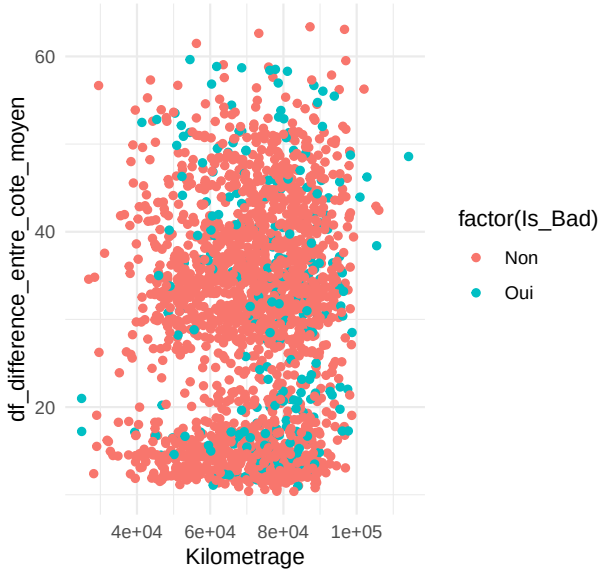
Ici toutes les variables différencient les deux groupes. Cela reste néanmoins une différenciation très faible. On a comparé les densités relatives à chaque groupe afin d'avoir des premiers présentiments sur nos futurs modèles et résultats.

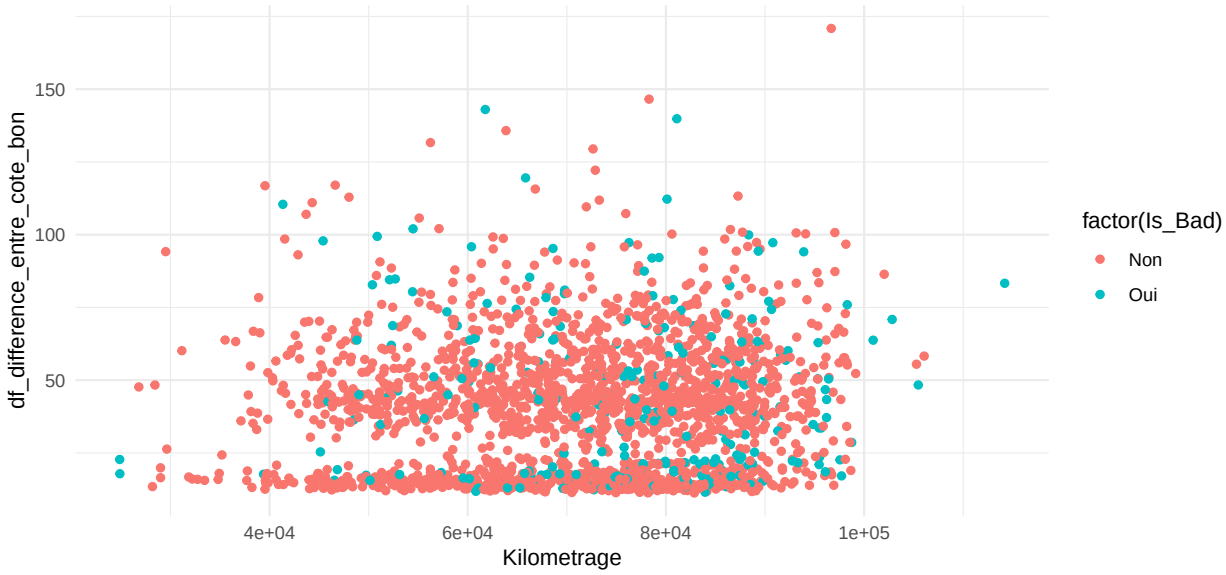
Au premier semestre nous avions déjà étudié la différence de kilométrage entre les deux groupes et avions conclu que les mauvais achats avaient un kilometrage relativement plus élevé que les bons achats.

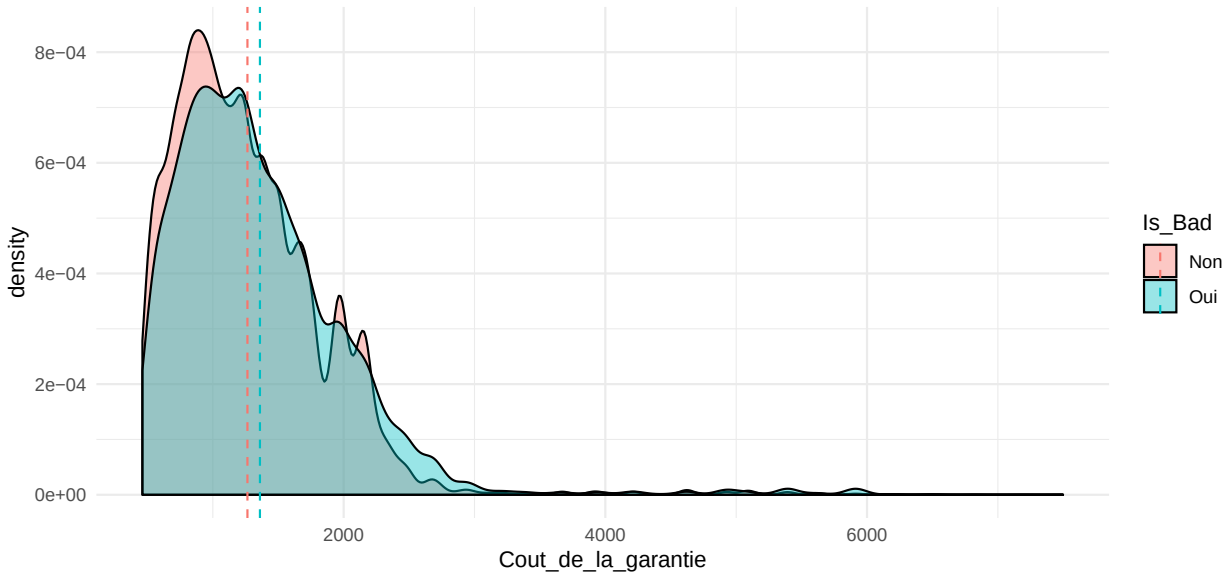


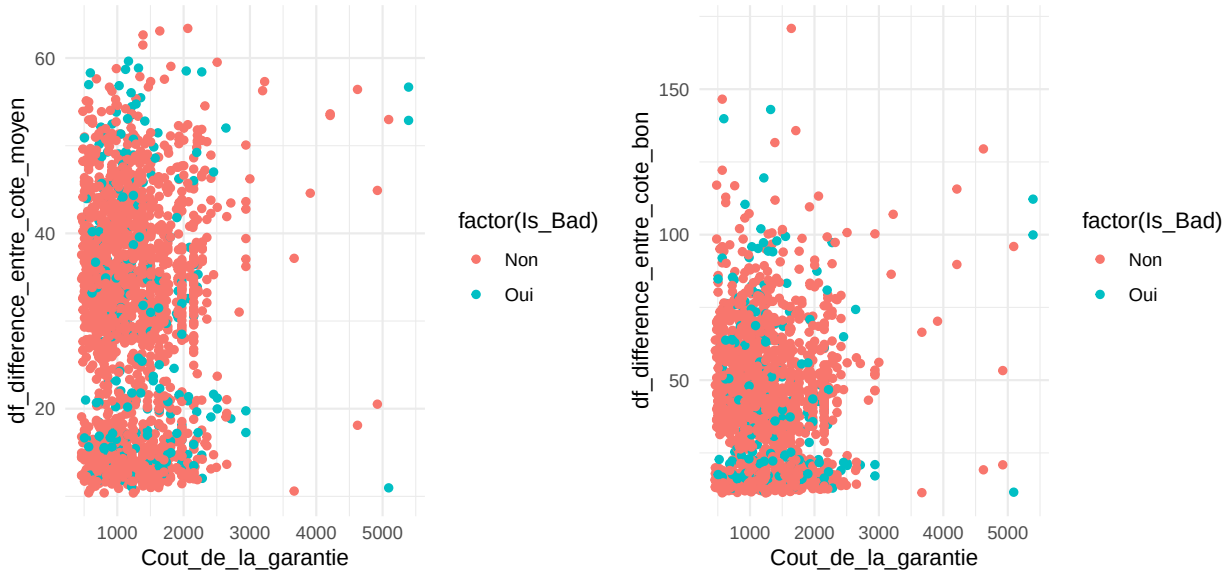
Quand on regarde les nuages de points du kilometrage en fonction des autres quantitatives on peut voir que les deux groupes sont très similairement répartis.



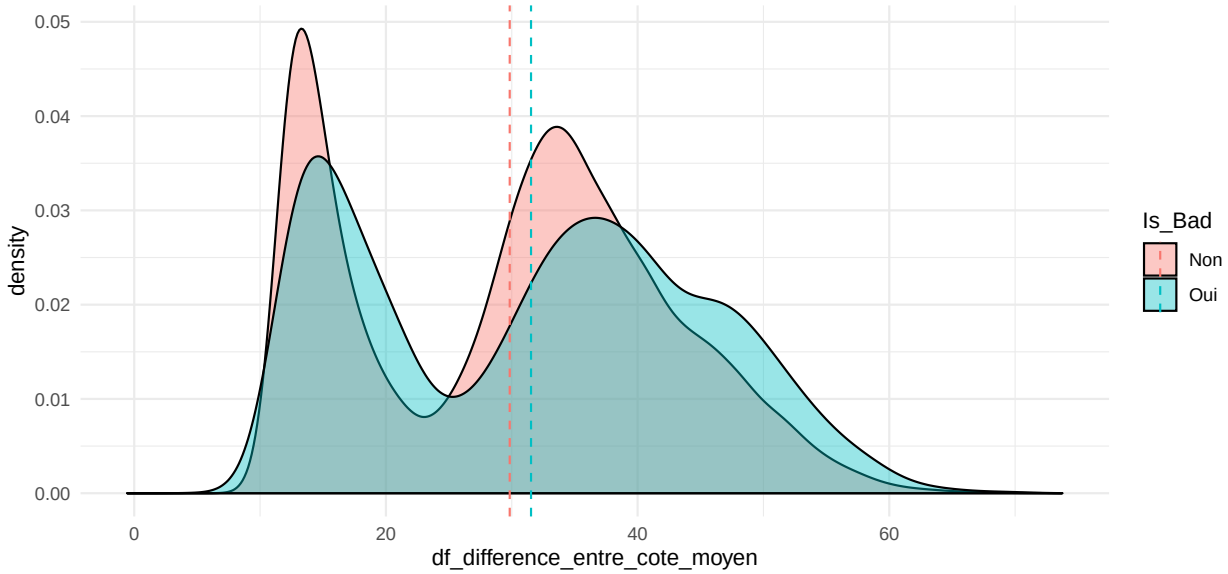


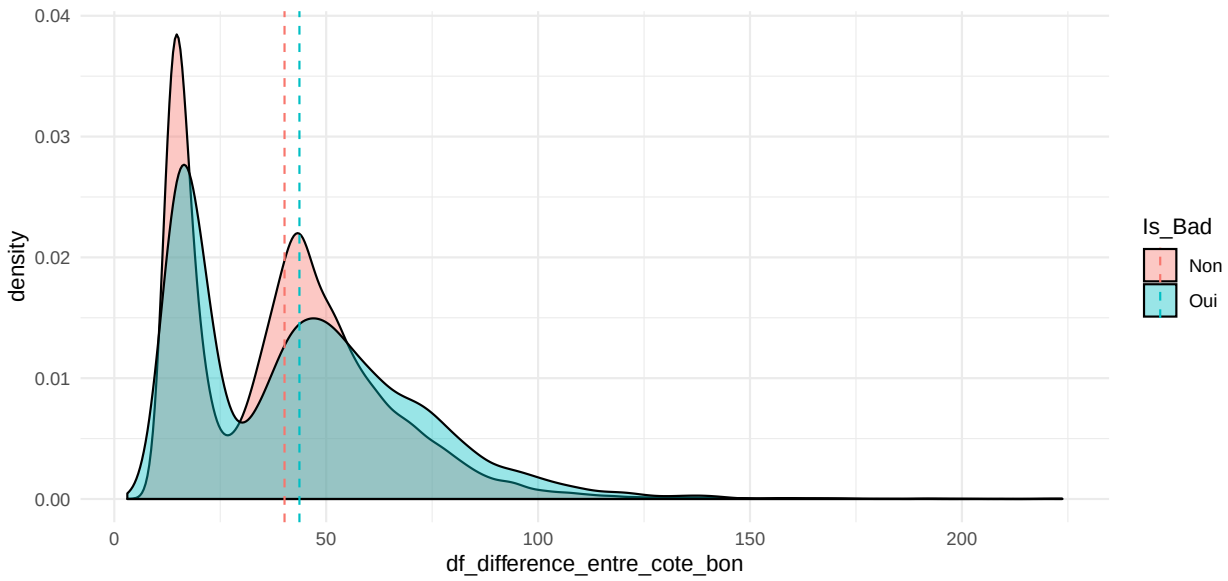

df_difference_entre_cote_bon
150
100
50
4e+04
6e+04
8e+04
1e+05
Kilometrage
factor(Is_Bad)
Non
Oui


density
8e−04
6e−04
4e−04
2e−04
0e+00
2000
4000
6000
Cout_de_la_garantie
Is_Bad
Non
Oui



Les deux graphiques suivants donnent des conclusions particulières, on peut voir que les véhicules mauvais achats, peu importe l'état ont une différence entre les cotes au détaillant et les cotes aux enchères plus grande que les bons achats. Ce qui est contre intuitif. On avait vu au premier semestre que les cotes aux enchères pour les véhicules mauvais achats sont moins élevées ce qui cause peut être cette différence.





En observant ces différents nuages de points et densités on peut déjà imaginer que les modèles de prédiction vont avoir du mal à repérer les différents groupes. On a déjà l'intuition que les analyses linéaires et quadratiques discriminantes ainsi que les k-plus proches voisins vont avoir du mal à bien prédire le groupe le moins représenté vu qu'il se confond avec le groupe le plus représenté.

Avant de commencer à construire nos modèles, il convient d'avoir une réflexion sur ce que l'on cherche. En effet, avec des problèmes de sous-représentation des mauvais achats et le fait que les deux classes ne semblent pas avoir de véritables facteurs discriminants, il semble compliqué de correctement prédire les mauvais achats. En revanche, nous allons essayer de construire des modèles qui prédisent le plus précisement possible les bons achats en essayant de minimiser l'erreur d'estimation globale et le classement de mauvais achats en bons achats.

2 Analyse linéaire discriminante

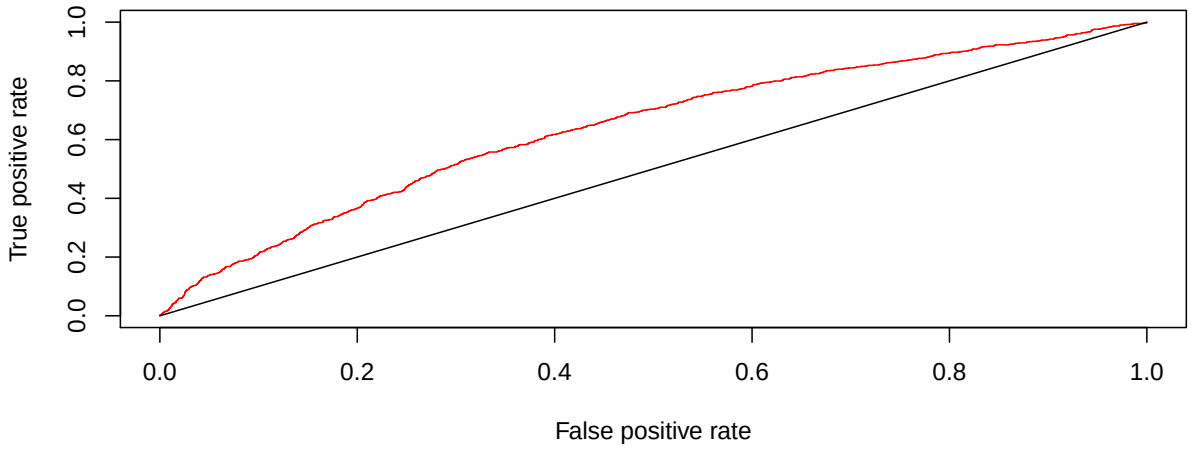
Table 1: Matrice de confusion

Réalité	Prediction		
	Non	Oui	Sum
Non	6299	0	6299
Oui	918	0	918
Sum	7217	0	7217

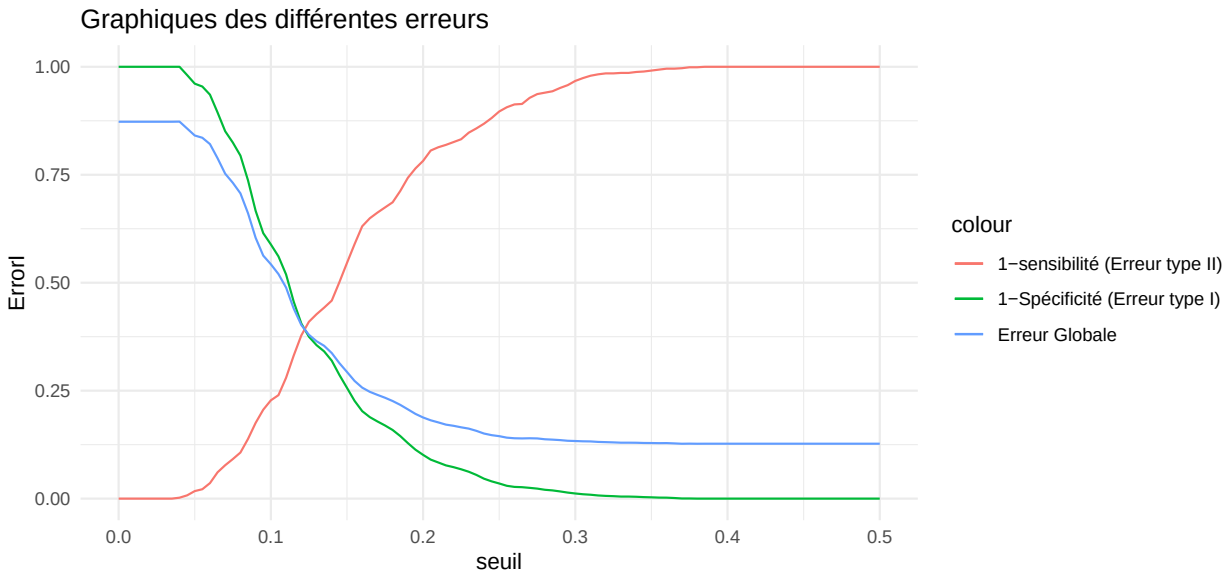
Table 2: Proportion en ligne de la Matrice de confusion

Réalité	Prediction	
	Non	Oui
Non	100	0
Oui	100	0

Après réalisation d'une LDA sur les variables quantitatives kilométrage, Cout de la garantie et Âge du véhicule, nous observons que le modèle n'est pas précis avec 0 mauvais achats de véhicule prédits. En sachant que l'intérêt d'une prédiction sur nos données est de prédire si un véhicule est un mauvais achat nous réalisons la courbe Roc dans l'espoir d'améliorer le modèle.



Malheureusement la courbe Roc, bien que légèrement concave est très proche de la bissectrice cela nous indique qu'il sera donc difficile de prédire correctement si un véhicule est un mauvais achat ou non.



Nous voulons diminuer l'erreur sur la sensibilité donc pour ce faire nous devons baisser le seuil de décision du modèle. Après analyse du graphique ci-dessus nous choisissons le seuil optimal de 0.14 qui permet de diminuer l'erreur de sensibilité tout en conservant une erreur globale et une erreur de spécificité correcte. Nous réalisons donc une LDA avec ce nouveau seuil.

Table 3: Matrice de confusion

Réalité	Prediction		
	No	Yes	Sum
Non	4287	2012	6299
Oui	421	497	918
Sum	4708	2509	7217

Table 4: Proportion en ligne de la Matrice de confusion

Réalité	Prediction	
	No	Yes
Non	68.1	31.9
Oui	45.9	54.1

```
## [1] 33.71207
```

En utilisant le nouveau seuil de décision on trouve un modèle prédictif bien plus sensible. En effet, la sensibilité est passé de 0 à 54.1 donc la chance de prédire un bon achat à la place d'un mauvais achat a nettement diminué. Cette diminution se fait au détriment de la spécificité ,même si nous sommes moins intéressés par cette dernière. On peut relever qu'elle passe de 100 à 68.1.

Bien que ce modèle ne soit pas très précis, c'est le mieux que nous pouvons faire avec une lda. Cela nous indique, comme le montre les nuages de points, que la discrimination linéaire n'est pas adaptée et que l'obtention de groupes ne se fera pas au travers d'une LDA.

3 Analyse quadratique discriminante

Nous avons réalisé une QDA de la variable `Is_Bad` sur les variables `Age_du_Vehicules`, `Cout_de_la_garantie`, `kiliometrage` et les differences entre les côtes.

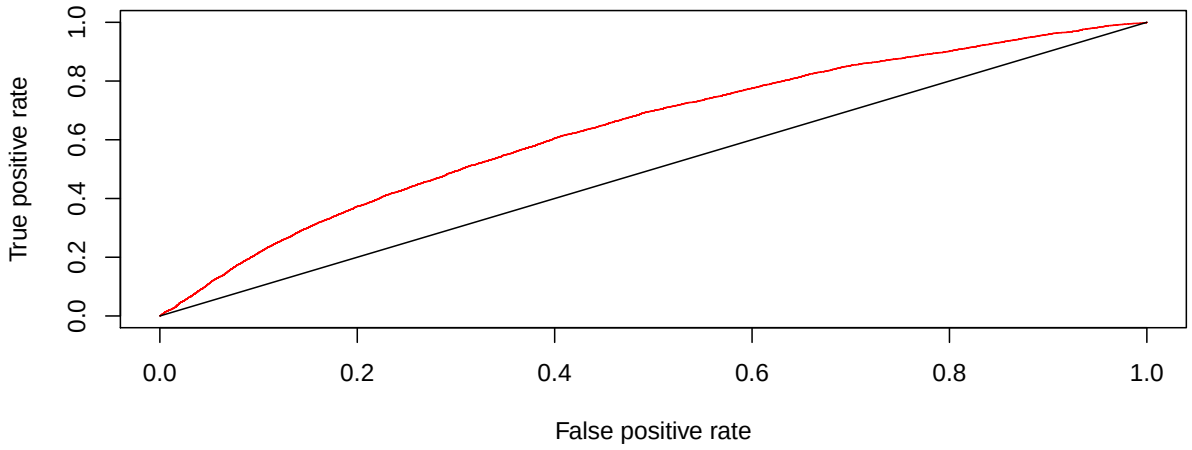
Table 5: Matrice de confusion

Réalité	Prediction		
	Non	Oui	Sum
Non	41141	1062	42203
Oui	5592	315	5907
Sum	46733	1377	48110

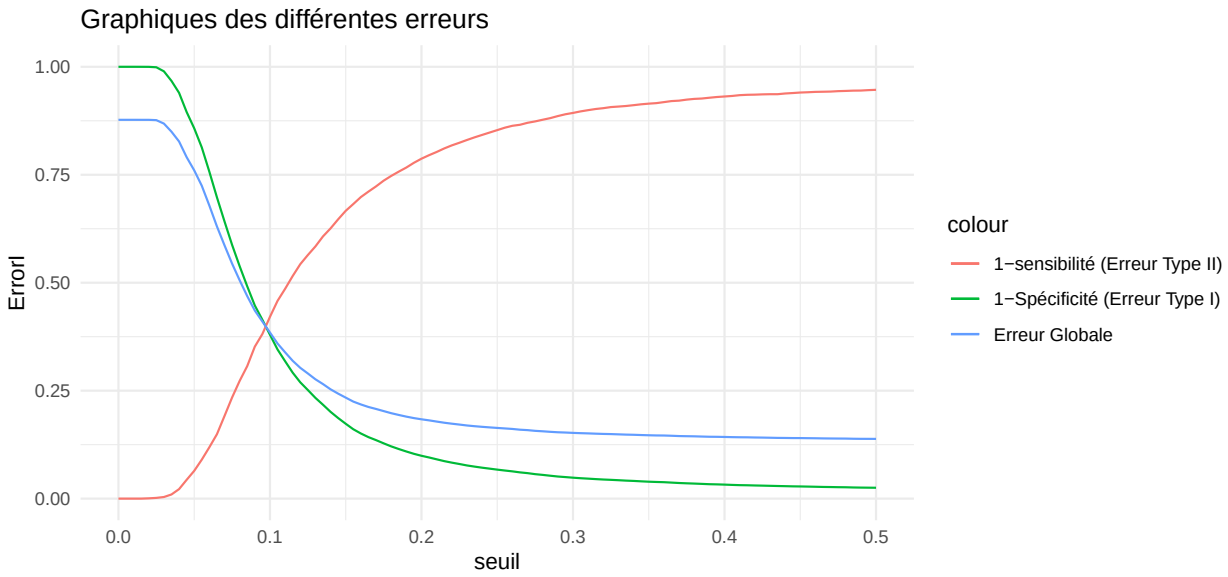
Table 6: Proportion en ligne de la Matrice de confusion

Réalité	Prediction	
	Non	Oui
Non	97.5	2.5
Oui	94.7	5.3

Le modèle prédit très mal les mauvais achats avec seulement 315 prédits correctement, cependant, le modèle est meilleur que la LDA. En revanche, la prédiction des bons achats est bonne. L'erreur globale du modèle est de 13.83%. L'intéret de notre modèle étant de prédire les voitures en mauvais état pincipalement, nous réalisons la courbe ROC.



Cette courbe ROC est assez proche de la bisséctrice. Elle nous montre finalement qu'en cherchant à augmenter le taux de vrai positifs (sensibilité), le taux de faux positifs augmentera dans des proportions quasi similaires, ce qui n'améliore pas le modèle significativement. Le meilleur résultat à obtenir est un taux de vrai positifs d'environ 60% mais le taux de faux positifs sera de 40% environ, ce qui est élevé.



Ce graphique nous permet de voir qu'une erreur globale faible entraine forcément un sensibilité faible. Or on cherche à avoir une sensibilité plutot élevée car il est important de ne pas prédire une voiture comme bon achat alors que c'est un mauvais achat. Cependant, si on veut augmenter cette sensibilité, l'erreur globale et la spécificité diminuent forcément.

On décide donc d'abaisser le seuil à 0.15 afin de faire augmenter la sensibilité tout en gardant une erreur globale raisonnable. On testera ce nouveau modèle sur les données tests.

Table 7: Matrice de confusion

Réalité	Prediction		
	No	Yes	Sum
Non	17454	3629	21083
Oui	2029	942	2971
Sum	19483	4571	24054

Table 8: Proportion en ligne de la Matrice de confusion

Réalité	Prediction	
	No	Yes
Non	82.8	17.2
Oui	68.3	31.7

Cet ajustement du seuil nous permet de prédire beaucoup plus de mauvais achats. La sensitivité passe de 5.3% à 31.7%. Ainsi, l'erreur de type II, c'est à dire le taux de voitures predites comme bon achats alors qu'elles sont de mauvais achats en réalité a diminué mais reste élevé (68.3%). De plus, on perd de la précision dans la prédiction des bon achats avec seulement 82.8% de bonnes predictions sur les bons achats contre 97.5% sur le modèle précédent. De plus, notre erreur globale augmente et passe à 11.7605487%.

En conclusion, cette QDA prédit assez mal si une voiture est un mauvais achat ou non. Il est difficille de connaitre en particulier les mauvais achats et lorsqu'on améliore cette prediction, l'erreur globale de prediction augmente car on predit mal les bons achats.

4 Les k-plus proches voisins

Pour réaliser notre modèle KNN nous avons séléctionné comme précedemment les variables quantitatives qui différencient le plus nos deux groupes, nous selectionnerons donc comme précédemment pour la LDA et la QDA:

- le coût du véhicule
- le kilométrage
- le coût de la garantie
- la différence entre les différentes cotes au moment de l'achat et le coût du véhicule

4.1 Un découpage

Nous testons d'abord avec un seul découpage aléatoire de nos données (66,7% données d'entrainement, 33,3% données de test). Nous décidons d'un nombre classique de 3 voisins.

Table 9: Matrice de confusion

Erreur globale	**Prédiction**		
	10.74		
	Non	Oui	Sum
Non	6210	89	6299
Oui	686	232	918
Sum	6896	321	7217
Non	98.59	1.41	
Oui	74.73	25.27	

L'erreur globale est de 10.74%, les bons achats sont très bien estimés, par contre les mauvais achats sont très souvent classés bons achats. Ce qui est la pire prédiction possible.

Pour améliorer le modèle la seule chose que l'on peut faire est de modifier le nombre de voisins, nous allons donc regarder les performances du modèle en fonction du nombre de voisins.

4.2 Tuning des knn

On tune maintenant les knn, en utilisant des échantillons générés par validation croisée, pour décider le nombre de voisins (entre 1 et 5) qui diminue le plus l'erreur globale.

Table 10: Choix du meilleur K

Erreur globale minimale	**13.91**	
K	Erreur globale	Dispersion
1	21.75	0.96
2	21.62	1.05
3	15.92	0.82
4	15.92	0.84
5	13.91	0.82

L'erreur globale est la plus faible avec 5 voisins, ainsi que la dispersion de l'erreur. Nous choisirons donc 5 voisins.

4.3 Validation croisée

Nous avons décidé de tester les knn sur 100 échantillons générés par validation croisée afin d'avoir potentiellement une meilleure erreur. Nous avons aussi testé les knn sur des échantillons bootstrap mais l'erreur étant

bien plus élevée par souci d'optimisation des capacités de calcul de l'ordinateur nous avons décidé de ne pas les inclure.

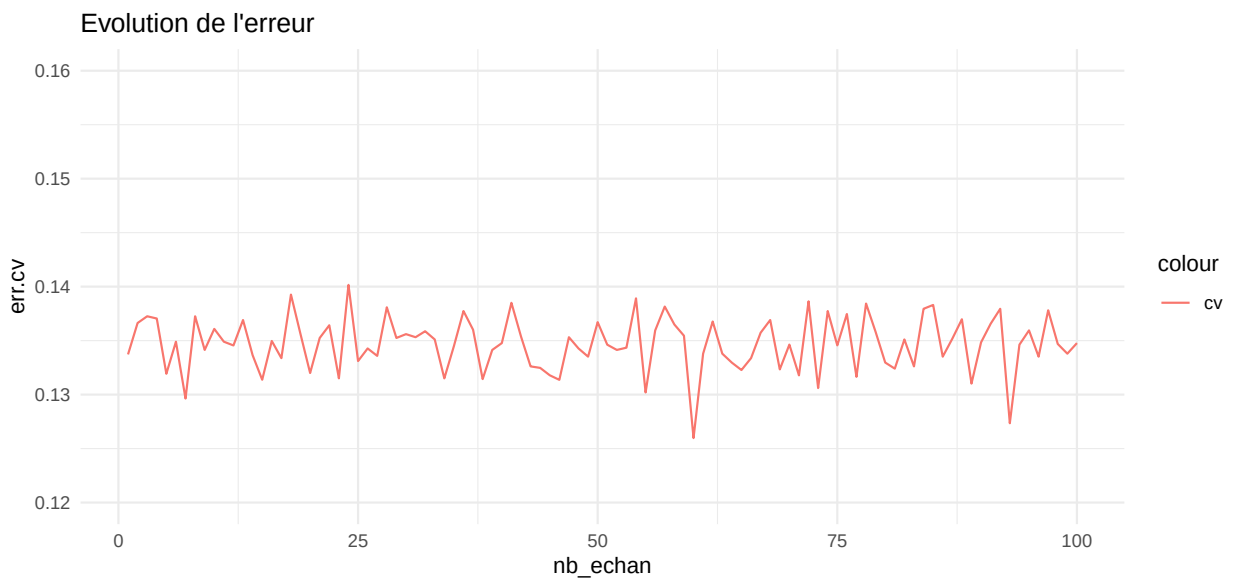


Table 11: Matrice de confusion avec validation croisée

Erreur globale	**Prédiction**		
	13.48		
	Non	Oui	Sum
Non	12442	202	12644
Oui	1743	45	1788
Sum	14185	247	14432
Non	98.4	1.6	
Oui	97.48	2.52	

On voit que l'erreur globale est stable sur la centaine d'échantillons générés par validation croisée, la variance de l'erreur est donc faible ainsi que le biais(bénéfice des échantillons croisés), le modèle n'en reste pas moins mauvais avec une erreur globale faible mais presque aucun mauvais véhicule repéré.

Nous pourrions faire le même commentaire qu'avec la LDA et la QDA, si on diminue le nombre de voisins, le modèle est légèrement plus spécifique mais beaucoup moins sensible, la précision du modèle en est grandement affectée. A l'inverse, si on cherche à optimiser l'erreur globale, la spécificité est très mauvaise.

5 Arbres de décisions

Réalisons les arbres de décisions sur les variables quantitatives en rajoutant certaines qualitatives intéressantes telles que les types de roues, les enchères et les constructeurs.

On essaie de trouver le meilleur élaguage avec l'arbre suivant :

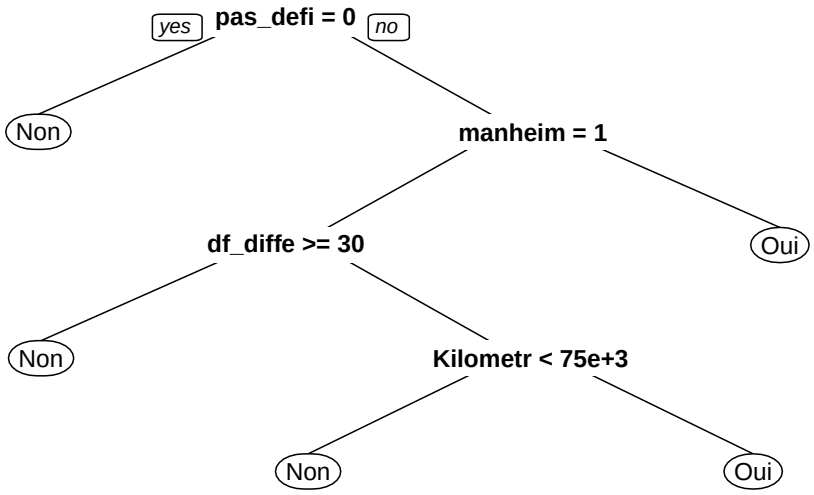
Table 12: Matrice de confusion

Prediction	Réalité		
	Non	Oui	Sum
Non	6247	721	6968
Oui	52	197	249
Sum	6299	918	7217

Table 13: Proportion en ligne de la Matrice de confusion

Prediction	Réalité	
	Non	Oui
Non	99.2	78.5
Oui	0.8	21.5

	x
Sensitivity	0.2145969
Specificity	0.9917447
Pos Pred Value	0.7911647
Neg Pred Value	0.8965270
Precision	0.7911647
Recall	0.2145969
F1	0.3376178
Prevalence	0.1271997
Detection Rate	0.0272967
Detection Prevalence	0.0345019
Balanced Accuracy	0.6031708



On a un arbre très court avec une erreur globale la plus faible qu'on ait obtenue jusqu'ici, qui repère beaucoup mieux les mauvais achats

On cherche maintenant l'arbre qui maximise l'aire sous la courbe ROC.

	cp
14	0.0008715

Table 14: Matrice de confusion

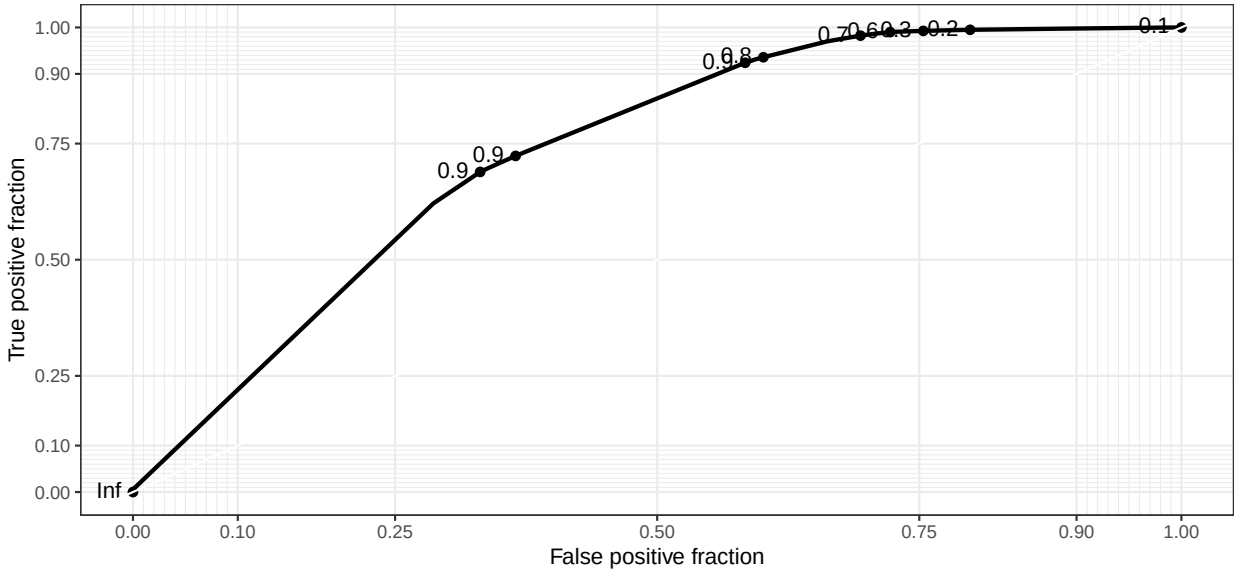
Prediction	Réalité		
	Non	Oui	Sum
Non	6236	663	6899
Oui	63	255	318
Sum	6299	918	7217

	x
Sensitivity	0.2777778
Specificity	0.9899984
Pos Pred Value	0.8018868
Neg Pred Value	0.9038991
Precision	0.8018868
Recall	0.2777778
F1	0.4126214
Prevalence	0.1271997
Detection Rate	0.0353332
Detection Prevalence	0.0440626
Balanced Accuracy	0.6338881

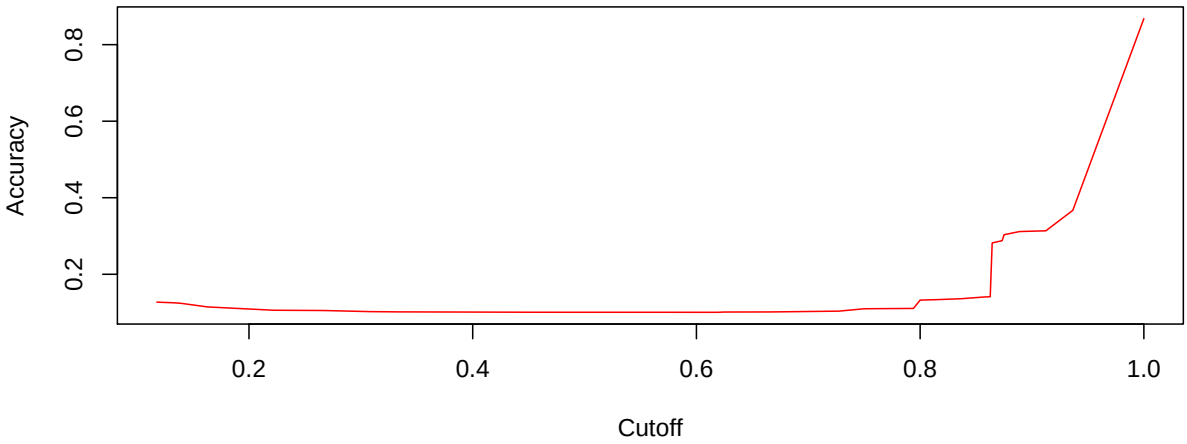
```
##  Accuracy
## 0.8994042
```

Les résultats de cet arbre sont plus satisfaisants que l'arbre précédent. En effet, la sensibilité passe à 27.8 et le taux de voitures prédites comme bons achats alors qu'elles sont de mauvais achats en réalité diminue par

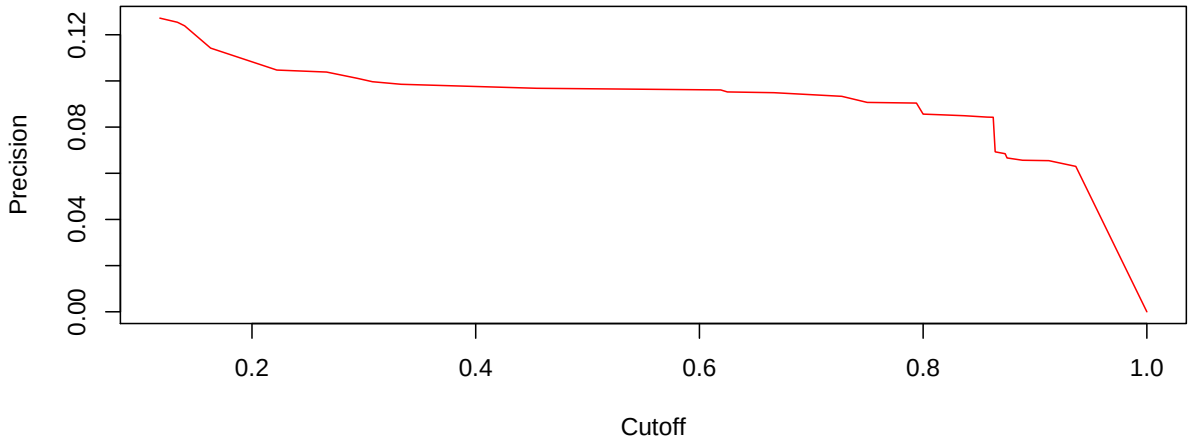
la même occasion en passant à 72.2% mais reste élevé tout de même.



La courbe ROC est légérement améliorée mais on voit bien que la prédiction ne sera pas très bonne tout de même.



L'erreur globale augmente avec la taille de l'arbre. Ceci ne change pas par rapport à l'arbre précédent.



Comme l'arbre précedent, la précision diminue jusqu'à un seuil puis augmente brusquement lorsque l'arbre est complet. Par conséquent, il est raisonnable de prendre un arbre élagué suffisement grand mais pas trop non plus afin d'avoir une erreur globale raisonnable et une bonne précision.

En conclusion, cette arbre prédit beaucoup mieux les mauvais achats que l'arbre précédent mais il les prédit tout de même assez mal, en témoigne la sensibilité de 27.8. De ce fait, on choisira plus cet arbre afin de prédire la variable `Is_Bad`.

5.0.1 Sous-échantillonage

On sait que nous avons un problème de tailles d'échantillons entre nos deux groupes : `prop[1]`% de nos données sont classées "Non" et `prop[2]`% sont classées "Oui".

Nous avons pu observer que nos précédents modèles ont du mal à repérer les mauvais achats. Nous allons utiliser des techniques de sous-échantillonage afin de rééquilibrer la représentation de nos classes et ainsi potentiellement mieux repérer les mauvais achats.

Table 15: Matrice de confusion

Prediction	Réalité		
	Non	Oui	Sum
Non	4560	294	4854
Oui	1739	624	2363
Sum	6299	918	7217

Table 16: Proportion en ligne de la Matrice de confusion

Prediction	Réalité	
	Non	Oui
Non	72.4	32
Oui	27.6	68

	x
Sensitivity	0.6797386
Specificity	0.7239244
Pos Pred Value	0.2640711
Neg Pred Value	0.9394314
Precision	0.2640711
Recall	0.6797386
F1	0.3803718
Prevalence	0.1271997
Detection Rate	0.0864625
Detection Prevalence	0.3274214
Balanced Accuracy	0.7018315

	cp
19	0.0029049

L'erreur globale est plus importante que les arbre précédents, par contre on repère beaucoup mieux les mauvais achats, la sensibilité étant la meilleure qu'on ait obtenu jusqu'ici. De plus on ne baisse pas trop la spécificité et l'erreur globale.

5.0.2 Sur-échantillonage

On va maintenant faire l'inverse et rééchantilloner les mauvais véhicules afin d'avoir la même taille que l'échantillon des bons véhicules.

	cp
19	0.0029049

Table 17: Matrice de confusion

Prediction	Réalité		
	Non	Oui	Sum
Non	4903	340	5243
Oui	1396	578	1974
Sum	6299	918	7217

Table 18: Proportion en ligne de la Matrice de confusion

Prediction	Réalité	
	Non	Oui
Non	77.8	37
Oui	22.2	63

	x
Sensitivity	0.6296296
Specificity	0.7783775
Pos Pred Value	0.2928065
Neg Pred Value	0.9351516
Precision	0.2928065
Recall	0.6296296
F1	0.3997234
Prevalence	0.1271997
Detection Rate	0.0800887
Detection Prevalence	0.2735209
Balanced Accuracy	0.7040036

5.0.3 Algorithme Smote

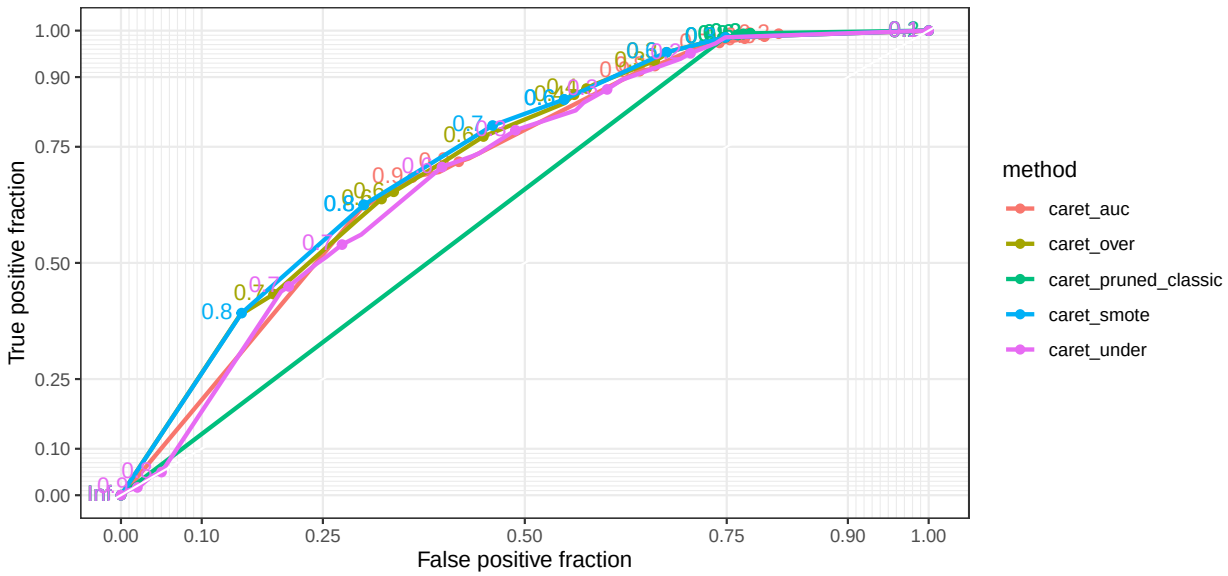
Essayons avec un modèle de réequilibrage d'échantillon un peu plus complexe dans sa réalisation et potentiellement meilleur.

	x
Sensitivity	0.3244975
Specificity	0.9533402
Pos Pred Value	0.4927509
Neg Pred Value	0.9099405
Precision	0.4927509
Recall	0.3244975
F1	0.3913043
Prevalence	0.1225615
Detection Rate	0.0397709
Detection Prevalence	0.0807120
Balanced Accuracy	0.6389189

L'erreur globale est plus faible que dans les deux modèles de réechantillonage précédents mais on ne repère pas assez bien les mauvais véhicules.

5.0.4 Comparaison des différents arbres

	arbre_elag	auc_mod	under_echan	over_echan	arbre_smote
Sensitivity	21.46	27.78	67.97	62.96	32.45
Specificity	99.17	99.00	72.39	77.84	95.33
Pos Pred Value	79.12	80.19	26.41	29.28	49.28
Neg Pred Value	89.65	90.39	93.94	93.52	90.99
Precision	79.12	80.19	26.41	29.28	49.28
Recall	21.46	27.78	67.97	62.96	32.45
F1	33.76	41.26	38.04	39.97	39.13
Prevalence	12.72	12.72	12.72	12.72	12.26
Detection Rate	2.73	3.53	8.65	8.01	3.98
Detection Prevalence	3.45	4.41	32.74	27.35	8.07
Balanced Accuracy	60.32	63.39	70.18	70.40	63.89
error_glob	10.71	10.06	28.17	24.05	12.37



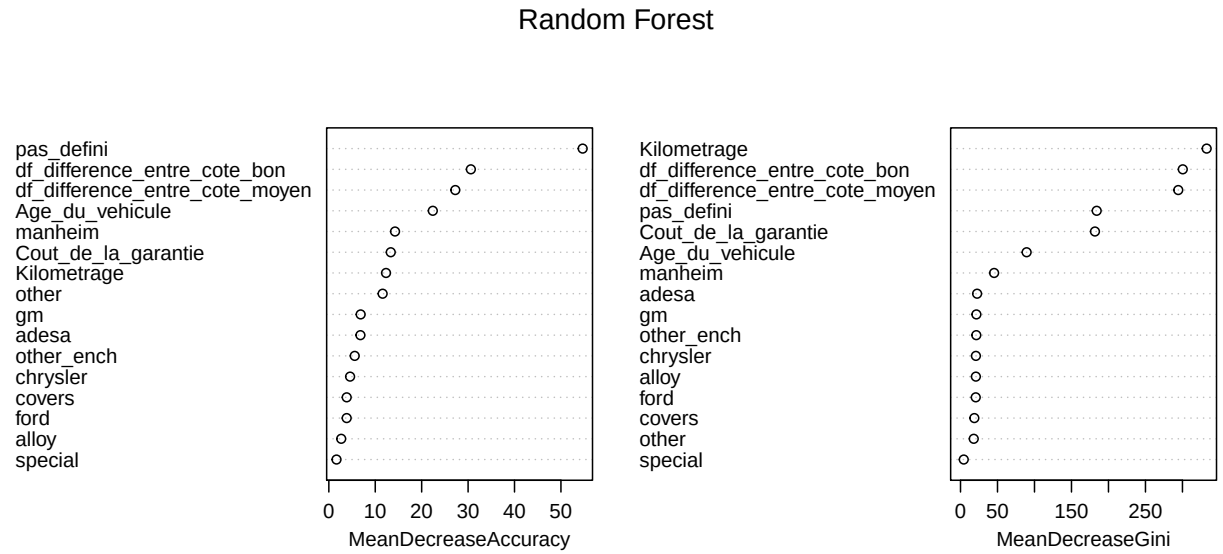

```
## $caret_pruned_classic
##         y_pred
## y_true     0    1
##   FALSE  197  721
##   TRUE    52 6247
##
## $caret_auc
##         y_pred
## y_true     0    1
##   FALSE  255  663
##   TRUE    63 6236
##
## $caret_under
##         y_pred
## y_true     0    1
##   FALSE  624  294
##   TRUE  1739 4560
##
## $caret_over
##         y_pred
## y_true     0    1
##   FALSE  578  340
##   TRUE  1396 4903
##
## $caret_smote
##         y_pred
## y_true     0    1
##   FALSE  281  637
##   TRUE   289 6010
```

5.1 Forêts aléatoires

On décide donc de faire une forêt aléatoire sur les mêmes variables afin d'améliorer le modèle.

Après analyse nous observons que l'erreur OBB ce stabilise aux alentours de 150 arbres. Nous choisissons donc

ce seuil pour réaliser note forêt aléatoire. De plus après avoir réalisé plusieurs test nous fixons le paramètre MTRY à 10.



La variable `pas_defini` est celle qui est le plus importante pour la random forest. Bien que nous n'ayons aucune information sur cette modalité, on peut émettre l'hypothèse qu'un véhicule ayant des roues à nues est un véhicule à ne pas acheter. On peut aussi souligné l'importance des variables difference entre les côtes et coût du véhicules.

Table 19: Matrice de confusion

Prediction	Réalité		
	Non	Oui	Sum
Non	56252	6097	62349
Oui	735	1863	2598
Sum	56987	7960	64947

	x
Sensitivity	0.2340452
Specificity	0.9871023
Pos Pred Value	0.7170901
Neg Pred Value	0.9022117
Precision	0.7170901
Recall	0.2340452
F1	0.3529077
Prevalence	0.1225615
Detection Rate	0.0286849
Detection Prevalence	0.0400018
Balanced Accuracy	0.6105738

```
##  Accuracy
## 0.8948065
```

On note une lègere baisse de la sensibilité par rapport aux arbres contrebalancé par une légère augmentation

de la spécificité. Ceci n'est pas très interressant pour notre prédiction bien que l'erreur globale diminue.

6 Conclusion

Dans l'optique de notre analyse, des compromis sont nécessaires. Les premiers modèles LDA, QDA et KNN ne sont absolument pas adaptés à nos besoins. Ils renvoient chacun une erreur globale acceptable mais ils ne repèrent pas correctement les mauvais achats. Quand on optimise le seuil afin de "mieux" séparer les groupes l'erreur globale en patit trop.

Les modèles d'arbres générés sont beaucoup plus adaptés. Nous pouvons rajouter les variables qualitatives et ainsi gagner de l'information. L'arbre élagué optimal génère une erreur globale la plus faible. Elle repère très bien les bons achats et améliore le classement des mauvais achats. L'arbre élagué qui maximise l'aire sous la courbe ROC a lui aussi une erreur globale faible, moins bonne que l'arbre élagué optimal. Il classe mieux les mauvais achats(sensibilité égale à 0,25 contre 0.22), il est donc préféré à l'arbre élagué qui diminue le plus l'erreur globale.

L'arbre qui pratique du sur-échantillonage renvoie une erreur globale plus élevée mais aussi une bien meilleure sensibilité. C'est le modèle qui classe le moins de mauvais achats en bons achats, et celui qui classe le mieux les mauvais achats. Il classe plus de bons achats en mauvais achats.

L'arbre qui pratique du sous-échantillonage renvoie une erreur globale plus élevée que le sur-échantillonage mais aussi une sensibilité plus élevée. C'est le modèle qui classe le mieux les mauvais achats parmis tous les modèles précédents.

La forêt aléatoire de 150 arbres génére une erreur globale très faible, mais aussi une sensibilité moins elevée, elle est moins intéressante dans l'optique de notre analyse.

Nous avons cherché tout au long de notre analyse à réduire au maximum le classement des mauvais achats en bons achats. Il faut maintenant pratiquer un arbitrage. Quel est le meilleur modèle à choisir qui prédirait au mieux les mauvais achats sans trop mal prédire les bons achats? Dans cette optique nous sommes prêts à sacrifier l'erreur globale et pencherions davantage vers du sous-échantillonage ou du sur-échantillonage. Parmis ces deux modèles, nous prenons aussi en compte la taille et la complexité de l'arbre qui sont relativement similaires pour les deux arbres.

Par souci d'optimisation de la sensibilité nous choisirons tout de même l'arbre qui sous-échantillone au détriment des % d'erreur globale perdus, mais nous préférons estimer trop de mauvais véhicules qui sont de bons véhicules que l'inverse.

Appliquons notre modèle final sur les données tests afin de voir sa performance.

Table 20: Matrice de confusion

Prediction	Réalité		
	Non	Oui	Sum
Non	40255	3164	43419
Oui	16732	4796	21528
Sum	56987	7960	64947

Table 21: Proportion en ligne de la Matrice de confusion

Prediction	Réalité	
	Non	Oui
Non	70.6	39.7
Oui	29.4	60.3

	x
Sensitivity	0.6025126
Specificity	0.7063892
Pos Pred Value	0.2227796
Neg Pred Value	0.9271287
Precision	0.2227796
Recall	0.6025126
F1	0.3252849
Prevalence	0.1225615
Detection Rate	0.0738448
Detection Prevalence	0.3314703
Balanced Accuracy	0.6544509

On a des résultats relativement satisfaisants : une erreur globale de 30.63%, une spécificité très correcte (on estime plutôt bien les mauvais véhicules) et une sensitivité qui ne diminue pas trop.