

Analyse de la Clientèle d'un Concessionnaire Automobile pour la Recommandation de Modèle

Rendu Final

Groupe 5



Quan ZHANG
Yajuan LUO
Yue ZHAO
Chaymae FAZAZI-IDRISSI

Table des matières

Contents

1	Description	4
2	Préparation	5
2.1	Nettoyage de la base de données	5
2.1.1	Nettoyage dans fichiers .csv	5
2.1.2	Nettoyage dans Oracle sous SQL	5
2.2	Charger les données dans R	9
2.3	Fusion des fichiers	9
2.3.1	Fusion entre Catalogue.csv et Immatriculations.csv	10
2.3.2	Fusion entre Clients.csv et la fusion comme 3.2.1	11
3	Hypothèse	13
3.1	Renommer tableFinal	13
3.2	Apprentissage supervisé grâce aux différents classifieurs	13
3.2.1	Arbre de Décision	13
4	Construire le modèle	15
4.1	Le modèle de la prédiction du PRIX	15
4.1.1	Fusion des données	15
4.1.2	Visuellement	16
4.1.3	Apprentissage supervisé grâce aux différents classifieurs	17
4.1.4	Application de la méthode	20
4.2	Le modèle de la prédiction du LONGUEUR	22
4.2.1	Preparation de donnees	22
4.2.2	Visuellement	22
4.2.3	Apprentissage supervisé grâce aux différents classifieurs	23
4.3	Le modèle de la prédiction du NBPORTES	27
4.3.1	Fusion des données	27
4.3.2	Visuellement	27
4.3.3	Apprentissage supervisé grâce aux différents classifieurs	28
4.3.4	Application de la méthode	32
4.4	Le modèle de la prédiction du COULEUR	34
4.4.1	Preparation de donnees	34
4.4.2	Visuellement	34
4.5	Le modèle de la prédiction du OCCASION	36
4.5.1	Fusion des données	36
4.5.2	Visuellement	36
4.5.3	Apprentissage supervisé grâce aux différents classifieurs	37
4.5.4	Application de la méthode	41
5	Intégration de modèle	42

6 Conclusion	44
6.1 Erreurs	44
6.2 NBPLACES	44
6.3 La manque de la mélangelement de la base de données: Cross-Validation	45

1 Description

Pour ce projet on doit concevoir un outil qui va permettre à un concessionnaire automobile de cibler les véhicules qui peuvent intéresser ses clients plus précisément :

- un outil rapide qui peut évaluer le type de véhicule qui peut intéresser ses clients en se basant sur les caractéristiques de chaque client et les différents besoins
- envoyer une documentation précise sur le véhicule le plus adéquat pour des clients sélectionnés par son service marketing

Pour cela on doit faire une méthode de gestion de projet et un plan de mise en œuvre, pour l'analyse aussi il va nous falloir les techniques de data mining, machine learning et deep learning qu'on va utiliser dans notre projet pour répondre à cette problématique en se basant sur des fichiers de données qu'on a à disposition (Catalogue, Immatriculations, Marketing) et les données des clients.

2 Préparation

2.1 Nettoyage de la base de données

2.1.1 Nettoyage dans fichiers .csv

Nous allons trouver et corriger les erreurs de syntaxe:

- Nous remplaçons (è, é) par (e, e) dans tous les fichiers .csv;
- Dans fichier clients.csv: Nous remplaçons (Masculin, Homme, Féminin) par (M, M, F, F);

Pour l'instant, nous ne traitons pas ces valeurs vides et inconnus (grâce à `filter()` dans `library(dplyr)` sous R ou bien nettoyage sous SQL, nous allons sélectionner les données correctes), par exemple:

535	48 F	159	En Couple		FALSE	9652 KH 90
536	36 M	1270	celibataire	0	FALSE	6249 OO 45
537	52 M	594	celibataire	0	FALSE	7653 EU 72
538	41 F	1385	En Couple ?		FALSE	5239 SI 80

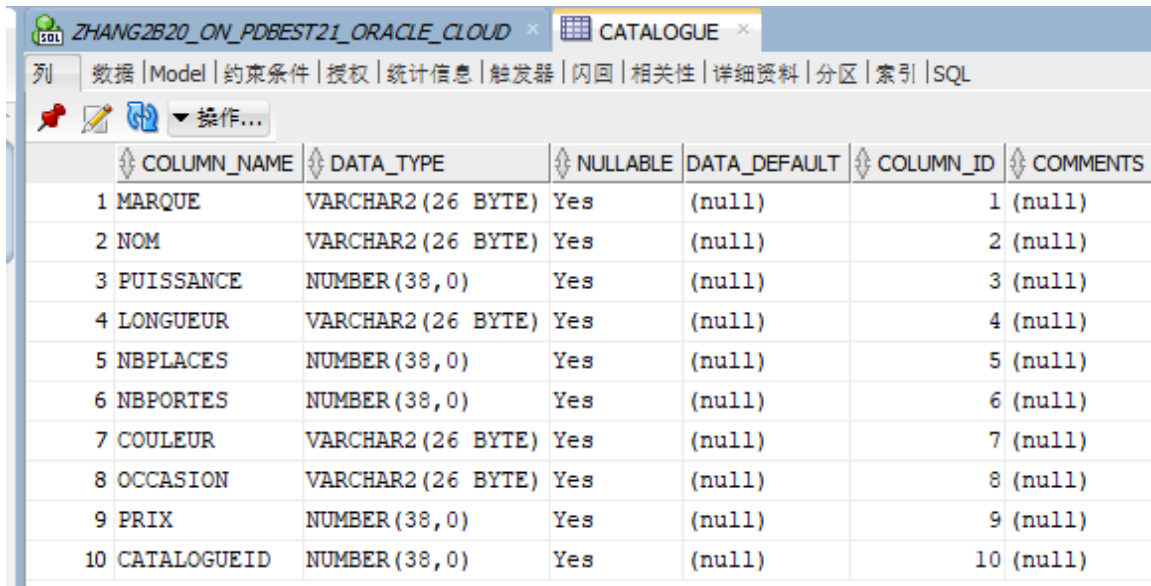
Figure 1: exemple des valeurs vides et inconnus

2.1.2 Nettoyage dans Oracle sous SQL

Nous changeons "2eme voiture" par "deuxiemeVoiture". Quand nous chargeons les données, nous allons cocher pour que les données puissent nullable(default null).

Ensuite, nous allons nettoyer chaque table:

CATALOGUE



	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	MARQUE	VARCHAR2 (26 BYTE)	Yes	(null)	1 (null)	
2	NOM	VARCHAR2 (26 BYTE)	Yes	(null)	2 (null)	
3	PUISSANCE	NUMBER (38, 0)	Yes	(null)	3 (null)	
4	LONGUEUR	VARCHAR2 (26 BYTE)	Yes	(null)	4 (null)	
5	NBPLACES	NUMBER (38, 0)	Yes	(null)	5 (null)	
6	NBPORTES	NUMBER (38, 0)	Yes	(null)	6 (null)	
7	COULEUR	VARCHAR2 (26 BYTE)	Yes	(null)	7 (null)	
8	OCCASION	VARCHAR2 (26 BYTE)	Yes	(null)	8 (null)	
9	PRIX	NUMBER (38, 0)	Yes	(null)	9 (null)	
10	CATALOGUEID	NUMBER (38, 0)	Yes	(null)	10 (null)	

Figure 2: Table: CATALOGUE

Il s'agit d'un tableau très basique, nous pouvons donc constater qu'il n'y a pas d'erreurs dans ce tableau. Donc nous changeons rien dedans.

CLIENTS

	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	AGE	NUMBER(38,0)	Yes	(null)	1 (null)	
2	SEXE	VARCHAR2(26 BYTE)	Yes	(null)	2 (null)	
3	TAUX	VARCHAR2(26 BYTE)	Yes	(null)	3 (null)	
4	SITUATIONFAMILIALE	VARCHAR2(26 BYTE)	Yes	(null)	4 (null)	
5	NBENFANTSACHARGE	NUMBER(38,0)	Yes	(null)	5 (null)	
6	DEUXIEMEVOITURE	VARCHAR2(26 BYTE)	Yes	(null)	6 (null)	
7	IMMATRICULATION	VARCHAR2(26 BYTE)	Yes	(null)	7 (null)	

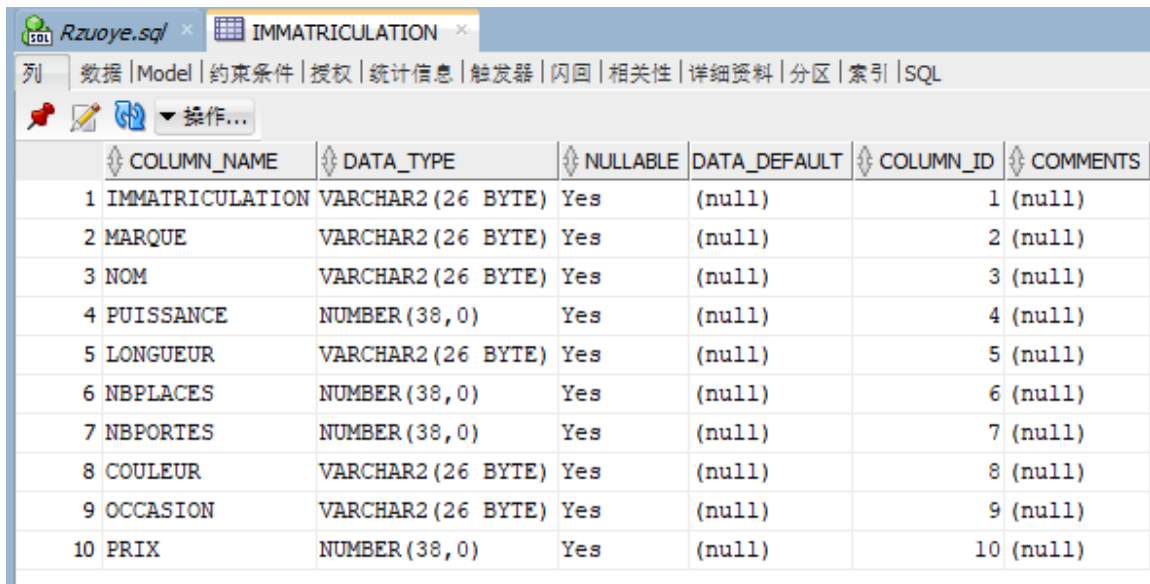
Figure 3: Table: CLIENTS

Les opérations dans table clients sous SQL:

```

1  --Pourque Age soit dans [18, 84]
2  select * from clients where age < 18 or age > 84;
3  DELETE from clients where age < 18 or age > 84;
4
5  --Domain de valeurs 'SEXE': 'F','M'
6  select * from clients where SEXE != 'M' and SEXE != 'F';
7  delete from clients where SEXE != 'M' and SEXE != 'F';
8
9  --Domain de valeurs 'taux': [544, 74185]
10 select * from clients where taux = ' ';
11 delete from clients where taux = ' ';
12 select * from clients where taux = '?';
13 delete from clients where taux = '?';
14 select * from clients where taux < 544 or taux > 74185;
15 delete from clients where TO_NUMBER(taux) < 544 or TO_NUMBER(taux) > 74185;
16
17 --Pour SITUATIONFAMILIALE, on fait group by pour voir les erreurs possibles
18 select sum(age),SITUATIONFAMILIALE from clients group by SITUATIONFAMILIALE;
19 delete from clients where SITUATIONFAMILIALE = '?';
20 delete from clients where SITUATIONFAMILIALE = ' ';
21 delete from clients where SITUATIONFAMILIALE = 'N/D';
22
23 --Domain de valeurs NBENFANTSACHARGE: [0, 4]
24 select * from clients where NBENFANTSACHARGE < 0 or NBENFANTSACHARGE > 4;
25 delete from clients where NBENFANTSACHARGE < 0 or NBENFANTSACHARGE > 4;
26
27 --DEUXIEMEVOITURE: TRUE or FALSE
28 select * from clients where DEUXIEMEVOITURE != 'TRUE' and DEUXIEMEVOITURE != 'FALSE';
29 delete from clients where DEUXIEMEVOITURE != 'TRUE' and DEUXIEMEVOITURE != 'FALSE';
30
31 -- on touche rien pour les valeurs IMMATRICULATION car il y a pas d'errers comme ' ', '?', 'N/D'

```

IMMATRICULATION


	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	IMMATRICULATION	VARCHAR2(26 BYTE)	Yes	(null)	1	(null)
2	MARQUE	VARCHAR2(26 BYTE)	Yes	(null)	2	(null)
3	NOM	VARCHAR2(26 BYTE)	Yes	(null)	3	(null)
4	PUISSANCE	NUMBER(38,0)	Yes	(null)	4	(null)
5	LONGUEUR	VARCHAR2(26 BYTE)	Yes	(null)	5	(null)
6	NBPLACES	NUMBER(38,0)	Yes	(null)	6	(null)
7	NBPORTES	NUMBER(38,0)	Yes	(null)	7	(null)
8	COULEUR	VARCHAR2(26 BYTE)	Yes	(null)	8	(null)
9	OCCASION	VARCHAR2(26 BYTE)	Yes	(null)	9	(null)
10	PRIX	NUMBER(38,0)	Yes	(null)	10	(null)

Figure 4: Table: IMMATRICULATION

Les opérations dans table IMMATRICULATION sous SQL:

```

1  --Afin d'utiliser Group by, on doit nettoyer au moins un column, par exemple 'PRIX'
2  --Domain de valeurs PRIX: [7500, 101300]
3  select * from IMMATRICULATION where prix < 7500 or prix > 101300;
4  --on touche pas si'l y a pas d'erreurs
5
6  --Pour les valeurs marque:
7  select sum(prix), MARQUE from IMMATRICULATION group by MARQUE;
8  --on touche pas si'l y a pas d'erreurs
9
10 --Pour les valeurs nom:
11 select sum(prix), nom from IMMATRICULATION group by nom;
12 --on touche pas si'l y a pas d'erreurs
13
14 --Domain de valeurs PUISSANCE: [55, 507]
15 select * from IMMATRICULATION where PUISSANCE < 55 or PUISSANCE > 507;
16 --on touche pas si'l y a pas d'erreurs
17
18 --Pour les valeurs longueur:
19 select sum(prix), longueur from IMMATRICULATION group by longueur;
20 --on touche pas si'l y a pas d'erreurs
21
22 --Domain de valeurs NBPLACES: [5, 7]
23 select * from IMMATRICULATION where NBPLACES < 5 or NBPLACES > 7;
24 --on touche pas si'l y a pas d'erreurs
25
26 --Domain de valeurs NBPORTES: [3, 5]
27 select * from IMMATRICULATION where NBPORTES < 3 or NBPORTES > 5;
28 --on touche pas si'l y a pas d'erreurs
29

```

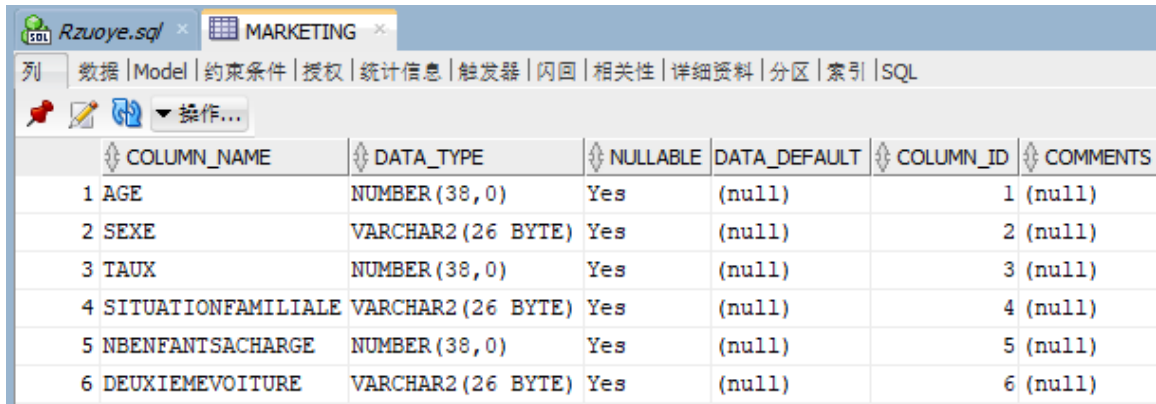
```

30 --Pour les valeurs COULEUR:
31 select sum(prix), COULEUR from IMMATRICULATION group by COULEUR;
32 --on touche pas si'l y a pas d'erreurs
33
34 --Pour les valeurs OCCASION:
35 select sum(prix), OCCASION from IMMATRICULATION group by OCCASION;
36 --on touche pas si'l y a pas d'erreurs

```

Après vérification, il n'y a aucune erreur dans ce tableau

MARKETING



	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	AGE	NUMBER(38,0)	Yes	(null)	1	(null)
2	SEXE	VARCHAR2(26 BYTE)	Yes	(null)	2	(null)
3	TAUX	NUMBER(38,0)	Yes	(null)	3	(null)
4	SITUATIONFAMILIALE	VARCHAR2(26 BYTE)	Yes	(null)	4	(null)
5	NBNFANTSACHARGE	NUMBER(38,0)	Yes	(null)	5	(null)
6	DEUXIEMEVOITURE	VARCHAR2(26 BYTE)	Yes	(null)	6	(null)

Figure 5: Table: MARKETING

Les opérations dans table MARKETING sous SQL:

```

1 --Domain de valeurs 'taux': [544, 74185]
2 select * from MARKETING where taux < 544 or taux > 74185;
3 delete from MARKETING where taux < 544 or taux > 74185;

```

IL FAUT PAS OUBLIER:

```

1 COMMIT;

```


2.2 Charger les données dans R

```

1 install.packages("RJDBC")
2 library(RJDBC)
3
4 drv <- RJDBC::JDBC(driverClass = "oracle.jdbc.OracleDriver", classPath = Sys.glob("C:/Users/12506/
   OneDrive/Desktop/ESTIA_3A/R/Oracle/drivers/*"))
5
6 ##classPath : add path to drivers jdbc
7
8 #Connexion OK
9 conn <- dbConnect(drv, "jdbc:oracle:thin:@(DESCRIPTION=(ADDRESS=(PROTOCOL=TCP) (HOST=144.21.67.201) (
   PORT=1521)) (CONNECT_DATA=(SERVICE_NAME=pdbest21.631174089.oraclecloud.internal)))", "ZHANG2B20",
   "ZHANG2B2001")
10
11 allTables <- dbGetQuery(conn, "SELECT owner, table_name FROM all_tables where owner = 'BABEAU2B20'")
12
13 tableCatalogue <- dbGetQuery(conn, "select * from Catalogue")
14 tableClients <- dbGetQuery(conn, "select * from Clients")
15 tableIm <- dbGetQuery(conn, "select * from IMMATRICULATION")
16 tableMar <- dbGetQuery(conn, "select * from MARKETING")
17 View(tableCatalogue)
18 View(tableClients)
19 View(tableIm)
20 View(tableMar)

```

Data		
▶ allTables	5 obs. of 2 variables	📄
▶ conn	Formal class JDBCConnection	🔍
▶ drv	Formal class JDBCDriver	🔍
▶ tableCatalogue	270 obs. of 10 variables	📄
▶ tableClients	11065 obs. of 7 variables	📄
▶ tableIm	1048575 obs. of 10 variables	📄
▶ tableMar	9 obs. of 6 variables	📄

Figure 6: Les tables apres les nettoyages

2.3 Fusion des fichiers

- Le fichier Clients.csv contient les informations sur les clients ayant les véhicules vendus cette année.
- Le fichier Immatriculations.csv contient les informations sur les véhicules vendus cette année.
- le fichier Catalogue.csv identifier des catégories de véhicules.

Afin d'éviter une prédiction introuvable (Différents types de voitures ont des paramètres différents) et donner la recommandation aux clients, nous prédisons un paramètre chaque fois (marque, nom, puissance....).

Pour la recommandation de modèle, au début de ce projet, nous avons ajouté un column "catalogueId" dans fichier Catalogue.csv. Grâce à ce column, la fusion entre Catalogue.csv et Immatriculations.csv nous permet d'obtenir une relation entre **IMMATRICULATION** et **CATALOGUEID** et la fusion entre Immatriculations.csv et Clients.csv nous permet d'obtenir une relation entre **les paramètres de clients** et **CATALOGUEID**.

Nous allons analyser les relations entre les informations sur les clients et CATALOGUEID et predire pour la Recommandation de Modèle pour les clients sélectionnés par le service marketing dans Marketing.csv.

2.3.1 Fusion entre Catalogue.csv et Immatriculations.csv

```
1 ##charger tous les libraries possible utilises
2 library(rvest)
3 library(ggplot2)
4 library(dplyr)
5 library(scales)
6 library(maps)
7 library(mapproj)
8 library(plotly)
9 library(rpart)
10 library(rpart.plot)
11 library(C50)
12 library(tree)
13 library(ROCR)
14 library(randomForest)
15 library(e1071)
16 library(naivebayes)
17 library(nnet)
18 library(kknn)
19
20 tableIm.f <- merge(tableIm, tableCatalogue, by = c("MARQUE", "NOM", "PUISSANCE", "LONGUEUR", "NBPORTES", "COULEUR", "OCCASION", "PRIX"))
21
22 tableIm.f <- tableIm.f[c(9, 12)]
23 View(tableIm.f)
```

	IMMATRICULATION	CATALOGUEID
654094	7820 YI 94	117
654095	1543 JP 74	117
654096	6692 ZE 81	117
654097	8969 DO 95	117
654098	4260 SV 64	117
654099	7166 PE 13	117
654100	9870 OA 14	117
654101	895 AW 29	117
654102	7120 LQ 69	117
654103	249 AO 23	117
654104	8500 NA 93	117
654105	6372 OG 71	117
654106	4467 FX 42	117
654107	7107 QQ 84	117
654108	895 KO 72	117
654109	7008 CI 08	117

Figure 7: Fusion entre Catalogue.csv et Immatriculations.csv

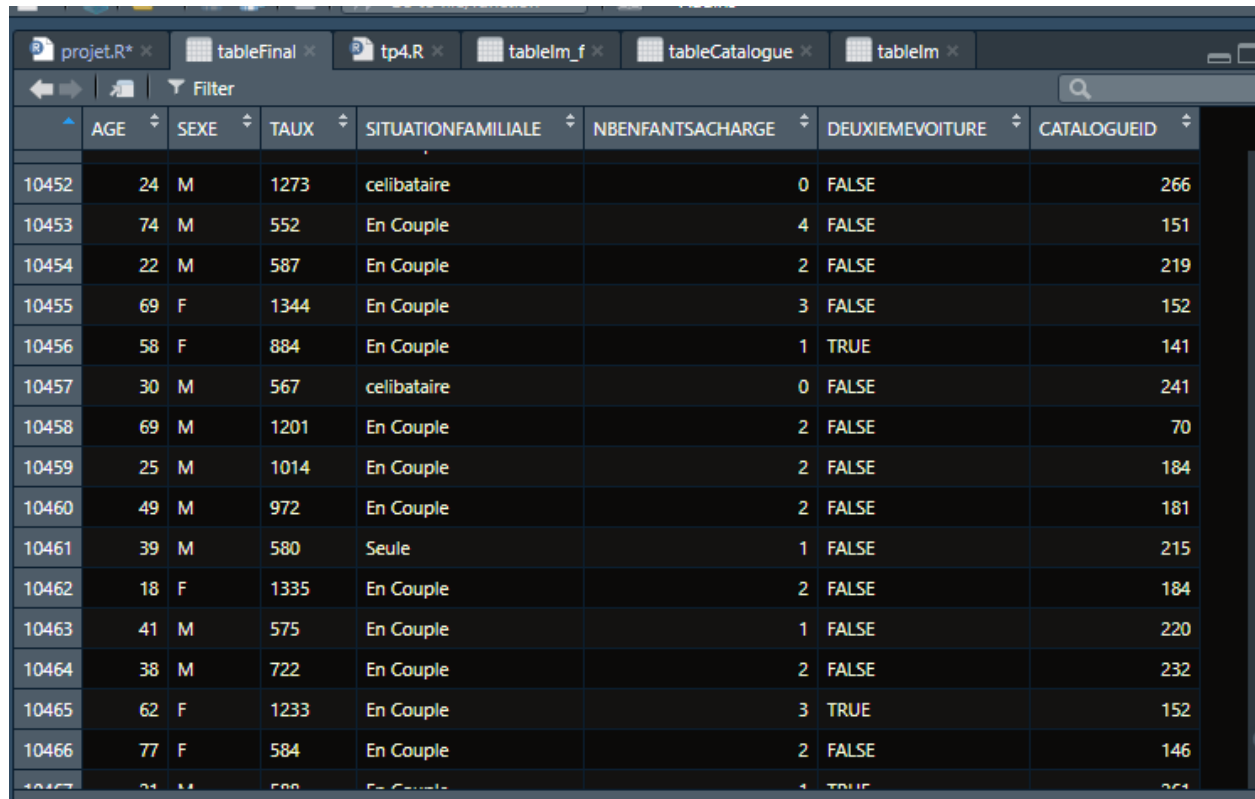
2.3.2 Fusion entre Clients.csv et la fusion comme 3.2.1

```

1 tableFinal <- merge(tableClients, tableIm_f, by = "IMMATRICULATION")
2 View(tableFinal)
3
4 ## supprimer column "IMMATRICULATION"
5 tableFinal <- subset(tableFinal, select=-IMMATRICULATION)
6 View(tableFinal)

```

Finalement, nous avons obtenu une table qui contient tous les champs nous interesent:



	AGE	SEXE	TAUX	SITUATIONFAMILIALE	NBENFANTSACHARGE	DEUXIEMEVOITURE	CATALOGUEID
10452	24	M	1273	celibataire	0	FALSE	266
10453	74	M	552	En Couple	4	FALSE	151
10454	22	M	587	En Couple	2	FALSE	219
10455	69	F	1344	En Couple	3	FALSE	152
10456	58	F	884	En Couple	1	TRUE	141
10457	30	M	567	celibataire	0	FALSE	241
10458	69	M	1201	En Couple	2	FALSE	70
10459	25	M	1014	En Couple	2	FALSE	184
10460	49	M	972	En Couple	2	FALSE	181
10461	39	M	580	Seule	1	FALSE	215
10462	18	F	1335	En Couple	2	FALSE	184
10463	41	M	575	En Couple	1	FALSE	220
10464	38	M	722	En Couple	2	FALSE	232
10465	62	F	1233	En Couple	3	TRUE	152
10466	77	F	584	En Couple	2	FALSE	146
10467	31	M	588	En Couple	1	TRUE	264

Figure 8: La table finale

Ensuite, nous allons analyser les relations parmi les informations de clients et modèle de voiture.

3 Hypothèse

Comme nous l'avons fait au début, nous avons ajouté une colonne ID à la catégorie. Notre idée est de créer directement un modèle, qui nous permet de prédire directement un ID, à travers différents algorithmes en apprentissage supervisé.

3.1 Renommer tableFinal

```
> table(tableEssaiFinal$CATALOGUEID)
```

1	2	3	6	7	22	23	24	26	28	29	36	37	38	39	40	46
72	74	54	72	62	36	52	1	50	46	54	1	3	2	1	2	21
47	48	49	50	61	62	63	64	65	66	67	68	69	70	76	77	78
21	23	30	29	31	40	30	31	39	84	84	84	94	74	60	79	67
79	80	86	87	88	89	90	92	96	97	98	99	100	112	113	115	117
78	58	21	28	39	26	27	1	41	30	36	46	35	41	44	46	41
120	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151
47	3	1	3	9	1	25	29	22	14	17	52	45	49	41	107	32
152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	168	169
85	104	88	121	39	40	14	17	26	63	9	16	35	41	5	9	7
170	173	181	182	183	184	185	186	187	188	189	190	206	207	208	209	210
3	4	337	34	40	357	36	37	46	313	328	327	35	29	39	42	34
215	216	218	219	220	221	224	231	232	233	234	235	236	237	238	239	240
37	33	37	44	44	1	1	59	424	429	404	61	53	407	68	407	71
241	246	247	248	249	251	252	253	254	255	258	259	260	261	262	263	264
29	26	23	36	34	3	8	8	3	13	9	8	3	63	547	561	64
265	266	267	268	269	270											
65	500	507	42	53	499											

Comme dans cette image, bien que nous ayons cent mille données, face à 270 modèles, il semble que ce ne soit pas suffisant pour obtenir un modèle mature. Nous avons quand même décidé de continuer.

3.2 Apprentissage supervisé grâce aux différents classifieurs

Creation des ensembles d'apprentissage et de test:

```
1 tableFinal <- merge(tableClients, tableIm_f, by = "IMMATRICULATION")
2 View(tableFinal)
3
4 # Creation des ensembles d'apprentissage et de test
5 id_EA <- tableEssaiFinal[1:7388,]
6 id_ET <- tableEssaiFinal[7389:11082,]
```

3.2.1 Arbre de Décision

```
1 tableFinal <- merge(tableClients, tableIm_f, by = "IMMATRICULATION")
2 View(tableFinal)
3 # Apprentissage du classifieur de type arbre de decision rpart
4 treeEssai1 <- rpart(CATALOGUEID~., id_EA)
5 prp(treeEssai1, type=4, extra=1, box.col=c("tomato", "skyblue")[treeEssai1$frame$yval])
6 ##Warning message:
```



4 Construire le modèle

Après avoir analysé les données du tableau du catalogue, nous avons décidé d'établir un modèle prédictif pour les paramètres de voiture suivants:

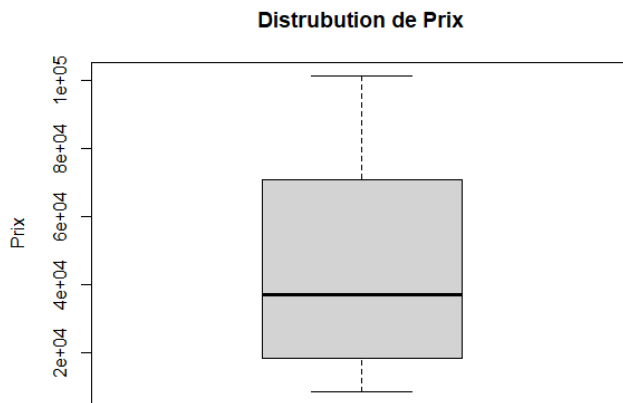
- PRIX
- LONGUEUR
- NBPORTES
- COULEUR
- OCCASION

4.1 Le modèle de la prédiction du PRIX

4.1.1 Fusion des données

```
1 tableImPrix <- tableIm_f[c(8, 9)]
2 tablePrixFinal <- merge(tableClients, tableImPrix, by = "IMMATRICULATION", incomparables = NA)
3 #requete pour la distribution des donnes: Prix
4 summary(tablePrixFinal$PRIX)
5 boxplot(tablePrixFinal$PRIX, data=tablePrixFinal, main="Distrubution_de_Prix", ylab="Prix")
```

On a obtenu la distribution de Prix:



PRIX	
Min.	: 8540
1st Qu.:	18310
Median :	37100
Mean :	44917
3rd Qu.:	70910
Max.	:101300

Nous avons divisé le prix en trois niveaux selon le valeur(j'ai pas encore trouvé la méthode de réaliser la meme fonction de update sous SQL, donc on a changé directement sous SQL):

```
1 update IMMATRICULATION set PRIX = 1 where PRIX <= 18310;
2 update IMMATRICULATION set PRIX = 2 where PRIX > 18310 and PRIX < 70910;
3 update IMMATRICULATION set PRIX = 3 where PRIX >= 70910;
4
5 update CATALOGUE set PRIX = 1 where PRIX <= 18310;
6 update CATALOGUE set PRIX = 2 where PRIX > 18310 and PRIX < 70910;
7 update CATALOGUE set PRIX = 3 where PRIX >= 70910;
8 commit;
```

Après recharger les tables, renommer le prix:

```

1 ##Changer
2 table(tableImPrix$PRIX)
3 ##J'ai pas trouve une methode qui nous permet de realiser la meme fonction de Update
4 tableImPrix <- within(tableImPrix,{
5   PRIX[PRIX == 3] <- "Luxe"
6   PRIX[PRIX == 2] <- "Moyen"
7   PRIX[PRIX == 1] <- "Economique"
8 })

```

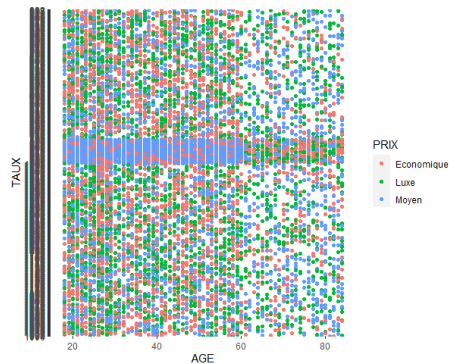
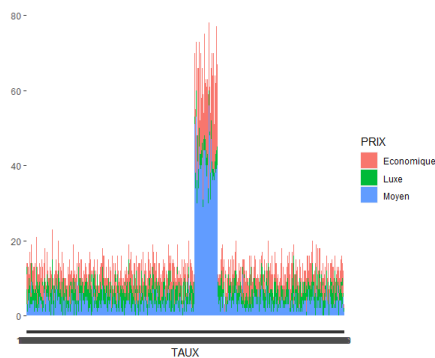
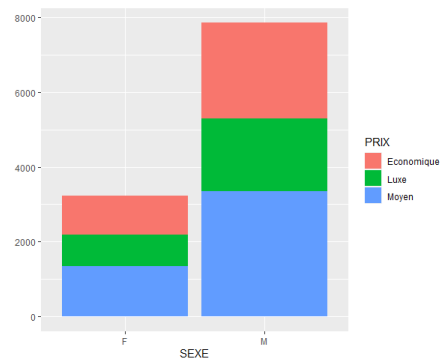
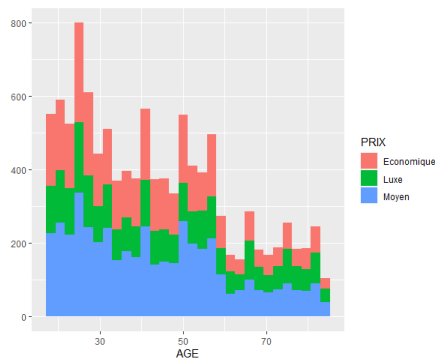
4.1.2 Visuellement

D'après notre jugement subjectif, les facteurs les plus importants affectant le prix d'une voiture sont age, sexe et taux(Capacité d'endettement du client en euros).

```

1 ## Visuellement
2 attach(tablePrixFinal)
3 tablePrixFinal <- subset(tablePrixFinal, select=-IMMATRICULATION)
4 qplot(AGE, data=tablePrixFinal, fill=PRIX)
5 qplot(SEXE, data=tablePrixFinal, fill=PRIX)
6 qplot(TAUX, data=tablePrixFinal, fill=PRIX)
7
8 table(SEXE,PRIX)
9
10 qplot(Age, TAUX, data=tablePrixFinal,color=PRIX)

```



4.1.3 Apprentissage supervisé grâce aux différents classifieurs

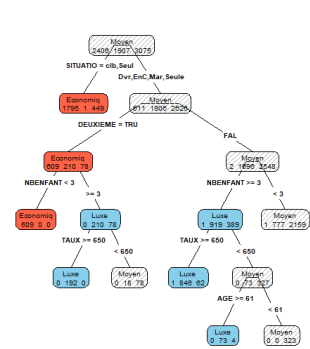
Nous allons tout d'abord choisir un arbre de decision:

Aprendissage

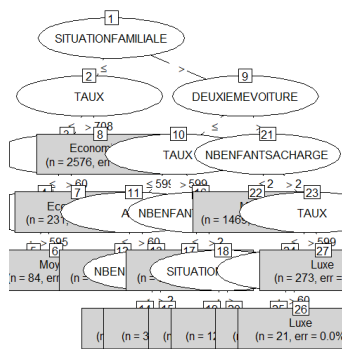
```

1  ### comparer arbre de decision
2  Prix_EA <- tablePrixFinal[1:7388,]
3  Prix_ET <- tablePrixFinal[7389:11082,]
4
5  ## Verifier tous les donnnes sont en bon format
6  str(Prix_EA)
7  Prix_EA$TAUX <- as.integer(Prix_EA$TAUX)
8  Prix_ET$TAUX <- as.integer(Prix_ET$TAUX)
9  str(Prix_EA)
10
11 ## Aprendissage
12 #rpart
13 Prixtree1 <- rpart(PRIX~, Prix_EA)
14 prp(Prixtree1, type=4, extra=1, box.col=c("tomato", "skyblue")[Prixtree1$frame$yval])
15
16 #C5.0
17 Prix_EA$PRIX <- as.factor(Prix_EA$PRIX)
18 Prixtree2 <- C5.0(PRIX~, Prix_EA)
19 plot(Prixtree2, type="simple")
20
21 #Classification and regression trees
22 Prixtree3 <- tree(PRIX~, data= Prix_EA)
23 plot(Prixtree3)
24 text(Prixtree3, pretty=0)

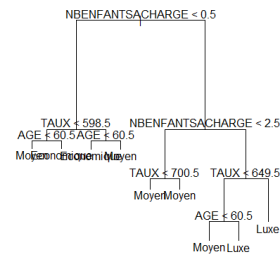
```



(e) a



(f) b



(g) c

Figure 10: Aprendissage: (a)rpart (b)C5.0 (c)Classification and regression trees

comparaison

```

1  #comparaison
2  predPrix.tree1 <- predict(Prixtree1, Prix_ET, type="class")
3  predPrix.tree2 <- predict(Prixtree2, Prix_ET, type="class")
4  predPrix.tree3 <- predict(Prixtree3, Prix_ET, type="class")
5  # Calcul des matrices de confusion

```

```

6 table(Prix_ET$PRIX, predPrix.tree1)
7 table(Prix_ET$PRIX, predPrix.tree2)
8 table(Prix_ET$PRIX, predPrix.tree3)

```

```

> table(Prix_ET$PRIX, predPrix.tree1)
      predPrix.tree1
      Economique Luxe Moyen
Economique    1187     1     1
Luxe           1    541    346
Moyen          250     32   1335
> table(Prix_ET$PRIX, predPrix.tree2)
      predPrix.tree2
      Economique Luxe Moyen
Economique    1074     1    114
Luxe           1    537    350
Moyen          115     1   1501
> table(Prix_ET$PRIX, predPrix.tree3)
      predPrix.tree3
      Economique Luxe Moyen
Economique     739     1    449
Luxe            94    551    243
Moyen           197     32   1388

```

Figure 11: matrices de confusion

ROC et Calcul de l'AUC

A cause de (**ROCR currently supports only evaluation of binary classification tasks.**), on calcule que le taux de reussi. Evidemment, C5.0 a la meilleure performance (0.842).

Ensuite, on utilise differents classifieurs suivants:

```

1  #-----#
2  # RANDOM FORESTS #
3  #-----#
4  # Apprentissage du classifieur de type foret aleatoire
5  rfPrix <- randomForest(PRIX~, Prix_EA)
6  # Test du classifieur : classe predite
7  rf_classPrix <- predict(rfPrix, Prix_ET, type="response")
8  # Matrice de confusion
9  table(Prix_ET$PRIX, rf_classPrix)
10 # Test du classifieur : probabilites pour chaque prediction
11 rf_probPrix <- predict(rfPrix, Prix_ET, type="prob")
12 # L'objet genere est une matrice
13 rf_probPrix
14 #-----#
15 # SUPPORT VECTOR MACHINES #
16 #-----#
17 # Apprentissage du classifieur de type svm
18 svmPrix <- svm(PRIX~, Prix_EA, probability=TRUE)
19 # Test du classifieur : classe predite
20 svm_classPrix <- predict(svmPrix, Prix_ET, type="response")
21 # Matrice de confusion
22 table(Prix_ET$PRIX, svm_classPrix)
23 # Test du classifieur : probabilites pour chaque prediction
24 svm_prob <- predict(svmPrix, Prix_ET, probability=TRUE)

```

```

25 # L'objet genere est de type specifique aux svm
26 svm_prob
27 # Recuperation des probabilites associees aux predictions
28 svm_prob <- attr(svm_prob, "probabilities")
29 # Conversion en un data frame
30 svm_prob <- as.data.frame(svm_prob)
31 #-----#
32 # NAIVE BAYES #
33 #-----#
34 # Apprentissage du classifieur de type naive bayes
35 nbPrix <- naive_bayes(PRIX~, Prix_EA)
36 nbPrix
37 # Test du classifieur : classe predite
38 nbPrix_class <- predict(nbPrix, Prix_ET, type="class")
39 nbPrix_class
40 table(nbPrix_class)
41 # Matrice de confusion
42 table( Prix_ET$PRIX, nbPrix_class)
43 # Test du classifieur : probabilites pour chaque prediction
44 nbPrix_prob <- predict(nbPrix, Prix_ET, type="prob")
45 # L'objet genere est une matrice
46 nbPrix_prob
47 #-----#
48 # NEURAL NETWORKS #
49 #-----#
50 # Apprentissage du classifieur de type perceptron monocouche
51 nnPrix <- nnet(PRIX~, Prix_EA, size=12)
52 nnPrix
53 # Test du classifieur : classe predite
54 nnPrix_class <- predict(nnPrix, Prix_ET, type="class")
55 nnPrix_class
56 table(nnPrix_class)
57 # Matrice de confusion
58 table(Prix_ET$PRIX, nnPrix_class)
59 # Test du classifieur : probabilites pour chaque prediction
60 nnPrix_prob <- predict(nnPrix, Prix_ET, type="raw")
61 # L'objet genere est un vecteur des probabilites de prediction
62 nnPrix_prob
63 #-----#
64 # K-NEAREST NEIGHBORS #
65 #-----#
66 # Apprentissage et test simultanes du classifieur de type k-nearest neighbors
67 knnPrix <- kknn(PRIX~, Prix_EA, Prix_ET)
68 # Resultat : classe predite et probabilites de chaque classe pour chaque instance de test
69 summary(knnPrix)
70 # Matrice de confusion
71 table(Prix_ET$PRIX, knnPrix$fitted.values)
72 # Conversion des probabilites en data frame
73 knnPrix_prob <- as.data.frame(knnPrix$prob)

```

```
> table(Prix_ET$PRIX, rf_classPrix)
      rf_classPrix
Economeque Luxe Moyen
Economeque 1184 1 4
Luxe 1 544 343
Moyen 247 14 1356
```

(a) RANDOM FORESTS

```
> table(Prix_ET$PRIX, svm_classPrix)
      svm_classPrix
Economeque Luxe Moyen
Economeque 1187 1 1
Luxe 1 516 371
Moyen 251 6 1360
```

(b) SUPPORT VECTOR MACHINES

```
> table(Prix_ET$PRIX, nbPrix_class)
      nbPrix_class
Economeque Luxe Moyen
Economeque 1060 103 26
Luxe 37 584 267
Moyen 301 269 1047
```

(c) NAIVE BAYES

```
> table(Prix_ET$PRIX, nnPrix_class)
      nnPrix_class
Economeque Luxe Moyen
Economeque 1189
Luxe 888
Moyen 1617
```

(d) NEURAL NETWORKS

```
> table(Prix_ET$PRIX, knnPrix$fitted.values)
      knnPrix$fitted.values
Economeque Luxe Moyen
Economeque 1105 1 83
Luxe 3 631 254
Moyen 142 203 1272
```

(e) K-NEAREST NEIGHBORS

Figure 12: Résultat des différents classifieurs

Taux de réussite:

Modèle	Taux de réussite
arbre de decision	0.842
RANDOM FORESTS	0.835
SUPPORT VECTOR MACHINES	0.829
NAIVE BAYES	0.728
K-NEAREST NEIGHBORS	0.814

4.1.4 Application de la méthode

On choisi arbre de decision (C5.0) Et on va appliquer cette méthode:


```
1 #-----#
2 # APPLICATION DE LA METHODE arbre de decision (C5.0) #
3 #-----#
4 # Visualisation des donnees a predire
5 View(tableMar)
6
7 #=== C5.0 ===#
8 class.treeC50 <- predict(Prixtree2, tableMar, probability=TRUE)
9 # L'objet genere est de type specifique aux svm
10 class.treeC50
11 # Recuperation des probabilites associees aux predictions
12 prob.treeC50 <- attr(class.treeC50, "probabilities")
13 # Conversion en un data frame
14 prob.treeC50 <- as.data.frame(prob.treeC50)
15 resultatPrix <- data.frame(tableMar$ID, class.treeC50, prob.treeC50)
16
17 #=== ARBRE DE DECISION C5.0 ===#
18 class.treeC50 <- predict(Prixtree2, tableMar, type="class")
19 prob.treeC50 <- predict(Prixtree2, tableMar, type="prob")
20 resultatPrix <- data.frame(tableMar, class.treeC50, prob.treeC50)
21 resultatPrix <- data.frame(tableMar, class.treeC50)
22
23 # Renommage de la colonne des classes predites
24 names(resultatPrix)[7] <- "PRIX"
25
26 #-----#
27 # ENREGISTREMENT DES PREDICTIONS #
28 #-----#
```

```

29 # Enregistrement du fichier de resultats au format csv
30 write.table(resultat1, file='predictions.csv', sep="\t", dec=".", row.names = F)

```

Enfin, on a obtenu:



	AGE	SEXE	TAUX	SITUATIONFAMILIALE	NBENFANTSACHARGE	DEUXIEMEVOITURE	PRIX
1	21	F	1396	celibataire	0	FALSE	Economique
2	59	F	572	En Couple	2	FALSE	Moyen
3	64	M	559	celibataire	0	FALSE	Economique
4	79	F	981	En Couple	2	FALSE	Moyen
5	55	M	588	celibataire	0	FALSE	Moyen
6	34	F	1112	En Couple	0	FALSE	Moyen
7	58	M	1192	En Couple	0	FALSE	Moyen
8	35	M	589	celibataire	0	FALSE	Moyen
9	59	M	748	En Couple	0	TRUE	Economique

Figure 13: Le modèle de la prédiction du PRIX

4.2 Le modèle de la prédiction du LONGUEUR

4.2.1 Preparation de donnees

```

1 tableImLG <- tableIm_f[c(4, 9)]
2 tableLGFinal <- merge(tableClients, tableImLG, by = "IMMATRICULATION", incomparables = NA)
3 #requete pour la distribution des donnees: LONGUEUR
4 summary(tableLGFinal$LONGUEUR)
5 ##changer les types de donnees
6 str(tableLGFinal)
7 tableLGFinal$TAUX <- as.integer(tableLGFinal$TAUX)
8 tableLGFinal <- subset(tableLGFinal, select=-IMMATRICULATION)

```

4.2.2 Visuellement

```

1 ##On trouve qu'ils y moins de voitures moyenne
2 library(ggplot2)
3 qplot(LONGUEUR, data=tableLGFinal)
4 table(tableLGFinal$DEUXIEMEVOITURE, tableLGFinal$LONGUEUR)
5 qplot(DEUXIEMEVOITURE, data=tableLGFinal, color=LONGUEUR)
6 qplot(TAUX, data=tableLGFinal, fill=LONGUEUR, bins = 5)
7 boxplot(AGE~LONGUEUR, data=tableLGFinal, col=c("red", "blue"))
8 qplot(SEXE, data=tableLGFinal, color=LONGUEUR)

```

Tout d'abord, nous avons constaté grâce aux statistiques qu'il y a peu de clients choisiraient d'acheter des modèles de longueur moyenne:

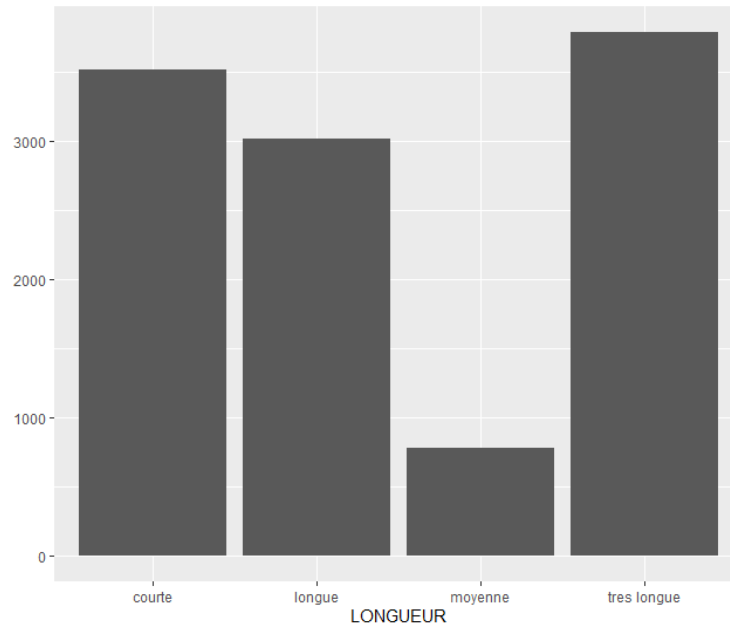
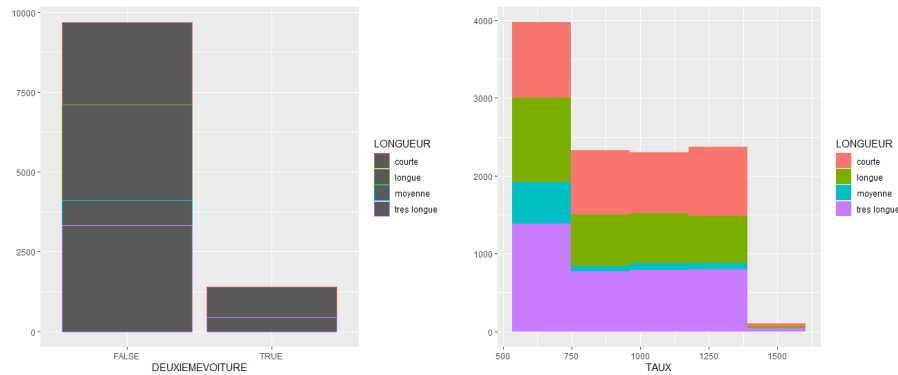
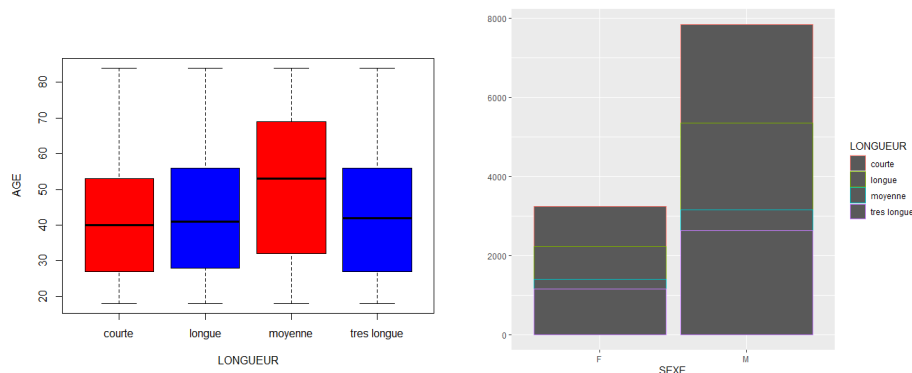


Figure 14: Statistiques de LONGUEUR

Par rapport à l'achat de longueur moyenne, l'intention d'achat des utilisateurs qui choisissent d'acheter d'autres modèles est plus équilibrée; Les gens qui ont grand capacité d'endettement du client ne préfèrent pas acheter les voitures qui ont la longueur moyenne:



Par rapport à l'achat d'autres modèles, l'âge moyen des personnes qui achètent de la longueur moyenne est plus élevé; Mais l'intention d'achat n'est pas affectée par le sexe:



4.2.3 Apprentissage supervisé grâce aux différents classifieurs

Apprendissage

```

1 #-----#
2 # Apprentissage #
3 #-----#
4
5
6 LG_EA <- tableLGFinal[1:7388,]
7 LG_ET <- tableLGFinal[7389:11082,]
8
9 #3 arbres de decision
10 LGtree1 <- rpart(LONGUEUR~., LG_EA)
11 LG_EA$LONGUEUR <- as.factor(LGtree1$LONGUEUR)
12 LGtree2 <- C5.0(LONGUEUR~., LG_EA)
13 LGtree3 <- tree(LONGUEUR~., data=LG_EA)
14 #RANDOM FORESTS
15 rfLG <- randomForest(LONGUEUR~., LG_EA)
16 # SUPPORT VECTOR MACHINES
17 svmLG <- svm(LONGUEUR~., LG_EA, probability=TRUE)
18 # NAIVE BAYES
19 nbLG <- naive.bayes(LONGUEUR~., LG_EA)
20 # NEURAL NETWORKS
21 nnLG <- nnet(LONGUEUR~., LG_EA, size=12)

```

```

22 # K-NEAREST NEIGHBORS
23 knnLG <- kknm(LONGUEUR~, LG_EA, LG_ET)

```

Text des classifieurs et matrice de confusion

```

1  ##Text des arbres et matrice de confusion
2  predLG.tree1 <- predict(LGtree1, LG_ET, type="class")
3  predLG.tree2 <- predict(LGtree2, LG_ET, type="class")
4  predLG.tree3 <- predict(LGtree3, LG_ET, type="class")
5  # Calcul des matrices de confusion
6  table(LG_ET$LONGUEUR, predLG.tree1)
7  #predLG.tree1
8  #courte longue moyenne tres longue
9  #courte 1166 1 0 1
10 #longue 1 1021 0 0
11 #moyenne 270 0 0 0
12 #tres longue 1 493 0 740
13 table(LG_ET$LONGUEUR, predLG.tree2)
14 #predLG.tree2
15 #courte longue moyenne tres longue
16 #courte 1166 1 0 1
17 #longue 1 1021 0 0
18 #moyenne 270 0 0 0
19 #tres longue 1 493 0 740
20 table(LG_ET$LONGUEUR, predLG.tree3)
21 ###
22 # predLG.tree3
23 #courte longue moyenne tres longue
24 #courte 940 227 0 1
25 #longue 304 675 0 43
26 #moyenne 270 0 0 0
27 #tres longue 163 310 0 761
28 ##Text des classifieurs et matrice de confusion
29 result.rfLG <- predict(rfLG, LG_ET, type="response")
30 table(LG_ET$LONGUEUR, result.rfLG)
31 # result.rfLG
32 #courte longue moyenne tres longue
33 #courte 1154 1 12 1
34 #longue 1 1015 0 6
35 #moyenne 258 0 12 0
36 #tres longue 1 491 0 742
37 result.svmLG <- predict(svmLG, LG_ET, type="response")
38 table(LG_ET$LONGUEUR, result.svmLG)
39 # result.svmLG
40 #courte longue moyenne tres longue
41 #courte 1166 1 0 1
42 #longue 2 1020 0 0
43 #moyenne 270 0 0 0
44 #tres longue 1 493 0 740
45 result.treeNaiveLG <- predict(nbLG, LG_ET, type="class")
46 table(LG_ET$LONGUEUR, result.treeNaiveLG)
47 # result.treeNaiveLG
48 #courte longue moyenne tres longue
49 #courte 332 0 684 152

```



```

50 #longue 12 661 2 347
51 #moyenne 5 0 265 0
52 #tres longue 7 320 0 907
53 result.treeNnetLG <- predict(nnLG, LG_ET,type="class")
54 table(LG_ET$LONGUEUR, result.treeNnetLG)
55 # result.treeNnetLG
56 #tres longue
57 #courte 1168
58 #longue 1022
59 #moyenne 270
60 #tres longue 1234
61 table(LG_ET$LONGUEUR, knnLG$fitted.values)
62 ## courte longue moyenne tres longue
63 #courte 1063 1 103 1
64 #longue 1 744 0 277
65 #moyenne 154 0 116 0
66 #tres longue 2 355 0 877
67 #-----#
68 # plot arbres de decisions#
69 #-----#
70
71 ##arbres de decisions
72 prp(LGtree1, type=4, extra=1, box.col=c("tomato", "skyblue")[LGtree1$frame$yval])
73 plot(LGtree2, type="simple")
74 plot(LGtree3)
75 text(LGtree3, pretty=0)

```

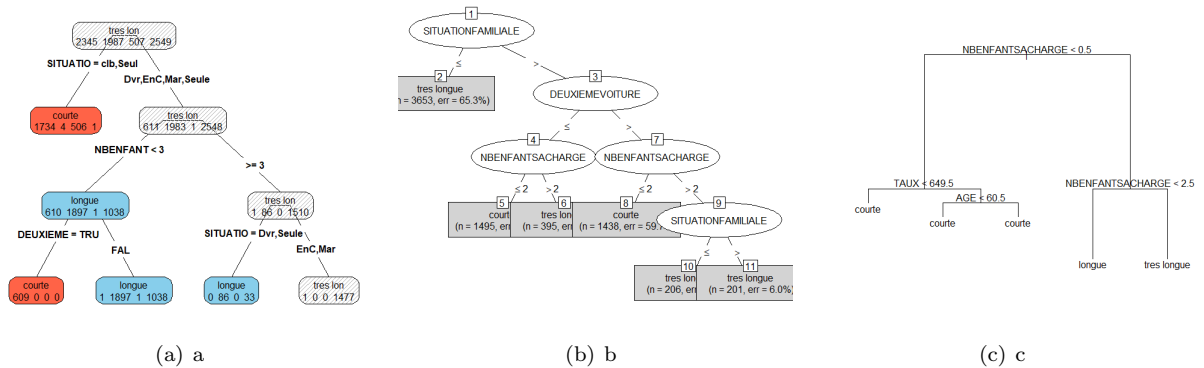


Figure 15: (a)rpart (b)C5.0 (c)Classification and regression trees

ROC et Calcul de l'AUC

A cause de (ROCR currently supports only evaluation of binary classification tasks.), on calcule que le taux de reussi.

Comparasion de taux de réussi:

Modèle	Taux de réussite
rpart	0.7924
C5.0	0.7924
tree	0.6432
RANDOM FORESTS	0.7913
SUPPORT VECTOR MACHINES	0.7921
NAIVE BAYES	0.5861
K-NEAREST NEIGHBORS	0.7580

Après de la comparaison, on a choisi d'utiliser rpart pour ce modèle:

```

1 #-----#
2 # APPLICATION DE LA METHODE rpart #
3 #-----#
4 # Visualisation des donnees a predire
5 View(tableMar)
6
7 #=== rpart ===#
8 class.trerplG <- predict(LGtree1, tableMar, type="class")
9
10 resultatLG <- data.frame(tableMar, class.trerplG)
11
12 # Renommage de la colonne des classes predites
13 names(resultatLG)[7] <- "LONGUEUR"

```

	AGE	SEXE	TAUX	SITUATIONFAMILIALE	NBNFANTSACHARGE	DEUXIEMEVOITURE	LONGUEUR
1	21	F	1396	celibataire	0	FALSE	courte
2	59	F	572	En Couple	2	FALSE	longue
3	64	M	559	celibataire	0	FALSE	courte
4	79	F	981	En Couple	2	FALSE	longue
5	55	M	588	celibataire	0	FALSE	courte
6	34	F	1112	En Couple	0	FALSE	longue
7	58	M	1192	En Couple	0	FALSE	longue
8	35	M	589	celibataire	0	FALSE	courte
9	59	M	748	En Couple	0	TRUE	courte

Figure 16: Le modèle de la prédiction du LONGUEUR

4.3 Le modèle de la prédiction du NBPORTES

4.3.1 Fusion des données

```

1 tableImNP <- tableIm_f[c(5,9)]
2 View(tableImNP)
3 tableFinal_np <- merge(tableClients, tableImNP, by= "IMMATRICULATION", incomparables =NA)
4 summary(tableFinal_np$NBPORTES)
5 tableFinal_np$NBPORTES <- as.factor(tableFinal_np$NBPORTES)
6 attach(tableFinal_np)
7 tableFinal_np <- subset(tableFinal_np, select = -IMMATRICULATION)
8 #
9 NBPORTES_EA <- tableFinal_np[1:7388,]
10 NBPORTES_ET <- tableFinal_np[7389:11082,]
11 NBPORTES_EA$TAUX <- as.integer(NBPORTES_EA$TAUX)
12 NBPORTES_ET$TAUX <- as.integer(NBPORTES_ET$TAUX)

```

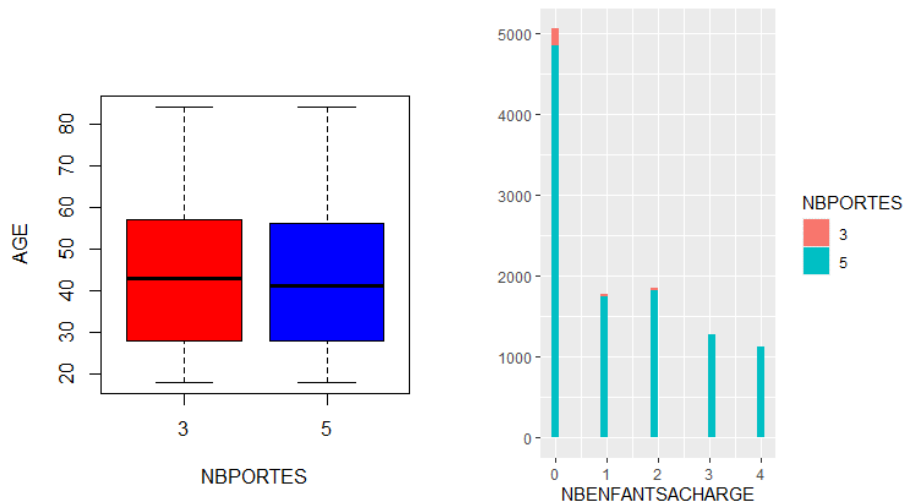
4.3.2 Visuellement

On va essayer de trouver les relations parmi les paramètres:

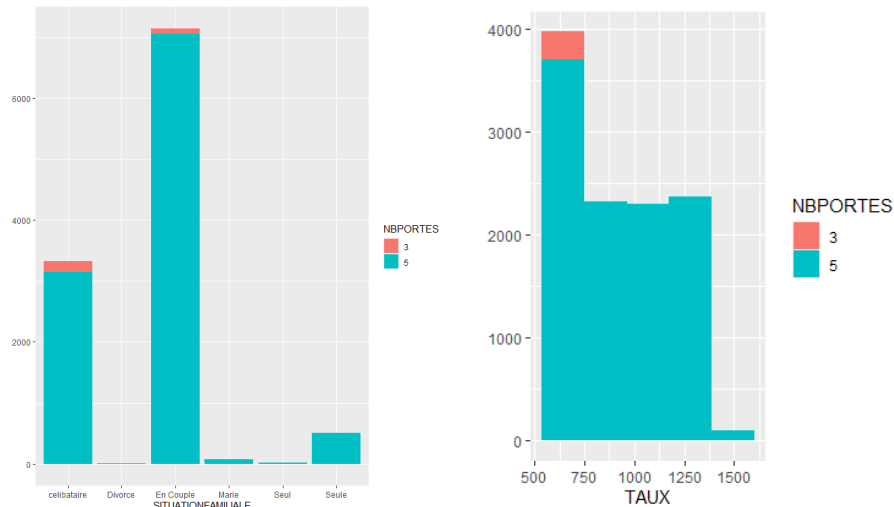
```

1 ##On trouve qu'ils y moins de voitures occasions
2 qqplot(NBPORTES, data=tableFinal_np)
3 qqplot(SEXE, data=tableFinal_np, fill=NBPORTES)
4 qqplot(NBENFANTSACHARGE, data=tableFinal_np, fill=NBPORTES)
5 qqplot(TAUX, data=tableFinal_np, fill=NBPORTES, bins=5)
6 qqplot(DEUXIEMEVOITURE, data=tableFinal_np, fill=NBPORTES)
7 qqplot(SITUATIONFAMILIALE, data=tableFinal_np, fill=NBPORTES)
8 boxplot(AGE~NBPORTES, data=tableFinal_np, col=c("red", "blue"))

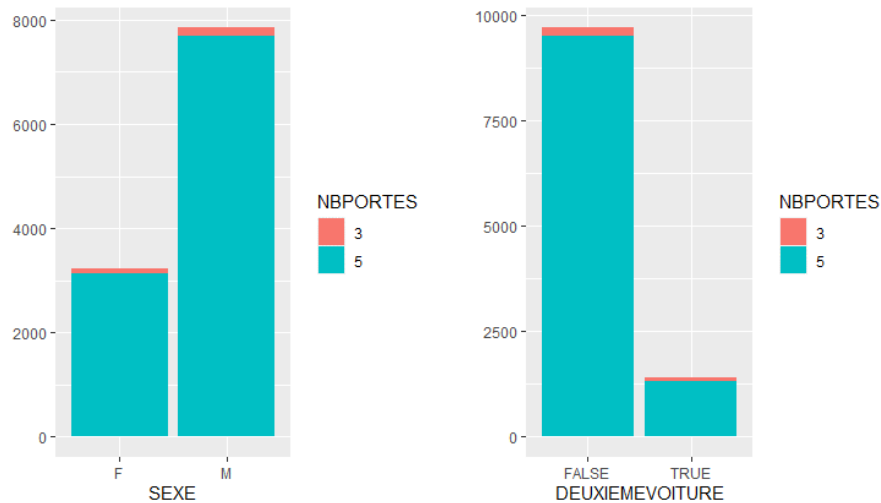
```



Vu que les gens ont tendance à acheter la voiture qui avec 5 portes, on propose le modèle plutôt de 5 portes si le client n'indique pas sa préférence de la voiture avec 3 portes.



Dans le graphique de SEXE et DEXIEMEVOITURE, on ne peut pas constater clairement qu'il y a de la règle ou la liaison évidente entre celles et le nombre de portes, donc globalement on ne prend pas compte de ces deux éléments. Par contre, on peut remarquer les liaisons entre le nombre de porte et les trois restes facteurs. Selon le graphique du nombre des enfant, on dirait que les gens qui ont plus des enfants ont plus de possibilités de choisir la voiture avec 5 portes, en d'autres termes, les gens qui choisissent la voiture avec 3 portes n'ont pas des enfants, c'est logique. Par le graphique de la situation familiale, il indique que les gens qui achètent la voiture de 3 portes sont plutôt célibataire. Et par le graphique de taux, on trouve que les personnes qui sont plus capables d'endetter n'achètent pas la voiture de 3 portes.(L'age)



4.3.3 Apprentissage supervisé grâce aux différents classifieurs

Apprentissage

```

1 #-----#
2 #rpart #
3 #-----#
4 NBPORTESTree1<-rpart(NBPORTES~, NBPORTES_EA)
5 prp(NBPORTESTree1,type=4,extra = 1, box.col

```

```

6 c("tomato","skyblue")[NBPORTESTree1$frame$yval])
7 #-----#
8 #C5.0 #
9 #-----#
10 #NBPORTES_EA$NBPORTES <- as.factor(NBPORTES_EA$NBPORTES)
11 NBPORTESTree2<-C5.0(NBPORTES~, NBPORTES_EA)
12 plot(NBPORTESTree2,type="simple")
13 #-----#
14 #Tree #
15 #-----#
16 NBPORTESTree3<-tree(NBPORTES~, data=NBPORTES_EA)
17 text(NBPORTESTree3,pretty = 0)
18 #-----#
19 #RANDOM FORETS#
20 #-----#
21 rfNP<-randomForest(NBPORTES~, NBPORTES_EA)
22 #-----#
23 #SUPPORT VECTOR MACHINES#
24 #-----#
25 svmNP<-svm(NBPORTES~, NBPORTES_EA,probability=TRUE)
26 #-----#
27 #NAIVE BAYES #
28 #-----#
29 tableFinal_np$NBPORTES <- as.factor(tableFinal_np$NBPORTES)
30 nbNP<-naive.bayes(NBPORTES~, NBPORTES_EA)
31 #-----#
32 #NEURAL NETWORKS #
33 #-----#
34 nnNP<-nnet(NBPORTES~, NBPORTES_EA,size=12)
35 #-----#
36 #K-NEAREST NEIGHBORS#
37 #-----#
38 knnNP<-kkn(NBPORTES~, NBPORTES_EA,NBPORTES_ET)

```

Prediction et Visuellement avec Courbe ROC

```

1 #-----#
2 #decision tree#
3 #-----#
4 predNBPORTES.tree1 <- predict(NBPORTESTree1, NBPORTES_ET, type="vector")
5 predNBPORTES.tree2 <- predict(NBPORTESTree2, NBPORTES_ET, type="class")
6 predNBPORTES.tree3 <- predict(NBPORTESTree3, NBPORTES_ET, type="class")
7 table(NBPORTES_ET$NBPORTES,predNBPORTES.tree1)
8 table(NBPORTES_ET$NBPORTES,predNBPORTES.tree2)
9 table(NBPORTES_ET$NBPORTES,predNBPORTES.tree3)
10 #rpart
11 pNP.tree1<-predict(NBPORTESTree1,NBPORTES_ET,type = "vector")
12 print(pNP.tree1)
13 rocNP.pred1<-prediction(pNP.tree1,NBPORTES_ET$NBPORTES)
14 print(rocNP.pred1)
15 rocNP.pref1<-performance(rocNP.pred1,"tpr","fpr")
16 print(rocNP.pref1)
17 plot(rocNP.pref1,col="green")
18 #c5.0

```

```

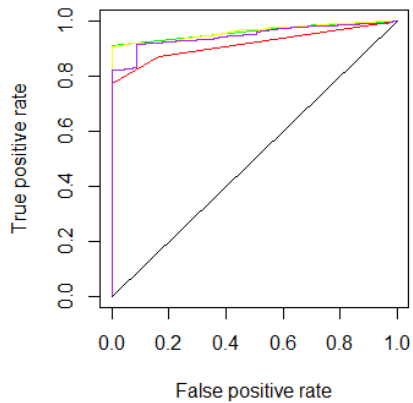
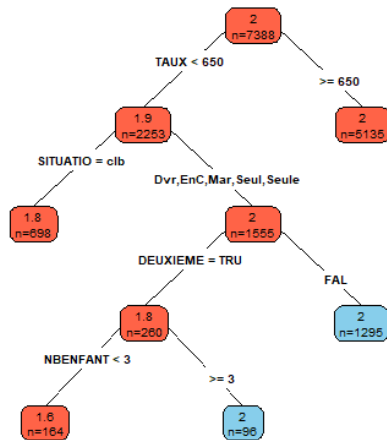
19 pNP.tree2<-predict(NBPORTESTree2,NBPORTES_ET,type = "prob")
20 rocNP.pred2<-prediction(pNP.tree2[,2],NBPORTES_ET$NBPORTES)
21 rocNP.pref2<-performance(rocNP.pred2,"tpr","fpr")
22 plot(rocNP.pref2,add = TRUE,col="blue")
23 #tree
24 pNP.tree3<-predict(NBPORTESTree3,NBPORTES_ET,type = "vector")
25 rocNP.pred3<-prediction(pNP.tree3[,2],NBPORTES_ET$NBPORTES)
26 rocNP.pref3<-performance(rocNP.pred3,"tpr","fpr")
27 plot(rocNP.pref3,add = TRUE,col="red")
28 #-----#
29 #RANDOM FORETS#
30 #-----#
31 rf_classNP<-predict(rfNP,NBPORTES_EA,type = "response")
32 table(NBPORTES_EA$NBPORTES,rf_classNP)
33 rf_probNP<-predict(rfNP,NBPORTES_ET, type = "prob" )
34 rf_probNP
35 rf_predNP<-prediction(rf_probNP[,2],NBPORTES_ET$NBPORTES)
36 rf_prefNP<-performance(rf_predNP,"tpr","fpr")
37 plot(rf_prefNP,add = TRUE, col="yellow")
38 #-----#
39 #SUPPORT VECTOR MACHINES#
40 #-----#
41 svm_classNP<-predict(svmNP,NBPORTES_ET,type = "response")
42 table(NBPORTES_ET$NBPORTES,svm_classNP)
43 svm_prob<-predict(svmNP,NBPORTES_ET,probability=TRUE)
44 svm_prob
45 svm_prob<-attr(svm_prob,"probabilities")
46 svm_prob<-as.data.frame(svm_prob)
47 svm_pred<-prediction(svm_prob[,2],NBPORTES_ET$NBPORTES)
48 svm_pref<-performance(rf_predNP,"tpr","fpr")
49 plot(svm_pref,add = TRUE, col="black")
50 #-----#
51 #NAIVE BAYES #
52 #-----#
53 nbNP_class<-predict(nbNP,NBPORTES_ET,type = "class")
54 nbNP_class
55 table(nbNP_class)
56 table(NBPORTES_ET$NBPORTES,nbNP_class)
57 nbNP_prob<-predict(nbNP,NBPORTES_ET,type="prob")
58 nbNP_prob
59 nbNP_pred<-prediction(nbNP_prob[,2],NBPORTES_ET$NBPORTES)
60 nbNP_pref<-performance(nbNP_pred,"tpr","fpr")
61 plot(nbNP_pref,add = TRUE, col="purple")
62 #-----#
63 #NEURAL NETWORKS #
64 #-----#
65 nnNP_class<-predict(nnNP,NBPORTES_ET, TYPE="class")
66 nnNP_class
67 table(nnNP_class)
68 table(NBPORTES_ET$NBPORTES,nnNP_class)
69 nnNP_prob<-predict(nnNP,NBPORTES_ET,type = "raw")
70 nnNP_prob
71 nnNP_pred<-prediction(nnNP_prob,NBPORTES_ET$NBPORTES)

```

```

72 nnNP_pref<-performance(nnNP_pred,"tpr","fpr")
73 plot(nnNP_pref,add = TRUE, col="orange")
74 #-----#
75 #K-NEAREST NEIGHBORS#
76 #-----#
77 table(NBPORTES_ET$NBPORTES,knnNP$fitted.values)
78 knnNP_prob<-as.data.frame(knnNP$prob)
79 knnNP_pred<-prediction(knnNP_prob[,2],NBPORTES_ET$NBPORTES)
80 knnNP_pref<-performance(knnNP_pred,"tpr","fpr")
81 plot(nnNP_pref,add = TRUE, col="black")

```



Calcul de l'AUC

```

1 #-----#
2 #decision tree#
3 #-----#
4 aucNP.tree1<-performance(rocNP.pred1,"auc")
5 attr(aucNP.tree1,"y.values")
6 aucNP.tree2<-performance(rocNP.pred2,"auc")
7 attr(aucNP.tree2,"y.values")
8 aucNP.tree3<-performance(rocNP.pred3,"auc")
9 attr(aucNP.tree3,"y.values")
10 #-----#
11 #RANDOM FORETS#
12 #-----#
13 rf_aucNP<-performance(rf_predNP,"auc")
14 attr(rf_aucNP,"y.values")
15 #-----#
16 #SUPPORT VECTOR MACHINES#
17 #-----#
18 svm_aucNP<-performance(svm_pred,"auc")
19 attr(svm_aucNP,"y.values")
20 #-----#
21 #NAIVE BAYES #
22 #-----#
23 nbNP_aucNP<-performance(nbNP_pred,"auc")

```

```

24 attr(nbNP_aucNP, "y.values")
25 #-----#
26 #NEURAL NETWORKS #
27 #-----#
28 nnNP_aucNP<-performance(nnNP_pred,"auc")
29 attr(nnNP_aucNP, "y.values")
30 #-----#
31 #K-NEAREST NEIGHBORS#
32 #-----#
33 knnNP_aucNP<-performance(knnNP_pred,"auc")
34 attr(knnNP_aucNP, "y.values")

```

Modèle	L'INDICE AUC
arbre de decision(tree)	0.9031576
RANDOM FORESTS	0.9097089
SUPPORT VECTOR MACHINES	0.063305
NAIVE BAYES	0.9345112
NEURAL NETWORKS	0.5
K-NEAREST NEIGHBORS	0.8545001

4.3.4 Application de la méthode

```

1 #-----#
2 # APPLICATION DE LA METHODE NAIVE BAYES #
3 #-----#
4 # Visualisation des donnees a predire
5 View(tableMar)
6
7 #=== NAIVE BAYES ===#
8 class.treerpartNB <- predict(nbNP, tableMar, probability=TRUE)
9 resultatNB <- data.frame(tableMar, class.treerpartNB)
10
11 # Renommage de la colonne des classes predites
12 names(resultatNB)[7] <- "NBPORTES"

```

Alors nous avons le résultat pour occasion:

<div> <div>projct.R ×</div> <div>resultatNB ×</div> <div>tablelm_f ×</div> </div>								
Filter								
	AGE	SEXE	TAUX	SITUATIONFAMILIALE	NBENFANTSACHARGE	DEUXIEMEVOITURE	NBPORTES	
1	21	F	1396	celibataire	0	FALSE	5	
2	59	F	572	En Couple	2	FALSE	5	
3	64	M	559	celibataire	0	FALSE	3	
4	79	F	981	En Couple	2	FALSE	5	
5	55	M	588	celibataire	0	FALSE	3	
6	34	F	1112	En Couple	0	FALSE	5	
7	58	M	1192	En Couple	0	FALSE	5	
8	35	M	589	celibataire	0	FALSE	3	
9	59	M	748	En Couple	0	TRUE	5	

Figure 17: Le modèle de la prédiction de NBPORTES

4.4 Le modèle de la prédiction du COULEUR

4.4.1 Preparation de donnees

```

1 #-----#
2 # Preparation de donnees #
3 #-----#
4
5 tableImCL <- tableIm.f[c(6, 9)]
6 tableCLFinal <- merge(tableClients, tableImCL, by = "IMMATRICULATION", incomparables = NA)
7 #requete pour la distribution des donnees: OCCASION
8 summary(tableCLFinal$COULEUR)
9 ##changer les types de donnees
10 str(tableCLFinal)
11 tableCLFinal$TAUX <- as.integer(tableCLFinal$TAUX)
12 tableCLFinal <- subset(tableCLFinal, select=-IMMATRICULATION)

```

4.4.2 Visuellement

```

1 ##On trouve qu'ils y moins de voitures occasions
2 library(ggplot2)
3 qplot(OCCASION, data=tableOCCAFinal)
4 table(tableOCCAFinal$DEUXIEMEVOITURE, tableOCCAFinal$OCCASION)
5 qplot(DEUXIEMEVOITURE, data=tableOCCAFinal, color=OCCASION)
6 qplot(TAUX, data=tableOCCAFinal, fill=OCCASION, bins =5)
7 boxplot(AGE~OCCASION, data=tableOCCAFinal, col=c("red", "blue"))
8 qplot(SEXE, data=tableOCCAFinal, color=OCCASION)

```

Après des recherches comparatives, nous avons constaté que la relation entre la couleur et toute variable n'est pas évidente.

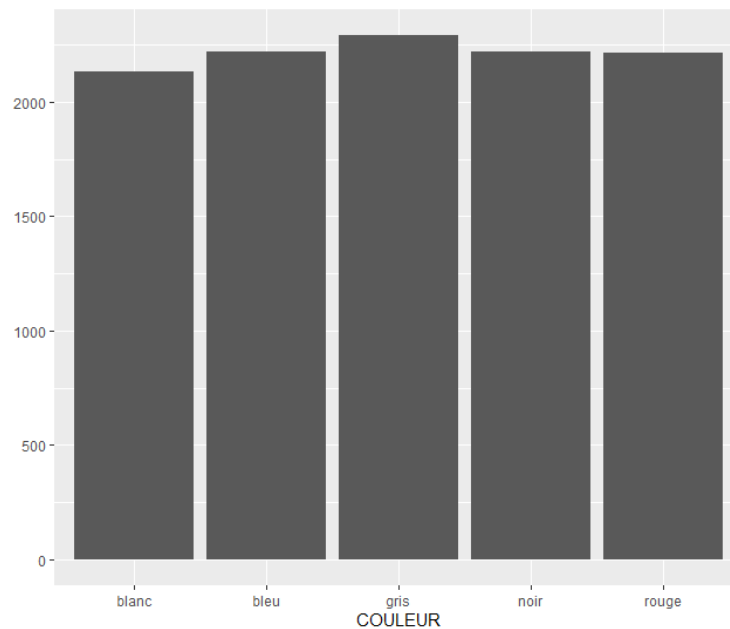
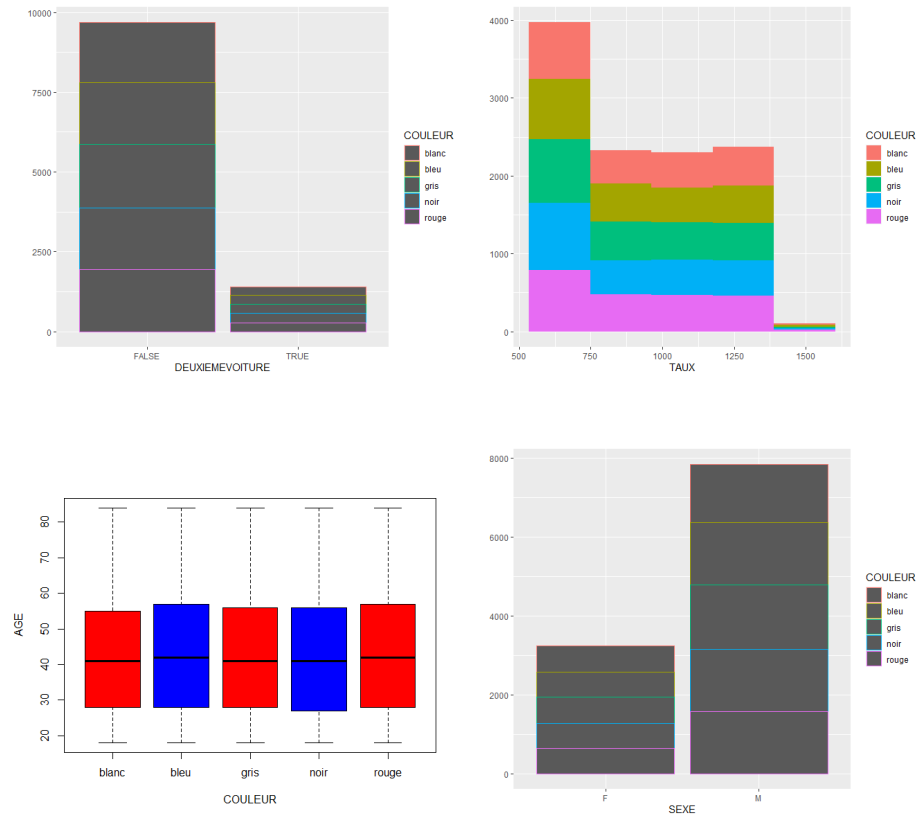


Figure 18: Statistiques de COULEUR



Par conséquent, nous n'étudions plus cette variable, car même si les résultats sont disponibles, ils ne sont pas très fiables.

4.5 Le modèle de la prédiction du OCCASION

4.5.1 Fusion des données

```

1 tableImOCCA <- tableIm_f[c(7, 9)]
2 tableOCCAFinal <- merge(tableClients, tableImOCCA, by = "IMMATRICULATION", incomparables = NA)
3 #requete pour la distribution des donnees: OCCASION
4 summary(tableOCCAFinal$OCCASION)
5 ##changer les types de donnees
6 str(tableOCCAFinal)
7 tableOCCAFinal$TAUX <- as.integer(tableOCCAFinal$TAUX)

```

4.5.2 Visuellement

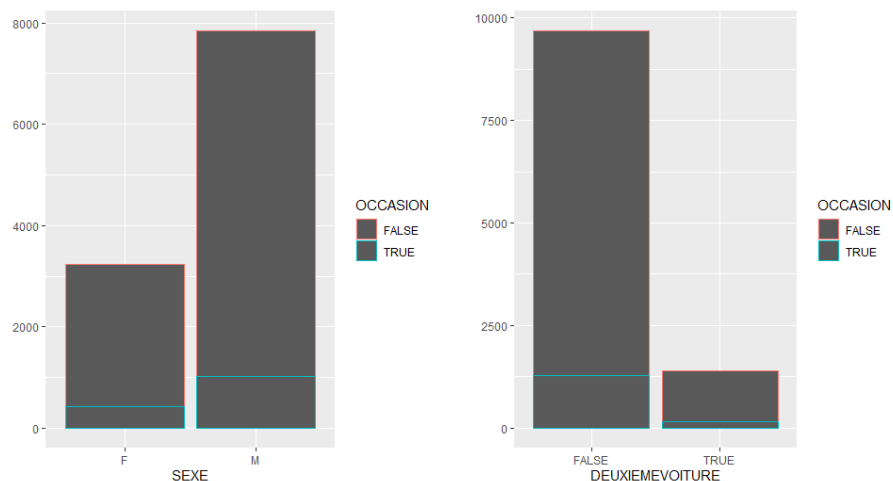
On va essayer de trouver les relations parmi les paramètres:

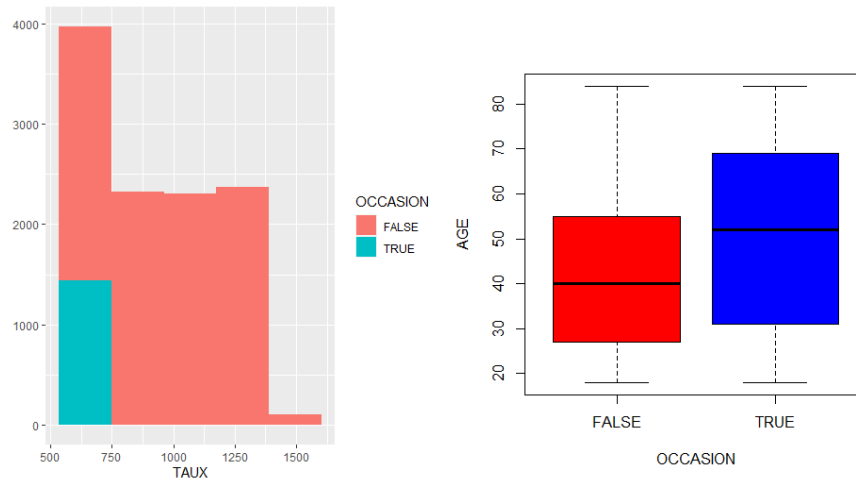
```

1 ##On trouve qu'ils y moins de voitures occasions
2 library(ggplot2)
3 qplot(OCCASION, data=tableOCCAFinal)
4 table(tableOCCAFinal$DEUXIEMEVOITURE, tableOCCAFinal$OCCASION)
5 qplot(DEUXIEMEVOITURE, data=tableOCCAFinal, color=OCCASION)
6 qplot(TAUX, data=tableOCCAFinal, fill=OCCASION, bins =5)
7 boxplot(AGE~OCCASION, data=tableOCCAFinal, col=c("red", "blue"))
8 qplot(SEXE, data=tableOCCAFinal, color=OCCASION)

```

Comme dans les figures au dessous, nous avons visualisé les relations entre le sexe, la deuxième voiture, l'âge et l'état de l'occasion. Pour les relations entre les sexes et l'état de l'occasion ou bien entre si c'est la deuxième voiture et l'état de l'occasion, nous n'avons pas pu trouver une règle ou bien une tendance. Dans figure 3, nous trouvons que les personnes qui ont bonne capacités d'endettements achètent moins de voitures d'occasions. Dans figure 4, nous trouvons que Les voitures d'occasion ont un meilleur marché auprès des jeunes.





4.5.3 Apprentissage supervisé grâce aux différents classifieurs

Nous allons tout d'abord choisir un arbre de decision:

Aprendissage

```

1 #-----#
2 # Apprentissage #
3 #-----#
4 tableOCCAFinal <- subset(tableOCCAFinal, select=-IMMATRICULATION)
5 OCCA_EA <- tableOCCAFinal[1:7388,]
6 OCCA_ET <- tableOCCAFinal[7389:11082,]
7
8 #3 arbres de decision
9 OCCAtree1 <- rpart(OCCASION~., OCCA_EA)
10 OCCA_EA$OCCASION <- as.factor(OCCA_EA$OCCASION)
11 OCCAtree2 <- C5.0(OCCASION~., OCCA_EA)
12 OCCAtree3 <- tree(OCCASION~., data=OCCA_EA)
13 #RANDOM FORESTS
14 rfOCCA <- randomForest(OCCASION~., OCCA_EA)
15 # SUPPORT VECTOR MACHINES
16 svmOCCA <- svm(OCCASION~., OCCA_EA, probability=TRUE)
17 # NAIVE BAYES
18 nbOCCA <- naive_bayes(OCCASION~., OCCA_EA)
19 # NEURAL NETWORKS
20 nnOCCA <- nnet(OCCASION~., OCCA_EA, size=12)
21 # K-NEAREST NEIGHBORS
22 knnOCCA <- kkn(OCCASION~., OCCA_EA, OCCA_ET)

```

Text des classifieurs et matrice de confusion

```

1 ##Text des arbres et matrice de confusion
2 predOCCA.tree1 <- predict(OCCAtree1, OCCA_ET, type="class")
3 predOCCA.tree2 <- predict(OCCAtree2, OCCA_ET, type="class")
4 predOCCA.tree3 <- predict(OCCAtree3, OCCA_ET, type="class")
5 # Calcul des matrices de confusion
6 table(OCCA_ET$OCCASION, predOCCA.tree1)
7 table(OCCA_ET$OCCASION, predOCCA.tree2)

```

```

8 table(OCCA_ET$OCCASION, predOCCA.tree3)
9 #####
10 # FALSE TRUE
11 #FALSE 3189 37
12 #TRUE 290 178
13 ##Text des classifieurs et matrice de confusion
14 result.rfOCCA <- predict(rfOCCA,OCCA_ET, type="response")
15 table(OCCA_ET$OCCASION, result.rfOCCA)
16 #####result.rfOCCA
17 #####FALSE TRUE
18 #FALSE 3181 45
19 #TRUE 290 178
20 result.svmOCCA <- predict(svmOCCA,OCCA_ET, type="response")
21 table(OCCA_ET$OCCASION, result.svmOCCA)
22 #####result.svmOCCA
23 #####FALSE TRUE
24 #FALSE 3186 40
25 #TRUE 307 161
26 result.treeNaiveOCCA <- predict(nbOCCA,OCCA_ET, type="class")
27 table(OCCA_ET$OCCASION, result.treeNaiveOCCA)
28 #####result.treeNaiveOCCA
29 #####FALSE TRUE
30 #FALSE 2518 708
31 #TRUE 2 466
32 result.treeNnetOCCA <- predict(nnOCCA, OCCA_ET,type="class")
33 table(OCCA_ET$OCCASION, result.treeNnetOCCA)
34 #####result.treeNnetOCCA
35 #####FALSE
36 #FALSE 3226
37 #TRUE 468
38 table(OCCA_ET$OCCASION, knnOCCA$fitted.values)
39 #####FALSE TRUE
40 #FALSE 3020 206
41 #TRUE 228 240

```

CALCUL DE COURBES ROC

```

1 # Test du classifieur : probabilités pour chaque prediction
2 p.treeRpartOCCA <- predict(OCCAtree1, OCCA_ET, type="prob")
3 # Courbe ROC
4 roc.predOCCA1 <- prediction(p.treeRpartOCCA[,2], OCCA_ET$OCCASION)
5 roc.perfOCCA1 <- performance(roc.predOCCA1,"tpr", "fpr")
6
7 p.treec50OCCA <- predict(OCCAtree2, OCCA_ET, type="prob")
8 # Courbe ROC
9 roc.predOCCA2 <- prediction(p.treec50OCCA[,2], OCCA_ET$OCCASION)
10 roc.perfOCCA2 <- performance(roc.predOCCA2,"tpr", "fpr")
11
12 p.treeTreeOCCA <- predict(OCCAtree3, OCCA_ET, type="vector")
13 # Courbe ROC
14 roc.predOCCA3 <- prediction(p.treeTreeOCCA[,2], OCCA_ET$OCCASION)
15 roc.perfOCCA3 <- performance(roc.predOCCA3,"tpr", "fpr")
16
17 rf_probOCCA <- predict(rfOCCA, OCCA_ET, type="prob")

```

```

18 # Courbe ROC
19 roc.predOCCA4 <- prediction(rf.probOCCA[,2], OCCA_ET$OCCASION)
20 roc.perfOCCA4 <- performance(roc.predOCCA4,"tpr","fpr")
21
22 svm_probOCCA <- predict(svmOCCA, OCCA_ET, probability=TRUE)
23 # Recuperation des probabilites associees aux predictions
24 svm_probOCCA <- attr(svm_probOCCA, "probabilities")
25 # Conversion en un data frame
26 svm_probOCCA <- as.data.frame(svm_probOCCA)
27 # Courbe ROC sur le meme graphique
28 roc.predOCCA5 <- prediction(svm_probOCCA[,2], OCCA_ET$OCCASION)
29 roc.perfOCCA5 <- performance(roc.predOCCA5,"tpr","fpr")
30
31 nb_probOCCA <- predict(nbOCCA, OCCA_ET, type="prob")
32 # Courbe ROC
33 roc.predOCCA6 <- prediction(nb_probOCCA[,2], OCCA_ET$OCCASION)
34 roc.perfOCCA6 <- performance(roc.predOCCA6,"tpr","fpr")
35
36 nn_probOCCA <- predict(nnOCCA, OCCA_ET, type="raw")
37 # Courbe ROC
38 roc.predOCCA7 <- prediction(nn_probOCCA[,1], OCCA_ET$OCCASION)
39 roc.perfOCCA7 <- performance(roc.predOCCA7,"tpr","fpr")
40
41 # Conversion des probabilites en data frame
42 knn_probOCCA <- as.data.frame(knnOCCA$prob)
43 # Courbe ROC
44 roc.predOCCA8 <- prediction(knn_probOCCA[,2], OCCA_ET$OCCASION)
45 roc.perfOCCA8 <- performance(roc.predOCCA8,"tpr","fpr")

```

CALCUL DE L INDICE AUC

```

1 auc.treeOCCA1 <- performance(roc.predOCCA1, "auc")
2 auc.treeOCCA2 <- performance(roc.predOCCA2, "auc")
3 auc.treeOCCA3 <- performance(roc.predOCCA3, "auc")
4 auc.treeOCCA4 <- performance(roc.predOCCA4, "auc")
5 auc.treeOCCA5 <- performance(roc.predOCCA5, "auc")
6 auc.treeOCCA6 <- performance(roc.predOCCA6, "auc")
7 auc.treeOCCA7 <- performance(roc.predOCCA7, "auc")
8 auc.treeOCCA8 <- performance(roc.predOCCA8, "auc")
9
10 attr(auc.treeOCCA1, "y.values")
11 attr(auc.treeOCCA2, "y.values")
12 attr(auc.treeOCCA3, "y.values")
13 attr(auc.treeOCCA4, "y.values")
14 attr(auc.treeOCCA5, "y.values")
15 attr(auc.treeOCCA6, "y.values")
16 attr(auc.treeOCCA7, "y.values")
17 attr(auc.treeOCCA8, "y.values")

```

Choisissons un classifieur qui a la meilleur performance

Après les calculs, nous allons choisir RANDOM FORESTS comme notre modèle d'apprentissage.

Modèle	L'INDICE AUC
arbre de decision	0.9241559
RANDOM FORESTS	0.9247706
SUPPORT VECTOR MACHINES	0.9208312
NAIVE BAYES	0.9246225
NEURAL NETWORKS	0.5
K-NEAREST NEIGHBORS	0.9033746

Après avoir comparé différents arbres de décision, nous avons constaté qu'ils avaient la même structure. Donc ils ont la même matrice de confusion, la même courbe ROC et la même l'indice AUC. En fait, dans 4.5.2, nous avons déjà trouvé les deux facteurs les plus influents: TAUX et AGE.

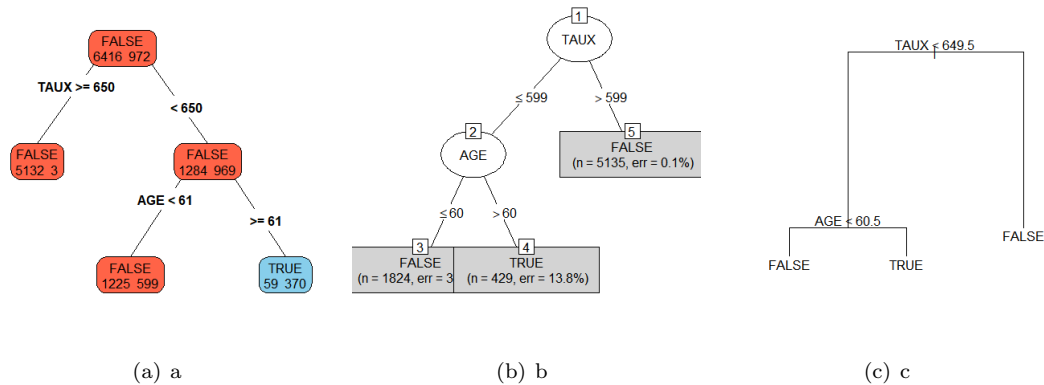


Figure 19: (a)rpart (b)C5.0 (c)Classification and regression trees

Nous pouvons aussi tracer la courbe ROC:

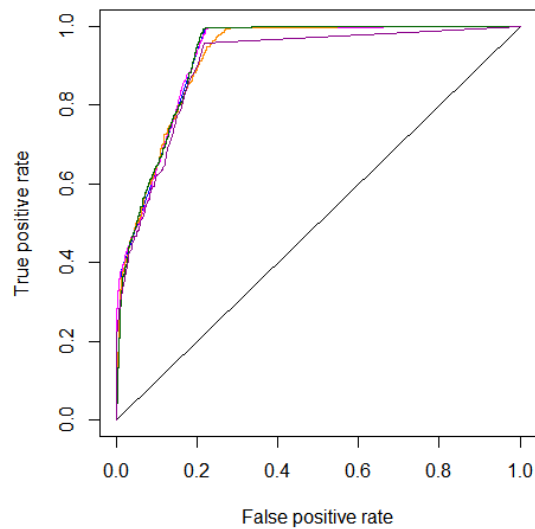


Figure 20: La courbe ROC

4.5.4 Application de la méthode

```

1 #-----#
2 # APPLICATION DE LA METHODE RANDOM FORESTS #
3 #-----#
4 # Visualisation des donnees a predire
5 View(tableMar)
6
7 #=== RANDOM FORESTS ===#
8 class.trerFOCCA <- predict(rfOCCA, tableMar, probability=TRUE)
9 resultatOCCA <- data.frame(tableMar, class.trerFOCCA)
10
11 # Renommage de la colonne des classes predites
12 names(resultatOCCA)[7] <- "OCCASION"

```

Alors nous avons le résultat pour occasion:

	AGE	SEXE	TAUX	SITUATIONFAMILIALE	NBENFANTSACHARGE	DEUXIEMEVOITURE	OCCASION
1	21	F	1396	celibataire	0	FALSE	FALSE
2	59	F	572	En Couple	2	FALSE	FALSE
3	64	M	559	celibataire	0	FALSE	TRUE
4	79	F	981	En Couple	2	FALSE	FALSE
5	55	M	588	celibataire	0	FALSE	FALSE
6	34	F	1112	En Couple	0	FALSE	FALSE
7	58	M	1192	En Couple	0	FALSE	FALSE
8	35	M	589	celibataire	0	FALSE	FALSE
9	59	M	748	En Couple	0	TRUE	FALSE

Figure 21: Le modèle de la prédiction de OCCASION

5 Intégration de modèle

On a décidé d'utiliser package "openxlsx" pour l'export de fichier car il y a des erreurs si on utilise .csv.

```

1 #-----#
2 # ENREGISTREMENT DES PREDICTIONS #
3 #-----#
4 # Enregistrement du fichier de resultats au format xlsx
5
6
7
8 install.packages("openxlsx")
9 library(openxlsx)
10 write.xlsx(resultatTOTAL,"predictions.xlsx")

```

	A	B	C	D	E	F	G	H	I	J	
1	AGE	SEXE	TAUX	SITUATIONFA	NBENFANTS	DEUXIEMEVO	LONGUEUR	NBPORTES	OCCASION	PRIX	
2	21	F	1396	celibataire	0	FALSE	courte	5	FALSE	Economique	
3	59	F	572	En Couple	2	FALSE	longue	5	FALSE	Moyen	
4	64	M	559	celibataire	0	FALSE	courte	3	TRUE	Economique	
5	79	F	981	En Couple	2	FALSE	longue	5	FALSE	Moyen	
6	55	M	588	celibataire	0	FALSE	courte	3	FALSE	Moyen	
7	34	F	1112	En Couple	0	FALSE	longue	5	FALSE	Moyen	
8	58	M	1192	En Couple	0	FALSE	longue	5	FALSE	Moyen	
9	35	M	589	celibataire	0	FALSE	courte	3	FALSE	Moyen	
10	59	M	748	En Couple	0	TRUE	courte	5	FALSE	Economique	
11											
12											

Figure 22: Le modèle final

Si nous voulons en savoir plus, nous pouvons rechercher directement dans la base de données

```

1 ---FINAL
2 select MARQUE from CATALOGUE WHERE LONGUEUR = 'courte' AND NBPORTES = 5 AND
   OCCASION = 'FALSE' AND PRIX = 1 group by MARQUE;
3 select MARQUE from CATALOGUE WHERE LONGUEUR = 'longue' AND NBPORTES = 5 AND
   OCCASION = 'FALSE' AND PRIX = 2 group by MARQUE;
4 select MARQUE from CATALOGUE WHERE LONGUEUR = 'courte' AND NBPORTES = 3 AND
   OCCASION = 'TRUE' AND PRIX = 1 group by MARQUE;
5 select MARQUE from CATALOGUE WHERE LONGUEUR = 'courte' AND NBPORTES = 5 AND
   OCCASION = 'FALSE' AND PRIX = 2 group by MARQUE;
6 select MARQUE from CATALOGUE WHERE LONGUEUR = 'longue' AND NBPORTES = 3 AND
   OCCASION = 'FALSE' AND PRIX = 2 group by MARQUE;
7 select MARQUE from CATALOGUE WHERE LONGUEUR = 'longue' AND NBPORTES = 5 AND
   OCCASION = 'FALSE' AND PRIX = 2 group by MARQUE;
8 select MARQUE from CATALOGUE WHERE LONGUEUR = 'courte' AND NBPORTES = 5 AND
   OCCASION = 'FALSE' AND PRIX = 2 group by MARQUE;
9 select MARQUE from CATALOGUE WHERE LONGUEUR = 'courte' AND NBPORTES = 3 AND
   OCCASION = 'FALSE' AND PRIX = 2 group by MARQUE;
10 select MARQUE from CATALOGUE WHERE LONGUEUR = 'courte' AND NBPORTES = 5 AND
   OCCASION = 'FALSE' AND PRIX = 1 group by MARQUE;

```

À partir de là, nous pouvons tirer le choix le plus approprié pour les clients. En raison des facteurs subjectifs du client, nous ne donnons pas de réponse précise. Au lieu de cela, donnez-leur quelques suggestions à prendre en considération. Donner aux clients plusieurs choix possibles peut attirer des clients aux goûts

uniques.

Nous l'avons remarqué dans tous les algorithmes. Notre objectif est de minimiser l'impact des données en dehors de l'hyperplan où se trouve le cluster. Par conséquent, il est irrationnel de donner aux clients une recommandation spécifique.

Hinge loss: $\max \left\{ 0, 1 - y_i (x_i^T \beta - \beta_0) \right\}$

(loss compared to a slab excluding observation x_i)

Interpretation:

- ▶ $\hat{\xi}_i > 0$: x_i is on wrong side of slab boundary
- ▶ $\hat{\xi}_i = 0$:
 - $y_i(x_i^T \hat{\beta} - \hat{\beta}_0) = 1$: x_i is on slab boundary
 - $y_i(x_i^T \hat{\beta} - \hat{\beta}_0) > 1$: x_i is on correct side

$$\hat{\beta}, \hat{\beta}_0, (\hat{\xi}_i)_{i=1}^n := \underset{\beta, \beta_0, (\xi_i)_{i=1}^n}{\operatorname{argmin}} \quad \frac{1}{n} \sum_{i=1}^n \xi_i + \lambda \|\beta\|^2$$

subject to:

$$\forall i \quad \xi_i \geq 0 \text{ and } y_i(x_i^T \beta - \beta_0) \geq 1 - \xi_i$$

(optimum turns one inequality into an equality $\rightsquigarrow \hat{\xi}_i = \max \left\{ 0, 1 - y_i(x_i^T \hat{\beta} - \hat{\beta}_0) \right\}$)

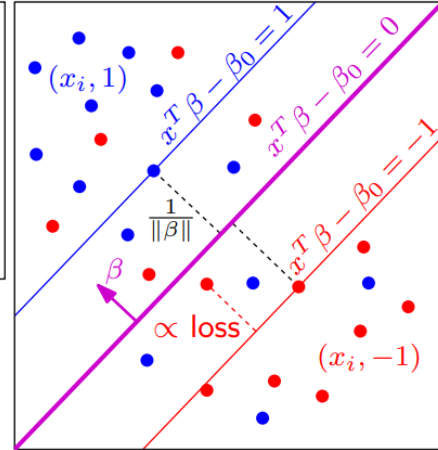


Figure 23: Algorithmme SVM

6 Conclusion

6.1 Erreurs

Après la dernière fusion, on obtient `tableFinal` qui contient plus de lignes que `tableClients`. C-à-d, il y a encore des erreurs à nettoyer. Mais nous allons ignorer ces erreurs (0.1% n'influence pas)

Data		
allTables	5 obs. of 2 variables	
conn	Formal class JDBCConnection	
drv	Formal class JDBCDriver	
tableCatalogue	270 obs. of 10 variables	
tableClients	11065 obs. of 7 variables	
tableFinal	11082 obs. of 7 variables	
tableIm	1048575 obs. of 10 variables	
tableIm_f	1048575 obs. of 2 variables	
tableMar	9 obs. of 6 variables	

6.2 NBPLACES

```
> summary(tableIm_f)
```

MARQUE	NOM	PUISSANCE	
Length:1048575	Length:1048575	Min. : 55.0	
Class :character	Class :character	1st Qu.: 75.0	
Mode :character	Mode :character	Median :150.0	
		Mean :198.9	
		3rd Qu.:245.0	
		Max. :507.0	

LONGUEUR	NBPORTES	COULEUR	
Length:1048575	Min. :3.000	Length:1048575	
Class :character	1st Qu.:5.000	Class :character	
Mode :character	Median :5.000	Mode :character	
	Mean :4.868		
	3rd Qu.:5.000		
	Max. :5.000		

OCCASION	PRIX	IMMATRICULATION	NBPLACES.x
Length:1048575	Min. : 7500	Length:1048575	Min. :5
Class :character	1st Qu.: 18310	Class :character	1st Qu.:5
Mode :character	Median : 25970	Mode :character	Median :5
	Mean : 35767		Mean :5
	3rd Qu.: 49200		3rd Qu.:5
	Max. :101300		Max. :5

NBPLACES.y	CATALOGUEID	
Min. :5	Min. : 1.0	
1st Qu.:5	1st Qu.: 78.0	
Median :5	Median :165.0	
Mean :5	Mean :154.7	
3rd Qu.:5	3rd Qu.:234.0	
Max. :5	Max. :270.0	

Y a que valeur 5 dans ce champs !!! Nous devrions nous demander s'il s'agit d'un problème de base de données, s'il s'agit d'un problème de base de données, nous devrions essayer de restaurer ou de supprimer cette variable directement.

6.3 La manque de la mélangelement de la base de données: Cross-Validation

Quand nous avons séparé notre base en 3 tiers et les 2 premiers tiers pour l'apprentissage et le dernier tier pour la prédiction, cela causera l'abaissement de la fiabilité du cadre (Les anciennes données ne sont pas aussi fiables que les données actuelles). En plus, nous devons répéter le processus Echantillonnage - Apprentissage - Prédiction plusieurs fois jusqu'à la performance de notre classifieur soit stable et le meilleur.

Si les taux de réussi ne sont pas des grandes différences, on ne peut pas confirmer lequel classifieur qui performance mieux vise à ce modèle. Afin de diminuer les erreurs et garantir ces petites différences sont la vérité au lieu d'être générées aléatoirement, il faut répéter l'expérimentation et diviser l'ensemble d'entraînement plusieurs fois, il y a une méthode pour résoudre ce problème, on appelle ce processus Cross-Validation.

On utilise plutôt K-fold Cross Validation : Notre ensemble de test ne contiendra plus qu'une seule donnée, mais plusieurs, et le nombre spécifique sera déterminé en fonction de la sélection de K. Par exemple, si K = 5, les étapes que nous suivons pour utiliser la validation croisée en cinq volets sont les suivantes :

- Divisez tous les ensembles de données en 5 parties
- Prenez une copie à chaque fois comme ensemble de test sans le répéter, et utilisez les quatre autres copies comme ensemble d'apprentissage pour entraîner le modèle, puis calculez la MSE du modèle sur l'ensemble de test.
- Prenez la moyenne des 5 fois pour obtenir la MSE finale

Méthode de réalisation concrète

```
47 set.seed(1234)
48 ind<-sample(2,nrow(tableFinal),replace=TRUE, prob=c(0.7,0.3))
49 table_EA <- tableFinal[ind==1,]
50 table_ET <- tableFinal[ind==2,]
51 |
```