Etude de cas UE STA211

Géolocalisation de véhicules professionnels

Maya Besma Le Corre

Table des matières

1Page de garde.....1
2Présentation de l'étude.....3
2.1Présentation générale.....3
2.1.1Présentation de la société et des données.....3
2.1.2Problématique métier.....3
2.2Préparation des données.....3
2.2.1Inventaires des données.....3
2.2.2Qualité des données.....4
2.2.3Nettoyage.....5
2.2.4Transformation.....5
2.3Objectif de l'étude.....7
3Exploration des données.....8
3.1Analyse statistique des variables.....8
3.1.1Description des données.....8
3.1.2Distribution des variables centrées et réduites.....15
3.1.2.1Variables initiales.....15
3.1.2.2Selection de variables.....17
3.1.2.3Sélection de la valeur cible entre satisfaction et target.....17
3.1.2.4Manuelle.....22
3.1.2.5Sélection automatique à l‘aide d‘un arbre de décision.....23
3.2Analyse multi-variée.....25
3.2.1ACP avec le tableau initial.....25
3.2.1.1Données en entrées.....25
3.2.1.2Analyse de correlation.....27
3.2.1.3Description des dimensions.....29
3.3Tableau individus x variables.....32
4Modélisation.....34
4.1Echantillonage pour l‘apprentissage.....34
4.2Description des prédicteurs et de la variable à prédire.....35
4.3Regression Logistique.....36
4.4KNN.....37
4.5Arbre de décision.....38
5validation des modèles.....39
6Conclusions.....41

1 Présentation de l'étude

1.1 Présentation générale

1.1.1 Présentation de la société et des données

- La société : La société qui a communiqué les données gratuitement souhaite que celle ci-restent confidentielles et donc ne souhaite pas divulguer son identité. Les données sont issues d'une solution complète de Géoproductivité adossée à un dispositif de géolocalisation de véhicule professionnel. Des rapports sur les déplacements et des analyses de la productivité des opérateurs sont proposés aux utilisateurs du service (sur la base des traces GPS enregistrées par les balises connectées au serveur et branchées aux véhicules de l'entreprise). Les données sont collectées par un système centralisé, hébergé sur le cloud. Le service englobe une interface web pour les utilisateurs sédentaires, une application smartphone et des balises GPS pour les utilisateurs mobiles.

- Les données : Le fichier extrait de la base centrale contient plus de $2x10^6$ relevés GPS en continue (24h/7j). L'échantillon de données retenu pour l'étude concerne 3 clients et une vingtaine de véhicules répartis sur 3 départements du Sud de la France sur la période du mois de janvier 2018. Cet échantillon constitue un tableau de plus de 160 000 observations après dédoublonage.
- Documentation sur les donées : La documentation technique sur le fonctionnement des balises est traduite du chinois vers un anglais approximatif. Donc l'analyse va chercher à mieux comprendre le comportement des balises.

1.1.2 Problématique métier

- Contexte :

Le service de géolocalisation doit réactualiser à intervalle régulier la position GPS des véhicules. Dans les faits l'intervalle de temps n'est pas toujours régulier ni satisfaisant. Cette situation engendre de nombreux appels et une défiance vis à vis de la fiabilité de la solution applicative.

- Problématiques :

Existe-il une explication qui pourrait être déduite des données elle-même ?
Existe-il un modèle qui permettrait de décrire ces situations atypiques ? Peut - on prédire l'insatisfaction ?

- Hypothèse métier :

Les délais s'allongent quand la couverture GSM est mauvaise en certains endroits.

1.2 Préparation des données

1.2.1 Inventaires des données

Le jeu de données est complet mais contient des mesures aberrantes liées à des évènements exceptionnels expliqués et d'autres mesures, dans une proportion à déterminer, sont qualifiées d'anomalies inexpliquées car

générant de l'insatisfaction client.

Le jeu de données ne permet pas de répondre directement à la question métier posé. Il faut le transformer pour calculer la variable à expliquer. Il faut également définir le tableau à analyser pour répondre à la problématique car plusieurs approches sont possibles. Le tableau de valeurs numériques de 16 variables et plus de 2,5 millions d'observations formées par 85 balises distinctes et indépendantes va être explorer pour déduire plusieurs points :

- à partir de quelle valeur considère t on qu‘un délai n‘est pas satisfaisant ?
- quel sera la définition des individus et des variables à étudier ?
  - Les observations peuvent être regroupées par balise pour l'analyse exploratoire formant ainsi des groupes d'individus. Au sein de chaque groupe les observations ne sont pas indépendantes. Il existe une relation temporelle puisque ce sont des mesures répétées dans un intervalle de temps court (quelques secondes).
  - On peut considérer les balises comme des individus et aggréger les valeurs de sorte à conserver les informations nécessaires à l'étude en ne tenant pas compte de la dimension temporelle.
  - On peut aussi considérer un ensemble d'observations dans l'ensemble sans distinction entre les balises.

  - Jeu de données initial : Info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2562750 entries, 0 to 2562749
Data columns (total 16 columns):
DateReception_HeureReception    datetime64[ns]
DateGPS_HeureGPS                datetime64[ns]
idBalise                        int64
ModemId                         int64
Longitude                       float64
Latitude                        float64
Vitesse                         int64
Direction                       int64
Altitude                        int64
Satellites                      int64
MessageId                       int64
Puissance                       float64
VoltageBatterie                 float64
Distance                        int64
Information1                    int64
Information2                    int64
dtypes: datetime64[ns](2), float64(4), int64(10)
memory usage: 312.8 MB
```

1.2.2 Qualité des données

- aucune donnée manquante
- ModemId et idBalise sont indentiques : informations redondantes
- vitesse (de 0 à 160 km/h) et direction (0 à 360°) ont des plages de valeurs réalistes
- l'altitude est parfois négative : peut - être un problème de réglage ou un soucis de précision de la mesure ou bien est-ce justifié ?
- le nombre de satellites varie de 0 à 12 : plus le nombre de satellite est grand plus la position est précise selon les informations techniques fournies par le fabriquant des balises

- Information 1 et 2 sont toutes présentes mais non documentées ni exploitées par l'éditeur de logiciel : l'interprétation est donc impossible

1.2.3 Nettoyage

Pour information, les données et traitement associés ont fait l'objet d'une déclaration à la CNIL par les employeurs et par la société qui fourni ces données.

Le nettoyage a simplement consisté à supprimer les observations du week-end car les trajets ayant un caractère personnel ne peuvent pas faire l'objet de traitement analytique en regard du respect du droit à la vie privée.

Quelques doublons ont été supprimés également sur une clé reprennant l'ensemble des variables issues du tableau initial sans tenir compte de la date / heure de reception (date d'enregistrement sur le serveur).

1.2.4 Transformation

- Le besoin métier est d'analyser le délai de transmission des données de chaque balise vers le serveur. Les données brutes date et heure d'enregistrement de la position et la date et heure de l'enregistrement sur le serveur doivent être transformées pour calculer un délai.

- Plusieurs délais peuvent être calculés : le délai entre le moment du relevé GPS et le moment où le serveur reçoit et l'enregistre ; le délai qui s'écoule entre deux relevés effectués par une balise ; le délai qui s'écoule entre deux enregistrements sur le serveur pour une balise (en seconde)

- On dispose également de la distance cumulée dans chaque relevé effectué par chaque balise. On peut donc calculer la distance parcourue entre deux relevés (en mètre).
-

La transformation consiste donc à caluler les variables supplémentaires à étudier en tant que variables explicatives pour le différentiel de distance et en tant que variable à expliquer pour le délai le plus qualitatif.

- Transformation pour la phase d'exploration :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 165969 entries, 0 to 165968
Data columns (total 23 columns):
idBalise             165969 non-null int64
vitesse              165969 non-null int64
direction            165969 non-null int64
altitude             165969 non-null int64
satellites           165969 non-null int64
messageId            165969 non-null int64
distance             165969 non-null int64
delaiServeur         165969 non-null int64
delaiBalise          165969 non-null int64
deltaDistance        165969 non-null int64
lng                  165969 non-null float64
lat                  165969 non-null float64
power                165969 non-null float64
volt                 165969 non-null float64
labelBalise          165969 non-null object
labelMessage         165969 non-null object
delaiTransmission    165969 non-null int64
jourBalise           165969 non-null int64
jourServeur          165969 non-null int64
rangDateGPS          165969 non-null int64
rangDateRec          165969 non-null int64
client               165969 non-null int64
labelClient          165969 non-null object
dtypes: float64(4), int64(16), object(3)
memory usage: 29.1+ MB
```

Les principales transformations sont les suivantes :

- recodage de variables quantitatives en variables qualitatives : idBalise, messageId (en regroupant les classes similaires), client ;
- discrétisation : lng, lat

Les données numériques calculés :

- delaiTransmisison : Date_Heure Serveur - Date_Heure Balise
- delaiBalise : Date_Heure de la position future - Date_Heure de la position courante
- delaiServeur : Date_Heure de l'enregistrement futur - Date_Heure de l'enregistrement courant
- client : deduit des 2 premiers digits du numéro de balise
- jourBalise et jourServeur : calcul du jour de la semaine à partir de la date de la balise et de la date du serveur

Les données qualitatives calculés :

- labelClient : C+code du client
- labelMessage : M+code du message
- labelBalise : B+code de la balise
- zone : concaténation de la classe longitude (lngBin) et la classe latitude (latBin)
- individu : concaténation du numéro de balise avec la zone du relevé
- satisfaction : 2 classes à déterminer (délais corrects et délais anormaux)
- target : : 2 classes à déterminer en fonction d’un seuil (à calculer)

1.3 Objectif de l'étude

Problématique data

Le premier problème est la détermination des “individus“ et des “variables“ à étudier pour vérifier l‘hypothèse métier. La phase exploratoire va permettre de le déterminer en étudiant les relations entre les variables. De même l‘étude aidera à définir les règles d‘étiquetage pour la classification supervisée.
La phase de modélisation va aborder la problématique dans une optique de classification binaire des observations qui induisent de l'insatisfaction ou pas afin de déterminer un modèle prédictif supervisé (problème de classification binaire).

En phase exploratoire

- Déterminer le tableau individus x variables
- Etiquetter les observations pour la classification supervisée
- Expliquer le délai (variable à expliquer)

En phase de modélisation

- Prédire la classe d'appartenance : “satisfaction“
  - Les méthodes choisies pour la prédiction sont les suivantes :
    - K plus proches voisins : car l‘aspect topologique présente un intérêt non négligeable avec des données géolocalisées mais aussi parce que nous supposons que des individus avec des variables similaires et donc proches vont donner un meilleur résultat.
    - Regression logistique binaire qui présente l‘avantage également d‘avoir un modèle lisible et donc en phase avec le caractère explicatif de l‘objectif de l‘étude.
    - Arbre de décision pour l‘aspect lisibilité
  - Un échantilonage en train / test / validation est largement possible vu la taille de l‘échantillon. Sur le tableau individus x variables la taille plus petite ne le permet pas.

2 Exploration des données

2.1 Analyse statistique des variables

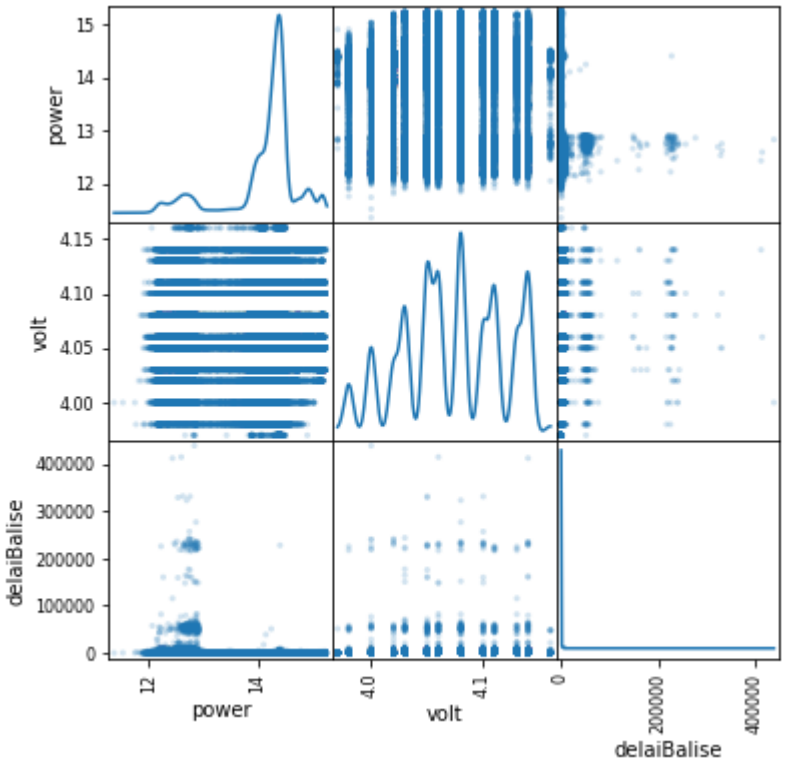
2.1.1 Description des données

• Description de “power“ (en volt), “volt“ (en Ah)

Ces variables caractérisent l'état du véhicule avec la puissance et le voltage de la baterie. Les valeurs sont plutôt concentrées sur une petite plage. La moyenne est significative.

	power	volt
count	161789.000000	161789.000000
mean	14.145729	4.073147
std	0.628871	0.045684
min	11.360000	3.970000
25%	14.060000	4.050000
50%	14.320000	4.080000
75%	14.430000	4.110000
max	15.260000	4.160000

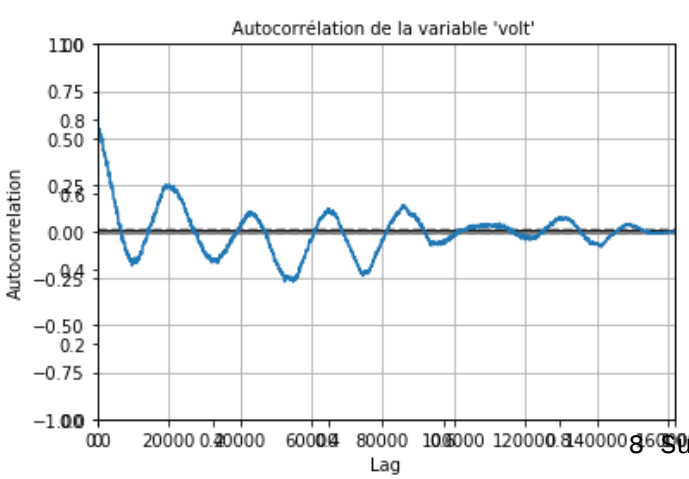
On observe les courbes de densité des variables power et volt sur la diagonale. La densité de la variable volt présente de nombreux pics typique des mesures de signaux electriques.

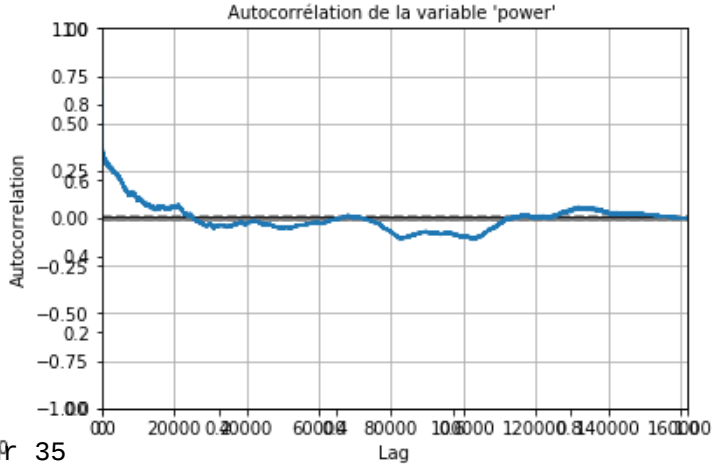


La variable volt est indépendante du délai de la balise alors que la variable power semble avoir un lien.

La puissance semble être répartie selon deux ordres de grandeurs : valeurs basses et valeurs hautes. La densité des mesure de délai d‘emission plus long se concentre pour les valeurs de puissance basses. Le rapport est inversé.

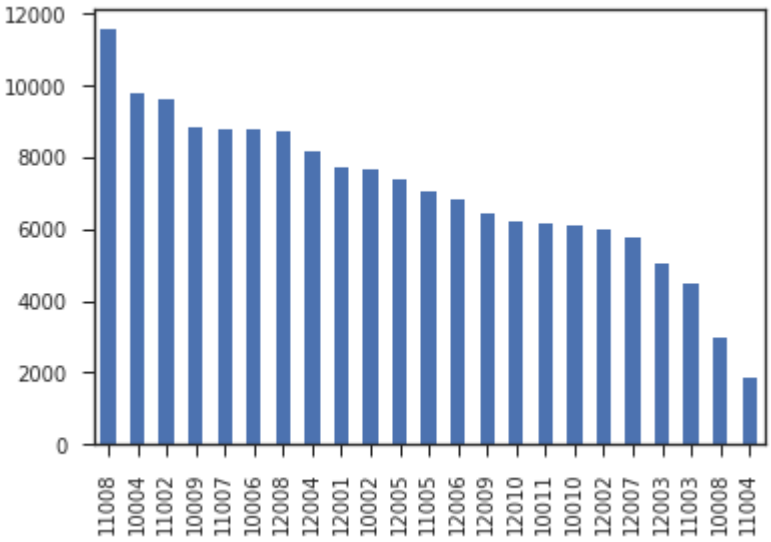
Les observations étant des mesures répétées l‘autocorrélation entre observations est représentée dans ces graphiques pour chacune des 2 variables :





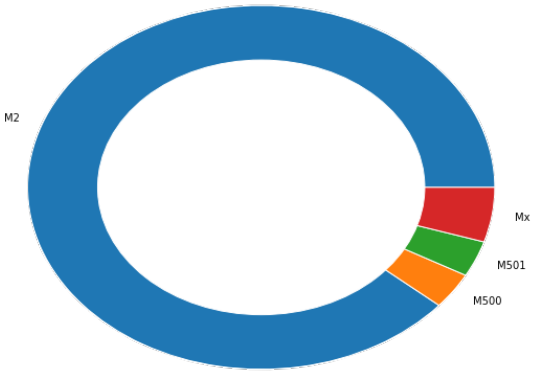
- idBalise

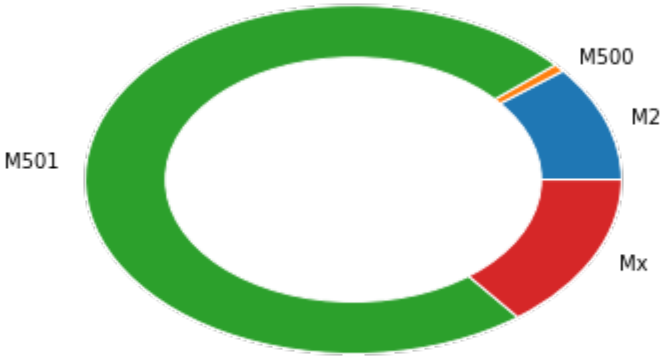
L'idBalise est une valeur numérique discrète qui identifie une balise donc dans les faits il s'agit d'une modalité. La répartition des observations entre chacune des balises est inégale. Le diagramme ci-contre regroupe le nombre d'observations par balise pour l'échantillon de l'étude.



Les balises retenues sont celles qui ont des mesures prises entre les 31/12/2017 et le 31/1/2018.

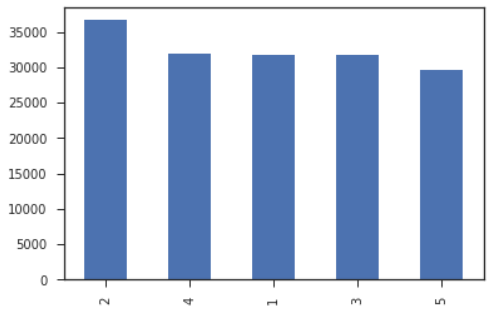
- Decsription de la variable messageId



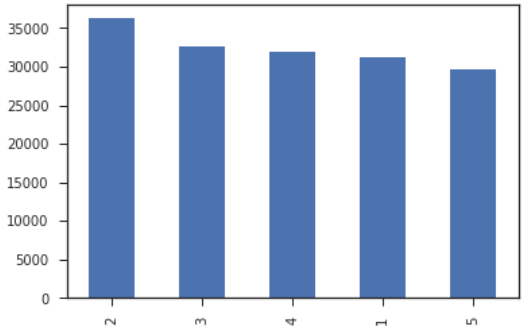


On constate une répartition inégale des observations par messageId en fréquence à gauche et à droite en délaiBalise cumulé.
D'après la documentation, le type 2 signifie que le véhicule est en cours de déplacement. Parmis les autres messages moins représentés, la moitié des autres messages représentés sont les types 500 et 501 (changement d'état de la balise selon la documentation technique). Un recodage des autres valeurs en une classe Mx regroupant les plus petites modalités en nombre.

- jourBalise et jourServeur



jourBalise



jourServeur

Les variables jourBalise et jourServeur donnent la meme information. On consate une répartition équilibrée entre les jours ouvrés.

- Description des variables : ‘lng','lat','altitude','satellites','direction','vitesse'

	lng	lat	altitude	satellites	direction	vitesse
count	161789.000000	161789.000000	161789.000000	161789.000000	161789.000000	161789.000000
mean	5.976357	43.440904	86.265265	8.056432	184.141073	45.853637
std	0.791364	0.184252	82.481048	2.067119	103.368115	31.752229
min	4.788180	43.079900	-338.000000	0.000000	0.000000	0.000000
25%	5.399390	43.287690	27.000000	7.000000	93.000000	22.000000
50%	5.488970	43.356660	63.000000	8.000000	183.000000	41.000000
75%	7.057930	43.656750	118.000000	9.000000	274.000000	69.000000
max	7.319920	43.831350	665.000000	12.000000	360.000000	154.000000

Les coordonnées GPS (lng et lat) présentent des écart-type assez faibles. Pour mieux étudier ces variables on discrétise les coordonnées et on les transforme en classe pour représenter une zone. Ceci en vue d‘étudier l‘hypothèse métier.

La moyenne de la longitude et de la latitude donne un centre géographique moyen autour duquel les véhicules effectuent leur parcours. La direction du véhicule est un angle variant de 0 à 360°. Toutes les valeurs possibles sont représentées et les métriques montrent une répartition équiprobable.

La vitesse moyenne est significative. Ce groupe de variables ne sauraient répondre à l‘étude de l‘hypothèse métier en aggrégeant par balise uniquement et en effectuant des moyenne car pour analyser l‘influence de la zone géographique sur le délai il faut plusieurs obervations. Le choix est donc de regrouper les variables par couple balise – zone géographie. Ce couple forme les individus à étudier.

- Description de la variable : ‘satellites‘

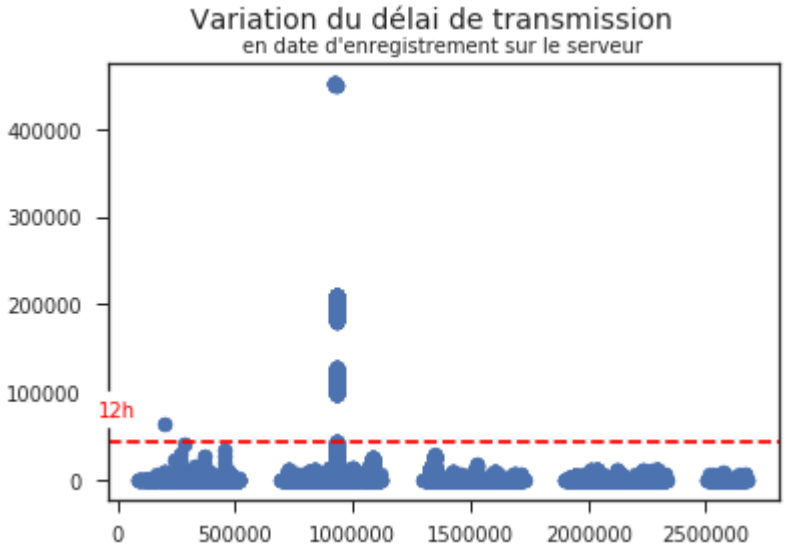
```
count 161789 ; mean 8 ; std 2 ; min 0 ; 25% 7 ; 50% 8 ; 75% 9 ; max 12
```

Le nombre de satellites permet de fiabiliser la position GPS. Dans cet échantillon, les positions GPS sont cohérentes malgrès l'absence de satellites signalés lors des relevés par les balises. La moyenne est significative smais dans le regroupement par individus la même information serait répété donc cette variable ne présente pas d‘intérêt.

- Description des variables : 'delaiTransmission', 'delaiBalise', 'delaiServeur', 'deltaDistance'

	delaiTransmission	delaiBalise	delaiServeur	deltaDistance
count	161789.000000	161789.000000	161789.000000	161789.000000
mean	1333.727590	323.709875	324.529857	267.964009
std	13650.271791	5644.235737	5766.628391	198.736905
min	-120.000000	0.000000	-286.000000	0.000000
25%	0.000000	20.000000	20.000000	153.000000
50%	1.000000	20.000000	20.000000	219.000000
75%	1.000000	25.000000	24.000000	378.000000
max	452207.000000	438085.000000	450017.000000	7274.000000

On trouve des valeurs très extrêmes pour le délai de transmisison :
Cette valeur maximale (452207 secondes ; presque 70h de délai) est liée à une situation anormale qui s'est produite en début de période (mois de janvier 2018).



Le phenomène observé est le cas d'une balise (la 10004) qui n'a pu transmettre les relevés pendant un certain temps (44h) et qui envoi tout l'historique de la trace GPS (892 relevés) en plusieurs paquets.

On observe ce phénomène uniquement sur le délai de transmission. Les délais d'emission et de réception sont inférieurs à 20 secondes.

Dans la transformation du tableau individus x variables, il faut aggréger les observations qui apportent une information sur le délai.

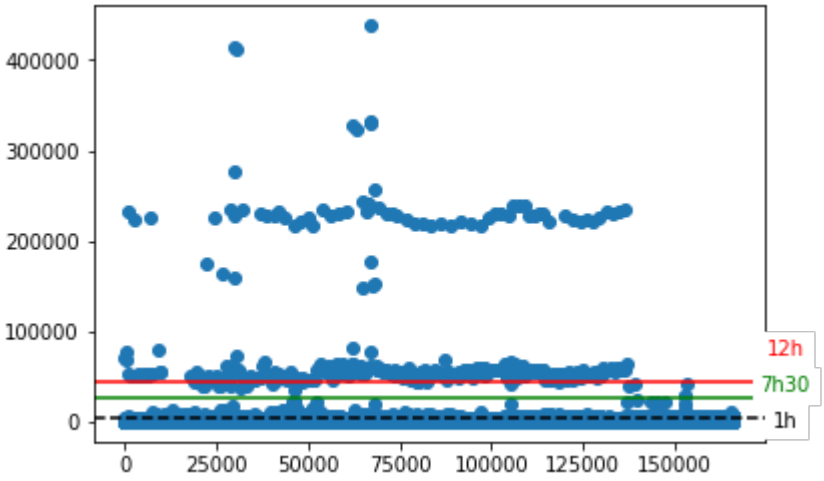
Comme nous sommes dans une approche de classification binaire, nous

conservons les valeurs extrêmes et nous les étiquetterons comme insatisfaisantes bien que liées à un évènement exceptionnel.

La proportion étant très faible cela impacte peu la proportion entre les deux classes.

On doit trouver la règle d'étiquetage pour permettre l'apprentissage supervisé.
Etudions donc les différents délais calculés :

Un délai de transmission négatif signifie que les horloges du serveur et des balises ne sont pas réglés de la même manière. Le délai ne peut être calculé de manière correcte sur cette base.
Un délai négatif entre deux enregistrements successifs sur le serveur (Date heure Serveur) signifie que le serveur n'a pas enregistré/reçu les positions des balises dans l'ordre chronologique d'enregistrement des positions par la balise GPS.

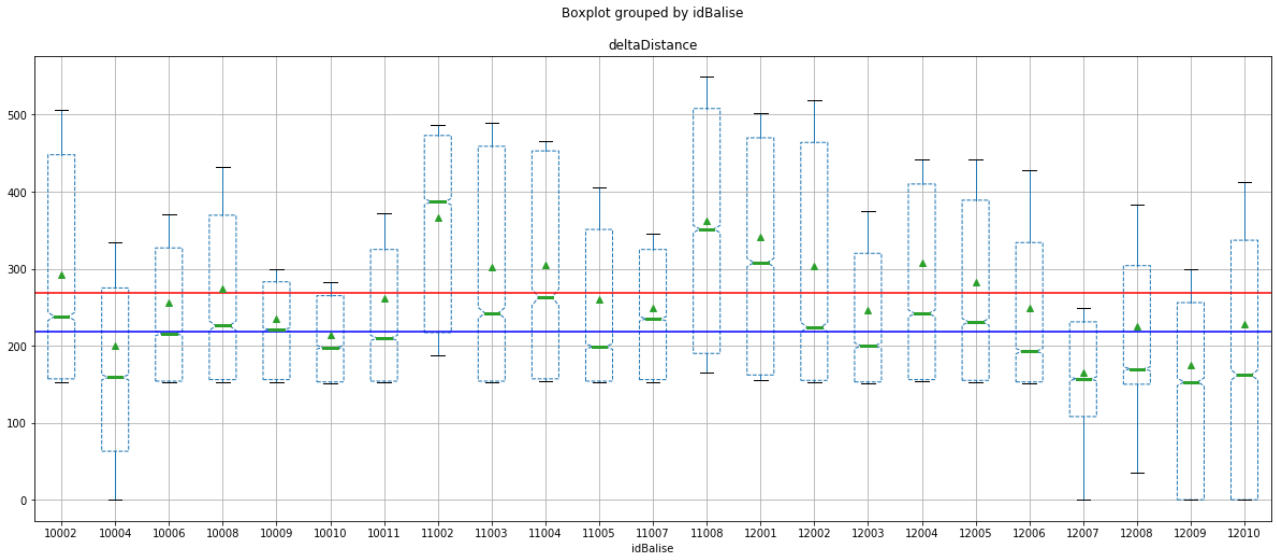


Nuage de points de la série "Delai d'emission de la balise" (delaiBalise)

Aucun délai négatif entre deux relevés succesifs puisque le tri des observations a été établi avant les calculs selon l'ordre d'enregistrement des positons GPS.
Le délai de transmission n'est pas une variable exploitable pour cette étude. Le délai serveur est légèrement décalé par rapport au délaiBalise mais fourni la même information.

Remarque : On considèrera pour la suite uniquement le délai calculé entre deux mesures effectuées par les balises indépendantes et basées sur la même horloge.

- Analyse du deltaDistance par Balise (distance parcourue entre 2 relevés) :



Les variations de deltaDistance restent comparables entre les balises bien qu'il existe des valeurs extrêmes assez disparates d'une balise à l'autre sans doute liées à des phénomènes exceptionnels.

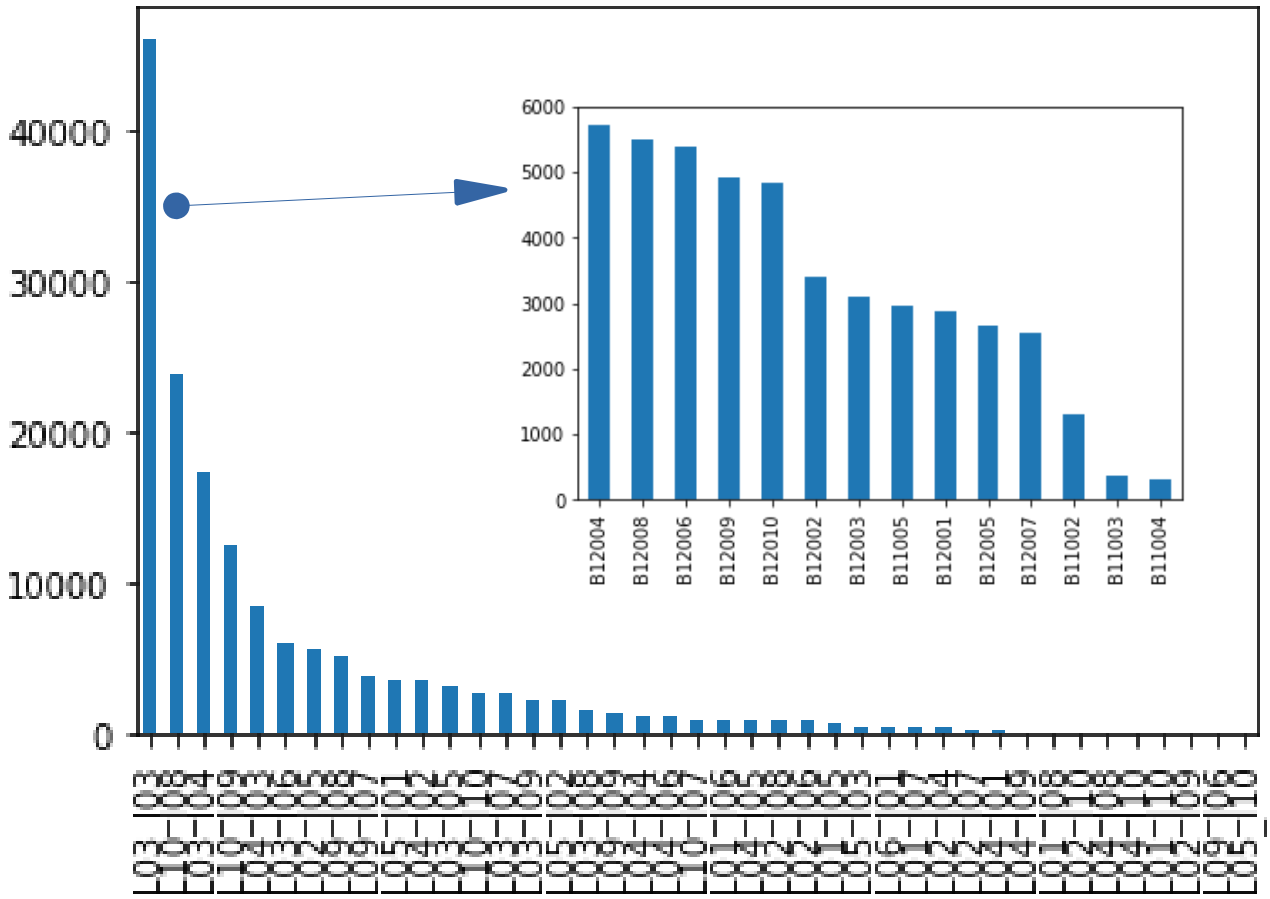
```
Count: 161789; mean:268(ligne rouge); std: 199; min : 0; max: 7274(en
mètre) ; 25%  : 153 ;  50% : 219 (ligne bleue) ; 75% : 378
```

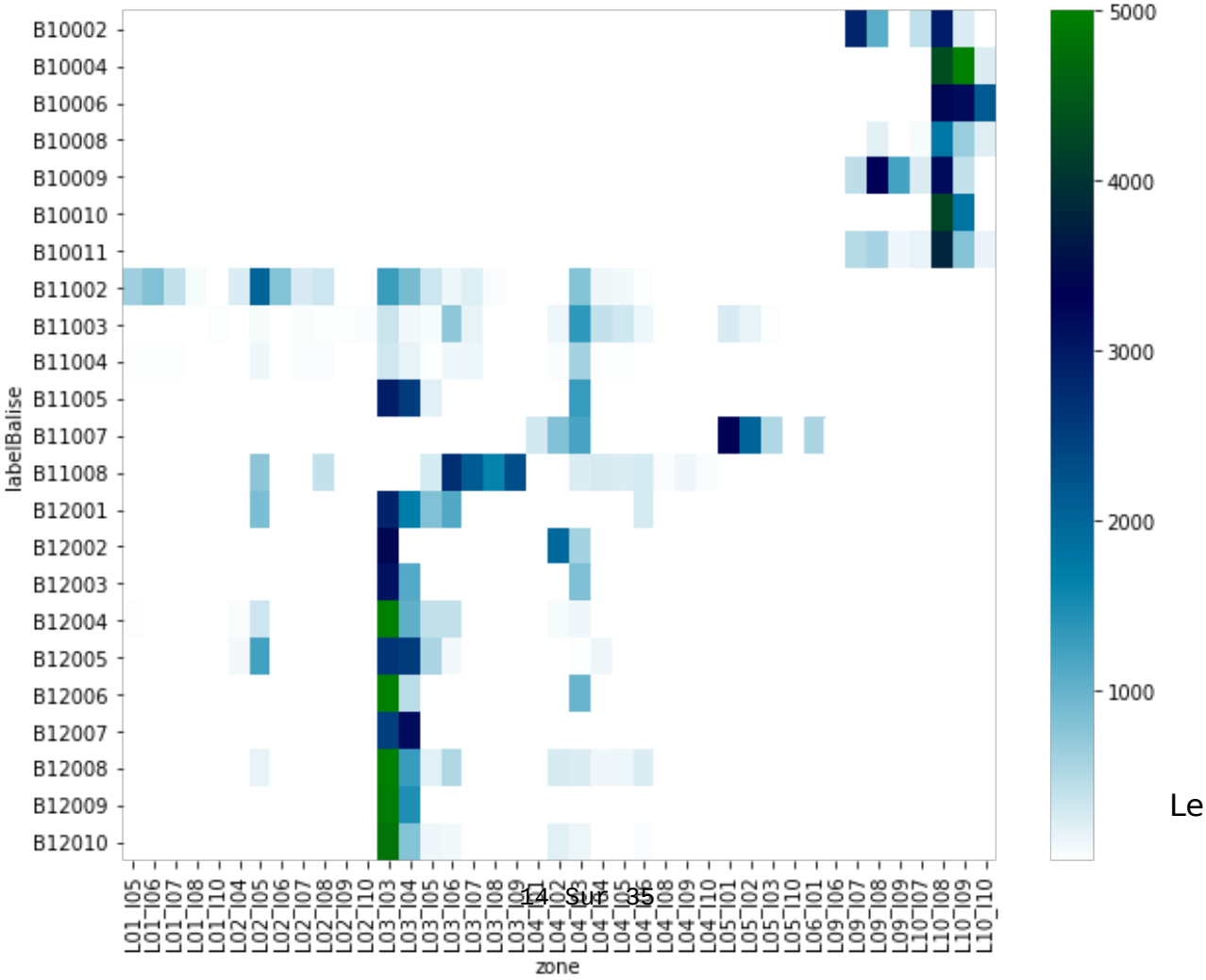
Cette valeur permet de créer deux classes en fonction de l'état : les observations pendant que le véhicule roule et les observations pendant l'arrêt du véhicule. La nouvelle variable état prend deux valeur : roule/stop (valeur stop est attribuée si la vistesse = 0 et delta de distance = 0 et messageId différent de 2 ; la valeur roule est attribuée sinon).

L'exploration des données montre une différence d'échelle entre les valeurs et de grands déséquilibres entre différents groupes d'observation (par message, par balise, par zone). Le jour de la semaine (jour ouvré uniquement) n'influence pas le délai d'emission de la balise. La moyenne de la variable à expliquer (le délai futur avant un nouveau relevé de la balise) n'est pas significative ni celles des coordinnées GPS.

Un regroupement par balise pour constituer un tableau d'individus oblige à aggréger les autres variables et la moyenne n'est pas toujours significative en particulier pour la variable à expliquer. La définition des individus repose principalement sur la volonté de conserver le maximum d'information par individu représentant aux mieux un groupes de mesures.

Observons la fréquentation des zones géographiques discrétisées :
La zone la plus fréquentées est la zone L03_l03 {Longitude : (5.295, 5.548] ; latitude : (43.23, 43.305]} par le trio de balises en tête du classement B12004, B12008 et B12006.


40000
30000
20000
10000
0
6000
5000
4000
3000
2000
1000
0
B12004
B12008
B12006
B12009
B12010
B12002
B12003
B11005
B12001
B12005
B12007
B11002
B11003
B11004

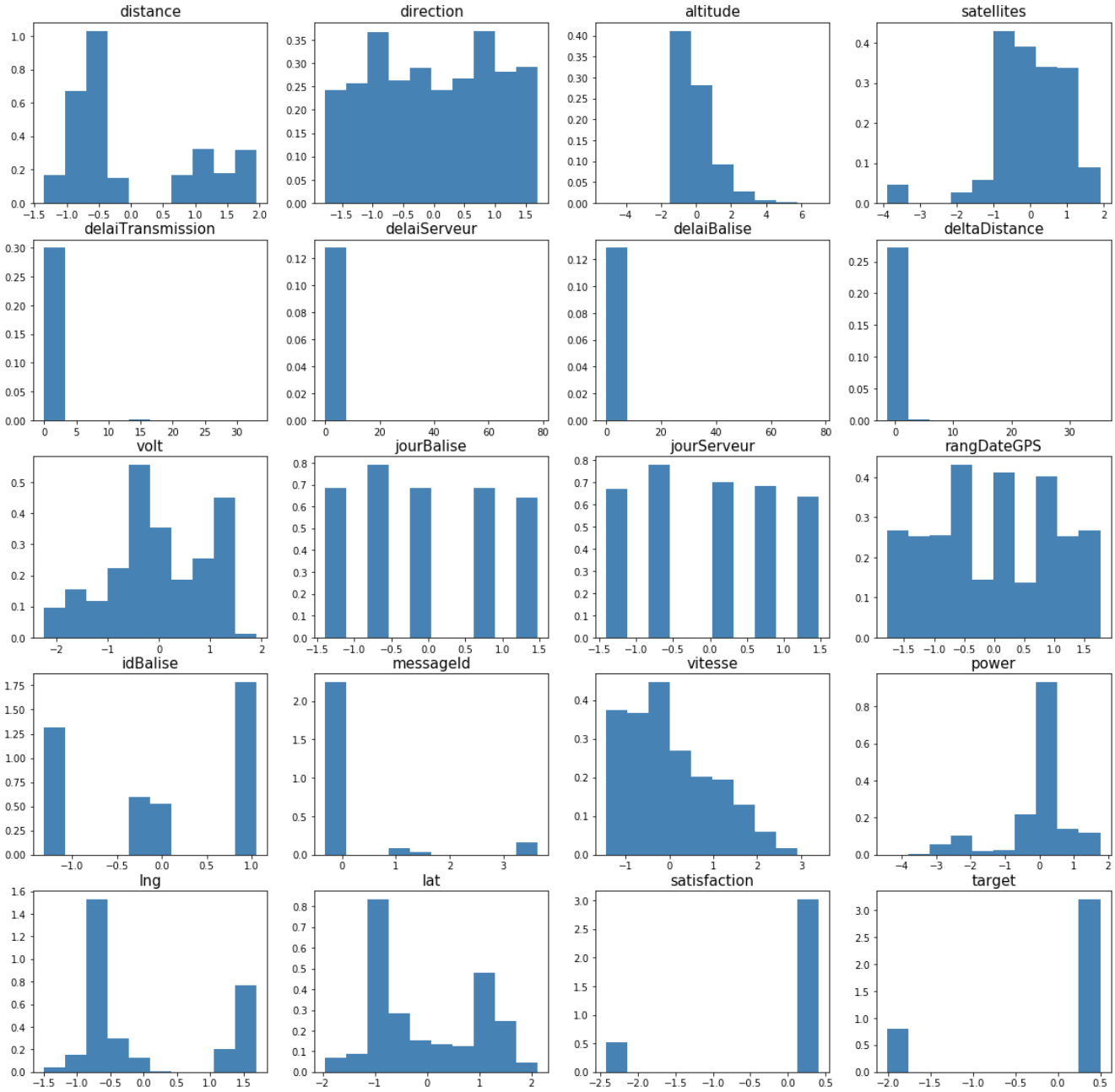

labelBalise
B10002
B10004
B10006
B10008
B10009
B10010
B10011
B11002
B11003
B11004
B11005
B11007
B11008
B12001
B12002
B12003
B12004
B12005
B12006
B12007
B12008
B12009
B12010
zone
5000
4000
3000
2000
1000

diagrame ci-contre compte pour chaque couple Balise / Zone géographique le nombre d'observation de l'échantillon. On constate une concentration géographique sur environ 40 zones différentes avec moins cinq zones très fréquentées.

2.1.2 Distribution des variables centrées et réduites

2.1.2.1 Variables initiales

Les variables explicatives centrées et réduites ne suivent pas une distribution gaussienne. La variable vitesse est la plus proche d'une loi normale.



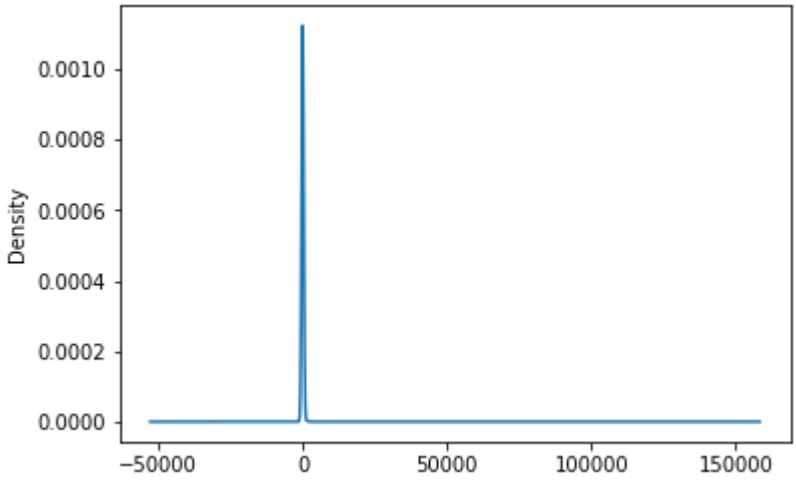
Calcul du coefficient d'applatissement et du coefficient d'asymétrie des variables :

	Kurtosis	Skewness
distance	-1.10271154722	0.722170378417
direction	-1.22337448914	-0.037547220342
altitude	4.16012500262	1.76156330856
satellites	4.18651251054	-1.53615150602
delaiTransmission	349.834612677	16.5731444628
delaiServeur	2002.99530541	40.6636945123
delaiBalise	1910.72310512	39.8826672776
deltaDistance	71.3441038119	4.14330923429
volt	-0.863693647295	-0.171416864673
jourBalise	-1.27167209814	0.0695726196964
jourServeur	-1.2604854504	0.0610219350038
rangDateGPS	-1.14649436732	0.0211500278918
idBalise	-1.57822934617	-0.212923992749
messageId	8.21273932866	3.10520850043
vitesse	-0.560429896814	0.489097704293
power	2.28044249246	-1.56772099566
lng	-1.31898058713	0.736428542228
lat	-1.39339507899	0.297269909897

2.1.2.2 Selection de variables

2.1.2.3 Sélection de la valeur cible entre satisfaction et target

La courbe de densité du délaiBalise est la suivante :



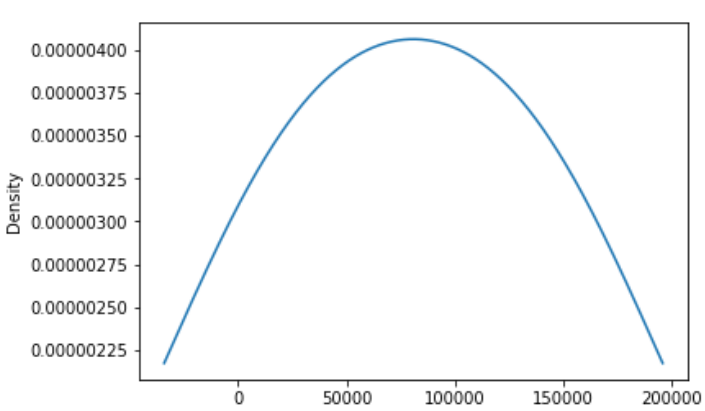
On voit qu'elle ne suit pas une distribution gaussienne.
On ne va donc pas utiliser directement cette variable pour la prédiction.
On la recode en fonction d'une règle d'étiquetage.

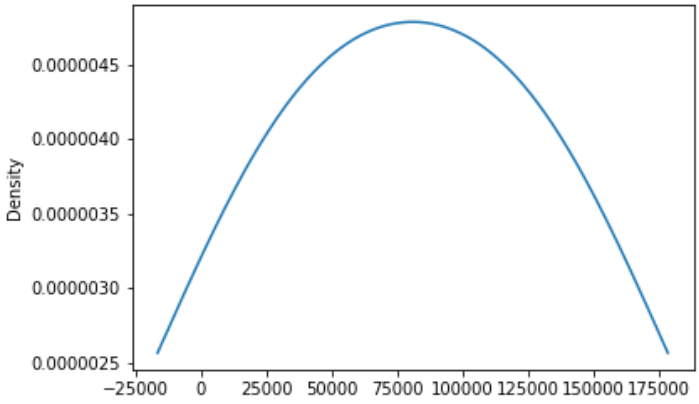
Deux valeurs cibles ont été calculées. Il faut en choisir une de la manière la plus objective possible. Deux possibilités :
- target : valeur 0 pour toutes les valeurs supérieures à 34 secondes (seuil fixé d'après l'étude de répartition 20-80 de l'ensemble de l'échantillon) et 1 en dessous.
- satisfaction : Même règle que la précédente sauf pour les observations quand le véhicule est à l'arrêt (état stop).

La variable "target" est choisie car elle présente deux avantages : plus proche d'une loi gaussienne symétrique (un peu plus aplatie) et plus neutre par rapport à la transformation des données.

Ci-dessous apparaissent les courbes de densité et les coefficients d'asymétrie et d'applatissement des deux variables :

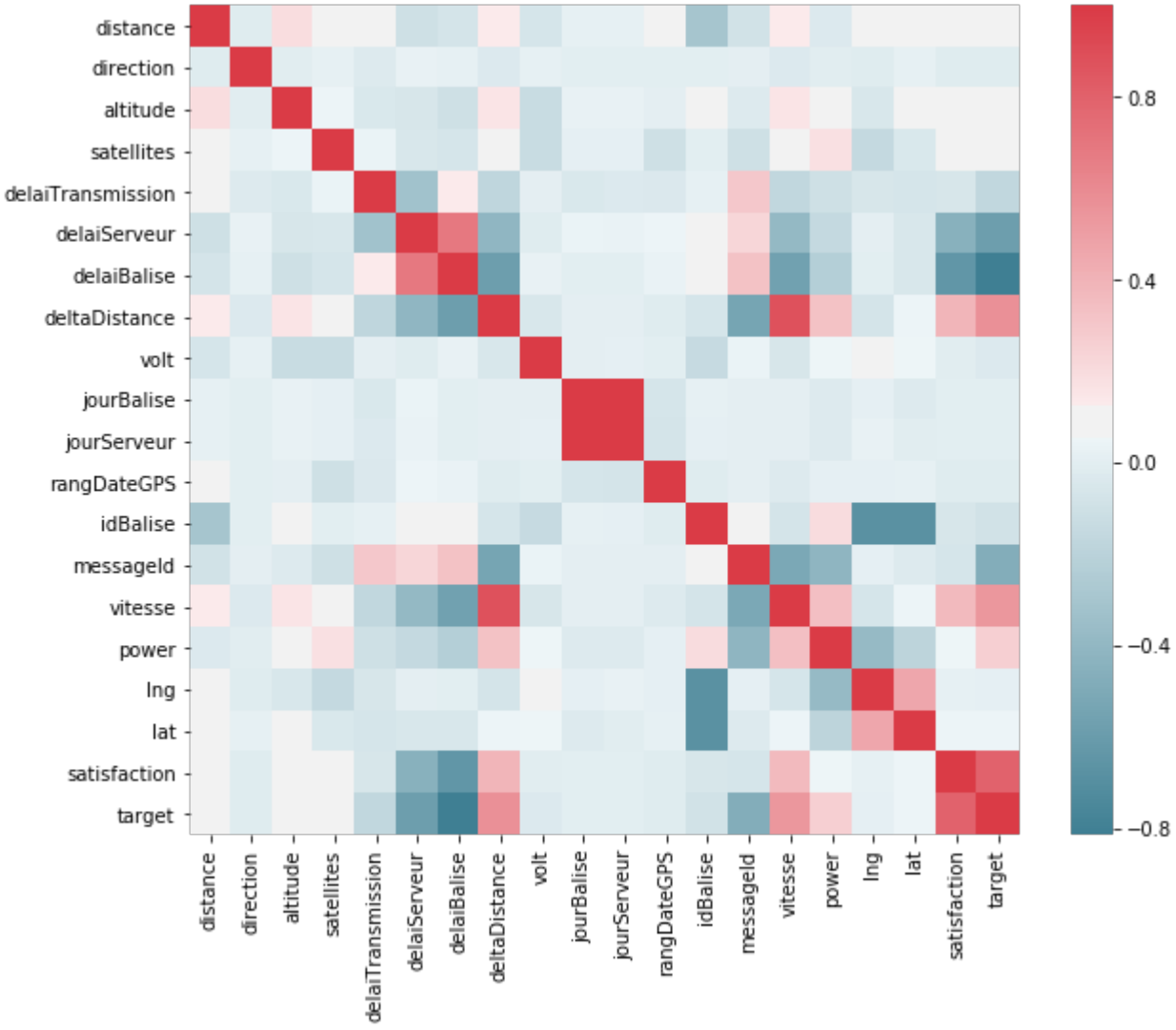
satisfaction : kurtosis = 2 et skew = -2 | target : kurtosis = 0 et skew = -1





2.1.2.4 Manuelle

Le calcul du V de cramer sur les variables initiales et une carte de chaleur avec la matrice de corrélation (test de Spearman) vont aider à selectionner des variables explicatives le plus indépendantes possibles entre elles pour éviter la redondance d'information dans l'équation du modèle.



Remarques importantes :
- DeltaDistance et vitesse sont positivement corrélées
- idBalise est trés corrélée à la longitude et la latitude
- DeltaDistance et delaiBalise sont aussi fortement corrélées

- target qui est un recodage de delaiBalise est corrélée à deltaDistance et vitesse

1. Coefficient de V de cramer (test de relation avec target et satisfaction) :

	Prédicteur	V_de_Cramer
[1,]	"delaiBalise"	"1"
[2,]	"target"	"1"
[3,]	"rangDateGPS"	"0.961200662743024"
[4,]	"distance"	"0.951941698292631"
[5,]	"delaiServeur"	"0.915862860312623"
[6,]	"satisfaction"	"0.800421324560655"
[7,]	"lng"	"0.731567706444927"
[8,]	"deltaDistance"	"0.713530940979035"
[9,]	"lat"	"0.637324435674061"
[10,]	"vitesse"	"0.627369308700483"
[11,]	"messageId"	"0.541431141902659"
[12,]	"power"	"0.429449125815572"
[13,]	"delaiTransmission"	"0.340187471254863"
[14,]	"balise"	"0.211586827565464"
[15,]	"altitude"	"0.159979875708145"
[16,]	"satellites"	"0.144298190104982"
[17,]	"direction"	"0.12174528291397"
[18,]	"volt"	"0.0557623558078282"
[19,]	"jourBalise"	"0.0178643665185515"
[20,]	"jourServeur"	"0.0163525152009615"

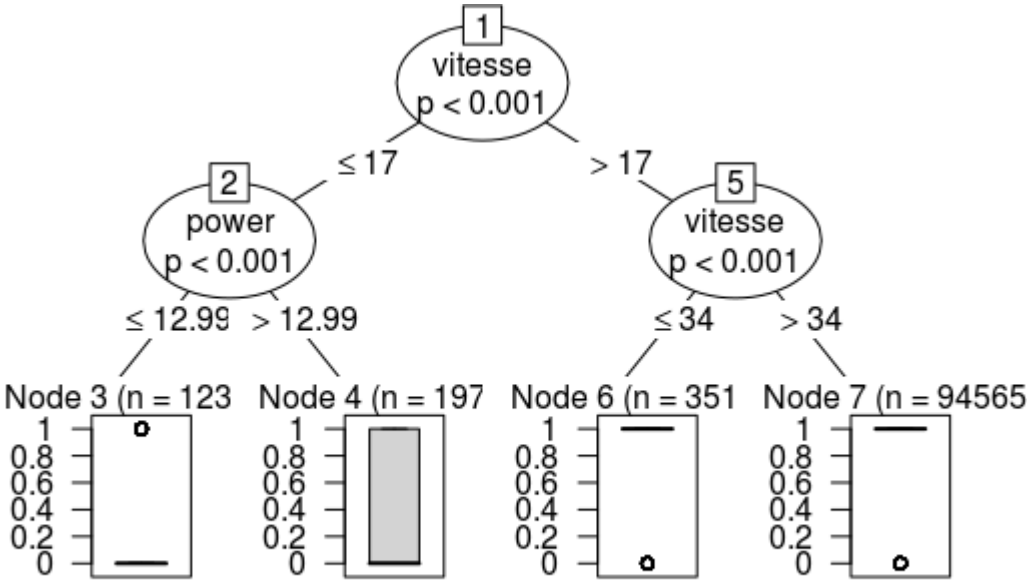
	Prédicteur	V_de_Cramer
[1,]	"satisfaction"	"1"
[2,]	"rangDateGPS"	"0.965626395954443"
[3,]	"distance"	"0.961071387528821"
[4,]	"delaiBalise"	"0.935080186643646"
[5,]	"delaiServeur"	"0.818310077934923"
[6,]	"target"	"0.800421324560655"
[7,]	"lng"	"0.68162117243938"
[8,]	"deltaDistance"	"0.634664699260478"
[9,]	"lat"	"0.584981804442235"
[10,]	"vitesse"	"0.489409100215287"
[11,]	"delaiTransmission"	"0.252849727147494"
[12,]	"messageId"	"0.192027899547548"
[13,]	"altitude"	"0.140434852503954"
[14,]	"power"	"0.140073941396213"
[15,]	"balise"	"0.127843924750387"
[16,]	"satellites"	"0.116634131053592"
[17,]	"direction"	"0.0994116504491581"
[18,]	"volt"	"0.0487156333805681"
[19,]	"jourBalise"	"0.0171119112158875"
[20,]	"jourServeur"	"0.0151736406694734"

2.1.2.5 Sélection automatique à l'aide d'un arbre de décision

En générant un arbre trés court on voit les variables les plus discriminantes. Avec un arbre fixé à une profondeur maximale de 2 on voit que les variables les plus discriminantes sont : vitesse et power de l'arbre ci-dessous à 4 feuilles.

```
Response:  target
Inputs:  delaiBalise, deltaDistance, volt, idBalise, messageId, vitesse, power, lng, lat
Number of observations:  161789

1) vitesse <= 17; criterion = 1, statistic = 40752.382
  2) power <= 12.99;
criterion = 1,
statistic = 3700.411
    3)*  weights = 12315
  2) power > 12.99
    4)*  weights = 19729
1) vitesse > 17
  5) vitesse <= 34;
 criterion = 1,
statistic = 9305.653
```

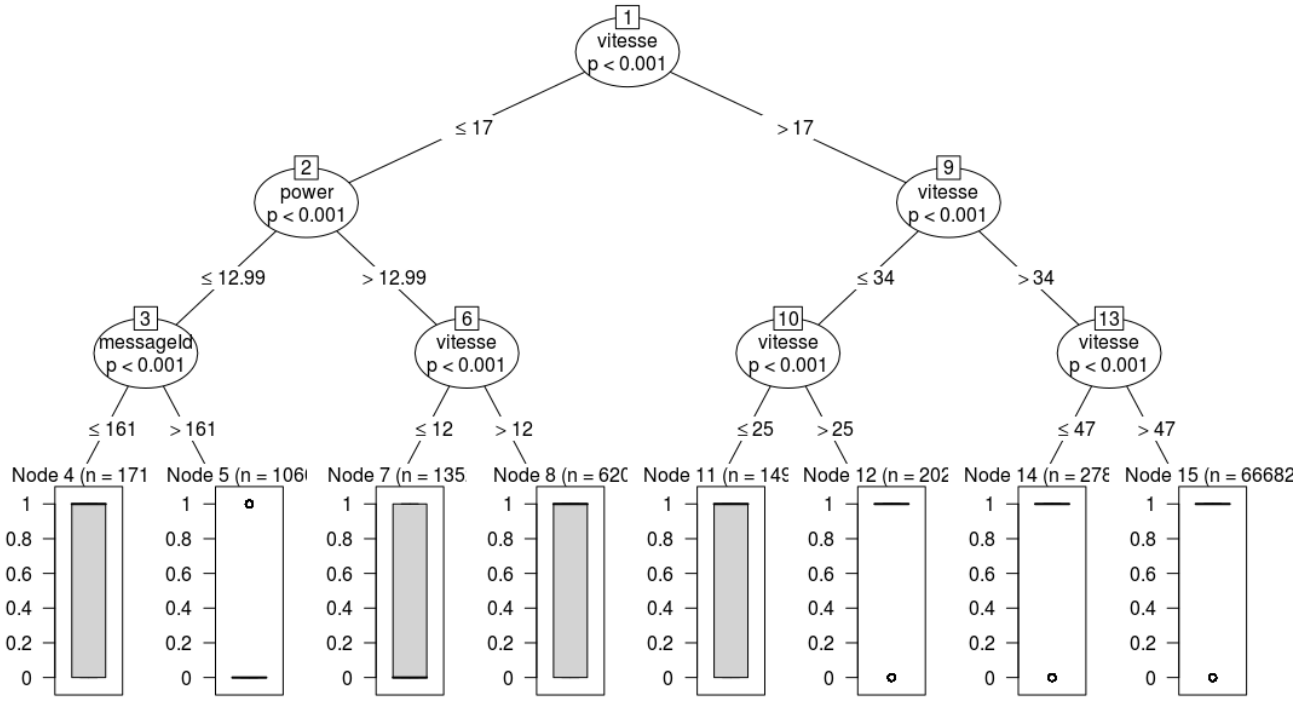
```
    6)*  weights = 35180
  5) vitesse > 34
    7)* weights = 94565
```



Les boites à moustaches au niveau des feuilles représentent la distribution des valeurs à partir du niveau 3.

On est curieux de voir avec un arbre de profondeur 3 :

```
Conditional inference tree with 8 terminal nodes

Response:  target
Inputs:  deltaDistance, messageId, vitesse, power, lng, lat
Number of observations:  161789

1) vitesse <= 17; criterion = 1, statistic = 40752.382
  2) power <= 12.99; criterion = 1, statistic = 3700.411
    3) messageId <= 161; criterion = 1, statistic = 1402.329
      4)*  weights = 1711
    3) messageId > 161
      5)*  weights = 10604
  2) power > 12.99
    6) vitesse <= 12; criterion = 1, statistic = 191.447
      7)*  weights = 13521
    6) vitesse > 12
      8)*  weights = 6208
1) vitesse > 17
  9) vitesse <= 34; criterion = 1, statistic = 9305.653
    10) vitesse <= 25; criterion = 1, statistic = 986.607
      11)*  weights = 14900
    10) vitesse > 25
      12)*  weights = 20280
  9) vitesse > 34
    13) vitesse <= 47; criterion = 1, statistic = 2554.888
      14)*  weights = 27883
    13) vitesse > 47
      15)*  weights = 66682
```



NB: en annexe se trouve le résultat détaillé avec une profondeur de 5.

2.2 Analyse multi-variée

2.2.1 ACP avec le tableau initial

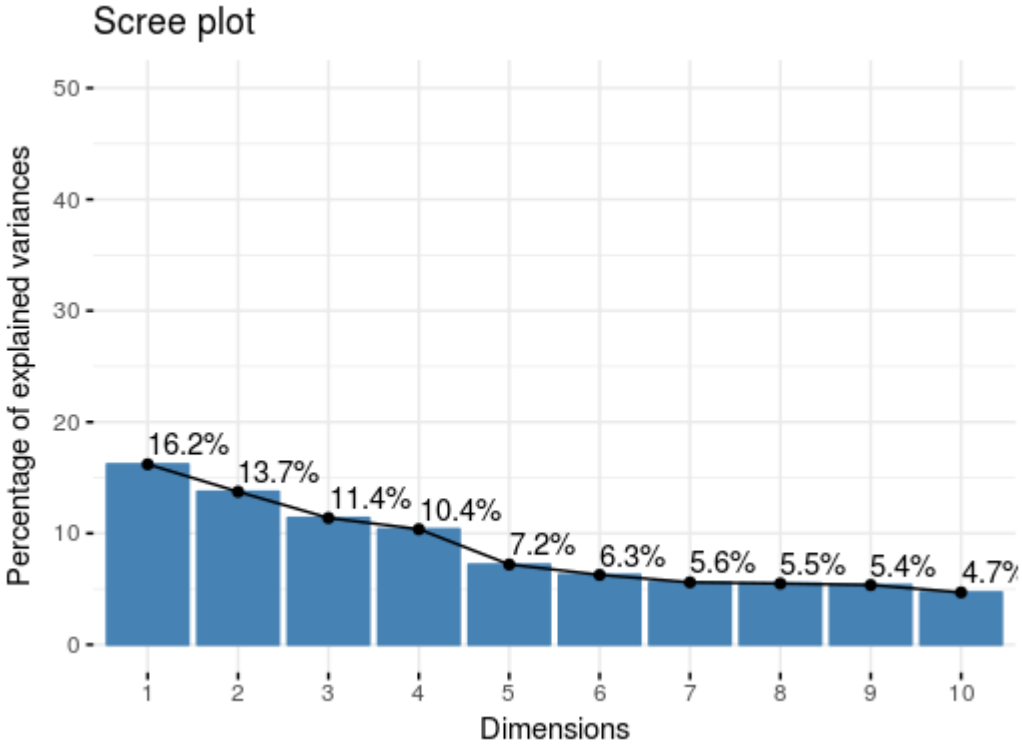
2.2.1.1 Données en entrées

Le tableau en entrée est constitué des 11 variables initiales ("volt","idBalise","messageId", "vitesse", "power", "lng", "lat") et 9 variables recodées et calculée. Les variables "satisfaction", "target" sont utilisées comme variables supplémentaires (même si le choix entre ces deux variables est tranché, on garde les 2 par curiosité). Je délègues à la méthode la sélection de variable.

Une première ACP avec la database complète sans tenir compte du résultat de l'arbre décision car on veut pouvoir comparer la selection de varaiable.

```
> names(database)
 [1] "distance" "direction" "altitude" "satellites" "delaiTransmission" "delaiServeur"
"delaiBalise"       "deltaDistance"
 [9] "volt"   "jourBalise" "jourServeur"       "rangDateGPS"       "idBalise"
"messageId"         "vitesse"           "power"
[17] "lng"               "lat"               "satisfaction"      "target"
```

L'ébouli des valeurs propres issue de l'ACP est représenté dans ce graphique :

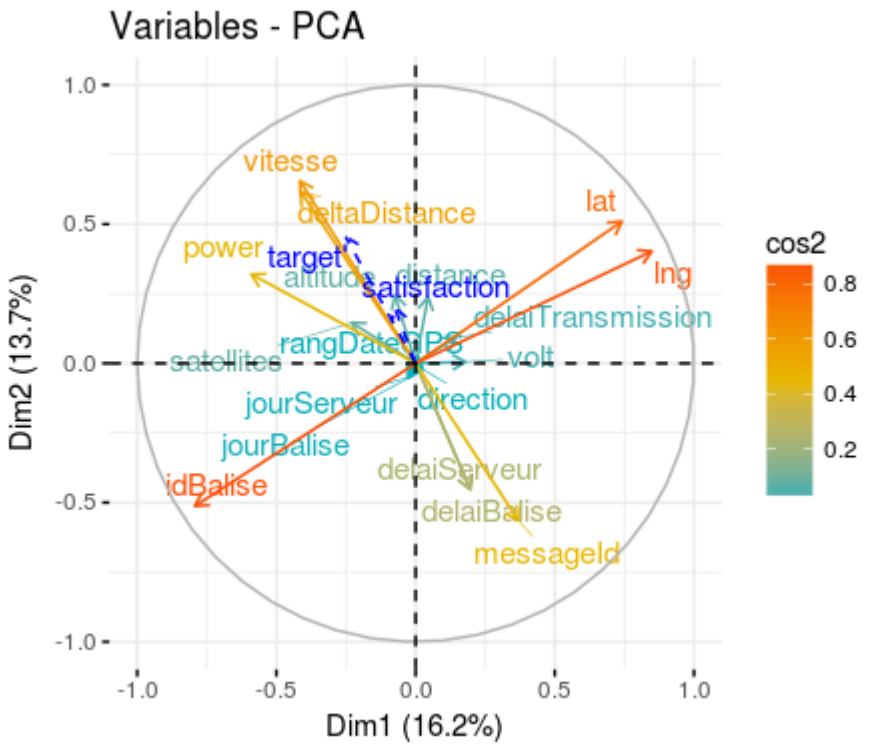


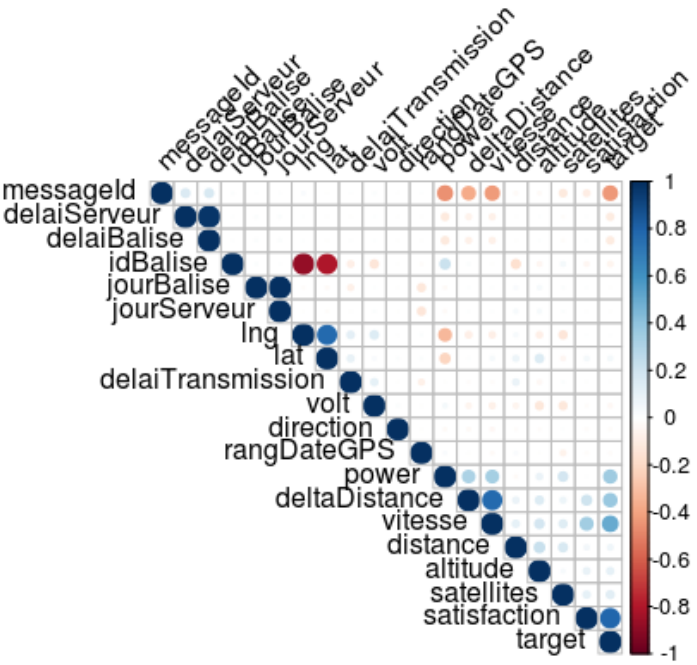
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.916464629	16.202581274	16.20258
Dim.2	2.472375356	13.735418644	29.93800
Dim.3	**2.048317364**	**11.379540909**	**41.31754**
Dim.4	1.866725705	10.370698359	51.68824
Dim.5	1.296115186	7.200639924	58.88888
Dim.6	**1.128410117**	**6.268945097**	**65.15782**
Dim.7	1.007859545	5.599219697	70.75704
Dim.8	0.988989455	5.494385863	76.25143
Dim.9	**0.964354231**	**5.357523505**	**81.60895**
Dim.10	0.844783363	4.693240906	86.30219
Dim.11	0.785074666	4.361525921	90.66372
Dim.12	0.681733277	3.787407095	94.45113
Dim.13	0.458554027	2.547522372	96.99865
Dim.14	0.218358226	1.213101253	98.21175
Dim.15	0.200628201	1.114601114	99.32635
Dim.16	0.096461270	0.535895943	99.86225
Dim.17	0.023110551	0.128391951	99.99064
Dim.18	0.001684831	0.009360171	100.00000

En considérant les 3 premiers axes on obtient 41% d'information. Ce qui ne permet pas de décrire une grande partie de l'échantillon. Avec 9 dimensions qui est la moitié moins de variable qu'au départ nous obtenons 81%.
On peut regarder la contribution des variables à chaque dimension pour choisir les variables les mieux corrélées et comparer avec les variables sélectionnées par un arbre de décision volontairement très court.

2.2.1.2 Analyse de correlation

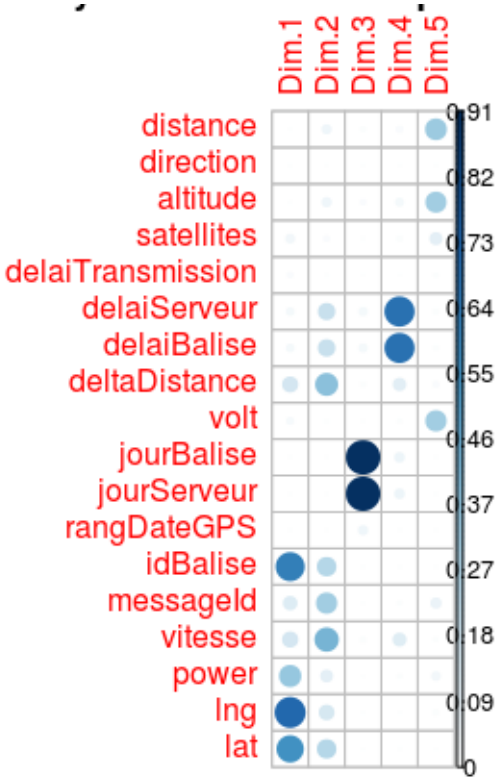
Le cercle de corrélation suivant met en évidence les relations intuitivement détectées.





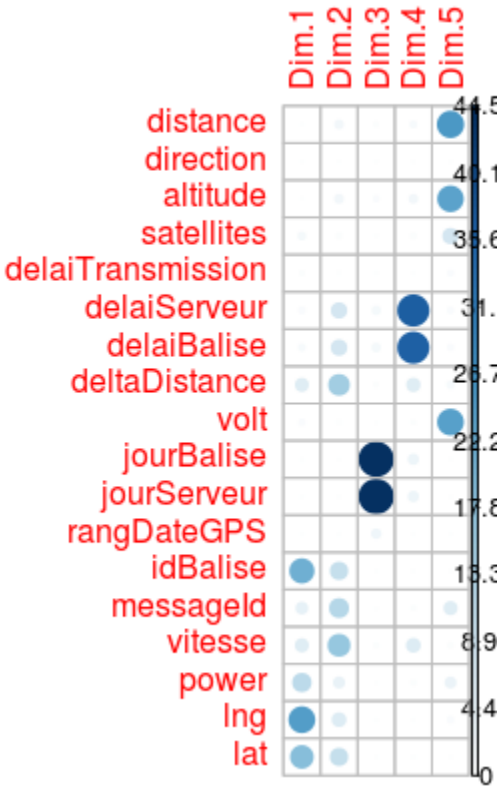
Les variables les mieux représentées sont les coordonnées GPS et l'identifiant de balise. Ce qui n'est pas étonnant étant donné que chaque balise poursuit régulièrement les mêmes parcours GPS.
Le deuxième groupe de variables concerne la dynamique du véhicule vitesse/distance (mécaniquement corrélées entre elles) suivi du troisième groupe formé avec la puissance et le message qui traduit un état du véhicule (roule-stop).
La matrice de corrélation ci contre donne une idée de l'intensité de la relation entre les variables de ces 3 groupes. Elle est trés forte et linéaire entre les coordonnées GPS et la balise, forte entre la vitesse et le deltaDistance. On détecte également un lien linéaire assez faible entre le message et le triplet vitesse-deltaDistance-power.

2.2.1.3 Description des dimensions

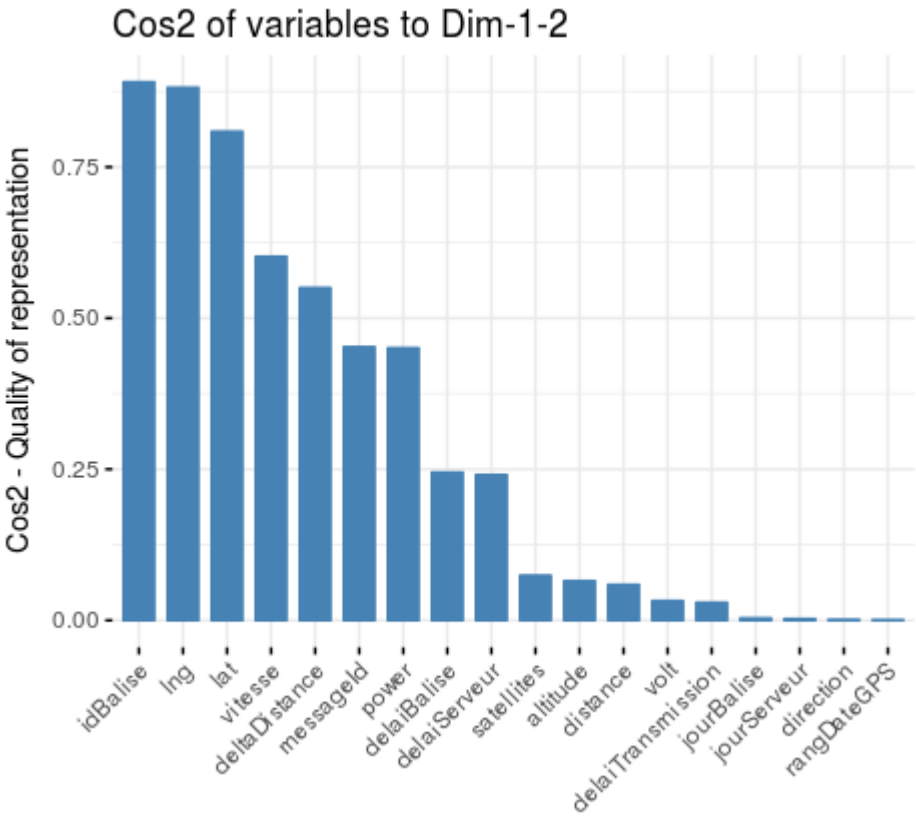


Le diagramme à gauche montre les contribution des variables aux dimensions

et le diagramme à droite montre la qualité de la représentation des variables dans les dimensions.

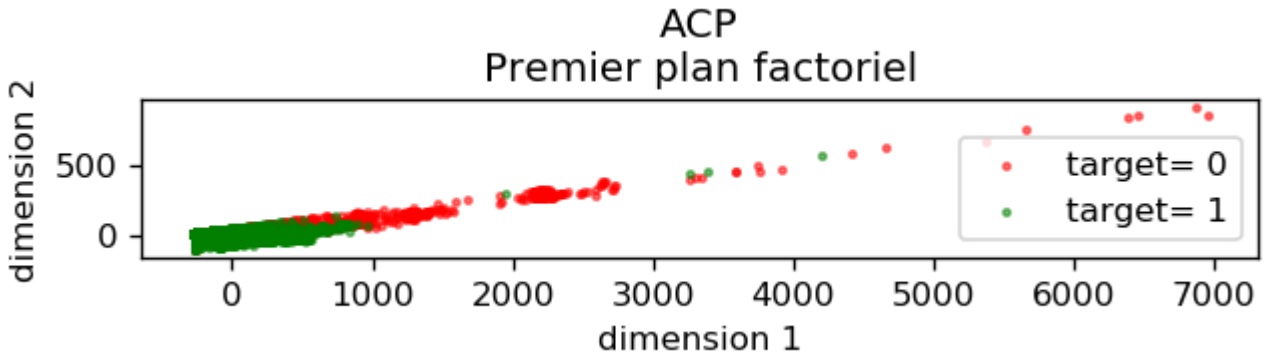


La dimension 3 est clairement la dimension temporelle qui n'a aucun lien avec les autres variables.
La dimension 4 est celle du délai (Balise ou serveur c'est la même information).
La dimension 5 concerne des variables qui ont peu d'influence sur la variable à expliquer.
Le premier plan factoriel regroupe plus de variables liées entre elles : le parcours, le délai et l'état du véhicule.
La qualité de la représentation n'est pas suffisante pour pourvoir généraliser sur cette seule base. Or, on constate malgès tout que l'intuition de départ est confortée par cette analyse.



Les variables deltaDistance, vitesse sont très corrélées et bien représentées. L'arbre de décision a mis en évidence les variables power et vistesse qui sont également corrélées. Mais la variable power est mal représentée sur le premier plan factoriel. L'ACP est refaite de manière plus académique avec recodage des variables discrètes en variable catégorielle. Le nuage de points est coloré par

rapport à la valeur cible et est visible ici :

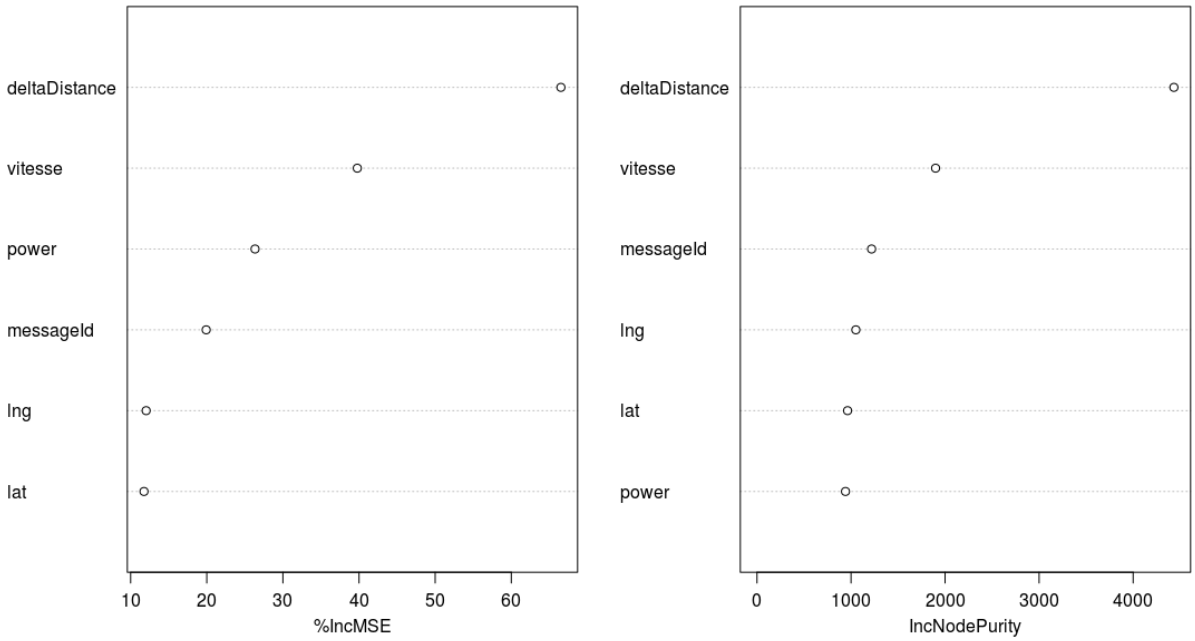


La première dimension est l'axe qui traduit le plus d'information du nuage de points.

Voyons ce que nous donne une forêt aléatoire (50 arbres en testant 3 variables puis 4).

- avec 3 variables : on a moins de variance expliquée (65.43% contre 66%)

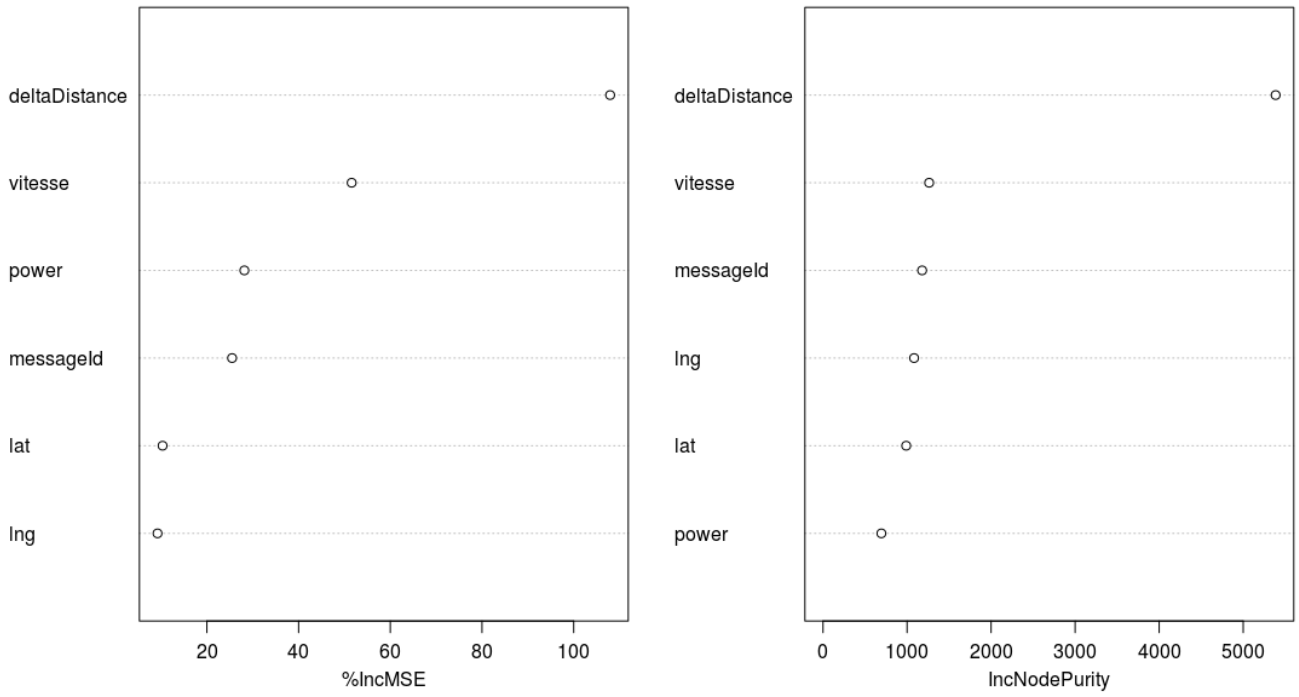
```
Call:
 randomForest(formula = target ~ ., data = da, ntree = 50, mtry = 3,      importance = TRUE,
replace = FALSE, na.action = na.roughfix)
                Type of random forest: regression
                      Number of trees: 50
No. of variables tried at each split: 3

          Mean of squared residuals: 0.0554156
                    % Var explained: 65.43
```



- Avec 4 variables : la forêt dégrade un petit peu la variance expliquée (65.22%)

```
Call:
```

```
 randomForest(formula = target ~ ., data = da, ntree = 50, mtry = 4,      importance = TRUE,
replace = FALSE, na.action = na.roughfix)
                Type of random forest: regression
                      Number of trees: 50
No. of variables tried at each split: 4

           Mean of squared residuals: 0.05575852
                     % Var explained: 65.22
```



En conclusion, avec la forêt aléatoire on retient deltaDistance qu'on avait bien identifié grâce à l'ACP mais qui n'était pas ressorti avec autant d'importance dans le trés court arbre de décision (il apparait au niveau 4 dans les arbres plus grands). Si on tolère le 1% de baisse de variance expliquée entre l'ACP et la forêt alétoire alors on accepte de prendre cette variable qui a tout de même une importante influence sur la variable cible. On perd un petit peu en explication et on y gagne certainement en prédiction.
On peut conclure que ce qui explique le délai n'est pas la zone géographique mais le fait que le véhicule roule ou pas.

L'analyse exploratoire montre que la variable la plus influente est DeltaDistance. On ne peut pas batir de modele sur cette seul variable. Pour la regression logistique en particulier (c'est mieux en général) il vaut mieux eviter les variables linéairement corrélées. Donc on garde lng mais pas lat (a longitude est préférable à la latitude car le test V de cramer la place en meilleure position. On ramène donc la matrice initiale à 4 variables actives). On garde power et messageId qui apparaissent assez haut dans les arbres. Donc on utilisera les variables actives suivantes pour établir le modèle :

DeltaDistance (distance géographique entre 2 relevés),
lng (longitude), power (puissance) et messageId

I

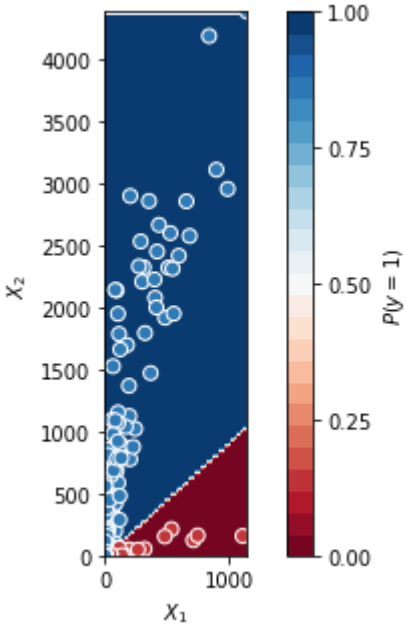
2.3 Tableau individus x variables

Le tableau variables individus est déterminé grâce aux analyse précédentes :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 640 entries, 0 to 639
Data columns (total 7 columns):
individu        640 non-null object
labelMessage    640 non-null object
niveauPower     640 non-null object
etat            640 non-null object
Target0         640 non-null int64
Target1         640 non-null int64
target          640 non-null int64
dtypes: int64(3), object(4)
memory usage: 35.1+ KB
```

Le nouveau tableau est issu d'un regroupement des observations suivant la clé suivante : individu – labelMessage – niveauPower – etat – valeur de target
- individu : regroupe les variables lng + lat + idBalise correspondand à la première composante principale des ACP.
- labelMessage : utilisation du labelMessage recodé pour rééquilibrer le nombre d'observation.
- niveauPower : recodage en classe des valeur des observations sur la base du seuil défini à l'aide des précédents arbre de décisions.
- etat : défini sur la base des valeurs deltaDistance et vitesse (si les 2 sont égale à 0 alors le véhicule est à l'arrêt la valeur ‚stop' est attribué sinon c'est la valeur 'roule')

A l'aide d'une ACM nous pourions essayer de mieux voir les individus et refaire une analyse pour confirmer ou infirmer ce qui a été établi jusqu'ici mais je pense que nous ne découvrirons rien de plus.



En fait j'ai juste fait manuellement une méthode de clustering trés spécifque à cet échantillon. Ce qui ne présente au final pas un grand intérêt.
Et pour se faire plaisir, on a le résultat de la classifiction manuelle en couleur ci-contre.
On voit les 160 milles observations résumées en 640 situations différentes décrites par les variables les plus représentatives. En bleu on voit les rapports de fréquence* de délai long proches de 1 et en rouge ceux qui sont proches de 0. X1 quantifie le nombre de fois où le délai est insatisfaisant et X2 le nombre de fois où le délai est satisfaisant. On voit qu'au delà de mille observations positives pour une situation donnée (balise-zone-message-power-etat) on ne trouve plus de situation négative.

Nous pouvons passer à la recherche du meilleur modèle prédictif désormais équipé des meilleurs prédicteurs.

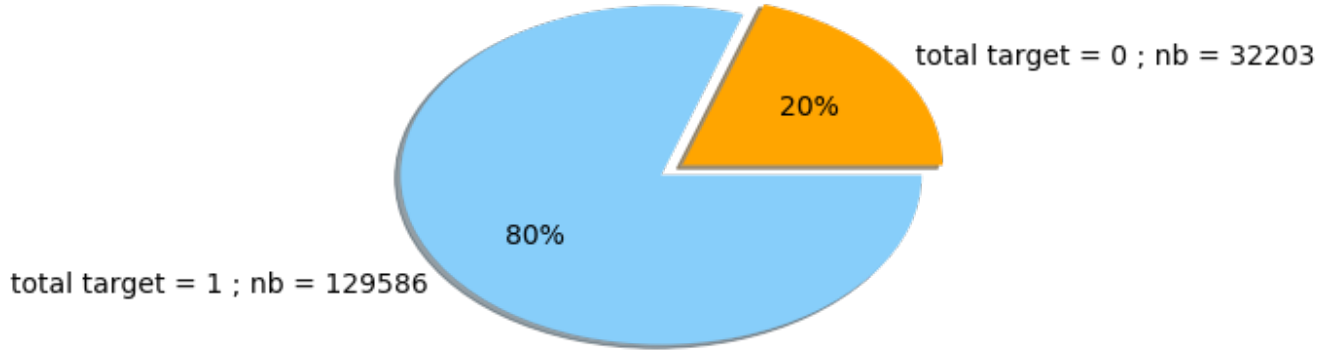
* Fréquence calculée en faisant le ratio nombre de délai long divisé par le nombre de délais compté sur la zone pour une balise donnée (recoder dans la variable individu)

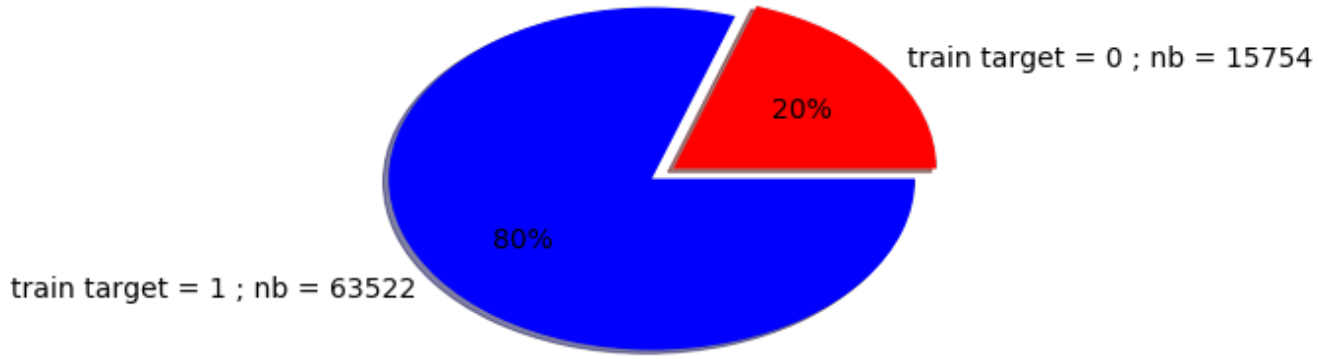
3 Modélisation

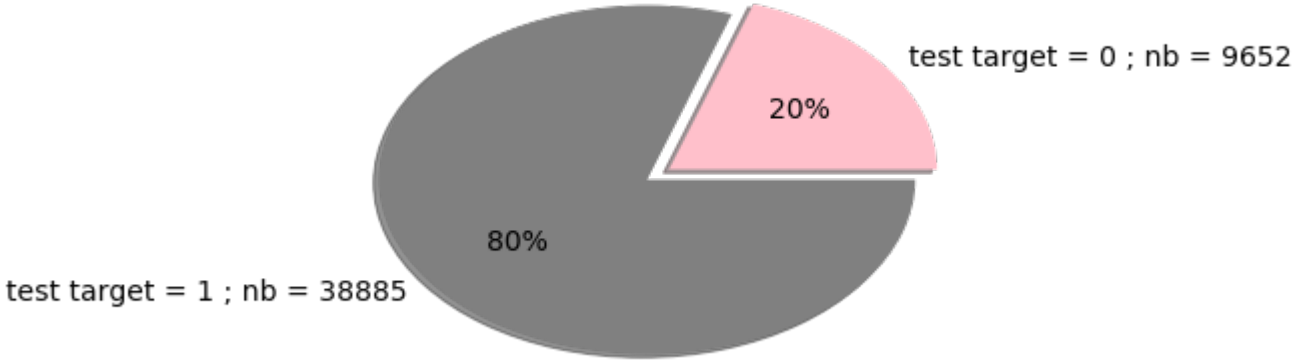
Notebook (D. Modélisation)

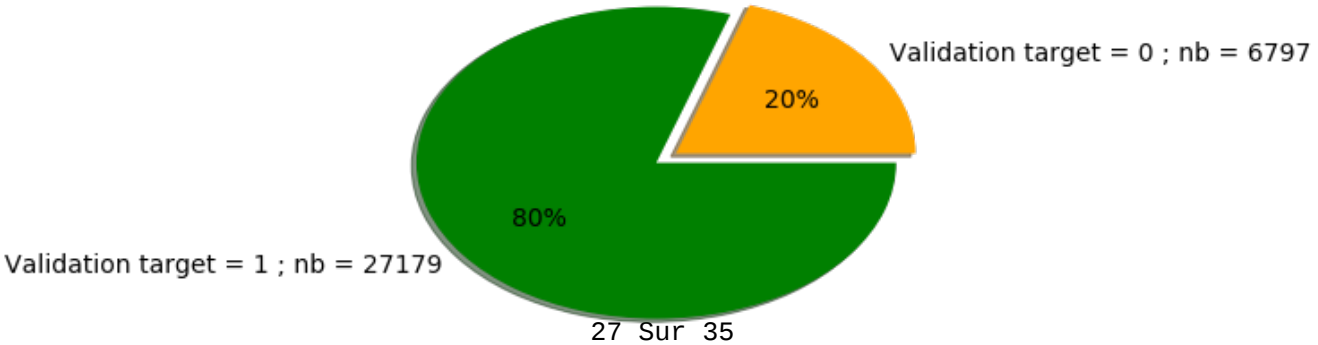
3.1 Echantillonage pour l'apprentissage

Les échantillons sont différents pour l'apprentissage, le test et la validation. La proportion de valeur cible est identique à l'échantillon total.

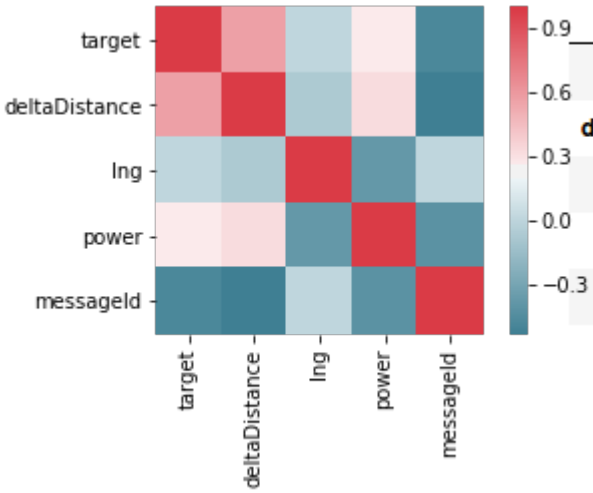






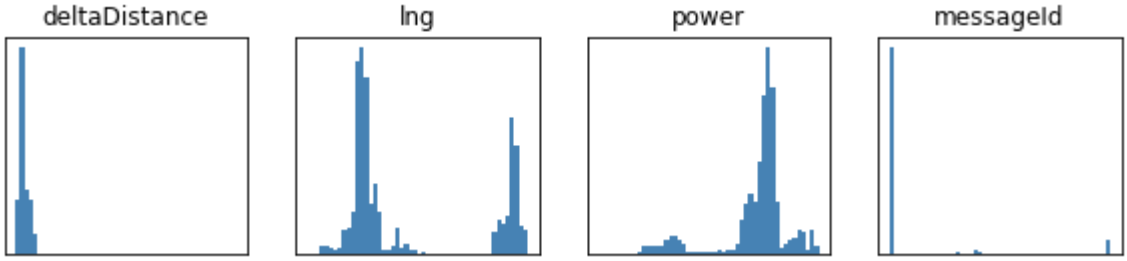


3.2 Description des prédicteurs et de la variable à prédire

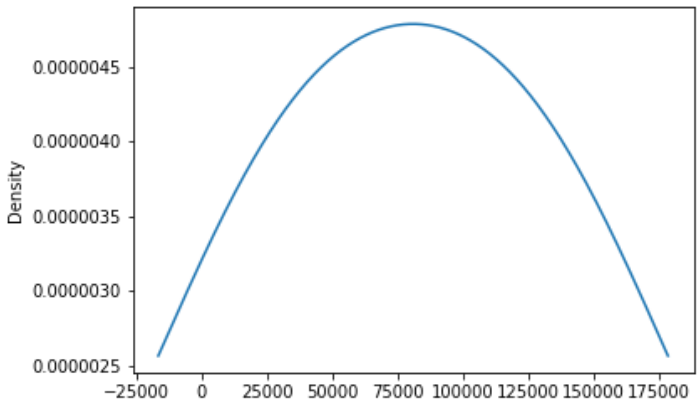


	target	deltaDistance	lng	power	messageId
target	1.000000	0.570628	0.013242	0.268091	-0.471623
deltaDistance	0.570628	1.000000	-0.070117	0.327717	-0.530888
lng	0.013242	-0.070117	1.000000	-0.374819	0.008253
power	0.268091	0.327717	-0.374819	1.000000	-0.410192
messageId	-0.471623	-0.530888	0.008253	-0.410192	1.000000

Les variables explicatives sont linéairement indépendantes (avec les méthodes Pearson et Sperman) entre elles et ne sont ainsi distribuées :

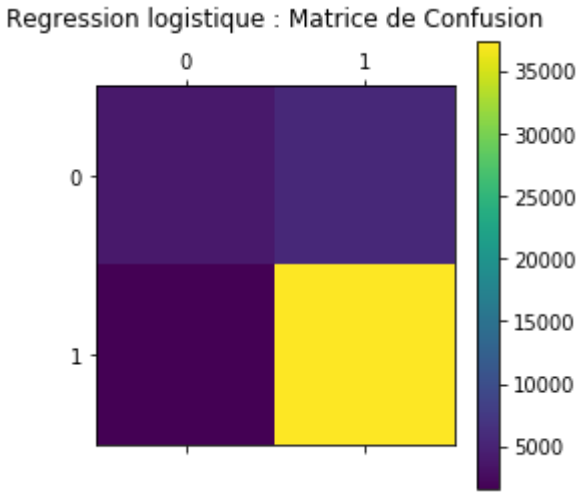


Les prédicteurs qui ont une dépendance forte avec la variable cible sont : deltaDistance et messageId. La variable cible suit une loi gausssienne un peu plus applatie et la valeur 1 représente la situation souhaitable.

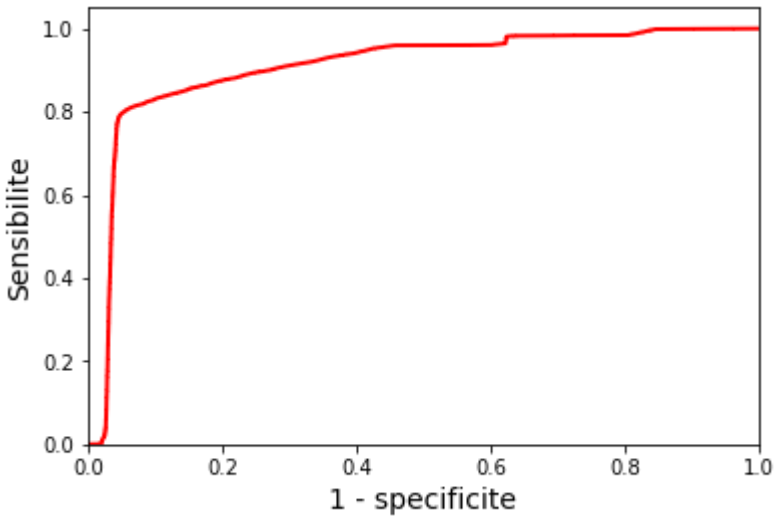


3.3 Regression Logistique

```
Meilleur(s) hyperparamètre(s) sur le jeu d'entraînement: {'C': 0.01}
_Regression logistique_ Résultats de la validation croisée :
        accuracy = 0.855 (+/-0.003) for {'C': 0.01}
        accuracy = 0.854 (+/-0.003) for {'C': 0.05}
        accuracy = 0.854 (+/-0.003) for {'C': 0.1}

Précision du classifieur Regression Logistique
sur le jeu de test : 0.852
```

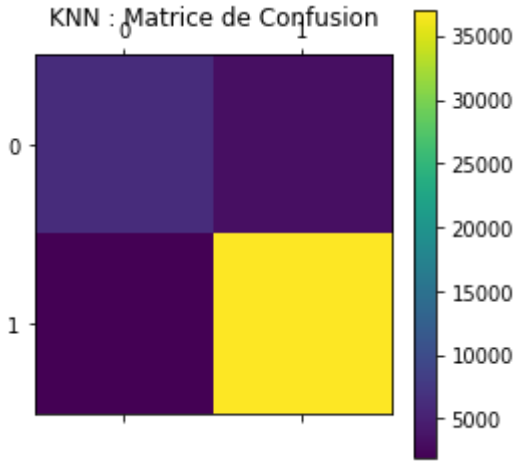


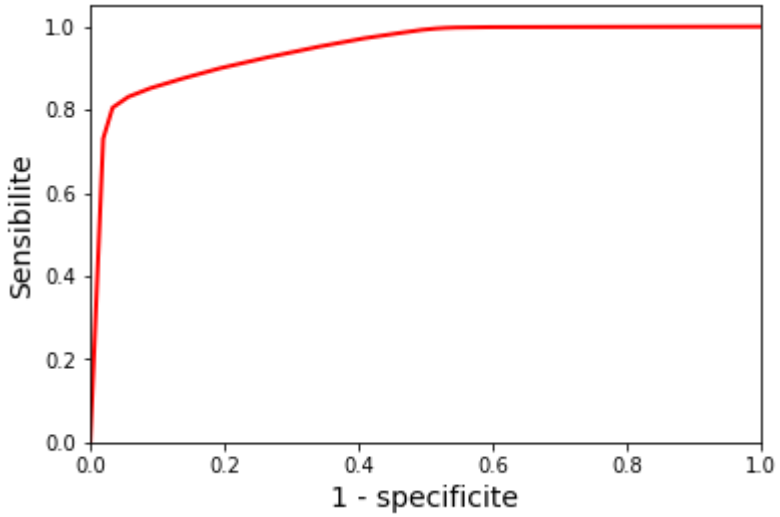

```
col_0      0      1
target
0       3993   5659
1       1550  37335

Aire sous la courbe ROC (AUC) : 0.909
Sensibilité : 0.95
1-Spécificité : 0.58
Seuil : 0.76

 Estimation de l'erreur de prévision : 0.15
```



3.4 KNN

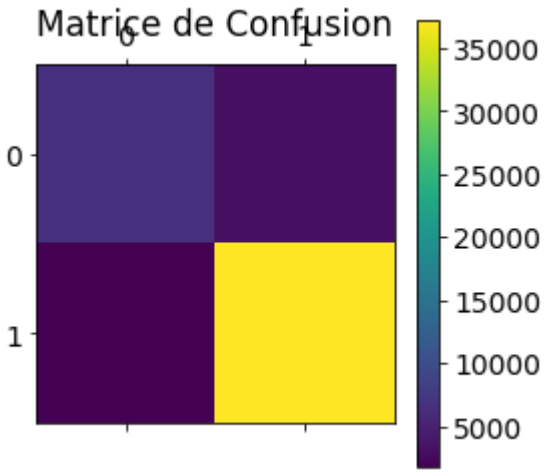
```
Meilleur(s) hyperparamètre(s) sur le jeu d'entraînement: {'n_neighbors': 15}
_KNN_ Résultats de la validation croisée :
        accuracy = 0.889 (+/-0.003) for {'n_neighbors': 3}
        accuracy = 0.892 (+/-0.003) for {'n_neighbors': 5}
        accuracy = 0.894 (+/-0.004) for {'n_neighbors': 9}
        accuracy = 0.893 (+/-0.003) for {'n_neighbors': 13}
        accuracy = 0.894 (+/-0.004) for {'n_neighbors': 15}

Précision du classifieur KNN sur le jeu de test : 0.894

col_0      0      1
target
0       6359   3293
1       1876  37009


Aire sous la courbe ROC (AUC) : 0.947
Sensibilité : 0.95
1-Spécificité : 0.66
Seuil : 0.53

 Estimation de l'erreur de
prévision : 0.11
```



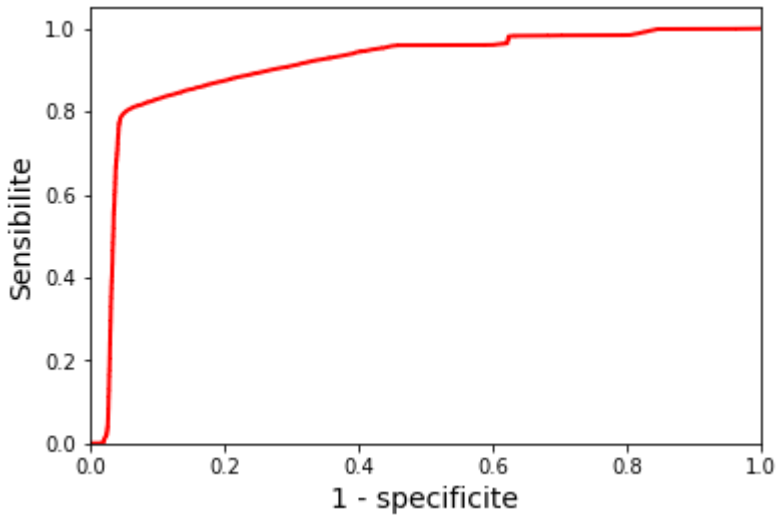


3.5 Arbre de décision

```
Meilleur profondeur de l'arbre de décision :
{'max_depth': 8}
Score du classifieur ARBRE sur le jeu de test : 0.899
_Abre de décision _ Résultats de la validation croisée :
        accuracy = 0.898 (+/-0.004) for {'max_depth': 5}
        accuracy = 0.898 (+/-0.004) for {'max_depth': 6}
        accuracy = 0.899 (+/-0.003) for {'max_depth': 7}
        accuracy = 0.899 (+/-0.004) for {'max_depth': 8}
        accuracy = 0.899 (+/-0.004) for {'max_depth': 9}
        accuracy = 0.883 (+/-0.004) for {'max_features': 1}
        accuracy = 0.885 (+/-0.006) for {'max_features': 2}
        accuracy = 0.886 (+/-0.005) for {'max_features': 3}
        accuracy = 0.886 (+/-0.008) for {'max_features': 4}

Précision du classifieur Arbre de décision sur le jeu de test : 0.866

col_0      0      1
target
0       6465   3187
1       1693  37192
```




```
Aire sous la courbe ROC (AUC) : 0.909
Sensibilité : 0.95
1-Spécificité : 0.58
Seuil : 0.76

 Estimation de l'erreur de prévision : 0.10
```

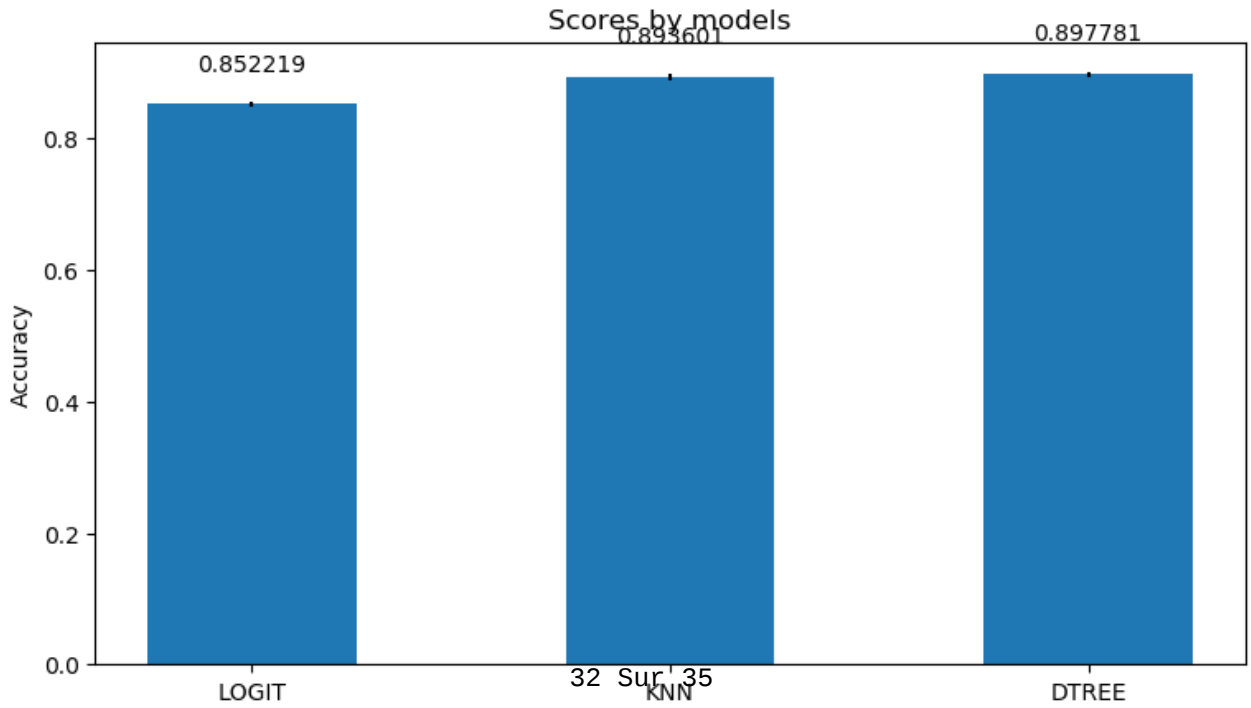


4 validation des modèles

L'arbre de décision l'emporte sur la phase d'apprentissage et test. Il a une aire sous la courbe ROC la plus proche de 1 (comme la regression logistique) il est tout aussi sensible que les autres et est le moins spécifique. Voyons si le classement se maintient sur la comparaison des modeles avec l'échantillon de validation :

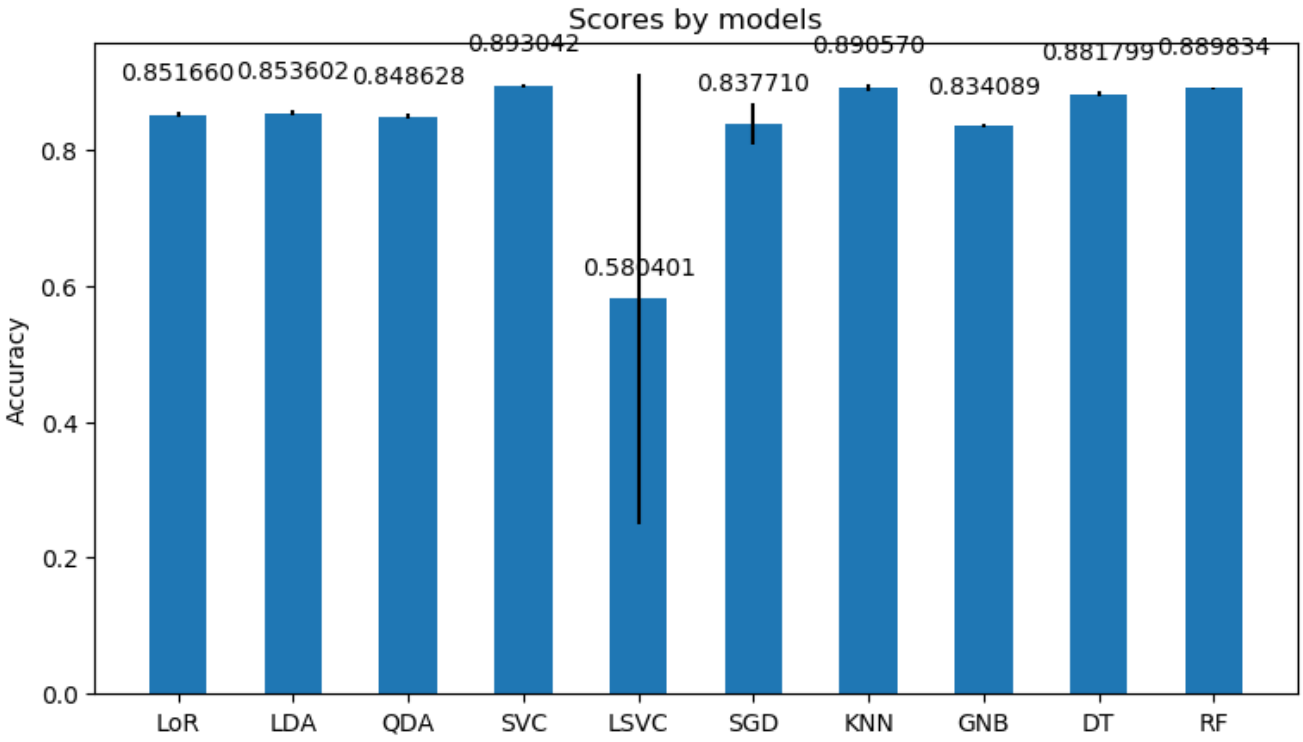
```
LOGIT (regression logistique)
Score sur echantillon de validation : 0.851
Aire sous la courbe ROC (AUC) : 0.915
Sensibilité : 0.95
1-Spécificité : 0.57
Seuil : 0.76
Mean: 0.852189644517
Standard Deviation: 0.00435883199146

KNN
Score sur echantillon de validation : 0.895
Aire sous la courbe ROC (AUC) : 0.949
Sensibilité : 0.95
1-Spécificité : 0.67
Seuil : 0.53
Mean: 0.893601412942
Standard Deviation: 0.00388344399216

DTREE (arbre de décision)
Score sur echantillon de validation : 0.902
Aire sous la courbe ROC (AUC) : 0.954
Sensibilité : 0.96
1-Spécificité : 0.68
Seuil : 0.51 Mean: 0.897780757499    Standard Deviation: 0.00398306132141
```

L'arbre obtient un meilleur score mais est le plus spécifique et sensible. Les moyennes obtenues et l'écart type sont représentés dans le diagramme comparatif ci-dessous :



Une évaluation superficielle d'autres méthodes ouvre des pistes de recherche pour optimiser la prédiction sans être trop spécifique. Sans trop entrer dans les détails les pistes qui se révèlent intéressantes sont la forêt aléatoire (pour contrer le caractère trop spécifique de l'arbre de décision) et la méthode SVC.

```
LoR
Results: [ 0.85697469  0.84930096  0.84812362  0.85562914  0.84827079]
Mean: 0.851659838861 Standard Deviation: 0.00383556605465
LDA
Results: [ 0.85579753  0.85253863  0.85033113  0.85901398  0.85033113]
Mean: 0.853602478366 Standard Deviation: 0.00336564607114
QDA
Results: [ 0.84932313  0.84841795  0.84282561  0.85459897  0.84797645]
Mean: 0.84862842316 Standard Deviation: 0.00374892206819
SVC
Results: [ 0.89287816  0.89080206  0.88977189  0.89595291  0.89580574]
Mean: 0.893042152225 Standard Deviation: 0.00252386280174
LSVC
Results: [ 0.86080047  0.15202355  0.81883738  0.19852833  0.87181751]
Mean: 0.58040144811 Standard Deviation: 0.331582066337
SGD
Results: [ 0.81695115  0.866078    0.86269316  0.78675497  0.85607064]
Mean: 0.837709582012 Standard Deviation: 0.0309624285694
KNN
Results: [ 0.89052384  0.88609272  0.88447388  0.89713024  0.8946284 ]
Mean: 0.89056981534 Standard Deviation: 0.00483152848938
GNB
Results: [ 0.83343143  0.83414275  0.83119941  0.83870493  0.83296542]
Mean: 0.83408878789 Standard Deviation: 0.00250416700402
DT
Results: [ 0.87992937  0.8785872   0.87932303  0.88874172  0.88241354]
Mean: 0.88179897191 Standard Deviation: 0.003701909295
RF
Results: [ 0.88758093  0.88933039  0.88682855  0.8928624   0.89256806]
Mean: 0.889834066786 Standard Deviation: 0.00249034071577
```



5 Conclusions

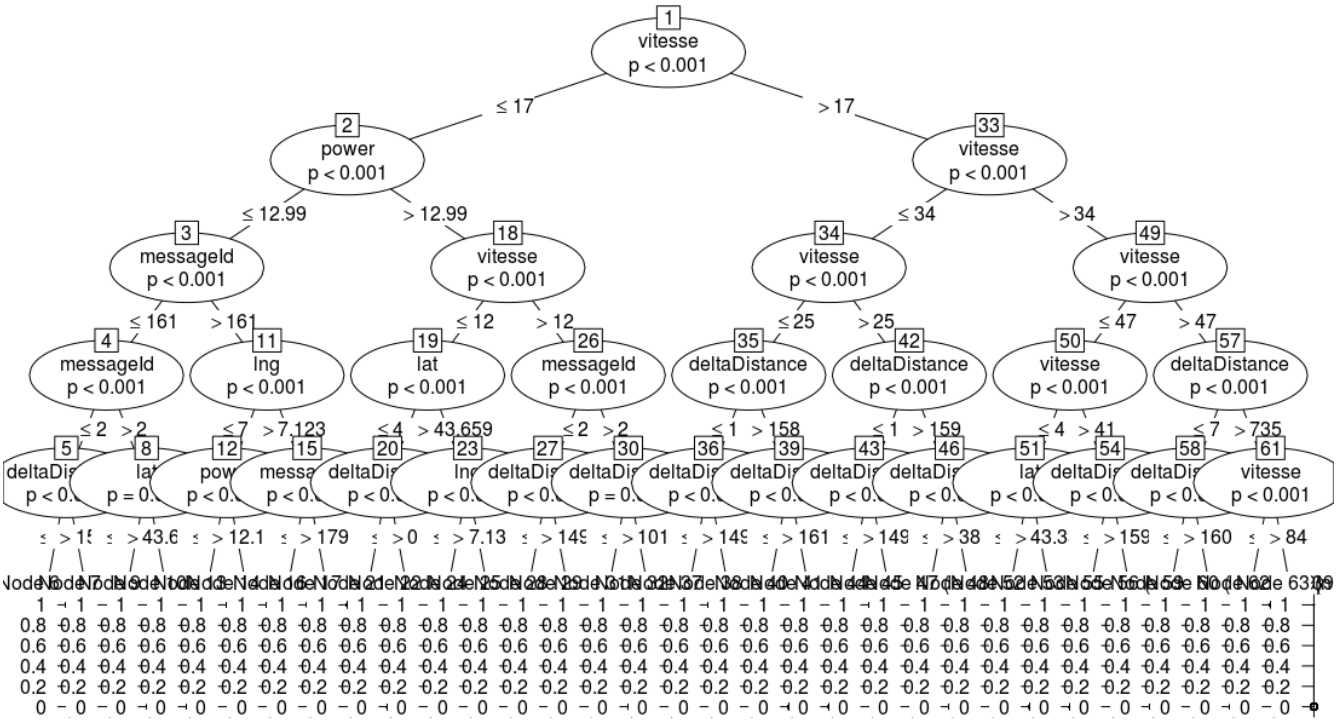
L‘analyse bivariée a laissé la place à l‘intuition confirmée par l‘ACP. Le recours à l‘ACP pour la sélection de variables est plus efficace que l‘analyse bivariée croisée avec le test V de cramer. Le recours aux arbres de décision et à la forêt aléatoire à mis fin au débat en quelques secondes. En tant qu‘apprentis analyste de données cette démarche m‘a paru rassurante et à caractère pédagogique également pour une restitution à la société qui va exploiter ces résultats seulement si elle est convaincue par les résultats.

Le résultat de l‘étude montre qu‘il n‘y a pas de lien direct entre la position GPS et le rallongement du délai d‘émission des balises. La cause commune aux deux faits : les délais longs sont localisées et certaines balises mettent plus de temps à émmettre en certains points que d‘autres, est que lorsque qu‘un véhicule est à l‘arrêt, la puissance diminue et la balise cesse d‘emettre (elle s‘endort). On infirme l‘hypothèse métier selon laquelle le rallongement du délai est lié à une mauvaise couverture GSM.

Pour étudier plus avant cette hypothèse métier, il faut ajouter des informations complémentaires comme les points GPS des arrêts fréquents pour considérer le rallongement en temps comme normal et étudier à nouveaux les corrélations.

La recherche de modèle prédictif a été frutuseuse mais d‘autres méthodes seraient sans doute meilleure. Il faudrait poursuivre l‘étude notamment avec l‘algorithme SVC qui a donné le meilleur score avec un apprentissage et test sur l‘échantillon de validation (sans chercher à optimiser les paramètres).
En terme d‘application, dans le contexte métier il faudrait décider s‘il est important de bien prédire les 0 quitte à en prédire trop pour faire le choix du meilleur modèle.

Annexes 1 : arbre de décision de profondeur 5 pour la sélection de variable



Annexe 2 : Résultat de la recherche de la meilleur profondeur de l‘arbre de décision

J‘avais envie de savoir si l‘arbre de la sélection de variable était si court que ça. Effectué à l‘aide d‘un programme python – j‘ai recherché la meilleure profondeur pour l‘arbre de décision. Pour la sélection de variable, on s‘est contenté d‘un arbre court mais la curiosité l‘a remporté pour estimer au mieux ce paramètre.

```
 Estimation de l'erreur de prévision : 0.14
Meilleur profondeur :
{'max_depth': 9}
GridSearchCV(cv=5, error_score='raise',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best'),
       fit_params={}, iid=True, n_jobs=-1,
       param_grid=[{'max_depth': [5, 6, 7, 8, 9]}, {'max_features': [1, 2, 3, 4]}],
       pre_dispatch='2*n_jobs', refit=True, scoring=None, verbose=0)
```