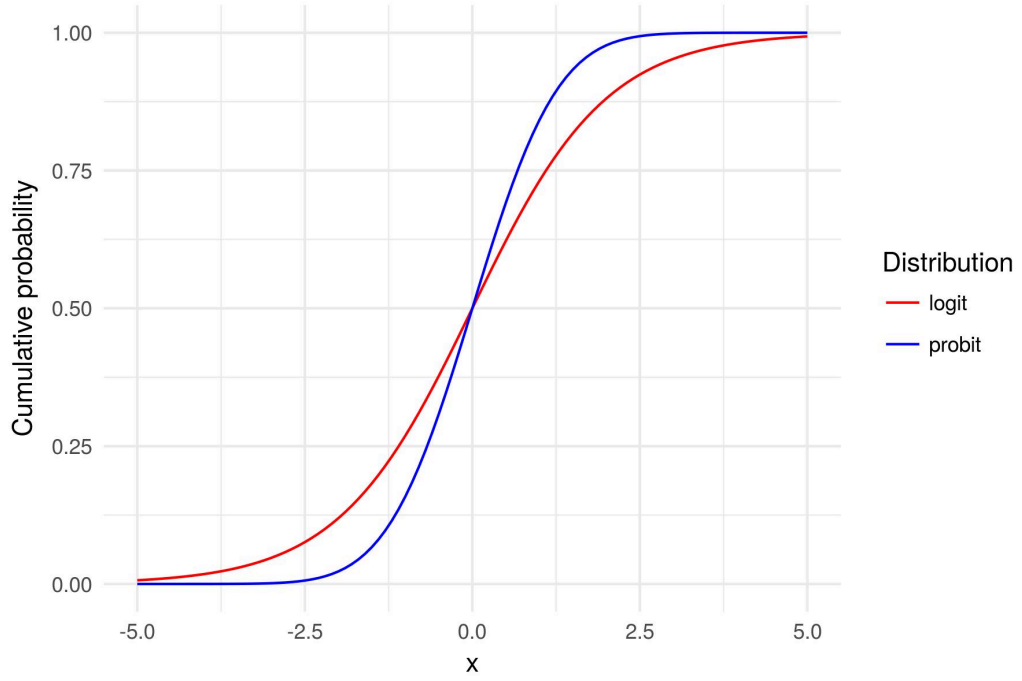
Microéconométrie

RAPPORT ÉCRIT

29/03/2024

Master 1 Ingénierie économique et financière



POUPONNEAU Guillaume
AGRAFFEL Nicolas

Sous la direction de M^me^ LARRIBEAU Sophie

Sommaire

29/03/2024 1
Introduction - Description des données 3
Variables: 4
PHASE DE NETTOYAGE 5
ANALYSE EXPLORATOIRE ET TRAITEMENT DES DONNÉES 5
2.1 Analyse univariée 6
2.2 Analyse bivariée 8
Modélisation 14
1.1 - Subdivision 15
1.2 - Modèle logit 16
1.2.2 : Logit plutôt que Probit 21
1.3 - Le déséquilibre de classe 22
1.4 - Interprétation de l'importance des variables : 26
2 - Modèle de prévision : 26

Introduction - Description des données

Notre mission s'est concentrée sur l'optimisation des survols du réseau pour le diagnostic de Enedis Bretagne. Ce projet ambitieux a commencé par une étape fondamentale et cruciale : la mise en perspective et la compréhension approfondie de la base de données. Avant de plonger dans les complexités techniques du nettoyage et de la normalisation des données, nous avons accordé une attention particulière à la réflexion stratégique afin de cerner efficacement l'objectif de notre analyse et anticiper les défis potentiels. *In fine*, notre but est de permettre à Enedis de réaliser des économies sur la maintenance de son réseau électrique.

Cette démarche initiale débute par une analyse méticuleuse de la structure des données, de leurs caractéristiques intrinsèques et des relations sous-jacentes entre les variables. L'importance de cette phase réflexive ne peut et ne doit pas être sous-estimée. Elle a posé les bases d'une approche méthodique et éclairée pour la préparation des données, garantissant que chaque étape suivante soit adaptée aux particularités des données de Enedis.

La préparation des données implique des tâches de nettoyage et une redéfinition de notre variable cible qui se définit par la nécessité ou non d'une maintenance compte tenu des informations de 2023. Notre approche vise à préserver la pertinence et l'exactitude des informations pour les analyses subséquentes.

Nous avons divisé notre base de données, pour obtenir une base d'entraînement (qui contient 75% des données de la base initiale), et une base de test avec les 25% restants.

Ainsi, notre problématique est la suivante : Grâce aux données de 2023, trouver les principales variables pouvant causer un besoin d'intervention , puis créer un modèle capable de prédire sur quelles lignes il y aurait un besoin d'intervention.

Dans un premier temps, après avoir nettoyé et analysé les variables, nous verrons quelles sont les variables les plus susceptibles de causer un incident, grâce à un modèle Logit, puis nous créerons le modèle prédictif dans un second temps avec un random forest.

Nos Variables

ID: Identifiant du tronçon

Nb_of_incident: Nombre d'incident depuis la mise en service

Electrical_length: Longueur du tronçon

Service_date: Date de mise en service

Length_climate_hazard_plan: Taille du tronçon soumis aux aléas climatique

Length_fragile_section: Longueur de tronçon fragile

Nb_of_anomaly: Nombre d'anomalie depuis la mise en service

Year_helicopter_flight: Année du dernier passage en hélicoptère

Last_treatment_PR_imobilized: Année de la dernière immobilisation/intervention de tronçon

Toutes les variables explicatives ont été transmises par Enedis. Elles ont été jugées les plus importantes par leur équipe quant à la réalisation de ce projet.

Les données étant anonymisées pour des contraintes légales, nous ne pouvons pas faire de recherches pour en ajouter davantage. Nous allons donc nous limiter à la base de données fournie dans le cadre de l'étude.

1. Phase de nettoyage

Grâce à la description que nous avons faite de notre base de données, nous pouvons nous rendre compte qu'il y a beaucoup de valeurs manquantes sur nos variables (153095). Il faut donc s'adapter à la nature des variables. Il y a 4 choix qui s'offrent à nous:

1. Remplacer par la médiane
2. Remplacer par la moyenne
3. Supprimer ou remplacer par un 0
4. Ne pas en tenir compte si cela ne gène pas

-*Nombre d'incidents* (36872 NAs): Le plus pertinent est de remplacer par un 0 les NAs.

-*Nombre d'anomalie* (50314 NAs): Idem

-*Dernière maintenance effectuée* (65909 NAs): c'est la variable dont on va se servir pour la variable cible, on ne l'utilisera que pour créer celle-ci, il n'y a donc pas besoin de la modifier.

Nous avons pris l'initiative de changer la date de création de la ligne en durée d'existence, une date est difficilement interprétable par un modèle.

On crée une variable pour remplacer la dernière année de passage en hélicoptère, elle indiquera la durée en année du dernier passage avec celui-ci. Elle indiquait au préalable la date du dernier passage en hélicoptère.
Par exemple: Dernier passage en hélicoptère : 2017, nouvelle variable = 6 (2023-2017)

Variable cible y: Elle n'existe pas encore, on va donc la créer à partir de la variable "Dernière maintenance effectuée" ("Last_treatment_PR_imobilized"), i.e: 0 si il n'y a pas eu d'intervention en 2023, 1 si il y en a eu une, dans le but de créer notre base de teste.

2. Analyse exploratoire et traitement des données

Après la phase de nettoyage, nous avons procédé à une analyse univariée, examinant chaque variable individuellement pour comprendre sa distribution et ses caractéristiques. Cette étape nous a permis de détecter des anomalies, des tendances et des insights sur chaque variable. Cela nous a permis notamment de dégager **la nature déséquilibrée de la variable cible**. Par la suite, nous avons utilisé l'analyse bivariée pour explorer les relations entre les paires de variables. Cette approche a révélé des corrélations et des interactions intéressantes qui n'étaient pas évidentes lors de l'analyse univariée. Cette analyse a été complétée par un tableau de corrélation et une ACP pour déterminer si les variables impactent notre variable cible, cela nous permet d'avoir une idée des performances futures de nos modèles.

2.1 Analyse univariée

Nous avons créé des classes pour faciliter l'interprétation de la durée d'existence du tronçon: "(04) > 67 ans","(03) 56-67 ans","(02) 31-55 ans", "(01) <= 30 ans"

Voici les processus que nous avons utilisé pour visualiser les données:

skim : résumé rapide des données

```
── Variable type: numeric ──────────────────────────────────────────────────────
   skim_variable                n_missing complete_rate      mean        sd       p0      p25      p50      p75
 1 ID_t                                 0             1    46420.     26811.       0     23141.    46414    69607.
 2 Nb_of_incident                       0             1        0.832      1.08     0         0        1        1
 3 Electrical_length                    0             1      294.       278.       0.0100   96.6    217.     416
 4 Length_climate_hazard_plan           0             1        2.95       3.66     0         0.569    1.79     3.88
 5 Length_fragile_section               0             1        3.96       5.72     0         0.799    2.06     4.64
 6 Nb_of_anomaly                        0             1        2.42       4.55     0         0        0        4
 7 Year_helicopter_flight               0             1     2020.         1.43  2018      2020     2020     2022
 8 Last_treatment_PR_immobilized    65909             0.112 2019.         2.91  2013      2016     2018     2021
 9 y                                    0             1        0.0116     0.107    0         0        0        0
10 Duree_annees                         0             1       39.8       10.5      0.816    35.2     39.2     45.2
11 Duree_annees_helico                  0             1        2.55       1.43     1         1        3        3
12 cl_Duree                             0             1        1.98       0.494    1         2        2        2
    p100 hist
```

Nb_of_incident : Nombre d'incidents, varie de 0 à 6, avec une moyenne proche de 1.

y : variable cible binaire (0 ou 1), avec une très faible moyenne suggérant une distribution déséquilibrée. Cela indique un possible besoin de rééchantillonnage pour plus tard.

Visualisation des distributions:

1.Distribution de la Longueur Électrique (Electrical_length)

- La majorité des valeurs sont proches de zéro, indiquant que la longueur électrique est très petite pour un grand nombre d'observations. Il y a quelques observations avec des valeurs plus élevées, mais elles sont beaucoup moins fréquentes.

2.Distribution du nombre d'incident (Nb_of_incident)

- La plupart des incidents ont un nombre de 0 ou 1, ce qui suggère que les incidents sont rares ou peu fréquents dans cet ensemble de données.

3.Distribution de Length_climate_hazard_plan

- La distribution des plans liés au climat est également fortement inclinée vers les valeurs plus faibles, ce qui indique que pour la plupart des observations, la longueur des plans de risque climatique est faible.

4.Distribution de Nb_of_anomaly (Nombre d'anomalies)

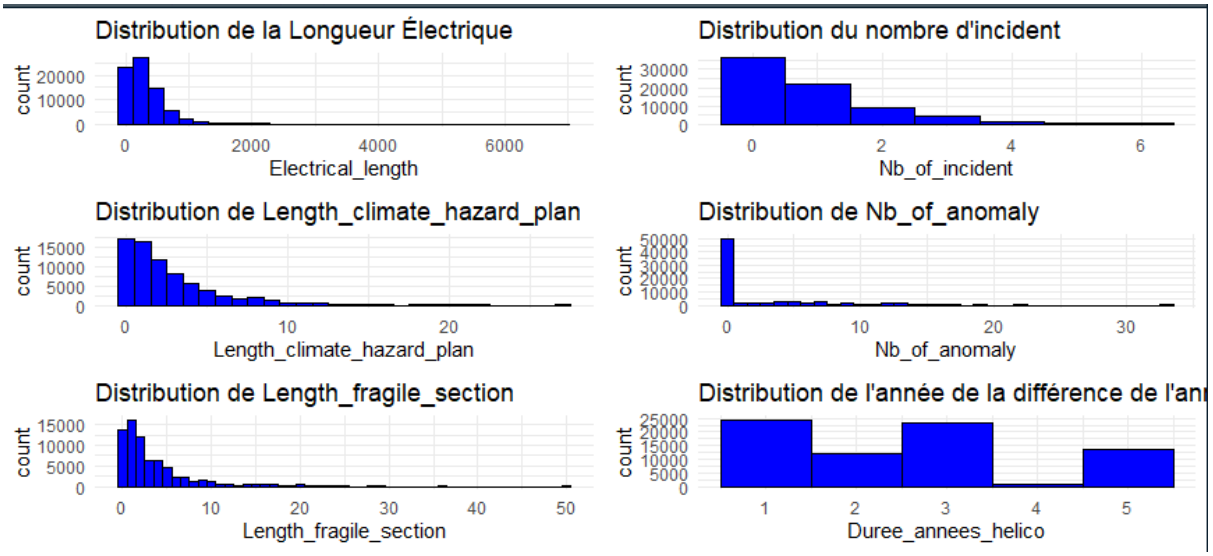
- Cette variable montre que la plupart des observations ont peu ou pas d'anomalies, avec une forte concentration à zéro.

5.Distribution de Length_fragile_section (Longueur des sections fragiles)

- Encore une fois, la distribution est concentrée vers les valeurs basses, indiquant que les sections fragiles sont généralement courtes.

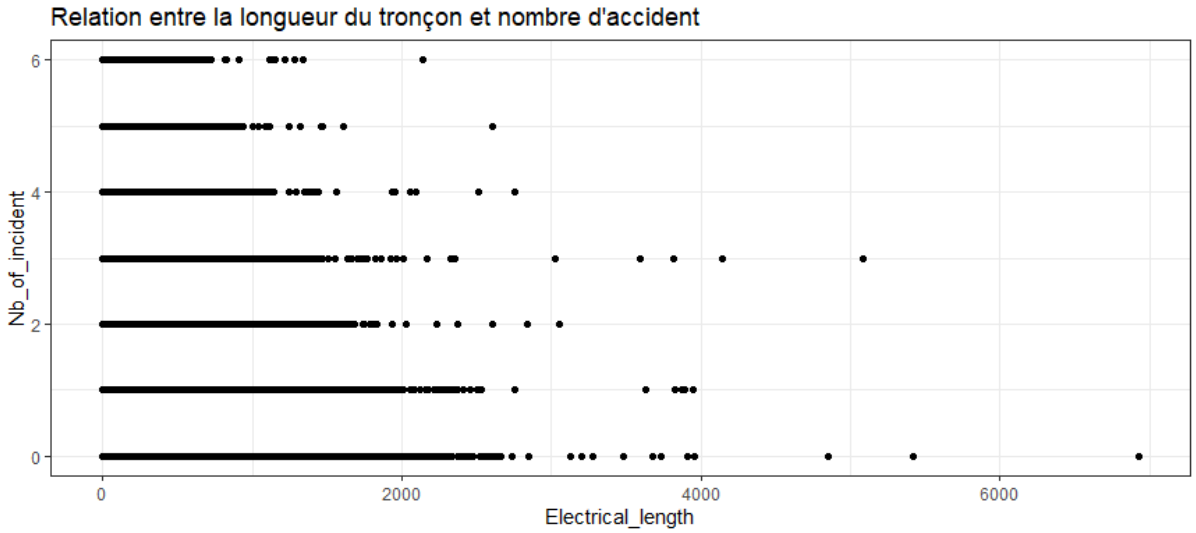
6.Distribution de l'année de la différence de l'année de passage de l'hélicoptère (Duree_annees_helico)

- Les années de passage de l'hélicoptère semblent être réparties plus uniformément par rapport aux autres variables, mais avec des pics, suggérant que certaines durées sont plus communes que d'autres.



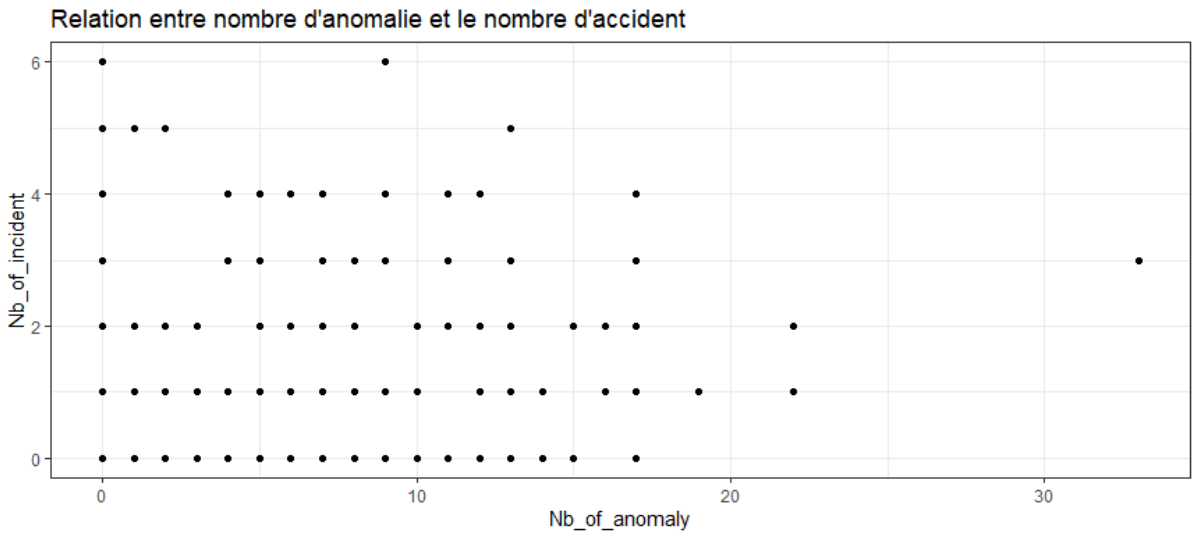
En résumé, les graphiques indiquent que les valeurs de ces variables dans l'ensemble de données sont majoritairement concentrées sur les valeurs basses avec quelques valeurs plus élevées qui sont beaucoup moins fréquentes. Cela pourrait suggérer que la plupart des observations dans cet ensemble de données représentent des situations "normales" ou "non critiques", avec quelques rares exceptions qui peuvent nécessiter une attention particulière, une analyse supplémentaire ou un choix de modèle capable d'en tenir compte.

2.2 Analyse bivariée

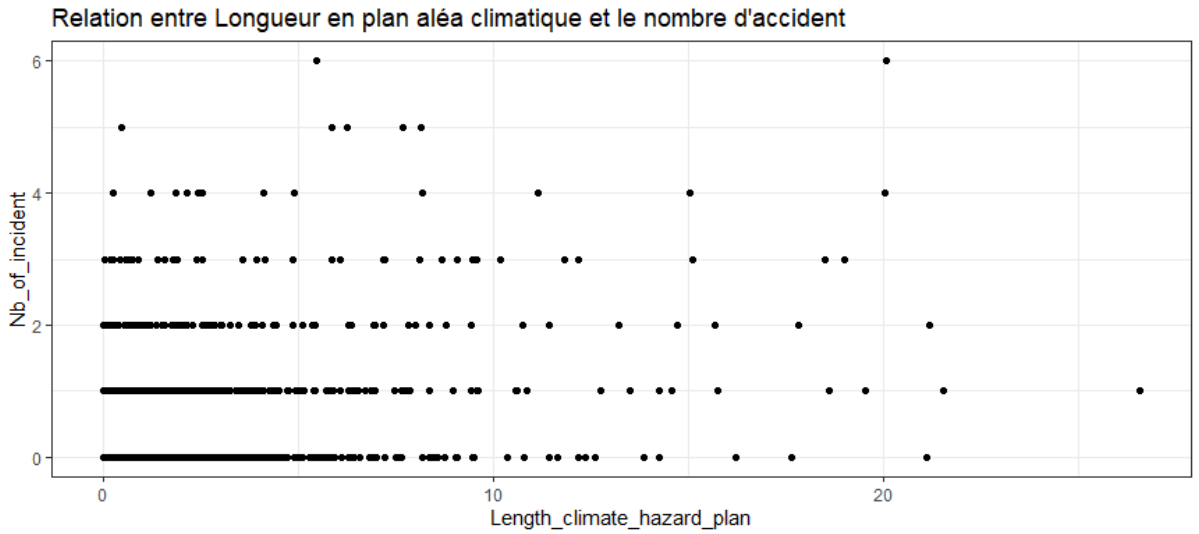


Il n'y a pas de tendance claire ou de motif évident qui indiquerait une relation forte entre la longueur du tronçon électrique et le nombre d'accidents. Les points semblent être répartis de manière assez aléatoire sur le graphique, suggérant qu'il n'y a pas de corrélation directe ou simple entre ces deux variables.

Selon ce graphique de dispersion, il n'y a pas de preuve visuelle d'une relation linéaire entre la longueur des tronçons électriques et le nombre d'accidents. Si une relation existe, elle pourrait être non linéaire ou influencée par d'autres facteurs non représentés dans ce graphique.

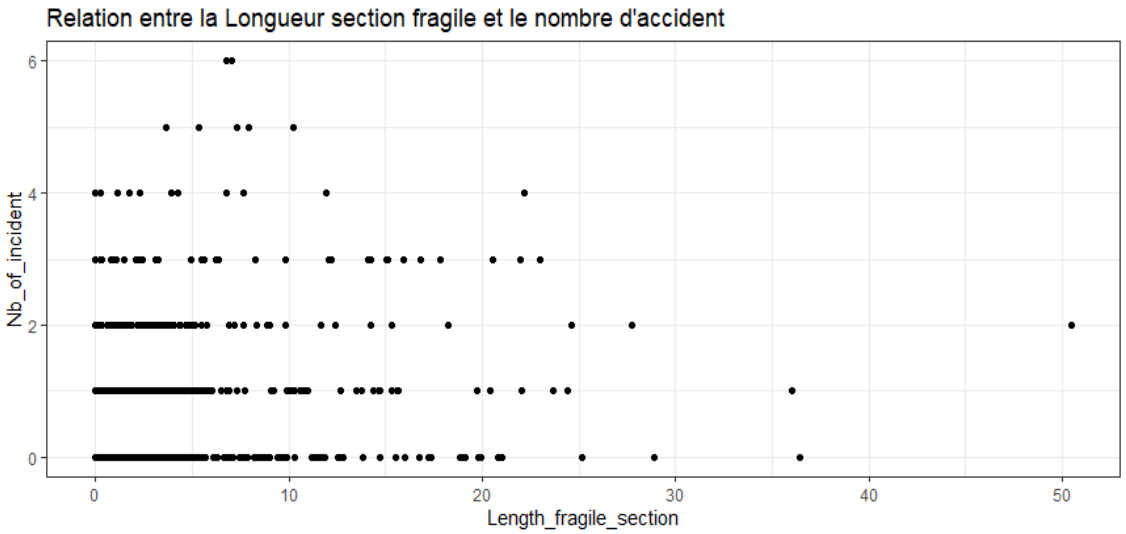


Il n'y a pas de tendance claire ou de corrélation évidente qui émerge du graphique. Le nombre d'accidents n'augmente pas de manière significative avec le nombre d'anomalies. Une anomalie ne se transforme pas forcément en accident, et un accident n'est pas forcément source d'une anomalie.



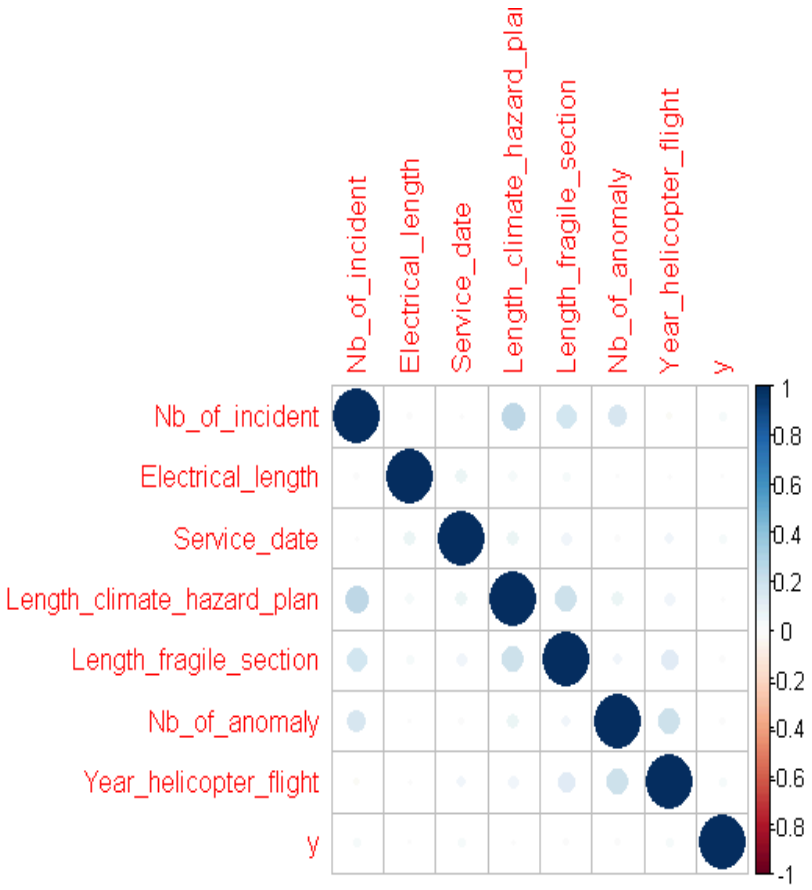
Les points semblent être répartis sans un motif clair ou une tendance qui indiquerait une relation directe entre la longueur des plans d'aléa climatique et le nombre d'accidents. Même pour des longueurs de plan plus importantes, le nombre d'accidents reste relativement bas (généralement inférieur ou égal à 2), ce qui suggère qu'il n'y a pas d'augmentation significative des accidents avec l'augmentation de la longueur du plan d'aléa climatique. La conclusion tirée de ce graphique est similaire à celle des graphiques précédents : il n'y a pas de corrélation claire et directe entre les deux variables étudiées. Cela peut indiquer que les plans d'aléa climatique, en tant que tels, ne sont pas un facteur prédominant dans la

survenue d'accidents, ou que d'autres variables non représentées sur ce graphique pourraient avoir un impact plus significatif sur le nombre d'accidents. Il est également possible que les plans soient efficaces pour mitiger les risques d'accidents liés au climat, ce qui expliquerait l'absence de corrélation.



La distribution des points n'indique pas de corrélation évidente ou de relation linéaire entre la longueur des sections fragiles et le nombre d'accidents. Cela signifie que, sur la base de ce graphique, on ne peut pas conclure que des sections fragiles plus longues conduisent à un nombre plus élevé d'accidents. Ce graphique ne montre pas de preuve visuelle d'une relation directe entre la longueur des sections fragiles et le nombre d'accidents. Cela pourrait suggérer que, bien que l'intuition puisse indiquer que des sections plus fragiles pourraient être associées à plus d'accidents, d'autres facteurs non représentés ici peuvent influencer la fréquence des accidents.

Matrice de corrélation

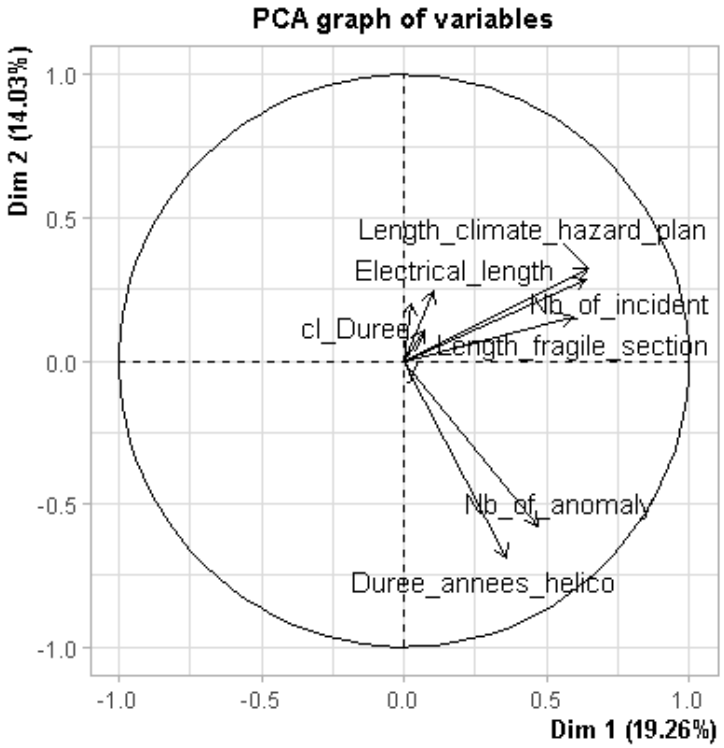


D'après la matrice de corrélation, il semble que certaines variables montrent une corrélation plus forte avec d'autres par rapport à ce qui a été interprété des graphiques de dispersion individuels. Par exemple, nb_of_anomaly pourrait montrer une corrélation plus forte avec nb_of_accident que ce que l'on pourrait conclure d'un seul graphique de dispersion. Cela ne contredit pas nécessairement l'analyse graphique précédente, mais cela ajoute une dimension supplémentaire à l'analyse en fournissant une mesure quantitative de la corrélation.

Les techniques de visualisation avancées comme **l'Analyse en Composantes Principales (PCA) et la méthode des K-means, avec le nombre de clusters K optimisé grâce à la méthode du coude**, ont été des outils analytiques essentiels dans notre processus d'exploration des données. Ces méthodes ont facilité notre compréhension profonde des caractéristiques sous-jacentes des données et ont mis

en exergue des structures et des groupements latents qui suggèrent des pistes de modèles prédictifs à évaluer. Grâce à ces visualisations, nous avons pu déceler et interpréter des patterns subtils au sein de nos données.

Analyse en composante principale



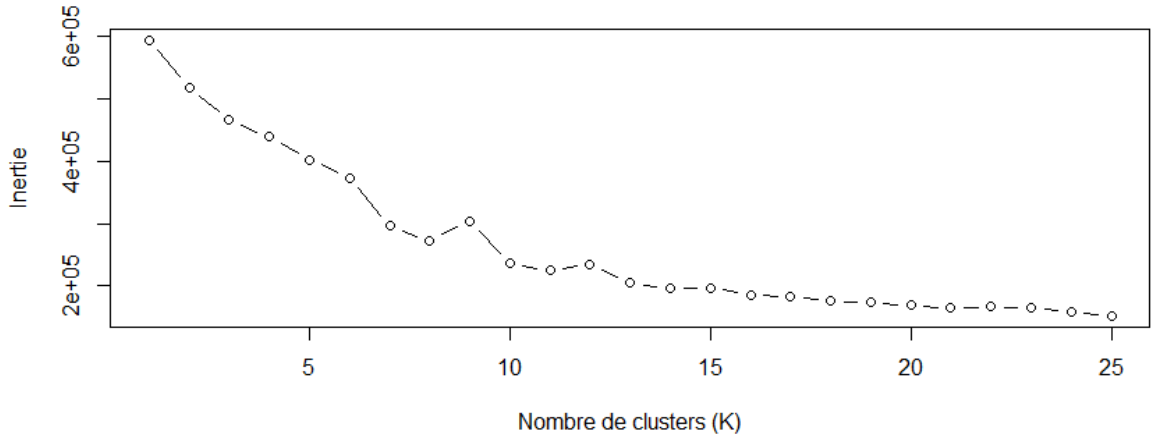
L'acp explique environ **34% de la variance** :

Le nombre d'anomalies et l'année du dernier passage en hélicoptère sont corrélés entre eux, ce qui semble logique. Si on tarde à inspecter une ligne électrique, des problèmes peuvent apparaître sur celle-ci. **Ainsi, plus ancienne est la dernière vérification des lignes, plus il existe une chance qu'elle connaisse un besoin d'intervention**.

"Length_fragile_section", "nb_of_incident" et "length_climate_hazard_plan" sont corrélées entre eux. **Une ligne fragile ou victime d'aléas climatiques connaît plus d'incident que les autres**.

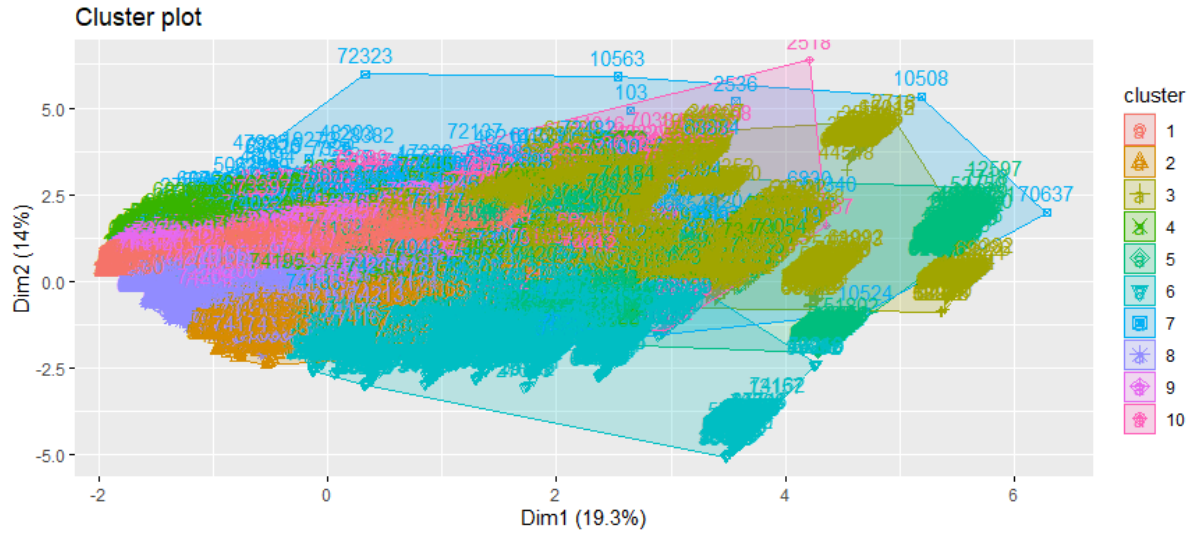
La date de mise en service et la longueur du tronçon électrique sont aussi corrélés entre eux. Il est difficile de savoir pourquoi, l'hypothèse la plus probable est que les premières lignes créées ont été placées sur les grands axes, et ont donc une taille supérieure aux autres.

Méthode du coude



La diminution de l'inertie ralentit sensiblement après 10 clusters, ce qui suggère que 10 pourrait être un choix raisonnable pour le nombre de clusters.

Clusters plot



Notre cluster plot est difficile à interpréter en détail au vu de la quantité de données. Cependant, les clusters ont l'air d'être regroupés, de taille identique, et se chevauchant. Cela indique qu'il n'y a pas de groupes qui se distinguent clairement des autres, les variables semblent homogènes. Il y a quelques variables qui sont à des extrémités, mais pas assez éloignées pour être interprétées comme des outliners. Enfin, le plus important, les clusters sont de forme allongés, ce qui pourrait indiquer des relations non linéaires, détail pouvant avoir son importance pour le choix des modèles futurs.

Modélisation

L'analyse statistique et la modélisation prédictive jouent un rôle essentiel dans la prise de décision et la résolution de problèmes dans de nombreux domaines, notamment en sciences sociales, en économie et en sciences de la santé. Dans le cadre de cette étude, nous avons entrepris de développer un modèle de **régression logistique** (logit) pour répondre à un problème de **classification binaire**. Cette introduction présente les raisons pour lesquelles nous avons choisi le modèle logit, ses points forts et ses limites.

Le modèle de régression logistique est un choix courant dans les cas où la variable cible est binaire ou dichotomique, c'est-à-dire qu'elle peut prendre deux valeurs distinctes telles que oui/non, succès/échec, etc. Dans notre étude, notre variable cible prend la valeur 0 ou 1, **nous avons donc opté pour ce modèle en raison de sa simplicité et de sa capacité à fournir des résultats interprétables**. En effet, le modèle logit est largement utilisé pour analyser les relations entre une variable binaire dépendante et un ensemble de variables **indépendantes** continues et/ou catégorielles.

Le modèle logit fournit des estimations directes des probabilités de classe, ce qui le rend particulièrement utile dans les cas où l'accent est mis sur la prédiction de probabilités plutôt que sur la simple classification. En outre, le modèle logit peut gérer efficacement des ensembles de données de taille moyenne à grande avec un temps de calcul raisonnable.

Cependant, malgré ses avantages, le modèle logit présente également certaines limitations. Tout d'abord, il suppose une relation linéaire entre les variables indépendantes et la log-odds de la variable dépendante, ce qui peut ne pas être approprié dans tous les cas. De plus, il peut être sensible à la présence de valeurs aberrantes et à la multicolinéarité entre les variables indépendantes. Enfin, le modèle logit peut rencontrer des difficultés avec des ensembles de données déséquilibrés où les classes de la variable dépendante sont fortement biaisées.

Malgré ces limitations, le modèle logit reste un outil puissant et largement utilisé dans de nombreux domaines de recherche et d'application. Dans cette étude, nous explorons son utilisation pour répondre à notre problème de classification binaire et analysons ses performances par rapport à d'autres méthodes de modélisation.

3. Subdivision

La subdivision de l'ensemble de données en un ensemble d'entraînement et un ensemble de tests est une pratique standard dans le domaine de l'apprentissage automatique et de la modélisation prédictive. Cette approche est essentielle pour **évaluer de manière fiable les performances d'un modèle et pour estimer sa capacité à généraliser à de nouvelles données non vues**. Dans notre étude, cette subdivision est cruciale pour plusieurs raisons.

Tout d'abord, en divisant nos données en un ensemble d'entraînement et un ensemble de test, nous nous assurons que le modèle est formé sur une partie des données et évalué sur une autre partie indépendante. Cela nous permet d'estimer de manière réaliste les performances du modèle sur des données qu'il n'a pas encore vues, ce qui est essentiel pour évaluer son utilité pratique.

De plus, la subdivision des données en un ensemble d'entraînement et un ensemble de test nous permet de détecter et de réduire le risque de surajustement (overfitting) du modèle. Le surajustement se produit lorsque le modèle s'adapte trop étroitement aux données d'entraînement spécifiques et ne parvient pas à généraliser correctement à de nouvelles données. En évaluant le modèle sur un ensemble de test distinct, nous pouvons identifier si le modèle a tendance à sur-ajuster les données d'entraînement et prendre des mesures pour y remédier, telles que la simplification du modèle ou l'utilisation de techniques de régularisation.

De plus, en divisant les données en ensembles d'entraînement et de test, nous nous assurons que nos résultats d'évaluation du modèle sont fiables et non biaisés. Sans cette subdivision, il existe un risque que le modèle soit évalué sur les mêmes données sur lesquelles il a été entraîné, ce qui peut conduire à une estimation trop optimiste de ses performances.

Nous avons fait le choix de subdiviser avec un taux de 0.75, c'est-à-dire que **75% de nos données seront dans l'ensemble train, et 25% dans l'ensemble test**.

4. Modèle logit

Dans le cadre de notre étude, nous cherchons à comprendre les déterminants de l'intervention sur la ligne électrique en 2023. Notre variable cible, **notée y, est binaire, représentant la décision d'intervenir ou de ne pas intervenir sur chaque tronçon de ligne électrique**. Notre objectif est d'utiliser un modèle logit pour modéliser cette décision d'intervention en fonction des caractéristiques des tronçons électriques.

Le modèle logit nous permettra de prédire la probabilité d'intervention sur chaque tronçon de ligne électrique en 2023 en fonction des variables explicatives disponibles.

En utilisant un modèle logit, **nous chercherons à identifier les variables explicatives qui sont significativement associées à la probabilité d'intervention.** Cela nous permettra de comprendre quels sont les facteurs qui influencent le plus la décision d'intervenir sur la ligne électrique en 2023. Ces informations pourront ensuite être utilisées pour orienter les décisions de maintenance et d'investissement dans le réseau électrique, en identifiant les tronçons à haut risque nécessitant une intervention prioritaire.

Dans le cadre de notre analyse avec le modèle logit, nous obtiendrons des coefficients associés à chaque variable explicative incluse dans le modèle. Ces coefficients nous fournissent des **informations sur la force et la direction de l'association entre chaque variable explicative et la probabilité d'intervention sur la ligne électrique en 2023.**

Plus précisément, chaque coefficient représente le changement logarithmique dans les cotes (odds) de l'intervention sur la ligne électrique pour un changement d'une unité dans la variable explicative correspondante, toutes les autres variables étant maintenues constantes. Un coefficient positif indique que l'augmentation de la valeur de la variable explicative est associée à une augmentation des chances d'intervention, tandis qu'un coefficient négatif indique une association inverse.

L'écriture du modèle logit peut être exprimée comme suit :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Où :

- p est la probabilité d'intervention sur la ligne électrique en 2023.
- β_0 est l'intercept du modèle.
- $\beta_1, \beta_2, \ldots, \beta_n$ sont les coefficients associés aux variables explicatives $x_1, x_2, \ldots, x_n$, respectivement.

Cette équation est ensuite utilisée pour estimer les probabilités d'intervention sur la ligne électrique en fonction des valeurs des variables explicatives. Les coefficients du modèle nous permettent d'évaluer l'importance relative de chaque variable explicative dans la prise de décision d'intervention et de quantifier l'impact de ces variables sur la probabilité d'intervention.

En interprétant les coefficients du modèle, nous pourrons identifier les facteurs les plus influents sur la décision d'intervention et comprendre comment ces facteurs contribuent à la probabilité d'intervention sur la ligne électrique en 2023. Cette information sera précieuse pour orienter les stratégies de maintenance et d'investissement dans le réseau électrique, en identifiant les facteurs clés à prendre en compte pour améliorer la fiabilité et la durabilité du système électrique.

En réalisant ces premières étapes, on peut avoir un aperçu du résultat de notre modèle:

```
Call:
glm(formula = y ~ ., family = "binomial", data = train)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.9102489  0.1786892 -33.076  < 2e-16 ***
Nb_of_incident             0.2319849  0.0337531   6.873 6.29e-12 ***
Electrical_length          0.0000979  0.0001349   0.726    0.468
Service_date               0.0206914  0.0037223   5.559 2.72e-08 ***
Length_climate_hazard_plan 0.0068049  0.0103928   0.655    0.513
Length_fragile_section     0.0061756  0.0058848   1.049    0.294
Nb_of_anomaly             -0.0546804  0.0106481  -5.135 2.82e-07 ***
Year_helicopter_flight     0.1698680  0.0259300   6.551 5.71e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7077.5  on 55660  degrees of freedom
Residual deviance: 6932.8  on 55653  degrees of freedom
AIC: 6948.8

Number of Fisher Scoring iterations: 7
```

Voici une interprétation des sorties clés :

- (**Intercept**) : L'ordonnée à l'origine est de -5.91, ce qui signifie que lorsque toutes les autres variables explicatives sont à 0, la log-odds de 'y' (intervention) est de -5.91. Le Pr(>|z|) indique que l'ordonnée à l'origine est significativement différente de 0.
- **Coefficients** : Chaque estimateur représente le changement dans la log-odds de 'y' pour une augmentation d'une unité de la variable correspondante, en maintenant constantes toutes les autres variables.
    - Par exemple, Nb_of_incident a un estimateur de 2.31e-01, ce qui indique qu'une augmentation d'un incident est associée à une augmentation de la log-odds de 'y' de 0.231.
    - Nb_of_anomaly a un estimateur négatif (-1.699e-01), ce qui suggère qu'une augmentation d'un an est associée à une diminution de la log-odds de 'y' de 0.1699.
- **Significativité** : Les étoiles (***, **, *) après les p-valeurs (Pr(>|z|)) indiquent le niveau de significativité statistique. Les variables avec des étoiles sont considérées comme ayant un impact statistiquement significatif sur la variable cible. Ici, Nb_of_incident, Nb_anomalies, Service_date, et Duree_annees_helico sont significatives au niveau 0.001.

- **Null deviance et Residual deviance** : Ces valeurs montrent la différence entre un modèle ne contenant que l'intercept (null) et le modèle complet avec toutes les variables (residual). Une grande différence indique que le modèle complet fournit un meilleur ajustement que le modèle nul.

- **AIC** : L'AIC (Akaike Information Criterion) est un indicateur de la qualité du modèle. Un modèle avec un AIC plus bas est généralement préféré. Ici, il est de 6948.8

Seulement, ces informations ne seront utiles que plus tard, **lorsque nous aurons confirmé la performance de ce modèle.**

Pour ce faire, nous allons appliquer le modèle entraîné sur l'ensemble de test. Cela va nous sortir des probabilités associées à y (=1, i.e nécessité d'intervention), puis nous les convertirons en 0 ou en 1.

-1 si p > 0,5

-0 sinon

En faisant une **matrice de confusion**, on obtient la sortie suivante:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 18341   212
         1     0     0

               Accuracy : 0.9886
                 95% CI : (0.9869, 0.9901)
    No Information Rate : 0.9886
    P-Value [Acc > NIR] : 0.5183

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.9886
         Neg Pred Value :    NaN
             Prevalence : 0.9886
         Detection Rate : 0.9886
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

Notre modèle a prédit 0 pour tous les cas, il y a **18341 vrais négatifs.**

Il y a **212 faux négatifs** (FN), où le modèle a incorrectement prédit la classe 0 alors que la vraie classe était 1. **Aucun vrai positif (TP) ou faux positif (FP) n'est noté**, signifiant qu'aucun cas positif (1) n'a été correctement ou incorrectement prédit.

L'exactitude du modèle est de 0.9886 ou **98.86%**, ce qui semble très élevé. Cependant, cela est trompeur car presque toutes les instances dans les données sont de la classe 0. Le modèle prédit simplement la classe majoritaire à chaque fois.

De plus, le coefficient Kappa est de 0, indiquant que l'accord entre les prédictions et les références **n'est pas mieux que le hasard.**

Nous sommes confrontés à un défi important, à savoir le déséquilibre de classes dans notre ensemble d'entraînement. Cette situation est particulièrement problématique car elle peut avoir un impact significatif sur les performances de notre modèle. En examinant de plus près les données, nous constatons que la classe "0", représentant l'absence d'intervention sur la ligne électrique, est fortement sur-représentée avec un total de **58 692 occurrences**, tandis que la classe "1", indiquant une intervention, est sous-représentée avec seulement **680 occurrences**. Cette disparité dans la répartition des classes peut entraîner des biais dans l'apprentissage du modèle, où il peut avoir tendance à privilégier la classe majoritaire au détriment de la classe minoritaire lors de la prise de décision. En conséquence, cela peut conduire à des prédictions biaisées et peu fiables, car le modèle peut avoir du mal à identifier correctement les exemples de la classe minoritaire. Il est impératif de prendre des mesures pour remédier à ce déséquilibre de classes afin d'améliorer la capacité du modèle à généraliser et à faire des prédictions précises sur les exemples de la classe minoritaire.

4.1 Logit plutôt que Probit

Avant de passer au rééchantillonnage, nous allons rapidement expliquer pourquoi nous avons choisi d'utiliser un modèle logit.

En effet, si l'on utilise un modèle Probit, voici ce que l'on obtient :

```
Call:
glm(formula = y ~ ., family = binomial(link = probit), data = train)

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.313e+02  2.046e+01   6.416 1.40e-10 ***
Nb_of_incident            9.247e-02  1.334e-02   6.930 4.21e-12 ***
Electrical_length         3.521e-05  5.242e-05   0.672    0.502
Length_climate_hazard_plan 3.008e-03  4.031e-03   0.746    0.456
Length_fragile_section    2.964e-03  2.364e-03   1.254    0.210
Nb_of_anomaly            -2.021e-02  3.904e-03  -5.178 2.25e-07 ***
Year_helicopter_flight   -6.631e-02  1.013e-02  -6.548 5.83e-11 ***
Duree_annees              8.443e-03  1.442e-03   5.855 4.76e-09 ***
Duree_annees_helico              NA         NA      NA       NA
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7077.5  on 55660  degrees of freedom
Residual deviance: 6928.6  on 55653  degrees of freedom
AIC: 6944.6

Number of Fisher Scoring iterations: 7
```

On peut voir que l'AIC est légèrement inférieur à celui obtenue pour le Logit (6948.8)
Cependant, un modèle Logit permet une interprétation plus poussée des coefficients :

Le modèle logit estime les log-odds d'un événement. Les log-odds sont le logarithme naturel du ratio entre la probabilité qu'un événement se produise et la probabilité qu'il ne se produise pas. Il est donc possible de les transformer (directement sur R, grâce au package "Broom") en odd-ratio, et donc de les interpréter (ce que nous ferons plus tard avec le modèle logit final).

Exemple : Si ma variable "x" a un odd-ratio de 2, une augmentation d' une unité de "x", double les chances d'avoir à intervenir sur la ligne.

Il est plus difficile de calculer cela pour un modèle Probit, et l'odd ratio, associé à la probabilité marginale, permet donc une interprétation des coefficients plus complète.

De plus, le modèle Logit prend mieux en compte les événements extrêmes. Sachant que les incidents (variable y) que l'on souhaite analyser sont des cas plutôt rares,et donc potentiellement "extrêmes", le Logit y sera mieux adapté.

5. Le déséquilibre de classe

Pour remédier au déséquilibre de classes dans notre ensemble de données d'entraînement, nous avons décidé d'utiliser la technique de sur-échantillonnage synthétique (**SMOTE - Synthetic Minority Over-sampling Technique**). SMOTE est une méthode efficace pour équilibrer les classes en générant de nouveaux exemples synthétiques de la classe minoritaire, basés sur des combinaisons linéaires des exemples existants. Cette approche permet de créer un ensemble de données d'entraînement plus équilibré sans répéter simplement les exemples de la classe minoritaire, ce qui réduit le risque de surajustement.

En utilisant SMOTE, nous allons **augmenter le nombre d'exemples de la classe minoritaire, dans notre cas les exemples d'intervention sur la ligne électrique, pour les rendre proportionnels à la classe majoritaire**. Cela permettra au modèle d'apprendre de manière plus équilibrée des deux classes et d'améliorer sa capacité à généraliser aux nouvelles données, en particulier aux exemples de la classe minoritaire. Cette approche devrait aider à atténuer les biais induits par le déséquilibre de classes et à améliorer les performances de prédiction globales du modèle.

Après avoir appliqué SMOTE pour rééquilibrer les classes, nous pourrons ré-entraîner notre modèle sur l'ensemble de données équilibré et évaluer ses performances à nouveau. En utilisant cette approche, **nous sommes confiants que nous pourrons améliorer la capacité de notre modèle à faire des prédictions précises sur les exemples de la classe minoritaire, ce qui est crucial pour résoudre notre problème de prédiction d'intervention sur la ligne électrique en 2023**.

Grâce à SMOTE, nous arrivons à une proportion de 1360 "0" et 1360 "1".

Nous pouvons alors appliquer de nouveau notre modèle logit, et nous avons les sorties suivantes:

```
Call:
glm(formula = y ~ ., family = binomial(link = logit), data = train)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.778e+00  2.065e-01  -8.610  < 2e-16 ***
Nb_of_incident             2.661e-01  3.828e-02   6.951 3.62e-12 ***
Electrical_length         -7.890e-06  1.472e-04  -0.054   0.9573
Service_date               2.851e-02  4.480e-03   6.363 1.98e-10 ***
Length_climate_hazard_plan 4.441e-03  1.165e-02   0.381   0.7031
Length_fragile_section     1.271e-02  7.564e-03   1.681   0.0928 .
Nb_of_anomaly             -5.742e-02  9.739e-03  -5.896 3.73e-09 ***
Year_helicopter_flight     1.605e-01  2.711e-02   5.922 3.18e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3604.4  on 2599  degrees of freedom
Residual deviance: 3449.5  on 2592  degrees of freedom
AIC: 3465.5

Number of Fisher Scoring iterations: 4
```

L’AIC (3465.5) obtenu est beaucoup plus bas, ce qui indique un bien meilleur modèle que précédemment.

Le pseudo R2 est une mesure de la qualité de l’ajustement du modèle. On va le calculer ici grâce au package “pscl”, et on obtient pour ce modèle :

```
fitting null model for pseudo-r2
         llh       llhNull            G2      McFadden          r2ML          r2CU
-1.724744e+03 -1.802183e+03  1.548773e+02  4.296937e-02  5.782872e-02  7.710496e-02
```

Nous pouvons voir que l’on obtient un **pseudo R2 de 0.0429**, ce qui est extrêmement faible. Notre modèle a donc une faible qualité d’ajustement. Sa capacité prédictive sera faible, il nous faudra donc utiliser un autre modèle dans ce but. Les coefficients restent cependant interprétables.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 11528    87
         1  6813   125

               Accuracy : 0.6281
                 95% CI : (0.6211, 0.6351)
    No Information Rate : 0.9886
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0131

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.62854
            Specificity : 0.58962
         Pos Pred Value : 0.99251
         Neg Pred Value : 0.01802
              Precision : 0.99251
                 Recall : 0.62854
                     F1 : 0.76966
             Prevalence : 0.98857
         Detection Rate : 0.62136
   Detection Prevalence : 0.62604
      Balanced Accuracy : 0.60908

       'Positive' Class : 0
```

La matrice de confusion affiche :
Vrais négatifs : Le modèle a bien prédit “y = 0” 11528 fois.
Faux positifs : Le modèle a mal prédit "y = 1” 87 fois, car en réalité, c’était des lignes ne nécessitant pas d’intervention.
Faux négatifs : Le modèle a mal prédit “y=1” 6813 fois, alors que c’était des cas où il fallait intervenir.
Vrais positifs : Le modèle a correctement prédit "y = 1" 125 fois.
L'accuracy est de 0.6281, ce qui signifie que le modèle a correctement prédit 63.5% des cas.
Sensitivity : La sensibilité est de 0.62854, indiquant que le modèle a correctement identifié 62.85% des cas réels de 'non-accident'.
Specificity : La spécificité est de 0.589, ce qui signifie que le modèle a correctement identifié 58.9% des cas réels d'accidents.

Ces résultats sont meilleurs que ceux obtenus précédemment, cependant le Kappa est de 0.0131, ce qui est très bas, indiquant que l'accord entre les prédictions et les vraies étiquettes n'est pas beaucoup mieux que le hasard.

Voici donc notre modèle Logit finale :

logit(p) = - 1.778

+ (0.27*Nombre d'incident)

- (0.000079*Longueur electrique)

+ (0.0285* Date de mise en service)

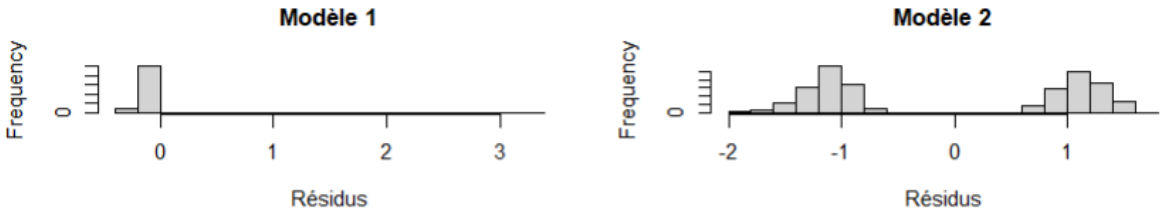
+ (0.00044* Longueur tronçon soumis au risque climatique)

+ (0.013* Longueur de la section fragile)

- (0.057* Nombre d'anomalies)

+ (0.16* Date du dernier passage en hélicoptère)

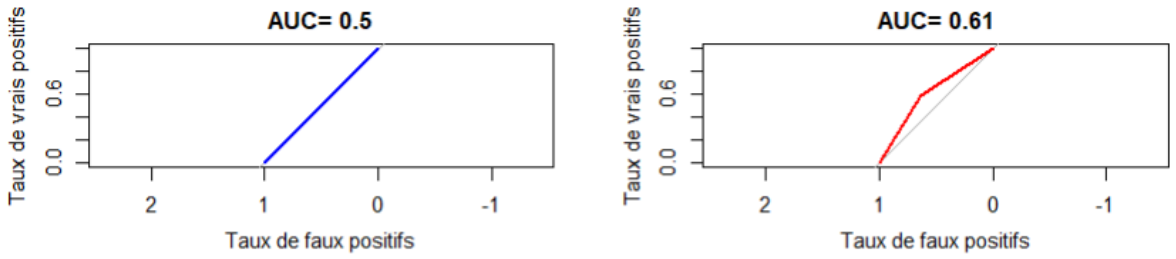
Voici un test sur les résidus. Le modèle 1 correspond au logit sans rééchantillonnage, le modèle 2 avec.



Nous pouvons voir que le modèle 1 a une distribution des résidus fortement asymétrique, avec une concentration autour de 0. La non-normalité et l'asymétrie des résidus suggèrent que le modèle pourrait ne pas bien capturer toute la structure des données.

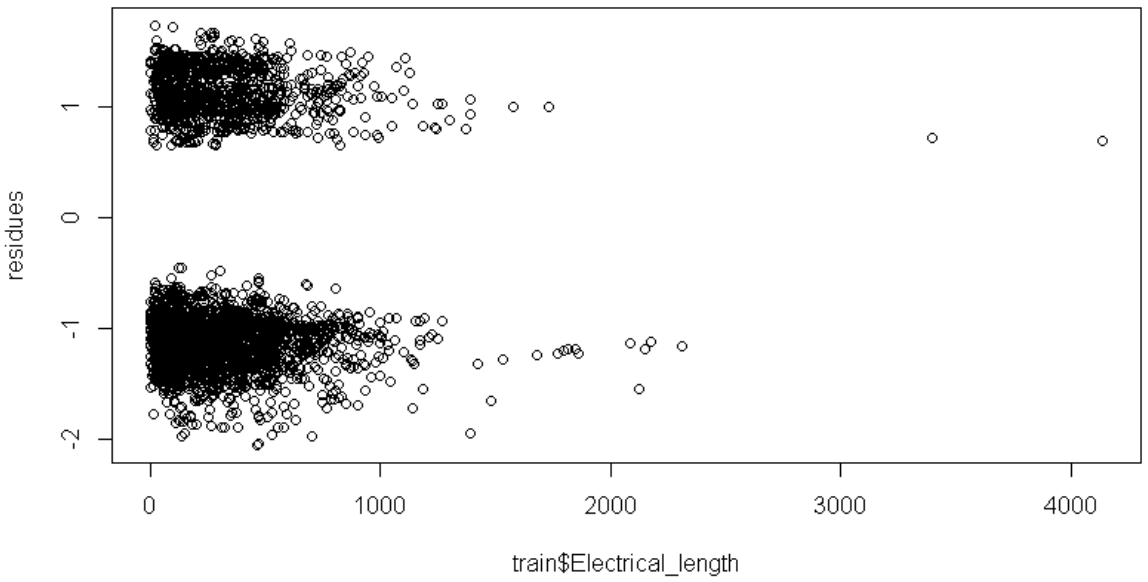
Le modèle 2 quant à lui à une distribution bien plus symétrique et ressemblant davantage à une loi normale. Il semble donc avoir un meilleur ajustement.

Passons maintenant à la courbe AUC : La courbe AUC, pour "Area Under the ROC Curve", est une mesure de performance pour les modèles de classification qui indique leur capacité à distinguer entre les classes. Une AUC de 1 représente un modèle parfait, tandis qu'une AUC de 0.5 indique une performance qui ne vaut pas mieux qu'un choix aléatoire. Plus l'AUC est proche de 1, meilleur est le modèle.

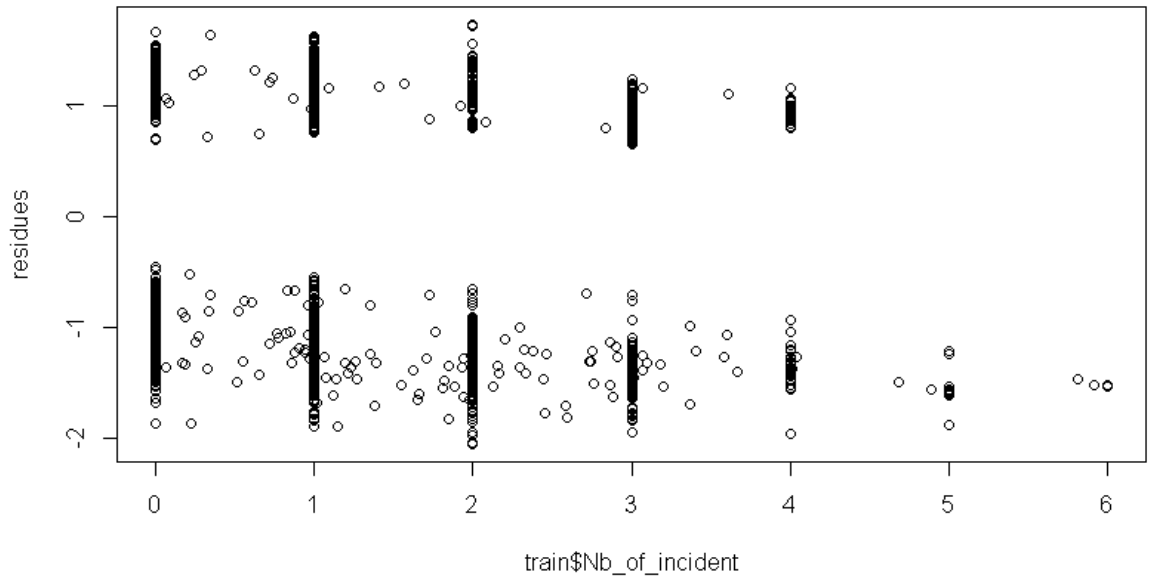


Nous pouvons voir que le modèle 1 (courbe bleue) a une AUC de 0.5, il ne prédit donc pas mieux que le hasard. Pour le modèle 2, l'AUC a une valeur de 0.61, il a donc une meilleure capacité prédictive que le hasard (et donc que le modèle 1). Cependant, cela reste faible. Sachant que nous utilisons ce modèle notamment pour l'analyse des coefficients, cette faible capacité prédictive est le point le plus important.

Enfin, lorsque l'on teste la relation entre une variable non significative (ici la longueur d'un tronçon) et notre variable y, voici ce que l'on obtient. Nous pouvons voir une absence de tendance linéaire. Les résidus sont regroupés en deux zones distinctes, ce qui peut indiquer que le modèle ne capture pas pleinement la relation entre notre variable et la variable cible. C'est une explication probable quant à la non significativité de la variable dans ce modèle, et donc de la limite du Logit dans notre cas.



Lorsqu'on le compare avec une variable significative (ici le nombre d'accidents), nous pouvons voir apparaître des résidus bien moins regroupés, ce qui montre que le modèle capte mieux les relations entre ces deux variables.



6. Interprétation de l'importance des variables

Signe des coefficients

Dans notre modèle, **la plupart des variables sont significatives**, seul la longueur du tronçon électrique ("electrical_lenght") et la portion de la ligne subissant des contraintes climatiques ("Lengh_climate_hazard_plan") ne le sont pas.

! Attention, cela ne veut pas dire que ces variables n'ont pas un réel impact sur la probabilité que la ligne subisse un incident ! En effet, comme dit précédemment, notre modèle **a du mal à capter toute la structuré des données**, notamment les relations non linéaires. Toutes les interprétations que nous allons faire ci-dessous ne vont donc **montrer que des tendances, et sont donc à prendre avec du recul**. La modélisation grâce à un random forest plus tard pourra nous aider à confirmer ou non ces interprétations.

Ainsi, les variables "date de mise en service", "nombre d'incident que la ligne a connu" et "temps depuis le dernier passage en hélicoptère" et "longueur fragile du tronçon" ont un coefficient positif et sont significatif. Il semblerait donc qu' une hausse d' une unité de ces variables augmente la probabilité d'avoir besoin

d'intervenir sur la ligne. Ceci est cohérent, car au plus un tronçon est vieux, plus il a connu de multiple incident, plus sa dernière inspection est ancienne ou encore plus la longueur fragile du tronçon est grande, au plus le tronçon à de chance d'avoir un besoin d'intervention.

La variable "nombre d'anomalie" a un coefficient négatif. Même si au premier abord cela peut sembler étrange, nos données ne nous fournissent pas le nombre d'anomalies signifiant réellement un besoin d'intervention.

Par exemple, une anomalie peut être un mouvement de la ligne très important détectés par des capteurs défaillants, traduisant une "fausse anomalie".

On en déduit que notre modèle a du mal à capter la complexité de cette variable, la variable "piégeant" notre modèle, qui n'est pas assez performant dans les relations non linéaires complexes comme celle-ci. (source : confirmé par l'équipe d'Enedis)

Effet marginaux

```
Nb_of_incident Electrical_length Service_date Length_climate_hazard_plan Length_fragile_section
       0.06267        -1.858e-06     0.006714                  0.001046                0.002994
Nb_of_anomaly Year_helicopter_flight
      -0.01352                0.0378
```

Prenons l'exemple du nombre d'incident:

une augmentation d'une unité du nombre d'incident : **passage de 1 à 2, implique une augmentation de 0.06267 de la probabilité d'avoir une intervention**. C'est cohérent avec notre sujet, au plus il y a d'incidents, au plus il y a nécessité d'intervenir sur la ligne.

Pour la longueur électrique:

Une augmentation d'une unité de la longueur (ex= 1m) **réduit la probabilité d'une intervention sur la ligne de 0.000001858**. Le coefficient étant très faible, l'impact de la longueur du tronçon n'impacte pas réellement la nécessité d'intervention sur la ligne.

Odd-ratio :

	term	estimate	std.error	statistic	p.value	conf.low	conf.high
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	0.169	0.207	-8.61	7.33e-18	0.112	0.252
2	Nb_of_incident	1.30	0.0383	6.95	3.62e-12	1.21	1.41
3	Electrical_length	1.00	0.000147	-0.0536	9.57e- 1	1.00	1.00
4	Service_date	1.03	0.00448	6.36	1.98e-10	1.02	1.04
5	Length_climate_hazard_plan	1.00	0.0117	0.381	7.03e- 1	0.982	1.03
6	Length_fragile_section	1.01	0.00756	1.68	9.28e- 2	0.998	1.03
7	Nb_of_anomaly	0.944	0.00974	-5.90	3.73e- 9	0.926	0.962
8	Year_helicopter_flight	1.17	0.0271	5.92	3.18e- 9	1.11	1.24

Nous pouvons voir que les odd-ratios confirment nos analyses précédentes. Le nombre d'incident sur les lignes et la date du dernier passage en hélicoptère sont les variables augmentant le plus les chances de devoir intervenir sur la ligne par rapport à la probabilité de ne pas à avoir à intervenir, tandis que “Nb_of_ Anomaly” baisse cette chance, ce qui est cohérent avec le signe négatif trouvé plus haut.

Par exemple, “Nb_of_incident” a un odd-ratio de 1.3. Cela signifie qu' une hausse d' une unité de cette variable augmente de 30% les chances d’avoir y=1 par rapport au chance d’avoir y=0 (Toutes choses égale par ailleurs)

7. Modèle de prévision

Afin de prédire nos futures valeurs, nous nous sommes dirigés vers un modèle “Random Forest”.

Un Random Forest est un algorithme d'apprentissage automatique qui combine de nombreux arbres de décision pour faire des prédictions plus précises et stables.

Ce modèle présente de nombreux avantages au vu de notre base de données. En effet, il est capable de capter les relations non linéaires, que l’on pourrait avoir comme l’a montré notre cluster plot. Il nécessite peu de préparation de données, et gère bien les valeurs extrêmes, comme ce que l’on peut avoir ici, les accidents sur les lignes étant rares. Enfin, il est quand même possible de voir l’importance des variables avec ce modèle, ce qui peut corroborer notre analyse vue avec le modèle Logit ci-dessus.

Afin de calibrer correctement notre modèle, nous avons mis en place une grille de paramètre ainsi qu’une validation croisée.

La grille de paramètre nous permet d’optimiser les performances de notre modèle, elle nous permet de tester plusieurs paramètres afin de trouver ceux qui nous donnent les meilleures performances. La validation croisée aide à estimer les performances du modèle sur des données non vues, ce qui peut réduire le risque de

surajustement. En utilisant la validation croisée, nous réduisons le risque de biais de sélection dans l'évaluation du modèle. Cela garantit que les performances rapportées ne sont pas simplement dues à une chance particulière dans la répartition des données.

Résultats du modèle

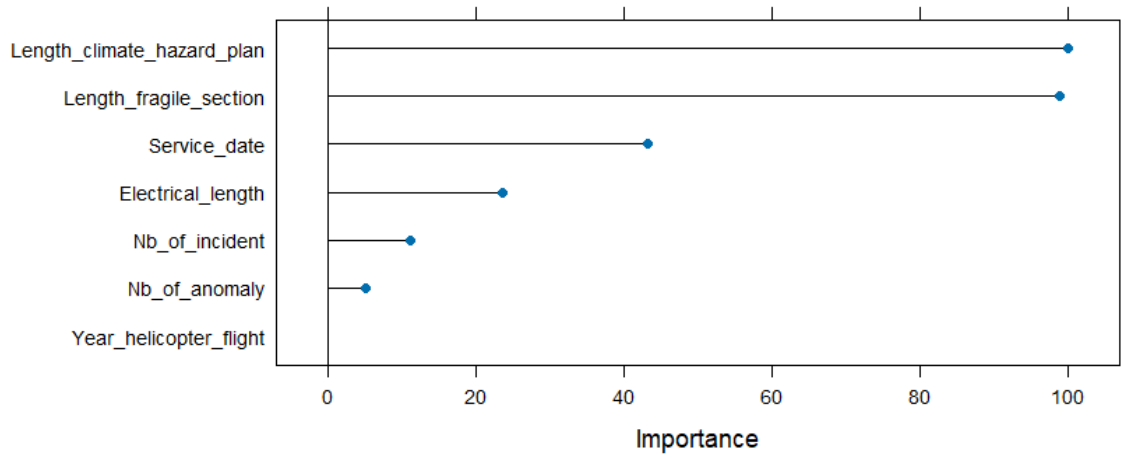
Prédictions	**0**	**1**
0	16838	27
1	1503	185

Accuracy	**Kappa**
0.9184615	0.8369231

Le modèle obtient une précision remarquable de 91.8%.

Le kappa de 0.8369231 indique un très bon niveau d'accord entre les prédictions de notre modèle et les valeurs réelles, avec une certaine correction pour les coïncidences aléatoires.

Importance des variables



On retrouve un résultat **complètement différent du modèle logit**:

“Length climate hazard plan” est la variable la plus importante de notre modèle, alors que dans le logit, elle n’était pas significative du tout. **Les variables les plus importantes sont maintenant les moins**.

Les différences d'importance des variables entre le modèle de forêt aléatoire et le modèle de régression logistique peuvent être attribuées aux **différences fondamentales dans la manière dont ces deux modèles fonctionnent et traitent les données.**

Le modèle de forêt aléatoire est un modèle d'apprentissage automatique basé sur un ensemble d'arbres de décision, où chaque arbre est construit en utilisant un sous-ensemble aléatoire des données et des variables. Dans ce cadre, l'importance des variables est calculée en mesurant l'impact de chaque variable sur la réduction de l'impureté des nœuds de l'arbre, généralement mesurée par le critère de Gini ou l'entropie. Ainsi, les variables qui contribuent le plus à la réduction de l'impureté sont considérées comme les plus importantes pour la prédiction.

En revanche, le modèle de régression logistique est un modèle statistique qui modélise la relation entre une variable dépendante binaire et un ensemble de variables indépendantes en utilisant une fonction logistique. Dans ce contexte, l'importance des variables est souvent mesurée à l'aide de statistiques telles que les coefficients de régression, qui indiquent la force et la direction de l'association entre chaque variable et la variable dépendante.

Ainsi, les différences dans les mesures d'importance des variables entre les deux modèles peuvent découler des méthodes de modélisation sous-jacentes et des mesures spécifiques utilisées pour évaluer l'importance des variables. Il est également important de noter que les deux modèles peuvent avoir des hypothèses différentes sur la relation entre les variables et la variable cible, ce qui peut influencer les résultats. Par conséquent, lors de l'évaluation et de la comparaison des résultats des deux modèles, il est crucial de prendre en compte les différences conceptuelles et méthodologiques entre eux.

Conclusion

En conclusion de ce projet de microéconométrie, nous avons élaboré un modèle de régression logistique robuste et un modèle de forêt aléatoire pour prévoir la nécessité d'intervention sur les lignes électriques d'Enedis en Bretagne. Notre objectif initial de comprendre les déterminants influençant la probabilité d'intervention a été atteint à travers l'application de ces techniques de modélisation avancées.

La régression logistique nous a permis de discerner l'influence marginale des différentes variables sur la probabilité d'une intervention, en mettant en lumière l'importance du nombre d'incidents, de la longueur des sections fragiles, et de l'année du dernier vol d'hélicoptère. Notre analyse a cependant révélé des limites dans la capacité prédictive de ce modèle, en partie dues au déséquilibre des classes et aux relations non-linéaires au sein de notre ensemble de données.

Nous avons adressé ce déséquilibre grâce à la technique SMOTE, qui a permis de rééquilibrer notre ensemble d'entraînement et de raffiner notre modèle logistique. Cependant, les résultats ont indiqué que, malgré une amélioration, la qualité d'ajustement du modèle restait modeste, comme le suggèrent le pseudo R^2 et l'AIC.

Selon le modèle logit, pour qu'Enedis réalise des économies dans ses déplacements de maintenance, les variables les plus importantes à considérer selon le modèle logit sont: le nombre d'incidents, la date du dernier passage en hélicoptère, la date de mise en service.

L'objectif de ce projet était avant tout de mieux appréhender les modèles probit et logit, mais face aux difficultés rencontrées, notamment sur la non linéarité des relations, nous avons pu en découvrir les limites.

Face à cette situation, nous avons implémenté un modèle de forêt aléatoire qui a surpassé la régression logistique en termes de précision et de capacité à gérer les non-linéarités et les valeurs extrêmes de notre jeu de données. L'importance des variables dégagées par le modèle de forêt aléatoire diffère significativement de celle observée dans la régression logistique, ce qui souligne l'importance de considérer différents modèles pour appréhender au mieux les données complexes.

Ainsi, selon le modèle Random Forest, les variables les plus importantes à considérer pour Enedis sont: **la longueur du tronçon impactée par les conditions météos, la longueur du tronçon fragile, et la date de mise en service**.

Les résultats des deux modèles ne sont pas les mêmes, mais il est important de les considérer de façon complémentaire afin d'obtenir différents points de vue. Il pourrait être intéressant de traiter la non linéarité des relations entre les variables (plusieurs tentatives ont été réalisées dans ce but, notamment le passage en log de certaines variables, mais sans résultat). Nous trouvions le projet très intéressant, mais face à ces difficultés, nous nous sommes rendus compte que les données étaient peut être insuffisantes, le fait que certaines variables ont été omises (anonymisation) ont rendu la tâche plus compliquées pour le modèle logit.