MAINTENANCE PRÉVENTIVE

1. Contexte

Enedis, acteur majeur dans le secteur de la **distribution d'électricité** en France, assure la gestion et l'exploitation des réseaux électriques. Le programme de surveillance et de maintenance de son réseau en Bretagne est une initiative cruciale permettant d'assurer la **fiabilité** et la **sécurité** de la distribution d'électricité dans la région.

Chaque année, l'entreprise effectue un survol aérien de 1/3 du réseau aérien moyenne tension en Bretagne. L'objectif : inspecter l'intégralité du réseau sur une période de trois ans. Généralement réalisé par hélicoptère, ce survol vise à détecter d'éventuelles anomalies structurelles. Suite à ce premier survol, le service responsable de la rénovation programmée entreprend un **dépouillement** des données collectées. Cela implique de définir les tronçons à visiter en utilisant des drones pour des diagnostics plus approfondis.

Dans le contexte du réseau électrique, un **tronçon** fait référence à une portion spécifique d'une ligne électrique qui nécessite une attention particulière en termes de surveillance, de diagnostic, et éventuellement de travaux de maintenance. Les critères de sélection incluent les **anomalies** détectées lors du survol, les **incidents** survenus au cours de l'année précédente, le **plan aléa climatique**, et l'état global du réseau.

Après l'obtention des diagnostics grâce aux survols et aux données recueillies, Enedis planifie les **maintenances** nécessaires pour l'année à venir. Cette phase est essentielle pour garantir la pérennité du réseau électrique breton, minimiser les risques d'incidents, et assurer une distribution stable et fiable de l'électricité dans la région.

Dans ce contexte, l'objectif est de planifier les maintenances nécessaires pour l'année suivante, anticipant ainsi les opérations de maintenance préventive.

Nous visons donc spécifiquement à identifier les tronçons à visiter en priorité, en se limitant à 25 km à inspecter dans le but de minimiser les coûts de déplacement.

2. Exploration de données

Les données mises à notre disposition ont préalablement été réparties dans deux bases dédiées à la modélisation. En l'occurrence, nous disposons d'une base d'apprentissage comportant 74214 tronçons et 9 caractéristiques dont un identifiant qui leur est propre. La seconde base, est dédiée au test des performances de nos modèles et comporte les 18854 tronçons restants.

Toutes les données, générées en 2022, à l'exception de l'immobilisation RP en 2023, sont disponibles. Les colonnes comprennent l'identifiant du tronçon, le nombre d'incidents, la date de mise en exploitation, la longueur de la section fragile, l'année du dernier vol, la longueur électrique du tronçon, la longueur en plan aléa climatique, le nombre d'anomalies, et l'année de la dernière maintenance. Les données manquantes indiquent l'absence d'événements.

2.1 Valeurs manquantes et doublons

Aux vues des recommandations d'ENEDIS, notre premier traitement portait sur la gestion des valeurs manquantes. En l'occurrence, toutes ces valeurs ont été imputées par 0. C'est le cas pour les variables suivantes :

- Nombre d'incident sur le départ : 49.7%
- Nombre d'anomalies sur le départ : 67.8%
- Année de la dernière maintenance sur le départ : 88.8%

Ensuite, nous nous sommes aussi assurés de l'absence de doublons dans la base ainsi que dans la variable « Identifiant du tronçon ».

2.2 Corrélation

Nous avons ensuite procédé à une évaluation des corrélations entre variables avec la cible et entre prédicteurs. Nous avons alors évaluer ces corrélations à 0.26 au plus haut entre la longueur du plan aléa climatique et le Nombre d'incidents. Nous observons ensuite une corrélation qui s'élève à 0.21 entre les prédicteurs « *Longueur de sections fragiles* » et « *Plan aléa climatique* ». En général, nous pouvons dire que ces corrélations sont faibles avec la cible autant qu'entre prédicteurs.

2.3 Traitement individuel de variable

Tout d'abord, nous avons binariser notre variable d'incidents.

Nous avons transformé « *Date de mise en exploitation* » en ancienneté en prenant comme référence l'année 2023. Nous avons également supprimé les observations pour lesquelles la variable « *Année du dernier vol d'hélicoptère* » prenait « 2022 » comme modalité. La raison, on ne sait pas si le survol en hélicoptère a été fait avant ou après l'incident.

Nous avons surtout remplacé les observations qui prenaient la modalité 1 pour « *Nombre d'incidents au départ* » mais qui n'avaient pas eu de maintenance car cela est impossible.

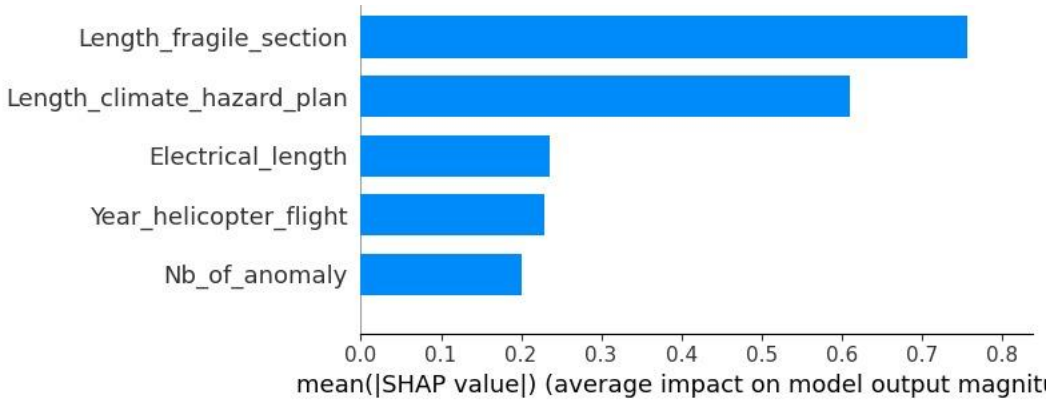
Dans le traitement de base, le plus important était la compréhension des incidents par tronçons. Les tronçons appartenant à la même zone avaient le même nombre d'incidents et d'anomalies. Cela s'explique par le fait qu'on coupe toute la ligne lors d'une intervention, de ce fait nous avons gardé seulement les incidents qui avaient eu une maintenance, cela veut dire que seuls les tronçons ayant eu une maintenance ont réellement présenté un incident.

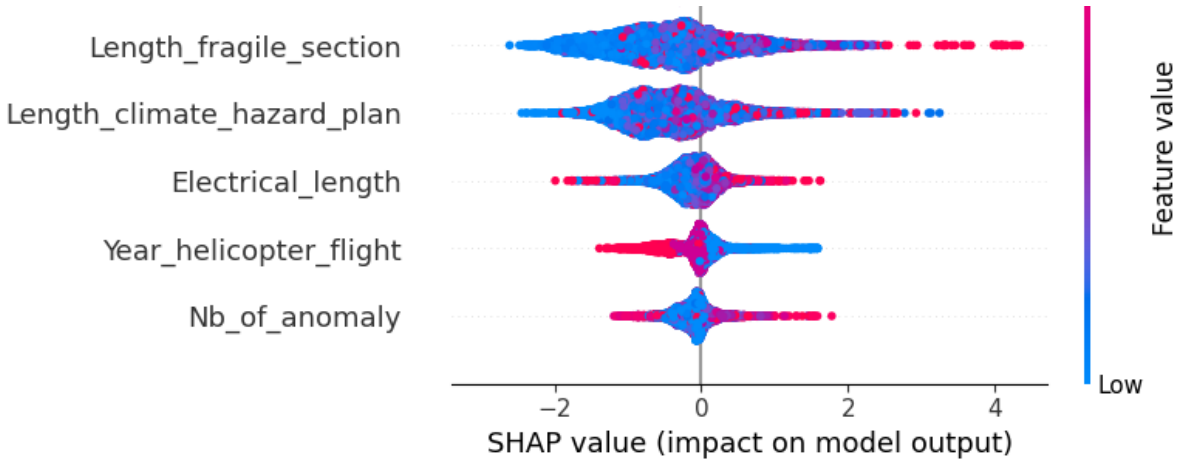
3. Modélisation

Liste de variables sélectionnées		
Longueur électrique	*Date de mise en exploitation*	*Plan de hasard climatique*
Longueur de sections fragiles	*Nombre d'anomalies*	

Nous avons comparé plusieurs modèles de machine learning et utilisé la validation croisée pour sélectionner la meilleure combinaison de paramètres pour chaque modèle. Le modèle nous fournissant les meilleurs résultats est le Gradient Boosting (Xgboost) avec lequel on obtient une MSE moyenne de 0.0380 pour les paramètres {'learning_rate' : 0.1, 'max_depth': 12, 'n_estimators': 150}. L'AUC obtenue est de 0,79% montre que notre modèle a une bonne capacité de discrimmination entre nos deux classes, et est robuste aux variations des métriques et aux traitements des données avec des classes déséquilibrées. Le temps de computation de notre modèle est de 4.4 secondes.

La matrice de confusion nous donne un taux de vrai positif, de 71,70% et un taux de vrai négatif de 71.86%. Donc 7 fois sur 10, notre modèle arrive à prédire les tronçons sur lesquels un incident survient véritablement.





4. **Que retenir ?**

La régression logistique a l'avantage d'être plus interprétable notamment avec les valeurs des coefficients et leur significativité. D'autre part nous pouvons faire des interprétations avec les rapports de cotes. En revanche, elle est en situation de sur-apprentissage puisque les performances sur l'échantillon d'apprentissage sont meilleurs que sur le test en termes de f1-Score et d'AUC. Enfin, comparativement avec le modèle xGBoost, ce modèle s'avère moins performant. D'où le défi d'arbitrer entre compréhension et performance du modèle.

BARRAUD Lorenzo
BRIAND Léo
MIRZA Simon
SONDEJI Karl