

Rapport d'analyse des temps de réponse de la London Fire Brigade



2025

Créé par : Léa Vauchel, Anne-Sixtine Lerebours, Alix Lavoipierre, Célia Taider

Cohorte : Novembre 2024 - formation continue

Mentor du projet : Elliott Douieb

Issues de milieux professionnels très différents les unes des autres, mais avec l'objectif commun de réussir notre projet de reconversion dans la data-analyse, nous nous sommes réunies autour d'un même sujet : la brigade des pompiers de Londres.

Le choix de ce thème a été motivé par nos propres centres d'intérêt, notre curiosité, mais aussi par le challenge du projet et l'amplitude des données à disposition, qui représentent un réel défi pour nous.

Contexte et objectif



L'objectif de notre sujet est d'analyser et/ou estimer les temps de réponse de la brigade des pompiers de Londres (la London Fire Brigade, que l'on appellera par son acronyme : LFB).

Afin de réaliser un travail de qualité, nous avons tout d'abord effectué quelques recherches sur la LFB. On peut considérer que les prémices de la LFB ont débuté à partir du 1er janvier 1833, là où 10 brigades d'assurance indépendantes se sont rejointes pour former le London Fire Engine Establishment. En 1862, elle est devenue la "Metropolitan Fire Brigade" à la suite des incendies de Tooley Street. En 1904, elle est renommée sous le nom de "London Fire Brigade". Pionnière dans l'évolution des méthodes modernes de lutte contre les incendies, elle a notamment joué un rôle majeur dans plusieurs événements historiques tels que : le Blitz (incendies massifs dus aux bombardements entre 1940 et 1941 durant la Seconde Guerre mondiale), l'incendie de la Grenfell Tower en 2017 ou le King Cross Fire en 1987. Il est d'ailleurs intéressant d'ajouter que la brigade de Londres a mis en place des plans quinquennaux notamment depuis l'incendie de la tour Grenfell le 14 juin 2017. Afin de mieux aborder notre étude, nous avons également réalisé quelques recherches sur le plan actuel (de 2024 à 2029). Nous savons par exemple que l'objectif de temps d'arrivée du premier camion (on parlera de première pompe) est de 6 minutes et 8 minutes pour le second. D'autres objectifs sont également définis dans ce plan, comme consacrer davantage de temps à la prévention contre les incendies (visites sur site, conseils gratuits sur la sécurité à domicile...) ou

encore réviser les politiques opérationnelles pour distinguer les appels qui demandent des conseils et ceux qui ont besoin d'être secourus. Avec plus de 100 casernes dans toute la capitale et environ 5000 pompiers, elle se distingue également par la diversité de sa flotte qui comprend des pompes (camion de pompiers), des échelles, des bateaux ou encore des véhicules spécialisés. Deux unités sont également dédiées aux opérations à haute risque (chimique, biologique ou nucléaire).

La population de Londres devrait atteindre près de 10 millions d'habitants d'ici les 10 prochaines années. Avec un taux de 6%, Londres reste l'une des villes où la croissance démographique est la plus forte. Il est intéressant d'ajouter que Londres possède 85% des gratte-ciel d'Angleterre, soit un des plus grands nombres de gratte-ciel en Europe, avec plus de 8 000 immeubles de grande hauteur, ce qui équivaut à 100 fois la moyenne nationale. Couvrant toute la région du Grand Londres, la LFB se doit d'être opérationnelle sur environ 1600 km² (sur 32 circonscriptions administratives, appelées Boroughs) afin de protéger et servir plus de 9 millions d'habitants.

Le projet que nous avons choisi a un intérêt à la fois sanitaire, logistique et économique (volet qui sera traité dans ce projet uniquement si le temps nous le permet) puisqu'il a l'ambition d'analyser des incidents sanitaires, d'optimiser l'organisation de la brigade afin de répondre à ces urgences.



Compréhension du sujet et manipulation des données

Compréhension du sujet :

Nous avions à disposition plusieurs jeux de données sous divers format (Excel et csv). Nous disposions de 3 jeux de données sur les mobilisations de la LFB, avec les informations concernant les camions de pompiers déployés, et 2 jeux de données répertoriant les incidents de la LFB. L'ensemble de ces données couvre la période de janvier 2009 à novembre 2024 ce qui représente presque 16 ans d'informations (il est intéressant de noter que la brigade de Londres a rendu ses données accessibles et les actualise annuellement). Nous pouvons donc considérer que l'intervalle est suffisamment grand pour pouvoir analyser correctement le set de données, mais aussi réaliser des estimations fiables sur les temps de réponse.

Nous avons, dans un premier temps, réalisé une concaténation de tous les fichiers concernant les incidents d'une part et de tous les fichiers concernant les mobilisations d'une autre part. Cette concaténation nous a donné un premier fichier des incidents comprenant 39 colonnes et 1 782 434 lignes. Le second fichier sur les mobilisations contient 24 colonnes et 2 496 800 lignes. Nous avions également à disposition deux documents sur la métadata nous permettant de mieux comprendre l'ensemble des informations contenues dans nos *DataFrames*. Nous avons donc fait une analyse exhaustive de celui-ci sous forme de tableau avec identification du nom de la colonne, sa signification, son type, le nombre de valeurs possibles (si variables catégorielles) et le taux en % de valeurs manquantes.

Afin de poursuivre notre étude, nous avons “mergé” nos deux fichiers en un seul *DataFrame* en prenant les identifiants d'incidents comme colonne commune. Le choix d'une fusion grâce à la méthode *inner* permet de conserver uniquement les données communes aux deux fichiers. Avant cela, nous avons réalisé une étude approfondie de nos données afin de lever les éventuels freins à cette fusion. Nous avons par exemple remarqué que le fichier 2009-2014 n'avait pas entré les identifiants de la même manière que les autres (format 1234.00 pour l'un et 1234 pour l'autre). De 2020 à 2024, le nom des *Boroughs* et des quartiers concernés par les incidents ont été ajoutés dans les formulaires. Le 9 Janvier 2014, dix casernes de pompiers ont été fermées dans le cadre du cinquième plan de sécurité de Londres (LSP5). Les zones des casernes ont été modifiées pour refléter ces fermetures, elles ont été réparties dans les zones des casernes adjacentes. Toutefois, ceci n'a pas été problématique car, pour fournir des données d'incident cohérentes, les terrains des casernes ont été modifiés pour tous les incidents de cet ensemble. Sur cette même période, les informations sur les dates et heures de retour à la station ne sont pas données. Nous avons également constater des doublons entre 2020 et 2024, en analysant ces lignes nous nous sommes aperçu que le code ressource était différent et donc qu'un même incident pouvait engendrer plusieurs

lignes en fonction du nombre de ressource sollicitées. Nous n'avons donc pas supprimé ces "faux doublons". Autant de problématiques auxquelles nous avons dû trouver des solutions et explications afin d'avancer sur notre projet tout en conservant une bonne qualité de données.

Le fichier final a pour dimensions : 2 488 987 lignes et 60 colonnes.

Notre DataFrame final ayant plus de 2 millions de lignes, il est important d'optimiser les performances en ajustant certains paramètres pour éviter les problèmes de mémoire. Nous avons donc lu les fichiers avec certains paramètres comme "low_memory = False".

Afin de gagner en efficacité et dans un souci de complémentarité, nous avons fait le choix de prendre connaissance du sujet et des données chacune de notre côté et de se partager nos recherches au cours de nombreuses réunions en tenant une "feuille de route". Ceci explique pourquoi nous n'avons pas forcément les mêmes codes car nous avons fait le choix de mutualiser nos recherches.

Nous avons ensuite ajouter d'autres colonnes afin de réaliser plus de statistiques comme "month" pour les mois de l'année, "weekday" pour les jours de la semaine ou encore "Inner/outer" pour distinguer les arrondissements de Londres et ceux de la périphérie. Nous avons également trouvé grâce au site "Historique Météo", l'ensemble des données météorologiques de l'année 2009 jusqu'à novembre 2024 afin de compléter notre jeu de données et ainsi analyser les temps de réponse de la LFB en fonction des conditions climatiques mais aussi en fonction de la visibilité au kilomètre.

Pour visualiser l'existence de relations entre différentes paires de variables catégorielles et numériques et mesurer la force de corrélation, nous avons réalisé une heatmap (pour les variables numériques) et des test statistiques. Voici trois exemples de nos tests statistiques :

- AttendanceTimeSeconds (temps d'assistance totale) et DelayCode_Description (description du code de retard) :

Résultat du test ANOVA : p-value de 0.0 (<5%), ce qui signifie qu'il y a une corrélation entre ces deux variables.

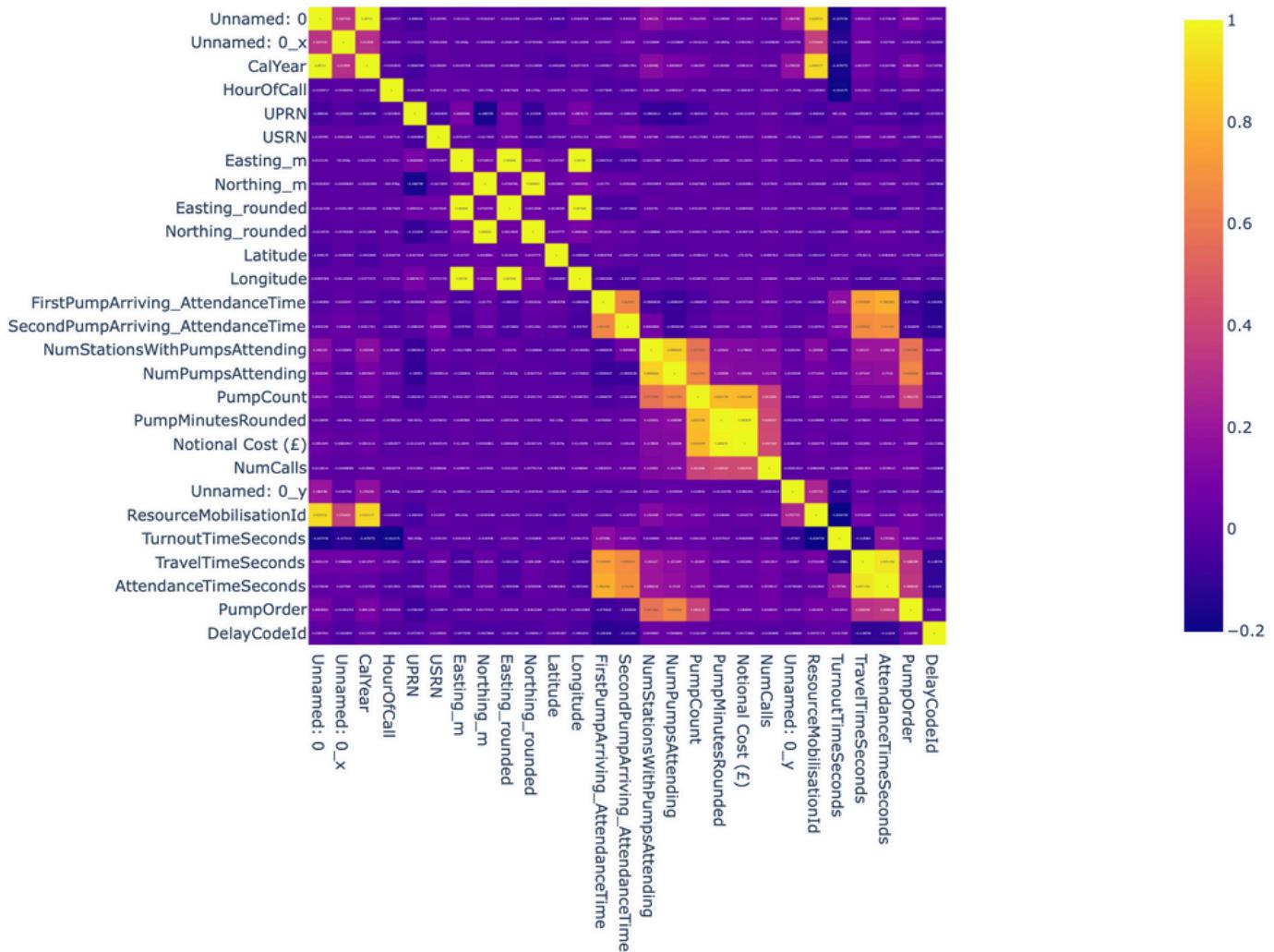
- FristPumpArriving_AttendanceTime (temps d'arrivée de la première unité) et DeployedFromStation_Name (nom de la station déployée) :

Résultat du test ANOVA : p-value de 0.0 (<5%), on conclut donc à une corrélation entre ces deux variables.

- StopCodeDescription (code correspondant à la catégorie de l'incident) et PropertyCategory (catégorie du type de propriété concerné par l'incident) :

Résultat du test du Khi-2 : p-value de 0.0 (<5%), il y a une corrélation significative entre ces deux variables.

Corrélation des variables numériques



Avec cette heatmap nous avons un aperçu global des corrélations entre les différentes variables numériques, nous pouvons donc déjà diriger notre étude.

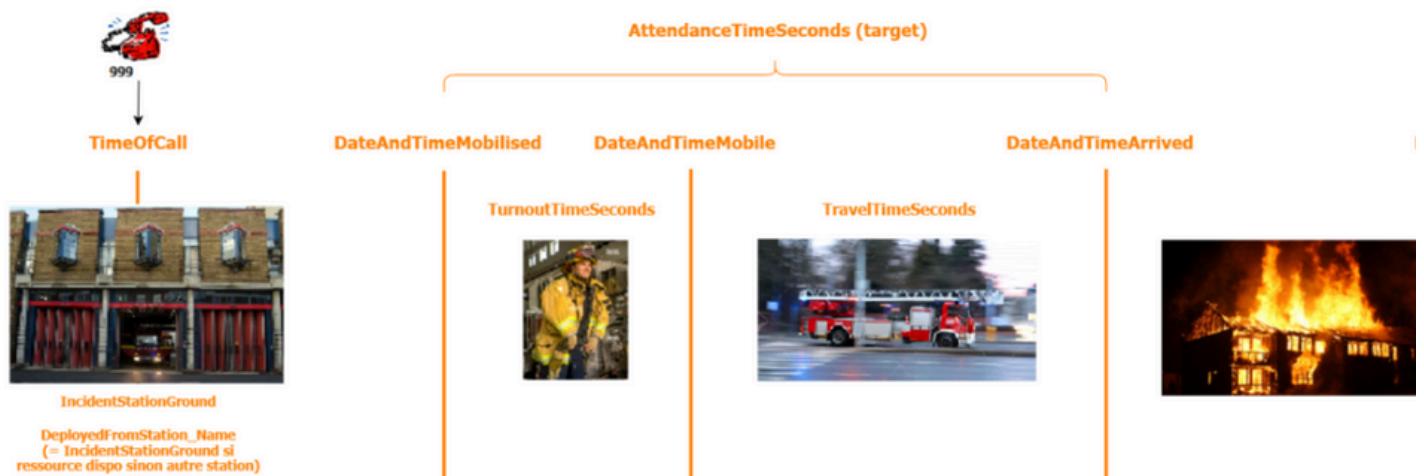
Choix des variables :

La variable cible :

Après étude et visualisation de nos données et en concertation avec notre mentor de projet, nous avons décidé de garder comme variable cible la “FirstPumpArriving_AttendanceTime”, soit le temps de réponse de la première pompe. Notre choix s'est porté sur cette variable, car la première pompe à être mobilisée est la plus importante parce qu'elle doit être la plus rapide pour assurer la sécurité des biens et des personnes.

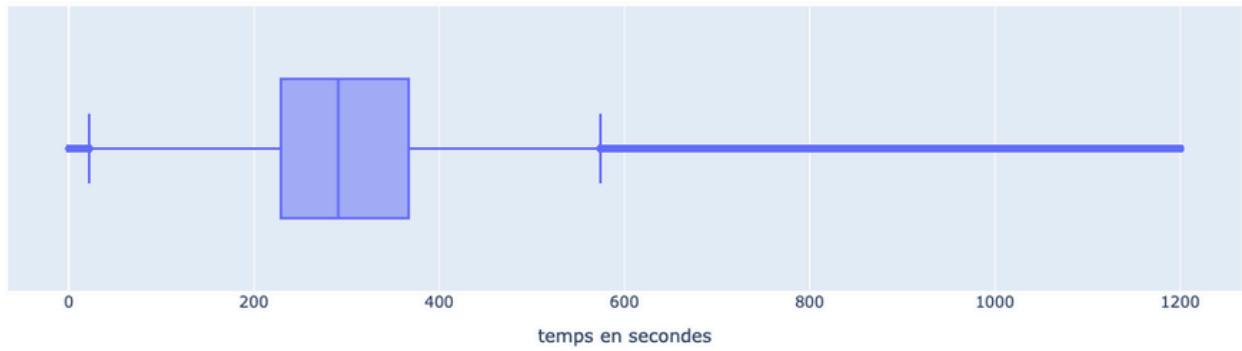
Afin d'éviter le risque de multicolinéarité et donc d'éventuellement perturber notre futur modèle de Machine Learning, nous avons supprimé toutes les variables ayant une distribution similaire à celle-ci comme : le temps de trajet, le temps de préparation après mobilisation et le temps d'assistance toutes pompes confondues.

Voici une représentation de ce qu'est “AttendanceTime” :



Cette variable est enregistrée en secondes, ce qui nous permet d'être extrêmement précises sur notre analyse et sur nos prédictions. Cette précision est nécessaire dans le milieu du service à la personne, car la moindre minute perdue peut avoir des conséquences importantes sur les incidents en cours. On estime qu'un incendie peut doubler sa taille en 30 à 60 secondes, une minute perdue auprès d'une personne en arrêt cardio-respiratoire représente 10% de chance de survie en moins.

Temps d'assistance de la première unité



On observe que les données sont comprises entre 19 et 579 secondes (moins de 10 minutes). On observe beaucoup d'outliers, ce qui nous donne des données allant de 1 seconde à 1200 secondes (20 minutes). 50 % des temps de réponse de la première équipe sont compris entre 229 et 369 secondes (3 minutes 49 secondes et 6 minutes et 9 secondes). Nous sommes sur des temps similaires, mais plus courts que le temps d'assistance globale. Nous pouvons donc en déduire que lorsqu'une équipe est la première à être mobilisée, elle se déplace de manière plus rapide. En moyenne, la première pompe à être déployée se déplace en 305 secondes, soit environ 5 minutes. Si nous faisons le comparatif entre le plan de la LFB et nos statistiques, on peut affirmer que les temps de trajets collent, la plupart du temps, aux objectifs de moins de 6 minutes de la LFB.

Les variables explicatives :

Tableaux récapitulatifs de l'ensemble des variables et justification de la sélection ou suppression de celle-ci :

Légende :



Variables explicatives sélectionnées



Variable cible

Variables administratives :

Colonnes :	Exemple :	Description :	Valeurs manquantes et gestion	Format et Distribution :	Justification de la sélection ou non :
IncidentNumber	000008-01012018	N° de l'incident	0	Int - /	Variable supprimée car n'explique pas la variable cible, simple identifiant administratif
NumStationsWithPumpsAttending	1	Nombre de stations ayant déployé une équipe	0,8 %	Float - /	Variable supprimée car nous cherchons à connaître le temps de réponse de la première équipe peu importe le nombre d'équipe déployé.
NumPumpsAttending	1	Nombre d'équipes déployées	0,8 %	Float - /	Variable supprimée pour les mêmes raisons que 'NumStationsWithPumpsAttending'.
PumpCount	1	Nombre d'équipes impliquées	0	Int - /	Variable supprimée pour les mêmes raisons que 'NumStationsWithPumpsAttending'.
Notional Cost (£)	328	Coût estimé de l'intervention (basé sur temps et ressources utilisées)	0	Int - /	Variable supprimée pour le moment qui sera remise après si le temps nous le permet afin d'analyser le volet économique de ce sujet. Mais au vu de notre objectif cette variable ne présente aucun intérêt pour notre analyse.
NumCalls	1	Nombre d'appel pour l'incident	0,1 %	Float - /	Variable supprimée car n'a pas d'incidence sur le temps de réponse de la LFB. Une équipe est mobilisée sur un incident dès le premier appel.
ResourceMobilisationId	5055153	Identifiant de mobilisation	0	Int - /	Variable supprimée car il s'agit d'un identifiant unique de mobilisation. Il

					ne nous apporte aucune justification sur le temps de réponse de l'équipe déployée.
Resource_Code	A392	Code ressource (matériel, camion)	0	Object - /	Cette variable aurait été intéressante à garder mais nous n'avions pas les données nécessaires pour l'exploiter. Les simples codes ne suffisent pas à déterminer de quelle catégorie de matériel il s'agit et nous n'avons pas trouvé les réponses à nos questions. Nous avons donc supprimé cette variable.
PerformanceReporting	1	Inclusion ou non dans des rapports de performance	0	Object - 1, 2, 'Not Used'	Variable supprimée car cette catégorie renseigne une décision prise après l'intervention et n'a donc pas d'influence sur le temps de réponse de l'unité.
PlusCode_Code	Initial	Code de la nature de la mobilisation	0	Object - /	Variable supprimée car lors de notre Dataviz nous avons vu qu'il s'agissait dans 97% des cas d'une mobilisation initiale. De plus, comme nous nous intéressons seulement à la première unité déployée sur l'incident, il s'agira toujours d'une mobilisation initiale.
PlusCode_Description	Initial Mobilisation	Description détaillée de la mobilisation	0	Object - /	Idem que pour 'PlusCode_Code'.

Variables temporelles :

Colonnes :	Exemple :	Description :	Valeurs manquantes et gestion	Format et Distribution :	Justification de la sélection ou non :
DateOfCall	2018-01-01 00:00:00	Date et heure précise de l'appel	0	Datetime - /	Variable supprimée car après la Dataviz nous nous sommes aperçues qu'il n'y avait pas d'incidence sur le temps d'assistance en fonction du jour, du mois ou de la période de l'année. Nous nous sommes cependant servie de cette variable pour étudier les temps de réponses en fonction des conditions météorologiques.
CalYear	2018	Année de l'appel	0	Int - Année de 2009 à 2024	Variable sélectionnée car nous avons observé des tendances en fonction des années. Nous avons toutefois supprimé toutes les lignes concernant les confinements liés au Covid19 (du 21 Mars 2020 au 21 Février 2021). Cet évènement étant un évènement exceptionnel, il

					viendrait fausser notre modèle de machine learning
TimeOfCall	00:04:25	Heure précise de l'appel	0	Object - /	Variable supprimée car trop précise pour l'usage que nous souhaitons en faire. Nous avons plutôt porté notre choix sur la variable 'HourOfCall'.
HourOfCall	0	Heure de l'appel	0	Int - de 0 à 23	Variable sélectionnée car l'heure d'intervention a un incident sur la variable cible (explication dans la DataViz)
DateAndTimeMobilised	01/01/2018 00:04:25	Date et heure d'ordre de mission(GMT)	0	Object - /	Variable supprimée car cette variable se retrouve indirectement dans la variable cible : délai de préparation des pompiers (=DateAndTimeMobilised - DateAndTimeMobilie) est additionnée au temps de trajet
DateAndTimeMobile	01/01/2018 00:05:38	Date et heure de mobilisation (GMT)	1,14 %	Object - /	Variable supprimée pour les mêmes raisons que 'DateAndTimeMobilised'
DateAndTimeArrived	01/01/2018 00:10:13	Date et heure d'arrivée sur les lieux (GMT)	0	Object - /	Variable supprimée car cette valeur se retrouve dans la valeur du temps de trajet

					(DateAndTimeArrived – DateAndTimeMobil e) qui est incluse dans la valeur de la variable cible
DateAndTimeLeft	01/01/2018 00:16:38	Date et heure de départ du lieu de l'incident (GMT)	1,86 %	Object - /	Variable supprimée car nous ne prenons en compte que les temps de réponse de la LFB et non la durée d'intervention, laquelle est indicative car elle correspond à l'heure de maîtrise de l'incident et non à sa résolution finale. Ne nous apporte pas d'informations sur notre variable cible.
DateAndTimeReturned	2018-01-01 00:22:45	Date et heure de retour à la station (GMT)	61,02 %	Object - /	Variable supprimée pour les mêmes raisons que 'DateAndTimeLeft'

Variables descriptives :

Colonnes :	Exemple :	Description :	Valeurs manquantes et gestion	Format et Distribution :	Justification de la sélection ou non :
IncidentGroup	False Alarm	Catégorie de l'incident	0,0003 %	Object - Special Service', 'Fire', 'False Alarm'	Variable supprimée car nous avons choisi de garder 'StopCodeDescription' qui est plus précis sur la catégorie de l'incident et donc sera plus pertinent pour notre modèle. De plus il n'y aucune valeur manquante dans 'StopCodeDescription'.
StopCodeDescription	AFA	Code détaillant la catégorie de l'incident	0	Object - cf tableau pour la distribution et la signification	Variable Sélectionnée car nous renseigne sur la catégorie des incidents de manière précise. Nous avons vu lors de la DataViz que le type d'incident pouvait impacter le temps de réponse de la LFB.
SpecialServiceType	RTC	Détails pour les unités spéciales, si nécessaire	67,46 %	Object - /	Variable supprimée car nous avons décidé de garder 'StopCodeDescription' qui détaille déjà les catégories d'incidents. Cette catégorie détaille plus spécifiquement la catégorie de service spécial (qui correspond au

					sauvetage ne relavant pas d'incendie) De plus, il y a énormément de données manquantes.
PropertyCategory	Non Residential	Type de propriété	0,0003 % - valeurs manquantes supprimées car ne représentent que 22 lignes sur l'ensemble de nos données.	Object - Road Vehicle', 'Outdoor', 'Dwelling', 'Outdoor Structure', 'Other Residential', 'Non Residential', 'Aircraft', 'Rail Vehicle', 'Boat'	Variable Sélectionnée car nous renseigne sur le type de « propriété » ou de bien et nous avons remarqué grâce à la Dataviz et aux tests statistiques que cette variable pouvait avoir un impact sur l'ampleur de l'incident et donc un impact possible sur le temps de préparation des équipes donc sur la variable cible.
.PropertyType	Mosque	Détails sur le type de propriété	0,0003 %	Object - /	Variable supprimée car détaille plus précisément le type de propriété avec énormément de valeurs uniques. Il s'agit d'une description individuelle et subjective
FirstPumpArriving _AttendanceTime	348	Temps total arrivée de la première équipe (en secondes)	7,64 %	Float - /	Variable cible - C'est ce que l'on cherche à prédire. (Détails au sein du rapport) - Suppression des lignes ne concernant pas le premier camion

					déployé (une pompe ne signifie pas un camion, c'est l'ordre de mission mais plusieurs camions peuvent être mobilisés au moment de la mobilisation initiale en fonction de l'ampleur de l'incident)
SecondPumpArriving_AttendanceTime	250.	Temps total arrivée de la deuxième équipe (en secondes)	64,22 %	Float - /	Variable supprimée car nous souhaitons nous concentrer sur le temps de réponse de la première équipe seulement.
PumpMinutesRounded	1	Temps passé sur l'incident en minutes (arrondi au plus proche)	0	Int - /	Variable supprimée car nous souhaitons connaître les temps de réponse de la LFB (c'est à dire le temps entre un appel et l'arrivée sur les lieux de l'incident). Le temps passé sur l'incident ne nous intéresse pas dans notre cas.
TurnoutTimeSeconds	73	Durée entre l'ordre de mission et le départ, en secondes	1,14 %	Float - /	Variable supprimée car créée une redondance avec notre variable cible.
TravelTimeSeconds	275	Durée de trajet, en secondes	1,15 %	Float - /	Variable supprimée car créer une redondance avec notre variable cible.

AttendanceTimeSeconds	348	Temps d'assistance (durée totale entre l'appel et l'arrivée sur les lieux, en secondes)	0	Int - /	Variable supprimée car créer une redondance avec notre variable cible.
PumpOrder	1	Ordre d'intervention (ex : première pompe à être déployée)	0	Int - /	Variable supprimée car ici on s'intéresse seulement à la première unité qui arrive sur les lieux donc tout notre jeu de donné ne portera que sur les premières unités.
DelayCodeId	9	Identifiant d'un code retard	75,06 %	Float - 3, 5, 6, 7, 8, 9, 10, 11, 12, 13	Variable supprimée après discussion avec notre mentor de projet car ces codes retards ne sont renseignés qu'après l'intervention et donc on ne peut pas s'en servir pour prédire un temps de réponse.
DelayCode_Description	Traffic, roadworks, etc	Description du code retard	75,06 %	Object - en entraînement lors de la mobilisation, Adresse incomplète ou erronée, défaut de l'appareil ou de l'équipement, arrivé mais retenu (autre raison), mesure de ralentissement de la circulation, Embouteillages et travaux..., conditions météorologiques, problèmes de	Idem 'DelayCodeId'.

communications/radio durant la mobilisation, pas retardé, en mission extérieure lors de la mobilisation

Variables de localisation :

Colonnes :	Exemple :	Description :	Valeurs manquantes et gestion	Format et Distribution :	Justification de la sélection ou non :
AddressQualifier	Within same building	Qualification de la localisation de l'incident	0,00005 %	Object - /	Variable supprimée car donnée administrative venant préciser le lieu d'intervention après l'incident (donnée qui ne nous sert pas pour prédire étant donné que cette information n'est disponible qu'une fois l'intervention terminée).
Postcode_full	N2 8AY	Code postal complet	50,03 %	Object - /	Variable supprimée car 50% de valeurs manquantes et manque de précision sur la localisation de l'incident (plusieurs colonnes correspondent à la localisation)
Postcode_district	N2	District postal	0	Object - /	Idem que postcode_full
UPRN	200220110	Numéro unique d'identifiant de la propriété	7,93 %	Int - /	Cette variable aurait pu être utile mais étant donné que certains incidents ne se situe pas au sein de propriété (ex : accident de la route, bateaux etc...) nous ne pouvons pas utiliser cette variable. De

					plus, il s'agit de milliers (ici même millions) de code unique ce qui viendrait perturber notre modèle de machine learning car difficile à généraliser. D'autre variable sur les emplacements des incidents ont été sélectionnées.
USRN	20013420	Numéro unique d'identifiant de la rue	9,14 %	Int - /	Idem que UPRN
IncGeo_BoroughCode	E09000003	Code administratif de l'arrondissement	0	Object - /	Variable supprimée car manque de précision sur le lieu de l'incident. D'autres variables indiquent l'emplacement de l'incident. Cette variable se retrouve par la localisation GPS avec les variables Easting_rounded et Northing_rounded
IncGeo_BoroughName	BARNET	Nom de l'arrondissement	0	Object - /	Idem IncGeo_BoroughCode
ProperCase	Barnet	Nom de l'arrondissement formaté	0	Object - /	Idem IncGeo_BoroughCode
IncGeo_WardCode	E05000049	Code administratif du quartier	0,03 %	Object - /	Variable supprimée pour les mêmes raisons que

					'IncGeo_BoroughCode'
IncGeo_WardName	EAST FINCHLEY	Nom du quartier actualisé	0,03 %	Object - /	Variable supprimée pour les mêmes raisons que 'IncGeo_BoroughCode'
IncGeo_WardNameNew	EAST FINCHLEY	Coordonnées exactes, système de projection cartographique nationale	0,03 %	Object - /	Variable supprimée pour les mêmes raisons que 'IncGeo_BoroughCode'
Easting_m	527184	Coordonnées exactes, système de projection cartographique nationale	50,03 %	Float - /	Variable supprimée car plus de 50% de données manquantes. Ces données manquantes étant des indications géographiques sur les incidents, nous ne pouvons pas les remplacer. Il était donc compliqué de garder cette variable.
Northing_m	189488	Coordonnées exactes	50,03 %		Idem que pour 'Easting_m'
Easting_rounded	527150	Coordonnées exactes arrondies à 50	0	Int - /	Variable sélectionnée car nous donne les informations sur le lieu de l'incident. De plus, ces variables (avec Northing_rounded) sont convertibles en latitude et longitude pour en extraire des lieux plus précis et ainsi calculer des distances entre les

					caserne et l'incident ou même définir quelle caserne est la plus proche
Northing_rounded	189450	Coordonnées exactes arrondies à 50	0		Idem 'Easting_rounded'.
Latitude	51.5899	Latitude	50,03 %	Float - /	Variable supprimée car trop de données manquantes pour être exploitable. Nous avons décidé de garder 'Easting_rounded' et 'Northing_rounded' à la place.
Longitude	-0.165453	Longitude	50,03 %		Idem latitude.
FRS	London	Région couverte par le service	0	Object - /	Variable supprimée car notre étude ne concerne que la région de Londres. Variable à valeur unique
IncidentStationGround	Finchley	LFB caserne (caserne locale responsable de la zone)	0	Object - /	Variable supprimée car nous avons vu que dans 38% des cas, ce n'est pas la caserne responsable de l'incident qui se déplace (ce sont les casernes de renfort). La caserne déployée (en premier, pour le premier camion) nous semblait plus appropriée

FirstPumpArriving _DeployedFromSt ation	Finchley	Station de déploiement de la première équipe	7,64 %	Object - /	Variable supprimée car redondance avec 'DeployedFromStati on_Name' (variable sélectionnée). Variable supprimée pour éviter le risque de multicolinéarité.
SecondPumpArriv ing_DeployedFrom Station	Clapham	Station de déploiement de la deuxième équipe	64,22 %	Object - /	Idem que pour 'SecondPumpArrivi ng_AttendanceTim e'.
DeployedFromStat ion_Name	Finchley	Nom de la station déployée	0,001 %	Object - /	Variable sélectionnée car nous renseigne sur la caserne déployée. Cette variable nous permettra de réaliser des calculs de distance et ainsi nous donnera des informations pour prédir les temps de réponses.
DeployedFromStat ion_Code	A39	Code de la station déployée	0,001 %	Object - /	Variable supprimée car nous avons choisi de garder 'DeployedFromStati on_Name' qui nous renseigne directement sur le nom de la station déployée (rend plus simple la lecture et l'interprétation).
DeployedFromLoc ation	Home Station	Lieu de déploiement	0,05 %	Object - /	Variable supprimée car nous avons vu

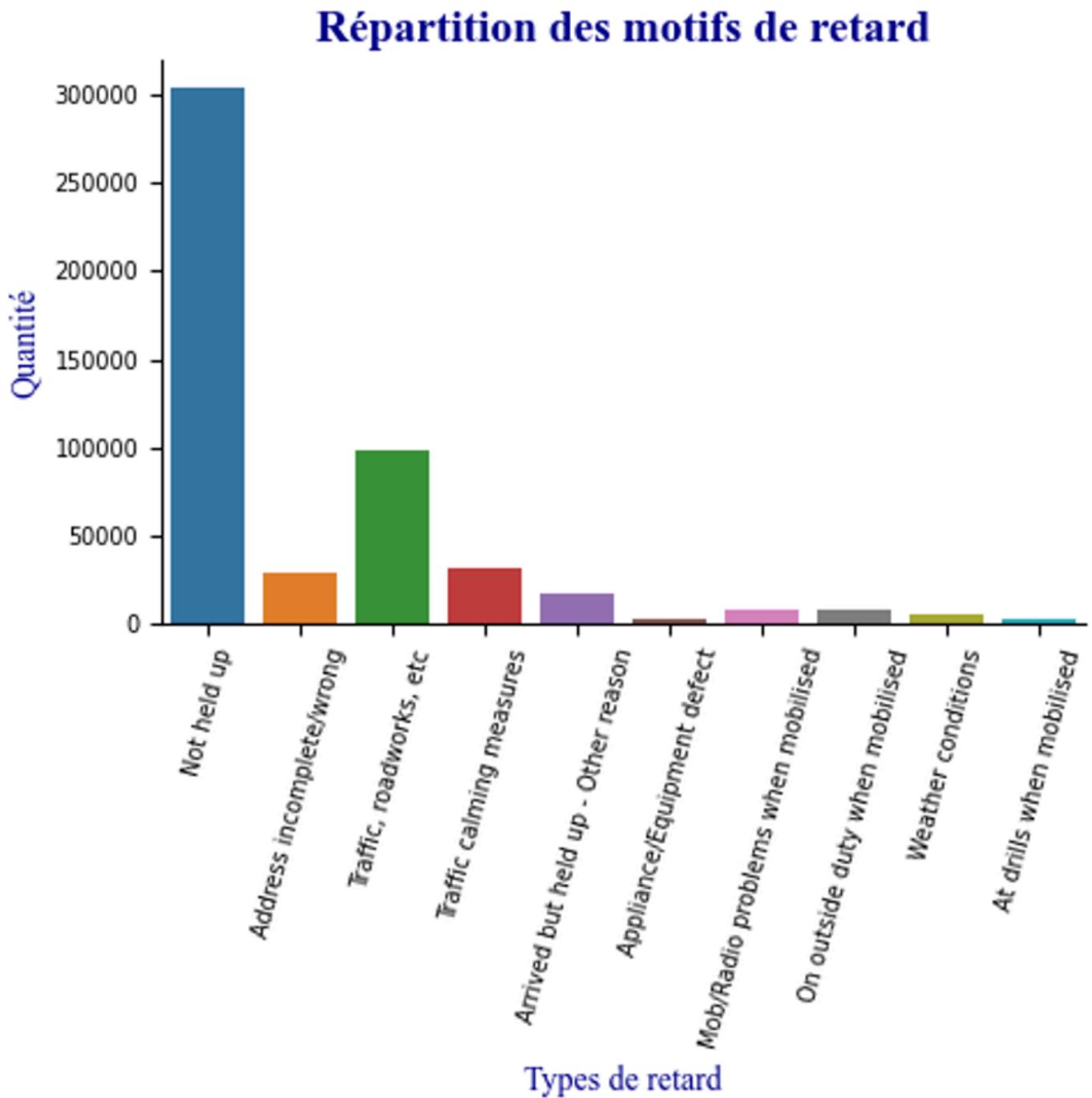
		(caserne, autre lieu d'incident, ...)			lors de notre DataViz que 38% des équipes ne partaient pas de la caserne mais cela correspond aux camions de renfort et non au premier camion. Utiliser cette variable n'aurait donc pas d'intérêt dans notre modèle de machine learning.
BoroughName	EALING	Nom de l'arrondissement	71,61 %	Object - /	Variable supprimée pour les mêmes raisons que 'IncGeo_BoroughCode'
WardName	MONKHAMS	Nom du quartier	71,64 %	Object - /	Idem 'BoroughName'.

*Le service spécial est le nom donné à l'ensemble des urgences auxquelles la LFB répond et qui ne sont pas des incendies ou des fausses alarmes. Il s'agit, par exemple, d'accidents de la route, de sauvetage et d'incidents liés à l'eau et les incidents impliquant des produits chimiques dangereux. En moyenne, un incident de service spécial prend environ une demi-heure. Ces dernières années, les incidents de service spécial les plus longs concernaient des incidents liés à des matières dangereuses, avec une durée d'intervention de deux heures.

Grâce à nos recherches et notre analyse des informations, nous avons réussi à dégager 7 variables explicatives : ‘HourOfCall’, ‘DeployedFromStation_Name’, ‘CalYear’, ‘Easting_rounded’, ‘Northing_rounded’, ‘StopCodeDescription’, ‘PropertyCategory’. Nous avons fait le choix de garder ‘DeployedFromStation_Name’ plutôt que ‘IncidentStationGround’ car dans 38% des cas, ce n'est pas la station de référence de l'incident qui est déployée.

Nous avons également décidé de supprimer “DelayCode_Description” car Londres a organisé de voies de bus facilement accessibles aux véhicules de secours en intervention. Elle dispose de feux de circulation intelligents sur certains carrefours qui peuvent donc donner la priorité aux véhicules de secours en temps réel. Les véhicules d'urgence bénéficient de règles de priorité leur permettant de franchir des feux rouges ou emprunter des voies non conventionnelles pour arriver sur les lieux d'incident le plus rapidement possible. On parle de “Traffic Management System” qui correspond à un ensemble de technologies et stratégies permettant d'optimiser le trafic routier. Grâce à ce système qui collecte les informations en temps réel et au GPS connecté des véhicules de secours, la LFB a les moyens de contourner directement les embouteillages et trouver les chemins les plus rapides. Le trafic est également régulé automatiquement et immédiatement avec l'adaptation des durées de feux

de circulation en fonction du trafic. De plus, la cause "non retardé" représente plus de 80% de nos données et ces motifs ne peuvent pas être connus à l'avance.



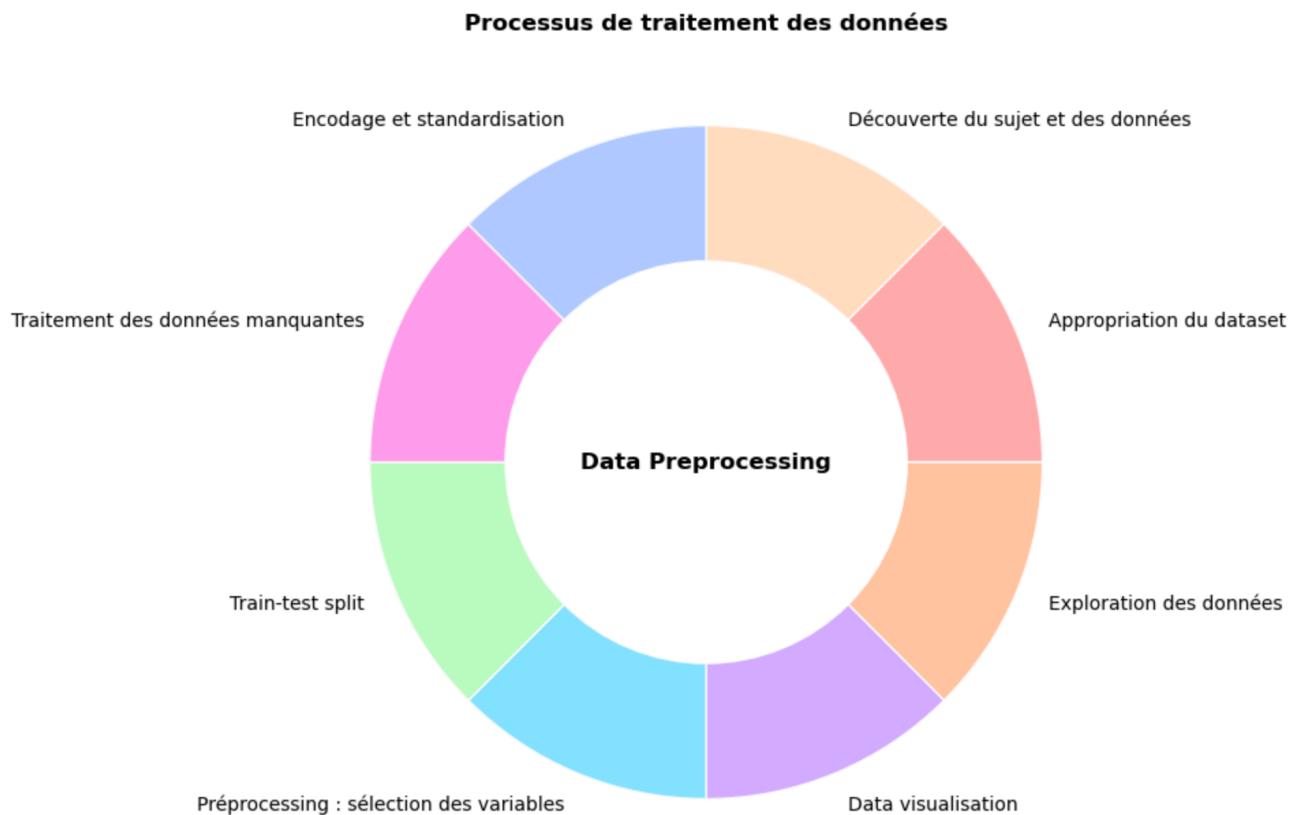
Si on fait abstraction des données manquantes (la datavisualisation a précédé le remplacement des valeurs manquantes), on observe que dans la plupart des cas il n'y a pas de retard. Les 3 premières causes de retard sont : les embouteillages ou travaux, une adresse erronée ou incomplète et des mesures de ralentissement de la circulation. Cet histogramme a conforté notre choix de transformer nos données manquantes par le mode.

Une fois notre sélection de variables réalisée, nous avons enrichi notre jeu de données avec deux variables représentant les conditions météorologiques du jour ainsi que la visibilité au kilomètre. Nous avons trouvé ces données en réalisant une concaténation des 16 fichiers csv du site "Historique météo" et en ne gardant que ces deux variables. Le merge de nos deux dataframes s'est fait sur la colonne commune 'DATE' (colonne 'DateAndTimeMoibilised' de notre fichier sur la LFB, renommée en 'DATE' pour la fusion). La variable concernant les données de visibilité a été modifiée pour en faire

une catégorisation en fonction de la distance de visibilité : de 0 à 2Km c'est une très mauvaise visibilité, de 2 à 10km c'est une visibilité moyenne et à plus de 10km c'est une très bonne visibilité. Nous avons donc créé une colonne supplémentaire avec ces catégories et supprimé l'ancienne colonne. La colonne "OPINION" comprenant les conditions météorologiques a été renommée en "Meteo" pour plus de lisibilité.

Nous avons fait le choix de supprimer toutes les lignes correspondantes aux différents confinements liés au Covid-19 afin de ne pas prendre en compte des données "exceptionnelles".

Nous verrons donc dans la prochaine partie quelles ont été nos méthodes de pré-processing.



Pré-processing et feature engineering :

Pré-processing :

Nous nous sommes concentrées, dans un premier temps, sur notre variable cible. Tout d'abord, nous avons supprimé toutes les valeurs manquantes, car un modèle incomplet peut donner lieu à des résultats biaisés ou inexacts. D'après nos données, cela représente 6 lignes seulement. Ayant un grand nombre d'outliers (plus de 100 000 lignes), nous les avons d'abord analysés. Nous avons pu observer que certains étaient liés à des codes particuliers de retard, d'autres représentaient des temps "normaux", car non retardés, donc probablement un rapport de distance avec la caserne mobilisée.

Par exemple, le 14 Juin 2017 a eu lieu l'incendie de la Grenfell Tower. Incendie ayant englouti un immeuble de logement sociaux de 23 étages situé dans l'un des quartiers les plus riches de Londres. Cet incident majeur permet d'expliquer certains outliers comme les 369 appels passés au 999 (centre de régulation). Un exemple plus récent est l'incendie de la Tower Hamlets le 7 Mai 2021, qui a mobilisé 13 pompes à la mobilisation initiale (ce qui signifie que 13 équipes ont été envoyées dès le départ), 46 camions de pompiers sur toute l'intervention de 14 casernes différentes. Dans la perspective de modélisation et de prédiction des temps de réponse, nous avons observé la distribution de ces valeurs extrêmes. Toutes les lignes supérieures à 75 % des outliers soit 780 secondes (13 minutes) ont été supprimées (22 000 lignes soit moins de 1 % de nos données). Nous n'avons pas supprimé tous les outliers car certains sont de vraies valeurs extrêmes et non des valeurs aberrantes, elles sont donc importantes pour notre modélisation.

Concernant les variables explicatives, la variable "DeployedFromStation_Name", contenait des valeurs manquantes avec 8 lignes ayant un NaN. Nous avons donc fait le choix de les supprimer, car cela n'impacterait pas notre jeu de données.

Les variables "Meteo" et "Visibility" contenaient elles aussi des valeurs manquantes (238). Après analyse de nos données, nous avons décidé de remplacer ces valeurs manquantes par leur mode. Pour éviter une fuite de données nous avons décidé d'effectuer ces modifications après la séparation des données.

Traitement et transformation des données :

Maintenant que nous avons un jeu de données nettoyé, nous sommes passées à l'étape de transformation de notre DataFrame. Les variables explicatives ont été séparées de la variable cible.

Nous avons divisé ensuite le jeu de données en set-train (80%) et set-test (20%) grâce à la fonction “train_test_split” de Scikit-Learn et nous avons fixé un random_state à 42. Après cette séparation, nous avons donc traité nos valeurs manquantes des variables “Meteo” et “Visibility” en les remplaçant par leur mode grâce au SimpleImputer de Sklearn, afin d'éviter une fuite des données.

Nous avons encodé nos variables catégorielles à l'aide de 3 méthodes : la *FrequencyEncoding*, le *OneHotEncoding* et le *OrdinalEncoding*.

Concernant les variables “DeployedFromStation_Name”, “Easting_rounded” et “Northing_rounded”, nous sommes parties sur le *FrenquencyEncoding*. En effet, devant le grand nombre de valeurs sur ces colonnes (une centaine pour la station déployée et plus de 500 pour les coordonnées GPS), un encodage par fréquence nous semblait plus pertinent car ceci évite la multiplication des colonnes qu'on aurait avec un OneHotEncoding et elle garde l'importance des valeurs les plus présente. Bien que les variables “Easting_rounded” et “Northing_rounded” soit des variables numériques discrètes, nous avons choisi de les traiter comme des variables catégorielles car elles représentent des données GPS et donc des lieux d'incident. Ce traitement nous semblait plus logique qu'une standardisation.

Pour les variables “PropertyCategory” et “StopCodeDescription” nous avons opté pour le OneHotEncoding car elles comportent toutes entre 10 et 16 valeurs seulement.

Concernant les variables “Meteo” et “Visibility” nous avons choisi de réaliser un OrdinalEncoding afin de préserver la notion d'ordre dans les valeurs.

Les variables “HourOfCall” et “CalYear” ont été standardisées car elles sont des variables numériques. Un autre type de transformation plus adéquate aurait pu être effectué si les variables cycliques représentaient une grande partie de nos variables. Ne représentant que 20% de nos variables sélectionnées, nous avons fait le choix de simplement utiliser une standardisation.

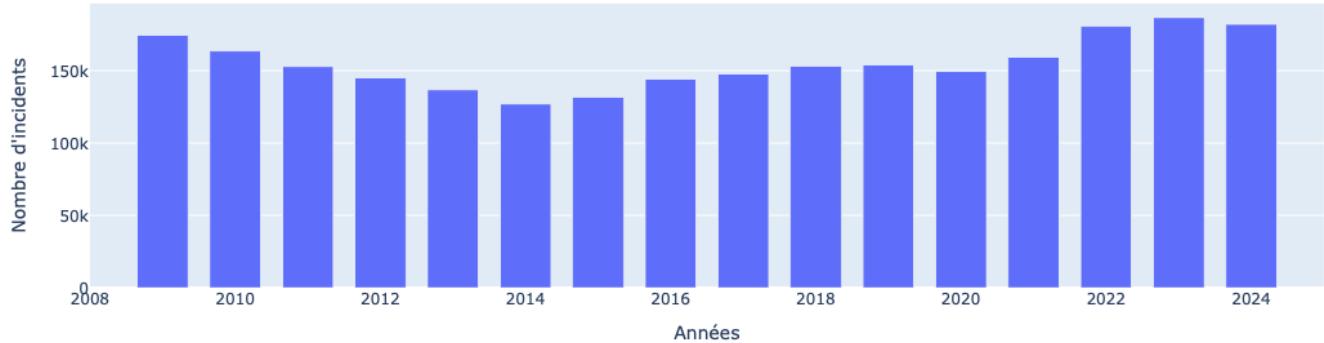
Une dernière éventuelle transformation reste à réfléchir : la réduction de dimension. En effet, nous disposons de tellement d'informations que nous ne savons pas si la modélisation sera efficace. Nous prendrons cette décision après concertation avec notre mentor de projet.



Dataviz et statistiques :

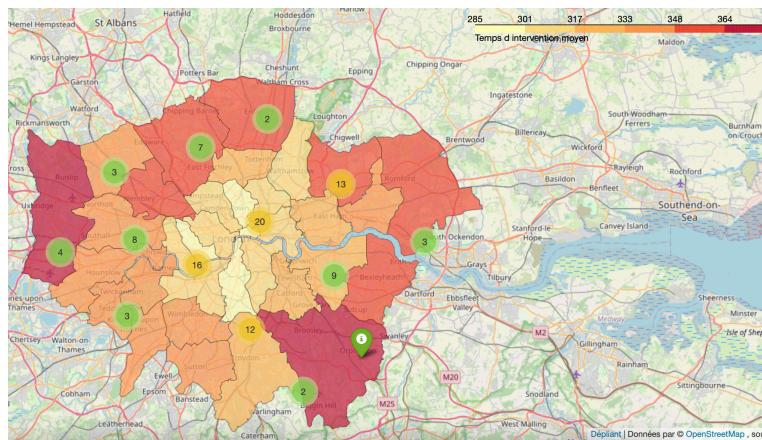
Voyons maintenant les différentes visualisation obtenues lors de l'analyse du sujet et les statistiques qui en découlent.

Distribution du nombre d'incidents par années

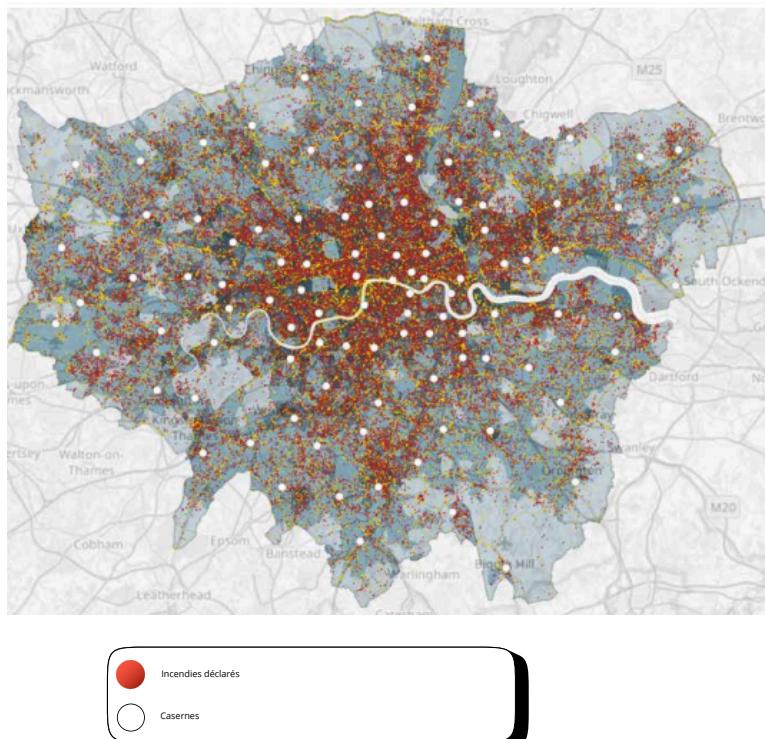


Pour commencer, voici une analyse du nombre d'incidents par année depuis 2009 . On constate une grande diminution des incidents entre 2009 et 2015 ceci peut avoir diverses raisons comme une promotion de la sécurité incendie à grande échelle, une amélioration des conditions de sécurité des infrastructures etc... Toutefois, on observe une augmentation du nombres d'incidents, surtout depuis les années 2020, ceci vient renforcer la nécessité de comprendre les temps de réponses de la LFB afin de les réduire.

Heatmap des temps de trajet de la LFB de Londres

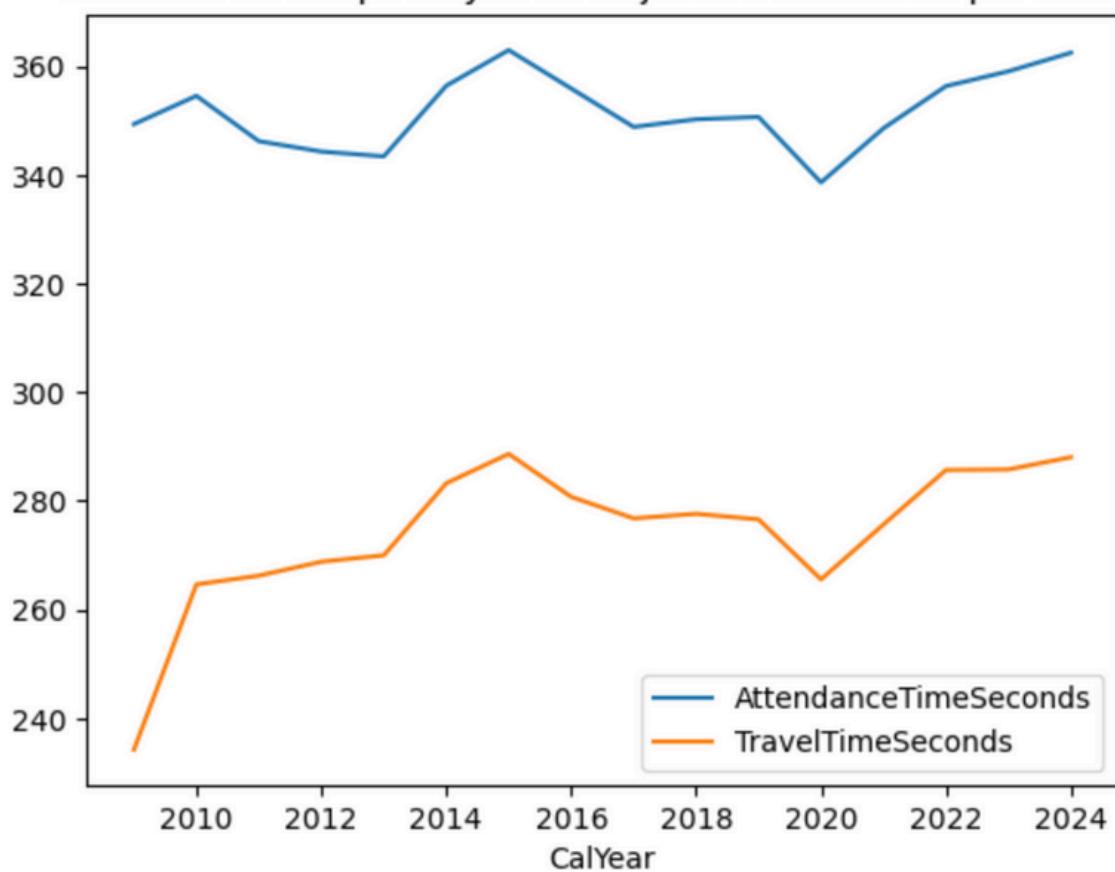


Répartition incendies et brigades de Londres (2016-2020)



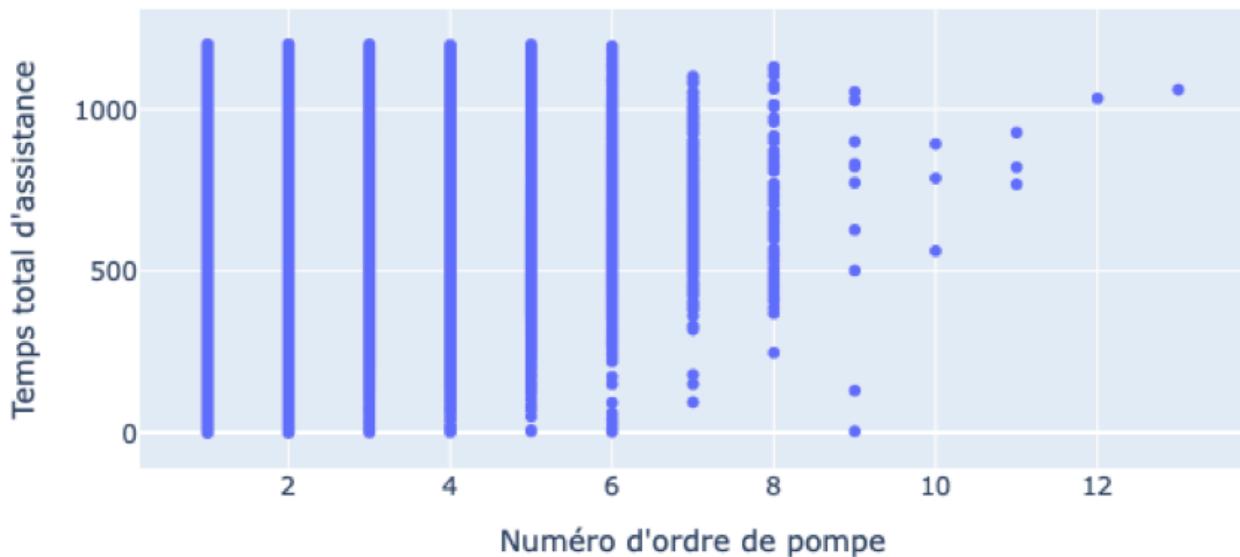
Ces deux cartes ici représentent les durées moyennes de trajet de la LFB (en haut) et la répartition des incidents et des brigades sur le territoire de Londres (en bas). On voit très bien ici que la plupart des incidents et des brigades se concentrent au centre de Londres et que cette zone correspond également aux temps les plus courts.

Evolution du temps moyen de trajet et d'assistance par année

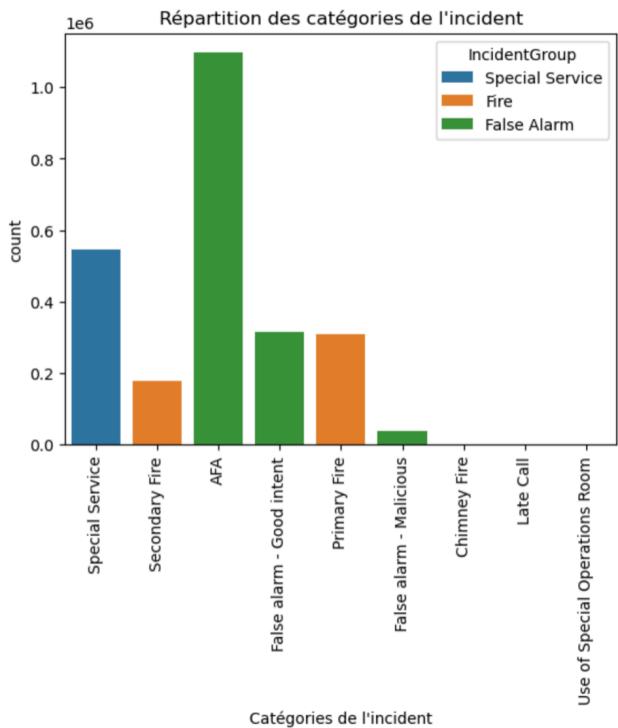


On peut observer grâce à notre jeu de données, une évolution croissante du temps de trajet et d'assistance au fil des ans. Cette évolution croissante peut notamment s'expliquer par l'accroissement de la population, mais aussi l'essor du tourisme (plus de monde = plus de trafic). On pourrait également corrélérer cela aux probables travaux d'aménagement de la ville de Londres. Avec cette représentation graphique, on observe bien tous les enjeux de la LFB concernant la réduction du temps de réponse. On observe bien une évolution similaire entre ces deux données, ceci paraît logique étant donné que, comme vu précédemment, les temps de préparation sont des temps très courts et que le temps de trajet est la partie la plus importante dans le temps d'assistance totale. On observe également une diminution du temps de trajet et d'assistance en 2020 que l'on peut associer au COVID-19 et au confinement.

Temps total d'assistance par ordre de pompe



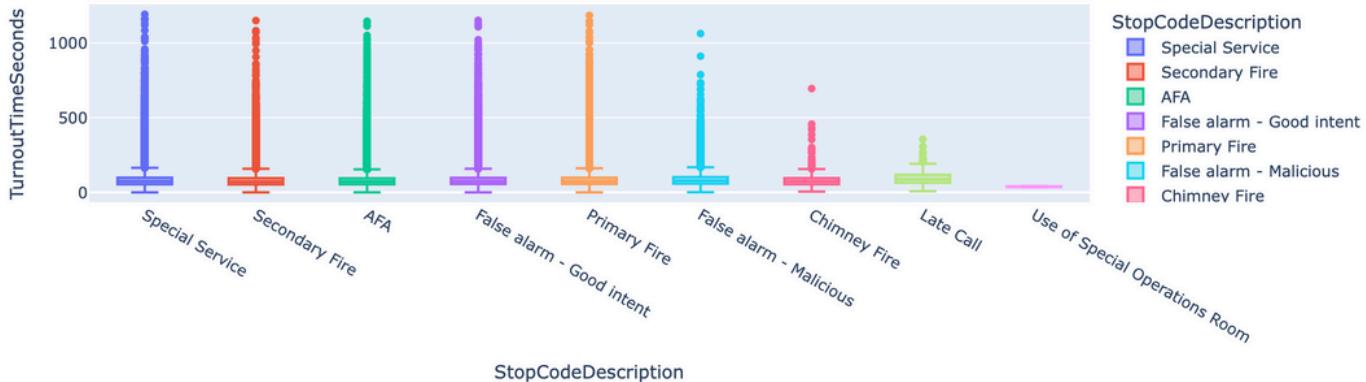
Notre hypothèse avant ce graphique était que plus une pompe était mobilisée en dernière intention, plus le temps d'assistance était élevé. Avec ce graphique on observe que c'est effectivement le cas mais à partir de la 6ème pompe. On peut alors énoncer plusieurs hypothèses : le temps de trajet est plus long car l'unité sait son ordre de mobilisation et sait donc que l'incident est déjà en train d'être géré par d'autres unités (la rapidité sur le trajet peut également engendrer des risques), plus une unité est mobilisée tardivement plus il y a de chance qu'elle soit déjà mobilisée sur un autre incident et donc mette plus de temps à se rendre sur le lieu de l'incident. On observe très peu de données entre la 9ème et 13ème pompe car en effet, peu d'incident nécessite autant de mobilisation.



StopCodeDescription	Description
Special Service	Incidents ne concernant pas directement un incendie (sauvetage, catastrophes naturelles, assistance technique)
Secondary Fire	Incendies mineurs
AFA	Alarme incendie automatique
Primary Fire	Incendies majeurs avec risque vie humaine
False alarm - Good Intent	Fausse alerte avec bonne intention
False alarm - malicious	Fausse alerte intentionnelle sans raison valable (objectif de perturber)
Chimney Fire	Feu de cheminée
Flood Call Attended - batch mobilised	Inondations où plusieurs équipes ou unités sont mobilisées pour répondre à des incidents multiples
Late call	Appel reçu en fin d'incident, nécessitant une vérification
Use of special Opérations Room	Événements complexes ou de grande envergure nécessitant la coordination d'une salle d'opération spéciale
Standby	En attente pendant qu'une autre intervention est en cours.

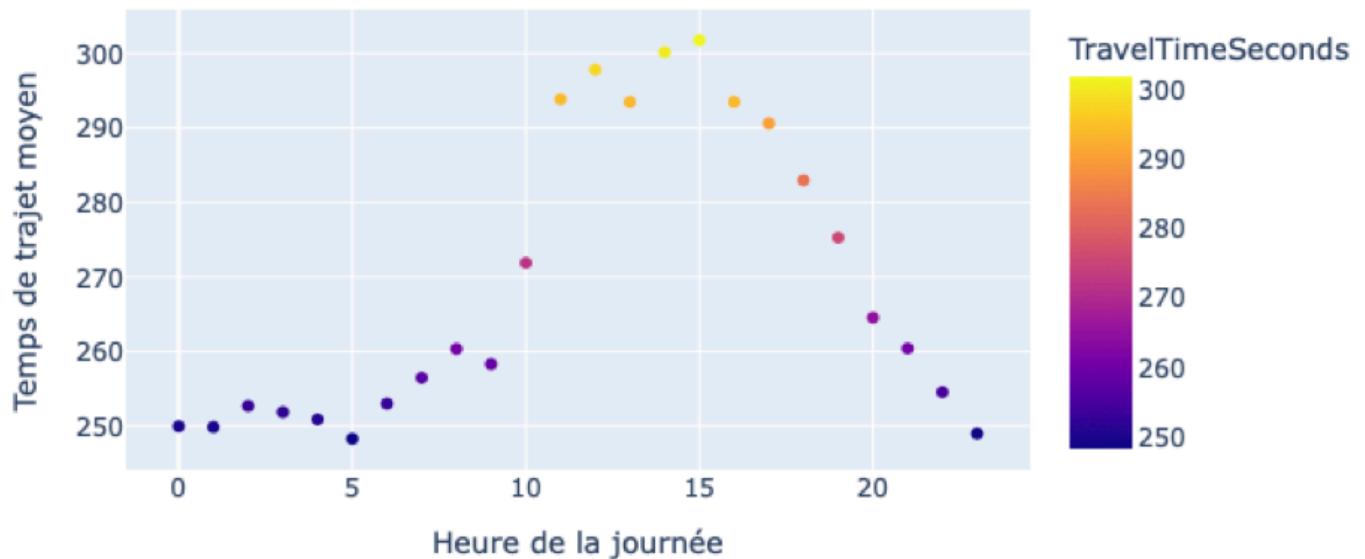
La plupart des incidents référencés concerne les alarmes-incendie automatiques. On remarque en seconde position, les services spéciaux (tout ce qui ne concerne pas directement un incendie comme les sauvetages humains, les incidents techniques...), puis en troisième position nous avons les fausses alarmes avec une bonne intention. On remarquera que les incendies majeurs avec risque pour la vie humaine se trouvent en 4ème position et les incendies mineurs en 5ème position.

Distribution des temps de préparation en fonction des incidents

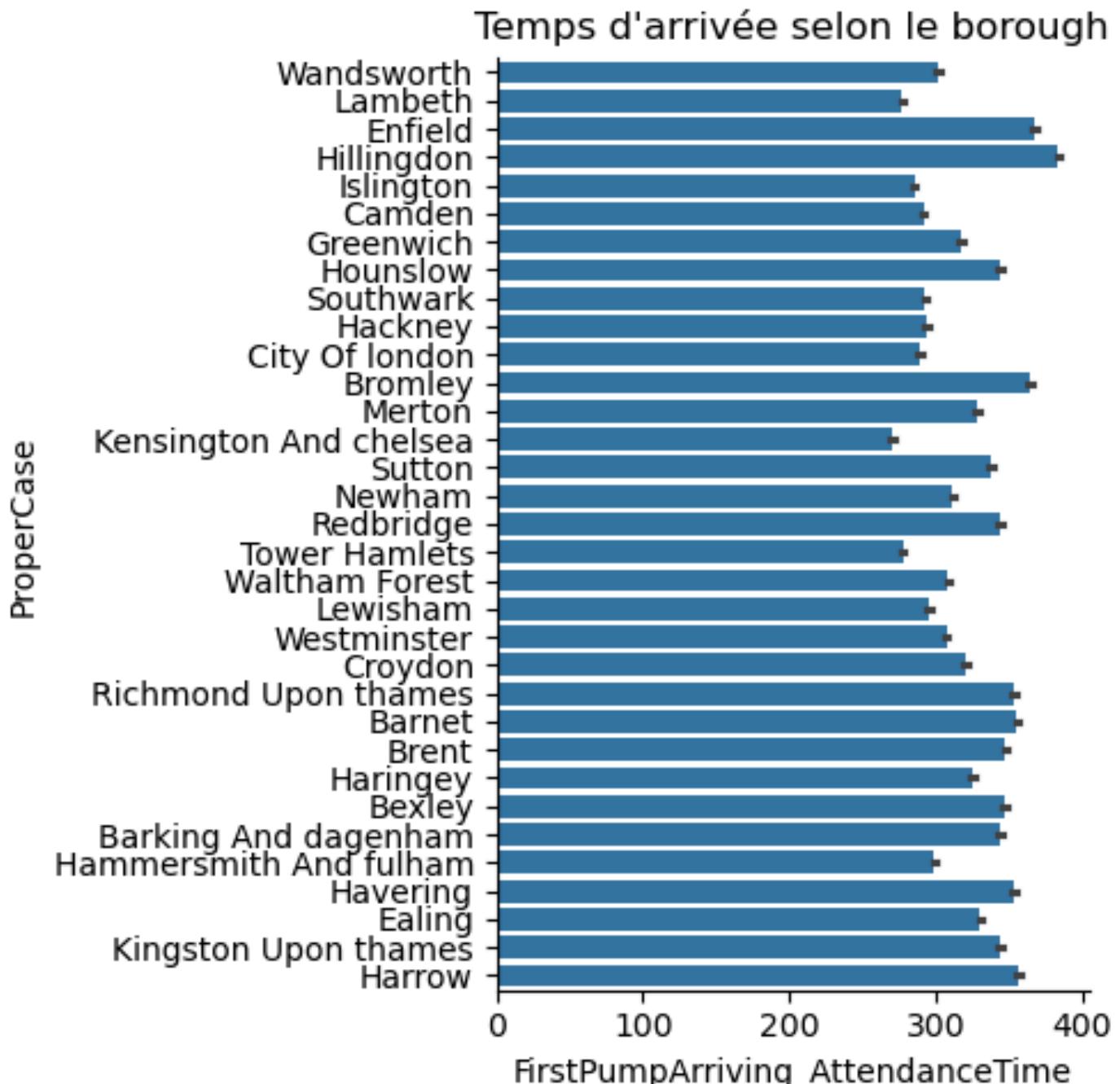


Avec ce graphique, nous étudions la relation éventuelle entre la catégorie de l'incident et le temps de préparation du camion. Ici on constate que la catégorie de l'incident peut ne pas avoir d'impact : la médiane ne varie quasiment pas entre les différentes catégories sauf pour les « special service » ou la médiane est plus basse 22 secondes ce qui n'est pas significatif pour conclure à une relation entre les deux variables. Toutefois, trois catégories se différencient : la « use of special operations room » qui correspond aux évènements complexes ou de grande envergure (catastrophe naturelle, attentats, incendie de grande envergure, évènements importants programmés...), les « late call » (vérification de fin d'incident) et les feux de cheminée. Les catégories 'late call' et 'chimney fire' sont les seules catégories avec très peu d'outliers et avec des données beaucoup moins dispersées. Les temps de préparation sont généralement plus courts. On pourrait analyser cela en partant du principe qu'une vérification de fin d'incident nécessite moins de préparation qu'une mobilisation pour un incendie par exemple. De même que pour les feux de cheminée.

Temps de trajet moyen en fonction des heures de la journée

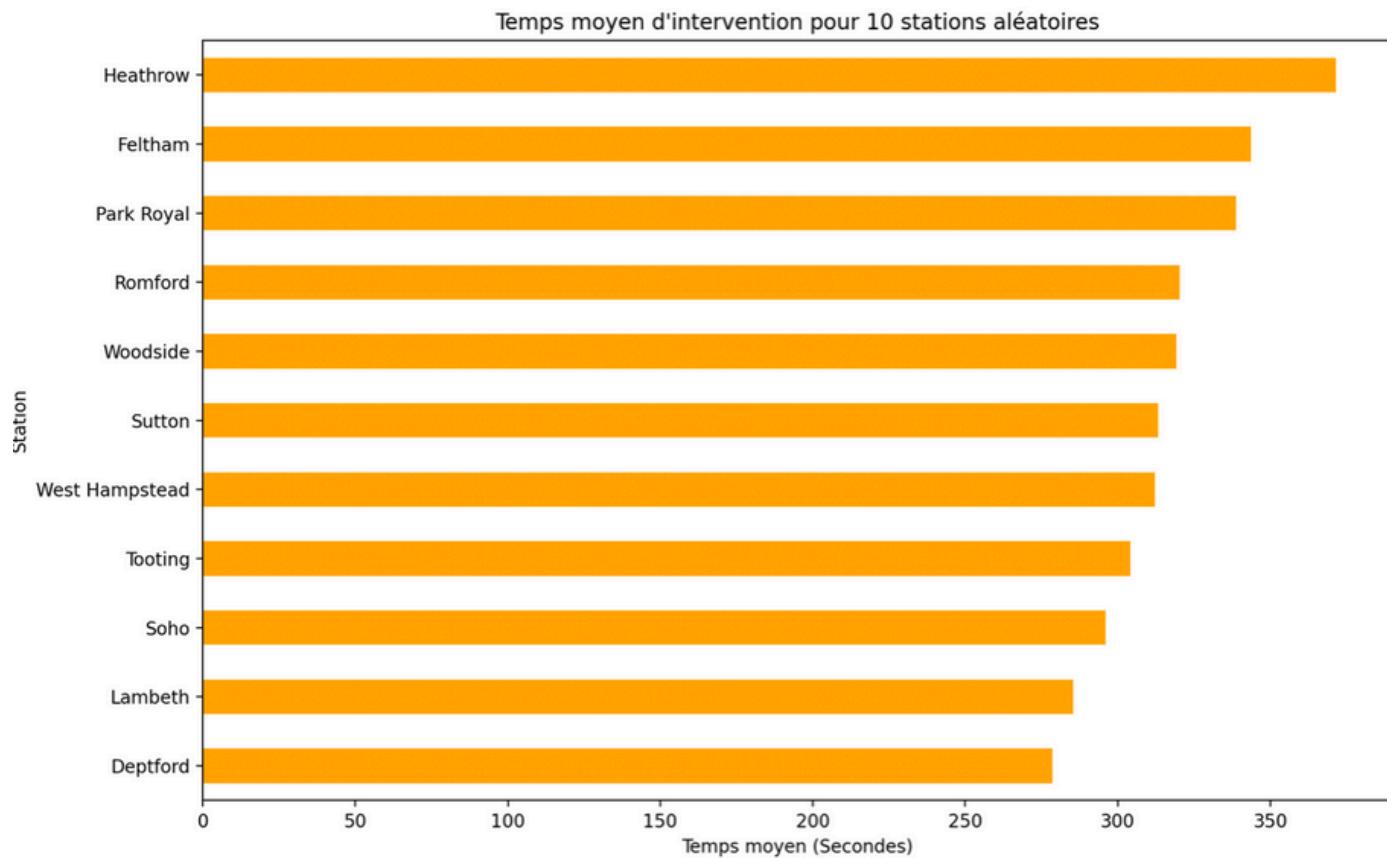


En analysant le temps de trajet moyen en fonction des heures de la journée on observe que la période 10h-19h est la période où les temps de trajet sont en moyenne plus long. Ceci s'explique par la présence humaine plus importante à ce moment de la journée. Bien que les différences ne paraissent pas significatives (moins d'une minute), dans le domaine du secours à la personne et aux biens, une minute peut faire une grande différence.

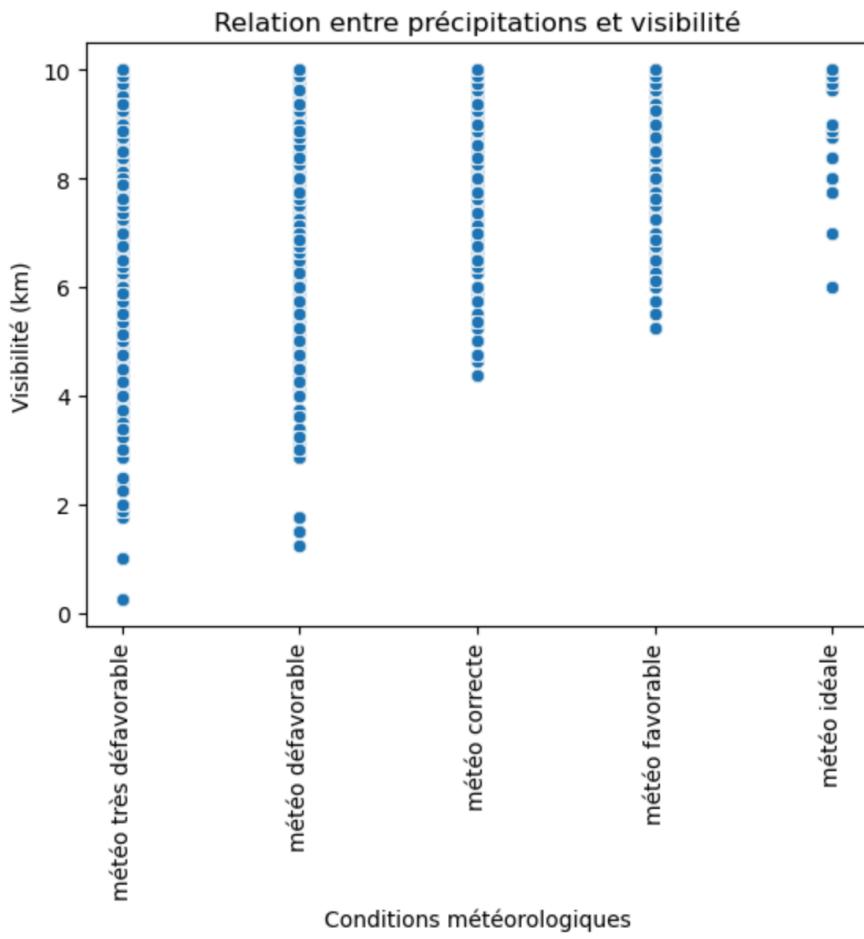


Ici nous pouvons observer une différence de temps d'assistance en fonction de l'arrondissement d'intervention. C'est pourquoi il nous semblait intéressant de garder les variables de coordonnées GPS dans notre future modélisation. De plus, elle se complète avec les informations sur la caserne ayant été déployée.

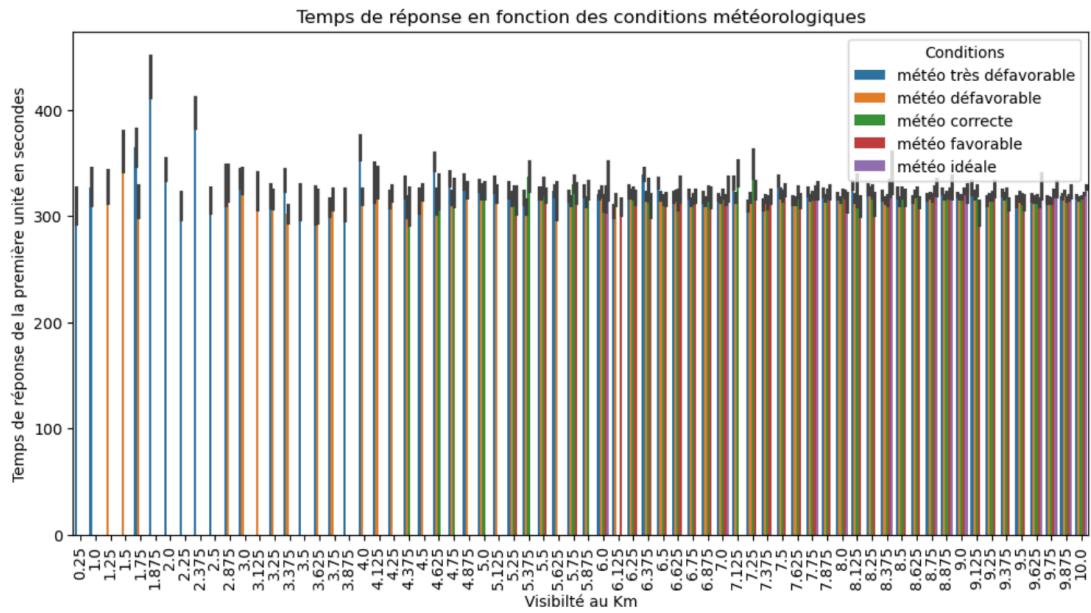
Étude de la moyenne du temps d'intervention sur un échantillon de 10% des casernes (sélection pondérée par la fréquence d'apparition)



Nous avions émis une hypothèse initiale selon laquelle l'emplacement des casernes pouvaient impacter le temps d'intervention de la première unité. Pour vérifier cette hypothèse nous avons choisi de représenter la moyenne du temps d'intervention par caserne. Ici nous avons pris échantillon de 10% d'entre elles (valeurs uniques) pondérées par leur fréquence d'apparition. On observe une différence qui peut osciller entre 270 et 370 secondes. Ceci permet de renforcer l'hypothèse selon laquelle l'emplacement des casernes impactent la variable cible.



Grâce à ce graphique, on observe bien la relation entre les conditions météorologiques et la visibilité au kilomètre. Plus la météo est favorables et plus les conditions de visibilité sont bonnes.



On observe avec ce graphique une corrélation entre les conditions météorologiques, la visibilité au kilomètre et le temps de réponse de la LFB. On remarque que plus la visibilité est mauvaise, plus les temps de trajet sont longs. On constate également que de 0 à 4 km environ de visibilité, les conditions météorologiques sont défavorables ou très défavorables.



Classification du problème

L'objectif de notre sujet étant de prédire les temps de réponse de la brigade des pompiers de Londres et notre variable cible étant une variable temporelle continue, nous sommes initialement parties sur une problématique de régression.

Modèles de Régression :

Nous nous sommes donc répartis plusieurs modélisations de Machine Learning afin de comparer nos résultats et évaluer la fiabilité de nos modèles. Au regard de nos résultats, nous nous sommes rapidement aperçus que nos modélisations ne pouvaient pas réaliser ces prédictions, ou du moins que nous n'en avions pas les capacités. Nous avons sélectionné les modèles en fonction de plusieurs critères, nous sommes partis de modèles plus classiques et simplistes jusqu'à des modèles plus complexes et plus optimaux pour notre cas de Machine Learning. Ces modèles sont :

- Linear Regression : modèle plus simple, supposant une relation linéaire entre la variable cible et les caractéristiques. Comme nous l'avions prédit, cela n'est pas notre cas.
- Decision Tree Regressor : modèle permettant de représenter des problématiques non linéaires.
- Random Forest Regressor : ce modèle peut être considéré comme une "optimisation" du Decision Tree Regressor car il représente une multitude de Decision Tree Regressor.
- SGD Regressor : modèle plus adapté à notre jeu de données de grande taille (initialement 1,5 GigaOctets de données)
- KNN Regressor : également adapté aux modèles de grande taille (non linéaire).

Afin de produire un travail de qualité et comparer les performances de nos modèles, nous avons fait le choix d'évaluer toutes les métriques de nos modèles : MAE (Mean Average Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) et R2 (score de détermination).

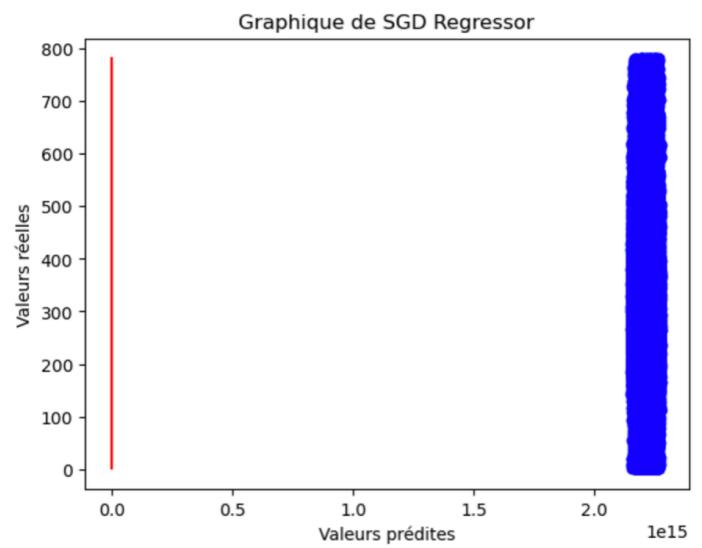
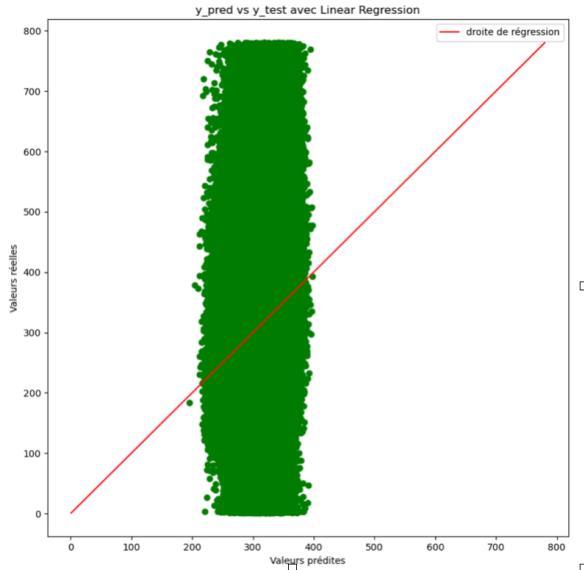
Voici nos résultats :

Modèles de Regression	MAE	MSE	RMSE	R2
Linear Regression	89.719590	14049.171200	118.529200	0.051449
Decision Tree Regressor	207.2518553	63919.102367	252.82227427	-3.319882904
Random Forest Regressor	88.149152936	13649.741853	116.832109685	0.0797185175
SGD Regressor	219791737230162	4.928111e+30	2219934929759535	3.32472727e+26
KNN Regressor	94.106070	15144.089260	123.061323	-0.022476

Nous pouvons constater ici que nos modèles de régression ne sont pas adaptés. Le score de détermination (R2) est autour de 0 pour chacun de nos modèles. Cela signifie qu'ils n'expliquent en rien la variance de la variable cible.

Par exemple, pour le Random Forest Regressor, la MAE de 88,1 signifie qu'il y a un écart de 88 secondes (soit environ 28% par rapport au temps moyen de 312,99 secondes). Au regard du contexte de notre projet, cet écart est beaucoup trop significatif pour être considéré comme pertinent.

Afin de représenter plus visuellement le fait que ces modèles ne parviennent pas à représenter la variance des données nous avons réalisé deux schémas : Linear Regressor et SGD Regressor :



Modèles de Classification binaire :

Nous avons donc tenté une approche différente en catégorisant notre problème de machine learning comme une classification binaire. En effet, grâce à nos recherches sur la LFB, nous savons que l'objectif du plan quinquennal en cours est de réaliser des interventions en moins de 6 minutes (temps entre l'appel et l'arrivée sur les lieux). Nous avons donc classé comme 0 les délais d'intervention supérieurs à 6 minutes (qui ne respectent pas l'objectif) et comme 1 les délais inférieurs à 6 minutes. Concrètement, nous avons réalisé une discréétisation de la variable cible.

Pour ce faire, nous avons entraîné plusieurs modèles du plus simple et classique au plus complexe :

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier
- KNeighbors Classifier
- Linear SVC

Toujours dans une optique de réaliser un rendu plus qualitatif, nous avons observé l'ensemble des métriques associées à ce type de modélisation : l'accuracy (exactitude des résultats), le recall (rappel), la précision, le F-1 Score et l'AUC. L'ensemble de ces métriques ont été obtenues grâce au "classification_report" du package "Scikit Learn".

Voici les résultats :

	ACCURACY	PRECISION		RECALL		F1-SCORE	
		CLASSE 0	CLASSE 1	CLASSE 0	CLASSE 1	CLASSE 0	CLASSE 1
LOGISTIC REGRESSION	0.71	0.54	0.71	0.00	1.00	0.00	0.83
DECISION TREE CLASSIFIER	0.70	0.49	0.79	0.49	0.79	0.49	0.79
RANDOM FOREST CLASSIFIER	0.74	0.61	0.77	0.31	0.92	0.41	0.84
X G B CLASSIFIER	0.73	0.68	0.74	0.15	0.97	0.24	0.84
K N N CLASSIFIER	0.69	0.37	0.72	0.09	0.93	0.15	0.81
LINEAR SVC	0.71	0.59	0.71	0.00	1.00	0.01	0.83

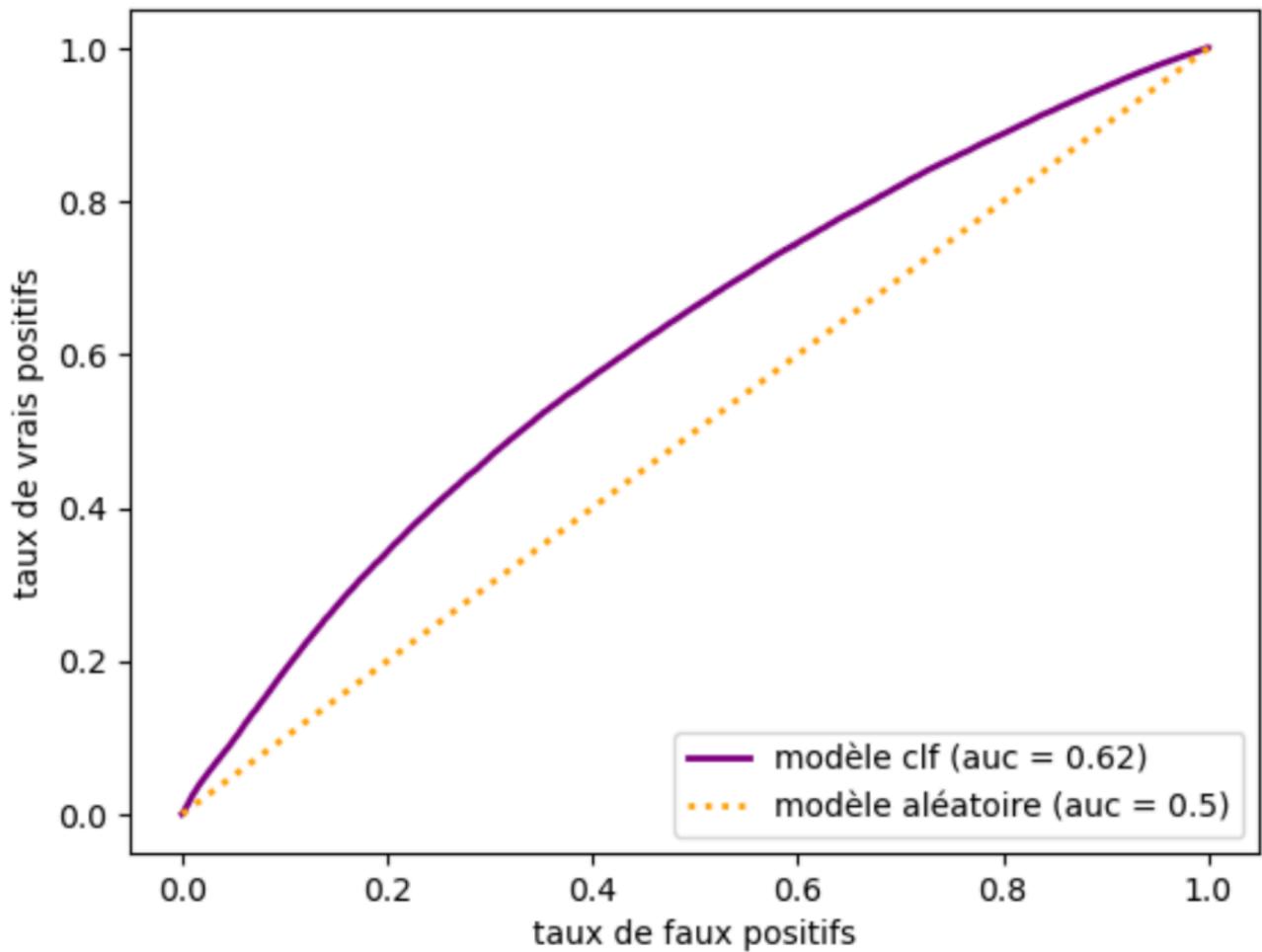
Comme nous pouvons le constater au premier abord, les scores d'accuracy apparaissent comme plutôt satisfaisants. Par exemple, nous obtenons un score de 74% avec le Random Forest Classifier (c'est-à-dire 74% de prédictions correctes sur l'ensemble des prédictions). Toutefois, d'après les

scores “recall” et “F1-score” on observe que la classe 0 n'est quasiment pas détectée par nos modèles excepté pour le Decision Tree Classifier (49% de prédictions) et le Random Forest Classifier (entre 30 et 40% de prédictions). Afin de vérifier le phénomène de surapprentissage, nous avons réalisé un accuracy sur le jeu d'entraînement. Il s'avère que la plupart de nos modèles ne présentent pas d'overfitting sauf pour le Decision Tree Classifier (0,99 sur le jeu d'entraînement et 0,70 sur le jeu de test). Nous avons donc modifié certains paramètres comme `max_depth` et `min_samples_leaf` afin de corriger le problème.

Afin d'optimiser nos résultats nous avons utilisé un `GridSearchCV` sur certains modèles afin de trouver les meilleurs hyperparamètres. Par exemple nous avons obtenu la metric “Minkowski” et le “`n_neighbors`” à 15. Malgré ces hyperparamètres, nos modèles restent non performants.

Afin d'illustrer nos propos, nous avons utilisé la courbe ROC et le score AUC sur quelques modèles, en voici une représentation graphique :

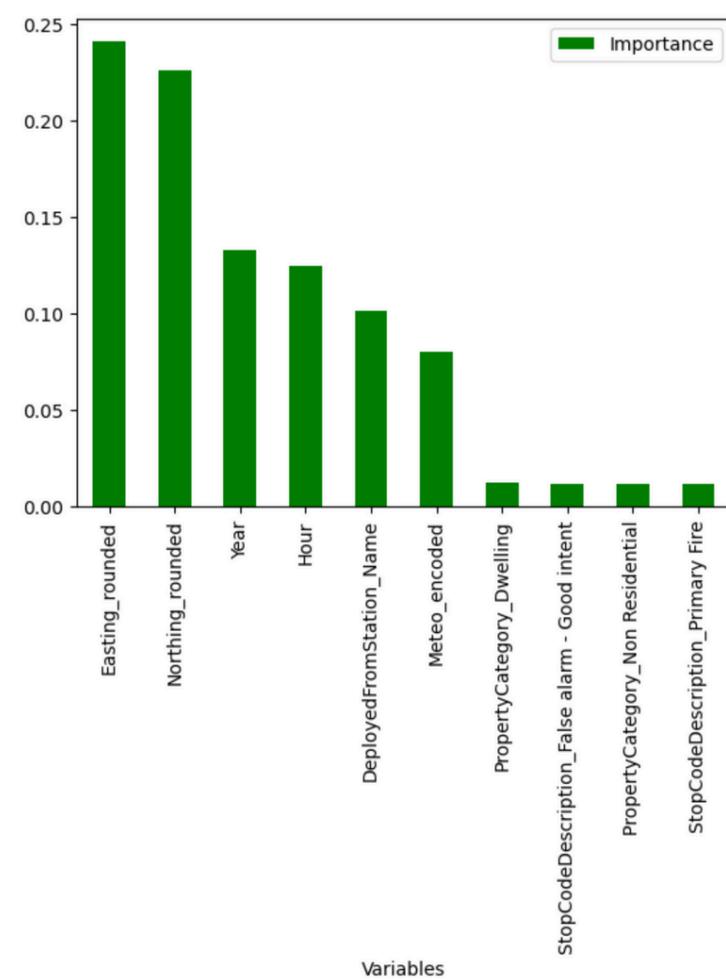
Courbe ROC de la Logistic Regression :



Grâce à cette courbe on observe que notre modèle n'est pas beaucoup plus efficace qu'un modèle aléatoire.

Nous avons réalisé un histogramme des variables les plus influentes afin de réduire notre jeu de données. Comme nous l'avions envisagé les données les plus importantes sont : les données GPS, les données temporelles et les conditions météorologiques.

Importance des variables explicatives :



Classification multi-classes :

En concertation avec notre mentor de projet, nous avons donc choisi de rester sur un modèle de classification mais de réaliser un modèle multi-classes afin d'être plus précis qu'un modèle binaire.

Face à l'étendue de nos données et à leur complexité, nous avons fait le choix de découper notre variable cible en fonction de ses quartiles. Le but étant d'obtenir des échantillons de taille similaire et ainsi permettre à notre modèle d'être plus performant. Dans un souci de compréhension, nous avons arrondi ces quartiles à la minute la plus proche :

- Classe 0 : inférieur ou égal à 4 minutes
- Classe 1 : entre 4 et 5 minutes
- Classe 2 : entre 5 et 6minutes 30secondes
- Classe 3 : entre 6minutes 30secondes et 13 minutes

Voici la répartition des classes :

- Classe 0 = 28.66%
- Classe 1 = 23.20%
- Classe 2 = 26.43%
- Classe 3 = 21.71%

Nous avons donc repris l'ensemble de nos modèles de classification binaire en les adaptant à notre problématique multi-classes. Les mêmes métriques ont donc été reprises mais ici nous nous intéressons surtout au RECALL. En effet, avec cette classification, l'objectif de notre modèle serait de prévenir la population du temps estimé d'assistance. Dans notre situation il vaut donc mieux éviter de prédire un temps court au sinistré si le temps d'assistance est en réalité long. De fait, le risque de prédire un temps inférieur à 5minutes à la personne sinistré, alors que l'unité mettra presque 10minutes (par exemple) risque d'engendrer plusieurs risques comme une plus grande détresse de la personne voire un risque vital et une perte de confiance vis à vis des services de secours. Pour simplifier, nous souhaitons éviter de sous estimer les temps d'attente des usagers.

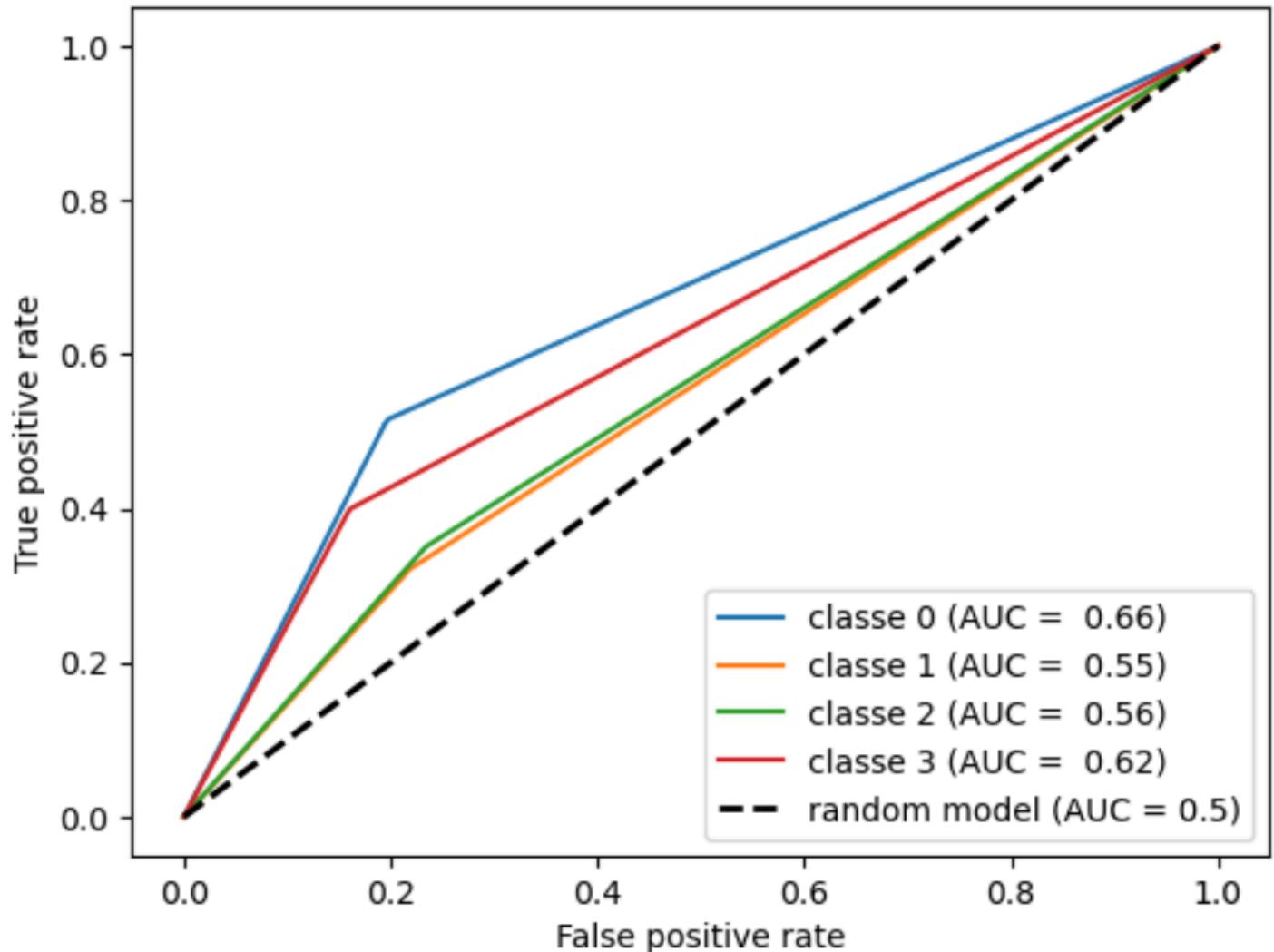
Voici les résultats :

	ACCURACY	PRECISION	RECALL	F1-SCORE
modèles de CLASSIFICATION		CLASSES 0,1,2,3	CLASSES 0,1,2,3	CLASSES 0,1,2,3
LOGISTIC REGRESSION	0.32	0.33	0.67	0.44
		0.25	0.00	0.00
		0.30	0.34	0.32
		0.32	0.19	0.24
DECISION TREE CLASSIFIER	0.40	0.51	0.51	0.51
		0.31	0.32	0.31
		0.35	0.35	0.35
		0.41	0.40	0.40
RANDOM FOREST CLASSIFIER	0.42	0.50	0.60	0.55
		0.33	0.27	0.30
		0.37	0.37	0.37
		0.45	0.41	0.43
XGB CLASSIFIER :	0.39	0.42	0.66	0.51
		0.36	0.08	0.13
		0.35	0.38	0.36
		0.41	0.39	0.40
KNN CLASSIFIER	0.28	0.31	0.45	0.37
		0.24	0.21	0.22
		0.28	0.23	0.25
		0.27	0.20	0.23

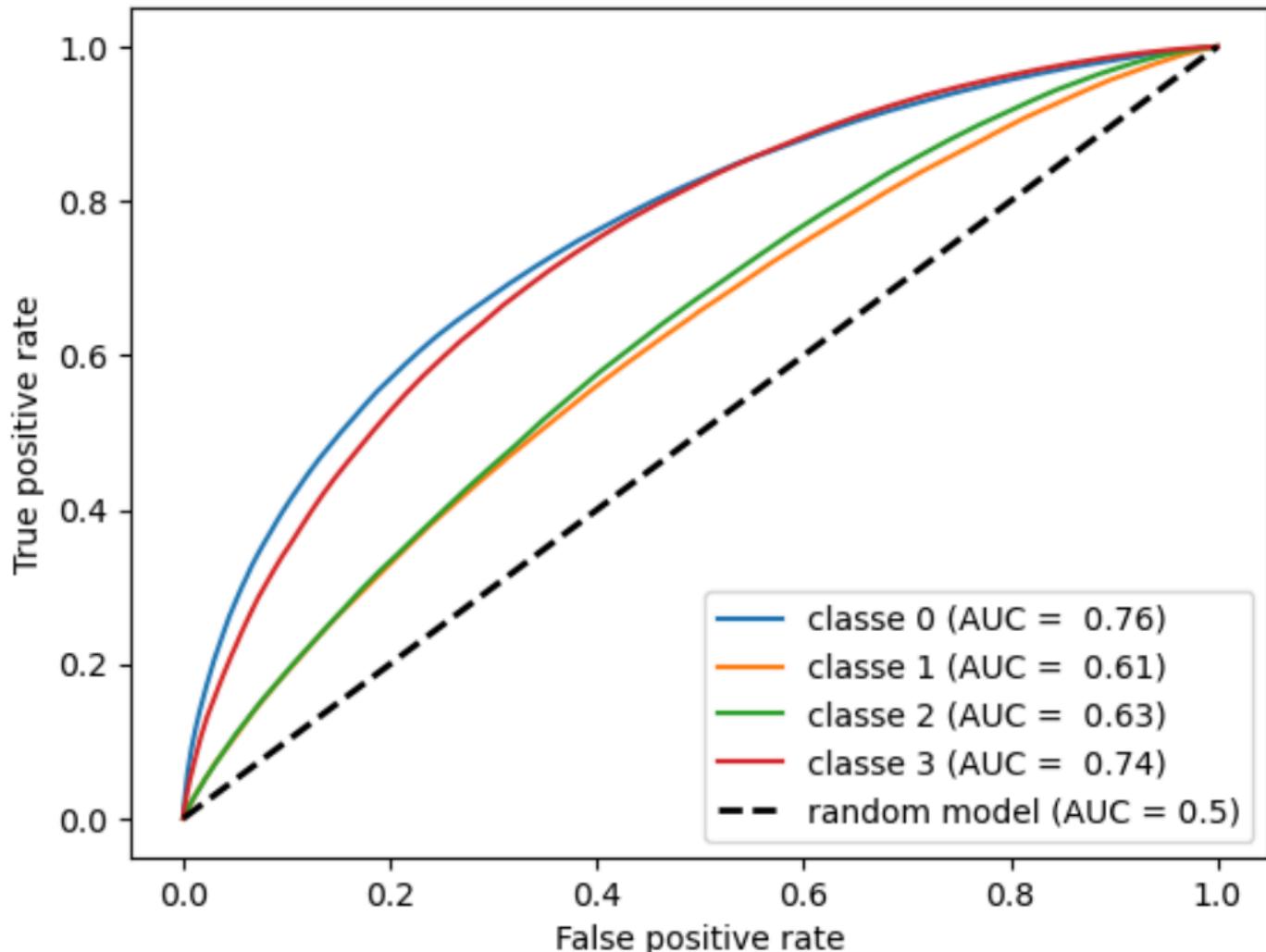
Nous constatons une baisse de l'accuracy sur nos modèles multi-classes mais un meilleur équilibre de détections entre elles.

Pour simplifier la lecture de ces résultats, voici une représentation graphique de nos deux modèles les plus performants grâce aux courbes ROC :

Courbe ROC et Score AUC du Decision Tree Classifier :

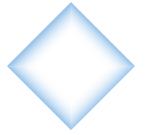


Courbe ROC et AUC du Random Forest Classifier :



Grâce à ces graphiques nous observons facilement que le Random Forest semble légèrement plus efficace qu'un modèle aléatoire. Nous verrons par la suite, l'ensemble de nos solutions d'optimisation de ces deux modèles et leurs résultats.

Choix et optimisation des modèles :



Outre la performance, notre préoccupation a aussi été l'interprétabilité des résultats et leur clarté pour une compréhension aisée à toute personne étrangère au projet. De plus, notre souhait était de réaliser un modèle qui aurait un réel intérêt pour la LFB et ses utilisateurs. C'est pour cette raison que nous avons fait le choix de ne pas garder les modélisations de classification binaire. Celles-ci n'ont d'intérêt que de prédire si les temps de réponse entreront dans les statistiques d'objectif. Il n'y a donc pas d'intérêt étant donné que ces données statistiques seront réalisées à postériori des interventions. En nous concentrant sur nos modélisations multi-classes, notre objectif serait de trouver des solutions de prédictions fiables afin de pouvoir informer l'utilisateur du temps estimé de réponse de la LFB. Aussi, grâce à ces modèles, nous pourrions aider la LFB à trouver l'unité la plus optimale pour répondre à un appel le plus rapidement possible.

Au vu de nos résultats, nous avons donc décidé de garder deux modèles et de les optimiser, notre choix s'est porté sur le Decision Tree Classifier et sur le Random Forest Classifier.

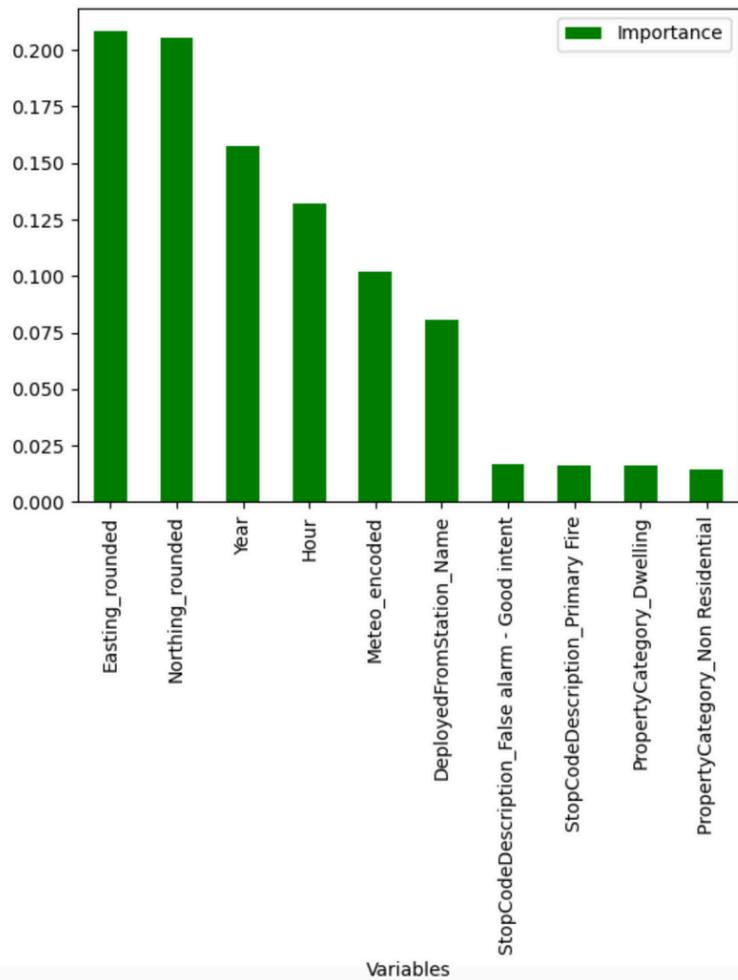
Le Decision Tree Classifier :

Dans un premier temps, nous avons d'abord tenté d'améliorer notre modèle à l'aide d'un GridSearch et d'une validation croisée. Les résultats obtenus à l'aide de cette méthode nous ont permis de fixer "max_depth" à 20 (c'est-à-dire une profondeur de l'arbre de 20 nœuds au maximum) et "min_samples_split" à 10 (c'est-à-dire un minimum de 10 échantillons requis pour séparer le nœud en 2). Toutefois, ces hyperparamètres n'ont pas suffi à améliorer notre modèle.

Voici les résultats :

	Accuracy	Precision	Recall	f1-score
Classe 0	0.42	0.49	0.62	0.55
Classe 1		0.33	0.25	0.29
Classe 2		0.36	0.39	0.38
Classe 3		0.47	0.36	0.41

Devant l'inefficacité de notre optimisation, nous avons tenté une approche différente. Nous avons réalisé une représentation graphique des variables les plus importantes :



Nous obtenons les mêmes variables que pour la classification binaire excepté les variables “Meteo_encoded” et “DeployedFromStation_Name” pour lesquelles l’importance a été “échangée”. Nous avons alors décidé de relancer notre modélisation Decision Tree Classifier en ne gardant que ces 6 variables.

Voici les résultats obtenus :

	Accuracy	Precision	Recall	f1-score
Classe 0		0.54	0.54	0.54
Classe 1		0.32	0.33	0.32
Classe 2	0.42	0.36	0.36	0.36
Classe 3		0.43	0.42	0.42

De cette manière, on observe une amélioration plus optimale de notre modèle mais nous sommes toujours loin d’un modèle satisfaisant. Nous avons donc décidé d’associer notre GridSearch et la validation croisée en ne gardant que nous 6 variables.

Voici les résultats obtenus :

	Accuracy	Precision	Recall	f1-score
Classe 0	0.44	0.51	0.65	0.57
Classe 1		0.34	0.28	0.31
Classe 2		0.38	0.41	0.39
Classe 3		0.51	0.37	0.43

Comme nous pouvons le remarquer ici, nous sommes parvenues à gagner quelques points de performance mais la classe 1 reste toujours sous représentée dans notre modèle. De plus, les résultats ne sont toujours pas satisfaisants et notre modèle est inexploitable en l'état.

Dans un second temps, nous avons donc décidé d'optimiser les paramètres du Random Forest Classifier qui avait obtenu de meilleurs résultats.

Le Random Forest Classifier :

En ce qui concerne le Random Forest Classifier nous avons tenté de trouver les meilleurs paramètres pour notre modèle, grâce au GridSearch et à la validation croisée, mais la puissance de nos ordinateurs ne nous a pas permis de les obtenir.

Afin de tenter tout de même une optimisation de notre modèle, nous avons essayé de relancer notre modèle de Random Forest Classifier en ne gardant que nos 6 variables précédemment sélectionnées.

Voici les résultats obtenus :

	Accuracy	Precision	Recall	f1-score
Classe 0	0.47	0.56	0.65	0.60
Classe 1		0.36	0.32	0.33
Classe 2		0.40	0.40	0.40
Classe 3		0.51	0.46	0.49

Bien que nous soyons parvenues à une amélioration de notre modèle, nous n'avons pas réussi à obtenir un modèle fiable, capable d'être utilisé pour la LFB.

Interprétations des résultats et proposition d'amélioration :



Comme évoqué précédemment, nos résultats ne sont pas satisfaisants et nos modèles manquent de fiabilité, malgré nos tentatives d'optimisation, après repérage des erreurs.

Par exemple, pour le Decision Tree Classifier, au meilleur de notre optimisation nous obtenons pour la classe 1 seulement 31% de prédictions correctes. On obtient 57% pour la classe 0, ce qui est notre meilleur résultat. On constate donc parfaitement que notre modèle n'est pas prédictif. Pour le Random Forest Classifier, on obtient un score plus ou moins satisfaisant pour la classe 0 (60%) mais totalement médiocre pour la classe 1 (33%).

Grâce aux hyperparamètres et à la sélection de nos 6 variables nous avons réussi à gagner quelques points de performance mais pas de manière significative.

Si le temps nous l'avait permis, nous aurions pu approfondir nos connaissances en machine learning et d'autres pistes d'optimisation plus performantes. Par exemple, nous avons tenté de mettre en oeuvre un modèle de GradientBoosting qui aurait sûrement été plus efficace mais nos ordinateurs n'ont pas réussi à le traiter.

Nous avons également observé que les variables sélectionnées sont peu discriminantes. Lors de l'exploration de notre dataset et du pré-processing nous avions tenté de créer une variable qui définirait la distance entre l'unité déployée et le lieu de l'incident. Malheureusement nous n'avions pas réussi à trouver les informations nécessaires (comme les adresses des casernes par exemple). Nous pensons qu'une telle variable serait bénéfique dans l'optimisation de notre modèle.

Malheureusement, en l'état actuel de nos connaissances et compétences techniques nous n'avons pas trouvé de solutions adaptées pour réaliser une amélioration réellement significative. Nous aimerais ajouter que la performance de nos ordinateurs ne nous a pas permis d'entraîner des modèles plus complexes.

Pour aller plus loin, il serait peut-être intéressant de revoir les formulaires donnés aux unités après leurs interventions. En effet, notre jeu de données comprenait énormément de variables (plus de 60 colonnes) dont beaucoup avaient le même objectif. Pour les variables de localisation par exemple, nous avions 26 colonnes. On observe que ces colonnes ne sont d'ailleurs pas toutes remplies, nous avons retrouvé énormément de valeurs manquantes dans ces fichiers. Peut-être serait-il nécessaire d'optimiser ce formulaire afin d'éviter les valeurs manquantes (trop de questions = chronophage pour les unités) et recentrer les informations essentielles aux statistiques et à la prédiction.

Bien que, en l'état, notre modèle ne puisse pas être exploité car il manque de fiabilité, nous avons tout de même réalisé une analyse approfondie de nos données et sommes parvenues à plusieurs conclusions métiers. Certaines pistes d'amélioration pourraient être exploitées par des professionnels plus qualifiés afin de proposer une solution optimale utile à la LFB mais également d'utilité publique.

Conclusion



L'ensemble de ce projet répond à une problématique essentielle. Certains événements marquants, comme les incendies historiques, les attentats ou encore la pandémie du COVID-19 nous rappellent l'importance d'une bonne gestion des services d'urgences. La brigade de Londres étant l'une des brigades les plus importantes du monde, elle est un exemple sur lequel il peut être intéressant de s'appuyer.

En réalisant ce projet, notre objectif était d'analyser les temps de réponse de la brigade des pompiers de Londres et éventuellement prédire ces temps de réponses. Cette analyse nous a permis de comprendre ce qui était directement lié avec les temps de réponse de la LFB et ainsi trouver des solutions et adaptations adéquates. Bien que nous ne soyons pas en capacité de prédire de manière fiable les futurs temps de réponse de la brigade pour le moment, il semblerait pertinent d'évaluer les outils à disposition de la LFB afin de proposer des actions concrètes, comme celles évoquées précédemment.

Afin de garantir la sécurité de la population et maintenir l'efficacité de la LFB, il est primordial que l'analyse de ces données s'inscrive dans une démarche de réévaluation et d'amélioration continue.

Références



- [Site de la LFB](#) (cliquez pour accéder au lien)
- [Site data de la LFB](#) (cliquez pour accéder au lien)
- Utilisation de chatGPT pour traduction et compréhension des métadatas
- [Site Historique météo](#) (cliquez pour accéder au lien)