

Rapport de projet de Data Science

Projet portant sur les données publiques de la



LONDON FIRE BRIGADE

disponible sur data.london.gov.uk

([UK Open Government Licence](#))

Introduction

La London Fire Brigade (LFB) est le corps de sapeurs-pompiers de Londres, né en 1833 de la fusion des différentes unités primitives de lutte contre l'incendie. Aujourd'hui avec 5992 employés dont plus de 5000 sapeurs-pompiers professionnels, cela en fait le service de secours et de lutte contre les incendies le plus actif du Royaume-Uni.

En outre, la LFB est la cinquième plus grande organisation de pompiers au monde, protégeant les biens et 8 millions d'habitants contre le feu sur les 1587 kilomètres carrés du Grand Londres.

En dehors de la partie opérationnelle, la brigade conduit la planification des secours et exécute des inspections et l'éducation au public.



Au cours des années 2015 et 2016, le LFB a reçu 171 488 appels d'urgence, dont 20 773 incendies, 48 696 fausses alertes d'incendie et 30 066 autres appels de service.

Les données liées aux incidents, appels d'urgences, et mobilisations de camions ont été récoltées et sauvegardées au fur et à mesure des activités de la LFB. Ce sont ces données qui nous aideront à répondre à notre problématique principale, prédire pour un incident précis un temps de réponse des pompiers associé.

Le code et les datasets de ce projet sont disponibles sur le dépôt [dec23-bds-pompiers](#) du compte GitHub DataScientest-Studio.

Compréhension et manipulation des données

Pour atteindre les objectifs de notre projet, nous avons utilisé deux jeux de données distincts :

Le premier jeu de données, accessible via le site Web gouvernemental des données de Londres, comprend des détails sur chaque incident traité depuis janvier 2009. Il fournit des informations sur la date, le lieu et le type de chaque incident.

Le deuxième jeu de données contient des détails sur chaque camion de pompiers envoyé sur les lieux d'un incident depuis janvier 2009. Il inclut des informations sur l'appareil mobilisé, son lieu de déploiement et les heures d'arrivée sur les lieux de l'incident.

Ces données sont disponibles librement sur le site Web gouvernemental des données de Londres.

La volumétrie de notre jeu de données est conséquente, avec un total de 2 277 517 entrées et 58 colonnes, et de 4.5Go. Ces colonnes comprennent divers types de données, des entiers (5), des flottants (19) et des objets (34).

La colonne "IncidentNumber" identifie de manière unique chaque incident, tandis que la colonne "CalYear" indique l'année de l'incident. Parmi les autres colonnes, on trouve des informations telles que l'heure de l'appel ("HourOfCall"), le temps de mobilisation ("DateAndTimeMobilised"), le temps de déplacement ("TravelTimeSeconds"), et le temps d'arrivée sur les lieux ("DateAndTimeArrived").

Il est à noter que certaines données peuvent être manquantes dans certaines colonnes. Par exemple, la colonne "DateAndTimeReturned" présente un taux élevé de données manquantes, atteignant environ 57,27%.

En ce qui concerne les statistiques descriptives, la plupart des colonnes ont des valeurs moyennes raisonnables et des écarts-types modérés, mais les valeurs maximales et minimales peuvent varier considérablement selon la colonne. Par exemple, le temps d'attente peut varier de 0 à 1200 secondes.

Enfin, il est important de noter que certaines colonnes ont un grand nombre de valeurs uniques, ce qui peut indiquer une diversité significative dans les données. Par exemple, la colonne "Resource_Code" compte 187 valeurs uniques, et la colonne "PerformanceReporting" en compte 3.

Cette description met en lumière la complexité et la richesse du jeu de données, nécessitant une analyse approfondie pour en tirer des informations significatives.

Les variables les plus pertinentes pour nos objectifs sont les suivantes :

- IncidentNumber
- CalYear
- HourOfCall
- ResourceMobilisationId
- Resource_Code
- PerformanceReporting
- DateAndTimeMobilised
- DateAndTimeMobile
- DateAndTimeArrived
- TurnoutTimeSeconds
- TravelTimeSeconds
- AttendanceTimeSeconds
- DateAndTimeLeft
- DeployedFromStation_Name
- DeployedFromLocation
- PumpOrder
- PlusCode_Code
- DateOfCall
- IncidentGroup
- PropertyCategory
- PropertyType
- AddressQualifier
- Postcode_district
- BoroughName
- NumCalls

La variable cible de notre étude est "**AttendanceTimeSeconds**", puisque notre objectif est d'estimer le temps de réponse.

Le jeu de données présente plusieurs particularités remarquables qui méritent d'être soulignées :

1. **Données temporelles précises** : Les données sont temporelles et fournissent des informations détaillées sur les événements, tels que l'heure de l'appel, le temps de mobilisation, le temps de déplacement et le temps d'arrivée sur les lieux. Cette précision temporelle est essentielle pour analyser les performances opérationnelles des services d'incendie.
2. **Hétérogénéité des variables** : Le jeu de données est composé de différentes variables, comprenant des types de données variés tels que des entiers, des flottants et des objets. Cette diversité offre une multitude d'angles d'analyse et de possibilités pour explorer les relations entre les différentes variables.
3. **Données manquantes** : Certaines colonnes présentent un pourcentage élevé de données manquantes, ce qui peut nécessiter des stratégies spécifiques pour les gérer lors de l'analyse. Par exemple, la colonne "DateAndTimeReturned" présente un taux de données manquantes d'environ 57,27%, ce qui peut influencer les résultats des analyses impliquant cette variable.

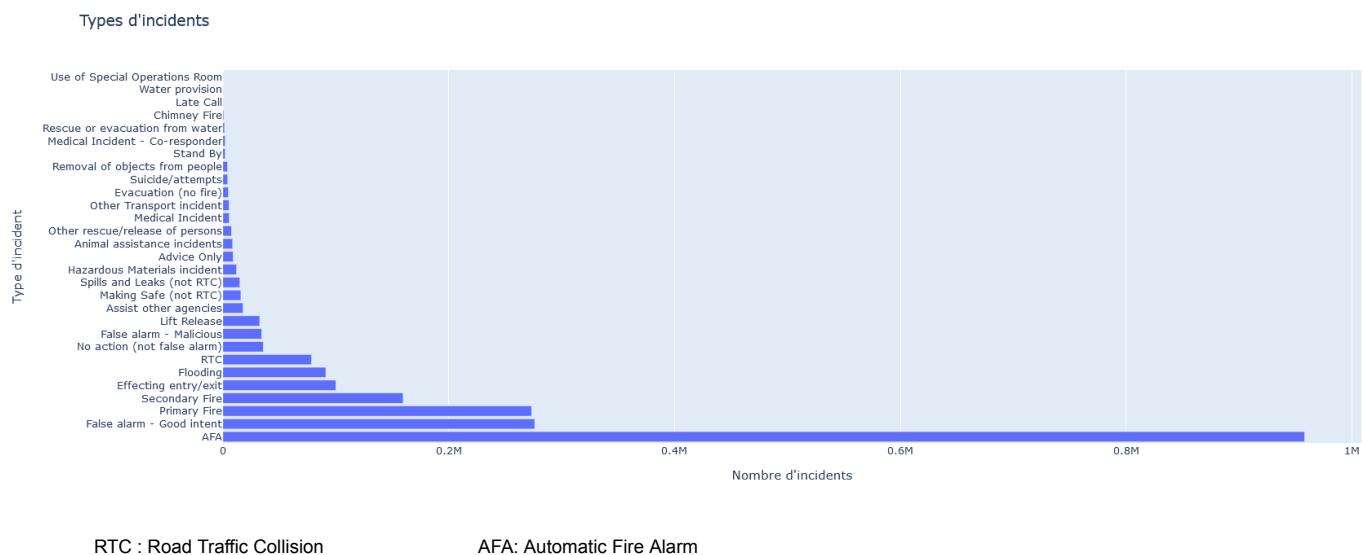
4. **Grande variabilité des valeurs** : Les statistiques descriptives révèlent une grande variabilité des valeurs dans certaines colonnes. Par exemple, le temps d'attente varie de 0 à 1200 secondes, ce qui indique des situations très différentes rencontrées lors des interventions des services d'incendie.
5. **Nombre élevé de valeurs uniques** : Certaines colonnes présentent un nombre élevé de valeurs uniques, ce qui témoigne de la diversité des situations auxquelles les services d'incendie sont confrontés. Par exemple, la colonne "Resource_Code" compte 187 valeurs uniques, ce qui reflète la multitude de ressources mobilisées lors des interventions.

En résumé, le jeu de données est caractérisé par sa richesse en données temporelles précises, sa diversité de variables, la présence de données manquantes, une grande variabilité des valeurs et un nombre élevé de valeurs uniques. Ces particularités offrent des opportunités passionnantes pour une analyse approfondie et une compréhension plus fine des opérations des services d'incendie.

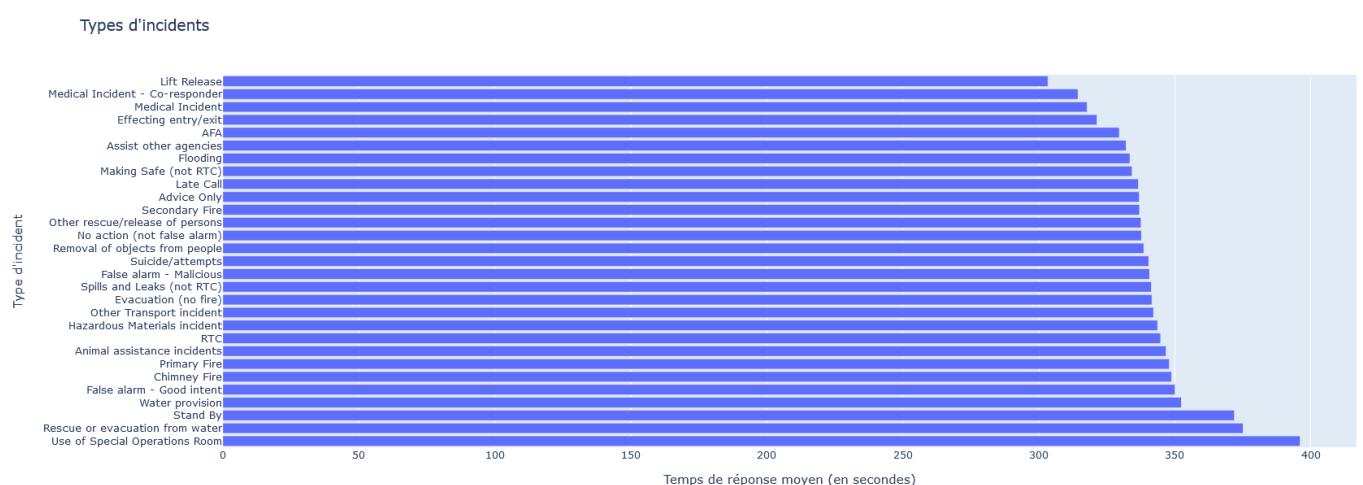
Exploration des données

Pour tenter de mieux comprendre nos données, nous explorerons de manière graphique les différentes distributions et relations entre nos variables.

Pour commencer, les incidents sont caractérisés par leurs types (incendie, malaise, accident de la route etc), et certains types sont bien plus présent dans notre jeu de données que d'autres :



On constate que les pompiers de Londres sont très actifs sur les incendies (et tests d'alarme associés). Cela peut sembler évident pour des pompiers, cependant à titre de comparaison, les incendies ne représentent que 3,1% des interventions de la Brigade de sapeurs-pompiers de Paris (pompierparis.fr).



Le temps de réponse des pompiers varie peu en fonction du type d'incidents, bien qu'on puisse constater un temps de réponse en général légèrement plus court pour les urgences médicales que pour les incendies par exemple. Cela pourrait éventuellement s'expliquer par la quantité de matériel à préparer et dont s'équiper.

Ensuite, étudions notre variable cible “AttendanceTimeSeconds”.

Ce temps en secondes correspond à la somme des variables suivantes :

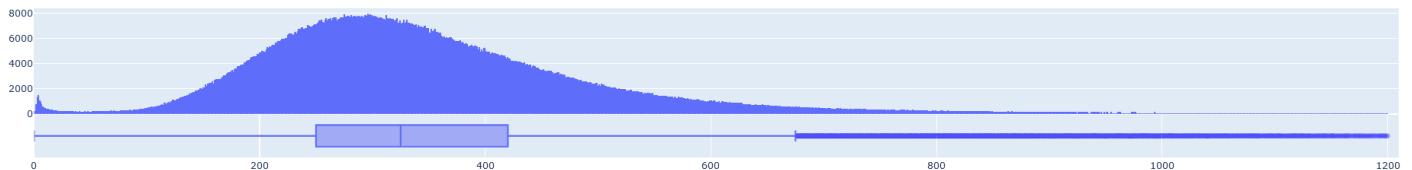
- “TurnoutTimeSeconds” : temps séparant l'instant où l'incident est confirmé par l'opérateur d'appel qui déclenche alors l'ordre de départ et l'instant où l'équipe quitte la caserne avec le camion.
- “TravelTimeSeconds” : temps séparant l'instant où l'équipe quitte la caserne avec le camion et l'instant où l'équipe arrive sur les lieux de l'incident et commence à traiter celui-ci.



“AttendanceTimeSeconds” correspond donc au temps séparant l'ordre de départ (étape 3) et l'arrivée du camion sur les lieux de l'incident (étape 5).

Représentons une distribution statistique de notre variable :

Distribution du temps de réponse en secondes des pompiers de Londres

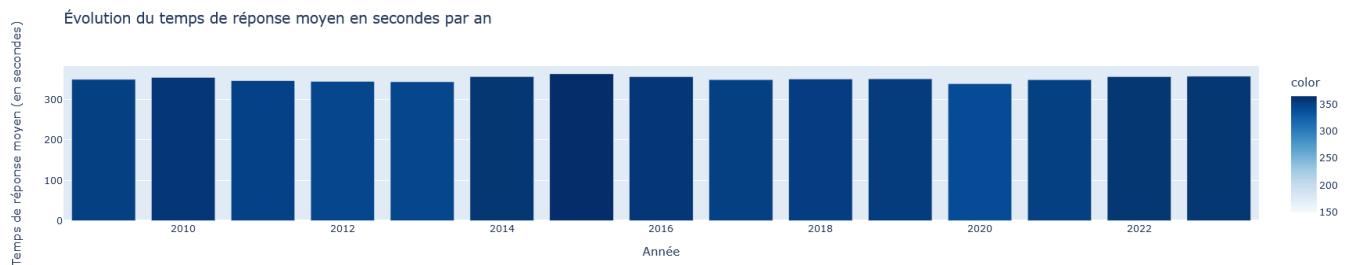


Description statistique :

25% des valeurs sont inférieures à 250 secondes et 25% supérieures à 420 secondes, avec une médiane à 325 secondes. On observe également des valeurs aberrantes au-delà de 675 secondes, avec un maximum de 1200 secondes.

Environ la moitié des mobilisations ont donc des temps de réponse allant de 4 à 7 minutes.

Explorons désormais l'évolution du temps de réponse des pompiers de Londres sur le long terme :



Nous observons ici le temps de réponse moyen de chaque année depuis 2009, et nous pouvons aisément constater que ces différences de délais ne semblent pas très significatives puisque le minimum et le maximum de temps réponse obtenu respectivement en 2020 et 2015 ne sont séparés que d'environ 24 secondes.

Pour préciser les variations du temps de réponse en fonction du temps, concentrons-nous sur une plus petite échelle, la variation fonction des heures de la journée :



On constate une faible variation du temps de réponse moyen au fil de la journée, notamment avec un légère augmentation en fin de nuit et l'après-midi, et un temps de réponse légèrement plus rapide le matin et en soirée.

Au total, on compte une certaine variation allant jusqu'à 43 secondes supplémentaires pour les interventions de 15h comparé aux interventions de 9h ou de 22h.

Preprocessing

Nos deux jeux de données possèdent un ID d'incident commun qui nous a permis de les fusionner entre eux pour obtenir un unique jeu de données comportant à la fois les informations liées aux incidents, ainsi qu'aux interventions de pompiers pour les traiter.

Un traitement a été effectué sur les colonnes comportant des données s'apparentant à des dates ou des horaires. Ainsi, quatre colonnes contenant respectivement la date (jour, mois et année), année, horaire (heure, minute et seconde) et l'heure de la journée, ont été fusionnées en une seule colonne "DateOfCall" de type "datetime" pour éviter tout doublon d'information et permettre d'utiliser un type de donnée dédié à cet usage.

Puisque chaque colonne se rapportant aux types d'incidents n'était qu'une précision des colonnes précédentes, les trois colonnes ont également été fusionnées en une colonne unique. De plus, nous avons supprimé les incidents de type "False Alarm", cette information n'étant disponible qu'à posteriori de la prédiction.

Certaines colonnes ont été renommées à des fins de lisibilité, comme le nom en minuscule des arrondissements où ont eu lieu les incidents nommé "ProperCase" qui a été renommé en "BoroughName".

Après réflexion et études des corrélations entre les différentes variables explicatives éventuelles, nous avons décidé de conserver les colonnes "IncidentNumber", "DateOfCall", "BoroughName", "IncidentType", "Station_Name", "Distance" et notre variable cible "AttendanceTimeSeconds".

Avec la suppression des valeurs manquantes (peu nombreuses après nos différents traitements), les mobilisations dont la caserne de départ est l'une des 6 suivantes ont été supprimées : Buckinghamshire, Dartford, Esher, Fordbridge, Hertfordshire et Staines. Pour cause, ces 6 casernes étaient associés en tout à 7 enregistrements de notre dataset, sur un total de 2 156 331.

Pour finir, nous avons considéré que les distances de plus de 10km étaient bien trop rares dans notre jeu de données pour qu'elles soient statistiquement représentative. Ces lignes, ainsi que celles dont le temps de réponse est inférieur à 93 secondes ou supérieur à 676 secondes ont été supprimées.

Feature Engineering

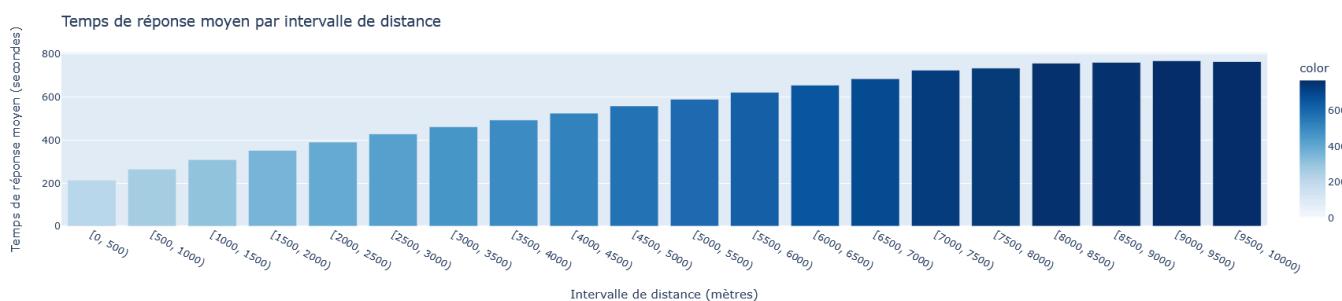
Il semblerait que la plupart des variables dont on pourrait instinctivement se dire qu'elles devraient avoir une influence sur le temps de réponse des pompiers n'en ont en réalité que très peu. Cela limite malheureusement notre marge de manœuvre pour réaliser un modèle de prédiction. Malheureusement une autre information ayant une très probable corrélation avec le temps de trajet des pompiers (et donc sur le temps de réponse total) sur notre variable cible n'existe pas dans notre dataset : la distance séparant le lieu de l'incident et la caserne d'où part le véhicule de pompier.

Cette variable est cependant calculable grâce à deux informations : les coordonnées GPS du lieu de l'incident et le nom de la première caserne concernée par l'intervention. Avec deux points GPS, on peut calculer grâce à la formule de Haversine ([wikipedia.org](https://en.wikipedia.org)).

La seule chose dont nous ne disposons pas pour calculer la distance est la latitude et la longitude des casernes de pompiers. Par chance, l'organisation Open Data Institute (ODI) a déjà travaillé sur une problématique similaire sur les mêmes données de la London Fire Brigade. Sur le GitHub de l'ODI est disponible un [fichier csv](#) comportant le nom de la centaine de casernes de pompiers de Londres ainsi que leurs coordonnées GPS.

En fusionnant le contenu de ce fichier avec notre dataset, puis en convertissant les coordonnées “easting” et “northing” des lieux d’incidents du format OSGB au format WGS84 (format standard de latitude et longitude), on peut appliquer la formule de Haversine pour calculer les distances en mètres entre le lieu d’incident et la caserne d'où part le camion de pompier.

En regroupant les distances par plages de 500 mètres, et en affichant les temps de réponse moyen des mobilisations, on observe ceci :



On obtient une évidente corrélation entre la distance séparant les casernes et les lieux d’incidents, une information qui nous sera utile pour notre modèle.

Modélisation

Préparation du jeu de données

Tout d'abord, nous récupérons notre jeu de données composé des colonnes "BoroughName", "IncidentType", "Station_Name", "Distance" et notre variable cible "AttendanceTimeSeconds", désormais sans la moindre valeur manquante, ayant un total de 2 156 331 lignes, soit autant d'exemples de mobilisations de camions de pompiers.

Nous effectuons d'abord une standardisation standard (StandardScaler) de nos variables numériques pour amener la distribution de nos distances en mètres et de nos temps de réponse en secondes sur une même échelle, de moyenne 0 et d'écart type 1.

Les autres variables (type d'incident, nom de l'arrondissement et nom de caserne) étant catégorielles, nous effectuons un encodage one-hot de manière à pouvoir les utiliser avec des modèles de régression ne pouvant pas gérer de telles variables.

Il ne reste alors plus qu'à séparer notre variable cible de nos variables explicatives, ainsi qu'à scindé en deux notre dataset pour créer un jeu d'entraînement et un jeu de validation.

Étapes de réalisation du projet

Identification du problème

Lorsqu'une personne est témoin d'un incident (incendies, malaises, etc) et appelle le 999 (numéro d'urgence du Royaume-Uni), un opérateur d'appel enregistre les différentes informations dont les pompiers ont besoin pour intervenir : lieu et type de l'incident etc. A partir de l'instant où cet opérateur constate l'état d'urgence de la situation et déclenche l'alerte, la brigade de pompiers est sollicitée pour se rendre sur les lieux.

L'objectif de ce projet est de créer un modèle capable de donner une prédiction du temps de réponse de la brigade de pompiers devant intervenir sur l'incident, en fonction des différentes informations données par le témoin à l'opérateur. Cette tâche peut être décrite par une régression, car elle vise à anticiper une mesure quantitative, à savoir le temps nécessaire pour que les pompiers répondent à un appel d'urgence en se rendant sur place, en se basant sur divers facteurs.

Choix des métriques

Ce modèle sera créé de manière à minimiser l'erreur quadratique moyenne (MSE) et maximiser le coefficient de détermination R^2 afin de rendre les prédictions les plus précises que possible. Nous avons choisi comme métrique principale de performance le Mean Squared Error (MSE), la Root Mean Squared Error (RMSE) et le coefficient de détermination R^2 . Ces métriques ont été sélectionnées en raison de leurs pertinences dans l'évaluation de la précision et de la qualité de nos modèles de régression.

Le MSE mesure la moyenne des carrés des écarts entre les valeurs prédites et les valeurs réelles, fournissant ainsi une indication de la précision globale du modèle.

Le RMSE, quant à lui, est une mesure de dispersion des erreurs de prédiction. En prenant la racine carrée du MSE, il fournit une estimation de l'écart moyen des résidus, ce qui le rend plus interprétable et plus facile à comparer à l'échelle des valeurs initiales.

Enfin, le coefficient de détermination R^2 mesure la proportion de la variance de la variable dépendante qui est prévisible à partir des variables indépendantes. Il offre une indication de l'adéquation globale du modèle aux données observées, exprimant ainsi la proportion de la variance totale de la variable dépendante qui est expliquée par le modèle.

En utilisant ces métriques, nous serons en mesure d'évaluer de manière exhaustive la performance de nos modèles de prédiction, ce qui nous permettra de sélectionner celui offrant les prédictions les plus précises et les plus fiables pour notre objectif spécifique.

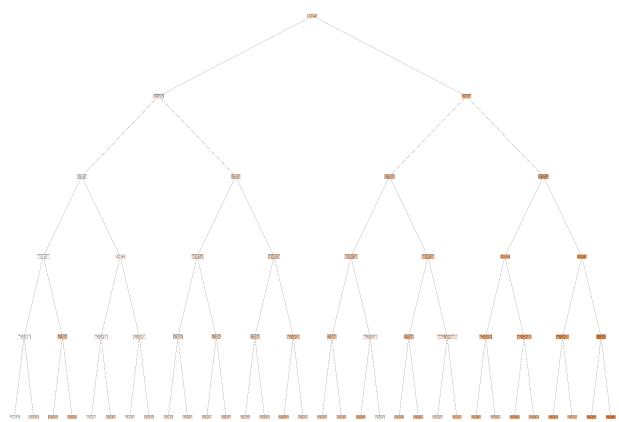
Choix du modèle et optimisation

Nous avons exploré plusieurs algorithmes dans notre étude, débutant par des modèles de régression linéaire simple et multiple à l'aide de LinearRegression. Nous avons également testé des modèles tels que Ridge (Ridge Regression), Lasso (Lasso Regression), ElasticNet (Elastic Net Regression), RandomForestRegressor (Random Forest), SGDRegressor (Stochastic Gradient Descent), DecisionTreeRegressor (Decision Tree) et GradientBoostingRegressor (Gradient Boosting). Tous ces modèles ont été générés avec des fonctions de la bibliothèque scikit-learn.

Parmi ces modèles, celui que nous avons retenu comme étant le plus performant est DecisionTreeRegressor. Malgré les nombreuses expérimentations et l'ajustement des hyperparamètres effectués sur les autres modèles, nous n'avons pas pu obtenir des résultats supérieurs à ceux obtenus avec DecisionTreeRegressor. Nous avons pu optimiser ce modèle en utilisant des techniques telles que la validation croisée et la recherche de grille (Grid Search CV), bien que cela puisse prendre un certain temps en raison de la taille importante de notre ensemble de données.

La robustesse et la flexibilité de DecisionTreeRegressor nous ont convaincus de le retenir comme modèle principal pour notre projet. Il a démontré une capacité à capturer les relations non linéaires entre les variables explicatives et la variable cible, ce qui est crucial dans notre contexte de prédiction du temps de réponse des services des pompiers d'urgence à Londres. Alors que tous les modèles essayés offrait des performances relativement similaires avec un R^2 d'environ 0.48 et un MSE variant entre 0.53 et 0.57, l'arbre de décision est le seul modèle à avoir montrer un R^2 supérieur, de 0.51, pour un MSE similaire. Ces performances ont été obtenues grâce aux hyperparamètres suivants : `max_depth : None`, `min_samples_split : 5`, `min_samples_leaf : 25`, obtenu grâce à une recherche par validation croisée.

Si on représente les 5 premiers niveaux de notre arbre de décision (disponible en grande taille dans le notebook “modélisation”, disponible sur le dépôt Github), les nœux des 4 premiers niveaux de cet arbre sont majoritairement composés de tests sur notre variable *distance*, ce qui semble bien confirmer l’importance de cette information pour déterminer un temps de réponse. On commence à voir des tests sur les noms de quartiers, de casernes et d’arrondissements seulement à partir du 5ème étage.

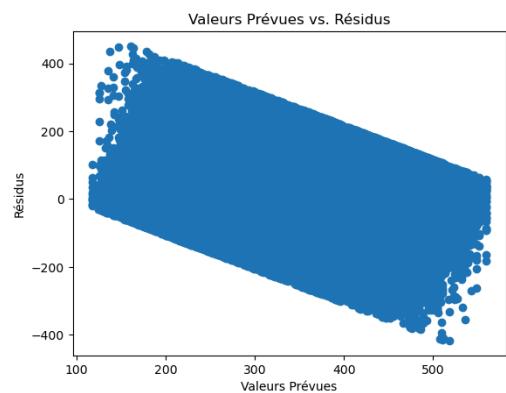


Points observés	Points prédits
323	305
231	222
626	499
663	416
259	230
148	266
381	454
395	393
484	434

Voici un échantillon des prédictions de notre modèle.

On constate aisément une certaine cohérence entre les résultats prédits et les véritables temps de réponse des brigade de pompiers. Cependant, on peut également constater une certaine imprécision sur les valeurs les plus faibles et les plus élevées, mettant en valeur la fiabilité de notre modèle sur la majorité des délais, mais une capacité de prédiction moindre sur les délais s’écartant de la norme.

Cette notion de sur-évaluation des valeurs faibles et de sous-évaluation des valeurs hautes peut être observée en affichant les résidus des valeurs prédites à l’aide d’un graphique en nuage de points.



Conclusion

Les performances obtenues par notre modèle peuvent éventuellement sembler relativement faibles comparées aux valeurs que nous avons régulièrement rencontrés dans notre formation, cependant nous faisons face ici à une problématique possédant de nombreux facteurs aléatoires rendant toutes prédictions très difficiles.

On peut notamment penser à l'activité des pompiers au moment de l'alerte, leurs rapidité pour préparer le matériel, leurs équipements individuels, le trafic routier, la météo, la précision et l'accessibilité de l'adresse de l'incident renseigné par le témoin, etc.

Finalement, ces résultats, bien que très probablement perfectibles, représentent généralement bien la plupart des délais nécessaires pour que les pompiers arrivent sur les lieux d'incidents signalés sur le numéro d'urgence à Londres.

Ce type de modèle pourrait permettre à terme à l'opérateur des pompiers de donner au téléphone un temps d'attente approximatif à l'interlocuteur au téléphone.

De plus, en mettant à jour ce modèle avec de nouvelles données, et en étudiant plus profondément les différents facteurs qui ont un impact sur ces délais, cela pourrait permettre à la ville de Londres d'optimiser ces différents facteurs via des modifications organisationnelles de la London Fire Brigade, voir des modifications d'infrastructure pour fluidifier les interventions.

Il est aussi important de préciser qu'il serait tout à fait possible d'adapter ce type de modèle à d'autres services publics, dans d'autres villes du monde, à condition que ces derniers effectuent un suivi rigoureux des données liées à leurs activités.