Temps de réponse de la Brigade des Pompiers de Londres

Angeline Duqueyroix
Aurélie Patron
Soliman Traboulsi

La brigade des pompiers de Londres, ou London Fire Brigade

- 5ème plus grand corps de sapeurs-pompiers au monde
- plus de 5 096 sapeurs-pompiers professionnels
- 103 casernes
- 33 subdivisions administratives de Londres

Notre objectif : analyser leur temps de réponse et prédire le délai d'intervention

SOMMAIRE

1 - Introduction

Contexte

Notre choix de projet s'est porté sur les pompiers de Londres. Mais pourquoi ? Les pompiers font partie d'une catégorie de métier considéré comme dangereux et ont une importance dans la société. Sans les pompiers, de nombreuses personnes périraient au cours d'incendies, d'accidents de la route, d'inondations... Leur rôle est important comme le montrent les actualités de nos jours ou ils sont très présents (incendies dans le monde, accidents de la route). Cependant notre connaissance sur le sujet est limitée, à part pour Soulayman.

Objectif

L'objectif de notre projet est d'analyser le temps de réponse des pompiers en fonction de différentes variables et de réussir à prédire le temps qu'ils mettront pour arriver sur les lieux des accidents.

2 - Les données

Cadre

Afin de mener à bien notre projet , deux fichiers ont pu être récupérés sur le site des pompiers de Londres (London Fire Brigade Incident Records - London Datastore) : un contenant des informations sur les incidents dans Londres depuis les années 2009 et un autre contenant les informations sur les mobilisation pour les mêmes incidents et sur la même période. Le site des pompiers de Londres met ces informations à disposition de toutes personnes intéressées.
La fusion de ces deux fichiers représente un volume important de données sur de nombreuses années. Nous avons donc décidé de nous limiter à une période de cinq ans pour la suite du travail : de 2017 à 2022, ce qui représente malgré tout 942502 incidents, décrits par 58 variables.

Pertinence

Parmi toutes nos données fournies, certaines se distinguent, notamment la variable cible. La variable cible n'est autre que le délai d'intervention appelé 'AttendanceTimeSeconds' dans notre jeu de données. Il prend en compte le temps de préparation des pompiers après avoir reçu l'appel et le temps de trajet pour se rendre sur les lieux de l'incident.
Exemple des colonnes présentes dans notre jeux de données : IncidentNumber, CalYear, TimeOfCall, HourOfCall, IncidentGroup, PropertyCategory, IncGeo_Borough_Name, Latitude, Longitude, DeployedFromStation_Name, etc…
Etant donné que nous avons fusionné 2 Data Frame pour construire le jeu de données, nous aurons certainement des données redondantes. Et certaines données ne sont pas exploitables car trop de valeurs manquantes.

Pre-processing

Les informations d'ensemble de notre jeu de données ont mis en évidence la présence de colonnes inexploitables et de valeurs manquantes, et également de quelques types de variables inadaptés.

Nous avons procédé au nettoyage des données comme suit :

1- Suppression des colonnes identifiées comme doubles, inutiles ou inexploitables

En observant le nombre de valeurs uniques prises par les variables ainsi que ces valeurs, nous avons identifié des colonnes contenant la même information sous des formes différentes ou des colonnes inutiles et jugé de celles que nous pouvions supprimer.

En observant les taux de valeurs manquantes, nous avons écarté les variables qui n'étaient pas exploitables (ici, nous avions de taux entre 39% et 99% pour certaines colonnes). Certaines de ces colonnes étant corrélées à 100% à d'autres colonnes n'ayant, elles, aucune ou peu de valeurs manquantes, elles ont également été supprimées. Les colonnes conservées contiennent des indications géographiques qui nous seront utiles pour la suite du projet.

2- Suppression de certaines lignes contenant des valeurs manquantes

Toujours en observant les taux de valeurs manquantes, mais cette fois les plus bas, nous avons identifié des variables ayant exactement le même taux et ayant également en commun les lignes (les incidents) contenant ces valeurs manquantes.
Nous avons choisi de supprimer ces lignes, ainsi que les autres lignes contenant des valeurs manquantes pour des variables catégorielles ayant un taux de valeur manquantes inférieur à 0,1%.

3- Gestion des valeurs manquantes pour les variables quantitatives

Après observation de la distribution de ces variables à l'aide de Box Plot, compte tenu de la présence de valeurs extrêmes sur l'ensemble de ces variables, nous avons choisi de remplacer les valeurs manquantes par les médianes.

4- Gestion des valeurs manquantes pour les variables qualitatives

1 seule variable a fait l'objet d'un remplacement de valeur manquante par une valeur arbitraire pour sa particularité : la colonne "SpecialServiceType" contient 78% de valeurs manquantes, mais les valeurs renseignées sont liées à la valeur "Special Service" dans la colonne "StopCode Description". Nous avons décidé de remplacer ces valeurs manquantes par la valeur "Not_concerned", car ces incidents ne sont pas spécifiques à un "Special Service".
Cette information sera utile dans la partie Visualisation.

5- Standardisation : Conversion des Colonnes Temporelles au Format Datetime

Nous travaillons sur des flux d'interventions qui sont datés et minutés à plusieurs étapes, donc notre jeux de données contient énormément d'informations temporelles.
Nous avons mis les variables temporelles au format datetime et réduit certaines au format time et pour faciliter leur manipulation et analyse ultérieure. Nous en avons notamment extrait une variable "Month" pour compléter ces variables temporelles.

A l'issue de ce pré-traitement, il reste 2 colonnes contenant des valeurs manquantes. Nous ferons attention à leur usage par la suite.
Notre jeu de données est maintenant de 941660 incidents décrits par 44 variables.
L'affichage d'une matrice de corrélation pour visualiser les coefficients de corrélation entre toutes les paires de variables quantitatives en fin de traitement nous confirme que nous n'avons plus de variables qui seraient trop corrélées.

Feature engineering

Nous avons beaucoup d'informations géographiques détaillant les lieux d'intervention, entre autres des coordonnées Easting et Northing.
Les latitudes et longitudes présentes à l'origine dans notre jeu de données n'étant pas exploitables, nous avons utilisé ces coordonnées pour les convertir en latitudes et longitudes déterminant le lieu d'intervention.

Nous avons également importé les latitudes et longitudes des casernes, qui manquaient dans notre jeu de données, depuis le site "London Fire Brigade" qui répertorie les casernes de Londres et renseigne sur leurs coordonnées géographiques.

Grâce à ces latitudes et longitudes pour les lieux d'intervention et pour pour les casernes depuis lesquelles les brigades sont déployées, nous avons ajouté à notre jeu de données la distance parcourue entre la caserne et le lieu d'intervention en utilisant la méthode Manhattan. Cette méthode repose sur un quadrillage pour estimer des distances parcourues en ville selon un schéma de déplacement citadin, mise au point en se servant de la ville de Manhattan comme référence.

Visualisations et Statistiques

1- Observation des variables pouvant influer sur le nombre d'interventions

- La représentation graphique via Plotly sous forme de courbe du nombre d'incidents par années entre 2017 et 2022, associé à l'utilisation du test statistique Anova, révèlent que l'année n'influe pas sur le nombre d'accidents.
  Mais il se peut que "les années Covid" aient fait baisser un peu ce taux !

  En revanche, cette représentation en fonction des mois de l'année permet d'observer que le nombre d'incidents varie selon les mois, avec des pics en juillet et en décembre.

- Nous avons également analysé le nombre d'incidents par années selon les 3 types répertoriés False Alarm, Fire et Special Service à l'aide d'un bar graphe : la tendance est la même quelle que soit l'année.
  En complément, l'observation du taux d'intervention par type d'incidents via un diagramme en camembert mène à la conclusion que près de 60% des interventions recensées correspondent à à des fausses alertes.

2- Observation des variables pouvant influer sur le délai d'intervention

Ce délai d'intervention est notre variable cible pour la réalisation du projet, il s'agit du temps écoulé entre la réception de l'appel et l'arrivée de la première brigade sur les lieux d'intervention. Nous recherchons les variables explicatives les plus pertinentes dans notre jeu de données.
Le délai d'intervention moyen est d'environ 6 minutes, et 50% des interventions se fait avec un délai compris entre 4 minutes et 7 minutes.

- Variables temporelles :

  Observation de ces variables par des courbes représentant le délai d'intervention en fonction de la variable temporelle.
  Le résultat du test statistique ANOVA montre une dépendance entre l'année d'intervention et le délai.
  Le résultat du test statistique de Kruskal-Wallis montre une dépendance entre le mois de l'intervention et le délai.
  Le résultat d'une régression OLS montre une dépendance entre l'heure de l'intervention et le délai.

- Variables géographiques :

  Observation de ces variables grâce à des cartes représentant les délais d'intervention en fonction des arrondissements et des districts de Londres.
  Le résultat du test statistique ANOVA montre une dépendance du délai d'intervention avec l'arrondissement concerné, mais pas avec le district.

Après avoir établi un Top 10 des casernes présentant le délai d'intervention le plus bas, nous avons établi avec un test ANOVA que le délai d'intervention dépend de la caserne qui prend en charge l'intervention.

La plus importante de ces variables est pour nous la distance calculée précédemment.
Nous avons observé sa distribution pour confirmer la pertinence des valeurs calculées et éliminer les valeurs aberrantes.
Ensuite, en affichant le relation Distances / Délai d'intervention nous avons établi des seuils sur les délais d'intervention au-delà desquelles les valeurs n'étaient pas exploitables. Le seuil minimal est basé sur le temps moyen que mettent les équipes à se préparer, et le seuil maximal est basé sur un délai qui devient trop élevé pour que la distance soit le seul facteur à entrer en jeu.
Nous avons également éliminé les distances au-delà desquelles la relation délai / distance devient inexploitable.

- Autres variables :

Observation du délai moyen d'intervention en fonction des 3 types d'incidents au moyen d'un bargraphe.
Le résultat du test statistique ANOVA montre une dépendance entre le type d'incident et le délai d'intervention .

Conclusion de cette phase d'exploration

Nous avons à prédire un délai d'intervention. Nous avons pour cela défini certaines variables comme étant suffisamment pertinentes et influentes pour être les variables explicatives du jeu de données que nous utiliserons pour les modélisations.
Notre variable cible étant une variable quantitative, nous prévoyons d'entraîner des modèles de régression.
Cependant, en observant la relation Distances / Délai d'intervention, nous avons également choisi d'établir deux variables binaires supplémentaires basées sur les délais d'intervention : un délai court et un délai long, pour pouvoir étudier cette relation sous forme de classification.

Nous exportons maintenant ce nouveau jeu de données constitué d'une quinzaine de colonnes sous forme d'un nouveau DataFrame pour passer aux étapes de modélisations.

3 - Réalisation

Une Régression

Notre objectif est de prédire un délai d'intervention.
Nous avons dans un premier temps traité ce problème en régression,la variable cible étant une variable quantitative.

- Modèles et optimisation

Le calcul du R2 de la méthode "score" nous a permis d'évaluer l'adéquation entre différents modèles et nos données. En appliquant le R2 sur les jeux d'entraînement et sur les jeux test, nous avons pu écarter les modèles présentant un overfitting.

Nous avons essayé les modèles de base de régression : une Linear Regression, un Decision Tree Regressor, et un Random Forest Regressor. Ces deux derniers présentant un réel overfitting, nous avons conservé le modèle de régression linéaire.
L'affichage de l'intercept et des coefficients des différentes variables sur ce modèle de régression linéaire nous confirme que la variable qui a le plus de poids sur cette régression est la distance, mais la prédiction n'est pas très précise.

Pour améliorer cette précision, nous avons également testé un modèle de Gradient Boosting Regressor puisqu'il combine les prédictions de plusieurs modèles de base.
Il semble plus performant que la régression linéaire sur notre jeu de donné car il présente un score r2 plus élevé (0,51). L'application d'un Feat Importances sur ce modèle a également confirmé le poids de la variable "Distances", mais la réduction du modèle aux 3 variables les plus importantes n'a malheureusement pas permis d'améliorer le score.

L'utilisation des métriques MAE, MSE et RMSE fournit une comparaison entre ces deux modèles sur leur capacité à prédire la variable cible :

- les valeurs sont assez proches entre les ensembles d'entraînement et de test pour les deux modèles, ce qui suggère que les modèles sont adaptés à nos données
- les valeurs absolues des erreurs (MAE) et les erreurs quadratiques (MSE, RMSE) sont plus faibles pour le modèle de Gradient Boosting Regressor, c'est donc ce modèle qui présente les meilleures performances pour notre prédiction.

Pour améliorer ces performances, nous avons optimisé les paramètres en appliquant un GridSeachCV avec une Validation Croisée sur le modèle de Gradient Boosting : un ajustement du nombre d'estimateurs, du taux d'apprentissage et de la profondeur maximale nous a permis d'améliorer le score ainsi que les MAE, MSE et RMSE.

- Interprétation des résultats

Ce modèle de Gradient Boosting Regressor est le plus adapté à nos données et le plus performant lorsqu'on aborde cette prédiction sous forme de régression. Il ne présente pas de surajustement ou de sous-ajustement, la variance et le biais ne sont pas élevés.
En ajustant le choix des colonnes, en fixant des seuils pour les valeurs extrêmes et en éliminant des valeurs aberrantes, nous avons pu constater une amélioration significative du score r2.

La façon d'encoder les variables qualitatives a aussi une influence significative : nous avons choisi un LabelEncoder pour les variables ayant 3 ou moins de 3 valeurs uniques, et un OneHotEncoder pour les autres.

En optimisant les hyperparamètres, nous avons également amélioré les performances du modèle, mais elles restent cependant plutôt modérées.

Une Classification

Nous avons ensuite choisi d'établir deux variables binaires supplémentaires basées sur le délai d'intervention pour pouvoir aborder ce problème sous forme de classification : les délais ont été reclassés en délais inférieurs ou égaux à 250 secondes prenant comme valeur 0 et en délais supérieurs à 250 secondes prenant comme valeur 1. Notre variable cible reste le délai d'intervention.

- Modèles et optimisation

Les classes 0 et 1 étant déséquilibrées, un Standard Scaler a été utilisé pour standardiser les données et les métriques ne devront pas se limiter à l'accuracy.

Trois modèles de classification ont été étudiés: le Decision Tree Classifier, le Random Forest Classifier et la Logistic Regression.
Parmi ces modèles, le modèle Random Forest se distingue. Il a la plus haute précision globale (accuracy) parmi les trois modèles (0.84). Il prédit donc correctement la classe des données dans une plus grande mesure. De plus, il présente la plus haute AUC-ROC (Area Under the Receiver Operating Characteristic Curve): 0.895, ce qui suggère de bonnes performances pour discriminer entre les classes.
Le modèle Random Forest présente également la plus haute valeur AUC-PR (Area Under the Precision-Recall Curve): 0.960. Cela signifie que le modèle est également performant pour la précision du rappel dans les données déséquilibrées.
En examinant les scores F1, qui combinent à la fois la précision et le rappel, le modèle Random Forest présente aussi les meilleurs scores pour les deux classes, ce qui indique un équilibre entre la précision et le rappel.
En ce qui concerne, les coefficient de détermination, ils ne sont pas les métriques les plus appropriées pour l'interprétation de nos modèles. Cependant, on peut noter que le modèle Random Forest a des coefficients de détermination relativement élevés, ce qui peut indiquer une bonne adaptation aux données.
Comme pour le modèle de régression, la variable influençant le plus le modèle est la distance.

- Interprétation des résultats

Le modèle Random Forest Classifier est le modèle le plus performant parmi les différents modèles testés pour estimer le délai d'intervention. Les scores obtenus, élevés, nous montre que notre modèle peut être utilisé pour déterminer pour chaque incident si le délai sera plutôt court (0) ou plutôt long. Basculer d'un modèle de régression à un modèle de classification a donc permis d'obtenir un modèle capable de prédire notre variable cible, le délai d'intervention.

4 - Conclusions

Difficultés

Si l'acquisition des données s'est faite assez facilement grâce aux éléments mis à disposition par la London Fire Brigade, la volumétrie du jeu de données a nécessité un certain temps pour distinguer et identifier correctement ces données, et ensuite pour procéder à un nettoyage précis de ces données.

Certaines variables étant inexploitables, nous avons aussi eu à alimenter notre jeu de données en collectant des données géographiques supplémentaires de façon à pouvoir calculer des variables pertinentes pour notre prédiction.
Nous avions peu travaillé sur des variables géographiques et temporelles lors de la formation, donc nous avons eu à utiliser des bibliothèques que nous n'avions pas vues dans les modules de formation.

La phase d'exploration nous a donc pris beaucoup de temps, et l'organisation d'un travail de groupe dans cette formation 100% distancielle n'a pas été évidente et a retardé le début de la phase de réalisation des modèles.

Pour la partie modélisations, nous avons peu abordé les méthodes d'évaluation et d'optimisation des hyperparamètres des modèles lors du cursus de formation, et nous n'avons pas vu les techniques d'interprétabilité.
Nous avons su entraîner facilement les modèles de base, mais l'approfondissement a été plus difficile à aborder, et les performances obtenues ne sont pas nécessairement celles escomptées.

La préparation de la soutenance a également nécessité un temps et un investissement non négligeables, notamment pour la préparation d'un Streamlit qui n'est pas évidente sans formation préalable.

Enfin, l'utilisation de Google Collab pour pouvoir travailler sur un espace partagé a aussi posé quelques difficultés. Nous avions beaucoup de données géographiques et temporelles, et ces données sont visiblement lourdes à traiter : pas assez de RAM pour afficher des cartes complexes, des temps de calculs très longs pour les distances, un entraînement des modèles et une optimisation des paramètres assez longs.

Bilan et suite du projet

Dans le cadre de ce projet, notre contribution principale a été de développer et d'optimiser le modèle de prédiction des temps d'interventions en utilisant les distances entre les casernes et les lieux d'intervention.

Nous avons abordé ce problème sous deux angles : d'abord une régression en utilisant nos variables explicatives, dont la distance calculée, pour prédire notre variable cible : le délai d'intervention.

En observant la relation distance/délai d'intervention, nous avons ensuite abordé ce problème sous forme de classification en séparant les délais d'intervention en deux variables binaires : délais courts et délais longs.

Cette approche a permis d'améliorer la performance des modèles ainsi que la précision des prédictions.

Compte tenu du volume du jeu de données et du temps imparti pour mener à bien cette analyse, nous nous somme concentrés sur certaines données géographiques , mais nous n'avons pas pu explorer d'autres pistes comme :

- Le temps passé par la brigade sur le lieux d'interventions : cette variable a un rôle sur les disponibilités des ressources mobilisées, aussi bien humaines que matérielles

- Les conditions de circulation selon les jours et les heures de la semaine, qui impactent directement le temps de trajet

- Les ressources disponibles des casernes

Annexes

- Notebook "Pompier.ipynb" : Phase exploratoire du projet
- Les 2 fichiers csv utilisés dans ce notebook : "LFB Incident data" et "LFB_Mobilisation_data"

- Notebook "Modelisation_Regression.ipynb" : Étude en régression
- Notebook "Modelisation_classification.ipynb" : Étude en classification
- Le fichier csv utilisé pour le modélisations :"df6c"