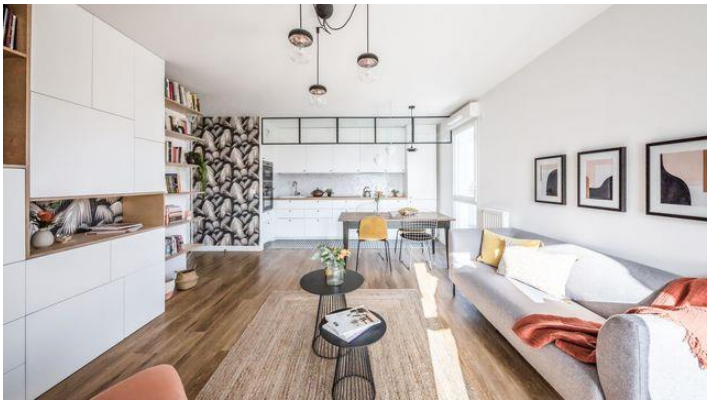
ML Project :

Lors de notre deuxième année en école d'ingénieurs, nous avons été amenés à découvrir un nouveau domaine : l'Intelligence Artificielle, et plus précisément le Machine Learning. A l'issu de cette année, nous devions remettre un projet de notre choix utilisant un dataset pour prédire quelque chose.

Ce projet devait être fait en groupe, ce qui ne nous a posé aucun problème, mon binôme et moi, car nous avons pris l'habitude de travailler ensemble. Cependant, le choix du sujet a été le premier problème rencontré dans ce long périple. En effet, le fait d'avoir le choix de notre sujet nous a amené à élargir nos horizons et sortir d'un projet qui est forcément en relation avec les sciences. Afin de trouver un sujet inspirant qui nous permettrait de s'accrocher même dans les moments les plus difficiles (car oui, il est facile de décrocher très rapidement !). Nous avons donc essayé de récolter des informations relatives aux problèmes que nous rencontrons dans notre quotidien. Nous avons commencé par nous poser la question à nous-même. En vain, aucun problème sérieux n'est ressorti de notre « brainstorming ». Nous nous sommes donc orientés vers les réseaux sociaux et notamment Facebook pour avoir plus de problématiques à étudier. Aucunes n'a retenu notre attention.

Cependant, une idée nous est venue à l'esprit grâce à un membre de notre entourage qui possède une startup dans la location courts et moyens termes. La startup loue des appartements à Londres afin de les reconditionner et de les relouer à d'autres personnes en facilitant tout le process en adaptant une technologie adéquate. Leur principale mission est de trouver l'appartement au meilleur prix afin de pouvoir augmenter leur marge.



En ce sens, nous avons donc organisé des appels téléphoniques afin de pouvoir poser quelques questions et se renseigner un peu plus sur la méthode de travail pour pouvoir proposer une solution adaptable à leur besoin.

HOSTELLAR

Grâce aux nombreux échanges, nous avons pu définir des objectifs et savoir comment l'on pouvait les aider. Notre connaissance (Raouf) et son associé (Michael) nous ont conseillé des sites pour avoir des dataset afin que l'on puisse entraîner notre machine. Nous nous sommes donc orientés vers Rightmove et Zoopla.

De ce fait, nous avons donc chercher où est-ce que l'on pouvait récupérer le dataset de Rightmove ou bien de Zoopla. Nous n'avons pas trouvé de dataset préétabli. Raouf, co-fondateur de la startup, nous a proposé une URL qui montre comment faire du web scraping. Nous avons utilisé les des fonctions python permettant de récolter les données du site Rightmove. Cependant, nous avons rapidement rencontrer un nouveau problème : Rightmove n'autorise qu'une récolte de 1050 données. Nous avons essayé de contourner ceci mais nous ne sommes pas parvenus car nous n'avions pas le temps et les compétences nécessaires pour faire cela. Michael et Raouf nous ont donc orienté vers un dataset de Airbnb.

Une fois le dataset trouvé, nous avons commencé à faire du data cleaning afin d'enlever les données non nécessaires pour entraîner la machine. Nous avions aussi créé de nouveaux tableaux à partir de deux fichier csv différents contenant les données pour chaque jour qui nous permettait d'avoir une variable prix pour chaque saison et le taux d'occupation de l'appartement durant les derniers mois.

Lorsque nous avons essayé de trouver des corrélations entre les variables et entraîner notre machine à partir de ces données, nous nous sommes rendu compte que la variance était minime et que la machine n'arrivait à bien saisir les données et se les approprier.

Après de nombreuses tentatives sans succès, nous nous sommes résiliés à dévier un peu de notre projet initial afin d'avoir un rendu pour notre soutenance et avions mis en standby le projet initial. Ce nouveau projet donne une prédiction sur le prix d'un appartement à l'achat. L'objectif principal de notre machine est d'aider les personnes, souhaitant vendre ou acheté un appartement, à prédire un prix de l'achat en fonction de ce qui se trouve sur le marché immobilier à Londres.

House Pricing

Pour ce faire, nous avons trouvé un nouveau dataset sur internet. Nous avons procédé de la même manière que pour le projet initial. Nous avons donc fait du data cleaning sur les 81 colonnes. Ensuite, nous avons affiché les cellules vides à l'aide de la librairie ***missingno*** qui permet d'avoir un visuel sur les données manquantes. De plus, afin d'avoir un pourcentage de données manquantes nous avons utilisé une fonction qui prend en entrée un tableau et retourne deux colonnes : le total de missing value par colonne et un pourcentage pour chaque colonne.

Afin de bien choisir nos variables avant de supprimer des colonnes, nous avons décidé de tracer des courbes en fonction du prix des appartements. Cela nous a permis de voir les corrélations entre le prix et différentes variables. Nous avons constaté que certaines courbes étaient linéaires et donc présentaient une forte corrélation. Pour affiner plus notre recherche de dépendance, nous avons aussi utiliser la fonction heatmap, qui nous a permis d'avoir une vue d'ensemble sur la dépendance des données au prix. A l'issue de cette étape, nous avons supprimé 71 colonnes pour alléger les inputs lors de la saisie de données dans l'application web.

Finalement, afin d'entrainer notre machine, nous avons utilisé la « Linear Regression ». Nous obtenons un train score de **76.63 %** et un test score de **77.52%**.