Plan oral Défi IA

* Présentation du contexte : compétition Kaggle, création de notre propre dataset grâce à un système de requêtes en fonction de différents avatar (différents langages, depuis un téléphone ou un ordi), prédire prix d’une nuit d’hôtel (différentes villes).
* Analyse descriptive : visualisation du jeu de données, de la distribution des variables, différents traitements en fonction variable qualitative/quantitative, étude corrélation entre les variables.
* Stratégie :
  + En premier lieu : on fixe une langue pour un avatar puis pour le reste des variables (ville, mobile, date) on choisit de manière random.
  + Ensuite, on observe la distribution des variables de notre datatrain et on compare à la distribution des variables dans le datatest et on fait nos requêtes en fonction de la différence entre les distributions afin de s’approcher au mieux de la distribution du datatest et ainsi augmenter notre score. On remarque une grosse proportion de language=hungarian dans le datatest et une distribution des dates assez singulière.
  + Test d’interprétabilité : difficile à lire et à interpréter.
  + Pour encoder les variables qualitatives : target encoding (on remplace valeur qualitative par la moyenne des valeurs de la variable cible correspondant à la valeur de notre variable qualitative.)
  + Ajout d’une variable index\_request afin de prendre en compte le nombre de requêtes effectuée par un même avatar. (Généralement augmentation des prix quand on a fait plusieurs requêtes pour une même ville)
  + Prédiction du log-price plutôt que du price.
  + Test de différents modèles et optimisation du random forest.
  + Partie submission
  + Création d’un DockerFile et d’une application gradio