Plan de travail prévisionnel

I. Motivation

Lors de ce projet de preuve de concept, je souhaite travailler sur des données tabulées et plus ou moins structurées afin d'essayer d'améliorer les performances accessibles grâce à des modèles « classiques » comme les modèles linéaires, les méthodes ensemblistes ou les arbres de décisions. Des recherches bibliographiques m'ont permis de découvrir que dernièrement des équipes utilisent des Transformers sur des données structurées afin d'améliorer les performances.[1], [2] Ces articles sont publiés à partir de fin 2020, il s'agit donc d'une nouvelle approche. Le principe consiste à appliquer des méthodes de traitements de données initialement prévues pour les données textuelles à des données tabulées et structurées. Dans ces articles les bases de données utilisées sont le plus souvent des bases de tests et très structurées ce qui favorise donc la performance du modèle. Mon projet serait donc d'utiliser cette méthode sur des données plus réelles, moins structurées afin de pouvoir analyser la performance sur ce type de données.

II. Choix de la base de données

Pour utiliser des données les plus réelles possibles, plusieurs bases de données sur recherchées sur le site data.gouv.fr. Le thème choisi est la prédiction du prix de l'immobilier, pour se rapprocher de la compétition Kaggle connue. Les bases sont filtrées en fonction de la provenance afin de s'assurer au maximum de la qualité de celles-ci : on favorisera au maximum des données publiées par un ministère ou une institution publique.

Les informations sur les mutations immobilières depuis 2017 sont disponibles sur tout le territoire français. On récupère également une base de données sur les communes pour avoir des informations pouvant influer sur la valeur immobilière. Il est possible également d'obtenir la liste de toutes les écoles afin de calculer la distance entre chaque bien et l'école la plus proche. On applique la même démarche pour les transports en communs pour déterminer la distance aux transports. Ces informations vont permettre de construire une base de données réelles, avec uniquement des données publiques disponibles en France.

III. Planning prévisionnel

Ce projet se décompose en plusieurs étapes, allant de la collecte des données jusqu'à la création de plusieurs modèles de prédictions. Pour chaque étape je prévois :

- Collecte et nettoyage des bases de données : 1 semaine
- Création de nouvelles features et compilation de toutes les données : 3 jours
- Développement des modèles (baseline, modèles classiques, Tab Transformers) : 1 semaine
- Rédaction des livrables : 3 jours

Cela représente au total environ 3 semaines de travail. Cela dépasse le temps estimé par OpenClassRooms pour le projet, mais il me semble nécessaire d'augmenter ce temps pour tester correctement cette méthode.

IV. Références

[1] X. Huang, A. Khetan, M. Cvitkovic, et Z. Karnin, « Tabtransformer: Tabular data modeling using contextual embeddings », *ArXiv Prepr. ArXiv201206678*, 2020.

[2] R. Cholakov et T. Kolev, « The GatedTabTransformer. An enhanced deep learning architecture for tabular modeling », *ArXiv Prepr. ArXiv220100199*, 2022.