Hackathon: Prédiction des Prix Immobiliers à Ames, Iowa

Contexte

Lorsque les acheteurs décrivent leur maison de rêve, ils se concentrent souvent sur des aspects comme le nombre de chambres ou la présence d'un jardin. Cependant, de nombreux facteurs cachés, tels que la hauteur du sous-sol ou la proximité d'une voie ferrée, influencent également le prix final de vente.

Ce hackathon met au défi les participants de **prédire les prix de vente des maisons à Ames, Iowa**, en utilisant un jeu de données contenant **79 variables explicatives** décrivant divers aspects des propriétés résidentielles. C'est une excellente opportunité pour les passionnés de **data science** et de **machine learning** d'appliquer leurs compétences et d'explorer des techniques avancées de régression.

Vous aurez accès à des **données fournies par Kaggle**, garantissant un jeu de données de haute qualité et représentatif du monde réel pour l'analyse et la modélisation.

Remerciements

Le jeu de données Ames Housing a été compilé par **Dean De Cock** à des fins éducatives en data science. Il constitue une alternative moderne et enrichie au célèbre jeu de données **Boston Housing**, souvent utilisé en analyse de prix immobiliers.

Objectifs du Hackathon

- **Prédire avec précision** le prix de vente des maisons en fonction des caractéristiques détaillées des biens immobiliers.
- **Expérimenter avec des modèles de machine learning avancés** comme *Random Forest, Gradient Boosting, XGBoost et LightGBM.*
- **Appliquer des techniques d'ingénierie des caractéristiques (feature engineering)** pour améliorer la performance des modèles.
- **Effectuer une analyse exploratoire des données (EDA)** afin d'identifier les tendances, les valeurs aberrantes et les principaux facteurs influents.

Critères d'Évaluation

Les soumissions seront évaluées selon les critères suivants :

1. **Présentation (30%)**
    - Clarté et efficacité du **pitch final**.
    - Capacité à expliquer les **insights**, défis rencontrés et conclusions.

2. **Performance du modèle (30%)**

    - Évaluée à l'aide de **l'Erreur Quadratique Moyenne (RMSE)** entre le **logarithme** du prix prédit et le **logarithme** du prix réel.

3. **Qualité de l'Analyse Exploratoire des Données (EDA) (20%)**

    - Analyse approfondie des variables influentes.
    - Visualisation des relations entre les variables.

- Détection et gestion des **valeurs manquantes et aberrantes**.

4. **Feature Engineering et Justification (15%)**

   - Création de **nouvelles variables pertinentes**.
   - Sélection des **meilleures caractéristiques** pour améliorer la précision du modèle.

5. **Qualité du Code & Documentation (5%)**

   - Code **bien structuré** et **commenté**.
   - Présentation **claire** des résultats et des insights.

Un **Notebook Jupyter** (ou équivalent) devra être soumis, détaillant l'approche suivie, y compris l'EDA, la sélection des caractéristiques, les modèles testés et leurs justifications.

Vos prédictions pour les données test.csv doivent être enregistrer dans le fichier **sample_submission.csv** sous la forme :

Id, SalePrice

1461, 169000.1

1462, 187724.1233

1463, 175221

...

Bibliothèques et Environnement

- **Seules des bibliothèques open-source sont autorisées**.
- Toutes les dépendances doivent être listées dans un fichier **requirements.txt** ou **environment.yml** .
- La solution soumise doit être **facilement reproductible** à l'aide du fichier d'environnement fourni.

Déroulement du Hackathon

- **Début** : **Mercredi 26 Février 2025 à 10h00**
- **Dernière soumission** : **Vendredi 28 Février 2025 à 10h00**

Ressources:

https://www.youtube.com/watch?v=82KLS2C_gNQ&list=PLO_fdPEVlfKqMDNmCFzQISI2H_nJcEDJq

https://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire

https://fr.wikipedia.org/wiki/Science_des_donn%C3%A9es

https://en.wikipedia.org/wiki/Exploratory_data_analysis