TP IFRISSE – Rodrique K

TP Science de Données IFRISSE 2020

Les données utilisées dans ce TP sont en relation avec les logements pour habitation. Elles décrivent de prime à bord, que le choix d'un logement par une personne dépend fortement de son prix que les autres variables telles que le nombre de chambres, les dimensions, etc.

La base de données est composée de 79 variables explicatives décrivant (presque) tous les aspects des maisons résidentielles à Ames.

Objectif Principal
L'objectif recherché dans cette étude, est de proposer des modèles de ML qui vous permettront de prédire avec la plus grande précision possible, le prix de chaque maison, en tenant compte des autres paramètres qui la composent. Ceci en mettant

Les données
Trois fichiers sont disponibles. Train.csv contenant les données d'entraînement, Test.csv contenantles données de Test et Description.txt qui décrit les différentes colonnes.

Compétence Pratiques
Ingénierie des Features
Techniques de régression avancées

Démarche à suivre

1. Nettoyage et formatage des données
2. Analyse exploratoire des données
3. Ingénierie et sélection de variables caractéristiques
4. Construction de modèles et Évaluation
    - Évaluer le meilleur modèle en vous basant sur la RMSE (l'erreur quadratique moyenne) (obligatoire)
    - Vous pouvez par la suite utiliser le.s metric.S de votre choix (précision, rappel, auc, etc.)

Étape 1 : Prétraitement des données et ingénierie des variables caractéristiques

1. Types de variables et valeurs manquantes
2. Analyse exploratoire des données
    - Variables numériques
    - Variables catégorielles
    - Corrélations entre variables numériques et la variable cible (prix de vente)

- Corrélations entre variables catégorielles et la variable cible (prix de vente)

3. Ingénierie et sélection des caractéristiques (feature ingineering)
    - Transformation logarithmique des variables numériques
    - Application d'un Hot-Encoding aux variables catégorielles
    - Concaténer les données de training et de test
    - Séparé en deux ensembles, données d'entraînement et de test (80 % et 20 %)
    - Imputation des valeurs manquantes

Étape 2 : Construction du modèle

1. Concaténer les données du train et du test et utiliser 80 % des données concaténées pour l'entrainement (*vous pouvez utiliser la fonction* ***train_test_split*** *de* ***scikitlearn*** *à cet effet*)
2. Choisissez au moins 3 modèles d'apprentissages différents
3. Évaluer chaque modèle en **1)** utilisant les paramètres par défaut et **2)** en utilisant la stratégie GridSearchCV afin de définir les paramètres essentiels.
4. Évaluer chaque modèle avec la métrique de votre choix et le justifier.

Rendu et dates

- Chacun devrait fournir son rendu soit 1) un fichier jupyter-notebook avec les commentaires et explications de resultats, soit 2) deux fichiers séparés, un contenant le code, et l'autre un rapport décrivant les étapes et résultats obtenus.
- Vous avez deux semaines à compté de la date du 18/12.

Remarque : Vous n'êtes pas obligé de suivre la démarche proposée ci-haut. Chacun est libre d'apporter d'autres types d'analyses, mais devrait préciser le pourquoi.