`

Conservatoire national des arts et métiers (CNAM)

Plan du sujet du

Projet STA201- Analyse Multivariée Approfondie

Nom du projet

---« Techniques Avancées de Régression pour

La Prévision des Prix Immobiliers »---

Présenté par :

----« Elio Bou Serhal »----

Professeur responsable :

---« Dr Zainab Assaghir »---

« La prédiction des prix immobiliers »

Introduction :

L'apprentissage automatique (Machine Learning : ML) est un sous-domaine de la fouille de données qui étudie les techniques automatiques pour apprendre à faire des prévisions précises en se basant sur des observations passées. Il utilise deux types de techniques :

1. L'apprentissage supervisé : (classification et régression), qui entraîne un modèle sur des données d'entrée et de sortie connues pour qu'il puisse prédire les sorties.
2. L'apprentissage non supervisé : (clustering), qui trouve des motifs cachés ou des structures intrinsèques dans les données d'entrée.

Ce qui nous intéresse dans ce projet, une base de données supervisée qui contient des informations sur **2 921 maisons** dans une province américaine Ames, Lowa (United state of America) et **81 variables** enregistrées pour chaque observation qui comprennent des informations démographiques sur les maisons, telles que :

1. **Id** : Identifiant pour chaque maison.
2. **MSSubClass, MSZoning, LotFrontage, LotArea** : Caractéristiques de la propriété et du zonage.
3. **Street, Alley, LotShape, LandContour** : Détails sur le terrain et les environs.
4. **Utilities, LotConfig, LandSlope** : Détails sur les services publics et la configuration.
5. **Neighborhood, Condition1, Condition2** : Quartier et proximité de diverses conditions.
6. **BldgType, HouseStyle** : Informations sur le type de bâtiment et le style de la maison.
7. **OverallQual, OverallCond, YearBuilt, YearRemodAdd** : Détails sur la qualité, l'état et l'année de construction.
8. **RoofStyle, RoofMatl, Exterior1st, Exterior2nd** : Matériaux du toit et de l'extérieur.
9. **MasVnrType, MasVnrArea** : Type et surface de parement en maçonnerie.
10. **ExterQual, ExterCond, Foundation** : Qualité extérieure et type de fondation.
11. **BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1** : Détails sur le sous-sol.
12. **BsmtFinType2, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF** : Autres détails sur le sous-sol.
13. **Heating, HeatingQC, CentralAir, Electrical** : Détails sur le chauffage et l'électricité.
14. **X1stFlrSF, X2ndFlrSF, LowQualFinSF, GrLivArea** : Détails sur la surface des étages.
15. **BsmtFullBath, BsmtHalfBath, FullBath, HalfBath** : Nombre de salles de bains.
16. **BedroomAbvGr, KitchenAbvGr, KitchenQual, TotRmsAbvGrd** : Détails sur les chambres et la cuisine.
17. **Functional, Fireplaces, FireplaceQu** : Détails fonctionnels et sur la cheminée.
18. **GarageType, GarageYrBlt, GarageFinish, GarageCars, GarageArea** : Détails sur le garage.
19. **GarageQual, GarageCond, PavedDrive** : Plus de détails sur le garage et l'allée.
20. **WoodDeckSF, OpenPorchSF, EnclosedPorch, X3SsnPorch, ScreenPorch** : Détails sur les porches et les terrasses.
21. **PoolArea, PoolQC, Fence, MiscFeature, MiscVal** : Piscine, clôture et caractéristiques diverses.
22. **MoSold, YrSold, SaleType, SaleCondition, SalePrice** : Détails des ventes.
23. **SalePrice** : le prix de chaque maison. (Variable cible)

On peut remarquer que l'importance des caractéristiques et les conséquences peuvent varier d'un modèle à un autre. Ce projet vise à analyser la performance de **3 modèles de Régression Linéaire** et **3 modèles de régularisations**, en choisissant les paramètres optimaux afin d'évaluer l'approche proposée.

1. Linear Regression (Forward, Backward et Step Wise)
2. Principal Component Regression (PCR)
3. Partial Least Squares Regression (PLS)
4. Ridge Regression (modèle 1 de régularisations)
5. LASSO Regression (modèle 2 de régularisations)
6. Elastic Net (modèle 3 de régularisations)

Objectif :

Notre objectif est de construire des modèles prédictifs dans le but de déterminer les caractéristiques les plus susceptibles pour estimer le prix d'immobilier en se basant sur les variables explicatives.

Étape à faire :

ÉTAPE 1 : « Importer la base de données »

1. Importer la base de données sur R-studio.
2. Importer les librairies nécessaires.

ÉTAPE 2 : « Analyse Descriptive »

1. Effectuer une analyse exploratoire pour comprendre les caractéristiques et la structure de notre base de données.
2. Détecter s'il existe des valeurs manquantes et si non, on doit remplacer quelques valeurs par « NA » pour les visualiser et les imputer en utilisant différentes méthodes de visualisation (histogramme, pattern…) et d'imputation (simple, multiple…) Pour pouvoir comparer la distribution des variables contenant ces valeurs manquantes avant et après imputation.

ÉTAPE 3 : « Prétraitement de la base de données »

1. Eliminer toutes les valeurs manquantes.
2. Faire une analyse descriptive de la base de données nettoyée.
3. Représenter graphiquement de chaque variable.
4. Explication théorique sur les conditions de la régression.
5. Vérification des conditions et application (Linéarité, Homoscédasticité ou Hétéroscédasticité, Normalité et Multi-colinéarité).

ÉTAPE 4 : « Transformation et Partitionnement de la base de données »

1. Transformer les variables qualitatives en variables quantitatives.
2. Diviser la base de données en ensemble d'entraînement et ensemble de test pour tester les performances de chaque modèle.

ÉTAPE 5 : « Application des modèles de régressions avancés »

1. Entraîner les 6 modèles différents en utilisant les données d'entraînement nettoyée.
2. Évaluer leur performance en utilisant deux mesures métriques RMSE et R^2.

ÉTAPE 7 : « Comparaison des résultats obtenus »

1. Créer un tableau contenant tous les résultats (Mesure métriques et Temps d'exécution).
2. Représenter graphiquement les mesures métriques.
3. Interpréter des résultats obtenus.
4. Indiquer les hyper paramètres du modèle le plus performant.
5. Tirer des conclusions sur les caractéristiques les plus importantes pour prédire le prix d'immobilier.

Références :

1. https://www.kaggle.com/code/apapiu/regularized-linear-models
2. https://www.kaggle.com/code/pmarcelino/comprehensive-data-exploration-with-python
3. https://www.kaggle.com/code/masumrumi/a-detailed-regression-guide-with-house-pricing#Fitting-model(simple-approach)
4. https://www.kaggle.com/code/gusthema/house-prices-prediction-using-tfdf
5. https://www.kaggle.com/code/pmarcelino/data-analysis-and-feature-extraction-with-python
6. https://www.kaggle.com/c/house-prices-advanced-regression-techniques