Concours de Régression Linéaire (House Prices - Advanced Regression Techniques - Kaggle)

Ismael Assani

STT2400 - Eté 2024

Contexte

Demandez à un particulier de décrire sa maison de rêve, et il ne commencera probablement pas par la hauteur du plafond du sous-sol ou la proximité d'une voie ferrée ou d'une plage. Mais l'ensemble de données de cette compétition prouve que bien plus de facteurs influencent les négociations de prix autant que le nombre de chambres ou une clôture blanche.

Avec 79 variables explicatives décrivant (presque) tous les aspects des maisons résidentielles à Ames en Iowa, cette compétition vous met au défi de prédire le prix final de chaque maison d'une part, et de déterminer les variables qui contribuent significativement à expliquer le prix des maisons.

Objectif

Votre mission, si vous l'acceptez, est de prédire d'une part le prix de vente de chaque maison, mais aussi de déterminer les variables qui expliquent ce prix en utilisant uniquement les méthodes apprises dans ce cours de régression linéaire.

Ensemble de Données

L'ensemble de données de vente de maisons à Ames en Iowa disponible sur Studium comprend les caractéristiques suivantes :

- **SalePrice** : Prix de vente de la propriété en dollars. C'est la variable cible que vous devez prédire.
- **MSSubClass** : Classe du bâtiment.
- **MSZoning** : Classification générale de la zone.
- **LotFrontage** : Pieds linéaires de rue connectés à la propriété.
- **LotArea** : Taille du terrain en pieds carrés.
- **Street** : Type d'accès routier.

- **Alley** : Type d'accès à l'allée.
- **LotShape** : Forme générale de la propriété.
- **LandContour** : Planéité de la propriété.
- **Utilities** : Type de services publics disponibles.
- **LotConfig** : Configuration du terrain.
- **LandSlope** : Pente de la propriété.
- **Neighborhood** : Emplacements physiques dans les limites de la ville d'Ames.
- **Condition1** : Proximité de la route principale ou du chemin de fer.
- **Condition2** : Proximité de la route principale ou du chemin de fer (si un second est présent).
- **BldgType** : Type de logement.
- **HouseStyle** : Style de logement.
- **OverallQual** : Qualité générale des matériaux et des finitions.
- **OverallCond** : Évaluation de l'état général.
- **YearBuilt** : Date de construction initiale.
- **YearRemodAdd** : Date de rénovation.
- **RoofStyle** : Type de toit.
- **RoofMatl** : Matériau du toit.
- **Exterior1st** : Revêtement extérieur sur la maison.
- **Exterior2nd** : Revêtement extérieur sur la maison (s'il y a plus d'un matériau).
- **MasVnrType** : Type de placage en maçonnerie.
- **MasVnrArea** : Surface du placage en maçonnerie en pieds carrés.
- **ExterQual** : Qualité du matériau extérieur.
- **ExterCond** : État actuel du matériau extérieur.
- **Foundation** : Type de fondation.
- **BsmtQual** : Hauteur du sous-sol.
- **BsmtCond** : État général du sous-sol.
- **BsmtExposure** : Murs de sous-sol de type walkout ou jardin.
- **BsmtFinType1** : Qualité de la zone finie du sous-sol.

- **BsmtFinSF1** : Pieds carrés finis de type 1.
- **BsmtFinType2** : Qualité de la deuxième zone finie (si présente).
- **BsmtFinSF2** : Pieds carrés finis de type 2.
- **BsmtUnfSF** : Pieds carrés non finis du sous-sol.
- **TotalBsmtSF** : Superficie totale du sous-sol en pieds carrés.
- **Heating** : Type de chauffage.
- **HeatingQC** : Qualité et état du chauffage.
- **CentralAir** : Climatisation centrale.
- **Electrical** : Système électrique.
- **1stFlrSF** : Superficie du premier étage en pieds carrés.
- **2ndFlrSF** : Superficie du deuxième étage en pieds carrés.
- **LowQualFinSF** : Superficie finie de faible qualité (tous les étages).
- **GrLivArea** : Superficie habitable au-dessus du sol en pieds carrés.
- **BsmtFullBath** : Salles de bains complètes au sous-sol.
- **BsmtHalfBath** : Salles de bains demi-complètes au sous-sol.
- **FullBath** : Salles de bains complètes au-dessus du sol.
- **HalfBath** : Salles de bains demi-complètes au-dessus du sol.
- **Bedroom** : Nombre de chambres au-dessus du niveau du sous-sol.
- **Kitchen** : Nombre de cuisines.
- **KitchenQual** : Qualité de la cuisine.
- **TotRmsAbvGrd** : Nombre total de pièces au-dessus du sol (ne comprend pas les salles de bains).
- **Functional** : Évaluation de la fonctionnalité de la maison.
- **Fireplaces** : Nombre de cheminées.
- **FireplaceQu** : Qualité de la cheminée.
- **GarageType** : Emplacement du garage.
- **GarageYrBlt** : Année de construction du garage.
- **GarageFinish** : Finition intérieure du garage.
- **GarageCars** : Capacité du garage en nombre de voitures.

- **GarageArea** : Superficie du garage en pieds carrés.
- **GarageQual** : Qualité du garage.
- **GarageCond** : État du garage.
- **PavedDrive** : Allée pavée.
- **WoodDeckSF** : Superficie de la terrasse en bois en pieds carrés.
- **OpenPorchSF** : Superficie de la véranda ouverte en pieds carrés.
- **EnclosedPorch** : Superficie de la véranda fermée en pieds carrés.
- **3SsnPorch** : Superficie de la véranda trois saisons en pieds carrés.
- **ScreenPorch** : Superficie de la véranda avec moustiquaire en pieds carrés.
- **PoolArea** : Superficie de la piscine en pieds carrés.
- **PoolQC** : Qualité de la piscine.
- **Fence** : Qualité de la clôture.
- **MiscFeature** : Caractéristique diverse non couverte dans d'autres catégories.
- **MiscVal** : Valeur de la caractéristique diverse.
- **MoSold** : Mois de vente.
- **YrSold** : Année de vente.
- **SaleType** : Type de vente.
- **SaleCondition** : Condition de vente.

Tâches recommandées

Exploration des Données (30 minutes)

- Bien comprendre le problème.
- Charger l'ensemble de données.
- Diviser l'ensemble de données en deux parties : "train" et "test" (par exemple 80% vs 20%). Utilisez la partie "train" pour construire vos modèles et la partie "test" pour évaluer leurs performances.
- Explorer les statistiques de base sur les variables (moyenne, variance, médiane, données manquantes, valeurs aberrantes, etc.).
- Analyser la normalité de la variable dépendante. A-t-on besoin de faire une transformation ? (log, exponentielle, boxcox etc.).
- Visualiser les relations entre les caractéristiques et la variable cible.

Construction du Modèle (1 heure)

- Construire un modèle de régression linéaire multiple initial en utilisant toutes les caractéristiques retenues.
- Est-il nécessaire de transformer certaines variables explicatives ?
- Vérifier la multicolinéarité entre les prédicteurs et/ou utiliser une ou plusieurs procédures de choix de modèle. (Attention ! certaines procédures peuvent demander beaucoup de temps.)
- Vérifier les résidus pour détecter tout motif suggérant des violations des hypothèses de régression dans vos modèles candidats (e.g., hétéroscédasticité, normalité, valeurs influentes).

Évaluation du Modèle (30 minutes)

- Évaluer le modèle en utilisant le RMSE sur la base *test* pour choisir votre modèle.

Soumission

- J'ai une autre base test avec moi et j'évaluerai le RMSE de votre modèle sur cette base test. Je vais également analyser la significativité de vos variables retenues comme explicatives du prix lorsque j'évaluerai le modèle sur la base test. Le groupe qui gagnera sera celui qui aura le RMSE le plus petit et dont toutes les variables (désignées) seront significatives sur ma base test.

Bonne chance !