Données du projet d'estimation des prix d'immobilier

GLO-7027 Analyse et traitement de données massives

Ali ASSAFIRI 111 054 128
Rhita OULIZ 111 082 917

14 février 2019

1 Introduction

Dans le cadre de ce cours, nous sommes mandatés à choisir et à réaliser un projet d'analyse et traitement de données massives de *Kaggle*. Toutefois, en fonction de l'intérêt commun de notre équipe, le projet *Kaggle* en question est intitulé "*House Prices : Advanced Regression Techniques* ".

1.1 Description

Le but de cette étude est l'amélioration de l'évaluation des valeurs d'immobilier dans la ville d'Ames aux États-Unis. Des études ont été réalisées afin de trouver une méthode alternative de prédiction de prix d'immobilier, autre que les méthodes conventionnelles. Pour ce faire, plusieurs variables doivent être prises en compte d'où la pertinence de l'analyse.
Les données pour cette étude comparent 79 variables pour 2919 échantillons dont 1460 immobiliers sont avec le prix d'achat connu. Les attributs de ces données comprennent les caractéristiques des immobiliers en termes de forme et qualité, l'an de construction et maintenance de ces composantes ainsi que leur localisation.

1.2 Objectifs

L'objectif principal de ce projet est la prédiction des prix de vente des immobiliers selon leur caractéristiques et leurs historiques. Et ce, dans le but de déterminer une méthode d'évaluation plus performante.
Cette étude comprend plus précisément trois objectifs :

1. Démystifier les données afin de déterminer les attributs qui influencent le plus la valeur d'immobilier.
2. Concevoir deux algorithmes de prédiction en utilisant les méthodes d'apprentissage automatique les plus efficaces pour ce sujet.
3. Établir une étude comparative des résultats des deux algorithmes. Et représenter des concussions sur les prix d'immobilier à Ames en Iowa.

1.3 Métrique de réussite

La première validation à faire est la soumission de nos valeurs prédites sur le site du *Kaggle* pour évaluer le taux de précision obtenu par nos modèles en se basant sur l'erreur quadratique moyenne (RMSE).
Par contre, cette évaluation ne nous informe point sur la robustesse de notre solution ni sur les variables et les méthodes à optimiser. À cet effet, avoir un graphe sur le taux d'apprentissage en fonction du *batch size* et le nombre d'*époch* sera utile pour détecter la présence de surapprentissage des modelés.
En fonction des résultats observés et pour donner suite aux deux étapes décrites ci-haut, nous serons en mesure de critiquer la méthode utilisée. Cependant, il est évident qu'une analyse logique sur les prix prédits doit être effectuée en prenant en compte le résultat des autres compétiteurs. Mais aussi, la revue de littérature doit être prise en considération.

1.4 Méthodologie

Cette étude comprendra l'exploration des données, la sélection des variables pertinentes, l'estimation des prix des immobiliers avec deux méthodes de prédiction et finalement une étude comparative des résultats. Ainsi, les trois principales méthodologies pour rependre aux objectifs du projet sont les suivantes :

— **La méthodologie pour l'objectif 1 :**
Pour avoir une intuition sur les algorithmes à étudier pour ce cas d'étude, nous explorerons les attributs des immobiliers en prenant en compte leurs distributions statistiques. Par la suite, nous déterminerons les anomalies (données manquantes et aberrantes) au niveau de données afin de sélectionner les algorithmes de prétraitement à utiliser pour les traiter.

— **La méthodologie pour l'objectif 2 :**
Nous allons étudier non seulement les données mais aussi les algorithmes de prédictions dans l'objectif d'en choisir que deux méthodes. Dans un premier temps, une méthode parmi *Random Forest* ou *Xgboost* sera choisie pour la première méthode. Par la suite, une architecture de réseaux de neurones sera à déterminer pour la deuxième méthode.

— **La méthodologie pour l'objectif 3 :**
Pour une étude comparative des deux méthodes, nous établirons une fonction de propagation d'erreur, évaluerons la justesse de résultats de chacune des méthodes pour en tirer une conclusion.

2 Les attributs

2.1 Les attributs

2.2 Les types d'attributs

2.3 Les attributs redondants

2.4 Nettoyage

3 Statistiques de base

3.1 Tendence centrale d'attributs

3.2 Dispersion des attributs

4 Étude des corrélations

4.1 observations

4.2 conclusion

5 Visualisation des données

6 Similarité et dissimilarité

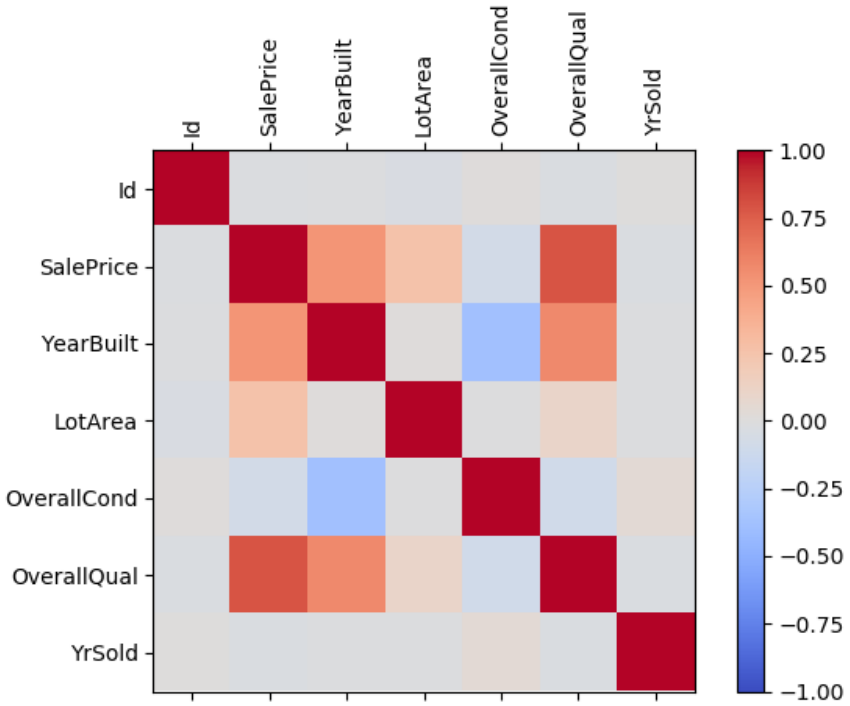

FIGURE 1 – Corrélations