

**Rapport**

*Projet 8 IML*

Participez à une compétition Kaggle !



***Thibault Roussel***

***SOMMAIRE***

1. Introduction
2. Compétition choisie
3. Cleaning effectué
4. Exploration des données
5. Modélisation
6. Modèle final
7. Résultats
8. Conclusion
9. Introduction

Créée en avril 2010 à San Francisco par Anthony Goldbloom, Kaggle est une plateforme qui organise des compétitions en data science et qui récompense les meilleurs analystes internationaux.

Dans ce huitième et dernier projet du parcours Ingénieur Machine Learning (IML), nous devons choisir une compétition, y participer et obtenir des résultats mesurables.

Kaggle est conçue comme une plateforme collaborative, ce projet est l’occasion de participer à cette communauté.

1. Compétition choisie

J’ai choisi la compétition « House Prices - Advanced Regression Techniques » dans lequel on doit prédire le prix de vente de maison et pratiquer le feature engineering.

Le dataset : - 79 variables explicatives décrivant tous les aspects de maisons résidentielles.

On a un fichier « train » sur lequel on va entrainer nos modèles, les maisons de ce fichier ont un prix associé. On a également un fichier « test », possédant une colonne de moins que « train » contenant les maisons dont on va chercher à estimer le prix de vente.

La métrique de ce projet est « Root-Mean-Squared-Error (RMSE) », nous devons rendre un fichier .csv contenant les identifiants des maisons à estimer ainsi que le prix que l’on a estimé.

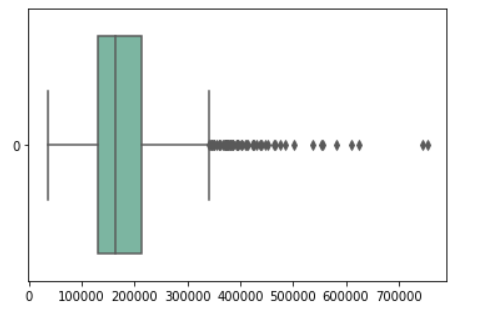
1. Cleaning effectué

**Pour train et test**

* Suppression de la variable « GarageYrBlt » et Id
* Gestion des variables dont les valeurs manquantes représentent une absence. Par exemple, ceci est indiqué dans le document texte data\_description.txt en

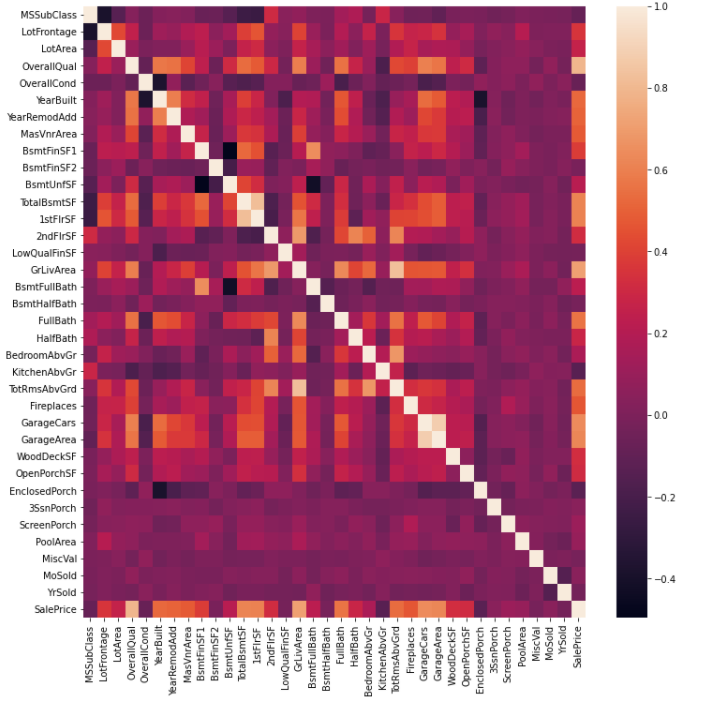
« NA No Basement» pour la variable Basement.

* OneHotEncoding des variables catégorielles
* Standardisation des variables quantitatives
* KNNImputer des valeurs manquantes sur test et dropna() sur train.

En sortie de ce nettoyage, certaines variables présentent dans train ne le sont pas dans test et inversement. On supprime ces variables.

Boxplot de la target « SalePrice »

J’ai choisi de conserver la target ainsi car certaines maisons peuvent avoir un prix élevé, je ne vois pas les losanges noirs comme des outliers.

1. Exploration des données

On constate que SalePrice est particulièrement correlée avec un grand nombre de variables de notre dataset.

1. Modélisation

Notre target est : **SalePrice**

Les modèles testés sont :

* RandomForestClassifier :

C’est un méta-estimateur qui ajuste un certain nombre de classificateurs d'arbres de décision sur divers sous-échantillons de l'ensemble de données et utilise la moyenne pour améliorer la précision prédictive et contrôler le surajustement.

* GradientBoostingRegressor :

Il construit un modèle additif par étapes ; il permet l'optimisation de fonctions de perte différenciables arbitraires. A chaque étape, un arbre de régression est ajusté sur le gradient négatif de la fonction de perte donnée.

* XGBRegressor :

XGBoost est une bibliothèque logicielle open source qui fournit un cadre de régularisation de gradient boosting.

Différentes modélisations ont été essayé, avec :

* Différents GridSearchCV
* Suppressions ou non des outliers dans train
* « Features importance » pour récupérer les 10 variables les plus utiles
* « Features importance » sans outliers dans train

1. Modèle final

Les résultats donnés pour la suppression des lignes contenant des outliers sont inférieurs à ceux dont les outliers ont été conservé.

Le modèle final sélectionné est un XGBRegressor.

**GridSearchCV :**

parameters = {

'n\_estimators': [100, 200, 500, 700],

'learning\_rate': [0.001, 0.01, 0.1, 0.2, 0.3],

'gamma': [0.5, 1, 1.5, 2, 5],

'subsample': [0.6, 0.8, 1.0]

}

On fait tourner ce modèle sur toutes les données d’entrainement avant de soumettre notre prédiction finale à Kaggle.

1. Résultats

Sur nos données test, notre meilleur modèle a une RMSE de 0.13222

Les paramètres de ce XGBRegressor sont les suivants :

{'gamma': 0.5, 'learning\_rate': 0.01, 'n\_estimators': 700, 'subsample': 0.8}

Nous sommes à la 4270eme position sur 12995.

Lien du kernel Kaggle : <https://www.kaggle.com/thibaultroussel/notebook-p8-tr>

1. Conclusion

* On a bien choisi une compétition kaggle
* On y a participé
* On a obtenu des résultats mesurables
* On a créé un kernel permettant de partager un élément intéressant avec la communauté