

Projet Machine Learning I

Souhaib BEN TAIEB - Tanguy BOSSER - Victor DHEUR

Altruy Alan & Delabie Xavier & Vanduynslager Estebane

Machine Learning 1



May 15, 2024

Table des matières

- Exploration et analyse des données
- Méthodologie
- Résultats

Exploration et analyse des données

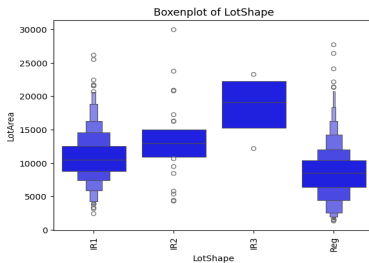
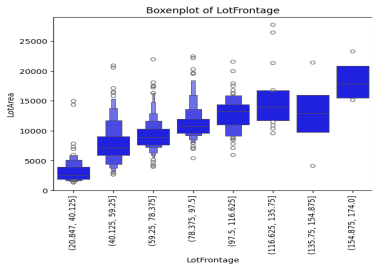
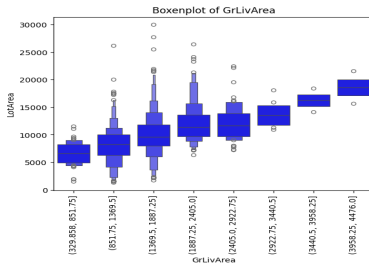
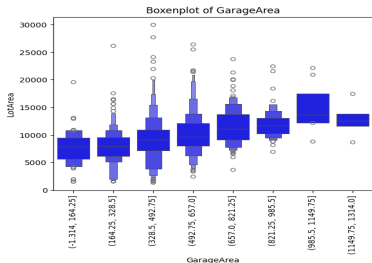
Analyse de plusieurs features pour garder les "meilleures"

- *Potentielles meilleures features*

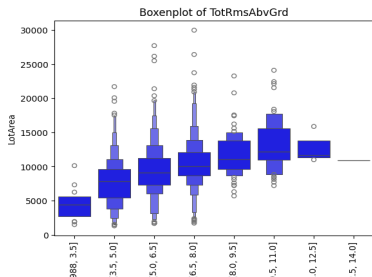
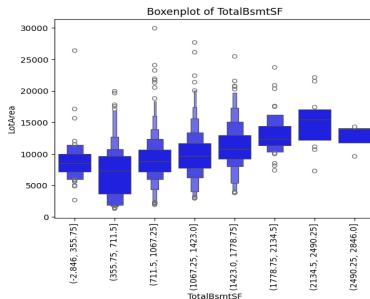
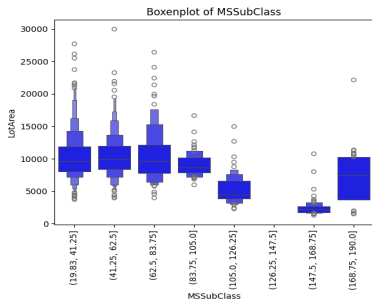
- *GarageArea*
- *GrLivArea*
- *LotFrontage*
- *LotShape*
- *MSSubClass*
- *TotalBsmtSF*
- *TotRmsAbvGrd*

Exploration et analyse des données

Boxenplots des features choisies



Exploration et analyse des données



Exploration et analyse des données

- *Autres "bonnes" features gardées*
 - *Alley*
 - *Heating*
 - *MSZoning*
 - *Neighborhood*
 - *Street*
 - *YearBuilt*
- Ces features ont été retenues suite à de bons retours de résultats après plusieurs tests.

Méthodologie

- Algorithme principal :
- *DecisionTree*
 - Choix des paramètres
 - `criterion = absolute_error`
 - `max_depth = 30000000`
 - `min_samples_split = 2`
 - `min_samples_leaf = 1`
 - `max_features = sqrt`
 - `max_leaf_nodes = None`
 - `random_state = 2`
 - Ignorer les valeurs abusives
 - Bagging (500 arbres)

Méthodologie

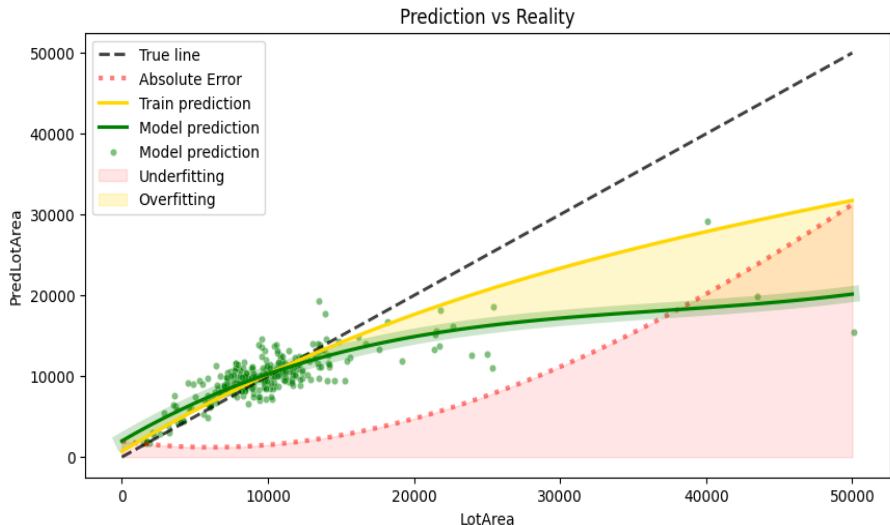
Autre algorithme testé :

StackingRegressor

- Implémentation du boosting
- Couplage avec Lasso

Ridge comme estimateur final

Résultats



Résultats

Résultats du Random Forest :

Time	2.00 s
Kaggle Score	0.24691
MSE Test	128.6303
MSE train	54.6578
To be near to 1	0.4249
Erreur absolue moyenne	1987.93
Erreur absolue maximale	34625.95
Erreur absolue minimale	11.56

Résultats du StackingRegressor :

Time	8.88 s
Kaggle Score	0.24677
MSE Test	112.9118
MSE train	66.9116
To be near to 1	0.5926
Erreur absolue moyenne	1975.71
Erreur absolue maximale	32407.32
Erreur absolue minimale	0.26

Discussion

Le calcul de ces résultats nous ont permis de tirer des conclusions

Interprétation :

- score : $\sqrt{\frac{1}{n} \sum_{i=1}^n (\log y_i - \log \hat{y}_i)^2}$
- Erreur carré moyenne
 - Test
 - Train
- ratio entre MSE Test et MSE Train
- erreur absolue
 - moyenne
 - maximale
 - minimale

Merci pour votre attention