PRESENTATION

PROJET 8 Ingénieur Machine Learning

RAPPORT

Presented by:
CHEIKH BADIANE

House Prices Advanced Regression Techniques

KAGGLE COMPETITION

OPENCLASSROOMS INGENIEUR MACHINE LEARNING

TABLE DES MATIERES

ANALYSE ET EXPLORATION DES DONNÉES: EDA

REDUCTION DES DIMENSIONNALITES ET MODELISATION

PREDICTION

INTRODUCTION

KAGGLE

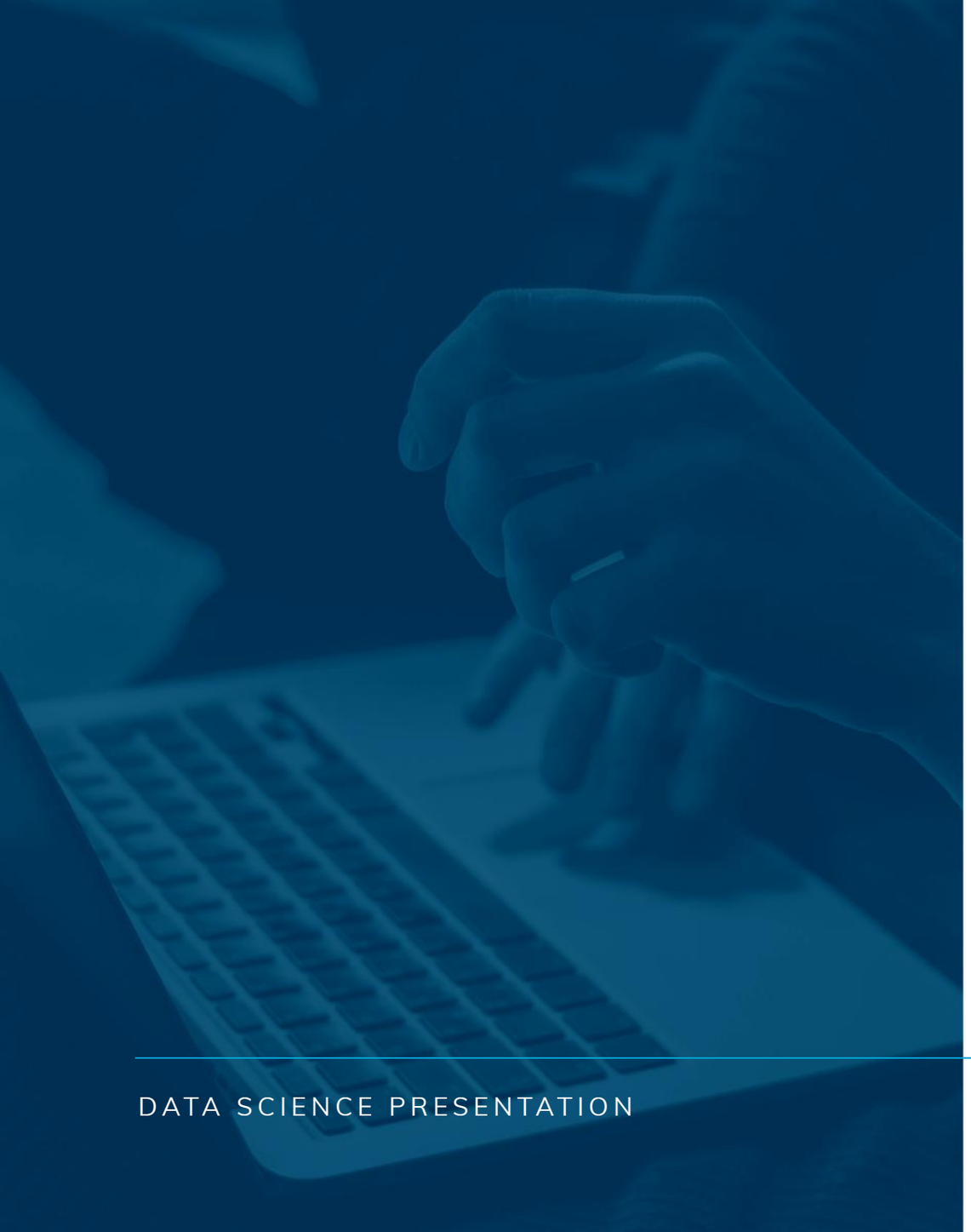
Kaggle est un plateforme qui réunit des datascientists du monde entier et propose des cours et surtout des compétition en machine learning.

COMPÉTITON

Sur des données de logement explorer et prédire le prix de vente des logements en utilisant les autres variables 79 au total qui constituent le logement.

Analyse et exploration des données (EDA)

1. Analyse des corrélation
2. Exploration des features
3. Features Engineering
4. Assemblage des données
6. Traitement des valeurs manquantes
7. Normalisation des données
8. Modélisation
9. Prédiction

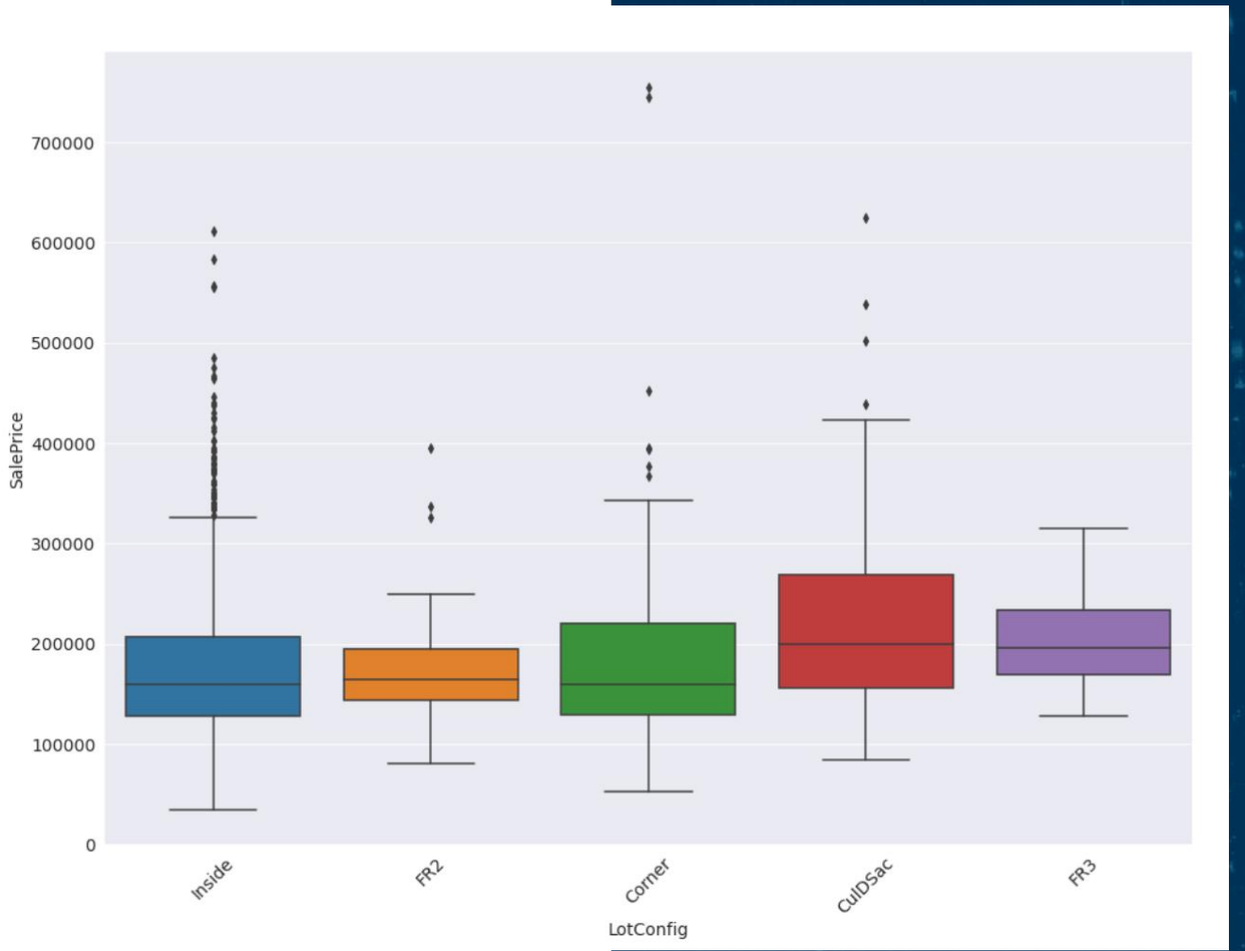


EDA

Dans cette partie, le travail consiste à checker les types de données analyser les relations qui pourraient exister entre le prix et les autres variables, encodage des variables catégorielle, imputation ou suppression des valeurs manquantes, traitement des multi colinéarités et standardisation des variables.

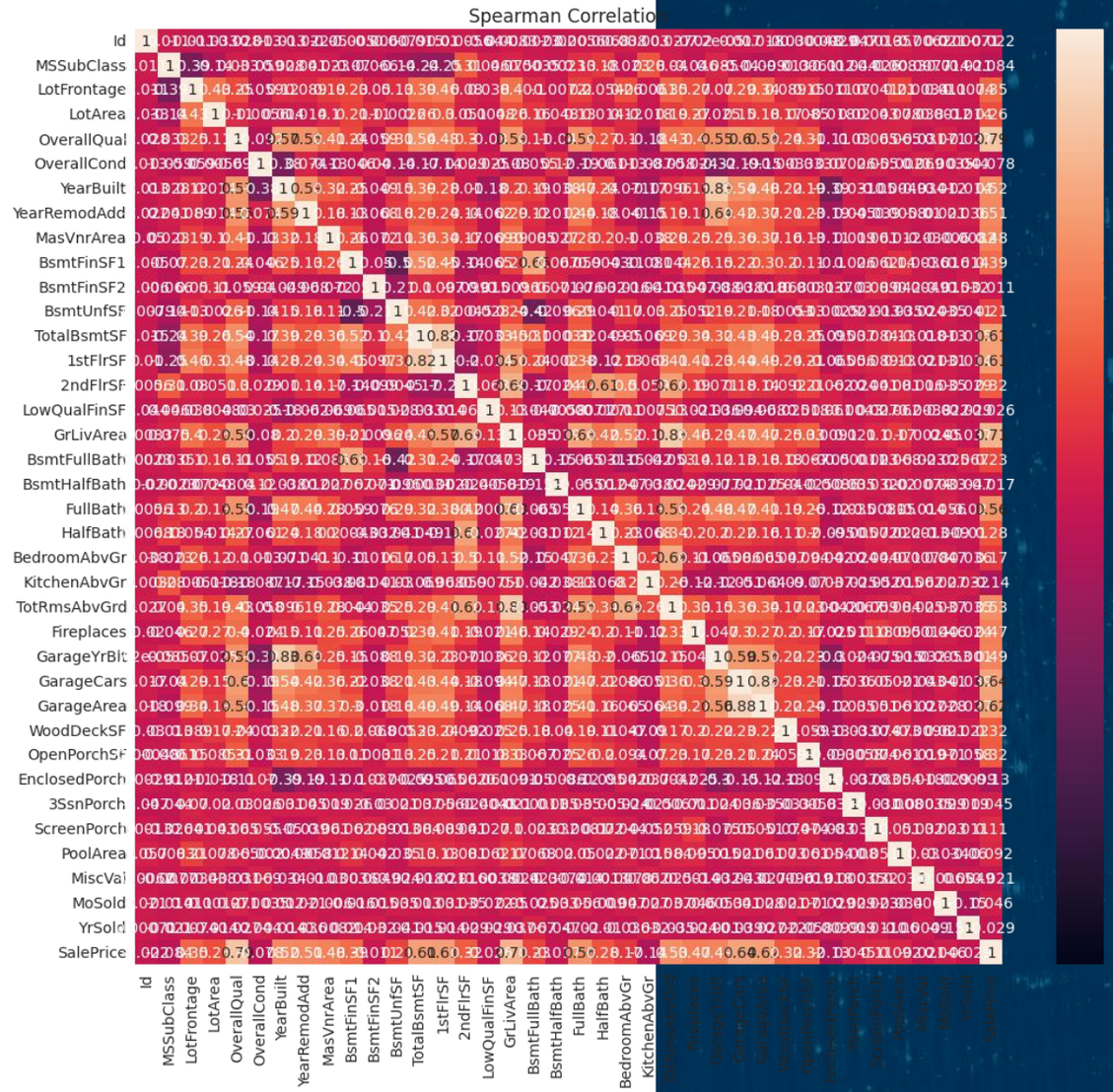
BOXPLOT DE LOTCONFIG

Distribution de la configuration du lot par rapport au prix du logement



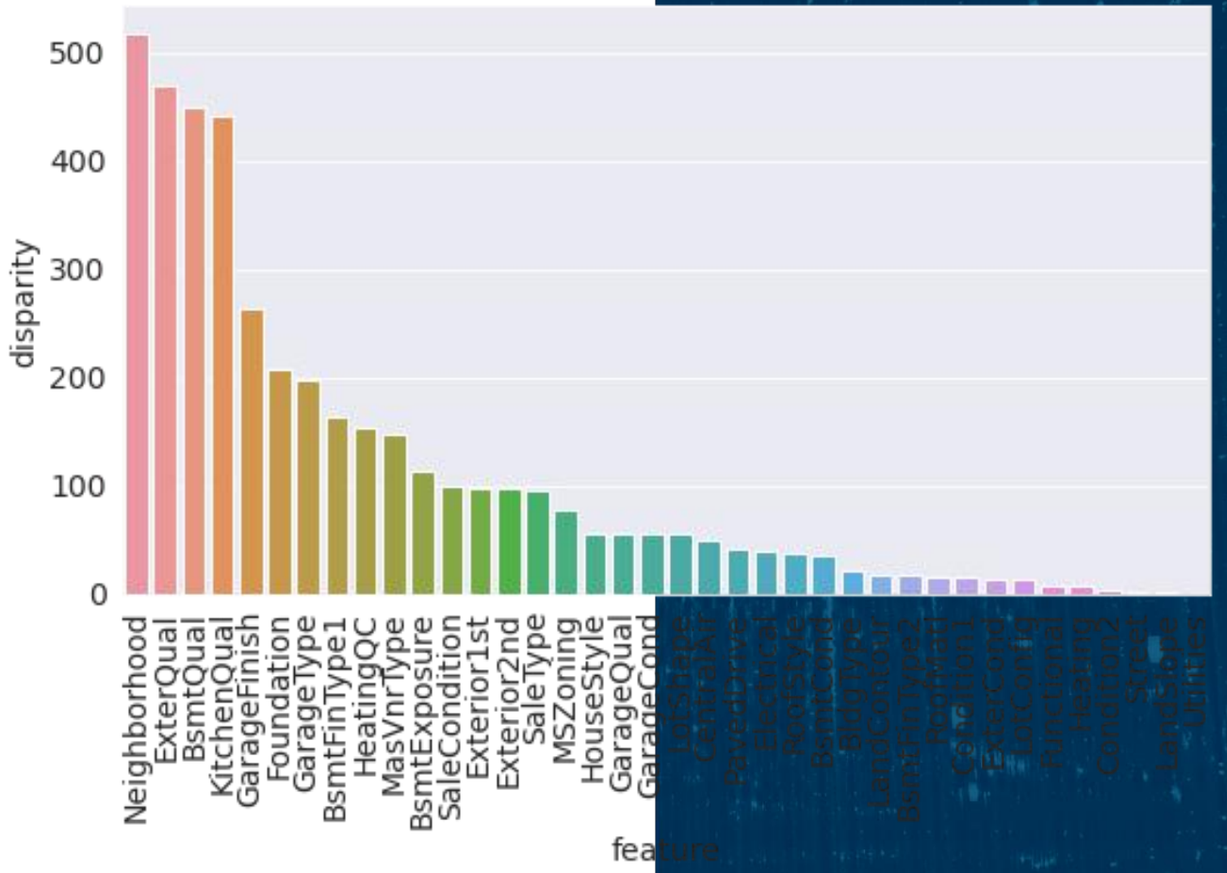
CORRÉLATION DES VARIABLES

La trice de corrélation des variable montre une forte coreeélation entre certaines variables, pour vant entraîner des problèmes de mult collinéarit



DISPARITÉ DES VARIABLES

L'analyse Anova permet de voir la disparité entre les variables avec le prix.



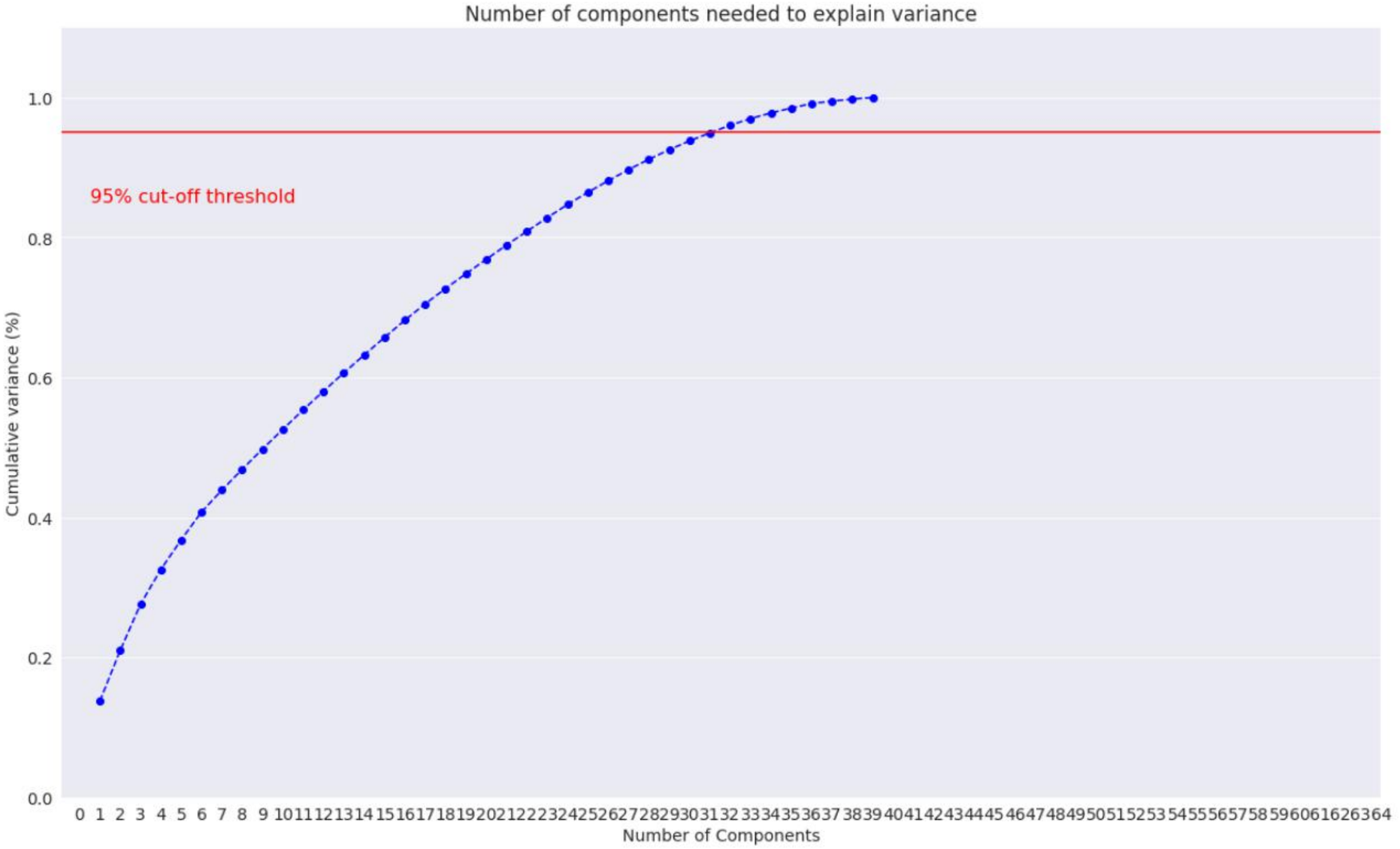
MULTI COLINÉARITÉ

toutes varibles ayant un coefficient de corrélation supérieur à 0.75 entre elles peuvent engendrer des problèmes de multi colinéarités donc on garde une des deux variables

	drop_feature	corr_feature	corr_values
0	Exterior1st	Exterior2nd	0.87
1	GrLivArea	TotRmsAbvGrd	0.81
2	TotalBsmtSF	1stFlrSF	0.80
3	YearBuilt	GarageYrBlt	0.79

REDUCTION DES DIMENTIONNALITES

Réduction des dimensionnalités avec une analyse des composantes principales PCA. Cette analyse nous a permis de conserver que 32 composantes qui expliquent à elles seules plus de 95% de la variance.


Number of components needed to explain variance
95% cut-off threshold
Cumulative variance (%)
0.0
0.2
0.4
0.6
0.8
1.0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
Number of Components

MODELISATION
REGRESSION LINEARE MULTIPLE

Timeline de la modélisation

6 ALGORITHMES

3 METRIQUES

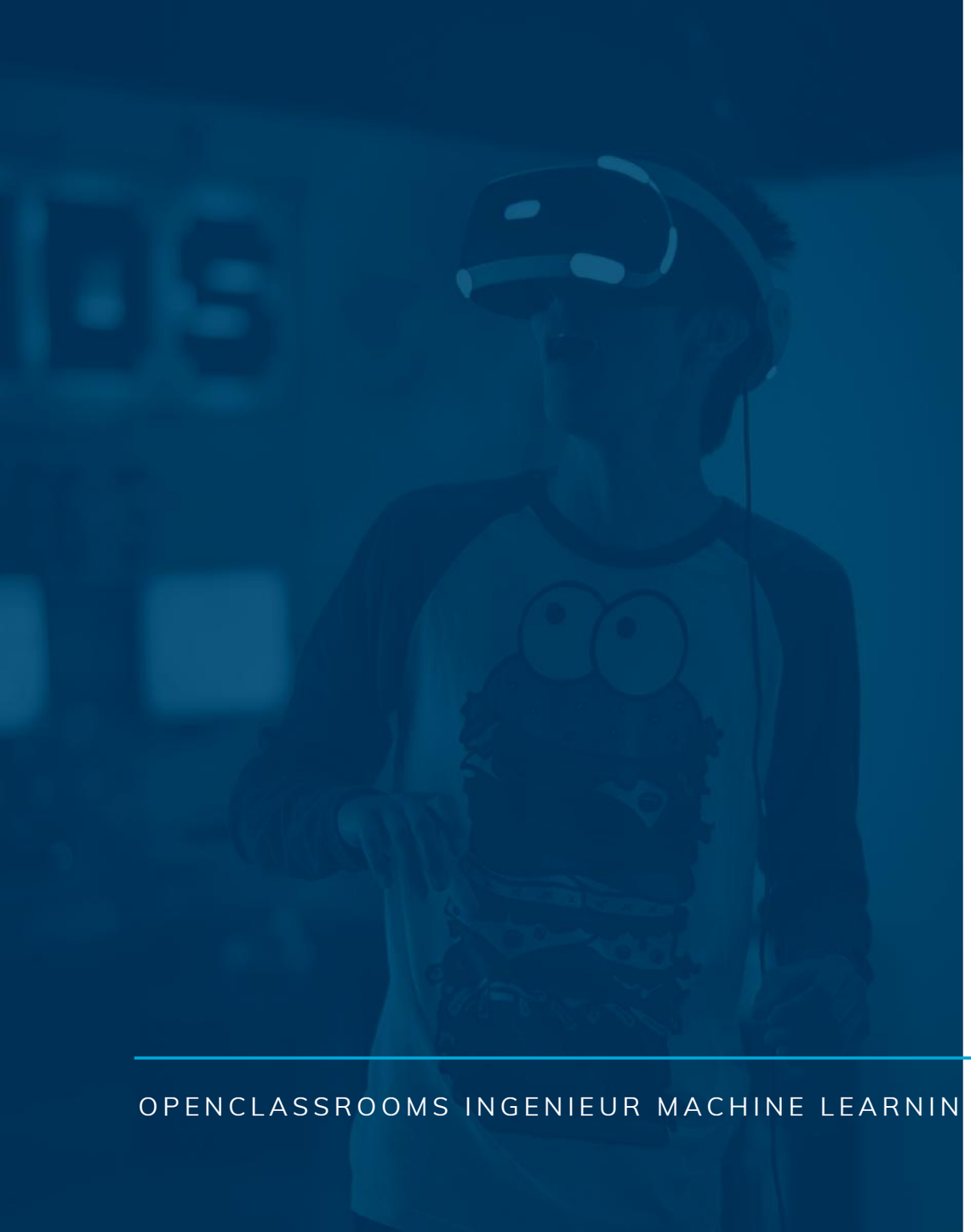
CROSSE VALIDATION

FINE-TUNING

Comparaison des modèles

Le meilleur modèle est celui qui aura le meilleur coefficient de détermination R2 suivi des autres métriques

	R2	MAE	MSE
XGB Boosting	0.8522	0.1124	0.0257
Linear Regression	0.8421	0.1202	0.0274
Random forest	0.8420	0.1142	0.0274
Light GBM	0.8408	0.1160	0.0276
Gradient Boosting	0.8309	0.1235	0.0293
Decision Tree	0.6616	0.1707	0.0587



PREDICTIONS

Prédiction avec le XGB Boosting après fine-tuning

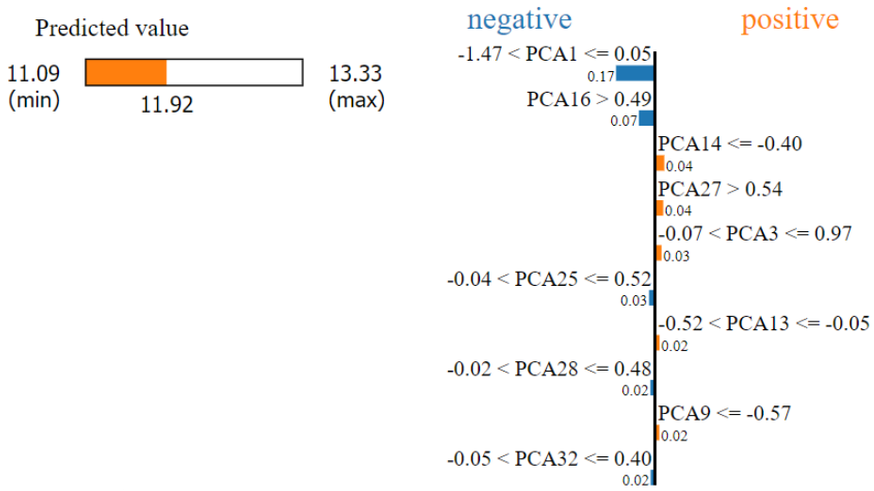
	Id	SalePrice
0	1461	136949.0
1	1462	145400.0
2	1463	175982.0
3	1464	204932.0
4	1465	190066.0

Explication du modèle avec Lime

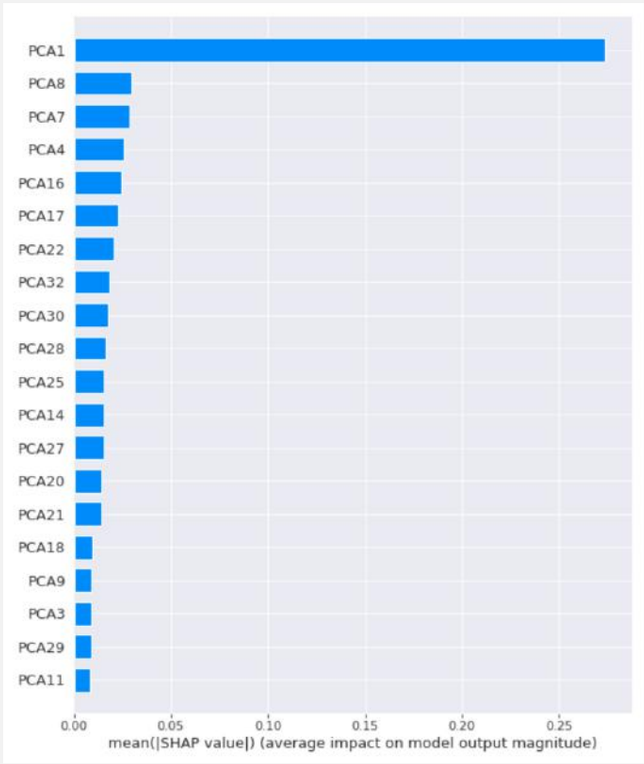
IMPORTANCE

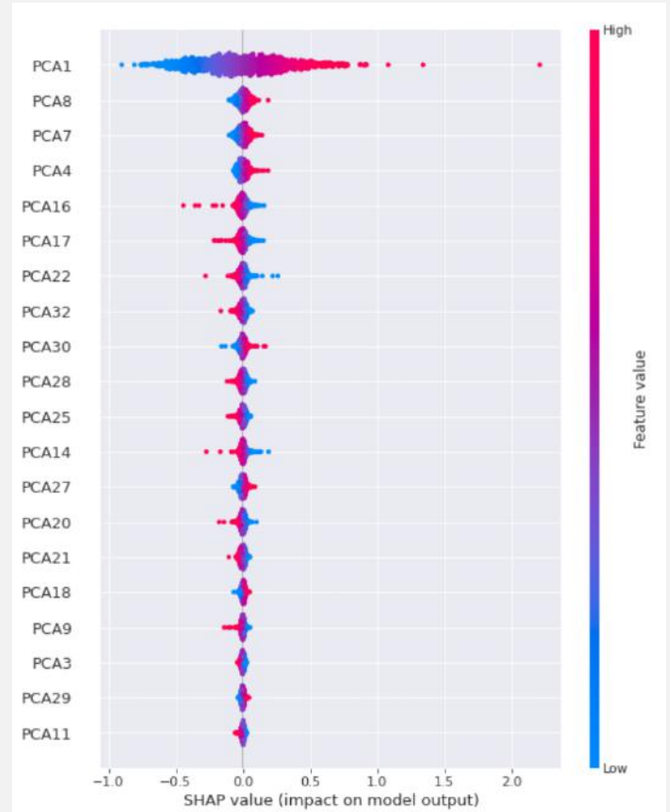
Feature	Value
PCA1	-0.73
PCA16	0.77
PCA14	-0.65
PCA27	1.23
PCA3	0.21
PCA25	0.49
PCA13	-0.36
PCA28	0.08
PCA9	-0.77
PCA32	0.20

INFLUENCE



L'inportance des variable dans le modèle avec Shap





CONCLUSION

La réduction des dimensionnalités n'a pas vraiment permis d'optimiser les modèles.
Il faudra revoir les techniques d'imputation mise en place et également revoir les
outliers; Peut-être la suppression automatique des variables avec de forte
corrélation a eu des impacts sur la précision et la qualité de mes modèles.
il faudra aussi revoir les paramètres des modèles et les optimiser d'avantage.

THANK YOU
Cheikh BADIANE