CHEIKH BADIANE

PROJET 8

PARTICIPATION A UNE COMPETITION KAGGLE

INTRODUCTION

Ce projet porte sur la participation à une compétition Kaggle dans la cadre du dernier projet du parcours Ingénieur Machine Learning de OpenClassRooms.

ROADMAP

EDA
Exploratory Data Analysis

Soumission
Choix du meilleur modèle et soumission

Modélisation
Choix des algos et modélisation

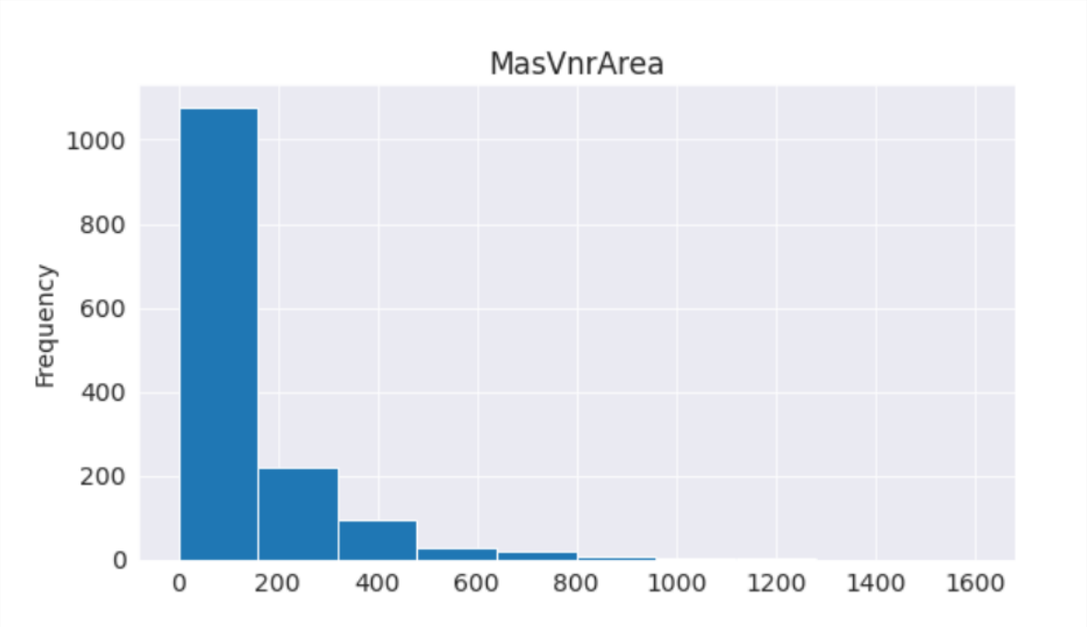
Analyse et exploration des données

mise en application des techniques d'analyse et exploration de données avant modélisation.

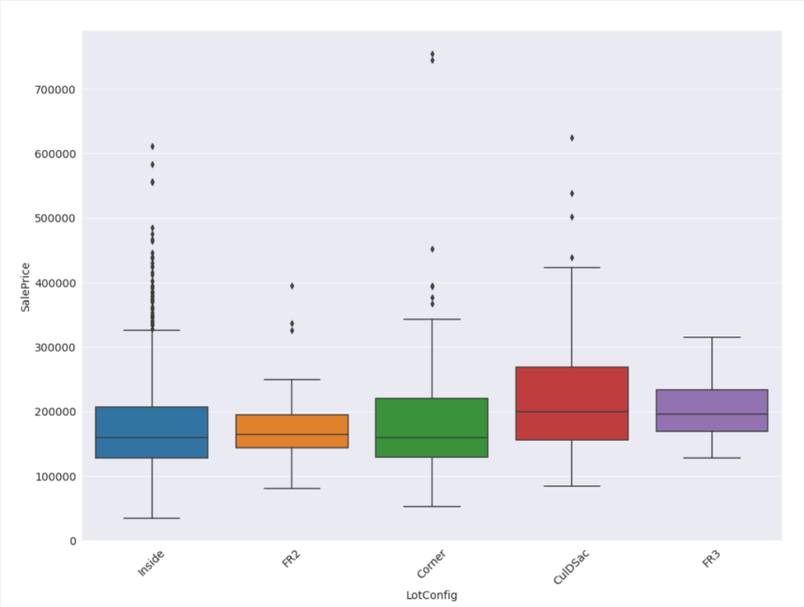
Technique	Mise en application
Vérification du type de chaque variable	Vérifier si chaque variable correspond au type attendu
Statistiques descriptives	analyse de la distribution des variables numériques
Analyse des valeurs manquantes et imputation	Suppression des variables avec plus de 40% de valeurs manquantes. Imputation par la moyenne pour les variables numérique et par Unknow pour les variables catégorielles.
Encodage	LabelEncoding des variables catégorielles
Standardisation	Standardisation des variables et transformation logarithmique de la variable target.
Multicollinéarité	Suppression des variables très corrélées (Coefficient de corrélation supérieur à 75%)

Distribution de quelques variables

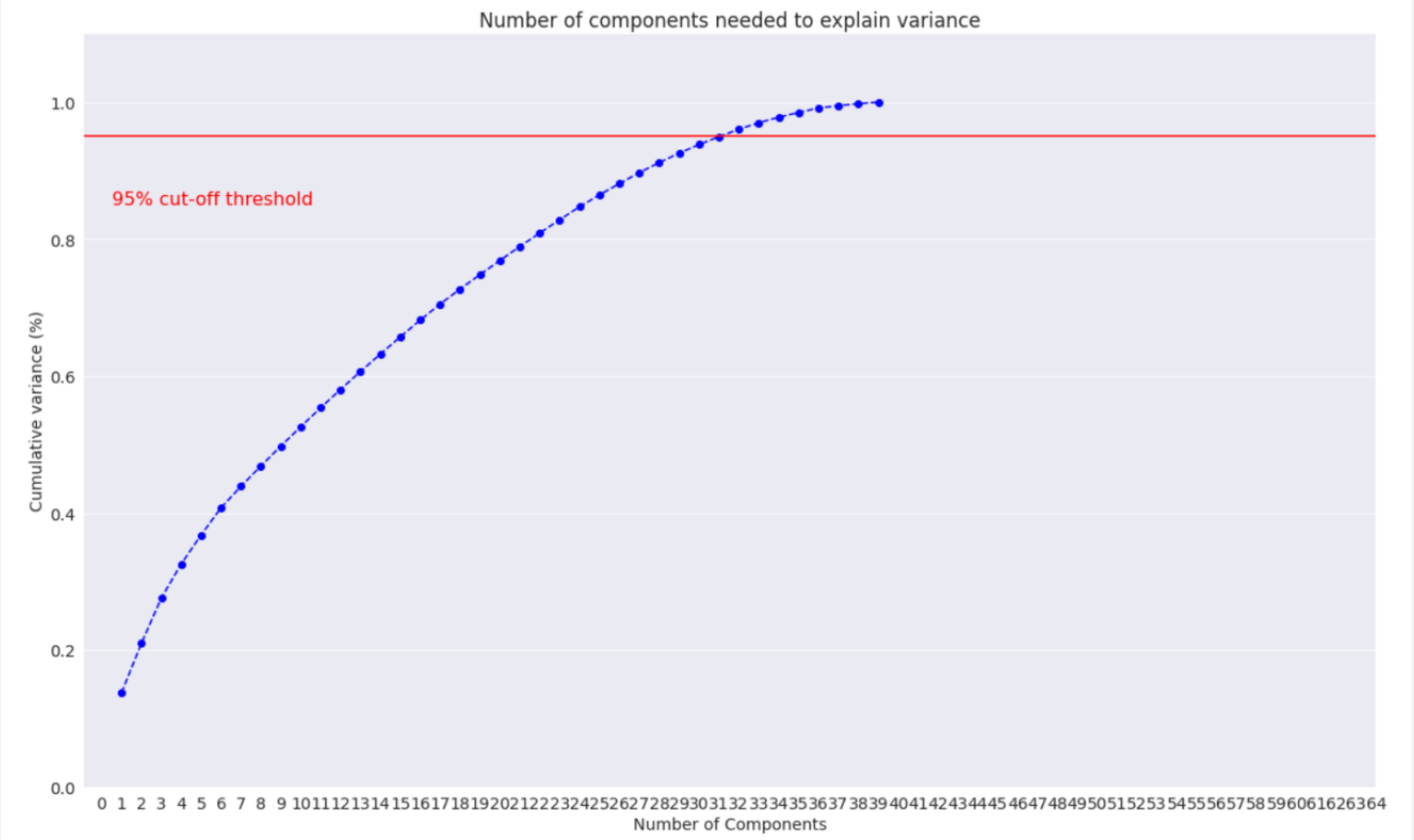
Variable numérique



Variable catégorielle

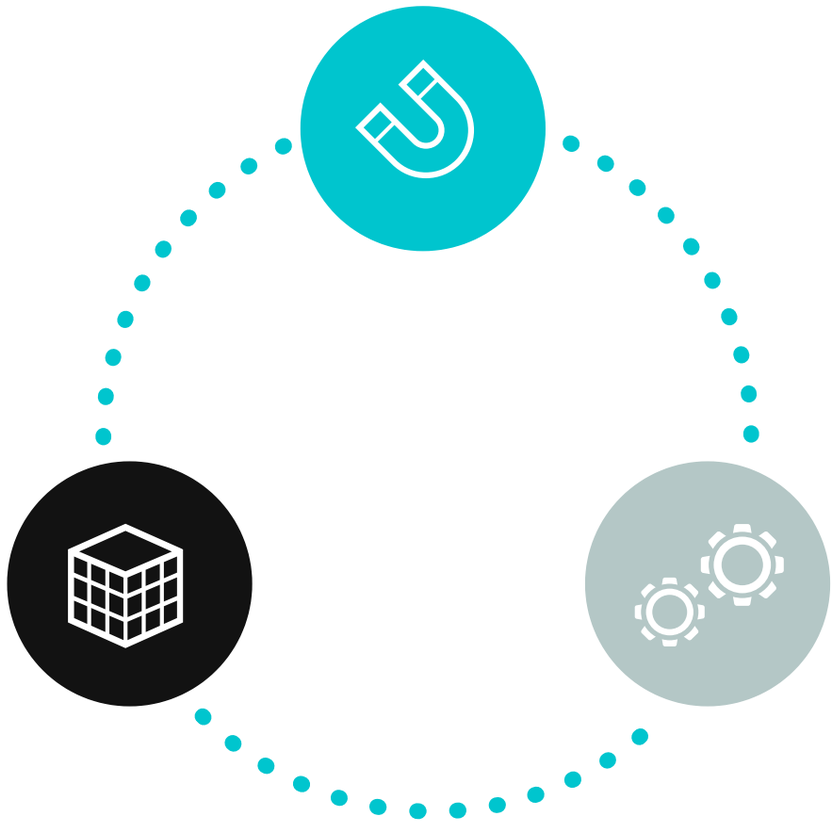


Réduction des dimentionnalités avec PCA



Modélisation

Application de plusieurs algorithmes et choix du meilleur algorithme en fonction de trois métriques



MÉTRIQUE1 : R2

La première métrique est le coefficient de détermination R2 qui permet de voir les Saleprice d'ajustement par rapport Saleprice d'origine

MÉTRIQUE2 : MAE

Erreur absolue moyenne, différence entre Salprice prédites et SalePrice d'origine

MÉTRTIQUE3 : MSE

Erreur quadratique moyenne, différence entre prédiction et valeurs d'origines par la différence moyenne au carré sur l'ensemble des données

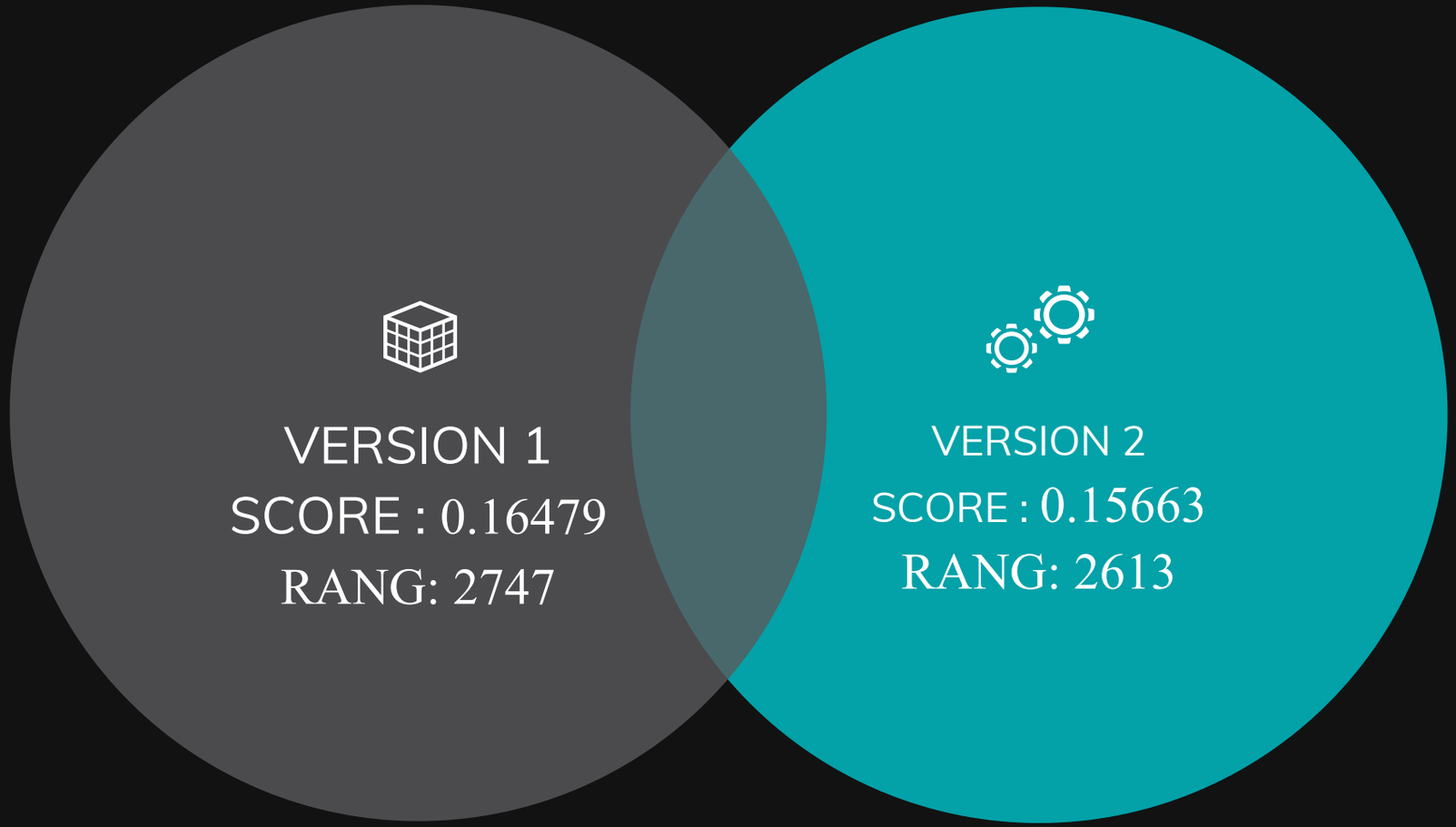
COMPARAISON DES ALGOS

Le meilleur modèle sera celui avec le plus grand R2 suivi des MAE et MSE. Et pour notre cas il s'agit de l'algorithme XGB Boosting. Et c'est ce dernier qui a été utilisé pour prédire la soumission.

	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

SOUMISSION

Deux soumissions au total



Conclusions

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.

OpenClassRooms
Cheikh Badiane

THE END

THANK YOU!