Projet Machine Learning

Axel Mazouth-laurol
Magatte LO
Grégoire Dihedhiou

February 2022

1 Introduction

L'achat d'un bien immobilier représente toujours une étape importante dans la vie d'un individu quelque soit sa catégorie socio-professionnelle. Et nécéssite un apport financier conséquent. De plus, son est bien souvent surévalué par son prorpiétaire. C'est pourquoi faire estimer le prix de vente d'un bien immobilier est capital, afin de s'assurer qu'il corresponde aux prix du marché. Dans le cadre de ce projet, nous participons à une compétition sur kaggle intitulée **"House Prices Advanced Regression techniques"** dont l'objectif est d'estimer au mieux le prix de 1460 maisons réparties dans l'ensemble de la ville d'Ames , dans l'état de l'Iowa aux États unis . Pour mener à bien ce projet, nous allons utilisé divers algorithmes de machine learning, tels que la régression linéaire, le XGBoost ou encore les forêts aléatoires. Au préalable, nous procéderons à une analyse fine et rigoureuse des données ainsi qu'à leur nettoyage, afin de les rendre exploitables et d'optimiser nos algorithmes futurs.

2 Analyse des données

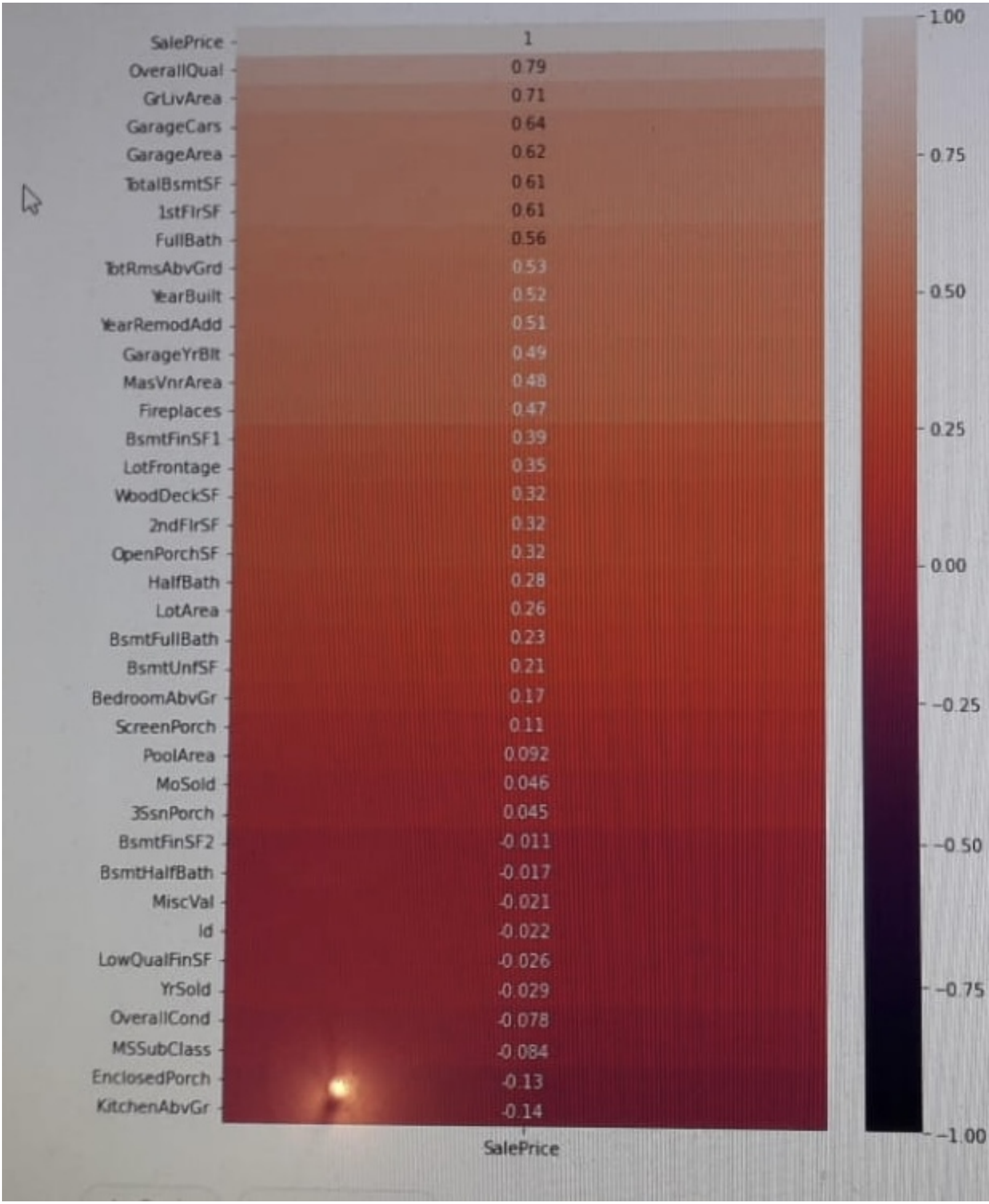
2.1 Analyse de forme

Nous avons à notre disposition deux jeux de données. Un jeu de données d'entraînement et un jeu de données de test. Le jeu de données d'entraînement comporte 1460 lignes et 81 colonnes tandis que le jeu de données de test comporte 1459 lignes et 80 colonnes. Notre variable "cible", celle que nous cherchons à prédire, est la variable **"SalePrice"** qui est présente dans notre jeu d'entraînement. Les autres variables sont les variables explicatives, qui serviront à prédire notre variable "target". Parmi ces variables, 43 variables sont de type catégorielles et 37 variables sont numériques. Parmi les variables catégorielles, nous sommes en présence de variables ordinales d'une part et nominales d'autre part. Cette distinction est importante lorsque nous procéderons au nettoyage des données. Enfin nos jeux de données respectifs comportent respectivement 19 et 33 colonnes dans lesquelles des données manquantes sont observées.

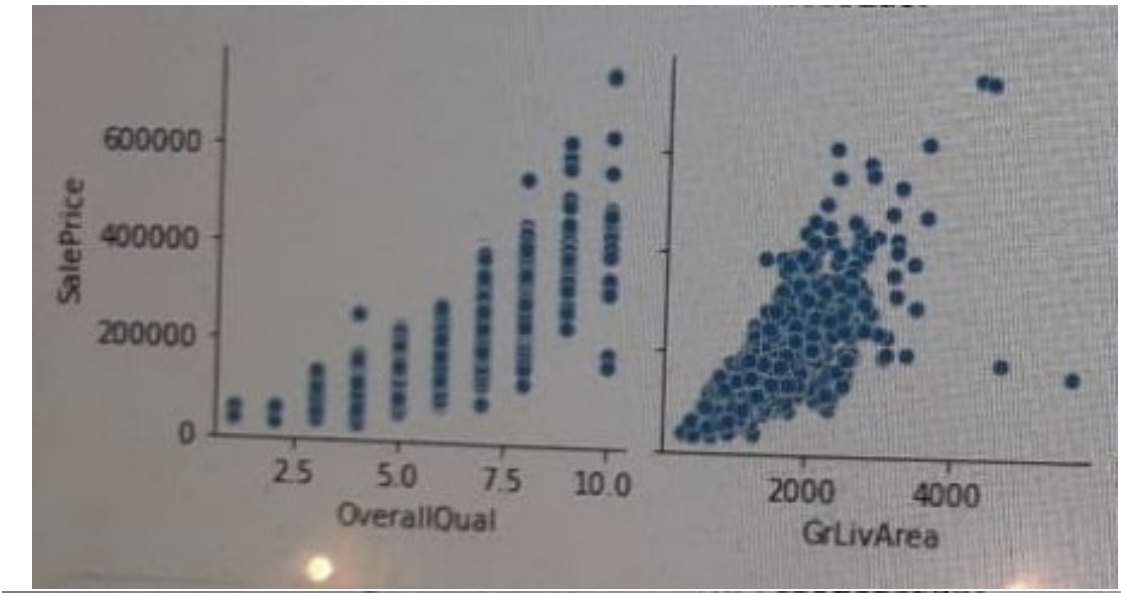
2.2 Analyse de fond

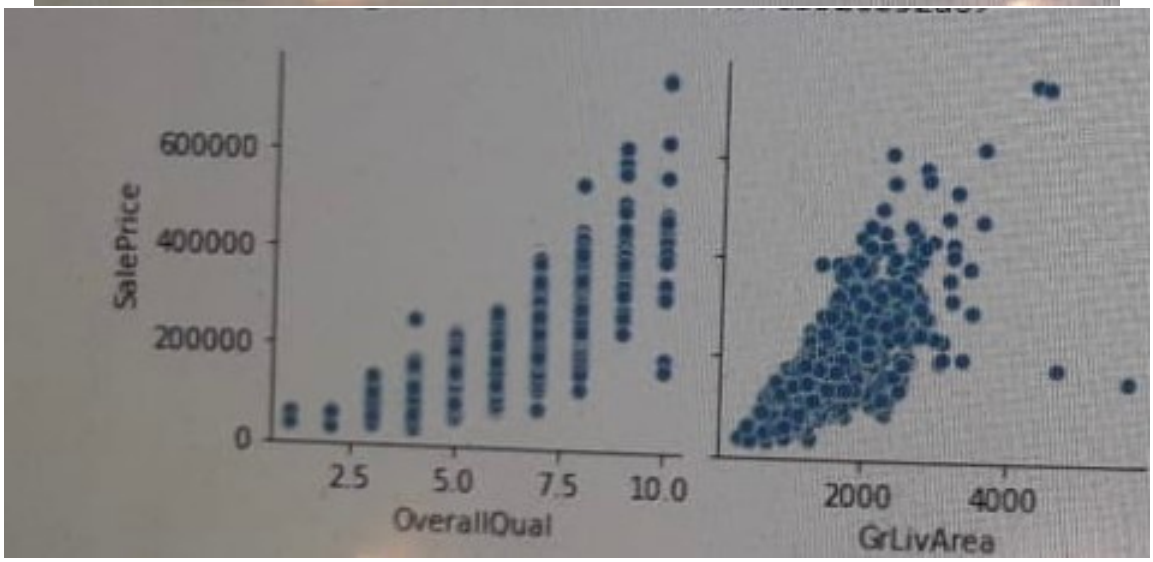
Notre analyse de la varible "Saleprice" révèle un prix moyen des maisons situé à **180921.195890** dollars, un prix médian. Le prix minimal de l'ensemble des biens est de **34900** dollars tandis que le prix du bien le plus élevé atteint la valeur de **755000** dollars. Enfin, près de 50% des biens immobiliers ont une valeur comprise dans l'intervalle [129975,214000] qui représentent respectivement le premier et le troisième quartile.

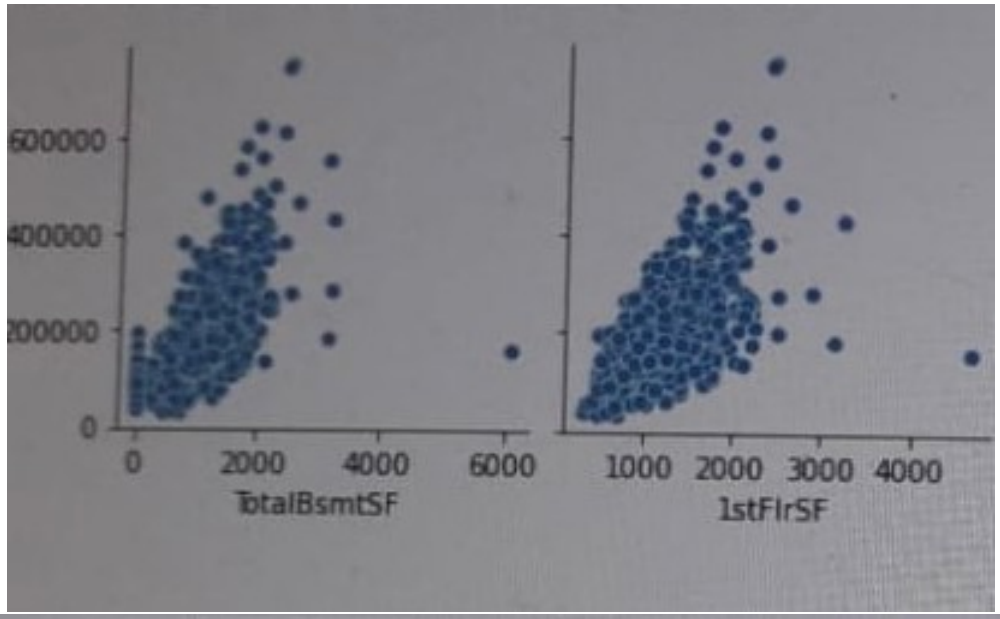
Le graphique traduisant la matrice de corrélation de la variable saleprice :

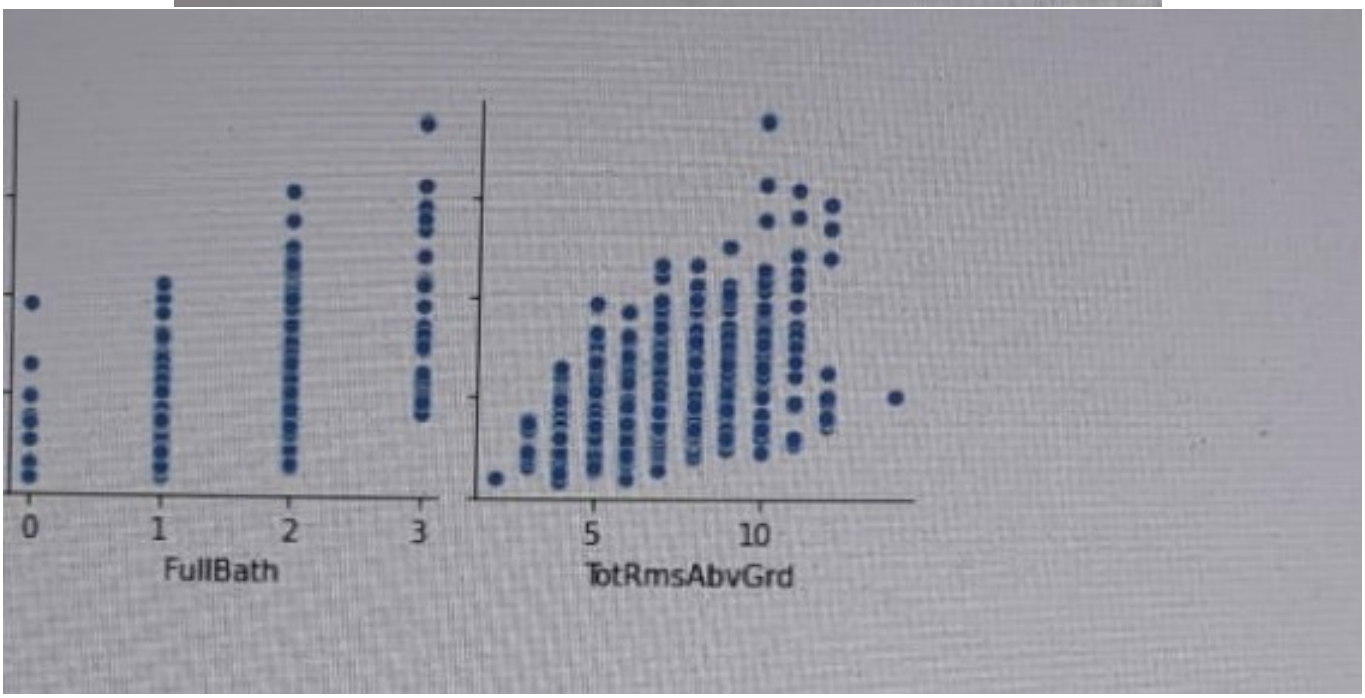

SalePrice 1
OverallQual 0.79
GrLivArea 0.71
GarageCars 0.64
GarageArea 0.62
TotalBsmtSF 0.61
1stFlrSF 0.61
FullBath 0.56
TotRmsAbvGrd 0.53
YearBuilt 0.52
YearRemodAdd 0.51
GarageYrBlt 0.49
MasVnrArea 0.48
Fireplaces 0.47
BsmtFinSF1 0.39
LotFrontage 0.35
WoodDeckSF 0.32
2ndFlrSF 0.32
OpenPorchSF 0.32
HalfBath 0.28
LotArea 0.26
BsmtFullBath 0.23
BsmtUnfSF 0.21
BedroomAbvGr 0.17
ScreenPorch 0.11
PoolArea 0.092
MoSold 0.046
3SsnPorch 0.045
BsmtFinSF2 -0.011
BsmtHalfBath -0.017
MiscVal -0.021
Id -0.022
LowQualFinSF -0.026
YrSold -0.029
OverallCond -0.078
MSSubClass -0.084
EnclosedPorch -0.13
KitchenAbvGr -0.14
SalePrice
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Les Graphes des nuages de points de la variable saleprice en foction des varaiables corrolées :









On distingue au travers de la matrice de corrélation une forte corrélation positive entre le prix de la maison et les variables suivantes : la taille du garage("GarageArea), le nombre de voitures pouvant être placées dans un garage("GarageCars"), l'état général de la maison("overallqual), la superficie du premier étage(1stFlrSF), ainsi que la surface totale du sous-sol("TotalBsmtSF"). Les nuages de points représentant la variable "SalePrice" en fonction de ces variables confirment cet état de fait.

On distingue au travers de la matrice de corrélation une forte corrélation positive entre le prix de la maison et les variables suivantes : la taille du garage("GarageArea), le nombre de voitures pouvant être placées dans un garage(“GarageCars”), l'état général de la maison("overallqual), la superficie du premier étage(1stFlrSF), ainsi que la surface totale du sous-sol(“TotalBsmtSF”). Les nuages de points représentant la variable “SalePrice” en fonction de ces variables confirment cet état de fait.

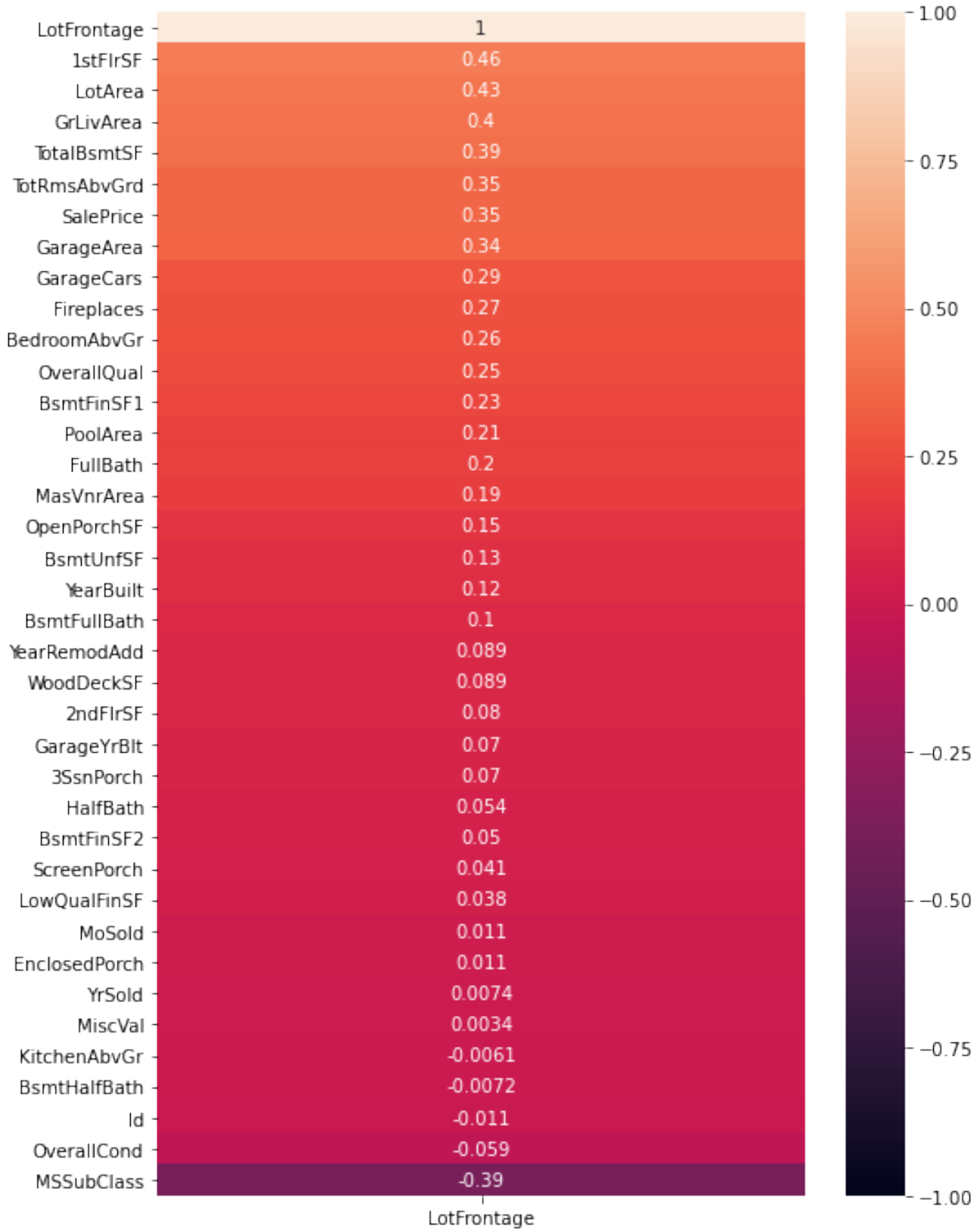
3 Preprocessing

Dans cette partie, nous allons procédé au nettoyage de données. Il s'agit de l'étape la plus importante car elle nous permet de garantir la cohérence de nos données

3.1 Traitement des valeurs manquantes

À travers notre analyse de données, on s'aperçoit que les variables "Alley","MiscFeature","PoolQC", et "Fence" concentrent chacune un pourcentage éléve de valeurs manquantes supérieur à 80. Nous avons donc pris la décision de ne pas les prendre en compte pour la suite du projet et par conséquent de la supprimer.De même, la variable ID ne représente aucun intérêt pour notre étude. Cependant d'autres variables possèdent elles aussi des valaurs manquantes mais en quantité limitée. L'idée première qui nous viendrait à l'esprit serait de supprimer pûrement et simplement toutes les lignes comportant au moins une valeur manquante. Toutefois, cette méthode nous ferait perdre de l'information sur nos données et les prédictions attendues ne seraient pas celles escomptées.
C'est la raison pour laquelle on a procédé à des méthodes d'imputation pour traiter les données manquantes restantes. Parmi les variables numériques présentant des données manquantes, les variables caractérisant l'année de

construction du garage, et la largeur du terrain seront imputés différemment des autres variables. L'absence de données dans la variable "GarageYrBlt" est liée au fait qu'il y ait une absence de garage pour les maisons concernées. On peut le constater en observant les autres variables représentant les caractéristiques des garages. Toutefois, après observation de cette variable, on s'aperçoit que la maison la plus récente date de 2207, ce qui constitue une valeur aberrante au premier abord. On pourrait être tenté de la supprimer et remplacer les valeurs manquantes par la moyenne ou la valeur médiane de l'année de construction. Cependant, la colonne GarageYrBlt correspond à l'année où le garage a été construit. Or, nous savons que dans certaines maisons, il n'y a pas eu de garage construit. Au lieu de remplacer les dates manquantes par la médiane, nous en avons conclu qu'il serait judicieux de les remplacer par une année future. On a donc remplacé les valeurs manquantes par l'année 2207. La variable "LotFrontage" présente 480 données manquantes. Ci-dessous le graphique représentant la matrice de corrélation lot-frontage :



FIGURE!!!

De plus, après analyse de la matrice de corrélation cette variable n'a pas de corrélations significatives avec les autres. Nous avons donc employé la méthode des k plus proches voisins pour imputer cette variable. Les autres variables catégorielles sont imputés par la valeur 0 car les valeurs manquantes sont liées de façon certaine à une

absence d'éléments.
En ce qui concerne les variables catégorielles, toutes les valeurs manquantes prendront la valeur "None".

3.2 Encodage de variables

Parmi nos variables qualitatives, se trouvent des variables qualitaitves ordinales et des variables qualitiatives nominales. La différence entre ces deux types de variables réside dans le fait que les variables nominales ne sont pas hierarchisées. Aucune valeur n'est supérieure à une autre. Comme par exemple le statut marital d'un individu. Á l'inverse, les variables qualitatives ordinales peuvent être classées les unes par rapport aux autres. On peut les classer dans un ordre logique selon une échelle de valeur. Le niveau de performance d'un sportif constitue un des très nombreux exemples possibles et imaginables.
En conséquence, nous avons encodé les variables nominale en implémentant une méthode d'encodage ordinale, où chaque valeur d'une variable sera associée à un entier. Tandis que les variables nominales ont été encodées à chaud. C'est à dire que pour une fonctionnalité donnée, nous avons crée autant de nouvelles colonnes qu'il y a de catégories possibles. Pour un échantillon donné, la valeur de la colonne correspondant à la catégorie est définie sur 1 tandis que toutes les colonnes des autres catégories sont définies sur 0.

3.3 Normalisation des variables

Avant de passer à l'étape de la modélisation, nous avons normalisé nos variables afin que les valeurs de toutes les variables soient mises à la même échelle. Pour cela, nous avons utilisé la fonction "RobustScaler" du package "Scikit-Learn". Contrairement à la standardisation, on ne soustrait pas nos données à la moyenne, mais à la médiane de chaque variable. Or la médiane est beaucoup moins sensible aux valeurs aberrantes et extrêmes que peut l'être la moyenne. Ensuite au lieu de diviser par l'écart type, on a divisé par l'inter-quartile de nos données. On applique l'expression suivante :
$X_s = \dfrac{X - mdiane}{IQR}$ où X_s désigne la variable X normalisée.

4 Regression linéaire

Comme cela a été dit précedemment, nous avons analysé et prétraité nos données avant tout entraînement de modèle sur notre dataset. Le dessein de ce projet étant de prédire la valeur d'une maison en fonction de variables explicatives. Nous avons pris la décision d'opter pour un modèle de regression linéaire régluarisé. C'est à dire que ces modèles vont réduire les coefficients de nos variables explicatives afin de sélectionner les variables les plus utiles à notre analyse. Voire même les supprimer. Nous avons effectuer trois régressions régulariséés, la régréssion Lasso, la régression Ridge, et la régréssion elasticnet. Pour chacune de ces méthodes, nous évaluerons la précision de notre modèle à l'aide de la valeur de l'erreur quatradique moyenne.

4.1 Régression Ridge

La régression Ridge est une extension de la régression linéaire où la fonction de perte est modifiée pour minimiser la complexité du modèle. Cette modification se fait en ajoutant un paramètre de pénalité équivalent au carré de la grandeur des coefficients. Le but de la régression ridge est d'estimer l'estimateur suivant :

$\hat{\beta^R} = argminL_\lambda(\beta)$ où $R_\mu(\beta) = \dfrac{\sum_{i=1}^{n}(Y_i - X_{i,.}\beta)^2}{2n} + \dfrac{\mu}{2}\|\beta|_2^2$

Dans la fonction de perte ci-dessus, beta est le paramètre que nous devons estimer. Une faible valeur de beta peut entraîner un sur-apprentissage du modèle, tandis qu'une valeur beta élevée peut entraîner un sous-apprentissage.

Dans scikit-learn, un modèle de régression de crête est construit en utilisant la classe Ridge. C'est celui-ci quenous avons utilisé. Après implémentation de l'algorithme, le résultat de notre algorithme affiche une précision de l'ordre de **0.8796201082907564** sur notre jeu de test.

4.2 Régression elastic$_net$

La régression linéaire elastic net utilise les pénalités des techniques de lasso et de crête pour régulariser les modèles de régression. Elle combine les deux termes de régularisation en un. les deux termes de régularisation en un. On cherche donc à résoudre le problème suivant : $\beta^{\hat{E}N} = argminF_\lambda(\beta)$ où $F_\lambda(\beta) = \frac{\sum_{i=1}^{n}(Y_i - X_{i,.}\beta)^2}{2n} + \lambda(\|\beta|_1 + \frac{1-\alpha}{2}\|\beta|_2^2)$ Dans ce cas, notre algorithme affiche une précision de **0.8217982077661392**.

4.3 Régression Lasso

La régression au lasso, ou l'opérateur de sélection et de retrait le moins absolu, est également une modification de la régression linéaire. Dans Lasso, la fonction de perte est modifiée pour minimiser la complexité du modèle en limitant la somme des valeurs absolues des coefficients du modèle (également appelée norme l1).

La fonction de perte pour la régression Lasso s'exprime de la façon suivante : $\hat{\beta^L} = argminL_\lambda(\beta)$ où $L_\lambda(\beta) = \frac{\sum_{i=1}^{n}(Y_i - X_{i,.}\beta)^2}{2n} + \lambda\|\beta|_1$

Dans la fonction de perte ci-dessus, alpha est le paramètre de pénalité que nous cherchons à estimer. L'utilisation d'une contrainte de norme l1 force certaines valeurs de poids à zéro pour permettre à d'autres coefficients de prendre des valeurs non nulles. L'algorithme nous affiche une précision finale égale à **0.8849585035423201**.

On en cocnlut donc que parmi les différentes méthodes de régression régularisé que l'on a étudiées, le modèle de régression lasso est le plus performant pour prédire le prix des maisons.

5 Forêt aléatoire

La forêt aléatoire est un ensemble d'arbres de décision.Les forêts aléatoires utilisent différents échantillons pour l'entraînement des données la variance observée dans les arbres de décision et permettent de construire plus d'arbres mais peu profonds. L'implémentation de nôtre forêt aléatoire fournit un résultat de **0.9841973639897152**.

6 Conclusion

Des différents algorithmes que l'on a implémentés, l'algorithme le plus peest celui des forêts aléatoires. Mais le modèle de régression Lasso révèle également des bonnes performances sur notre jeu de données. Les modèles de XGBOOST auraient également pu constituer une piste d'amélioration de nos modèles.