Master Modélisation Statistique

UNIVERSITÉ DE FRANCHE-COMTÉ

Faculté des sciences et techniques de Besançon

Rapport de projet Python

Année scolaire 2019/2020

Prédiction du prix de vente d'une maison à Boston, dans les années 70

Péparé par :
Yassine ABDOU

Encadré par :
Pr. Nicolas ASIN-HILAIRE
asin.nicolas@gmail.com

Master 2 modélisation statistique

5 février 2020

Table des matières

1 Introduction.

L'objectif de ce projet sera de construire un modèle de prévision du prix de vente d'une maison à Boston, dans les années 70 à partir de diverses données telles que la classe de construction, le pieds linéaires de rue reliés à la propriété, le type de logement, et d'autres variables.

Nous travaillerons sur des données issues d'un challenge du site Kaggle. Nous présenterons donc la table de données, le recodage de la variable d'intérêt puis nous modéliserons cette dernière qui représente l'échelle de prix de vente suivant toutes les variables données et enfin nous ferons la prévision afin de choisir le modèle le plus adapté à notre jeu de données.

2 Présentation de la table de données.

Ce jeu de données contient **1460** observations de **81** variables.
On donne ci-dessous une description des variables.

- SalePrice : le prix de vente de la propriété en dollars. Il s'agit de la variable cible à prédire.
- MSSubClass : la classe de construction
- MSZoning : la classification générale de zonage
- LotFrontage : pieds linéaires de rue reliés à la propriété
- LotArea : Taille du lot en pieds carrés
- Street : Type d'accès routier
- Alley : type d'accès à l'allée
- LotShape : Forme générale de la propriété
- LandContour : Planéité de la propriété
- Utilities : type d'utilitaires disponibles
- LotConfig : configuration du lot
- LandSlope : Pente de la propriété
- Neighborhood : emplacements physiques dans les limites de la ville d'Ames
- Condition 1 : Proximité de la route principale ou du chemin de fer
- Condition2 : Proximité de la route principale ou du chemin de fer (si une seconde est présente)
- BldgType : Type de logement
- HouseStyle : Style de logement
- OverallQual : Qualité globale du matériau et de la finition
- OverallCond : évaluation de l'état général
- YearBuilt : Date de construction d'origine
- YearRemodAdd : date de rénovation
- RoofStyle : Type de toit
- RoofMatl : matériau du toit
- Exterior1st : Revêtement extérieur de la maison
- Exterior2nd : Revêtement extérieur de la maison (si plusieurs matériaux)
- MasVnrType : Type de placage de maçonnerie
- MasVnrArea : surface de placage de maçonnerie en pieds carrés
- ExterQual : Qualité des matériaux extérieurs
- ExterCond : état actuel du matériau à l'extérieur
- Foundation : Type de fondation
- BsmtQual : Hauteur du sous-sol
- BsmtCond : État général du sous-sol
- BsmtExposure : Passerelle ou murs de sous-sol au rez-de-jardin
- BsmtFinType1 : Qualité de la surface finie du sous-sol
- BsmtFinSF1 : pieds carrés finis de type 1
- BsmtFinType2 : Qualité de la deuxième zone finie (si présente)
- BsmtFinSF2 : pieds carrés finis de type 2
- BsmtUnfSF : Pieds carrés de sous-sol non finis
- TotalBsmtSF : Superficie totale en pieds carrés du sous-sol
- Heating : Type de chauffage
- HeatingQC : Qualité et état du chauffage
- CentralAir : Climatisation centrale
- Electrical : Système électrique
- 1stFlrSF : Pieds carrés du premier étage

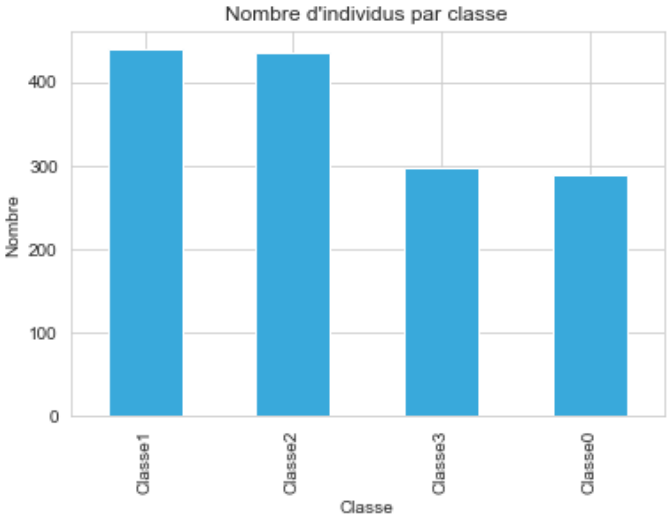
- 2ndFlrSF : Pieds carrés du deuxième étage
- LowQualFinSF : Pieds carrés finis de faible qualité (tous les étages)
- GrLivArea : surface habitable au -dessus du sol (pieds carrés)
- BsmtFullBath : salles de bain complètes au sous-sol
- BsmtHalfBath : Demi-salles de bain au sous-sol
- FullBath : salles de bain complètes au-dessus du sol
- HalfBath : demi-bains au-dessus du niveau du sol
- Bedroom : Nombre de chambres au-dessus du sous-sol
- Kitchen : Nombre de cuisines
- KitchenQual : Qualité de la cuisine
- TotRmsAbvGrd : Nombre total de chambres au-dessus du niveau du sol (n'inclut pas les salles de bains)
- Functional : cote de fonctionnalité de la maison
- Fireplaces : Nombre de cheminées
- FireplaceQu : Qualité du foyer
- GarageType : Emplacement du garage
- GarageYrBlt : Année de construction du garage
- GarageFinish : Finition intérieure du garage
- GarageCars : Taille du garage en capacité de voiture
- GarageArea : Taille du garage en pieds carrés
- GarageQual : Qualité de garage
- GarageCond : état du garage
- PavedDrive : allée pavée
- WoodDeckSF : Surface de terrasse en bois en pieds carrés
- OpenPorchSF : porche ouvert en pieds carrés
- EnclosedPorch : porche fermé en pieds carrés
- 3SsnPorch : porche trois saisons en pieds carrés
- ScreenPorch : surface du porche de l'écran en pieds carrés
- PoolArea : Espace piscine en pieds carrés
- PoolQC : qualité de la piscine
- Fence : qualité de la clôture
- MiscFeature : Fonctionnalité diverse non couverte dans d'autres catégories
- MiscVal : $ Valeur de la fonction diverse
- MoSold : Mois vendu
- YrSold : Année de vente
- SaleType : Type de vente
- SaleCondition : Condition de vente

2.1 Recodage de la variable d'intérêt en classes.

Dans cette partie, nous recodons la variable "**SalePrice**" (prix de vente), qui est une variable quantitative, en classes afin de simplifier l'étape de modélisation. Le but est de prédire dans quelle classe de prix sera le prix de la maison.

La première valeur ($**34900**) représente la plus petite valeur des prix et la dernière ($**775000**) représente la plus grande.
Cette variable deviendra celle à expliquer et "SalePrice" sera supprimée.

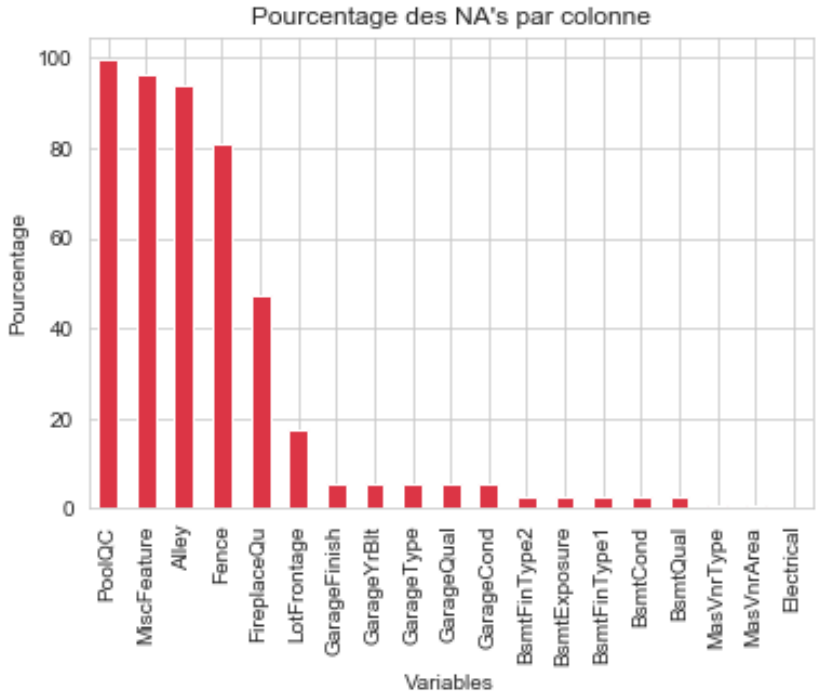
Ainsi, on peut représenter graphiquement la répartion des individus par classes.



On remarque que la **Classe1** et la **Classs2** contiennent beaucoup plus d'individus. La **Classe0** et la **Classs3** ont moins d'individus, très inférieurs aux deux autres Classes. Ceci est tout à fait normal car la **Classe0** correspond aux individus qui ont une maison à très bas prix, alors ils seront génés de donner les prix de leurs maisons. La **Classs3** correspond aux individus qui ont une maison très chère, alors beaucoup ne vont pas donner les prix de leurs maisons en ne voulant pas qu'on sache leurs revenus qui sont très élevés. La **Classe1** et la **Classs2** correspondent à ceux qui ont des maisons à prix moyen, donc il n'y a aucune raison de ne pas donner les prix de leurs maisons.

2.2 Traitement des données manquantes.

Notre table de données comporte plusieurs données manquantes. Certaines variables n'apportent pas une grande information. Ainsi, nous allons imputer les variables qui sont pertinentes et supprimer celles qui ont moins d'informations, c'est à dire plus de **50% de données manquantes**. Pour cela, voyons le pourcentage des valeurs manquantes de chacunes de ces variables.



Cette figure ci-dessus représente le pourcentage des valeurs manquantes de chacune des variables contenant des NaN. Nous avons 4 variables (**Alley, PoolQC, Fence et MiscFeature**) qui ont un pourcentage très élevé (supérieur à **50%**), nous pouvons les supprimées. Beaucoup de variables ont très peu de données manquantes,

la plupart sont des variables qualitatives. On voit aussi que certaines ont le même pourcentage de données manquantes.
Nous avons utilisé deux méthodes d'imputation :
-Le mode (Pour variables qualitatives) et le médiane (Pour variables quantitative)
-Imputation multiple en utilisant MissForest.

2.3 Statistiques descriptives.

Nous allons afficher une partie des statistiques descriptives de notre table de données.

	LotFrontage	LotArea	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	69.863699	10516.828082	1971.267808	1984.865753	103.117123	443.639726	46.549315	567.240411	1057.429452
std	22.027677	9981.264932	30.202904	20.645407	180.731373	456.098091	161.319273	441.866955	438.705324
min	21.000000	1300.000000	1872.000000	1950.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	60.000000	7553.500000	1954.000000	1967.000000	0.000000	0.000000	0.000000	223.000000	795.750000
50%	69.000000	9478.500000	1973.000000	1994.000000	0.000000	383.500000	0.000000	477.500000	991.500000
75%	79.000000	11601.500000	2000.000000	2004.000000	164.250000	712.250000	0.000000	808.000000	1298.250000
max	313.000000	215245.000000	2010.000000	2010.000000	1600.000000	5644.000000	1474.000000	2336.000000	6110.000000

	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath	BedroomAbvGr	TotRmsAbvGrd
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	1162.626712	346.992466	5.844521	1515.463699	0.425342	0.057534	2.866438	6.517808
std	386.587738	436.528436	48.623081	525.480383	0.518911	0.238753	0.815778	1.625393
min	334.000000	0.000000	0.000000	334.000000	0.000000	0.000000	0.000000	2.000000
25%	882.000000	0.000000	0.000000	1129.500000	0.000000	0.000000	2.000000	5.000000
50%	1087.000000	0.000000	0.000000	1464.000000	0.000000	0.000000	3.000000	6.000000
75%	1391.250000	728.000000	0.000000	1776.750000	1.000000	0.000000	3.000000	7.000000
max	4692.000000	2065.000000	572.000000	5642.000000	3.000000	2.000000	8.000000	14.000000

Dans la table ci-dessus, nous avons fait un résumé de certaines variables quantitatives du jeu de données.

Nous allons visualiser la corrélation entre les variables quantitatives. On représente ainsi cette corrélation sous forme de triangle.

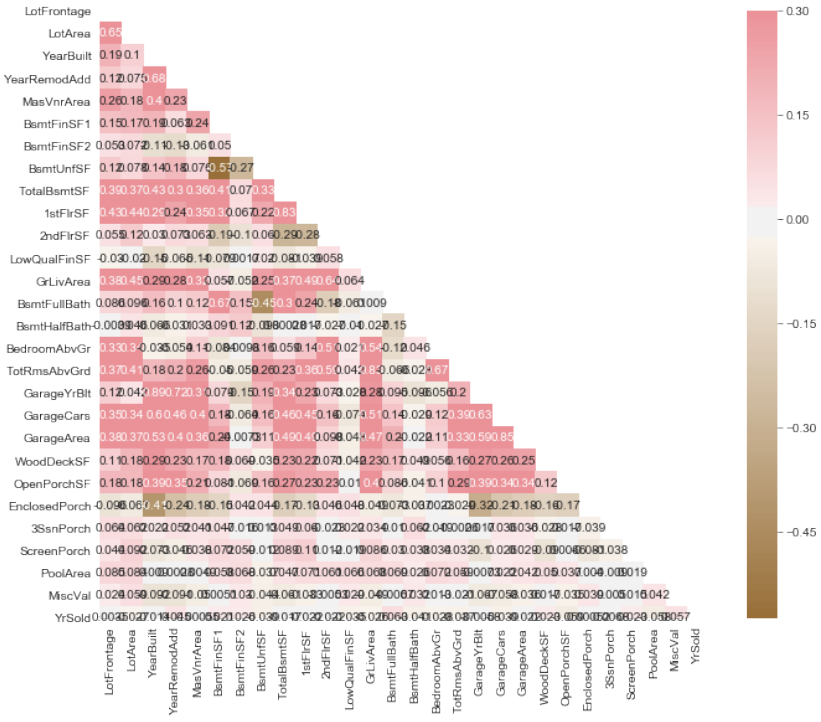
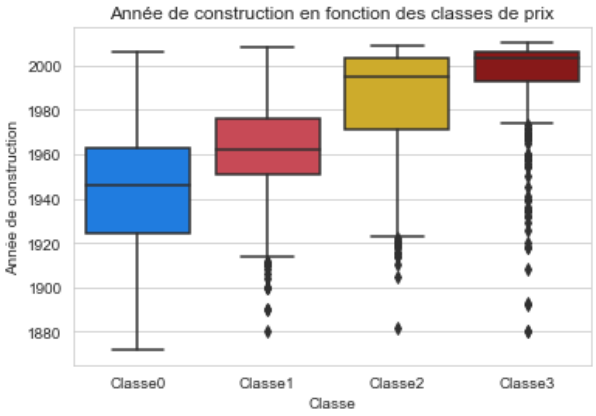


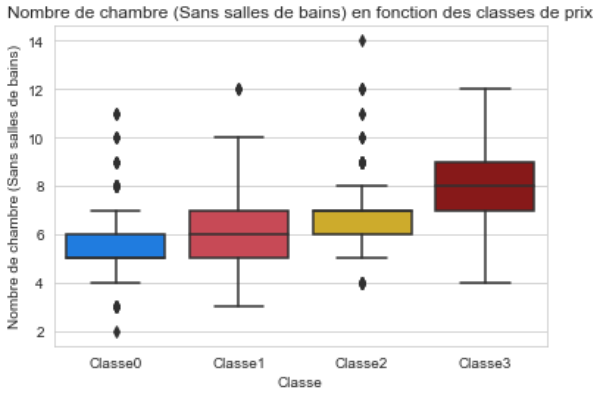
Figure 1: Matrice de corrélation (**spearman**)

Vu que nous avons beaucoup d'observation et d'après cette figure ci-dessus, nous pouvons affirmer que les variables sont corrélées entre elles. Certaines négativement et d'autres positivement. En outre il y'a un petit

nombre de variables qui sont faiblement corrélées. Dans cette figure, le marron représente les variables qui sont corrélées négativement et le rouge positivement. On voit clairement que les variables "**Lotfrontage**" et "**LotArea**" sont les plus fortement corrélées et les variables "**YrSold**" et "**MiscVal**" les plus faiblement corrélées.

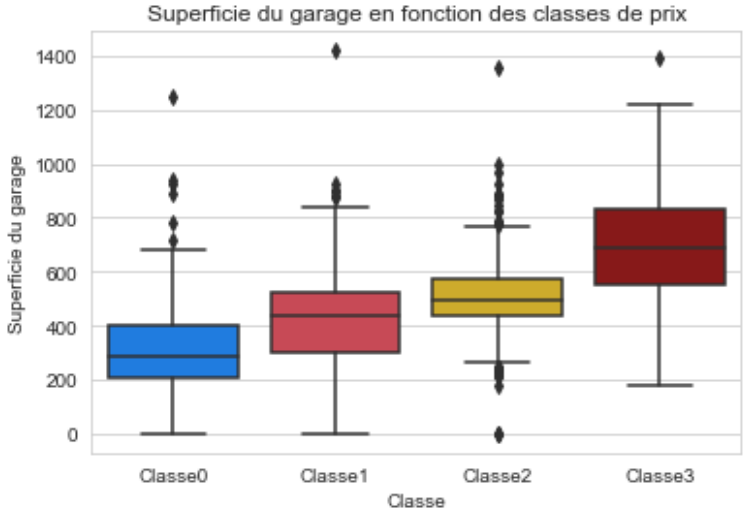
Nous allons chercher maintenant l'influence de certaines variables sur les différentes classes de prix. Pour cela, on utilise les box-plot afin d'identifier les valeurs extrêmes et voir la répartition des observations selon les classes.





La figure à gauche ci-dessus représente l'année de construction en fonction des classes prix. On remarque immédiatement que les épaisseurs médianes de ces quatre groupes sont différentes. L'année de construction correspondant à la Classe3 possède la plus grande médiane de l'ordre de 2005 environ, ensuite s'en suit la Classe2, Classe1 et Classe0. On voit aussi que dans la Classe0, 100% des observations se trouvent à l'intérieur de la boîte. On voit aussi un nombre importants de valeurs aberrantes à la Classe3 ainsi qu'au Classe2.

La figure à droite ci-dessus représente le nombre de chambre (sans salles de bains) en fonction des classes prix. On remarque que les épaisseurs médianes de ces quatre groupes sont aussi différentes. Le nombre de chambre médiane corresponadant à la Classe3 est égale à 8 et 100% des observations se trouvent à l'intérieur de la boîte. La Classe1 possède 99% des observations à l'intérieur de la boîte. On voit aussi peu de valeurs aberrantes à la Classe2 ainsi qu'au Classe0.



Cette figure ci-dessus représente la superficie du garage en fonction des classes prix. On remarque aussi que les épaisseurs médianes de ces quatre groupes sont aussi différentes. La superficie médiane correspondant à la Classe3 est de 680 environ et 99% des observations se trouvent à l'intérieur de la boîte. On observe beaucoup de valeurs aberrantes à la Classe3.

Illustrons maintenant la relation entre la classe de prix et la zone du bien :

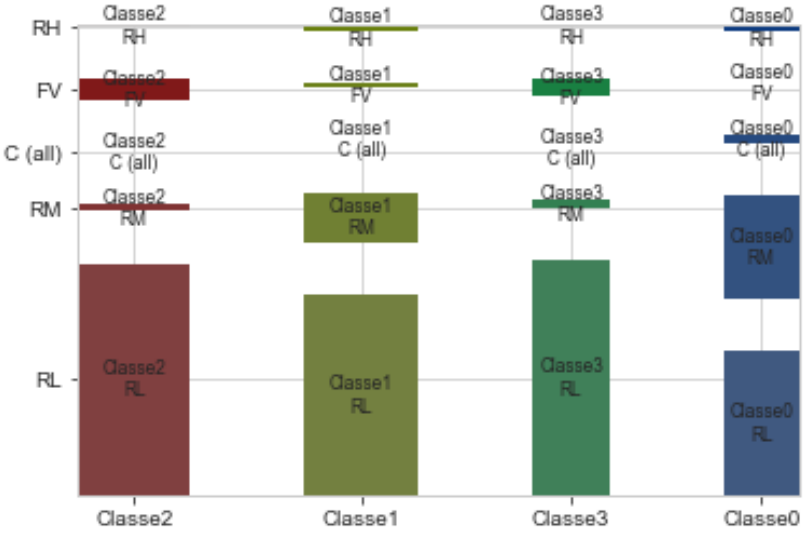


Figure 2: Classification générale de zonage en fonction des classes de prix

En classe 0, la plupart des biens sont soit des Résidentiels à basse densité (RL) ou à moyenne densité (RM), et très peu de biens commerciaux sont accessibles dans cette classe. Plus on augment la classe plus la part des Résidentiels à basse densité augmente et la part des Résidentiels à moyenne densité diminue.
Concernant les villages flottant résidentiels (PV), ils augmentent aussi en augmentant la classe, tout en restant relativement faible dans toutes les classes.

3 Modélisation.

3.1 Sélection des variables

Le but de la sélection des variables est de réduire le modèle aux variables explicatives les plus pertinentes.Pour cela nous allons implémenter la méthode **RFE** de **scikit-learn** : elle élimine au fur et à mesure les coefficients les plus faibles en valeur absolue, et s'arrête quand on arrive à la moitié ou à un nombre spécifié de variables.

Ainsi, on donne un tableau des **10 premières** variables sélectionnées avec la méthode d'imputation correspondante.

	Imputation_mode_mediane	Imputation_MissForest
0	MSZoning	MSZoning
1	Street	Street
2	LotShape	LotShape
3	LandContour	LandContour
4	Utilities	LandSlope
5	LandSlope	Condition1
6	Condition1	Condition2
7	Condition2	BldgType
8	BldgType	OverallQual
9	OverallQual	OverallCond

Figure 3: Variables sélectionnées après l'imputation par les deux méthodes

Avec cette méthode de sélection, nous avons passé de **76 variables** à **37 variables**. On observe aussi que les variables sélectionnées ne sont pas les mêmes, on a **22 variables** identiques pour les deux méthodes d'imputation.

3.2 K plus proches voisins.

Nous allons utiliser la fonction K plus proche voisin (**KNN**) avec les paramètres par défauts. Ensuite le nombre de plus proches voisins optimaux en utilisant la validation croisée.

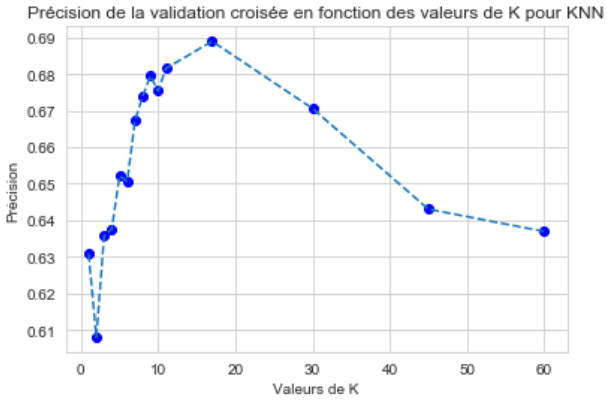


Figure 4: Imputation par médiane

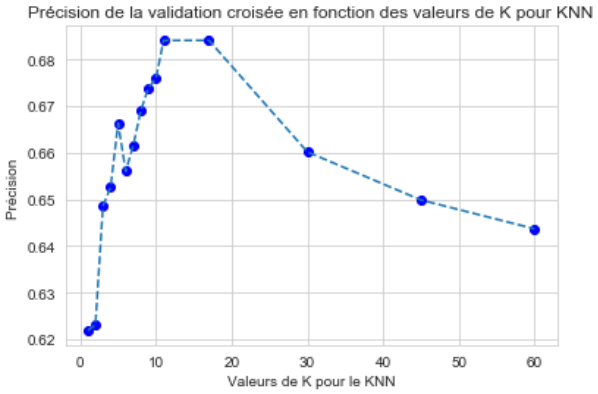


Figure 5: Imputation par MissForest

Les deux graphes ci-dessus représentent la précision de la validation croisée en fonction des paramètres K du K plus proches voisins. Pour les deux méthodes, nous avons obtenus le même paramètre optimal qui est égal à **17**.

Imputation par médiane	Train	Test
Nombre de plus proches voisins	Précision	Précision
5	0.7476	0.6690
17	0.7074	0.6804

Imputation par MissForest	Train	Test
Nombre de plus proches voisins	Précision	Précision
5	0.7613	0.6758
17	0.7123	0.6598

Dans la table ci-dessus est donnée les résultats de précision ainsi que le nombre de paramètres par défaut qui est **5** et le nombre optimal qui est égal à **17** pour chaque méthode d'imputation.
On obtient une meilleur précision en imputant les données manquantes avec le mode/médiane et en utilisant **17** voisins.

$$\begin{bmatrix} 50 & 41 & 2 & 0 \\ 5 & 92 & 17 & 1 \\ 1 & 27 & 100 & 4 \\ 0 & 1 & 41 & 56 \end{bmatrix}$$

Imputation-Mode/médiane

$$\begin{bmatrix} 50 & 42 & 1 & 0 \\ 7 & 91 & 16 & 1 \\ 0 & 29 & 96 & 7 \\ 0 & 2 & 43 & 53 \end{bmatrix}$$

Imputation-MissForest

Dans ces deux matrices de confusions ci-dessus, nous remarquons sur la diagonale que la plupart des biens sont classés dans la bonne classe de prix. On voit sur la matrice triangulaire supérieure et la la matrice triangulaire inférieure que les valeurs sont relativement grandes, ce ne modèle distingue pas le passage entre une classe et la classe précédente et une classe et la classe supérieure. Sur la diagonale, on a les bonnes prédictions et tout ce qui est hors diagnale ce sont les mauvaises prédictions. Sur un total de **438**, nous avons une prédiction de **298** (**68%**) dans la première matrice et **284** (**66%**) dans la deuxième matrice.

3.3 Forêts aléatoires.

Dans cette partie, nous utilisons la méthode Random Forest (**randomForest**) avec les paramètres par défauts. Ensuite le nombre de paramètres optimaux en utilisant la validation croisée.
Regardons l'importance des variables dans le modèle de forets aléatoires (seules les variables dont l'importance est supérieure à **3%** seront représentées) :

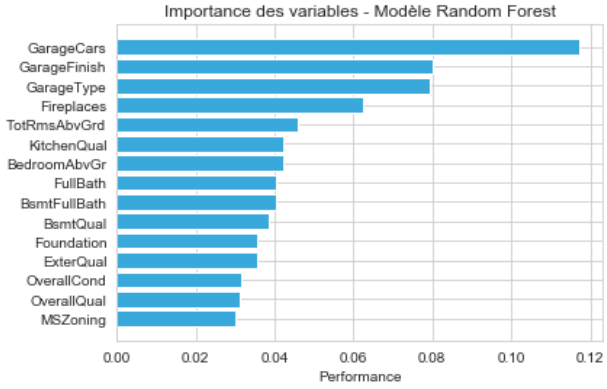


Figure 6: Imputation par médiane

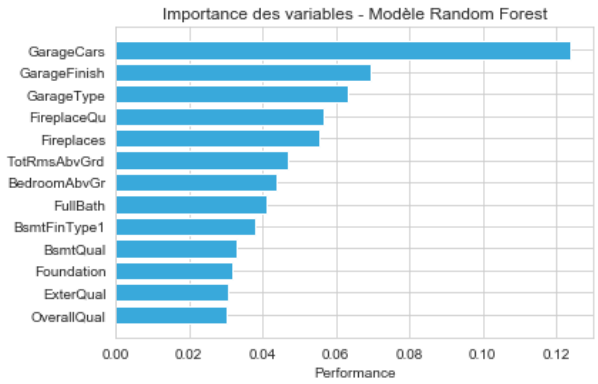


Figure 7: Imputation par MissForest

On remarque que les variables importantes sont différentes d'une méthode d'imputation à l'autre. On voit aussi qu'on a pas le même nombre de variables à imputer selon chaque type de méthode d'imputation.

Imputation par médiane	Train	Test
Nombre de paramètres	Précision	Précision
10	0.9775	0.7101
950	0.9941	0.7283

Imputation par MissForest	Train	Test
Nombre de paramètres	Précision	Précision
10	0.9843	0.7055
550	0.9990	0.7352

Dans la table ci-dessus est donnée les résultats de précision ainsi que le nombre de paramètres par défaut qui est **10** et le nombre de paramètres optimaux pour chaque méthode d'imputation.
Pour les deux méthodes, nous avons obtenus des paramètres optimaux différents. La méthode d'imputation par médiane donne un paramètre optimal égal à **950** alors que celle de MissForest donne **550**.
On obtient une meilleur précision en imputant les données manquantes avec la méthode MissForest et en utilisant 550 arbres.

$$\begin{bmatrix} 65 & 27 & 1 & 0 \\ 12 & 92 & 10 & 1 \\ 1 & 34 & 89 & 8 \\ 0 & 2 & 17 & 79 \end{bmatrix}$$

Imputation-Mode/médiane

$$\begin{bmatrix} 62 & 30 & 1 & 0 \\ 8 & 93 & 13 & 1 \\ 0 & 34 & 92 & 6 \\ 0 & 1 & 23 & 74 \end{bmatrix}$$

Imputation-MissForest

3.4 Gradient Boosting.

Dans cette partie, nous utilisons l'algorithme de Gradient Boosting (**GB**) avec les paramètres par défauts. Ensuite le nombre de paramètres optimaux en utilisant la validation croisée.
Regardons l'importance des variables dans le modèle de Gradient Boosting (seules les variables dont l'importance est supérieure à **3%** seront représentées) :

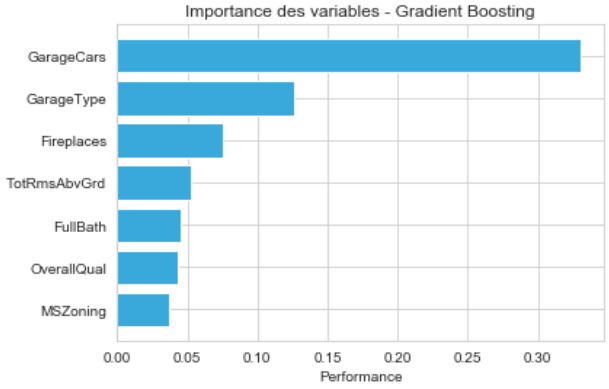


Figure 8: Imputation par médiane

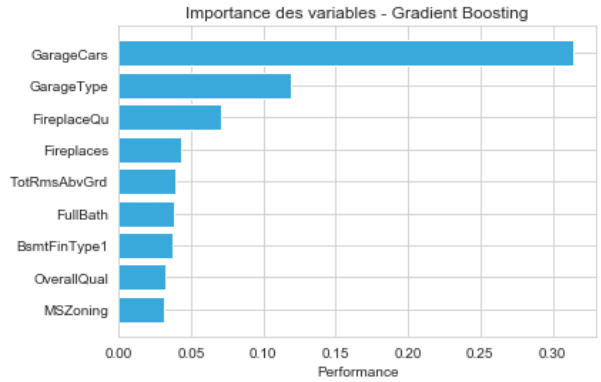


Figure 9: Imputation par MissForest

On remarque que les variables importantes sont différentes d'une méthode d'imputation à l'autre. On voit aussi qu'on a pas le même nombre de variables à imputer selon chaque type de méthode d'imputation.

Imputation par médiane	Train	Test
Nombre de paramètres	Précision	Précision
100	0.9207	0.7374
60	0.8845	0.7192

Imputation par MissForest	Train	Test
Nombre de paramètres	Précision	Précision
100	0.9325	0.7443
60	0.8982	0.7352

Dans la table ci-dessus est donnée les résultats de précision ainsi que le nombre de paramètres par défaut qui est **100** et le nombre optimal qui est égal à **60** pour chaque méthode d'imputation.
On obtient une meilleur précision en imputant les données manquantes avec la méthode MissForest et en utilisant 100 étapes de boosting.

$$\begin{bmatrix} 67 & 25 & 1 & 0 \\ 12 & 78 & 22 & 3 \\ 1 & 29 & 94 & 8 \\ 0 & 3 & 20 & 75 \end{bmatrix}$$

Imputation-Mode/médiane

$$\begin{bmatrix} 64 & 28 & 1 & 0 \\ 10 & 88 & 14 & 3 \\ 0 & 26 & 97 & 9 \\ 0 & 0 & 23 & 75 \end{bmatrix}$$

Imputation-MissForest

3.5 Adaboost.

Dans cette partie, nous utilisons la fonction **Adaboost** avec les paramètres par défauts. Ensuite le nombre de paramètres optimaux en utilisant la validation croisée.

Imputation par médiane	Train	Test
Nombre de paramètres	Précision	Précision
50	0.6135	0.6233
70	0.6184	0.6187

Imputation par MissForest	Train	Test
Nombre de paramètres	Précision	Précision
50	0.6253	0.6096
60	0.6145	0.6164

Dans la table ci-dessus est donnée les résultats de précision ainsi que le nombre de paramètres par défaut qui est **50** et le nombre de paramètres optimaux pour chaque méthode d'imputation.
Pour les deux méthodes, nous avons obtenus des paramètres optimaux différents. La méthode d'imputation par médiane donne un paramètre optimal égal à **70** alors que celle de MissForest donne **60**.
On obtient une meilleur précision en imputant les données manquantes avec le mode/médiane et en utilisant 50 estimateurs.

$$\begin{bmatrix} 80 & 12 & 1 & 0 \\ 52 & 41 & 20 & 2 \\ 3 & 21 & 70 & 38 \\ 0 & 1 & 17 & 80 \end{bmatrix}$$

Imputation-Mode/médiane

$$\begin{bmatrix} 68 & 23 & 2 & 0 \\ 38 & 52 & 24 & 1 \\ 4 & 30 & 86 & 12 \\ 0 & 3 & 20 & 75 \end{bmatrix}$$

Imputation-MissForest

4 Conclusion et points à améliorer.

L'objectif de ce projet était de prédire dans quelle classe de prix se trouverait le prix d'une maison. On a utilisé deux méthodes d'imputation et quatres modèles de classification et celui du gradient boosting donne la meilleure prédiction avec **73%** sur l'échantillon test en utilisant l'imputation mode/médiane.
Les variables les plus influentes sur la classe prix sont l'emplacement du garage, la taille du garage (en capacité de voiture), le nombre de cheminés, le nombre total de chambre au-dessus du niveau du sol (n'inclu pas les salles de bains), salles de bains complètes en-dessus du sol, la qualité globale du matériau et de la finition et la classification générale de zonage.

Point à améliorer

Nous aurions pu améliorer les modèles utilisés en intégrant la localisation des biens, car, cette dernière influence le prix du bien (*Le prix d'un bien au centre ville restera toujours plus cher qu'un autre dans les alentours*).
En intégrant de nouvelles variables comme les sous-marchés de Boston, le taux de vacances du bien, la localisation des monuments de la ville, nous pouvons améliorer notre prédiction.