École nationale supérieure d'informatique pour l'industrie et l'entreprise

House Prices: Advanced Regression Techniques

Réalisé par:

Malek BEN NEYA et Karim ASSAAD

21/04/2017

Table des matières

Table des figures

I
Introduction

L'apprentissage automatique ou apprentissage statistique (machine learning en anglais), champ d'étude de l'intelligence artificielle, concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques à remplir par des moyens algorithmiques plus classiques.
L'apprentissage automatique est la science permettant aux ordinateurs d'accomplir des tâches sans avoir été explicitement programmé dans ce sens. Dans les dernières décennies, l'apprentissage automatique a donné naissance aux véhicules sans conducteurs, à la reconnaissance de la parole, la recherche web performante et a largement contribué à l'amélioration de la compréhension du génome humain. L'apprentissage automatique est tellement présent aujourd'hui que vous l'utilisez probablement des dizaines de fois par jour sans même vous en rendre compte. Beaucoup de chercheurs pensent également qu'il s'agit du meilleur moyen de progresser vers une intelligence artificielle au niveau des humains.
L'apprentissage est un axe de recherche très étendu qui porte l'intérêt de plusieurs communautés de scientifiques. La théorie de l'apprentissage englobe un très grand nombre de modèles que l'on pourrait décrire de la manière suivante : Ayant à notre disposition un ensemble d'observations sur un phénomène, ici, notre base de départ est le projet Kaggle intitulée « House Prices : Advanced Regression Techniques » .
nos observations se présentent sous la forme du couple (xi,y) où xi représente la variable d'entrée et y la variable de sortie ou de réponse. Notre objectif le plus ultime est de fournir une prédiction qui se rapproche le plus des données de test en adoptant une méthodologie bien déterminée.

I.1 Présentation de la problématique

Cette compétition nous permet de déterminer un prix de vente de la maison de nos rêves en fonction de plusieurs critères et plusieurs aspects qui décrivent les maisons qui sont les variables utilisées dans cette base de données et qu'on décrira plus tard. Nous avons choisi comme nom de l'équipe " Team Sparta" .
En analysant 79 variables qui décrivent chaque aspect des résidences à lowa, en faisant le nettoyage adéquat, et en créant des modèles différents, nous allons parvenir à spécifier la variable à prédire qui est le prix de vente de chaque maison.
Nos données contiennent des données qualitatives et d'autres quantitatives :

- **SalePrice** le prix de vente de la propriété en dollars : c'est la variable à prédire.
- **MSSubClass :** La classe du batiment.
- **MSZoning :** Le classement général du zonage
- **LotFrontage :** Pieds linéaires de rue connectés à la propriété
- **LotArea :** Taille du terrain en pieds carrés
- **Street :** Type d'accèsroutier
- **Alley :** Type d'accès d'allée
- **LotShape :** Forme générale de la propriété
- **LandContour :** La planéité de la propriété

— **Utilities :** Types d'utilités valables
— **LotConfig :** Configuration du lot
— **LandSlope :** Pente de la propriété
— **Neighborhood :** Emplacements physiques dans les limites de la ville d'Ames
— **Condition1 :** Proximité de la route principale ou du chemin de fer
— **Condition2 :** Proximité de la route principale ou du chemin de fer (si la deuxième est présente)
— **BldgType :** Type d'habitation
— **HouseStyle :** Style d'habitation
— **OverallQual :** Matériel global et qualité de finition
— **OverallCond :**Évaluation de l'état général
— **YearBuilt :** Date originale de construction
— **YearRemodAdd :** Date de remodelage
— **RoofStyle :** Type de toiture
— **RoofMatl :** mtériel de Toiture
— **Exterior1st :** Revêtement extérieur sur maison
— **Exterior2nd :** Revêtement extérieur sur maison (si plus d'un matériel)
— **MasVnrType :**Type de plaçage de maçonnerie
— **MasVnrArea :**Zone de placage de maçonnerie en pieds carrés
— **ExterQual :** Qualité du matériau extérieur
— **ExterCond :** Condition actuelle du matériau à l'extérieur
— **Foundation :** Type de fondation
— **BsmtQual :** Hauteur du sous-sol
— **BsmtCond :** Etat général du sous-sol
— **BsmtExposure :** Parois de sous-sol au niveau de la marche ou du jardin
— **BsmtFinType1 :** Qualité du sous-sol fini
— **BsmtFinSF1 :** pieds carrés finis de type 1
— **BsmtFinType2 :** Qualité de la deuxième zone finie (si présent)
— **BsmtFinSF2 :** pieds carrés finis de type 2
— **BsmtUnfSF :** Pieds carrés inachevés de sous-sol
— **TotalBsmtSF :** Total pieds carrés de sous-sol
— **Heating :** Type de chauffage
— **HeatingQC :** Qualité et condition de chauffage
— CentralAir : Climatisation centrale
— Electrical : système électrique
— 1stFlrSF : Premier pied carré
— 2ndFlrSF : Deuxième étage pieds carrés

— LowQualFinSF : Pieds carrés finis de qualité (tous étages)
— GrLivArea : Au dessus de la catégorie, surface habitable pieds carrés
— BsmtFullBath : Sous-sol salle de bains complètes
— BsmtHalfBath : moitié salle de bains sous-sol
— FullBath : Salles de bains complètes au-dessus de la qualité
— HalfBath : Demi-bains au-dessus de la qualité
— Bedroom : Nombre de chambres au-dessus du niveau du sous-sol
— Kitchen : Nombre de cuisines
— KitchenQual : qualité de cuisines
— TotRmsAbvGrd : Total des pièces au-dessus de la catégorie (ne comprend pas les salles de bains)
— Functional : Évaluation de la fonctionnalité à domicile
— Fireplaces : Nombre de cheminées
— FireplaceQu : qualité de cheminées
— GarageType : Emplacement du garage
— GarageYrBlt : Année de construction de garage
— GarageFinish : Finition intérieure du garage
— GarageCars : Taille du garage dans la capacité de la voiture
— GarageArea : Taille du garage en pieds carrés
— GarageQual : qualité du garage
— GarageCond : condition du garage
— PavedDrive : Voie pavée
— WoodDeckSF : Zone de pont en bois sur pieds carrés
— OpenPorchSF : Aire de porche ouverte en pieds carrés
— EnclosedPorch : Zone de porche fermée en pieds carrés
— 3SsnPorch : Zone de portique de trois saisons en pieds carrés
— ScreenPorch : Surface de portique d'écran en pieds carrés
— PoolArea : Zone de piscine en pieds carrés
— PoolQC : qualité de piscine
— Fence : Qualité de la clôture
— MiscFeature : Fonctionnalités diverses non couvertes par d'autres catégories
— MiscVal : Valeur de la fonctionnalité diverse
— MoSold : Mois vendu
— YrSold : année vendu
— SaleType : type de vente
— SaleCondition : condition de vente

II

Analyse descriptives

II.1 Etude univarié

II.1.1 Variable Réponse SalePrice

Notre variable prix des maison est une variable quantitative qui indique le prix associé à chaque maison selon plusieurs critéres.
Durant notre étude nous allons exposer les caractéristique de cette variable dans un premier temps, puis nous allons nous intérésser aux différentes dépendence et linéarités avec le reste des variable.

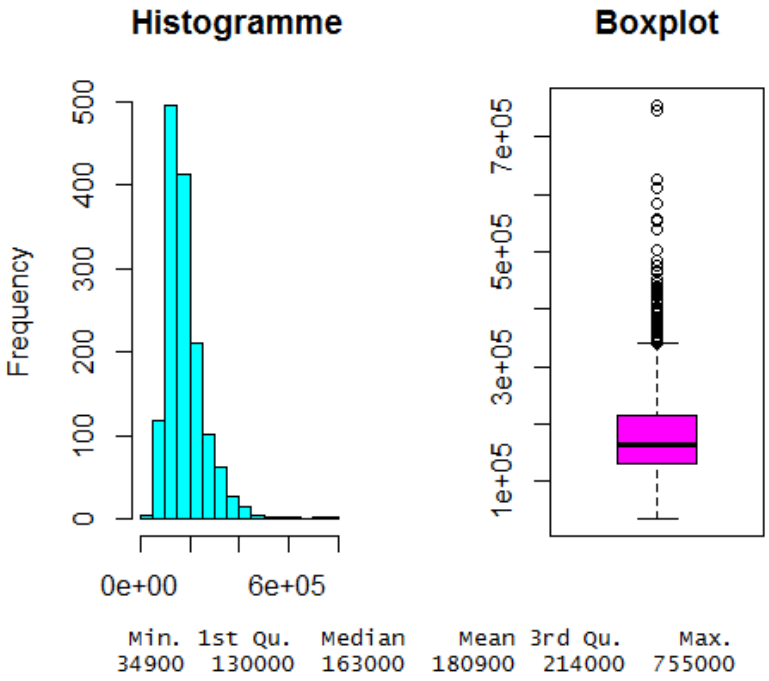

Figure II.1: Caractéristique de la variable Prix

D'aprés l'allure de l'histogramme on peut dire que notre variable réponse n'est pas exactement gaussienne, une transformation en log impliquera peut être une amélioration de cet allure.

II.1.2 Boxplots unidimensionnels

Nous avons visualisé les boxplots de quelques variables qui sont presque sur la même échelle pour bien mettre en évidence leurs valeurs caractéristiques (quartiles, médianes, valeurs aberrantes). vu l'importante quantité de variable nous avons choisi alétoirement les plus pertinant selon notre logique.

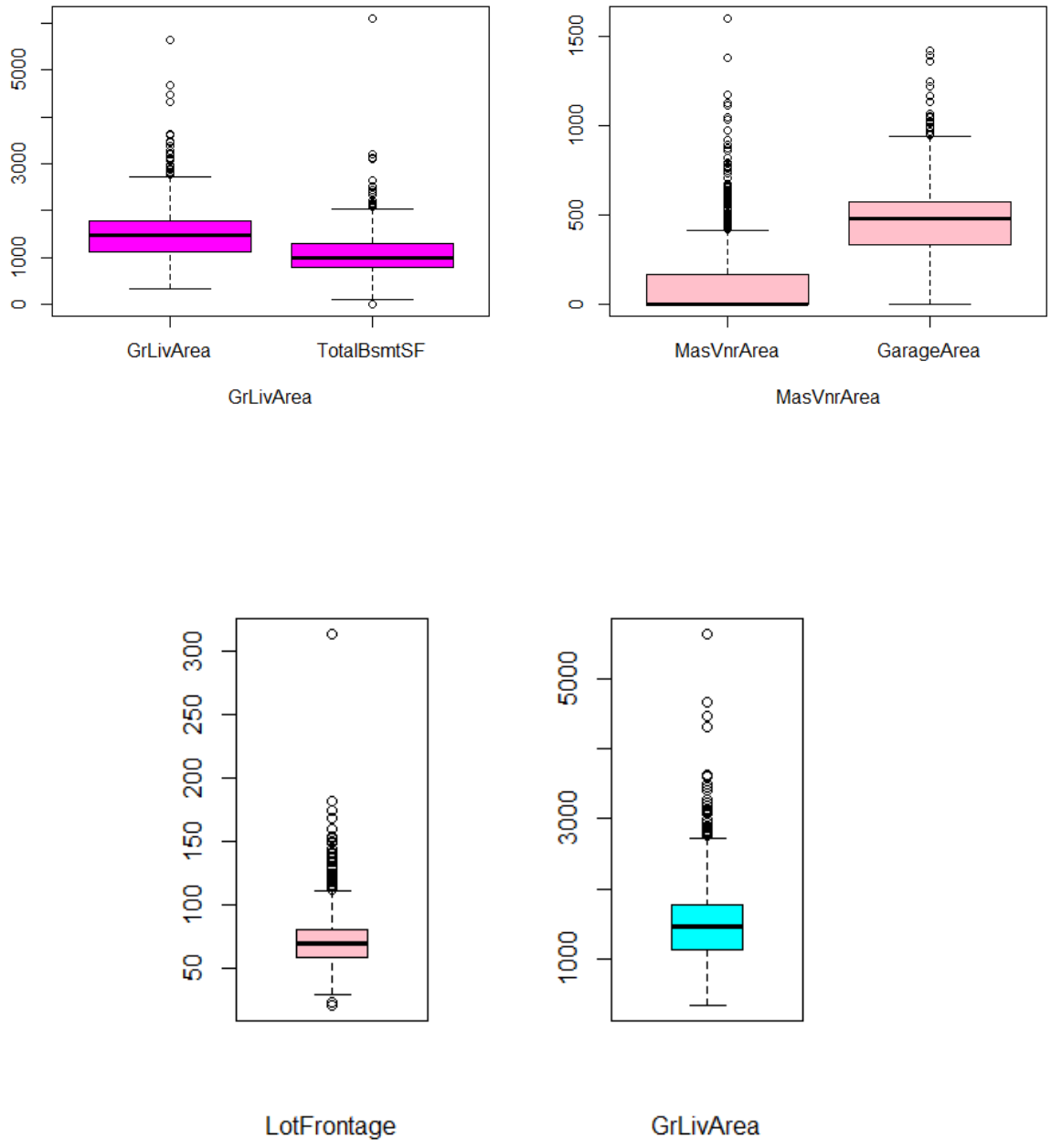

Figure II.2: Boxplots

On trouve que la moyenne de zone de placage de maçonnerie est presque nulle avec quelques cas qui dépasse les 500 métres carré. Le plupart des maisons ont la surface de garage à 500 métres carré.

II.1.3 Diagrammes circulaires

D'après ces graphiques on peut dire que la qualité et la condition de la plus grande majorité des maisons est une qualité moyenne de 5 et 6.

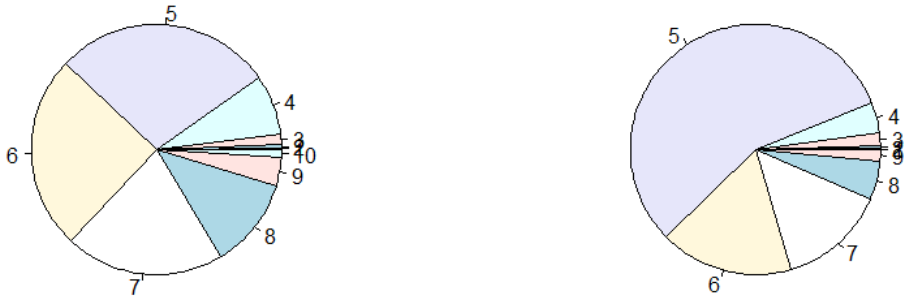

Figure II.3: overkal et overcond

Figure II.4: kitchenkal et bsmtkal

Concernant la qualité du 'basment' et qualité de cuisine la plus part des maisons sont attribuées à un niveau bien et moyen.

II.2 Etude Bivarié

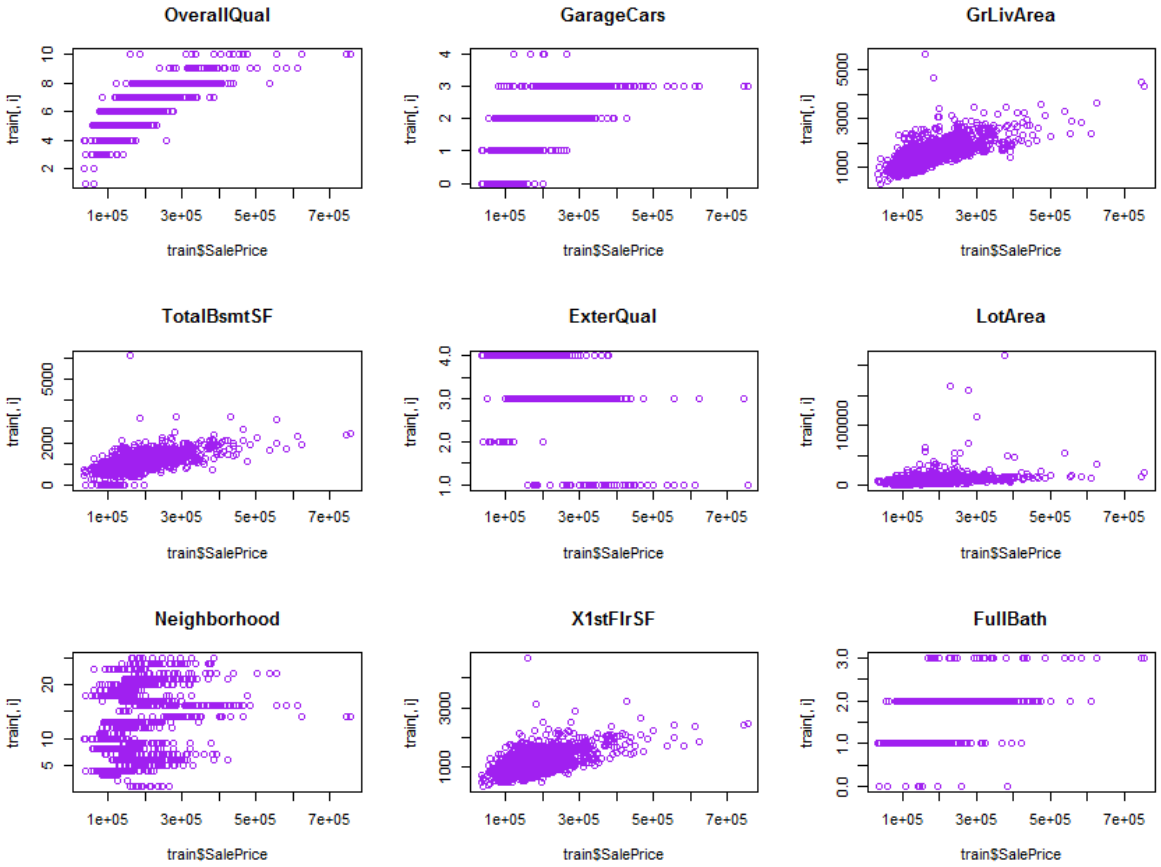

Figure II.5: Neighberhood

D'une part , on remarque que la distribution de nuages de points pour quelques variables est linéaire par rapport à la variation de SalePrice avec différence de densité de répartition de nuages de points , parmi les variables qui ont une distribution linéaire on trouve : OveralQual , GarageCars , ExterQual , Neighborhood, FullBath .

D'autre part , certaines variables ont une distribution de nuages de points qui évolue exponentiellement en fonction de variation de SalePrice mais avec une différence de vitesse de croissance , on trouve comme exemple : GrLibArea , TotalBsmSF , X1stFlrSF.

Pour la variable LotArea , on remarque sur le nuage des points que la distribution de valeurs de variables LotArea est très dense aux alentours de 0 c'est à dire pour différentes valeurs de SalePrice , notre variable LotArea garde une valeur nulle .

II.3 Diagramme de chaleur

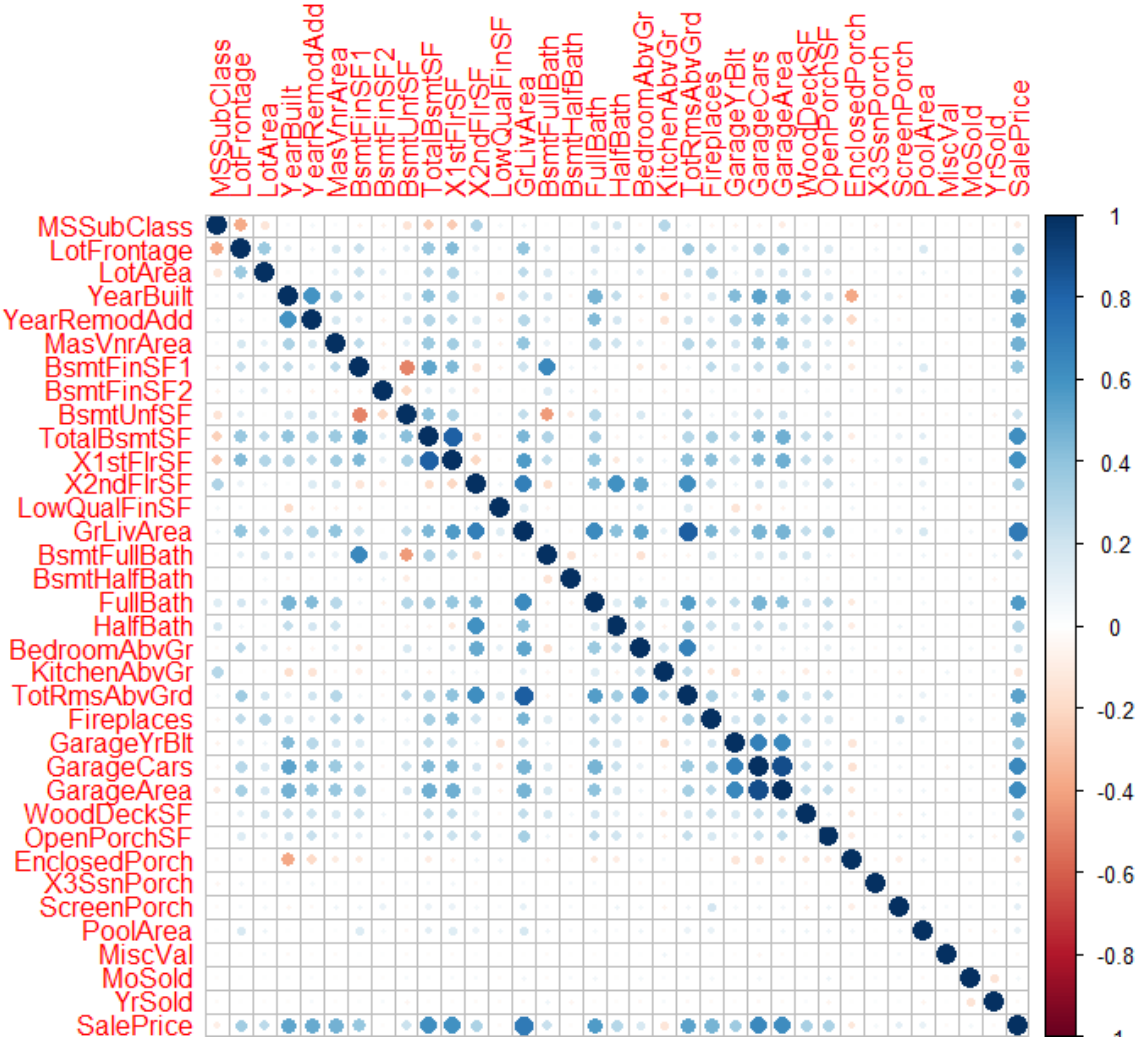

FIGURE II.6: Neighberhood

Le diagramme de chaleur joue un rôle trés important dans la séléction des variables pertinantes par la suite dans la construction des modéles car il démontre la corrélation entre les différentes varibles et surtout leurs correlation avec la variable cible qui est les prix des maisons, ici nous apercevons une forte corrélation positive de l'ordre de 0.8 entre le 'SalePrice' et la 'GrLivArea'.
La 'TotRmsABvGrd' est fortement corrélée avec la 'GrLivArea' avec une valeur de 0.7 donc ces deux varibles peuvent apporter la même information. Le 'BsmtUnfSF'est anti corrélé avec la 'BsmtFinSF1' donc logiquement ces variables ont des comportements symétriques , nous constatons aussi que le prix dépond fortement de la surface du garage avec une corrélation de 0.6.
d'une autre part il existe des prédicteurs dont la corrélation est presque nulle avec la variables réponse comme 'MoSold' et 'YrSold' donc on va peut être ne pas les prendre

en compte dans la partie modélisation.

III

Prétraitement des données

Le prétraitement des données est essentiel pour la suite du travail, il est constituée d'une partie remplacement de valeurs manquantes et d'une partie transformations de variables.

III.1 Remplacement des valeurs manquantes

III.1.1 Remplacement par des valeurs logique

A premiére vue, le nouveaux jeux de données présentent plusieurs valeurs manquantes, certaines colonnes sont même quasimenet vides, donc il a fallu comprendre la signification de ces variables bien particuliéres et trouver une solution logique pour les remplacer.

Pour les variables qualitatives qui ont des valeurs manquantes et qui correspondent à des qualités , nous avons remplacé chaque niveau de qualité par des chiffres qui montrent la différences de niveaux et par la suite nous avons mis 0 à la place des valeurs 'NA' qui signifient qu'il n'existe pas tel ou tel composants, nous avons appliqué cette méthode de remplacement pour les variables suivantes :
poolQC, fence, alley, FireplaceQu, MiscFeature, GarageCondition , GarageType, GarageFinish , GarageQual, BsmtQual, BsmtCond, BsmtExposure , BsmtFinType1, BsmtFinType2,BsmtFullBath, BsmtHalfBath, BsmtFinSF1 BsmtFinSF2, BsmtUnfSF, TotalBsmtSF.

III.1.2 Remplacement par moyenne ou fréquence

Cette méthode n'est pas trés conseillée pour ne pas biaiser l'information donc on l'a appliqué pour les variables qui ont un petit nombre de valeurs manquantes. Pour les variables qualitatives comme KitchenQual, Functional, SaleType, Utilities, BldgType, LotShape, MSZoning nous avons remplacé par la modalité la plus fréquente pour chaque variable. Pour les variables quantitatives comme 'GarageYrBlt', on a remplacé les inconnus donc les maisons qui n'ont pas de garage par une année qui n'existe pas dans le jeu de donnée (1600) pour éviter la perte de l'ordre de grandeur.

III.1.3 Remplacement par prédiction

Pour les variables quantitatives qui ont un énorme nombre de valeurs manquantes nous avons utilisé les arbres de décisions comme le cas de la variable MasVnrType.
Pour les variables quantitatives LotFrontage et MasVnrArea nous avons utilisé les arbres de régression pour faire la prédiction de l'échantillon manquant.

III.2 Transformations effectuées sur les variables

III.2.1 Combinaison de variables

Nous avons remarqué que quelques variables avaint exactement les memes modalités donc on à penser à les combiner ensemble et créer une troisiéme nouvelle variable.

Nous avons appliqué cette démarche sur les variables qui désignent le même objet comme Exterior1st et Exterior2nd donc on a introduit un '+' qui référe à un 'ou binaire' pour créer une nouvelle variable Exterior à 3 modalité :// 0, 1 et 2.

III.2.2 Transformation dummy

Nous avons traité les variables qualitatives en transformant une variable à n modalités à n variables binaires, de cette façon on assure la conservation de l'information et la facilité de son exploitation par les différents modèles.

IV
Construction des modèles

Une fois les données nettoyées, elles sont prêtes à être utilisées exploités par les algorithmes de prédiction, pour ce faire nous avons réalisé plusieurs modèle en commençant par le plus simple en allant jusqu'au plus complexe, le but étant de réaliser la meilleure performance.
Nous avons choisi de nous basées sur la qualité des données et non la quantité, en effets nous avons préféré fournir à nos modèles à chaque étapes les variables les plus pertinentes qui sont nés soit de notre sélection basé sur l'analyse descriptive des données ou selon les critères retournés par les modèles eux-mêmes donc le nombre de variables en entrée différent d'un modèle à un autre.

IV.1 Pénalisation Lasso

La pénalisation lasso appartient à la famille des méthodes de régularisation qui appliquent un rétrécissement sur l'espace des solutions pour empêcher les valeurs très grandes, pour permettre de centrer les valeurs au alentour de zéro ce qui favorise la standardisation des données au préalable. Cette méthode va donc modifier un petit peu la fonction de coût du problème initiale en la complétant par une fonction de pénalité. Cette pénalisation aura pour but de diminuer la distance entre les solutions possibles en se basant sur la distance de Manhattan.
Nous avons commencé par appliquer le modèle lasso, pour ce faire nous avons transformé notre data frame en une matrice de donnée et notre variable réponse en un vecteur, le taux d'erreur retourné par ce modèle est 0.15345.

IV.2 Random Forest

Nul doute que le Random Forest est l'un des algorithmes les plus populaires dans le monde de machine learning, c'est une méthode ensembliste qui se base sur une affectation aléatoire de sous échantillons ou on attribue à chaque arbre une vision parcellaire du problème, tant sur les observations en entrée que sur les variables à utiliser.
La prédiction dans notre problème de régression se fait par l'ensemble des arbres de décision en choisissant la moyenne des valeurs obtenues pour chaque nœud.
En appliquant cet algorithme nous avons du choisir minutieusement les paramètres d'entrées comme le nombre d'arbres optimal Ntree qui s'est stabilisé après la validation croisée à 500
Nous avons aussi implémenté notre modèle avec les variables les plus pertinentes selon le retour du paramètre 'importance' du model full et qui sont visualisé ci-dessous.

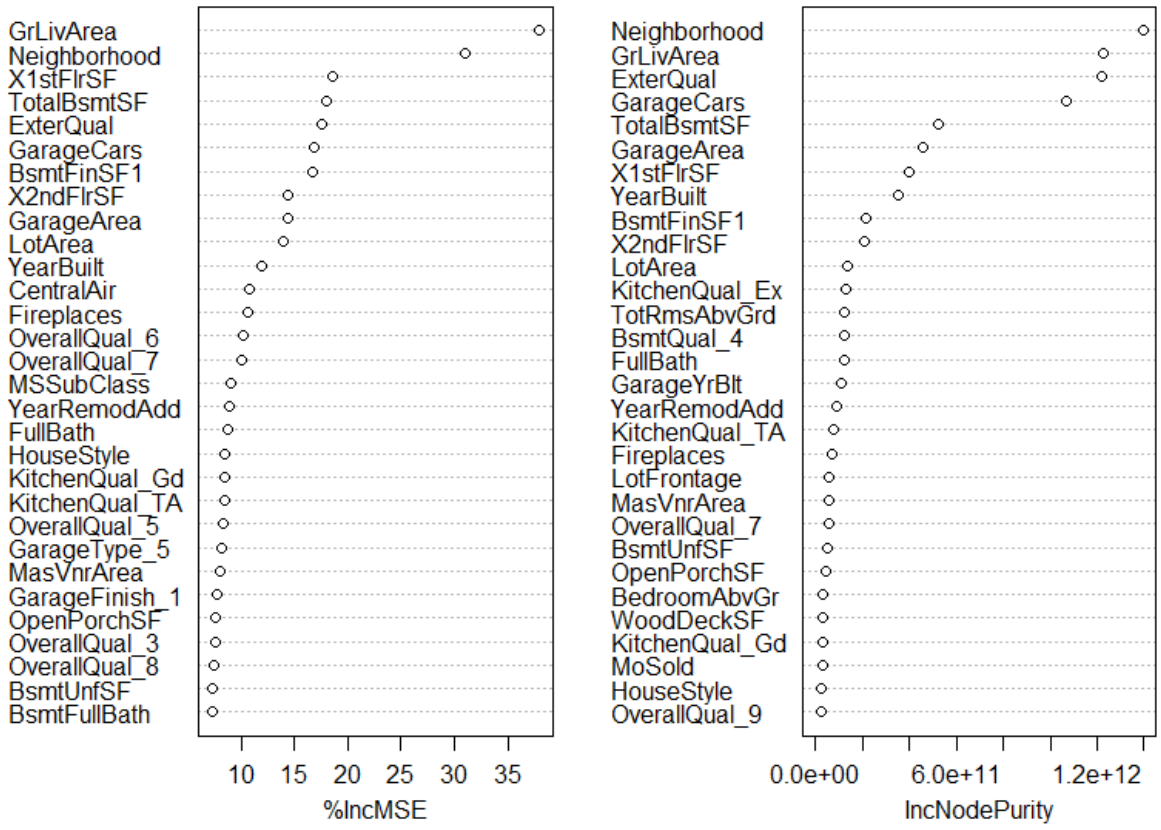

Figure IV.1: Importance des valeurs selon Random Forest

Le score obtenue en prenant en compte la première partie des variables plus importante selon le critère de Gini et un Ntree=500 est de 0.14603

IV.3 Conditional Random Forest

Les champs aléatoires conditionnels (conditional random forest ou CRFs) sont une classe de modèles statistiques utilisés en reconnaissance des formes et plus généralement en apprentissage statistique. Il sert à coder des relations connues entre observations et à construire des interprétations cohérentes. Il est souvent utilisé pour l'étiquetage ou l'analyse de données séquentielles, telles que le traitement du langage naturel ou les séquences biologiques et dans la vision par ordinateur. Plus précisément, les CRF trouvent des applications dans le marquage POS, l'analyse peu profonde, la reconnaissance de l'entité nommée, la recherche de gènes et la découverte de la région fonctionnelle critique des peptides parmi d'autres tâches, étant une alternative aux modèles cachés liés de Markov (HMM). Le score obtenue pour ce modéle est de 0.14783.

IV.4 Support Vector Machine

C'est un puissant algorithme qui repose essentiellement sur deux principes : La maximisation de la marge entre la frontière de décision et les exemples les plus proches, appelles vecteurs de support.
Le choix d'un hyperplan séparateur dans un nouvel espace de combinaison non linéaires entre les variables, dans lequel une séparation linéaire des individus sera possible.
Avant d'appliquer ce modèle nous avons centralisé et réduit les variables numériques pour améliorer la performance du SVM.
Le méta paramètres utilisées dans ce modéle est le cost qu'on a fixé à 10 ainsi que gamma à 0.5 selon les sortie de la fonction tune qui aide à choisir le méta paramètres.
Nous avons aussi attribué la valeur polynomiale au kernnel vu que le problème n'est pas tout à fait linéaire.
Le score rendu par ce modéle dans la plateforme kaggel est de 0.13653.

IV.5 Gradient Boosting

Pour essayer de simplifier le principe de cet algorithme, nous dirons qu'il est la combinaison du Descente de Gradient et du Boosting.
Pour bien paramétrer ce modéle il est important de bien choisir les méta paramétres et de comprendre leurs rôle dans l'implémentation du Gradient Boosting, pour ce fait on a itéré sur plusieurs modéles pour arriver à un compromis satisfaisant.
on a choisi une distributin gaussienne puisque notre variable prix est numérique,on a laissé n.tree à sa valeur par défaut donc à 100.
Concernant le paramétre interaction.depth qui signifie la profondeur des arbres et puisque le paramétre par défaut est 1 et qu'un seul arbre n'est pas logiquement sufissant pour prendre les désicions nous avons augmenté la valeur de ce paramétre à 10,pour le paramétre shrinkage c'est un peu délicat de trouver le bon paramétre sans tomber dans le sur apprentissage, pour cette raison nous lui avons attribué la valeur 0.1 qui reste une valeur acceptable.
Le retour de ce modéle est le plus performant car il a donnée une erreur égale à 0.12912 et un rang de 1106 sur Kaggel.

IV.6 Comparaison des modèles

Algorithme	Erreur
Pénalisation lasso	0.15345
Random Forest	0.14603
ConditionalRandom Forest	0.14783
Support Vector Machine	0.13653
Gradient Boosting	0.12912

La meilleure performance est obtenue par le modèle gradient boosting, en général tous les modèles sont assez robustes et dépondent fortement de leurs méta paramétres et de la pertinance des variables utilisées.

V
Conclusion

Au cours de notre projet, nous avons déployé une stratégie qui consiste tout d'abord à analyser les données à travers une étude descriptive sur les données et à nettoyer ces données en estimant les valeur manquantes .Enfin nous avons construit des modèles non linéaires ce qui nous as permis de prédire et d'améliorer les erreurs et les résultats de prédiction. Ce projet nous a donné l'opportunité de mettre en œuvre des différents outils d'apprentissage automatique.

Annexe

```r
#transofrmation binaire en modalite

binaryK<-function(base_data,variable){
  base<-base_data
  nb<-length(levels(base[[variable]]))
  interval<-1:nb
  for (i in levels(base[[variable]])){
    k=which(levels(base[[variable]])==i)
    var=base[[variable]]
    levels(var)[k] <- c('1')
    levels(var)[interval[which(interval!=k)]] <- rep(0,nb-1)
    base<-cbind(base,var)
    names(base)[names(base)=='var']<-paste(variable,"_",gsub(" ", "", i, fixed = TRUE), sep="")
  }
  base[[variable]]<-NULL
  return(base)
}
```

FIGURE V.1: Transofrmation binaire en modalite

```r
rf1<-randomForest(SalePrice~OverallQual+Neighborhood+  GrLivArea+  ExterQual+  GarageCars+ TotalBsmtSF+X1stFlrSF+
                    GarageArea+ KitchenQual+X2ndFlrSF+  BsmtQual+   YearBuilt+  BsmtFinSF1+ LotArea+
                    FullBath+   TotRmsAbvGrd+  YearRemodAdd+  FireplaceQu+MasVnrArea+ LotFrontage+GarageYrBlt+
                  GarageFinish+ GarageType+ BsmtUnfSF+OpenPorchSF+WoodDeckSF+ BsmtFinType1+  Fireplaces+
                    MoSold+ OverallCond+BedroomAbvGr+  HouseStyle+ BsmtExposure+  MSSubClass+ SaleCondition+
                  LandContour+BsmtFullBath+  MSZoning+CentralAir+ HeatingQC+  YrSold+ LotConfig+
                    Fence+ MasVnrType+ LotShape+HalfBath+ SaleType ,  data=train, importance=TRUE, ntree=500)
imp <-(rf1$importance[,2])
featureImportance <- data.frame(Feature=names(imp), Importance=imp)
```

FIGURE V.2: Random Forest

```r
svm3<-svm(SalePrice~ OverallQual    + Neighborhood+   GrLivArea+ GarageCars+ ExterQual+
                TotalBsmtSF+    X1stFlrSF+      GarageArea+     X2ndFlrSF+      BsmtFinSF1+
                YearBuilt+      FullBath+       LotArea+        GarageYrBlt+    TotRmsAbvGrd+
                YearRemodAdd+   KitchenQual_Ex+ BsmtQual_4+     MasVnrArea+     Fireplaces+
                LotFrontage      +   BsmtUnfSF+      KitchenQual_TA+ OpenPorchSF+    WoodDeckSF+
                OverallCond+    KitchenQual_Gd+ MoSold+         BedroomAbvGr+   Foundation+
                HouseStyle+     MSSubClass+     CentralAir+     SaleCondition+  BsmtFullBath+
                LandContour+    GarageType_5+   LotConfig+      HalfBath+       HeatingQC+
                ScreenPorch+    YrSold     +    MasVnrType+     BsmtExposure_2+ RoofStyle+
                Exterior1st_HdBoard +BsmtExposure_4+ RoofMatl+       MSZoning_RM+    Condition1+
                LandSlope+  SaleType_WD+    Exterior1st_VinylSd +GarageFinish_1+ GarageFinish_2+
                EnclosedPorch+  BsmtQual_3+  Exterior1st_BrkFace+KitchenAbvGr+   MSZoning_RL+
                FireplaceQu_4+ GarageFinish_3+ ExterCond+   SaleType_New+ LotShape_Reg+
                PoolArea+ BsmtFinSF2+ BsmtFinType1_6+ LotShape_IR1+ BsmtFinType1_3+
                BsmtHalfBath+BldgType_1Fam +  Exterior1st_MetalSd +LotShape_IR2+  Exterior1st_WdSdng +
                BsmtFinType1_5+ GarageType_4+BsmtFinType1_2+ PavedDrive+BsmtFinType1_1+
                FireplaceQu_5+  PoolQC_3+  Fence_3+  Exterior1st_CemntBd + LowQualFinSF+
                Functional_Typ+ Exterior1st_Plywood +Alley_2+ BsmtFinType1_4+ BldgType_Duplex
              +Exterior1st_WdShing +Electrical  +
                BsmtCond_4    + X3SsnPorch+     BsmtCond_2+     BsmtFinType2_4  +
                BsmtExposure_1+Alley_1+       Exterior1st_Stucco + Heating+
                BsmtFinType2_1 + MSZoning_C.all. + Functional_Mod+BsmtFinType2_3  +
                LotShape_IR3+  Condition2+    BsmtFinType2_5+BsmtFinType2_2 +
                Fence_4+       BldgType_TwnhsE   +  Fence_2+       MiscVal+
                BsmtFinType2_6+GarageQual_5+  GarageQual_3+  Exterior1st_BrkComm+
                Exterior1st_ImStucc +MSZoning_FV+   KitchenQual_Fa+GarageType_6+
                GarageQual_4+  GarageType_3+  Functional_Min2  +   SaleType_COD+
                BsmtQual_1+    Functional_Maj2   +  GarageCond_4+  MiscFeature_2+
                FireplaceQu_1+ FireplaceQu_3+ GarageCond_3+  Functional_Min1 +
                Exterior1st_AsbShng +SaleType_CWD+  MSZoning_RH+   GarageType_1+
                SaleType_ConLI+Exterior1st_Stone  + BldgType_Twnhs+Functional_Maj1 +
                Functional_Sev+BldgType_2fmCon   +  Street+  SaleType_ConLD +
                SaleType_Con+  MiscFeature_3+ Exterior1st_AsphShn +GarageCond_5, data=train, cost= 4 ,gamma=0.5, kernel="polynomial")
```

FIGURE V.3: SVM : Machine à vecteurs de support

```
g3<- gbm(SalePrice ~ OverallQual    + Neighborhood+   GrLivArea+ GarageCars+ ExterQual+
          TotalBsmtSF+    X1stFlrSF+      GarageArea+     X2ndFlrSF+      BsmtFinSF1+
          YearBuilt+      FullBath+       LotArea+        GarageYrBlt+    TotRmsAbvGrd+
          YearRemodAdd+   KitchenQual_Ex+ BsmtQual_4+     MasVnrArea+     Fireplaces+
          LotFrontage      +   BsmtUnfSF+      KitchenQual_TA+ OpenPorchSF+    WoodDeckSF+
          OverallCond+    KitchenQual_Gd+ MoSold+         BedroomAbvGr+   Foundation+
          HouseStyle+     MSSubClass+     CentralAir+     SaleCondition+  BsmtFullBath+
          LandContour+    GarageType_5+   LotConfig+      HalfBath+       HeatingQC+
          ScreenPorch+    YrSold      +   MasVnrType+     BsmtExposure_2+ RoofStyle+
          Exterior1st_HdBoard +BsmtExposure_4+ RoofMatl+       MSZoning_RM+    Condition1+
          LandSlope+  SaleType_WD+    Exterior1st_VinylSd +GarageFinish_1+ GarageFinish_2+
          EnclosedPorch+  BsmtQual_3+  Exterior1st_BrkFace+KitchenAbvGr+   MSZoning_RL+
          FireplaceQu_4+ GarageFinish_3+ ExterCond+   SaleType_New+ LotShape_Reg+
          PoolArea+ BsmtFinSF2+ BsmtFinType1_6+ LotShape_IR1+ BsmtFinType1_3+
          BsmtHalfBath+BldgType_1Fam +  Exterior1st_MetalSd +LotShape_IR2+  Exterior1st_WdSdng +
          BsmtFinType1_5+ GarageType_4+BsmtFinType1_2+ PavedDrive+BsmtFinType1_1+
          FireplaceQu_5+  PoolQC_3+  Fence_3+  Exterior1st_CemntBd + LowQualFinSF+
          Functional_Typ+ Exterior1st_Plywood +Alley_2+ BsmtFinType1_4+ BldgType_Duplex
        +Exterior1st_WdShing + Electrical  +
          BsmtCond_4    + X3SsnPorch +     BsmtCond_2+     BsmtFinType2_4  +
          BsmtExposure_1+Alley_1 +       Exterior1st_Stucco + Heating+
          BsmtFinType2_1+MSZoning_C.all.    + Functional_Mod+BsmtFinType2_3  +
          LotShape_IR3+  Condition2+    BsmtFinType2_5+BsmtFinType2_2 +
          Fence_4+       BldgType_TwnhsE   +  Fence_2+       MiscVal+
          BsmtFinType2_6+GarageQual_5+  GarageQual_3+  Exterior1st_BrkComm+
          Exterior1st_ImStucc +MSZoning_FV+   KitchenQual_Fa+GarageType_6+
          GarageQual_4+  GarageType_3+  Functional_Min2  +   SaleType_COD+
          BsmtQual_1+    Functional_Maj2   +  GarageCond_4+  MiscFeature_2+
          FireplaceQu_1+ FireplaceQu_3+ GarageCond_3+  Functional_Min1 +
          Exterior1st_AsbShng +SaleType_CWD+  MSZoning_RH+   GarageType_1+
          SaleType_ConLI+Exterior1st_Stone  + BldgType_Twnhs+Functional_Maj1 +
          Functional_Sev+BldgType_2fmCon   +  Street+  SaleType_ConLD +
          SaleType_Con+  MiscFeature_3+ Exterior1st_AsphShn +GarageCond_5+
          GarageQual_1+  SaleType_Oth+  SaleType_ConLw+MiscFeature_4+
          GarageCond_1+  MiscFeature_1+ Fence_1+       PoolQC_1+
          Utilities_AllPub   +Utilities_NoSeWa  +  BsmtCond_1+    Exterior1st_CBlock +
          PoolQC_2+      Alley_0+       FireplaceQu_2+ GarageCond_2+
          GarageCond_0 +  GarageType_2+  GarageQual_2+  BsmtQual_2 , data = train, distribution="gaussian",interaction.depth=10,shrinkage=0.1)
```

FIGURE V.4: Gradient Boosting