***Présentation du sujet*** :

L’objectif du projet qui nous a été soumis est la réalisation d’une analyse exploratoire sur le jeu de données portant sur les demandes de valeurs foncières. La demande de valeur foncière est un document produit par la direction générale des finances publiques qui permet de retracer l’ensemble des transactions intervenues dans le marché immobilier français sur une période donnée. Le jeu de données, sur lequel nous avons travaillé, se compose des 6 fichiers texte, contenant la totalité des transactions immobilières entre le deuxième semestre de l’année 2016 et le premier semestre de l’année 2021. L’ensemble de ces fichiers ont été directement téléchargé depuis le site « data.gouv.fr ».

Nous avons décidé de nous poser un objectif supplémentaire, consistant dans la construction du jeu de donnée le plus optimal possible pour construire un modèle efficace. Pour y parvenir, nous avons avant tout mené une analyse exploratoire afin d’étudier la data ainsi qu'à identifier d'éventuelles valeurs extrêmes ou incohérences dans le jeu de données. Puis, dans une première phase de « Pre-Processing », nous avons nettoyé les données et traité les valeurs extrêmes et les incohérences de certaines variables, afin de pouvoir convenablement construire les premiers modèles. Finalement, après la construction des modèles, nous sommes replongés de manière plus approfondie dans le traitement des valeurs extrêmes et des incohérences que l’analyse exploratoire nous a révélé, afin d’améliorer la performance de nos modèles.

**I. Récupération des données** :

Nous avons décidé de concaténer verticalement ligne par ligne les datasets sur l’ensemble de la période de 2016 à 2021. Le résultat de cette concaténation est dataframe contenant plus de 16 millions de données. Pour ne pas être confronté à des temps d’exécution du code trop grands et a des problèmes de surcharge du noyau sur Jupyter, nous avons décidé de réduire la taille de notre jeu de données.

Afin de réduire le nombre de données nous procéderons à un premier nettoyage des données, dont les étapes sont les suivantes :

* Nous détectons les colonnes vides ou quasi-vides ou autrement dit les colonnes qui ont un certain pourcentage de données manquantes. On remarque que presque la moitié de colonnes ont un taux de valeurs manquantes > 90 ou voir 100% ;
* Nous supprimons ces colonnes ainsi que la colonne *1er lot* malgré un taux de valeurs manquantes égal à 0.69 car cette colonne est sûrement liée avec la colonne *surface carre 1er lot* pour laquelle on a moins 10% de données. Au final cela revient donc à supprimer toutes les colonnes pour lesquelles il y plus de 30% de valeurs manquantes.

Ces opérations correspondent plutôt à une étape de pré-traitement des données qui devrait surgir après l'analyse des données mais que nous avons choisi de réaliser dès maintenant afin de travailler avec un DataFrame moins volumineux. De plus ces colonnes, étant donné le peu de valeurs saisi, ne sont pas exploitables.

Nous allons également restreindre notre étude à une zone géographique, car nos recherches sur le net indiquent que le prix de l’immobilier est très corrélée à la zone géographique. Cela nous permet à la fois de réduire la taille des données, afin que le temps d’exécution du code soit convenable, et de respecter la réalité du marché de l’immobilier. En particulier nous nous focaliserons dans un premier temps sur la région d’Ile-de-France pour laquelle on a créé un csv comportant uniquement les données de cette région ;

**II. Analyse des données IDF :**

L'objectif de cette partie est de mener une analyse exploratoire de données afin d'identifier d' éventuels problèmes dans nos données (valeurs extrêmes, incohérences, etc...) et les traiter par la suite. Cette étape est indispensable pour la bonne construction d'un modèle de prédiction de la valeur foncière.

Nous commençons par charger le csv des données concernant l’Ile-de-France.

***A. Analyse de la forme :***

* Target : Valeur foncière
* Dimensions : 2260817 lignes et 23 colonnes initialement
* Types de variables : 11 Qualitatives et 12 quantitatives une fois les variables converties
* Analyse de valeurs manquantes : on a déjà préalablement supprimé les colonnes ayant plus de 69% de valeurs manquantes. Heatmap=> Les variables surface terrain et nature culture ont plus de valeurs manquantes mais on décide de les garder: car le pourcentage de val manquantes ne dépasse pas 65%. En particulier pour surface terrain et nature culture les valeurs manquantes sont peut-être dû au fait qu’ils agissent d’appartement ( qui n’ont pas de terrain).

Nous nous apercevons que le type de certaines colonnes n’est pas cohérent. On commence, du coup, par vérifier le type de colonnes et convertir celles qu’on estime pertinentes pour la suite :

* Date Mutation => doit être en datetime ;
* Valeur foncière => doit être un float ;

Globalement nous remarquons que notre Target (la valeur foncière) présente 0.01% de valeurs manquantes. Nous constatons également qu’il y d'autres colonnes ayant des valeurs manquantes. Le traitement de ces colonnes fera l'objet d'une partie suivante.

***B. Analyse du fond  :***

Analyse du **Target**  : ***Valeur foncière*** => on retire les lignes correspondant aux valeurs manquantes de la variable valeur foncière => car il y a que 1% de valeurs manquantes et qu’à des fins de modélisation il est plus pertinent d’avoir un Target sans valeur manquantes.

Commentaire de stats :

Valeurs assez élevées ; minimum à 0.1 euros et maximum à 1 milliard d’euros => des biens vendu à ces valeurs ne sont pas réels (un immeuble même sur Paris ne peut pas valoir 1 milliard). Moyenne à 350 000 euros environ ce qui semble cohérent avec la réalité du marché de l’immobilier. De plus la médiane vaut environ 283000 euros et celle à 75% vaut 587000 euros. Cela implique que 75% des valeurs foncières sont inférieures à 587000. Etant donnée la taille du maximum et la concentration des données cela fait présager à la présence de valeurs extrêmes. Le boxplot ne fait que confirmer notre soupçon. La présence de valeurs extrêmes écrase le graphique en le rendant illisible. Par exemple, la médiane est écrasée, par les valeurs extrêmes supérieures, à 0, alors que ce n’est pas réellement le cas.

**Traitement effectué** **sur le Target:**

1. Nous avons éliminé les valeurs manquantes pour cette colonne
2. Nous avons identifié la présence de valeurs extrêmes. Nous procédons donc à l’élimination des valeurs extrêmes supérieures en utilisant la règle suivante : toute valeur supérieure à la somme entre le quantile à 75% et 1,5 fois la distance interquartile (différence entre quartile à 75% et la médiane) sont supprimés. Certaines de ces valeurs sont certes extrêmes mais pas aberrantes. En effet la vente d’une maison à 1 million d’euro par exemple est courante dans l’immobilier de luxe. Malgré cela nous les avons tout de meme supprimées. Pour les valeurs extrêmes inférieures on décide de retenir cette règle : nous supprimons toutes les valeurs foncières inférieures à 5000 euros car cela ne respecte pas la réalité du marché de l’immobilier.
3. Nous supprimons les maisons dont la valeur foncière est inférieure à 30000 euros car cela n'est pas en ligne avec la réalité des prix du marché de l’immobilier.
4. De plus dans la documentation il est expliqué que si une transaction concerne plusieurs locaux construits sur un terrain qui comporte plusieurs mentions différentes de la colonne *nature culture* alors le dataset comportera plusieurs lignes pour cette même transaction. La chose qui diffère entre ces lignes sera la valeur de nature culture. Donc nous décidons de supprimer ces doublons.
5. Nous nous apercevons également de lignes exactement identiques : nous supprimons donc ces doublons

Une fois ces opérations réalisées nous nous sommes intéressés à la distribution du Target. Selon le graphique, la distribution de notre Target n'est pas centrée sur sa moyenne et la queue distribution gauche de cette distribution est assez importante. Nous pouvons en conclure que la courbe de distribution de la valeur foncière ne présente ni l’allure ni les propriétés de la courbe de la loi normale.

**Analyse des autres variables** :

Certaines variables sont redondantes comme Code Commune et Commune par exemple. Pour ces colonnes nous décidons d’en garder qu’une seule et de supprimer l’autre. En comprenant les sens de variables on s’aperçoit également que certaines colonnes n’ont pas de lien avec la valeur foncière. Ainsi dans le but de vous proposer un modèle de prédiction de la valeur foncière nous supprimons également les variables qui n’ont aucun lien avec notre target. Dans cette partie nous focalisons donc sur l’analyse de variables restantes.

Pour analyser les différentes variables, nous avons affiché un certain nombre de graphiques sur le Jupyter. Nous avons décidé de en pas les reporter ici pour ne pas rendre notre manuscrit trop lourd.

*Nature Mutation* :

* variable qualitative ;
* valeurs possibles = Adjudication, Echange, Expropriation, Vente, Vente en l’état futur d'achèvement, vente terrain à bâtir ;
* Distribution : grande majorité de vente suivie par des ventes en état de futur achèvement
* Hypothèse pour le modèle : les valeurs foncières obtenues par adjudication sont généralement plus faibles que les prix obtenues par vente simple du bien immobilier

**Remarque**: certaines ventes en état de futur achèvement comportent beaucoup des valeurs manquantes. A supprimer par la suite.

*Type Voie* :

* variable qualitative ;
* plusieurs valeurs possibles
* Distribution : grande majorité de Rue (RUE) suivie par avenues(AV) et boulevard(BD)
* Hypothèse pour le modèle : un bien immobilier localisé sur une avenue ou boulevard à une valeur foncière plus élevé

Aucune incohérence à remarquer pour cette variable

*Code Commune* :

* Variable quantitative discrète ;
* valeurs possibles = de 1 à 692 ;

Aucune incohérence à remarquer pour cette variable

*Code département* :

* Variable quantitative discrète ;
* valeurs possibles = uniquement les département de l’ile de France ;
* Distribution : répartition assez équitable entre les différents départements, légère surreprésentation du département 77 ;
* Hypothèse pour le modèle : les valeurs foncières pourraient dépendre également du département => par exemple 92 peut être plus chère car plus proche de Paris

Aucune incohérence à remarquer pour cette variable

*Code type local* :

* Variable qualitative ;
* valeurs possibles = Maison (=1), Appartement (=2), Dépendance(=3), Local industriel ou commercial et assimilé(= 4) ;
* Distribution : presque 45% d’appartement, suivi par Dépendance environ 26% , Maison un peu plus de 20% et enfin les locaux industriel et assimilés;
* Hypothèse pour le modèle : les locaux industrielles sont susceptibles de coûter plus chères ; maison vaut plus qu’un appartement ; Donc la valeur foncière dépend du type de local.

**Remarques**: Nous constatons qu’il y a des lignes pour lesquelles le code type local, la surface terrain bâti, le nb pièces sont manquants. La surface terrain en revanche est correctement saisie. Nous pensons qu’ils s’agissent alors de terrains que ce soit des terrains à bâtir, à cultiver ou simplement des terrains à vendre.

Nous apercevons du même constant pour les ventes en état futur achèvement, pour lesquelles nous avons en plus l’absence de valeur pour la surface terrain. Mais cela semble logique car la transaction n’est pas terminée et il n’y a encore rien de construit. Donc aucun traitement à faire.

**Traitement à réaliser**: remplacer dans la colonne code type local les lignes discutés ci-dessus par la valeur 5 qui symbolise un terrain.

*Nombre pièces principales* :

**Remarques** : d’après la documentation la cuisine, les salles d’eau ne sont pas considérés comme des pièces principales. De plus pour les dépendances il n’y a pas de pièces principales.

* Variable quantitative discrète ;
* valeurs possibles = de 0 à 112 ;
* Distribution : La plupart des valeur sont concentrées entre 0 et 9 pièces principales. Le quartile à 75% vaut 3 ce qui implique que 75 % des biens ont au plus 3 pièces principales ce qui semble être cohérent. Mais quantile à 25% vaut 0 . En effet on remarque à partir du graphique que le nombre de biens avec 0 pièces principales est prépondérant (plus de 30% des données); de plus présence de valeurs extrêmes => 112 pièces principales c pas conforme à la réalité ; on soupçonne donc l’existence d’erreurs liée à cette colonne.
* Hypothèse pour le modèle : plus un bien a de pièces principales plus il est cher

**Incohérences :**

* il y a beaucoup de 0 pièces. Dans le cas des locaux cela est logique, un local industriel n’a pas de pièces principales. En revanche un appartement ou une maison avec 0 pièces n’est pas cohérent. Nous s’apercevons de l’existence de 1300 maisons et apparts avec 0 pièces.

**Traitement à faire :** supprimer ces valeurs ou remplacer par un valeur construit avec des recherches sur le net. Par exemple, souvent les appart avec une surface entre 80 et 120 m carré ont 2 pièces principales sans compter la cuisine et les salles de bain. Nous déciderons pour lesquels des 2 traitements opter en fonction du score des premiers modèles.

* Présence de valeurs extrêmes supérieures. Il y a certain biens immobiliers avec plus de 20 pièces principales ce qui ne reflète pas la réalité.

**Traitement à faire :** supprimer ces valeurs extrêmes en fonction du score du modèle.

* Il y a des ventes de terrains à bâtir pour lesquelles un code type local, un nombre de pièces et surface terrain à bâtir sont indiqués. Or nous nous intéressons aux simples transactions qui dans ce cas est la vente d’un terrain donc ces colonnes pour lesquelles une valeur est indiqué malgré qu'il s'agisse d’un terrain, sont à retraiter.

**Traitement à faire :** traiter ces colonnes en changeant donc leur valeurs de manière à respecter le fait qu’ il s’agit d’un terrain.

*Nb lots* :

**Remarque** : Un lot désigne les autres éléments achetés avec l’appartement. Ex : Cave, parking…

Variable quantitative discrète ;

* valeurs possibles = de 0 à 198 ;
* Distribution : les valeurs sont essentiellement concentré entre 0 et 4 (graphique).  Quantile à 75% vaut 1. La valeur max et le graphique témoignent de l’existence de valeur extrêmes qu’il faudra retraiter ;
* Hypothèse pour le modèle : plus de lots plus la valeur foncière est élevé

**Incohérences :**

Présence de valeurs extrêmes plus nous constatons qu’ il y a des même maisons mais avec des nombre de lots différents

**Traitement à faire :** supprimer les doublons et les valeurs extrêmes

*Nature culture* :

* Variable qualitative;
* valeurs possibles = cf documentation ;

Aucune incohérence à remarquer pour cette variable

*Surface réelle bâti* :

* Variable quantitative ;
* Distribution : valeur assez concentré et présence des valeurs extrêmes. Moyenne à 54 m2, max à 57470 m2 ( pas cohérent) et 75 % des valeurs sont en dessous de 75 m2 ;
* Hypothèse pour le modèle : plus la surface bâti est vaste plus la valeur foncière est élevé.

**Incohérences** : l’analyse des valeurs extrêmes révèle que la majorité des biens ayant une surface réelle bâti supérieure à 500 m2 correspondent à des locaux industriels, commerciens ou assimilés. Parmi les biens à plus de 500 m2 on remarque aussi l’existence de quelques maisons ce qui n’est pas normal étant donné le prix et la surface trop grande pour une maison. De plus il y une maison avec un surface réel bâti et une surface terrain non renseigné.

**Traitement à faire :** supprimer les transactions qui concernent des maison avec une surface bâti à plus de 500 m2, et supprimer la ligne correspondant à la maison sans surface réelle bâti et surface terrain.

*Surface terrain* :

* Variable quantitative ;
* Distribution : valeur assez concentré et présence des valeurs extrêmes. Moyenne à 1072 m2, max à 611662 m2 ( pas cohérent) et 75 % des valeurs sont en dessous de 577 m2 ;
* Hypothèse pour le modèle : plus la surface terrain est vaste plus la valeur foncière est élevé.

**Incohérences**: nous avons des valeurs extrêmes (1183 lignes dont la surface terrain est supérieur à 50000 m2). Nous remarquons que pour les valeurs foncières ayant des surfaces terrain supérieures à 50000, le code type local est toujours non renseigné. Il s'agit très probablement de terrains à vendre.

Il y a aussi des maisons pour lesquelles la surface terrain n’est pas renseignée et des valeurs extrêmes inférieures.

**Traitement à faire :**

1. Pourles valeurs foncières ayant des surfaces terrain supérieures à 50000 remplacer le code type local manquant par la valeur 5, qui indique qu’il s’agit d’un terrain à vendre.
2. Retraiter la colonne surface terrain pour les maisons pour laquelle il n’y a pas de valeur de surface terrain et remplacer par exemple par la moyenne de la surface par département.
3. Enlever les valeurs extrêmes.

***C. Visualisation de relations Target / variables :***

**Remarque :** nous nous focalisons uniquement sur les variables pour lesquelles nous avons conjecturé à un lien avec la valeur foncière. Pour analyser une éventuelle relation entre les différentes variables et le Target nous avons opté pour l’affichage de la distribution de la valeur foncière en fonction des valeurs uniques des variables lorsque celle-ci sont des variables qualitatives ou quantitatives discrètes. Dans le cas des variables quantitatives continues, telles que surface terrain par exemple, nous avons découpé la colonne en 4 sous-colonnes en fonction des quartiles de la variable. Nous avons décidé d’opter pour ce type de représentation plutôt qu’un nuage de points (qui serait plus approprié pour étudier la relation entre 2 variables) car la présence de valeurs extrêmes rendait ce type de représentation graphique imprésentable.

Pour analyser les différentes variables, nous avons affiché ces graphiques de distribution sur le Jupyter. Nous avons décidé de en pas les reporter ici pour ne pas rendre notre manuscrit trop lourd.

*Target / Type Local*

La courbe de la distribution de la valeur foncière des maisons est plus à droite de celle des appartements et des dépendances. On peut en déduire donc que les maisons coûtent généralement plus que les appartements et les dépendances. Nous constatons également que les fourchettes de prix pour les appartements et les dépendances se rejoignent car les courbes de distribution sont superposées. Les locaux industrielles sont certes minoritaires par rapport aux appartements, dépendances et maisons mais on remarque que les valeurs foncières plus élevées sont associées souvent aux locaux industrielles et assimilés.

**Conclusion :** notre hypothèse initiale semble être confirmée. Variable à tenir en compte pour le modèle.

*Target / Date mutation*

L’ensemble des courbes de distribution sont superposées. La date de mutation n’a donc pas d’influence significative sur la valeur foncière.

**Conclusion :** le graphique indique la présence d’aucun lien avec le Target.

*Target / Code département*

Les départements 75,92 sont caractérisés par une courbe de distribution plus décalée à droite synonyme que les prix dans ces départements sont généralement plus élevés. Le département 78 et 94 présentent aussi des valeurs foncières légèrement plus élevées. Le 75 correspond à Paris et les départements 78,92,94 sont plus proches de Paris que le reste. Il semble bien y avoir un lien entre la valeur foncière et le département d’appartenance. En particulier, plus le département est proche du centre, plus la valeur foncière est élevée .

**Conclusion :** lien avec le Target. Notre hypothèse initiale semble être confirmée. Variable à tenir en compte pour le modèle.

*Target / Type de voie*

L’ensemble des courbes de distribution sont superposées. Le type de rue ne semble pas influencer la valeur foncière.

**Conclusion :** le graphique indique la présence d’aucun lien avec le Target.

*Target / Surface réelle bâti*

Les courbes de distribution sont ordonnées de gauche à droite dans l’ordre de quartiles. Cela implique que plus la surface réelle bâti est grande, plus la valeur foncière l’est aussi. On conclut qu’il y a bien un lien entre les deux variables.

**Conclusion :** lien avec le Target. Notre hypothèse initiale semble être confirmée. Variable à tenir en compte pour le modèle.

*Target / Surface terrain*

Les courbes de distribution ne sont pas exactement superposées mais ne sont pas non plus ordonnées en fonction des quartiles comme c’était le cas pour la surface réelle bâtie. Nous remarquons que pour les quartiles à 25, 50 et à 75%, il semble qu’une surface terrain vaste soit synonyme d’un prix élevé mais cela n’est pas aussi accentué que pour la surface réel bâti. De plus, les biens sur les surfaces terrain les plus vastes ont des prix comparables aux autres biens de surface plus petite ( en effet la courbe rouge du quartile 4 est superposé aux restes des courbes).

**Conclusion :** Lien possible mais peut être faible. Variable à tenir en compte mais possiblement à supprimer ou modifier par la suite en fonction du score du modèle.

*Target / Nombre de lots*

L’ensemble des courbes de distribution sont assez superposées. Les valeurs foncières élevées semblent être associées avec un nombre de lots plus élevé, mais cela n’est pas non plus très net. Cette variable ne semble pas influencer significativement la valeur foncière.

**Conclusion :** aucun lien visible avec le Target. Variable à tenir en compte mais possiblement à supprimer ou modifier par la suite en fonction du score du modèle.

*Target / Nature culture*

Les valeurs foncières plus élevés sont le plus souvent associées aux mention suivantes : S (sols), AG(terrains d’agrément), J(jardins). Il semble qu'il y ait donc un lien entre le type de culture et la valeur foncière.

**Conclusion :** lien avec le Target. Variable à tenir en compte pour le modèle.

*Target / Nature mutation*

Les courbes de distribution ne sont pas superposées. On remarque par exemple que l’expropriation et l’échange sont associés aux valeurs foncières les plus faibles et que la vente en état du futur achèvement est associée aux prix plus élevés. Une vente simple semble amener une transaction dont le prix est légèrement plus élevé que celle d’une adjudication comme conjecturé initialement. La nature de mutation influence bien la valeur foncière.

**Conclusion :** lien avec le Target. Variable à tenir en compte pour le modèle.

**Remarque** : l’analyse du corrélogramme des variables révèle cependant que les corrélation avec la valeur foncière sont assez faibles mais cela n'exclut pas l’existence d’une relation. La plus élevée des corrélations concerne la relation valeur foncière / nb de pièces principales et vaut seulement 0,26. Au regard de l’ensemble des corrélation on peut dire que 3 variables ont une corrélation plutôt significative avec valeur foncière. Il s’agit de Code Type Local, Surface et Nombre de pièces. Malgré ces corrélations nous poursuivons dans l’élaboration du modèle de prédiction de la valeur foncière.

**III. Pre-processing**

***A. Mettre les données dans un format propice à la modélisation***

**Remarque** : dans un premier temps nous nous concentrons sur l’élimination des valeurs manquantes et l’encodage de certaines variables, afin de pouvoir convenablement créer, dans la partie suivante, des premiers modèles de prédiction de la valeur foncière. Puis nous reviendrons ultérieurement sur le traitement des données afin d’améliorer le score de performances des modèles que nous avons construits.

**Etapes effectués :**

**Traitement de la colonne type de Voie**

* Nous supprimons la colonne Type de voie car l’analyse précédente n’a pas conclu à l’existence d’un lien avec le Target.

**Traitement des Nan de surface réelle bâti, code type local et nombre de pièces**

Les variables surface réelle bâti, code type local et nombre de pièces principales ont toutes le même taux de valeurs manquantes. Cela est liée essentiellement aux ventes de terrain pour lesquelles les variables énumérées ci-dessus ne sont pas renseignées. Donc dans un premier temps on s’intéresse à réduire les Nan pour ces variables.

1. Nous supprimons les ventes en futur état d’achèvement pour lesquelles le code type local n’est pas renseigné
2. Nous supprimons tous les lignes dont le code type local et la surface terrain n’est pas renseigné

Les valeurs manquantes restantes à ce stade pour ces trois variables correspondent à des terrains.

1. Nous remplaçons le code type local de ces lignes par la valeur 5 qui désormais indique qu’il s’agit d’un terrain ;
2. Nous remplaçons par 0 les valeurs de surface réelle bâti et nombre de pièces pour ces lignes ;

**Traitement des Nan de surface terrain et nature culture**

1. Nous remplaçons les valeurs manquantes de nature culture par la valeur NR
2. Nous mettons à 0 la surface terrain pour l’ensemble des lignes ou cette information est manquante ; mais on pourrait être amené à modifier en fonction du score du modèle => par exemple remplacer par la moyenne de la surface terrain en fonction du code type local d’appartenance

Remarque : après l’ensemble de ces opérations notre dataset comporte 1370775 lignes 11 colonnes et plus aucune valeur manquante.

**Encodage des colonnes Nature mutation, Code commune, Nature culture, Nature mutation, Code département**

Nous avons décidé d’encoder ces colonnes afin d’obtenir une continuité dans les valeurs de ces variables. Pour ce faire nous avons l’encodeur du module « scikit learn ». A chaque catégorie d’une variable est associé un chiffre. Les chiffres commencent à 0.

***B. Création des premiers modèles***

Dans cette partie nous nous intéressons à la création des modèles de prédiction du prix de l’immobilier. Nous avons décidé de nous concentrer sur deux types de modèle : un modèle simple de régression linéaire et un modèle d’arbre de décision. Ce type de modèle permet la prédiction de la valeur cible en construisant des règles de décision à partir des caractéristiques des différents facteurs. En d’autres termes, à partir du jeu de données d'entraînement, le modèle parvient à construire un arbre de décision qui est utilisé pour attribuer une valeur à la variable à expliquer. Un arbre de décision se compose d’un nœud racine ou les données rentrent et des nœuds de décision caractérisé par une série de questions et réponses qui conditionnent le passage au nœud de décision suivant. Ainsi en fonction de la série de questions – réponses, le modèle d’arbre de décision parvient à donner une valeur à la variable cible.

Dans le Jupyter vous retrouverez les lignes de code pour normaliser les données. Néanmoins pour le modèle d’arbre de décision il n’est pas nécessaire de réaliser cette opération et pour le modèle de régression nous ne constatons pas d’effet significatif de cette normalisation.

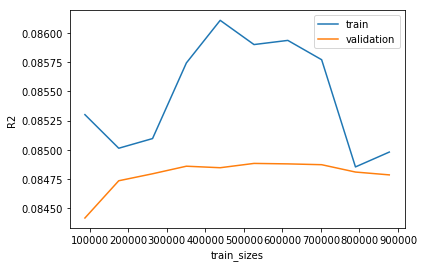
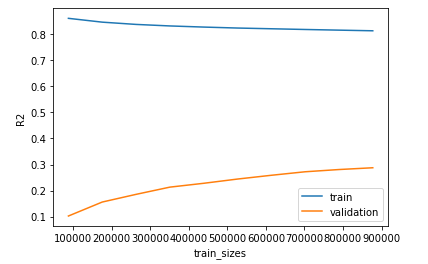
Pour construire ces modèles, nous avons créé la fonction « Evaluation ». Cette fonction permet, à partir d’un modèle de prédiction et d’un data Frame, contenant la variable cible et les facteurs explicatifs, d'entraîner et construire ainsi qu’afficher les statistiques de performance du modèle choisi. Il est également possible d’afficher la courbe d’apprentissage du modèle en mettant à True le paramètre correspondant dans la fonction. Pour ce faire nous avons réalisés les étapes suivantes :

1. Séparation du data frame initial en un data frame d'entraînement et un de validation (ou autrement dit contenant les données sur lesquelles nous allons tester le modèles une fois entraînés).
2. Entraînement des modèles sur les données d'entraînement.
3. Prédiction des valeurs foncières sur les données de validation
4. Construction des statistiques, stockage et affichage de ces dernières
5. Affichage de la courbe d’apprentissage si demandé ;

La courbe d’apprentissage peut se définir comme la représentation graphique de la performance du modèle en fonction de la taille des données d'entraînement. Il s’agit d’un excellent outil de diagnostic de la performance du modèle construit sur les données d'entraînement et de validation. L’indicateur de performance choisi dans la représentation de la courbe d’apprentissage est le R2.

Les statistiques de performance que nous avons affiché sont les suivantes :

* MAE : l'erreur absolue moyenne représente la moyenne de la différence absolue entre les valeurs réelles et prédites de la variable cible. Il mesure la moyenne des résidus sur l’ensemble des données.
* MSE : l'erreur quadratique moyenne représente la moyenne du carré de la différence les valeurs réelles et prédites de la variable cible. Il est un indicateur de la variance des résidus.
* RMSE : il correspond à la racine carré de l’MSE. Il mesure l’écart-type de résidus.
* Coefficient de détermination R2: il indique la part de la variance de la variable cible qui est effectivement expliqué par les facteurs explicatifs dans le modèle choisi.

**Analyse de résultats**

*Courbes d’apprentissage du modèle d’arbre de décision*

*Courbes d’apprentissage du modèle de régression linéaire*

Le modèle de régression linéaire n’aboutit pas à des résultats satisfaisants. Le pouvoir explicatif de ce type de modèle est très fiable : le coefficient de détermination vaut seulement 8,7%. Nous remarquons, à partir du graphique reprenant les courbes d’apprentissage, que le score de performance de ce modèle est très mauvais aussi bien pour les données d'entraînement que les données de validation.

En revanche, le modèle d’arbre de décision affiche une meilleure performance. Le coefficient de détermination vaut environ 31%. Le modèle affiche un score de plus de 80% pour les données d'entraînement, ce qui signifie qu’il parvient très bien à prédire les valeurs foncières sur la base des données sur lesquelles il s’est entraîné. Néanmoins, lorsque le modèle est soumis à un jeu de données inédit ( les données de validation) , il ne parvient pas aussi bien à déterminer la valeur foncière.

Etant données les résultats non satisfaisants du modèle de régression linéaire, nous décidons de le laisser de coté et de se focaliser sur l’amélioration du score du modèle d’arbre de décision.

**Remarque** : le score du modèle d’arbre de décision change lorsqu’on se focalise uniquement sur une catégorie de bien immobilier. Par exemple, en évaluant le modèle à partir d’un jeu de données qui contient uniquement les maisons, nous passons d’un R2 à 31% à un R2 de 21,6%. En revanche, si nous nous concentrons uniquement sur les appartements le R2 augmente, en passant à environ 60%. Cela témoigne de l’importance du type de bien immobilier dans la valeur foncière.

***C. Amélioration du modèle***

**1. Gestion des valeurs incohérentes et des valeurs extrêmes**

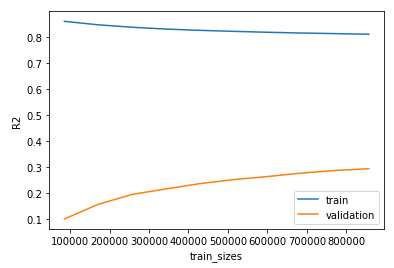
Dans cette partie nous allons traiter les valeurs extrêmes et incohérents qui restent dans notre jeu de données afin d’améliorer la performance de notre modèle

Traitement des valeurs extrêmes :

* Nous avons commencé les valeurs extrêmes supérieures de la variable nb de pièce principales. Pour ce faire nous avons bouclé sur le nombre de pièces ( en commençant par 4 pièces) et affiché à chaque itération le score du modèle d’arbre de décision. Nous remarquons qu’ à partir de 10 pièces principales le coefficient de détermination du modèle reste figé autour de 30-31%. Ainsi nous von supprimer toutes les lignes ayant plus de 10 pièces principales.
* Pour le nombre de lots nous gardons uniquement les valeurs inférieurs à 7. Plus de 7 lots n’est pas possible, étant donné qu’un lot correspond à un garage ou une cave ou etc. … ;
* Pour la surface réelle bâti nous supprimons toutes habitations ayant une surface supérieure à 500 mètres carré. Ces valeurs ne sont pas conformes à la réalité.
* Pour la surface terrain nous supprimons toutes les lignes ayant une surface supérieure à 100000 mètres carré. Ces valeurs ne sont pas conformes à la réalité.

Traitement des incohérences :

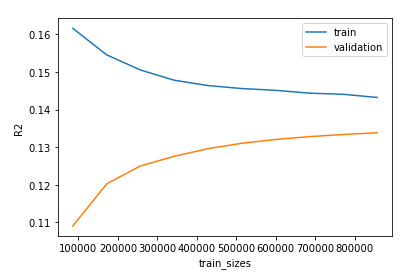
* Nous avons supprimé l’ensemble des habitations ayant un nombre de pièces principales supérieur à 7 mais une surface réelle bâti inférieure à 100 mètres carré. En effet, un nombre si élevé de pièces principales n’est pas compatible avec une si petite surface habitable.
* Nous avons supprimé toutes les maisons (code type local=1) et les appartements (code type local=2) ayant 0 pièces principales car il n’est pas possible d’avoir une habitation sans au moins une pièce principale.
* Nous avons éliminé tous le terrains ( code type local =5) dont un nombre de pièces principales est renseigné, car un terrain ne peut pas avoir de pièces.
* Nous supprimons toutes les maisons ayant une surface terrain nulle

Apres tous ces traitements notre jeu de données comporte encore beaucoup de données : 1338720 lignes et 11 colonnes. Le nombre de lignes que nous avons supprimé est ainsi très petit par rapport à la taille total du jeu de données.

*Courbes d’apprentissage du modèle d’arbre de décision*

La performance du modèle est très peu impacté par l’ensemble des opérations que nous avons réalisé sur le dataframe. Le coefficient de détermination R2 a légèrement baissé en passant de 31% à 30,79% et les courbes d’apprentissage n’ont pas changé. L’absence d’amélioration du modèle peut s’expliquer par le petit nombre de lignes que nous avons supprimé par rapport à la taille total du jeu de données.

**2. Sélection des variables**

L’étude des corrélogrammes des variables, dans la section analyse des relations Target / variables, nous a indiqué que 3 variables ont une corrélation plutôt significative avec valeur foncière. Il s’agit de Code Type Local, Surface et Nombre de pièces. Nous avons donc commencé par évaluer le modèle d’arbre de décision, en considérant uniquement ces trois variables

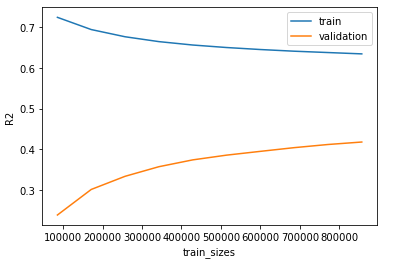
*Courbes d’apprentissage du modèle d’arbre de décision*

Comme le montre le graphique, les résultats sont pas du tout satisfaisants. Le R2 vaut seulement 13,3%. Nous abondons donc l’utilisation des corrélations comme critère de sélection des variables à garder dans le modèle.

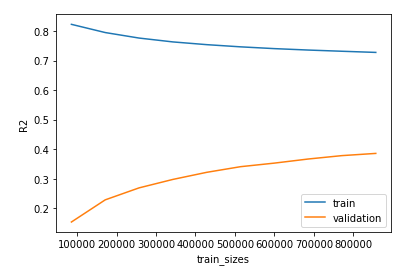
Afin de sélectionner les variables les plus pertinentes pour notre modèle nous avons créé la fonction « SelectVar » . Cette fonction prend en argument le jeu de données, un test de sélection optimale des variables, le nombre de variables que l’on souhaite garder dans le modèle et un paramètre indiquant si nous voulons ou pas que la courbe d’apprentissage soit affiché. A l’aide de ces arguments la fonction « SelectVar » permet de sélectionner le nombre indiqué de variables, les plus pertinentes pour prédire la valeur foncière, et de construire le modèle d’arbre de décision uniquement sur les variables sélectionnées. Elle permet également d’afficher les statistiques de performance et les courbes d’apprentissage, à travers l’appel de la fonction « Evaluation ». La sélection de variables à était obtenue en utilisant les modules « SelectKBest » et le test « f-regression » du package « sickit learn.FeautureSelection ». Le test « f-regression » est un test économétrique utilisant une statistique de Fisher, permettant de conclure à la significativité d’une ou plusieurs variables dans une régression.

Puis nous avons bouclé sur le nombre de variables à sélectionner, en affichant à chaque itération le score de performance du modèle à l’aide de la fonction « Evaluation », appelé dans la fonction « SelectVar ».

**Analyse de résultats**

****Nous remarquons qu’à partir de 6 variables sélectionnées dans le modèle, le coefficient de détermination commence à baisser. Néanmoins nous nous apercevons (voir graphique ci-dessous) que la performance du modèle, en sélectionnant uniquement 6 variables, sur les données d’entrainement est diminué en passant de 80-90% avant sélection des variables, à 70-80%. Pour ce qui concerne les données de test, la performance a bien augmentée , en passant de 31% à 42%.

*Courbes d’apprentissage du modèle d’arbre de décision, en sélectionnant que 6 variables*

Si on considère le modèle avec 8 variables, nous remarquons que le score sur les données d’entrainement est inchangé, mais que la performance sur les données de test a bien augmenté, en passant de 31%, sans sélection des variables, à environ 40%.

*Courbes d’apprentissage du modèle d’arbre de décision, en sélectionnant que 8 variables*

Nous décidons d’opter pour le modèle avec le meilleur score sur les données de test. Nous avons donc décidé de sélectionner le modèle à 6 variables. Pour ce modèle, les variables sélectionnés (classés selon l’ordre d’importance dans la prédiction de la valeur foncière) sont : 'Commune', 'Code département', 'Code type local', 'Surface réelle bâti', 'Nombre pièces principales' et 'Nature culture'. Donc nous pouvons conclure que d’après le modèle d’arbre de décision que nous avons conçu, la zone géographique, ainsi que le type de bien et la surface habitable influencent majoritairement la valeur foncière d’un bien immobilier.

***D. Analyse Covid***

Pour analyser l’effet du covid, nous avons commencé par garder seulement les données entre 2019 et le premier semestre 2021. Nous avons réaffiché le graphique décrivant la relation entre la valeur foncière et la date de mutation seulement sur ces 3 années afin d’identifier un éventuel impact de la pandémie. En analysant le graphique nous ne pouvons pas conclure à un effet significatif du Covid sur la valeur foncière. En effet, les courbes de distribution sur les 3 années sont quasiment superposées.

Nous avons ensuite séparé nos données en deux data set. Un pour la période de 2019 et l’autre pour la période 2020-202 à fin d’étudier les statistiques de ces deux jeux de données. En ce qui concerne le nombre de ventes, on voit que l’année 2020 a connu une baisse par rapport aux autres années. En effet nous étions en général aux alentours des 280 000 transactions contre 243 000 en 2020. En ce qui concerne les autres statistiques descriptives, nous n’observons pas de réels différences. En effet, les moyennes et les quantiles des différentes facteurs et de la valeur foncière sont relativement proches entre les deux périodes.

Nous avons également afficher les histogrammes des deux périodes superposées pour un ensemble de variables. Nous nous sommes surtout intéressés à la variable code département car nous anticipions une baisse des ventes à Paris avec une augmentation des ventes dans les départements plus éloignées. En effet, avec le covid il a souvent été communiqué que de nombreux parisiens quittaient Paris pour s’installer dans le 78 ou le 77. Néanmoins les données que nous avons ne nous permettent d’affirmer cette conjecture. Deux raisons peuvent expliquer ce phénomène. La première est que le Data Set se termine à la moitié de 2021 qui correspond à la fin d’un confinement. On peut anticiper que c’est plus sur les périodes suivantes que les gens ont acheté en dehors de Paris de manière plus importante. Cela est d’autant plus vrai que la recherche d’une maison et la procédure de vente prend du temps. La deuxième raison est que cet exode concerne principalement les ménages à haut revenu avec des prix de vente très élevées. Or dans la gestion des valeurs extrêmes de la valeur foncière nous n’avons gardé que les données qui étaient inférieures à 1,5 IQ. Ainsi nous avons supprimés toutes les ventes qui étaient supérieur au million d’euros. Nous ne remarquons rien pour les autres variables.

Pour finir sur le covid, en dehors du nombre de ventes qui a diminué sur la période 2020-2021, notre jeu de données ne nous a pas permis d’identifier un effet significatif du Covid.

***Conclusion***

L’objectif de ce projet était d’étudier le marché de l’immobilier en France. Devant le nombre conséquent de données nous nous sommes limités à une analyse sur la région Ile-De-France. Nous avons commencé par éliminer les facteurs qui comportaient un nombre de valeurs manquantes trop importants et qui donc n’était pas utiles à l’analyse. Dans un premier temps nous nous sommes intéressé à la variable cible : la valeur foncière. Nous avons supprimé toutes les valeurs incohérentes et extrêmes puis étudié le statistiques de cette variable.

Cette analyse avait pour but de construire le meilleur jeu de données afin de pouvoir prédire les prix de l’immobilier.