Mémoire sur l'Exploration et l'Analyse Prédictive: Prédiction du Prix des Maisons 'SalePrice' dans le cadre d'un Data Challenge Kaggle.

Réalisé par **Galaye Mbengue**
Présenté pour l'obtention du diplôme de Master 1 en Ingénierie Statistique Data Scientist (**ISDS**)

Etablissement : Institut Statistique de Sorbonne Université (**ISUP)**
Année Académique : 2022-2023

Remerciement

J'adresse d'abord mes chaleureux remerciements à Monsieur Olivier Wintenberger mon suiveur de mémoire pour nous avoir exposés à des défis passionnants tels que le Data Challenge Kaggle.Cette opportunité m'a permis de mettre en pratique mes connaissances et de développer des compétences essentielles dans le domaine de l'apprentissage automatique et de l'analyse de données.

Je suis également très reconnaissant envers toute l'équipe pédagogique et administrative de l'ISUP.

Je remercie aussi à tout le corps professoral de m'avoir apporté des connaissances et compétences qui m'ont permis d'approfondir mes études et recherches pour la rédaction de ce mémoire.

Également un grand merci à mes collègues de la promotion ISDS pour les échanges fructueux qui m'ont aidé à rédiger ce mémoire.

SOMMAIRE

INTRODUCTION GENERALE

Les données que nous utiliserons dans notre mémoire proviennent du concours de terrain de jeux (Kaggle) qui met au défi les scientifiques des données de prédire le prix final de chaque maison à Ames, dans l'Iowa. Les données sont fournies sous forme de fichier CSV, comprenant un ensemble d'entraînement (Train.csv) et un ensemble de test (Test.csv). Chaque ligne du fichier d'entraînement représente une maison avec 79 variables explicatives décrivant différents aspects de la propriété.

Certains des variables explicatives disponibles dans les données comprennent des informations sur la taille du terrain, la configuration du lot, la qualité des matériaux et des finitions, la présence de certaines caractéristiques telles que les cheminées et les piscines, le type de chauffage et de climatisation, ainsi que des détails sur le garage et les revêtements extérieurs, entre autres.

La préparation des données sera une étape cruciale dans notre analyse. Nous devrons effectuer des tâches telles que la sélection des valeurs pertinentes, la transformation des données si nécessaire (par exemple, la normalisation ou l'encodage des variables catégorielles), et l'imputation des valeurs manquantes.

Pour la sélection des variables, nous utiliserons des techniques telles que l'analyse exploratoire des données pour identifier les variables ayant le plus d'influence sur le prix des maisons. Nous pourrons également utiliser des méthodes de sélection de variables basées sur l'importance des caractéristiques dans les modèles, telles que l'importance des variables dans la Forêt aléatoire.

Concernant l'imputation des valeurs manquantes, nous pourrons utiliser des approches telles que l'imputation par la moyenne ou la médiane pour les variables numériques, et l'imputation par le mode pour les variables catégorielles.

Pour évaluer les performances de nos modèles, nous utiliserons plusieurs métriques appropriées pour les problèmes de régression. Nous mesurerons l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (MSE), la racine carrée de l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R2). Ces métriques nous permettront d'évaluer la précision de nos prédictions et de comparer les performances des différents modèles que nous avons choisis.

En utilisant ces techniques d'évaluation des modèles, nous pourrons déterminer le modèle qui donne les meilleures prédictions pour le problème spécifique de prédiction des prix des maisons à Ames, dans l'Iowa, et ainsi fournir des informations utiles pour les acheteurs, les vendeurs et le marché immobilier en général.

PARTIE I : PRÉPARATION ET NETTOYAGE DES DONNÉES

1. PRÉPARATION DES DONNÉES

1.1 Origine et composition de l'ensemble de données

L'ensemble de données utilisé dans ce mémoire contient des informations détaillées sur les maisons résidentielles à Ames, dans l'Iowa. Il est composé de 79 variables explicatives décrivant divers aspects des propriétés. La variable cible est le prix final de chaque maison (SalePrice), que nous cherchons à prédire à l'aide de nos modèles d'apprentissage automatique.

1.2 Variables explicatives

Les variables explicatives comprennent des informations telles que:
- Superficie du terrain (LotArea): Cette variable représente la taille en pieds carrés du terrain sur lequel la maison est construite. Une plus grande superficie du terrain peut généralement être associée à une maison plus spacieuse et à des aménagements extérieurs plus vastes.

- Superficie habitable (GrLivArea): Il s'agit de la superficie habitable totale de la maison, mesurées en pieds carrés. Cette variable est essentielle pour évaluer la taille réelle de la maison, ce qui peut avoir un impact significatif sur son prix.

- Nombre de chambres à coucher (BedroomAbvGr): Cette variable indique le nombre total de chambres à coucher situées au-dessus du niveau du sol. Cette variable est souvent un critère important pour les acheteurs potentiels, car il peut refléter la capacité d'accueillir une famille plus grande.

- Nombre de salles de bains (FullBath, HalfBath): Ces variables représentent respectivement le nombre de salles de bains complètes (avec douche ou baignoire) et le nombre de salles de bains partielles (avec seulement un lavabo et des toilettes). Le nombre de salles de bains est un facteur déterminant pour le confort et la commodité des occupants.

- Année de construction (YearBuilt): Cette variable indique l'année de construction de la maison. L'âge de la maison peut influencer sa valeur, car les maisons plus anciennes peuvent nécessiter plus de rénovations ou d'entretien.

- Matériau de la façade extérieure (Exterior1st, Exterior2nd): Ces variables fournissent des informations sur les matériaux utilisés pour la façade extérieure de la maison. Le type de matériau de revêtement peut contribuer à l'esthétique et à la durabilité de la propriété.

- Type de système de chauffage (Heating): Cette variable indique le type de chauffage utilisé dans la maison. Certains systèmes de chauffage peuvent être plus efficaces ou plus coûteux que d'autres, ce qui peut affecter le coût de possession de la maison.

- Nombre de garages (GarageCars): Cette variable représente le nombre de voitures que peut accueillir le garage de la maison. Un garage avec plus de places de stationnement peut être un avantage pour les propriétaires de plusieurs véhicules.
etc.

2. NETTOYAGE DES DONNÉES

2.1 Identification des valeurs manquantes

Lors de l'analyse approfondie de l'ensemble de données, nous avons repéré des valeurs manquantes dans certaines variables. Les valeurs manquantes peuvent être problématiques car elles peuvent entraîner des erreurs lors de l'entraînement des modèles d'apprentissage automatique et potentiellement biaiser les résultats.

Pour identifier les valeurs manquantes dans notre ensemble de données, nous avons utilisé la méthode **'Train.info()'**, qui nous a fourni une vue d'ensemble des informations sur les variables et nous a permis de détecter les valeurs manquantes. La méthode **'Train.info()'** nous a donné une liste des noms de colonnes, le nombre total de valeurs non nulles dans chaque colonne, ainsi que le type de données de chaque colonne.

```
In [9]: train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 80 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   MSSubClass     1460 non-null   int64
 1   MSZoning       1460 non-null   object
 2   LotFrontage    1201 non-null   float64
 3   LotArea        1460 non-null   int64
 4   Street         1460 non-null   object
 5   Alley          91 non-null     object
```

En utilisant cette méthoide, nous avons pu facilement repérer les variables avec des valeurs manquantes. Nous avons observé que les colonnes "Alley", "PoolQC", "Fence" et "MiscFeature" présentaient un nombre élevé de valeurs manquantes, dépassant les 80% du total des données. Etant donné que ces variables ne semblaient pas contenir des informations essentielles pour notre analyse et qu'elles étaient fortement incomplètes, nous avons pris la décision de les supprimer de notre ensemble de données.

```
In [18]: # Convertir la liste en un dataframe pandas
train_NULL = pd.DataFrame(train_NULL, columns=["column", "percentage"])
# Afficher les variables avec plus de 80% de valeurs manquantes
train_NULL[train_NULL["percentage"] > 80]
```

Out[18]:

	column	percentage
5	Alley	93.767123
71	PoolQC	99.520548
72	Fence	80.753425
73	MiscFeature	96.301370

Cette décision de suppression des variables avec un grand nombre de valeurs manquantes a été prise pour éviter l'introduction de bruit dans notre modèle de prédiction des prix des maisons. En éliminant ces variables, nous avons pu préparer notre ensemble de données de manière plus robuste pour la construction de nos modèles d'apprentissage automatique et ainsi améliorer la qualité de nos prédictions.

2.2 Traitement des valeurs manquantes

Dans cette section de notre mémoire, nous détaillons les étapes que nous avons suivies pour préparer les données avant de les utiliser pour entraîner nos modèles d'apprentissage automatique. La préparation des données est une étape cruciale dans tout projet d'apprentissage automatique, car elle garantit que les données sont adaptées à l'entraînement des modèles, assurant ainsi des résultats précis et fiables.

Tout d'abord, nous avons entrepris une analyse approfondie de l'ensemble de données pour identifier les valeurs manquantes dans les variables catégorielles et numériques. Basés sur la description des données, nous avons identifié certaines variables catégorielles, telles que 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1'; 'BsmtFinType2', 'GarageFinish', 'GarageQual' et 'GarageCond', qui contenaient la modalité "NA" indiquant "No Basement" ou "No Garage". Pour traiter ces valeurs manquantes, nous les avons remplacées par la valeur "None", signifiant l'absence de sous-sol ou de garage pour ces maisons.

De même, pour la variable 'MasVnrType', représentant le type de revêtement maçonné de la maison, nous avons remplacé les valeurs manquantes par "None", indiquant l'absence de revêtement maçonné pour ces maisons. Concernant la variable catégorielle 'Electrical', présentant également des valeurs manquantes, nous avons choisi de les remplacer par la valeur la plus fréquente dans cette variable, en utilisant la méthode de la valeur modale. Cette approche nous a permis de combler les valeurs manquantes de manière appropriée, en préservant la distribution existante des modalités dans la variable 'Electrical', tout en évitant d'introduire un biais potentiel dans notre analyse.

In [27]:
```
# Liste des variables à compléter avec la valeur "None"
none_vars = ['FireplaceQu', 'MasVnrType', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2'

# Compléter les valeurs nulles avec ma valeur "None"
train_cat[none_vars] = train_cat[none_vars].fillna('None')

# compléter la valeur nulle de Electrical avec la valeur la plus fréquente
train_cat['Electrical'] = train_cat['Electrical'].fillna(train_cat['Electrical'].mode().iloc[0])
```

2.3 Transformation des données catégorielles

Après avoir effectué l'imputation des valeurs manquantes dans notre ensemble de données, nous nous sommes attaqués à la transformation des données catégoriques en données numériques. En effet, les modèles d'apprentissage automatique requièrent généralement des données numériques en entrée, il était donc essentiel de convertir les variables catégorielles en un format approprié pour l'entraînement de nos modèles. Pour cela, nous avons utilisé des techniques d'encodage appropriées, telles que le label encoding, en fonction du type de variable catégorique et du modèle choisi pour l'entraînement.

2.4 Fusion des ensembles de données

Enfin, après avoir traité les variables catégorielles et numériques séparément, nous avons fusionné les deux ensembles de données pour former l'ensemble d'entraînement final. Cet ensemble de données préparé est maintenant prêt à être utilisé pour entraîner nos modèles d'apprentissage automatique.

La préparation minutieuse des données joue un rôle essentiel dans la qualité et la précision de nos modèles. En garantissant que les données sont complètes, adaptées aux modèles et dépourvues de valeurs manquantes, nous pouvons être confiants dans la robustesse de nos résultats lors de l'entraînement des modèles de prédiction des prix de l'immobilier.

PARTIE II : Revue de la littérature

La prédiction des prix des maisons est un sujet de recherche très étudié dans le domaine de l'apprentissage automatique et de l'économie immobilière. Plusieurs études ont été réalisées pour explorer différentes méthodes d'apprentissage automatique et évaluer leur efficacité dans la prédiction des prix immobiliers.
Dans l'article "Housing Price Prediction Based on Multiple Linear Regression" publié dans Scientific Programming en 2021, les auteurs ont utilisé la régression linéaire multiple pour prédire les prix des maisons en analysant les principaux facteurs affectant les prix. Cependant, ils ont conclu que des méthodes plus avancées pourraient améliorer les prédictions.

Une autre référence "Random Forests for Real Estate Price Prediction" publiée dans le journal of Real Estate Research en 2019, a utilisé la méthode de la forêt aléatoire pour prédire les prix des maisons dans une région métropolitaine. Les résultats ont montré que la forêt aléatoire donne des prédictions plus précises que d'autres méthodes.

Dans l'article "A Gradient Boosting Method for Effective Prediction of Housing Prices in Complex Real Estate Systems" publiée dans la conférence IEEE sur les technologies et applications de l'intelligence artificielle (TAAI) en 2020, les auteurs ont proposé un modèle optimisé basé sur l'amplification du gradient pour améliorer la prédiction des prix des maisons dans les systèmes immobiliers complexes. Les résultats expérimentaux ont montré que cette méthode peut être utilisée efficacement pour prédire les prix des maisons et offrir de meilleures performances par rapport aux autres modèles.

En résumé, notre revue de la littérature met en évidence l'importance de prédire les prix des maisons et l'impact que cela peut avoir sur le marché immobilier. Nous nous appuyons sur des études antérieures pour comparer et évaluer les performances de différentes techniques d'apprentissage automatique dans le contexte de notre mémoire. Ces recherches contribuent à renforcer notre méthodologie et à fournir des perspectives précieuses pour l'avenir de la prédiction des prix immobiliers.

PARTIE III : APPRENTISSAGE AUTOMATIQUE

1. SÉLECTION DES MODÈLES

Nous avons choisi trois modèles d'apprentissage automatique pour prédire les prix de l'immobilier: la Régression Linéaire, la Forêt Aléatoire et l'Amplification du Gradient (Gradient Boosting). Chaque modèle présente des caractéristiques spécifiques qui peuvent influencer ses performances dans la prédiction des prix immobiliers.

1.1 Régression Linéaire

Nous avons choisi ce modèle comme une approche de référence car il est simple et facile à interpréter. Il permet de capturer des relations linéaires entre les variables explicatives et la variable cible (prix de l'immobilier). Cependant, nous sommes conscients que la Régression Linéaire peut ne pas saisir des relations complexes dans les données.

1.2 Forêt Aléatoire

Ce modèle est une extension de l'arbre de décision et est capable de capturer des relations non linéaires dans les données. Il utilise un ensemble d'arbres de décision et agrège leurs prédictions pour fournir une prédiction globale. La Forêt Aléatoire est réputée pour sa robustesse aux valeurs aberrantes et sa capacité à gérer un grand nombre de variables. C'est pourquoi nous l'avons choisi pour sa polyvalence et sa capacité à produire des prédictions précises.

1.3 Amplification du Gradient

L'Amplification du Gradient est une technique d'ensemble similaire à la Forêt Aléatoire, mais elle construit les arbres de décision de manière séquentielle, en accordant plus d'importance aux erreurs résiduelles du modèle précédent. Cette méthode est très puissante pour la prédiction de données complexes et peut surpasser la Forêt Aléatoire en termes de performance prédictive. Nous l'avons choisi pour sa capacité à améliorer la prédiction des modèles de base et à réduire le biais

2. AJUSTEMENT DES HYPERPARAMÈTRES

2.1 Paramètres et Hyperparamètres

Dans le processus d'apprentissage automatique, il est essentiel de comprendre la distinction entre les paramètres et les hyperparamètres, car ils jouent des rôles différents dans la construction et l'optimisation des modèles.

-PARAMÈTRES

Les paramètres sont les valeurs internes du modèle qui sont apprises automatiquement à partir des données d'entraînement pendant le processus d'apprentissage. Ils déterminent les relations entre les variables explicatives et la variable cible dans le modèle. Par exemple, dans la Régression Linéaire, les paramètres seraient les coefficients attribués à chaque variable explicative dans l'équation linéaire. Ces valeurs internes sont ajustées de manière

itérative pendant l'entraînement pour minimiser l’erreur de prédiction sur les données d'entraînement.

-HYPERPARAMÈTRES
Les hyperparamètres, en revanche, sont des valeurs externes au modèle et ne sont pas apprises directement à partir des données. Ils doivent être définis avant le début du processus d’apprentissage et jouent un rôle essentiel dans le comportement global du modèle. Les hyperparamètres contrôlent la manière dont le modèle est construit et influencent ses performances. Par exemple, dans les modèles d’ensemble tels que la Forêt Aléatoire et l’Amplification du Gradient, les hyperparamètres peuvent inclure le nombre d’arbres dans la forêt, la profondeur maximale des arbres, le taux d’apprentissage, etc

2.2 Méthodes d’ajustement des hyperparamètres

L’optimisation des performances du modèle implique de trouver les valeurs optimales des hyperparamètres pour obtenir les meilleurs résultats de prédiction sur les données inconnues. Pour cela, on utilise des techniques telles que la recherche par validation croisée, qui permet d’évaluer différentes combinaisons d'hyper paramètres en utilisant des sous-ensembles de données d'entraînement et de validation.

En résumé, les paramètres sont appris automatiquement par le modèle pendant l'entraînement pour capturer les relations internes, tandis que les hyperparamètres sont définis manuellement avant l'entraînement pour contrôler le comportement global du modèle. Leur ajustement correct est crucial pour obtenir des modèles d’apprentissage automatique performants et précis.

3. DIVISION DES ENSEMBLES DE DONNÉES ET VALIDATION CROISÉE

3.1 Division en ensemble d'Entraînement et de Test

Pour évaluer la performance de nos modèles de manière rigoureuse et éviter le surapprentissage, nous avons divisé l’ensemble de données en deux parties distinctes :

. Ensemble d'entraînement
Cet ensemble représente environ 80% des données totales, il a été utilisé pour ajuster les paramètres internes des modèles lors du processus d’apprentissage. En utilisant les données d'entraînement, les modèles apprennent à capturer les relations entre les variables explicatives et la variable cible, afin de pouvoir faire des prédictions sur de nouvelles données.

. Ensemble de test
Cet ensemble représente environ 20% des données totales et contient des données inconnues pour les modèles, c’est-à-dire des exemples qu’ils n’ont jamais rencontrés pendant l'entraînement. L’ensemble de test a été complètement séparé de l’ensemble d'entraînement pour simuler des données réelles sur lesquelles le modèle sera évalué. Il permet de mesurer la performance réelle du modèle sur des données qu’il n’a pas vu auparavant, ce qui est essentiel pour évaluer son aptitude à généraliser à de nouvelles observations.

In [90]:
```
# Séparer les variables explicatives (X) et la variable cible (y) dans l'ensemble train
X = train.drop(["SalePrice"], axis=1)
y = train["SalePrice"]

# Diviser l'ensemble train en sous-ensemble d'entrainement et de test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

3.2 Utilisation de la validation croisée pour l'évaluation

En plus de la division en ensembles d'entraînement et de test, nous avons utilisé une technique de validation croisée pour renforcer la robustesse de l'évaluation des performances du modèle. Plus précisément, nous avons opté pour la validation croisée par plis (k-fold cross-validation), ou l'ensemble d'entraînement est divisé en k sous-ensemble (ou plis) de taille égale. Le modèle est ensuite entraîné k fois, chaque fois en utilisant k-1 plis comme données d'entraînement et un pli différent comme données de validation. Ce processus est répété k fois, et les performances du modèle sont évaluées en moyenne sur les k essais.

La validation croisée par plis nous permet de mieux utiliser nos données d'entraînement et de tester notre modèle sur plusieurs jeux de données différents, ce qui donne une évaluation plus robuste de la performance. Cela permet également de mieux détecter si notre modèle souffre de surapprentissage ou s'il est capable de généraliser correctement à de nouvelles données.

En conclusion, la division de l'ensemble de données en ensemble d'entraînement et de test, ainsi que l'utilisation de la validation croisée par plis, nous permettent d'obtenir une évaluation fiable des performances de nos modèles, tout en garantissant qu'ils peuvent généraliser efficacement à de nouvelles données.

```
# Validation croisée par plis (k-fold cross-validation)
scores = cross_val_score(model, X_train, y_train, cv=5, scoring='neg_mean_squared_error')
rmse_scores = np.sqrt(-scores)
print('Cross-validated RMSE:', rmse_scores)
```

4. EVALUATION DES PERFORMANCES

4.1 Métriques d'Évaluation

L'évaluation des performances des modèles nécessite l'utilisation des métriques appropriées pour mesurer la précision de leurs prédictions. Dans notre étude, nous utilisons plusieurs métriques couramment utilisées, telles que le MAE (Erreur Absolue Moyenne), le MSE (Erreur Quadratique Moyenne), le RMSE (Racine Carrée de l'Erreur Quadratique Moyenne) et le R2 (Coefficient de Détermination).

. Le MAE quantifie la moyenne des écarts absolus entre les valeurs prédites et réelles, offrant une idée de l'erreur moyenne de prédiction.

. Le MSE mesure la moyenne des carrés des écarts entre les prédictions et les valeurs réelles, donnant plus de poids aux erreurs importantes.

. Le RMSE représente la racine carrée du MSE, fournissant une mesure de l'erreur moyenne sous une forme plus compréhensible.

. Le R2 évalue la proportion de la variance de la variable cible expliquée par le modèle. Une valeur proche de 1 indique une bonne adéquation du modèle.

4.2 Visualisation des Résultats

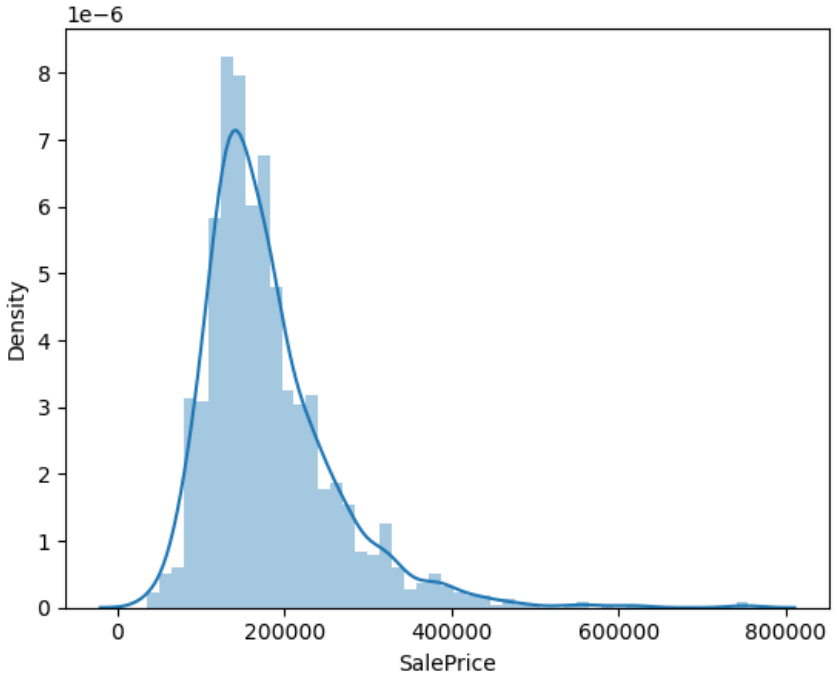
La visualisation joue un rôle essentiel dans la compréhension des performances des modèles. Pour cette raison, nous utilisons des graphiques de dispersion (Scatter Plots) pour illustrer visuellement la concordance entre les valeurs prédites et les valeurs réelles. De plus, les graphiques résiduels (Residual Plots) nous aident à examiner la distribution des résidus, fournissant des informations sur les tendances et les erreurs du modèle.

En combinant les métriques d'évaluation et les visualisations, nous obtenons une image complète de la performance de chaque modèle, ce qui nous permet de prendre des décisions éclairées pour la sélection du modèle final.

5. Analyse de la distribution de la variable cible 'SalePrice'

5.1 Distribution de 'SalePrice'

Pour mieux comprendre la distribution des prix des maisons dans notre ensemble de données, nous avons réalisé une analyse approfondie de la variable cible 'SalePrice'. Initialement, nous avons créé un graphique de distribution pour visualiser la répartition des valeurs de cette variable.

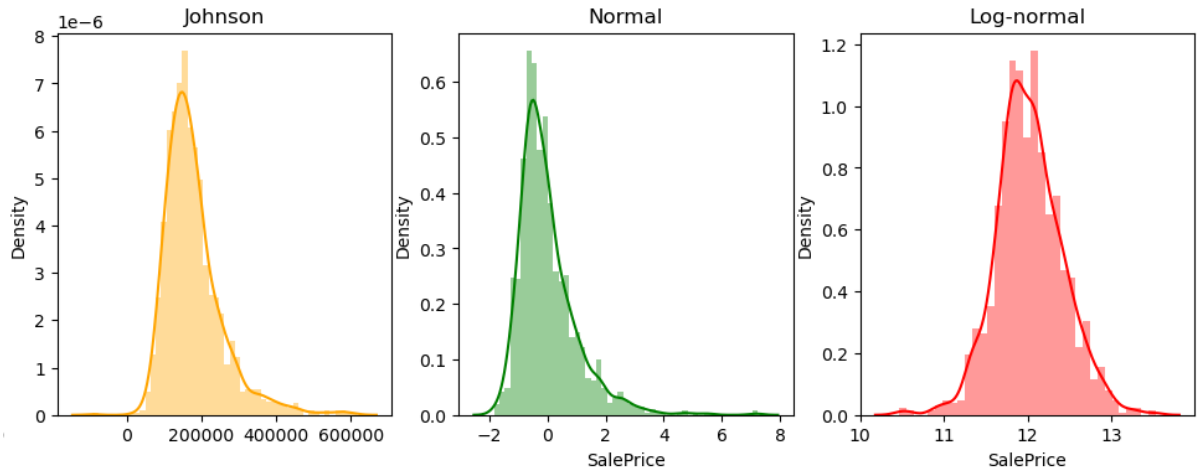


Le graphique initial a montré que la distribution de ‘SalePrice’ ne suit pas une distribution normale et présente une asymétrique à droite (positive skewness). Cela signifie que la majorité des maisons ont des prix de vente relativement bas, avec quelques maisons ayant des prix beaucoup plus élevés, créant ainsi une queue longue vers la droite. Cette caractéristique peut avoir des implications sur la performance de nos modèles d’apprentissage automatique, car de nombreux modèles supposent une distribution normale des données.

5.2 Techniques de Transformation

Pour remédier à cette asymétrie et améliorer la normalité des données, nous avons effectué différentes transformations sur la variable ‘SalePrice’. Nous avons utilisé les transformations basées sur la distribution Johnson, la Standardisation (central et réduction), et la transformation logarithmique.

En traçant les courbes de distribution pour chaque transformation, nous avons pu observer les effets de chaque méthode sur la distribution de ‘SalePrice’.



Les résultats ont montré que la transformation basée sur la distribution Johnson a donné le meilleur résultat pour se rapprocher d'une distribution normale, suivie de près par la transformation logarithmique.

En utilisant le test de Shapiro-Wilk, nous avons également confirmé que toutes les transformations ont significativement amélioré la normalité des données par rapport à la distribution initiale.

```
In [44]: # Effectuer le test de Shapiro-Wilk pour évaluer la normalité des données transformées
         print('Johnson:', stats.shapiro(johnson))
         print('Normal:', stats.shapiro(normal))
         print('Log-normal:', stats.shapiro(log_normal))

         Johnson: ShapiroResult(statistic=0.8976470828056335, pvalue=3.989466454826363e-30)
         Normal: ShapiroResult(statistic=0.869672954082489, pvalue=3.2072044604461286e-33)
         Log-normal: ShapiroResult(statistic=0.9912080764770508, pvalue=1.1514231346154702e-07)
```

En conclusion, la transformation basée sur la distribution Johnson ou la transformation logarithmique peut être utilisée pour préparer la variable cible 'SalePrice' avant d'entraîner nos modèles d'apprentissage automatique. Ces transformations permettront à nos modèles de mieux s'adapter aux données et d'améliorer la précision de nos prédictions de prix de vente des maisons à Ames, Iowa.

PARTIE IV : RESULTATS ET DISCUSSION

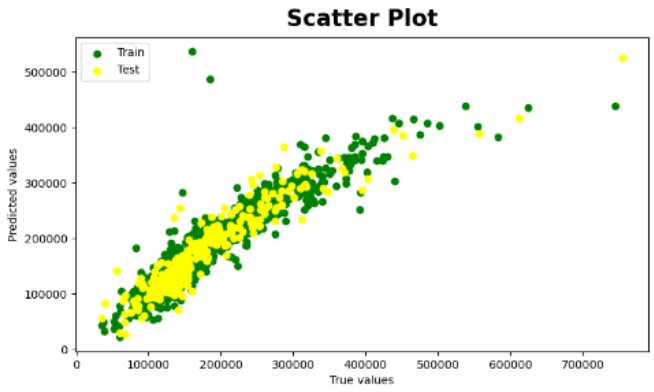
1. Présentation des performances détaillées de chaque modèle

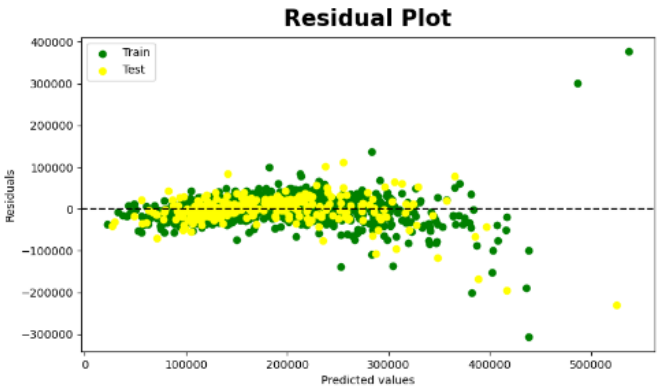
Après avoir entraîné et testé nos modèles d'apprentissage automatique sur l'ensemble de données, nous avons évalué leurs performances en utilisant différentes mesures d'évaluation appropriées, telles que l'Erreur Absolue Moyenne (MAE), l'Erreur Quadratique Moyenne (MSE), la Racine Carrée de l'Erreur Quadratique Moyenne (RMSE) et le coefficient de Détermination (R2). Nous analysons également les avantages et inconvénients respectifs de chaque modèle, ainsi que les raisons pour lesquelles l'un d'entre eux présente de meilleures prédictions dans ce scénario spécifique.

1.1 Régression Linéaire

La Régression Linéaire est une approche simple et facile à interpréter pour prédire le prix des maisons. Nous avons entraîné le modèle sur l'ensemble d'entraînement et évalué ses performances sur l'ensemble de test. Voici les résultats obtenus:

. Mean Absolute Error (MAE): 21692.13630392917
. Mean Squared Error (MSE): 1157933966.5260959
. Root Mean Squared Error (RMSE): 34028.42879896302
. Coefficient de détermination (R2): 0.849037194084248
. Cross-validated RMSE: [38275.70544673 36550.85867595 52625.40501331 27695.18539337
23766.83





-**Scatter Plot**
Le Scatter Plot montre la distribution des valeurs prédites par rapport aux vraies valeurs du prix des maisons. Nous observons une répartition générale de long d'une diagonale, indiquant une correspondance approximative entre les prédictions et les vraies valeurs. Cependant, nous pouvons remarquer des écarts importants par rapport à la ligne diagonale, suggérant que le modèle peut avoir des difficultés à prédire avec précision certaines valeurs.

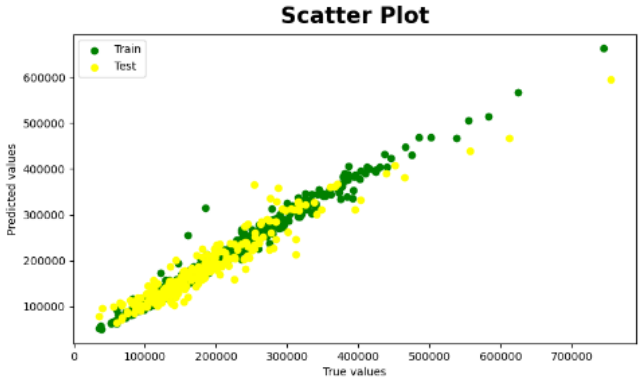
-Residual Plot
Dans le Residual Plot, nous pouvons observer la distribution des résidus (différence entre les valeurs prédites et les vraies valeurs) par rapport aux valeurs prédites. On constate une légère tendance ascendante des résidus indiquant que le modèle peut avoir une tendance à sous-estimer les valeurs les plus élevées. Cela peut être dû à la simplicité du modèle qui ne peut pas capturer des relations complexes entre les variables explicatives et la variable cible.

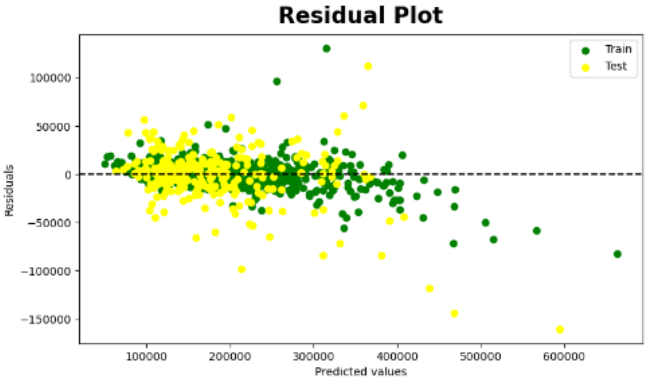
1.2 Forêt Aléatoire

La Forêt Aléatoire est un modèle d'ensemble qui combine plusieurs arbres de décision pour améliorer la précision des prédictions. Voici les résultats obtenus:

. Mean Absolute Error (MAE):17299.080342465753
. Mean Squared Error (MSE): 762024664.2900171
. Root Mean Squared Error (RMSE): 27604.794226547263

. Coefficient de détermination (R2): 0.9006529000584099
. Cross-validated RMSE: [27326.39480468 35747.51441071 37105.49906175 24865.4406738
25208.72012617]





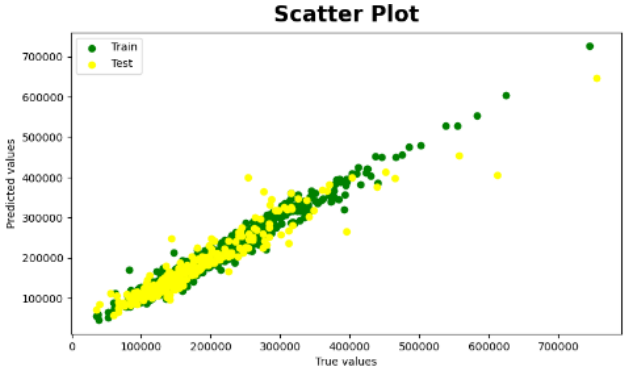
-Scatter Plot
Dans le Scatter Plot, nous pouvons voir une meilleure dispersion des valeurs prédites le long de la ligne diagonale, indiquant une meilleure correspondance avec les vraies valeurs. Cela démontre que la Forêt Aléatoire a réussi à prédire de manière plus précise les prix des maisons.

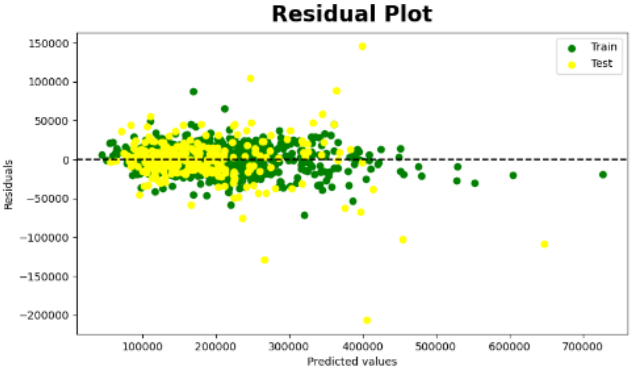
.b Residual Plot
Le Residual Plot montre une légère tendance ascendante des résidus, indiquant que le modèle peut avoir une certaine difficulté à prédire avec précision des valeurs plus élevées.

1.3 Amplification du Gradient

L'Amplification du Gradient est un modèle d'ensemble qui combine plusieurs arbres de décision. Voici les résultats obtenus:

. Mean Absolute Error (MAE):17308.477418144972
. Mean Squared Error (MSE): 803114920.597763
. Root Mean Squared Error (RMSE): 28339.282287979047
. Coefficient de détermination (R2): 0.8952958584935747
. Cross-validated RMSE: [32341.69411388 34347.14289802 32808.33325615 24022.79101659
23738.84570601]





-Scatter Plot
Le Scatter Plot présente des prédictions qui se rapprochent de la ligne diagonale, cela indique une bonne corrélation entre les prédictions et les valeurs réelles.

-**Residual Plot**
Le Residual Plot montre une répartition équilibrée des résidus autour de zéro, sans schéma identifiable. Cela indique que l'Amplification du Gradient a réussi à capturer de manière plus précise les variations entre les prédictions et les vraies valeurs, sans tendance à sous-estimer ou surestimer certaines valeurs.

2. COMPARAISON DES PERFORMANCES DES MODÈLES

La comparaison des performances entre les différents modèles offre un aperçu approfondi de leurs capacités à prédire avec précision les prix des maisons dans notre ensemble de données. Les métriques d'évaluation fournissent des indications quantitatives sur la qualité des prédictions de chaque modèle. Voici un résumé des performances de chaque modèle :

	MAE	MSE	RMSE	R2 \
Régression Linéaire	21692.136304	1.157934e+09	34028.428799	0.849037
Forêt Aléatoire	17299.080342	7.620247e+08	27604.794227	0.900653
Amplification du Gradient	17308.477418	8.031149e+08	28339.282288	0.895296

D'après les métriques évaluées, il est clair que la Forêt Aléatoire a obtenu les meilleures performances globales, affichant un RMSE plus bas et un R2 plus élevé par rapport aux autres modèles. Cela indique sa capacité à capturer les relations complexes entre les variables explicatives et la variable cible, conduisant à des prédictions plus précises. Toutefois, il est important de noter que l'Amplification du Gradient a également montré des résultats compétitifs, avec des performances solides sur ces métriques.

La sélection finale du modèle dépendra des priorités du projet. Si une interprétation simple est cruciale, la Régression Linéaire peut être préférée. Si la performance est primordiale et que l'interprétabilité peut être sacrifiée, la Forêt Aléatoire peut être la meilleure option. Cependant, compte tenu de nos résultats favorables avec l'Amplification du Gradient sur Kaggle, cette option offre un équilibre entre performances compétitives et complexité du modèle. La décision finale devrait tenir compte de ces facteurs ainsi que des objectifs spécifiques du projet.

PARTIE V : CONCLUSION

Au terme de cette étude, nous pouvons affirmer avec certitude l'importance cruciale de la préparation minutieuse des données dans le contexte des modèles d'apprentissage automatique. Cette phase initiale a permis d'appréhender la complexité de l'ensemble de données et de le rendre apte à être utilisé dans divers modèles prédictifs.

En résumé, nos efforts se sont concentrés sur la description de l'ensemble de données, le nettoyage des valeurs manquantes et la transformation des valeurs catégorielles en données numériques. Nous avons également identifié la distribution non normale de la valeur cible 'SalePrice', ce qui nous a conduits à explorer des transformations appropriées pour ajuster cette caractéristique. L'analyse comparative des modèles a ensuite été entreprise, avec un accent particulier sur la Régression Linéaire, la Forêt Aléatoire et l'Amplification du Gradient.

Nos résultats ont révélé des performances différentes pour chaque modèle en termes de mesures d'évaluation telles que le MAE, le MSE, le RMSE et le R2. La Régression Linéaire, bien qu'étant une approche simple, a montré des limitations dans la capture de relations complexes. En revanche, la Forêt Aléatoire et l'Amplification du Gradient ont montré une meilleure capacité à s'adapter aux nuances non linéaires des données, fournissant ainsi des prédictions plus précises pour les maisons avec des prix plus élevés.

Toutefois, il convient de noter que chaque modèle présente ses avantages et ses inconvénients, et le choix du modèle dépendra des priorités spécifiques de l'application. Nos résultats servent de base solide pour prendre des décisions éclairées en matière de sélection de modèle, en fonction des objectifs du projet.

L'Amplification du Gradient se positionne comme le choix privilégié pour anticiper les prix des maisons dans notre jeu de données, en raison de ses performances remarquables sur Kaggle qui surpassent l'ensemble des autres modèles. Sa capacité à capturer de manière précise les relations complexes entre les variables en fait une option de prédiction.

Enfin, des perspectives d'amélioration s'offrent à nous pour d'éventuelles études futures. L'exploration d'autres algorithmes et techniques avancées d'apprentissage automatique pourrait permettre d'obtenir des performances encore meilleures. De plus, une enquête plus approfondie sur les caractéristiques spécifiques du marché immobilier à Ames, Iowa, pourrait conduire à des modèles plus spécifiques et adaptés à ce domaine.

Notre participation à ce data challenge a été une opportunité d'apprentissage enrichissante. Nous avons compris que la maîtrise de la préparation des données est la pierre angulaire du succès de tout projet de modélisation. De plus, l'importance de l'exploration minutieuse de la distribution des variables et de l'adaptation appropriée aux modèles ne peut être sous-estimée. La comparaison des modèles a également révélé que la simplicité ne préjuge pas toujours de la performance, et qu'il est essentiel de choisir des modèles adaptés à la complexité des données.

Pour des travaux futurs, plusieurs pistes d'amélioration et d'extension se dessinent. L'exploration de méthodes d'apprentissage avancées, telles que les réseaux neurones ou les méthodes de séries temporelles, pourrait offrir des perspectives encore plus élevées en matière de performance prédictive. De plus, une compréhension plus fine des spécificités du marché immobilier à Ames, Iowa, pourrait guider la conception de modèles plus adaptés et plus précis.

En conclusion, cette étude nous a permis de plonger dans le monde complexe mais fascinant de la modélisation des prix immobiliers. Elle a renforcé notre conviction en l'importance de la préparation des données et a ouvert la voie à une amélioration continue et à de futurs développements dans le domaine de l'apprentissage automatique appliqué à l'évaluation immobilière.