

Projet d’estimation des prix d'immobilier

Rapport final du traitement des données

GLO-7027 Analyse et traitement de données massives

Ali ASSAFIRI 111 054 128

Rhita OULIZ 111 082 917

06 avril 2019

**Table des matières**

[**1.** **Introduction 10 pages** 3](#_Toc4658580)

[**2.** **Méthodologie** 3](#_Toc4658581)

[**2.1.** **Description de l’algorithme** 3](#_Toc4658582)

[**2.2.** **Description du fonctionnement de la méthode** 3](#_Toc4658583)

[**2.3.** **Décisions de design et d’implémentation** 3](#_Toc4658584)

[**3.** **Expérimentation** 3](#_Toc4658585)

[3.1. **Description des tests** 3](#_Toc4658586)

[**3.2.** **Résultats obtenus et statistiques pertinents** 3](#_Toc4658587)

[**3.3.** **Description des résultats et discutions** 3](#_Toc4658588)

[**4.** **Discutions** 3](#_Toc4658589)

[**4.1.** **Attributs utilisés** 3](#_Toc4658590)

[**4.2.** **Attributs importants** 3](#_Toc4658591)

[**4.3.** **Comparaison avec les attributs importants du rapport 3** 3](#_Toc4658592)

[**4.4.** **Problèmes identifiés et solutions** 3](#_Toc4658593)

[**5.** **Étude comparative des deux méthodes** (2 points) 3](#_Toc4658594)

[**5.1.** **Comparaison des deux solutions réalisées** 3](#_Toc4658595)

[**5.2.** **Condition de performance de chacune des deux méthodes** 3](#_Toc4658596)

[**5.3.** **Différences entre les deux méthodes** 4](#_Toc4658597)

[**6.** **Conclusion** (1 point) 4](#_Toc4658598)

[**6.1.** **Rétrospective sur le projet** 4](#_Toc4658599)

[**6.2.** **Améliorations et recommandation** 4](#_Toc4658600)

Table des tableaux

Table des figures

**Aucune entrée de table d'illustration n'a été trouvée.**

# **Introduction 10 pages**

Le présent rapport décrit notre deuxième traitement de données dans le but d’estimation de la valeur immobilière des maisons de la ville d’Âmes aux États-Unis. En plus de l’analyse et la description de ce traitement et ses résultats, une comparaison avec le traitement effectué lors du rapport 3 sera à la cinquième partie de ce rapport. Pour conclure avec des recommandation améliorations et rétrospectives.

À cette étape de traitement, deux méthodes seront implémentées en plus de la methode Random Forest implémenté précédemment. Il s’agit de la méthode de GBR (Gradient Boosting Regressor Model) et sa version légère en termes de temps de calcul LGBM (Light Gradient Boosting Regressor Model).

Une étude d’attributs les plus important pour chaque méthode sera implémenter également afin de découvrir les combinaisons intéressantes.

# **Méthodologie**

### **Description de l’algorithme**

### **Description du fonctionnement de la méthode**

### **Décisions de design et d’implémentation**

Afin d’améliorer notre apprentissage, une augmentation de données a été effectuer en donnant aux données manquantes des valeurs représentatif.

La figure ci-dessus représente le pourcentage de données manquantes. L’erichicement des données à été selon la strategie suivante :

A titre d’exemple, pour les

# **Expérimentation**

### **Description des tests**

* + 1. **Transformation logarithmique du prix**

Diagramme Quantile-Quantile un outil graphique permettant d'évaluer la pertinence de l'ajustement d'une [distribution](https://fr.wikipedia.org/wiki/Loi_de_probabilit%C3%A9) donnée à un modèle théorique.

L’exploration de la distribution des données via un diagramme quantile-quantile à révéler que les l’alignement doit toujours être sur la première bissectrice

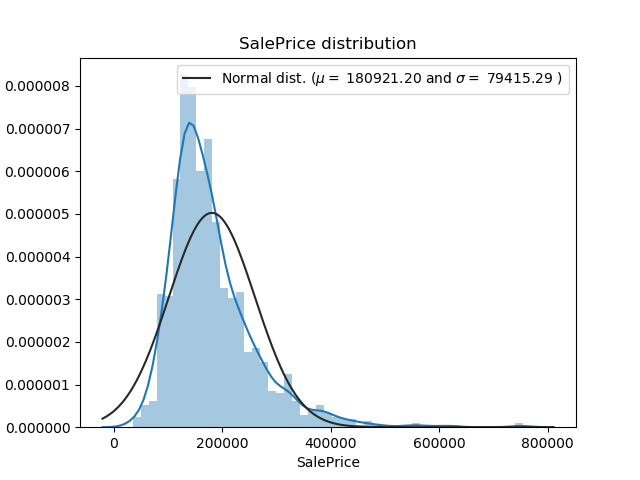
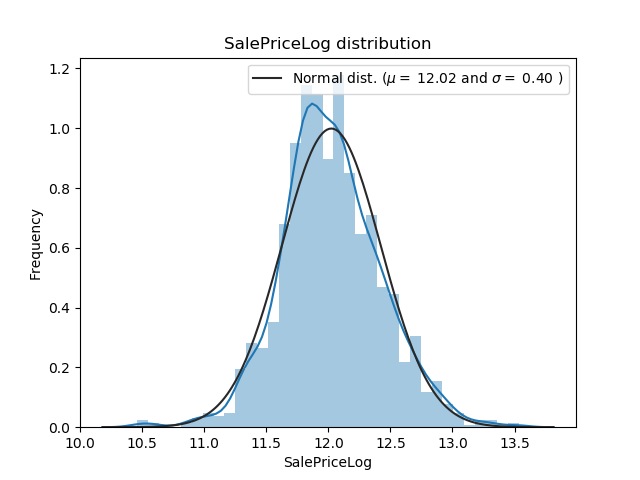
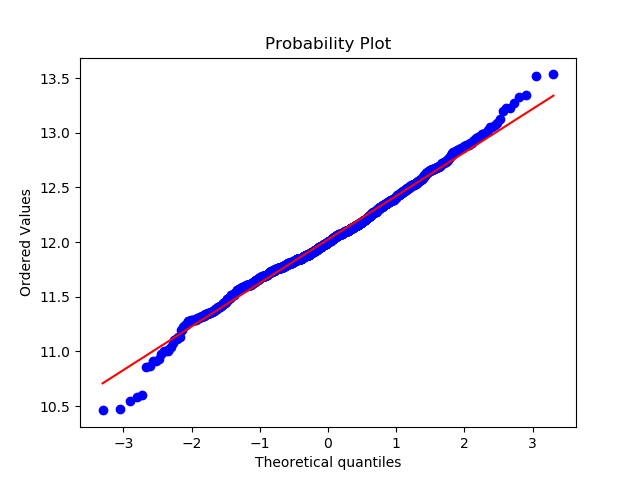
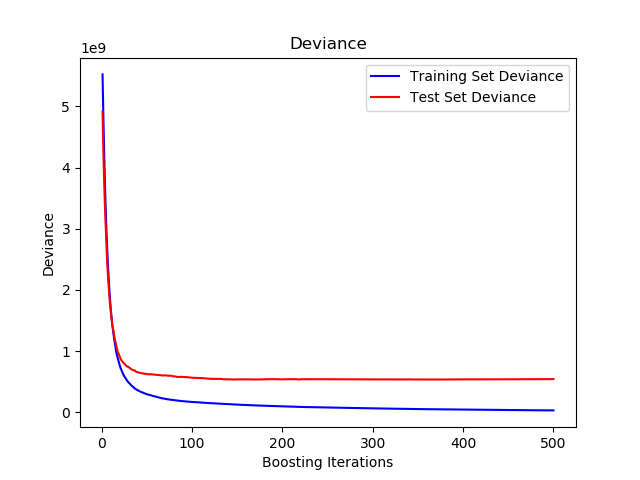
Tables des résultats avec et sans transformation log

La méthode LGBM a bien fonctionné avec les données. Une amélioration des résultats a été identifier. Cependant, les résultats de la méthode GBR appliqué sur les données de prix de vente sont plus précis que le LGBM.

* + 1. **Réduction de dimensionnalité**
    2. **Stratégie de recherche de l’architecture performante du modèle**

Description des paramètres qui influence grandement la performance de la méthode de régression par descente de gradient.

### **Résultats obtenus et statistiques pertinents**

1. 
2. Figure : SalePrice\_distrubution sans transformation logarithmique
3. 
4. Figure : SalePriceLog\_distrubution
5. 
6. Figure 3: SalePriceLog\_QQ-plot
7. 
8. Figure : Gradient Boosting Regressor Model\_Deviance

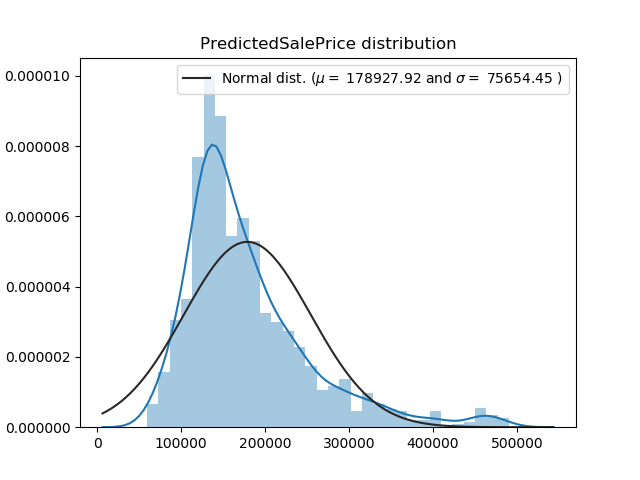


Figure : PredictedSalePrice\_distrubution

Les résultats de 500 itérations d’entrainement du modèle GBR montrent que généralisation d’estimation de prix d’immobilier est acceptable. La médiane de cette distribution de taux de precision est eguel à 90% avec un maximum qui attient 0.94.

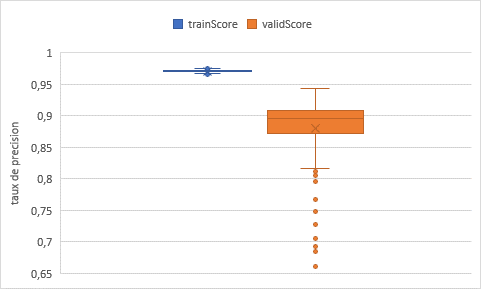


Figure : Taux de précision de GBR

### **Description des résultats et discutions**

Cas de maisons avec une mauvaise estimation

# **Discutions**

### **Attributs utilisés**

### **Attributs importants**

* + 1. Méthode GBR

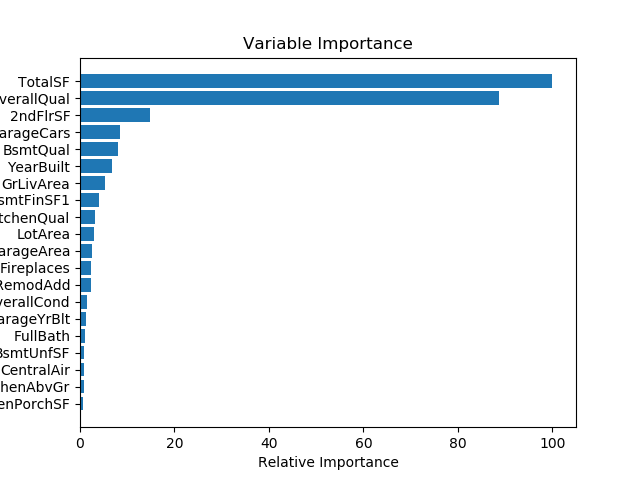


Figure 7: Gradient Boosting Regressor Model\_20importantFeatures

* + 1. Méthode LGBM

Voir Annexe B

Ajout d’un tableau plus clair

### **Comparaison avec les attributs importants du rapport 3**

Pour la méthode Randome forest, lors du rapport 3 on a travaillé avec les attributs …….

Une nouvelle étude sur la méthode nous a données. Voir l’annexe B

### **Problèmes identifiés et solutions**

# **Étude comparative des deux méthodes** (2 points)

### **Comparaison des deux solutions réalisées**

Selon

### **Condition de performance de chacune des deux méthodes**

Dans quelles conditions est-ce qu’une est préférable à l’autre, et pourquoi?

### **Différences entre les deux méthodes**

Discutez autant les différences provenant de la nature des algorithmes, que les différences résultant des leçons prises dans le rapport précédent.

# **Conclusion** (1 point)

### **Rétrospective sur le projet**

En quoi avez-vous eu raison dans votre plan initial, et en quoi avez-vous eu tort?

### **Améliorations et recommandation**

Si le projet était à refaire, que feriez-vous différemment?

**Annexe A : Graphes de visualisation des données**

Soumettez votre résultat final à Kaggle et indiquez votre score et position.

Incluez également la version finale de votre code (lien vers votre dépôt GIT ou fichier zip) avec votre soumission.

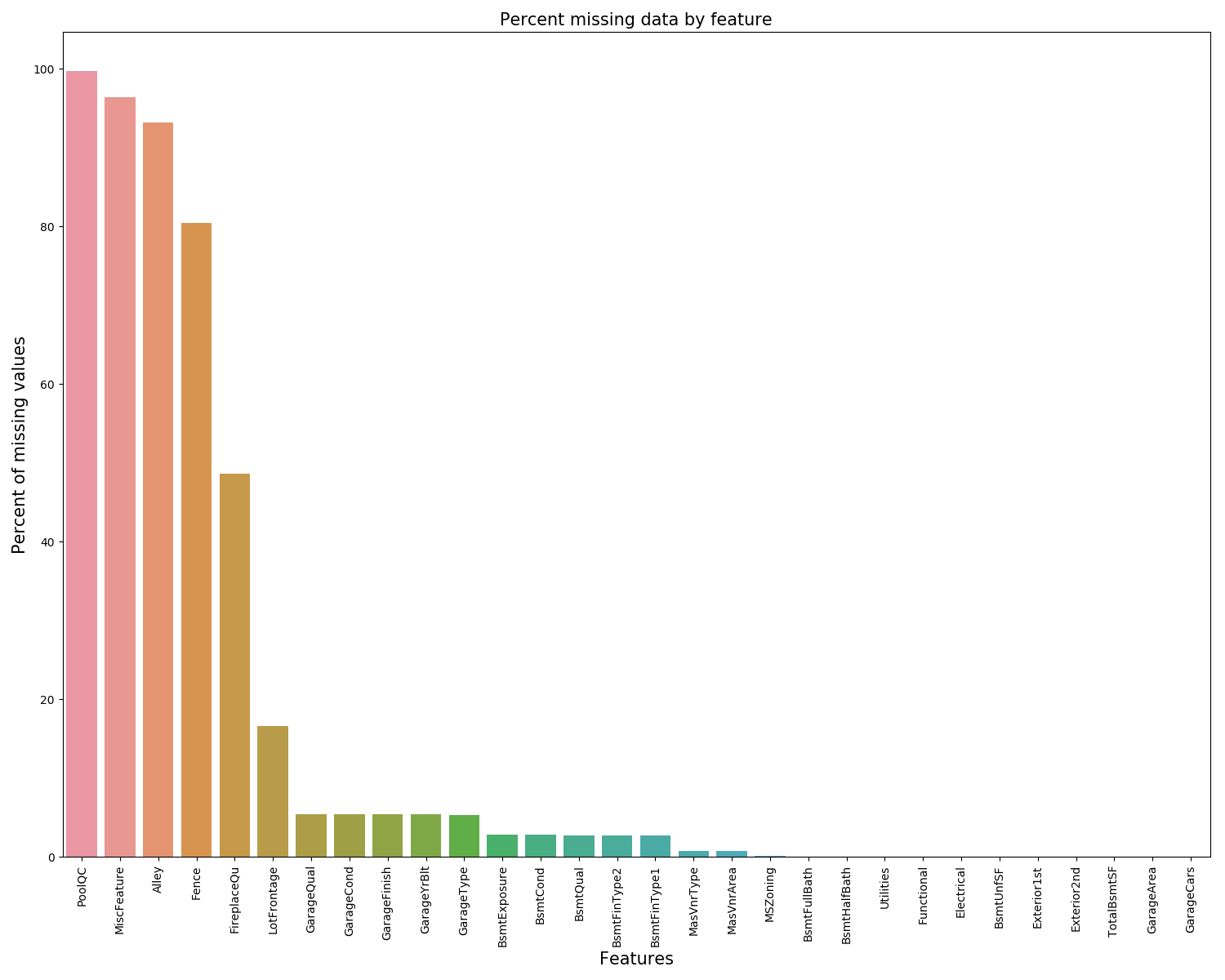


Figure : missingData

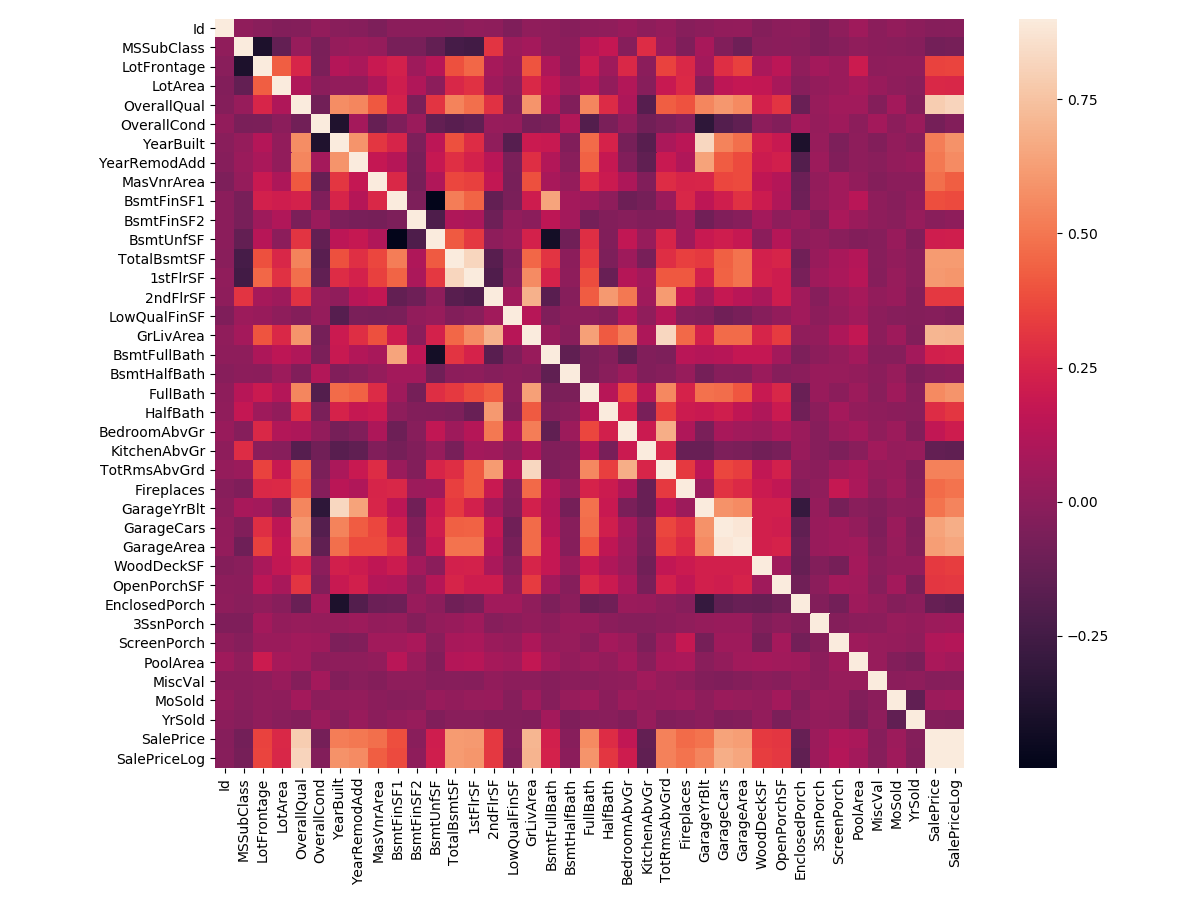
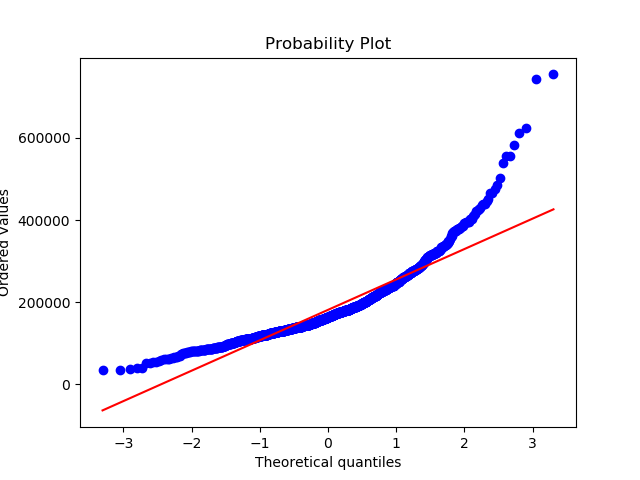


Figure : Corrélation entre les attributs après transformation des données

Figure : SalePrice\_QQ-plot

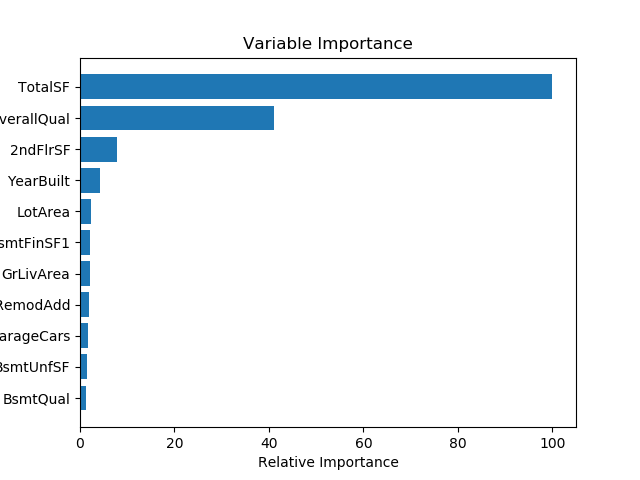


Figure 11: Random Forest Regressor Model\_10importantFeatures

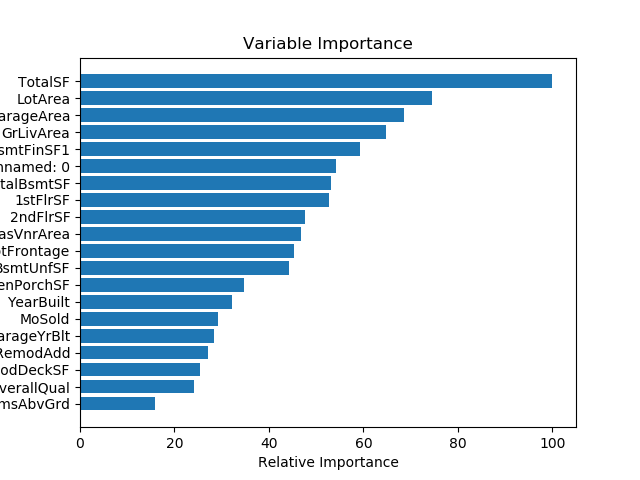


Figure 12: Light Gradient Boosting Regressor Model\_20importantFeatures

**Annexe B : Resultats des tests**

Feature selection from models

\*\*\*\*\*\* Random Forest Regressor Model \*\*\*\*\*\*\*\*\*

Train score: 0.9803

Validation score: 0.9051

All training data score: 0.9658

Random Forest Regressor Model time: 20.20s

New train\_X shape (1168, 19)

Feature selection from model

new features shape (1168, 2)

\*\*\*\*\*\* Random Forest Regressor Model White Importent Fetures \*\*\*\*\*\*\*\*\*

Train score: 0.9806

Validation score: 0.9037

Random Forest Regressor Model White Importent Fetures time: 8.12s

Feature selection from model

new features shape (1168, 17)

\*\*\*\*\*\* Gradient Boosting Regressor Model \*\*\*\*\*\*\*\*\*

Train score: 0.9941

Validation score: 0.9344

All training data score: 0.9826

Gradient Boosting Regressor Model time: 5.22s

New train\_X shape (1168, 17)

Feature selection from model

new features shape (1168, 2)

\*\*\*\*\*\* Gradient Boosting Regressor Model White Importent Fetures \*\*\*\*\*\*\*\*\*

Train score: 0.9913

Validation score: 0.9235

Gradient Boosting Regressor Model White Importent Fetures time: 0.97s

Feature selection from model

new features shape (1168, 39)

\*\*\*\*\*\* Light Gradient Boosting Regressor Model \*\*\*\*\*\*\*\*\*

Train score: 0.9994

Validation score: 0.9260

All training data score: 0.9852

Light Gradient Boosting Regressor Model time: 2.13s

New train\_X shape (1168, 39)

Feature selection from model

new features shape (1168, 17)

\*\*\*\*\*\* Light Gradient Boosting Regressor Model White Importent Fetures \*\*\*\*\*\*\*\*\*

Train score: 0.9995

Validation score: 0.9153

Light Gradient Boosting Regressor Model White Importent Fetures time: 1.43s

**Annexe C**

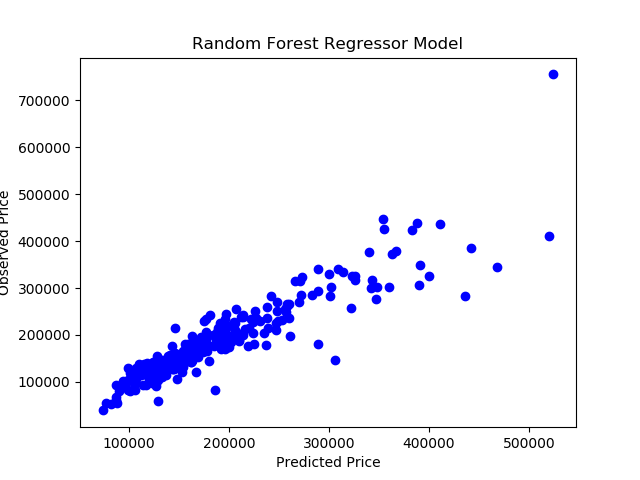
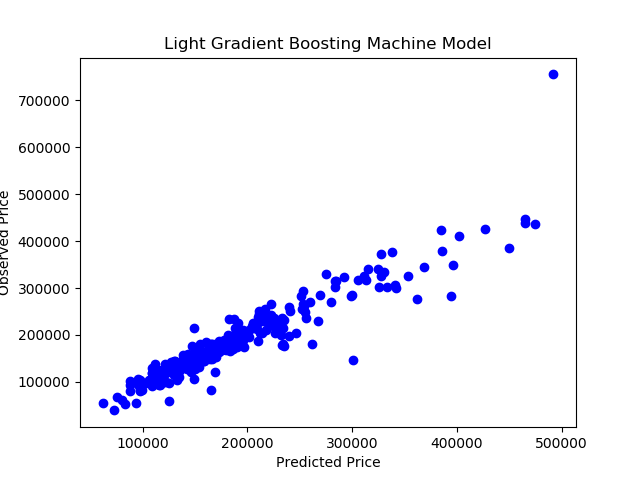


Figure 13: Random Forest Regressor Model\_distrubution

Figure 14: Light Gradient Boosting Machine Model\_distrubution

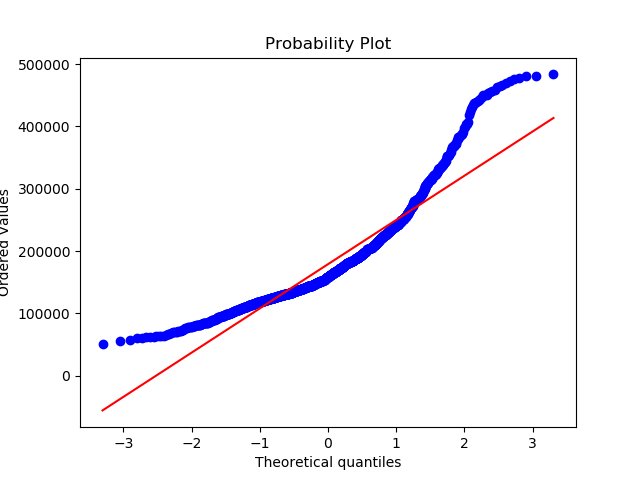


Figure : PredictedSalePrice\_QQ-plot