PRÉDICTION DES PRIX IMMOBILIERS

King County
USA

fitec SEGULA TECHNOLOGIES

Djamel GHARBI, Le 21/12/2018

Sommaire

Équipe

Projet

Architecture Big Data

Intelligence Artificielle

Résultats

Conclusions & Perspectives

L'ÉQUIPE PROJET

Mohammed Ghiles

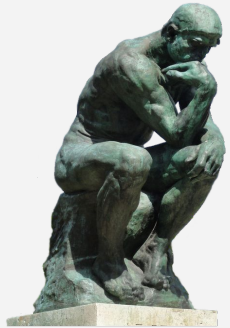
Anastasia Kornikova

Djamel Gharbi

Khadija Ajimi

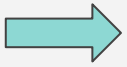
Equipe

Introduction

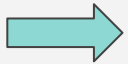


La création de valeur

Artiste



Ingénieur



Data Scientist

Le Projet

1. Contexte
2. Motivations

Contexte et Motivation

- Projet Projet Fil Rouge CBDATA 4
- Dataset : https://www.kaggle.com/harlfoxem/housesalesprediction/data
- L'objectif de cette analyse est de prédire les prix des maisons dans ce comté.
- Le client est une entreprise de construction du comté de King qui cherche à acheter des propriétés et à les revendre.
- Elle utilisera ce modèle pour trouver des maisons moins chères à acheter.

Contexte et Motivation

Plus de 21000 de biens recensés

20 Paramètres

Transactions immobilières sur 2014-2015

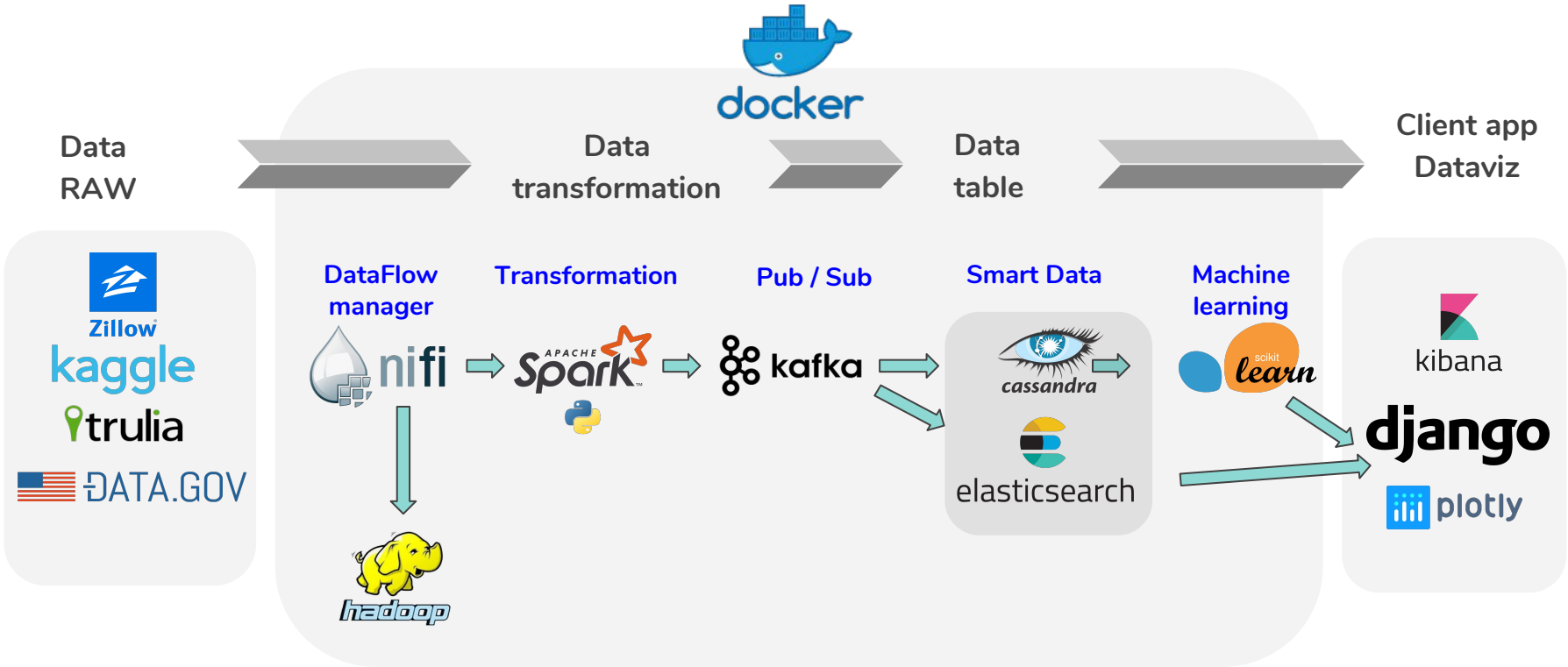
Demo 1

https://shrouded-scrubland-74851.herokuapp.com/homepage/

Architecture Big Data

1. Architecture / choix des outils
2. Scalabilité

1. Architecture / outils



Scalabilité

Docker Pull Command

```
docker pull degharbi/bigdata_datasci
```

Owner

degharbi

Data ingestion, transformation, data table, Machine Learning

Data RAW

Zillow
kaggle
trulia
DATA.GOV
⋮

Cluster 1

docker
Node 1

⋮

docker
Node n

Cluster n

docker
Node 1

⋮

docker
Node n

Client app
Dataviz

kibana
django
plotly

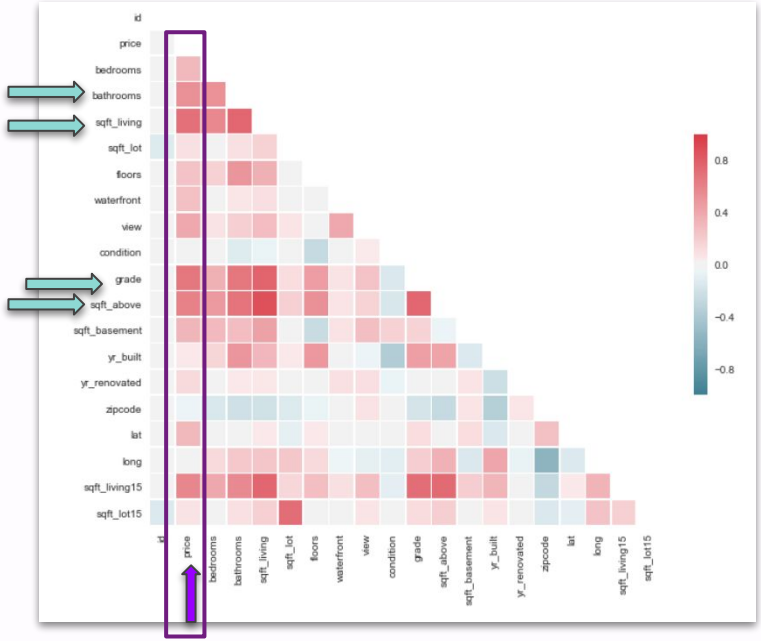
Machine Learning

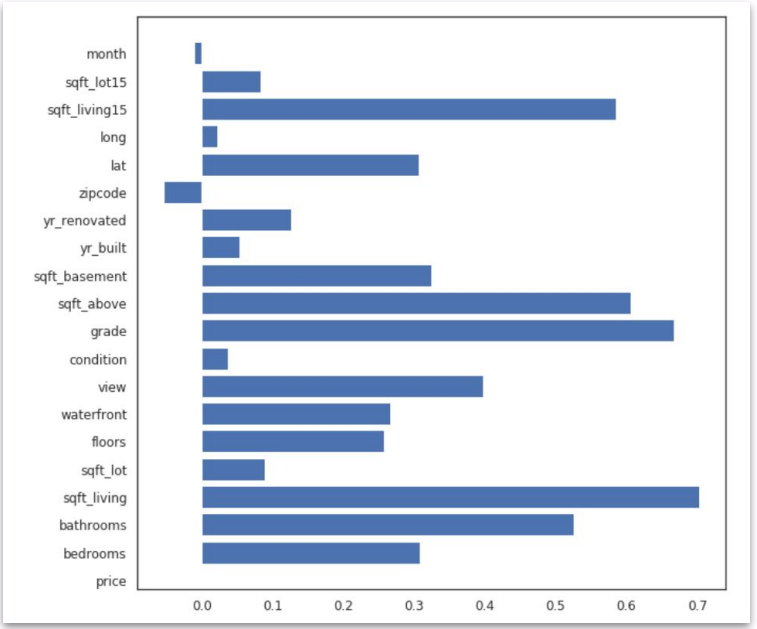
1. Analyse exploratoire
2. Data processing
3. Modèles de machine learning
4. Enrichissement du dataset

Analyse exploratoire : Check list

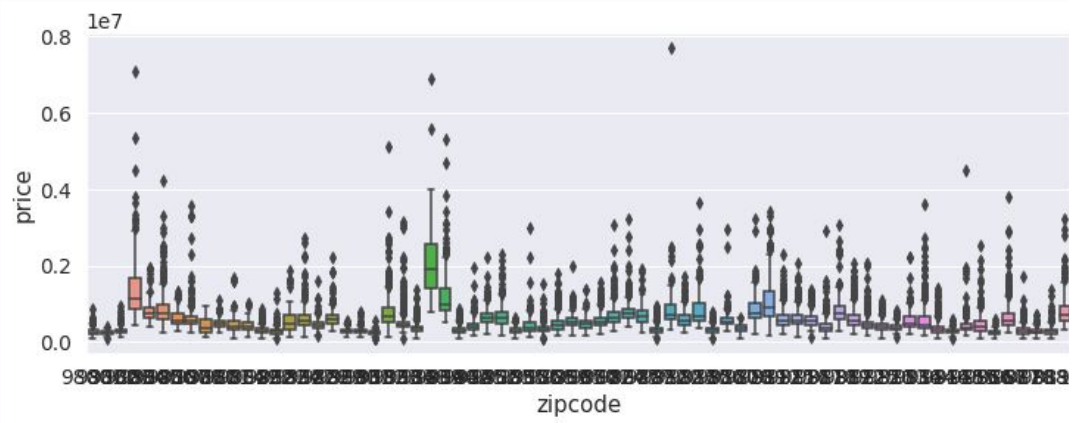
- ➔ Données manquantes: Oui/Non
- ➔ Type des données: numériques, catégorielles
- ➔ Transformer les variables: Oui/Non
- ➔ Recherche des corrélations entre les variables
- ➔ Selection des variables pertinentes

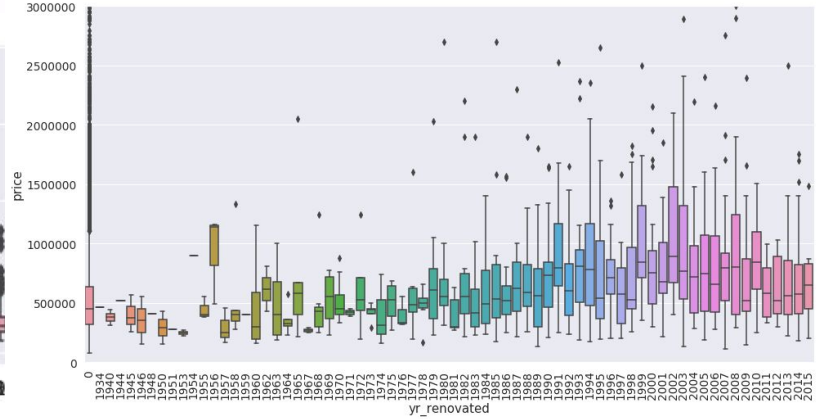
Matrice de corrélation : Identifier les variables significatives





Prix vs zipcode, année de rénovation…

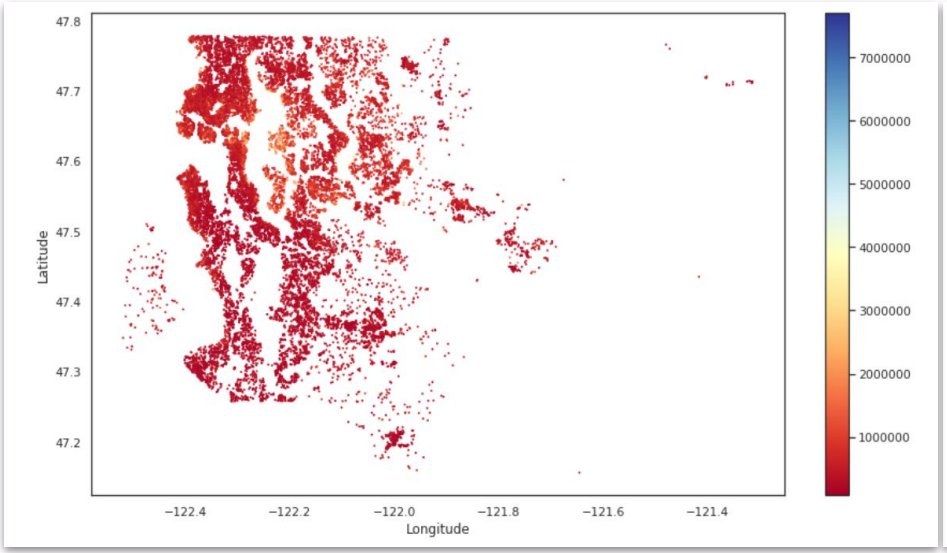




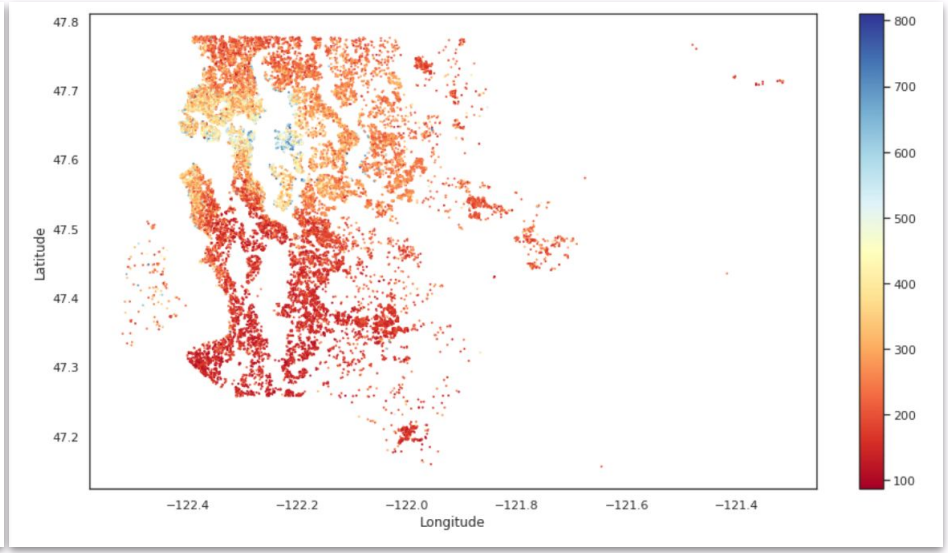
Intelligence Artificielles

Transformation de variables

Prix

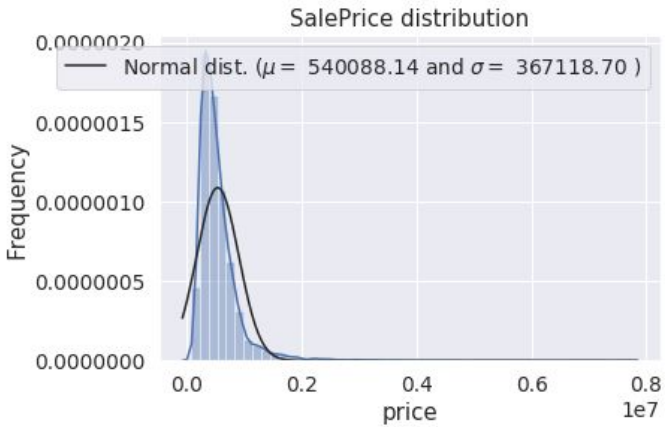


Prix par pied carré

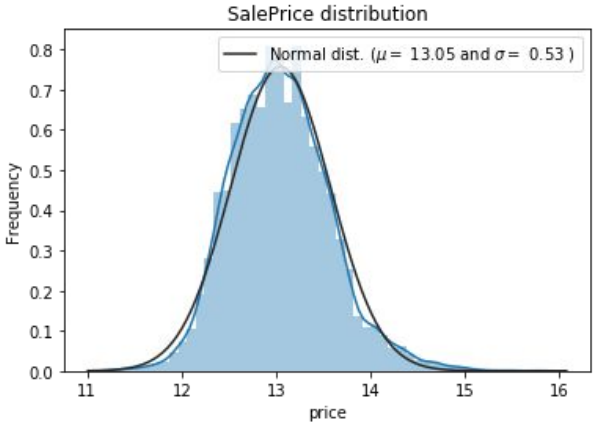


Data processing
Normalisation

mu = 540088.14 and sigma = 367118.70

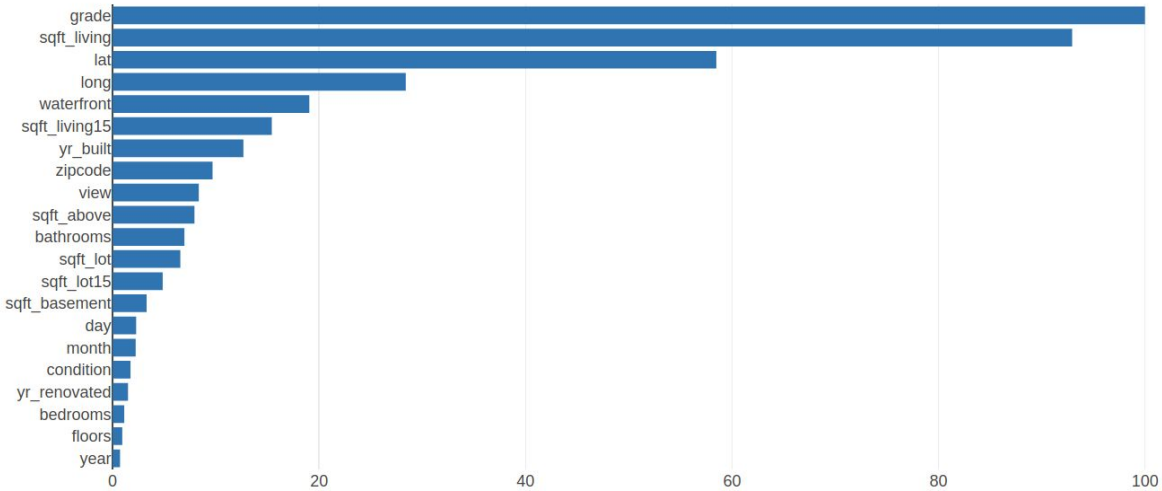


mu = 13.05 and sigma = 0.53



Data processing
Selection des variables

- ➔ Sélection basé sur les arbres de décisions
- ➔ Sélection univarié
- ➔ Variance faible ?



Modèles de Machine Learning

- ➔ Régressions Linéaires
- ➔ Arbre de décision / Random Forest
- ➔ Gradient Boosting

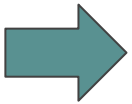
Modèles de Machine Learning

- Données numériques, 21 variables, petit dataset

- Régressions (linéaire, Lasso, Ridge, polynomiale)
- Arbres de décision/Random forest
- Gradient boosting/XGBoost/AdaBoost

Check List pour tous les algorithmes:

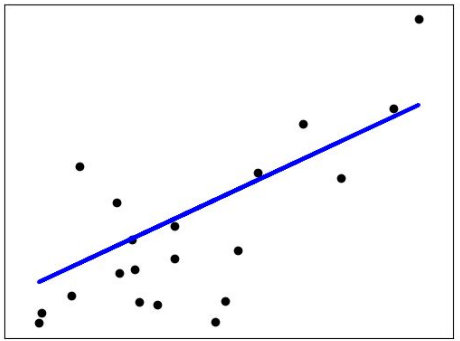
- Grid search
- Normalisation des variables
- Cross validation



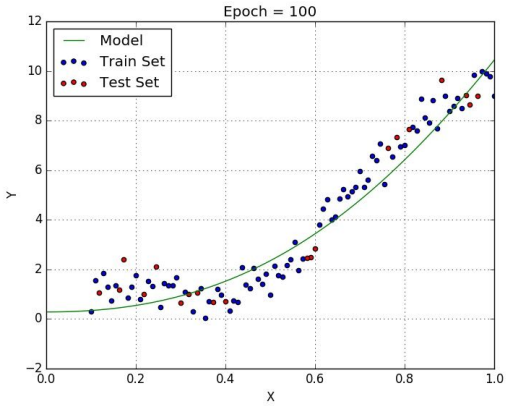
Score

Regression lineaire, polynomiale, Lasso, Ridge

Lineaire



polynomiale



Lasso : parametre de penalisation L1

Ridge : parametre de penalisation L2

Régression Linéaire : Résultats

Algorithmes	R2
Lasso	0.67743
Ridge	0.67748
Polynomiale (degré 2)	0.81804

Arbre de décision

Définition :

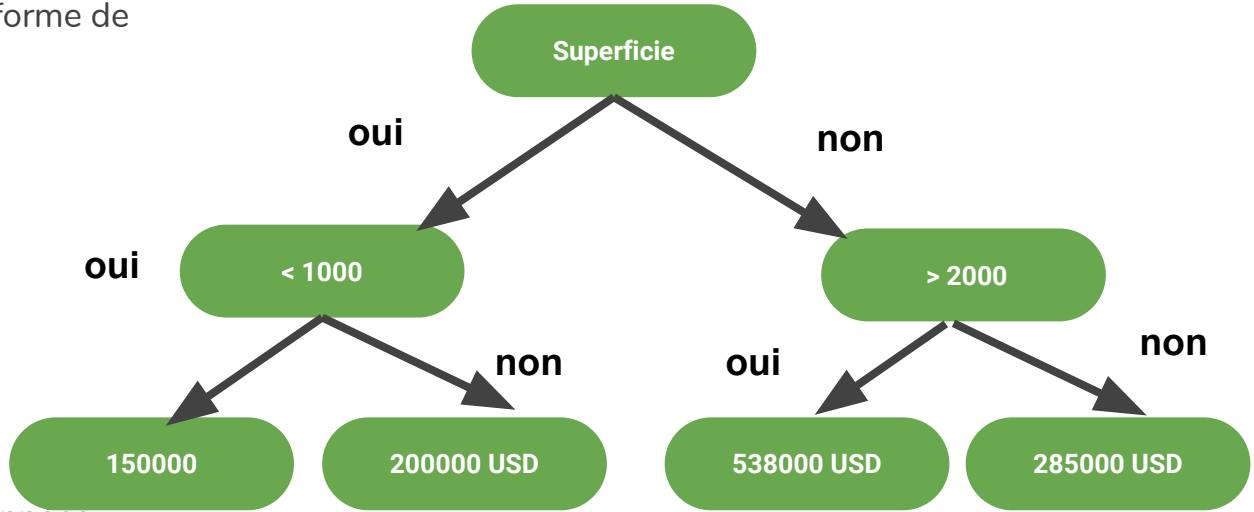
La structure des données sous forme de séquences de décisions (Arbre)

Objectif ?

Prédire un résultat

Avantages :

- Peu de préparation de données
- Logique oui/non
- Performant sur de grands jeux de données



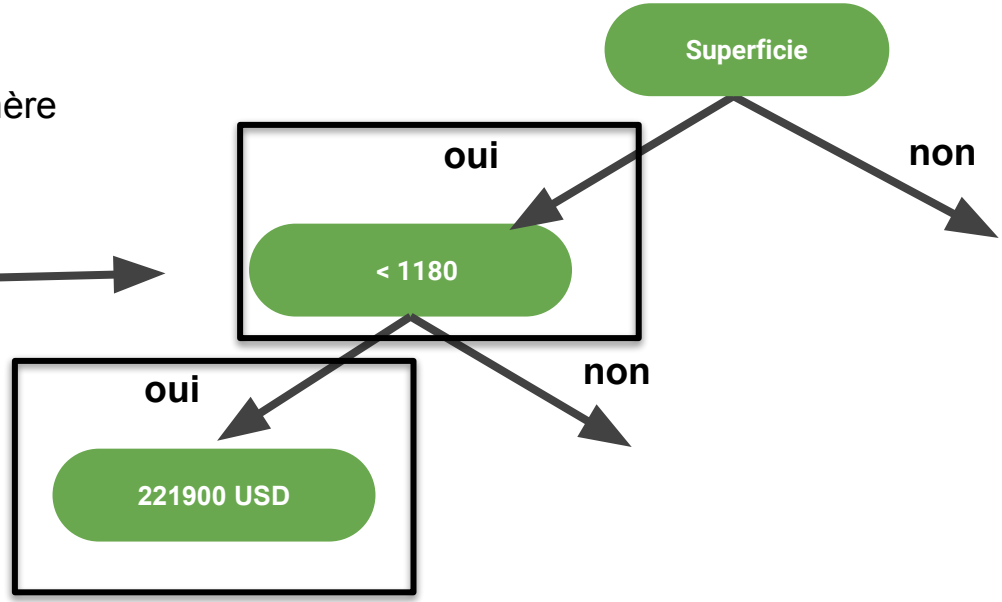
Arbre de décision

Chaque nœud interne correspond à un attribut

Chaque nœud teste l'attribut correspondant et génère Plusieurs branches

Variable numérique : test sur la valeur

Feuilles : prix du bien immobilier



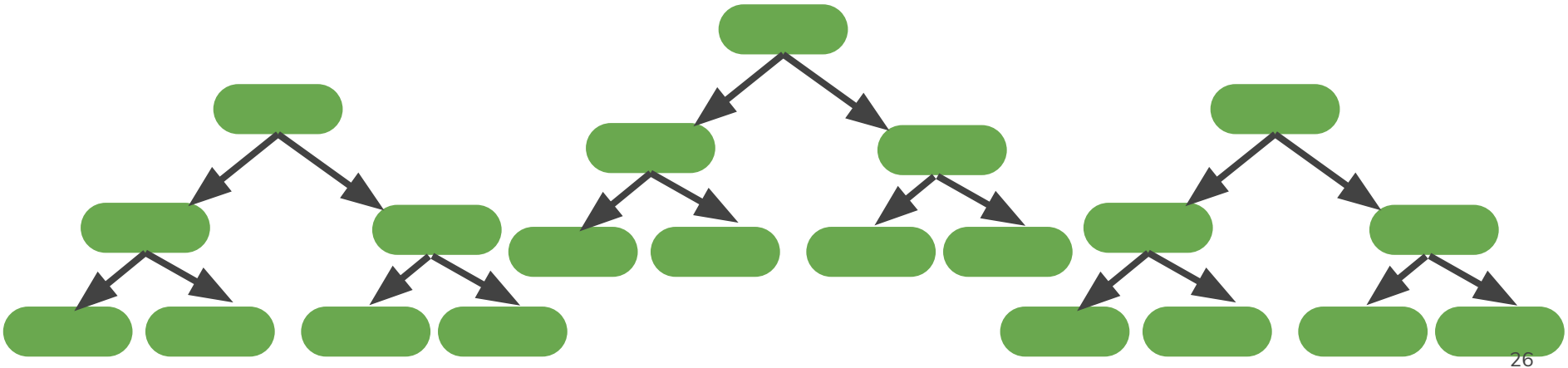
S'arrête quand les éléments d'un nœud ont la même valeur pour la variable cible (homogénéité)

Random Forest

- Algorithme **d'apprentissage supervisé**

- Le même type d'algorithme plusieurs fois pour former un modèle de prédiction plus puissant

- Combinaison de plusieurs arbres de décision : "Forêt aléatoire"



Arbre de décision / Random Forest

Résultats :

Arbre de décision	Random Forest
79.2 %	88.3 %

Avantages du Random Forest :

- Plus précis,
- Très stable : puissance de la “foule” car plusieurs arbres
- Fonctionne bien même avec les données manquantes

Inconvenients du Random Forest :

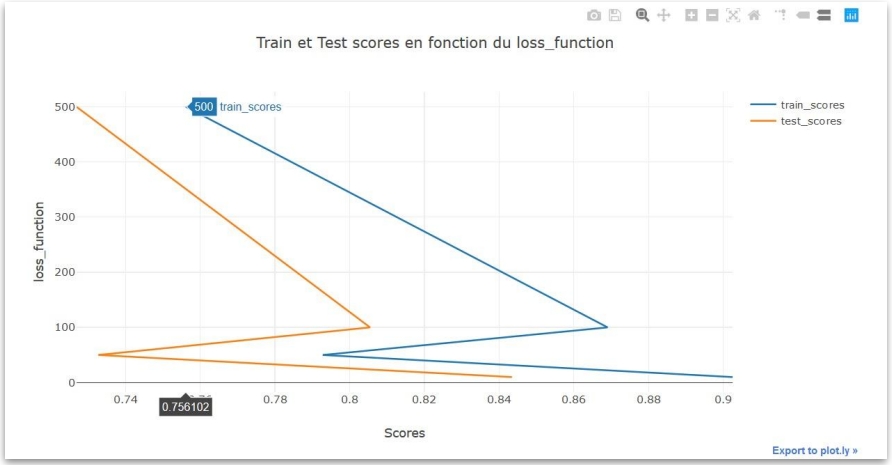
- Ressources de calcul ++++
-

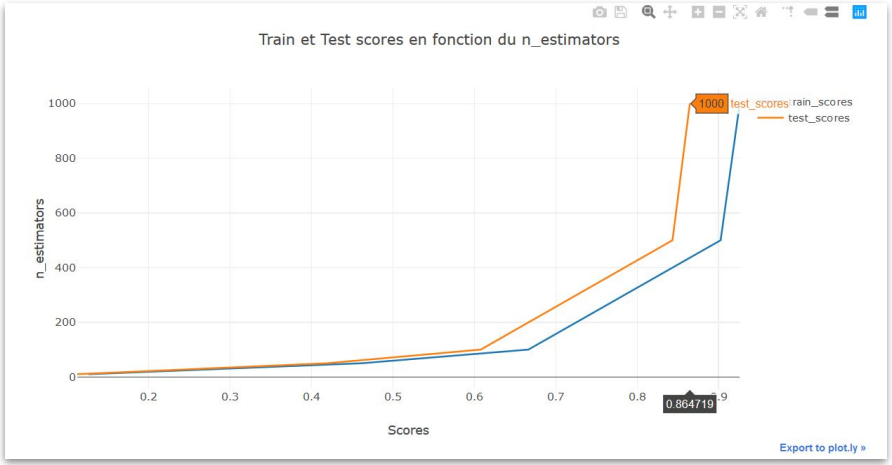
Gradient boosting

+ Très efficace sans beaucoup de préparation
+ Données numérique et catégorielles
+ Robuste aux outliers

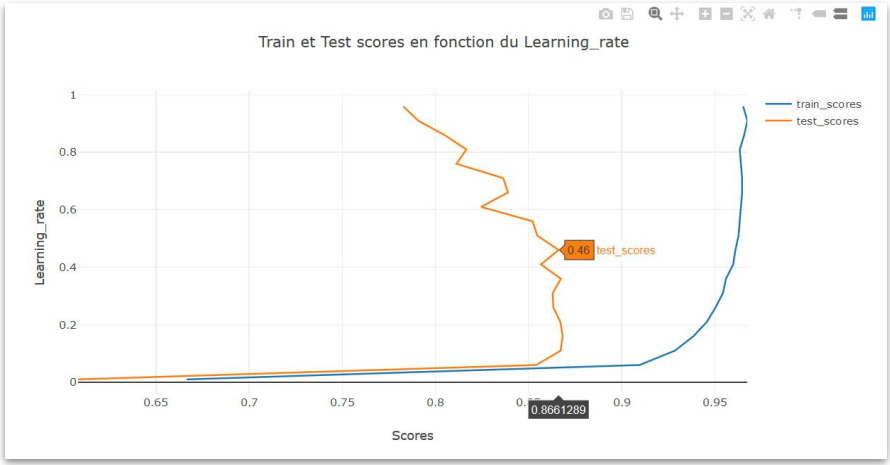
- Difficile à paralléliser suite aux process itératif

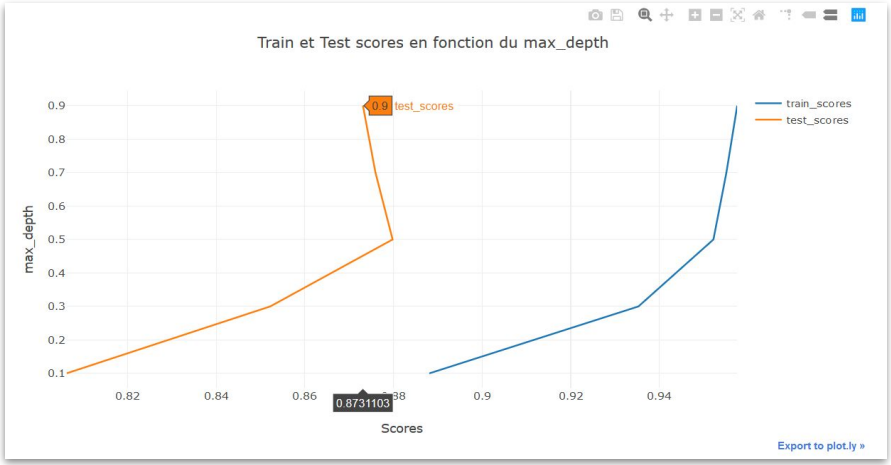
Gradient Boosting : Train vs Test Scores en fonction de fonction de coût et n_estimators





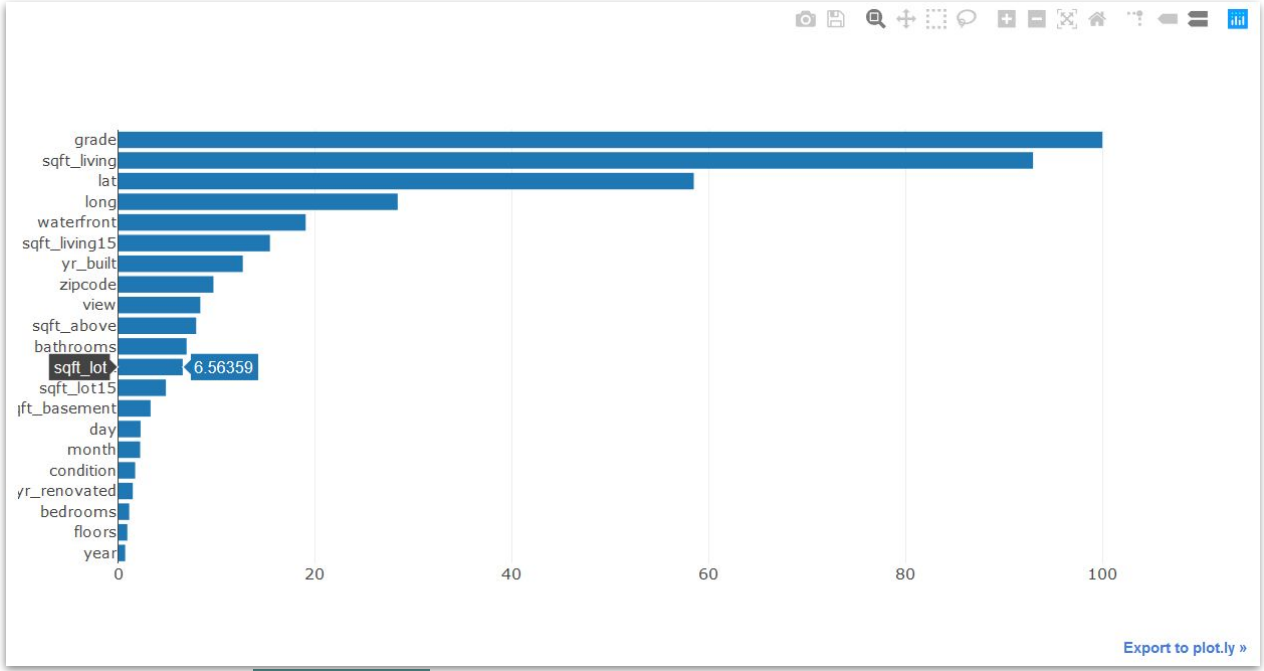
Gradient Boosting : Train vs Test Scores en fonction de learning_rate et max_depth





Intelligence Artificielles

Gradient Boosting : Meilleurs Paramètres



Intelligence Artificielles

Resultats

Algorithme	**R^2**
Gradient boosting	0.89069
Adaboost	0.7160
XGBoost	0.8907

Résultats et choix du meilleur algorithme

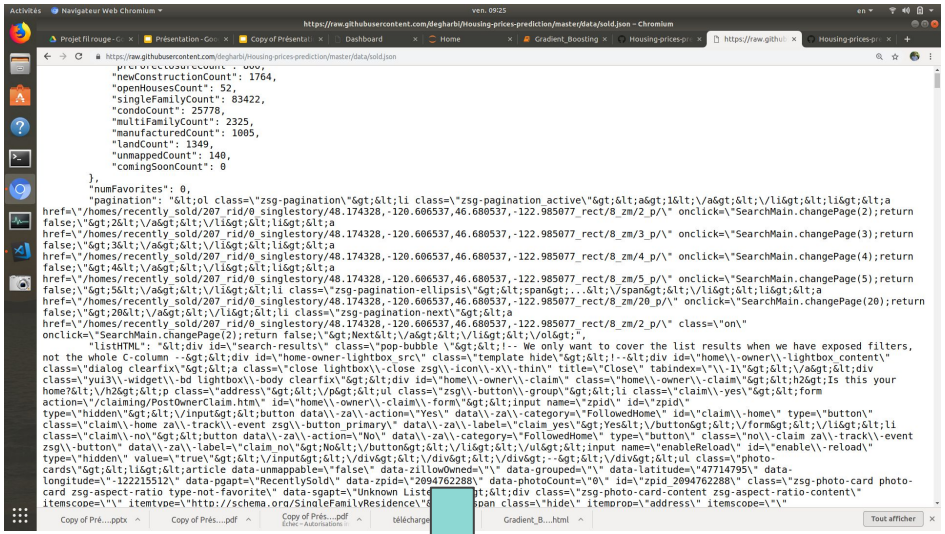
King County
USA

Meilleur algorithme

Algorithmes	R2
Regression	0.81804
Random Forest	0.883
Gradient Boosting	0.89069

Conclusion

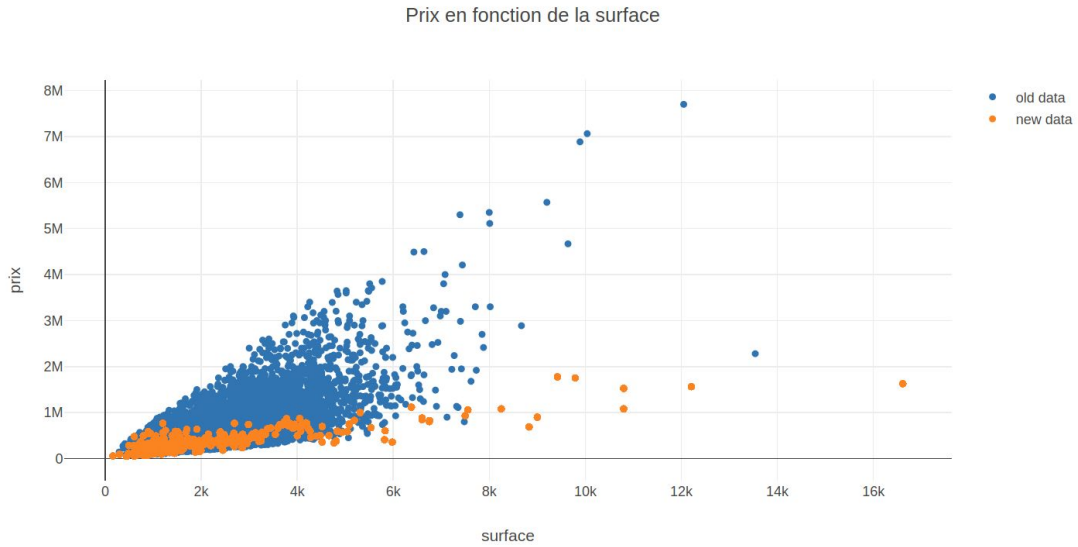
Test sur des données nouvelles




```
id,date,price,bedrooms,bathrooms,sqft_living,sqft_lot,floors,waterfront,view,condition,grade,sq
6123415060, '', '', 3.0, 2.0, 1810.0, 6198.0, '', '', '', '', '', '', '', -1, '', u'98177', 477
7488078255, '', '', 4.0, 4.0, 3080.0, 36154.0, '', '', '', '', '', '', '', -1, '', u'98027', 47
9844329224, '', '', 2.0, 1.0, 690.0, 4721.0, '', '', '', '', '', '', '', -1, '', u'98106', 475.
588296013, '', '', 4.0, 2.75, 2250.0, 17859.0, '', '', '', '', '', '', '', -1, '', u'98058', 47
4549248854, '', '', 4.0, 5.0, 4930.0, 10454.0, '', '', '', '', '', '', '', -1, '', u'98040', 47
```

Test sur des données nouvelles

Ajout des nouvelles lignes correspondant aux biens (scrapping Zillow) en 2018



Conclusions & Perspectives

King County
USA

Conclusion

- Parfois les modèles simples bien paramétrés donnent d'excellent résultats
- Preprocessing est trés important : garbage in .... Garbage out

Perspectives

- Enrichir les données avec données externes : open data, criminalité, infrastructure public...
- Trouver la meilleurs façon de compléter les données manquantes
- Construire un script automatisé qui choisit les meilleurs algorithmes
- Deep learning ?
- Généralisation du modèle sur d'autres régions des US
- Appliquer la même approche sur l'immobilier en france

Demo 2

https://shrouded-scrubland-74851.herokuapp.com/homepage/

Questions ?

King County
USA