



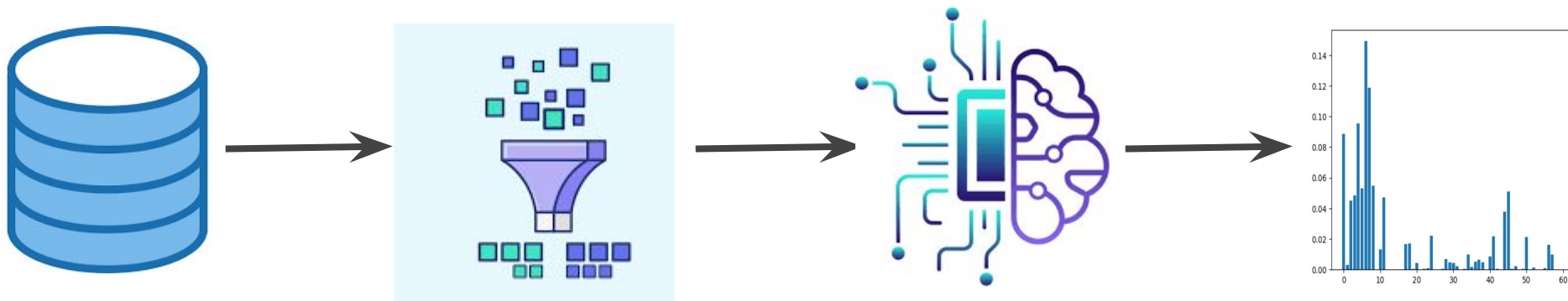
Seattle

OPENCLASSROOMS

Objectif : Ville neutre en émissions de carbone en 2050

Anissa TALEB
Projet n°4
Août 2023

Mission : Prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation



Data Engineering :

Explorer, transformer, nettoyer et préparer les données



Préparation des données: Métier

⇒ Filtrer les individus concernés par notre étude: Bâtiment non résidentiel



Préparation des données:Métier

⇒ Filtrer les individus concernés par notre étude: Bâtiment non résidentiel

⇒ Analyse énergétique conforme : 'ComplianceStatuts' = 'Compliant'



Préparation des données:Métier

⇒ Filtrer les individus concernés par notre étude: Bâtiment non résidentiel

⇒ Analyse énergétique conforme : 'ComplianceStatuts' = 'Compliant'

⇒ Exclure les features directement liée au calcul d"energy Star Score' : effet de Data leak



Préparation des données: Technique :

⇒ Binariser les colonnes de mesures énergétique = lecture plus simple

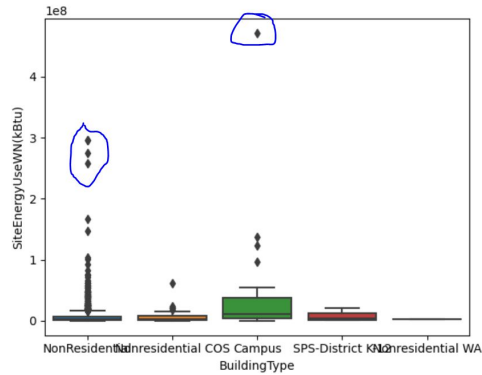


Préparation des données: Technique :

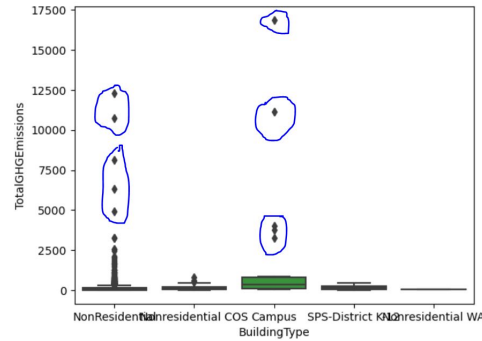
- ⇒ Binariser les colonnes de mesures énergétique = lecture plus simple
- ⇒ Catégoriser la colonne année



Nettoyage de jeu de donnée : Traitement des Outliers



Energy : 4/1548



CO2 : 10/1548

Nettoyage de jeu de donnée : Traitement des données manquantes

dfP4ENERGY.isnull().sum()

NumberOfFloors	0
NumberOfBuildings	0
BuildingType	0
PrimaryPropertyType	0
Latitude	0
Longitude	0
YearBuilt	0
PropertyGFATotal	0
PropertyGFAParking	0
PropertyGFABuilding(s)	0
LargestPropertyUseType	4
LargestPropertyUseTypeGFA	4
SiteEnergyUseWN(kBtu)	0
SteamUse(kBtu)	0
Electricity(kBtu)	0
NaturalGas(kBtu)	0
dtype: int64	

4 / 1548

dfP4CO2.isnull().sum()

NumberOfFloors	0
NumberOfBuildings	0
BuildingType	0
PrimaryPropertyType	0
Latitude	0
Longitude	0
YearBuilt	0
PropertyGFATotal	0
PropertyGFAParking	0
PropertyGFABuilding(s)	0
LargestPropertyUseType	4
LargestPropertyUseTypeGFA	4
TotalGHGEmissions	0
SteamUse(kBtu)	0
Electricity(kBtu)	0
NaturalGas(kBtu)	0
dtype: int64	

VS

dfP4ENERGYSTAR.isnull().sum()

NumberOfFloors	0
NumberOfBuildings	0
BuildingType	0
PrimaryPropertyType	0
Latitude	0
Longitude	0
YearBuilt	0
PropertyGFATotal	0
PropertyGFAParking	0
PropertyGFABuilding(s)	0
LargestPropertyUseType	4
LargestPropertyUseTypeGFA	4
SiteEnergyUseWN(kBtu)	0
SteamUse(kBtu)	0
Electricity(kBtu)	0
NaturalGas(kBtu)	0
ENERGYSTARScore	551
dtype: int64	

551/1548

dfP4CO2STAR.isnull().sum()

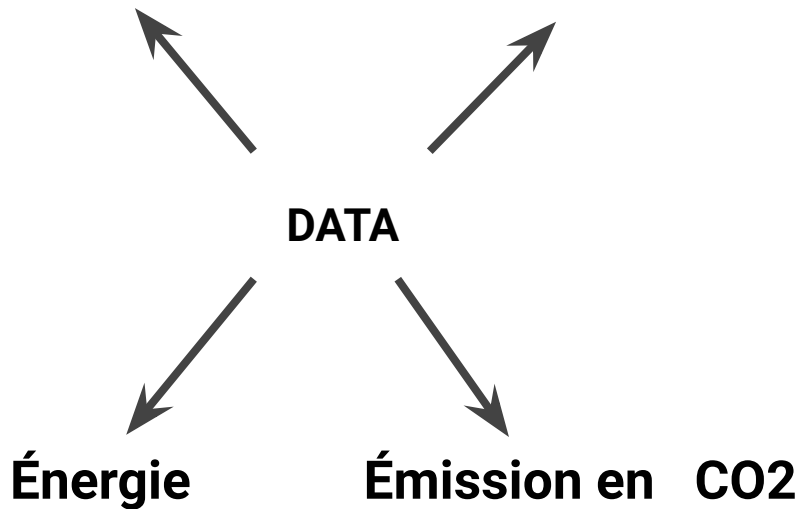
NumberOfFloors	0
NumberOfBuildings	0
BuildingType	0
PrimaryPropertyType	0
Latitude	0
Longitude	0
YearBuilt	0
PropertyGFATotal	0
PropertyGFAParking	0
PropertyGFABuilding(s)	0
LargestPropertyUseType	4
LargestPropertyUseTypeGFA	4
TotalGHGEmissions	0
SteamUse(kBtu)	0
Electricity(kBtu)	0
NaturalGas(kBtu)	0
ENERGYSTARScore	550
dtype: int64	

⇒ réduction significative de nombre d'individu avec energy starscore

Preparation de jeu de données à l'analyse :

Énergie+Energy Star Score

Émission en CO2 +Energy Star Score



Machine learning :

Créer et sélectionner des modèles basés sur les données



Préparation de données :

Preprocessing



Stratification



Standardisation



Stratification CV



Énergie:

RIDGE

Validation score r2: -0.5527231310562278

Test r2: 0.5090958988899592

ELASTIC NET

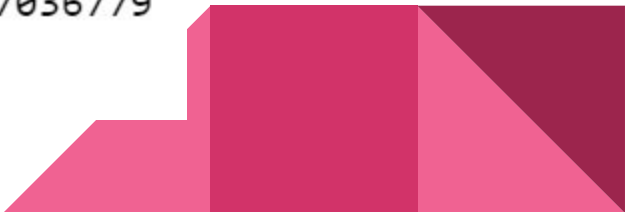
Validation score r2: 0.3099751376146085

Test r2: 0.42804301752083695

RANDOM FOREST

Validation score r2: 0.6641367137036779

Test r2: 0.7157614352093589



Énergie:

RIDGE

Validation score r2: -0.5527231310562278
Test r2: 0.5090958988899592

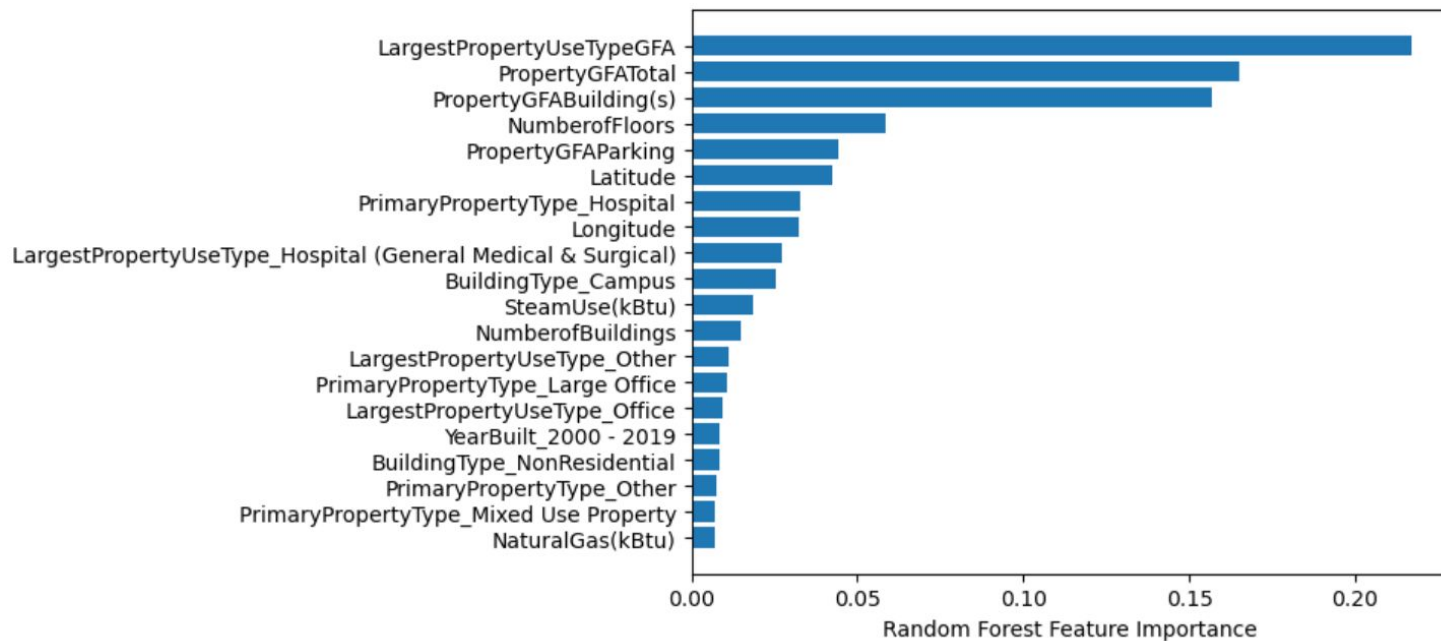
ELASTIC NET

Validation score r2: 0.3099751376146085
Test r2: 0.42804301752083695

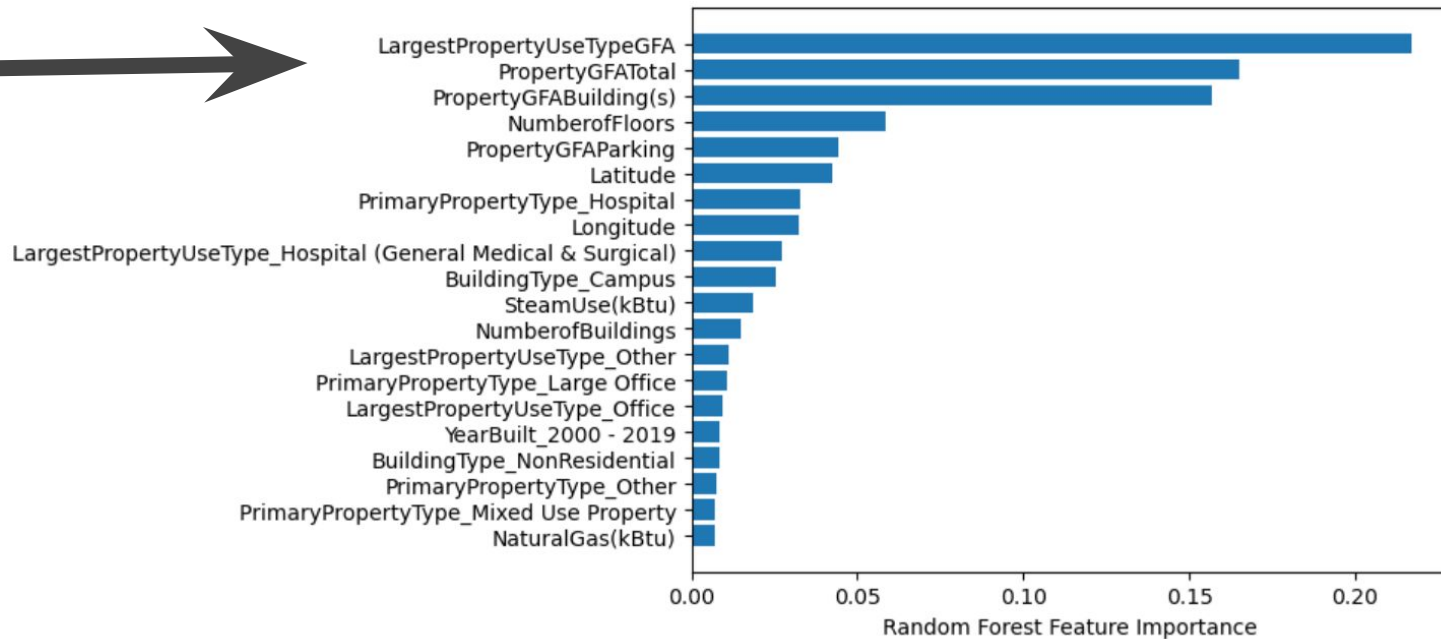
RANDOM FOREST

Validation score r2: 0.6641367137036779
Test r2: 0.7157614352093589

Énergie: feature importance



Énergie: feature importance



Énergie: shapley



CO2:

RIDGE

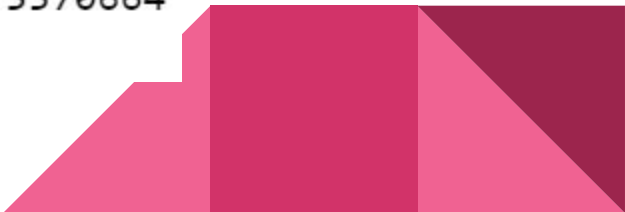
Validation score r2: 0.4930010337372187
Test r2: 0.2902345402107812

ELASTIC NET

Validation score r2: 0.5158386775570664
Test r2: 0.31932898356894324

RANDOM FOREST

Validation score r2: 0.5158386775570664
Test r2: 0.31932898356894324



CO2:

RIDGE

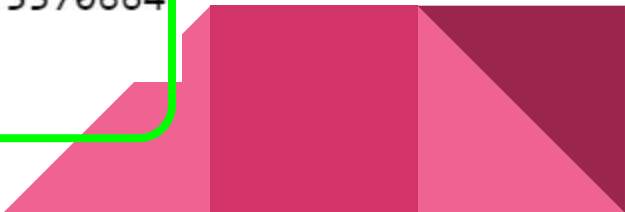
Validation score r2: 0.4930010337372187
Test r2: 0.2902345402107812

ELASTIC NET

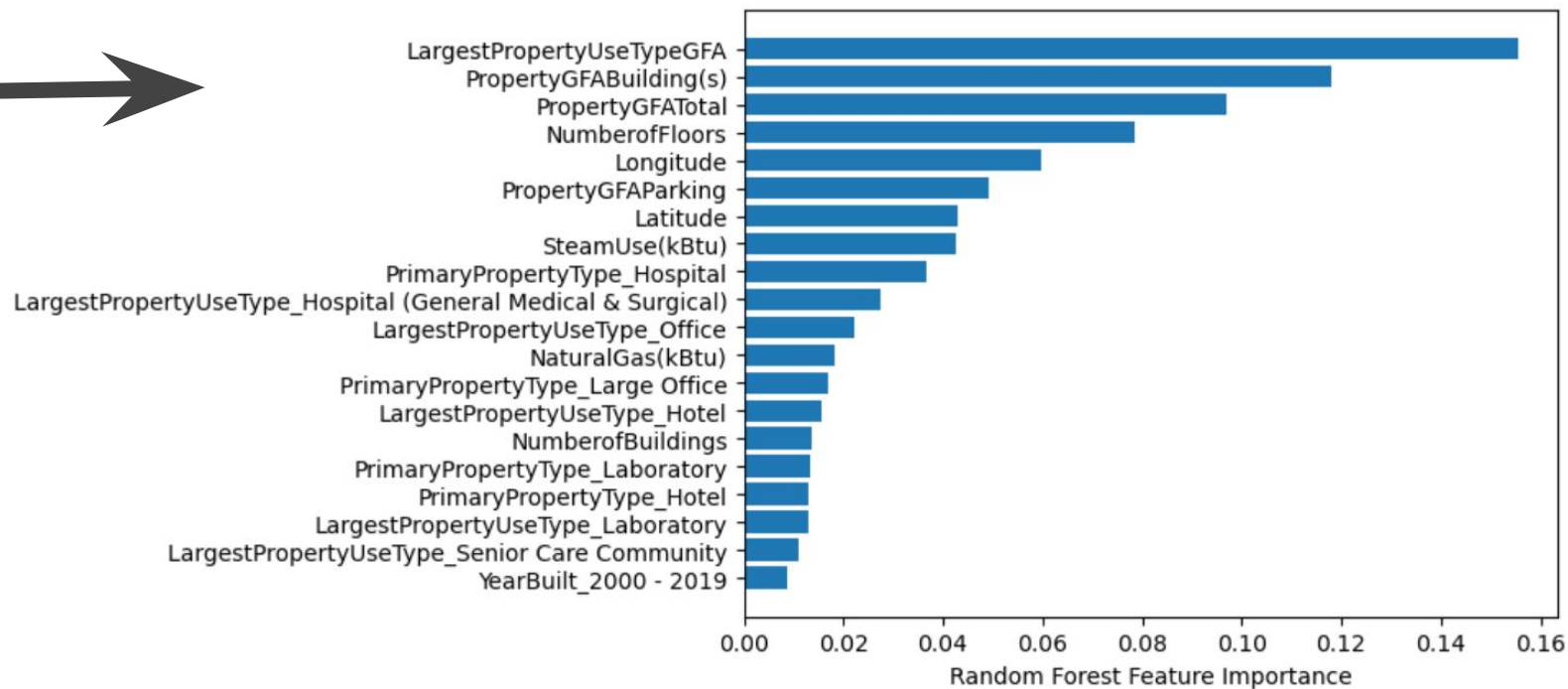
Validation score r2: 0.5158386775570664
Test r2: 0.31932898356894324

RANDOM FOREST

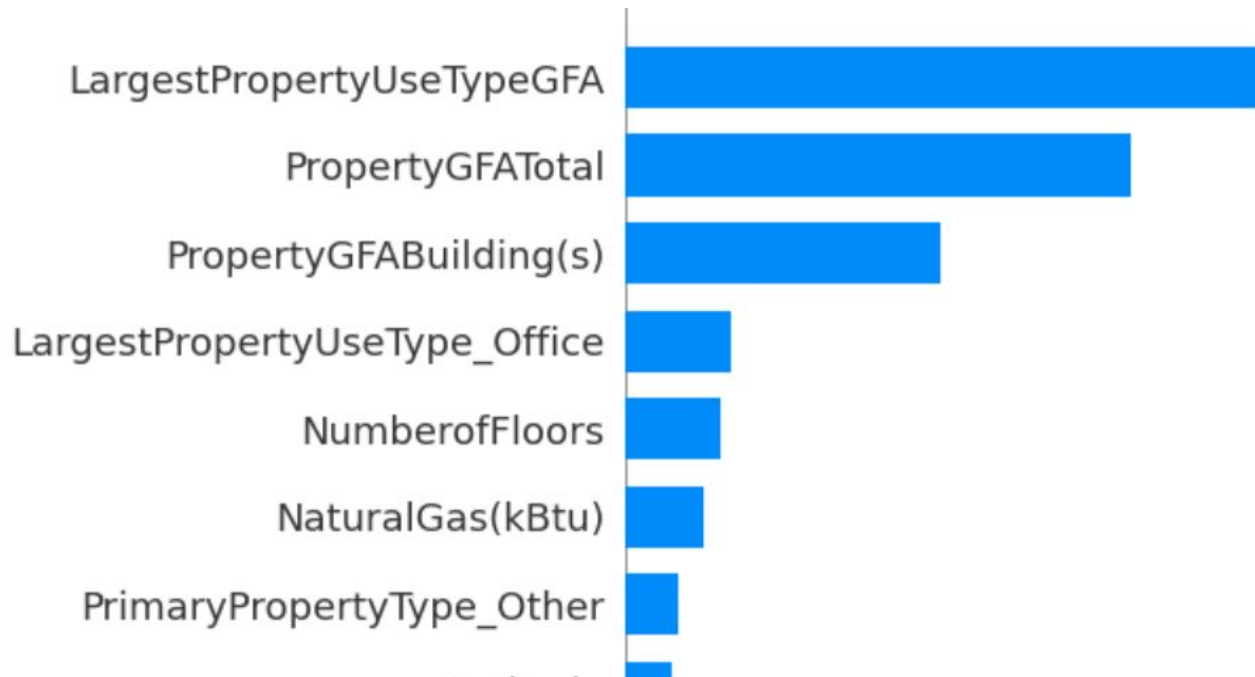
Validation score r2: 0.5158386775570664
Test r2: 0.31932898356894324



CO2: feature importance



CO2: shapley



Influence de ENERGYSTAR Score sur la prédiction d'énergie et de CO2: (Random Forest)

	données C.	données P.	données P. + Star score
Energie	r2_V : 0.664 r2_T: 0.715	r2_V : 0.682 r2_T: 0.647	r2_V : 0.713 r2_T: 0.777
CO2	r2_V : 0.515 r2_T: 0.319	r2_V : 0.567 r2_T: 0.515	r2_V : 0.539 r2_T: 0.686



Influence de ENERGYSTAR Score sur la prédiction d'énergie et de CO2: Top 3 feature importance

	données C.	données P.	données P. + Star score
Energie	'LargestPropertyUseTypeGFA' PropertyGFATotal PropertyGFABuilding	'LargestPropertyUseTypeGFA' PropertyGFATotal NumberofFloors	"LargestPropertyUseTypeGFA" PropertyGFATotal PropertyGFABuilding
CO2	LargestPropertyUseTypeGFA' PropertyGFATotal PropertyGFABuilding	"LargestPropertyUseTypeGFA" Latitude PropertyGFATotal	LargestPropertyUseType PrimaryPropertyType LargestPropertyUseType

Influence de ENERGYSTAR Score sur la prédiction d'énergie et de CO2: Top 3 Shapley

	données C.	données P.	données P. + Star score
Energie	"LargestPropertyUseTypeGFA PropertyGFABuilding PropertyGFATotal	"LargestPropertyUseTypeGFA PropertyGFABuilding NumberofFloors	"LargestPropertyUseTypeGFA PropertyGFATotal PropertyGFABuilding
CO2	"LargestPropertyUseTypeGFA PropertyGFATotal PropertyGFABuilding	"LargestPropertyUseTypeGFA Latitude PropertyGFATotal	PropertyGFATotal "LargestPropertyUseTypeGFA PropertyGFABuilding

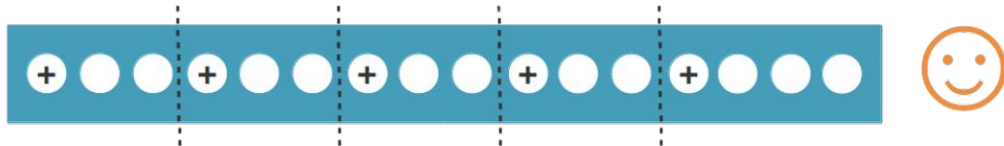
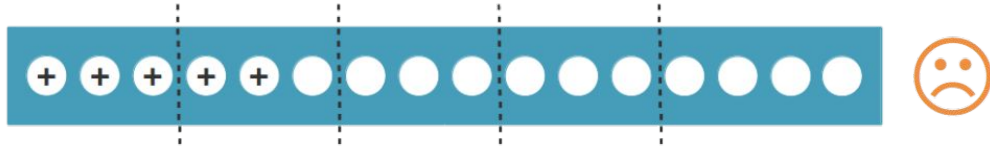
Conclusion :

TOP 3 des variables les plus influentes dans la prédiction sont :

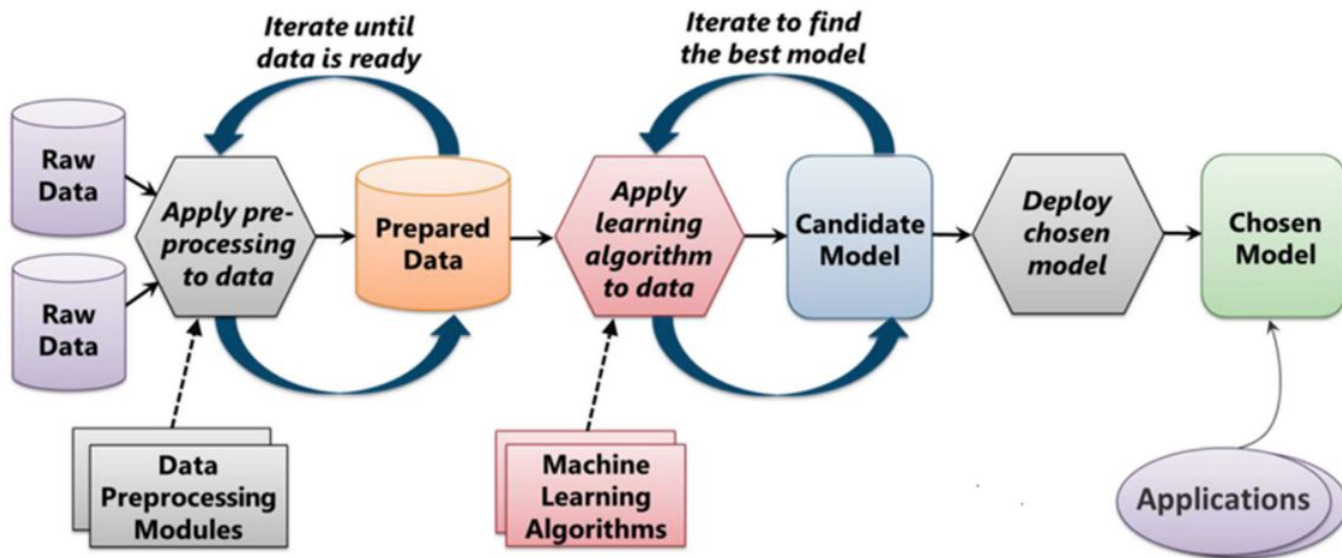
- PropertyGFATotal
- LargestPropertyUseType
- PropertyGFABuilding(s)



Stratification



The Machine Learning Process



From "Introduction to Microsoft Azure" by David Chappell

