Python for Data Analysis

- Prédiction du temps de clôture d'incidents

AOUES GAYA
IBO1

Découverte du dataset

- Journal d'évènements d'un outil de gestion des incidents
- 141 712 évènements avec 34 colonnes d'attributs pour les caractériser

	number	incident_state	active	reassignment_count	reopen_count	sys_mod_count	made_sla	caller_id	opened_by	opened_at	sys_created_by	sys_creat
0	INC0000045	New	True	0	0	0	True	Caller 2403	Opened by 8	29/2/2016 01:16	Created by 6	29/2/2016
1	INC0000045	Resolved	True	0	0	2	True	Caller 2403	Opened by 8	29/2/2016 01:16	Created by 6	29/2/2016
2	INC0000045	Resolved	True	0	0	3	True	Caller 2403	Opened by 8	29/2/2016 01:16	Created by 6	29/2/2016

Prédire le temps avant complétion d'un incident

La variable à prédire

L’objectif est de prédire à quel moment un incident va être clôturé.

- Dans le dataset, on a la date de clôture qui est **closed_at**

Problème:

Closed_at est une date et non une quantité.

Solution:

Prendre comme target la durée totale de l’incident en heures. Autrement dit, la durée entre le moment d’ouverture et le moment de clôture.

DureeOpenToClose= closed_at - opened_at

Colonne créée: Temps_Resolution

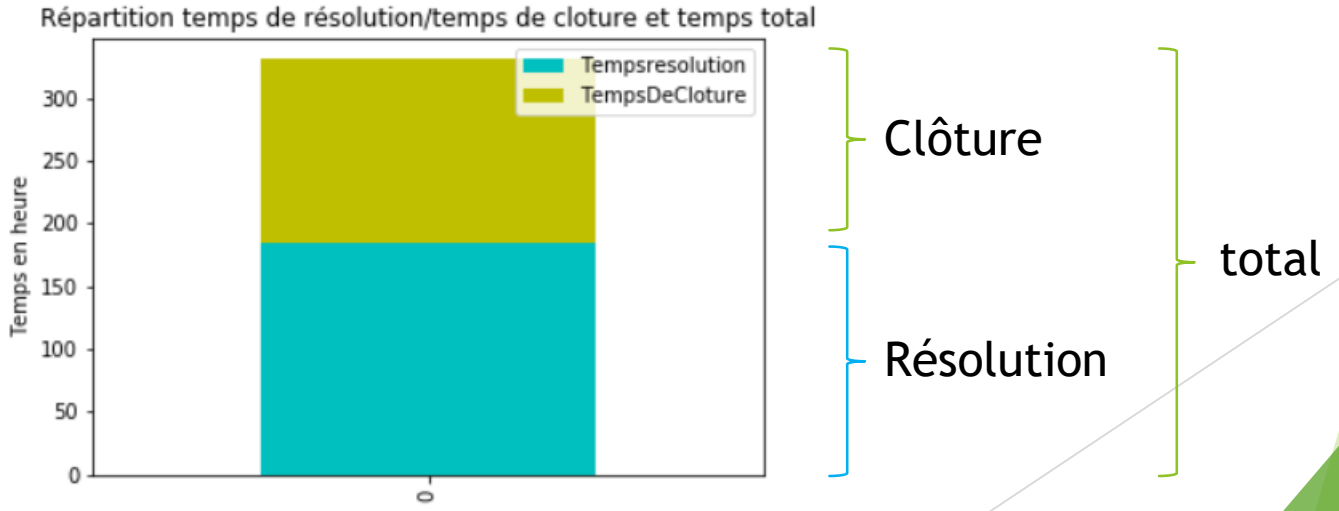
Le temps entre la résolution et la cloture est parfois conséquent.

=> On crée d'abord cette durée: Duree_ResolvedToClosed= closed_at – resolved_at

La vraie durée de résolution correspond au temps total - ce temps entre la résolution :

Temps_resolution= (Closed_at - opened_at) - Duree_ResolvedToClose

Ex: pour le premier incident= 130 h entre l'ouverture et la clôture mais il y'a 120 h entre le moment où l'incident est déjà résolu et sa clôture. Donc, le temps de résolution est 10 h = 130 h - 120 h



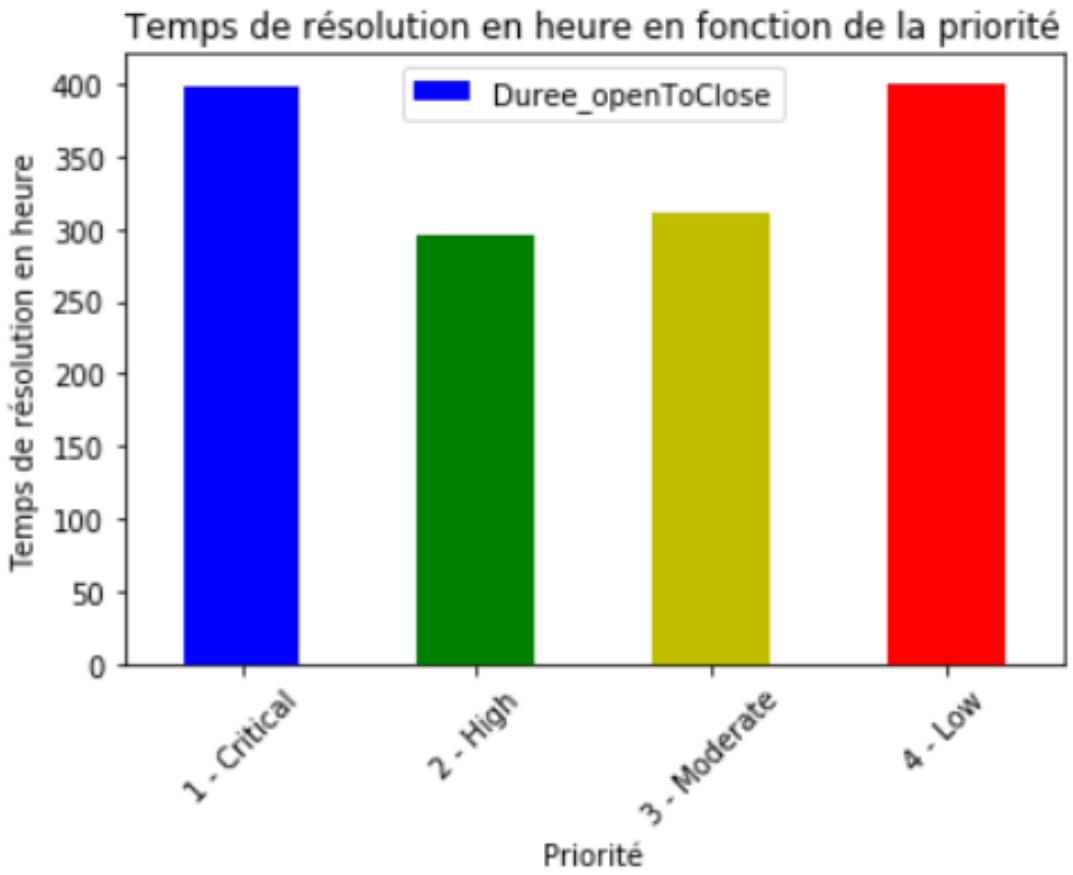
Les colonnes du dataset

- Sur les 34 colonnes: 3 seulement sont des variables numériques et 4 sont des booléens
- La majorité des colonnes sont des variables catégoriques.
- Certaines colonnes ont plus de 95% de données inconnues.

Traitements à faire:

- Colonnes numériques: Remplacer par la médiane ou la moyenne pour les valeurs manquantes
- Booléens: Compris par l'algorithme comme des entiers
- Colonnes catégoriques: Dummification pour binariser les valeurs.

Visualisation d’une variable catégorique



Agrégation ou non ?

Dans le dataset, un incident est représenté par une suite d'évènements. Autrement dit, plusieurs rows pour un incident.

	number	incident_state	active	reassignment_count	reopen_count	sys_mod_count	made_sla	caller_id	opened_by	opened_at	sys_created_by	sys_crea
0	INC0000045	New	True	0	0	0	True	Caller 2403	Opened by 8	29/2/2016 01:16	Created by 6	29/2/2016
1	INC0000045	Resolved	True	0	0	2	True	Caller 2403	Opened by 8	29/2/2016 01:16	Created by 6	29/2/2016
2	INC0000045	Resolved	True	0	0	3	True	Caller 2403	Opened by 8	29/2/2016 01:16	Created by 6	29/2/2016
3	INC0000045	Closed	False	0	0	4	True	Caller 2403	Opened by 8	29/2/2016 01:16	Created by 6	29/2/2016
4	INC0000047	New	True	0	0	0	True	Caller 2403	Opened by 397	29/2/2016 04:40	Created by 171	29/2/2016

- Les 4 premières lignes représentent le même incident
- Les 4 ont la même date d'ouverture et de fermeture

Agrégation ou non ?

- Des modèles sans agrégations dans un premier temps seront tentés avant d'essayer avec une agrégation sur la colonne number qui correspond à l'id de l'incident.

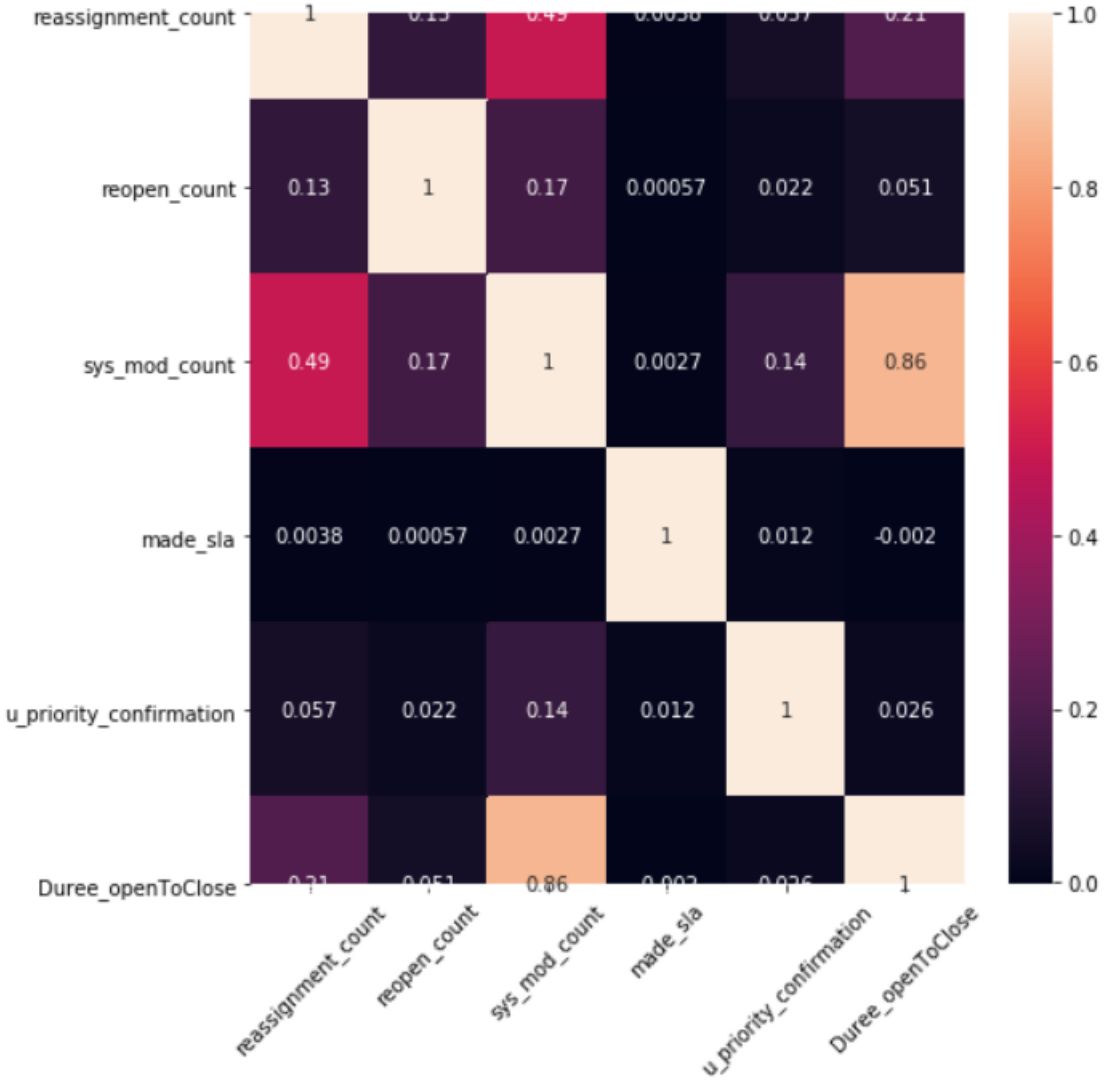
- Dans cette agrégation le plus intéressant est de retenir le maximum des colonnes numériques (**reassignment_count, reopen_count, sys_mod_count**)

- Le max de **sys_mod_count** correspond au nombre de fois au total qu'un incident a été mis à jour. Donc plus ce chiffre est élevé plus l'incident prend du temps à être clôturé.

```
dfvisu3=dfvisu3.groupby('number').agg({'reassignment_count':'max',
                                       'reopen_count':'max',
                                       'sys_mod_count':'max',
                                       'made_sla':'max',
                                       'u_priority_confirmation':'max',
                                       'Duree_openToClose': 'max'})
```

Matrice de Corrélation

On voit bien dans cette matrice de corrélation que la variable la plus liée à la durée de clôture d'un incident est la variable **Sys_mod_count**



Préprocessing

- Transformation des colonnes de temps en DateTime
- Calcule de la durée OpenToClose qui est notre target
- Remplacement des valeurs manquantes pour les colonnes numériques par la Moyenne de la colonne
- Agrégation de incidents sur la colonne number (id de l'incident)
- Dummification des colonnes catégoriques

Dummification

La dummification des colonnes catégoriques crée autant de colonnes que de catégories pour la colonne dummifiée.

⇒ Il faut donc dummifier uniquement les colonnes qui n'ont pas beaucoup de catégories au risque de voir notre dataset passer à plusieurs centaines voir milliers de colonnes

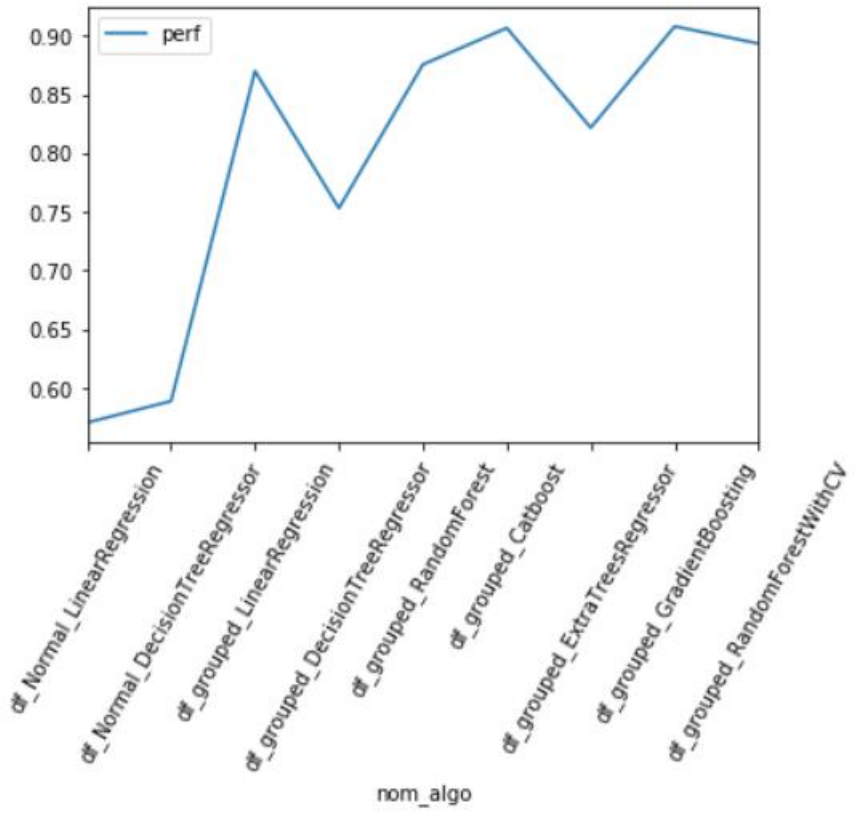
Les colonnes dummifiées: [category, impact, urgency, vendor, closed_code, priority, contact_type, caused_by]

Exemple de code de dummification:

```
dfP=dfGrouped['priority'].copy()
dum_dfp = pd.get_dummies(dfP, columns=["priority"], prefix='priority' )
dfGrouped=dfGrouped.join(dum_dfp)
```

Comparaison des modèles essayés

- Régression linéaire sans agrégation
- DecisionTreeRegressor sans agrégation
- Régression linéaire
- DecisionTreeRegressor
- RandomForest
- Catboost
- GradientBoosting
- Random Forest après Grid Search



Grid Search

Par le biais d'un GridSearch appliqué à un RandomForest, les meilleures paramètres du modèle ont été trouvés

```
Entrée [816]: grid.best_params_

   Out[816]: {'bootstrap': True,
              'max_features': 'auto',
              'min_samples_split': 8,
              'n_estimators': 20}
```

Avec ces paramètres, une amélioration du score de 6 % a été obtenue

Api Django

Requête Post envoyée sur le endpoint http://localhost:8000/prediction/predict/


```
{
    "reassignment_count":0,
    "reopen_count":0,
    "sys_mod_count":4,
    "made_sla":1,
    "u_priority_confirmation":1,
    "knowledge":0,
    "Duree_openToClose":null
}
```


Prédiction sous forme JSON reçue:


```
{
    "reassignment_count": 0.0,
    "reopen_count": 0.0,
    "sys_mod_count": 4.0,
    "made_sla": 1.0,
    "u_priority_confirmation": 1.0,
    "knowledge": 0.0,
    "Duree_openToClose": 283.36677444058523
}
```