PREDICTION RETARD AVION

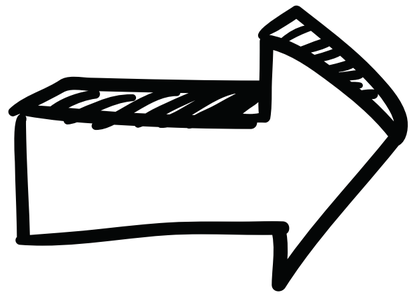
RYAD ET MATHIEU

• 2024

NOTRE OBJECTIFS

Prédire si un vol d’avion sera en retard ou non

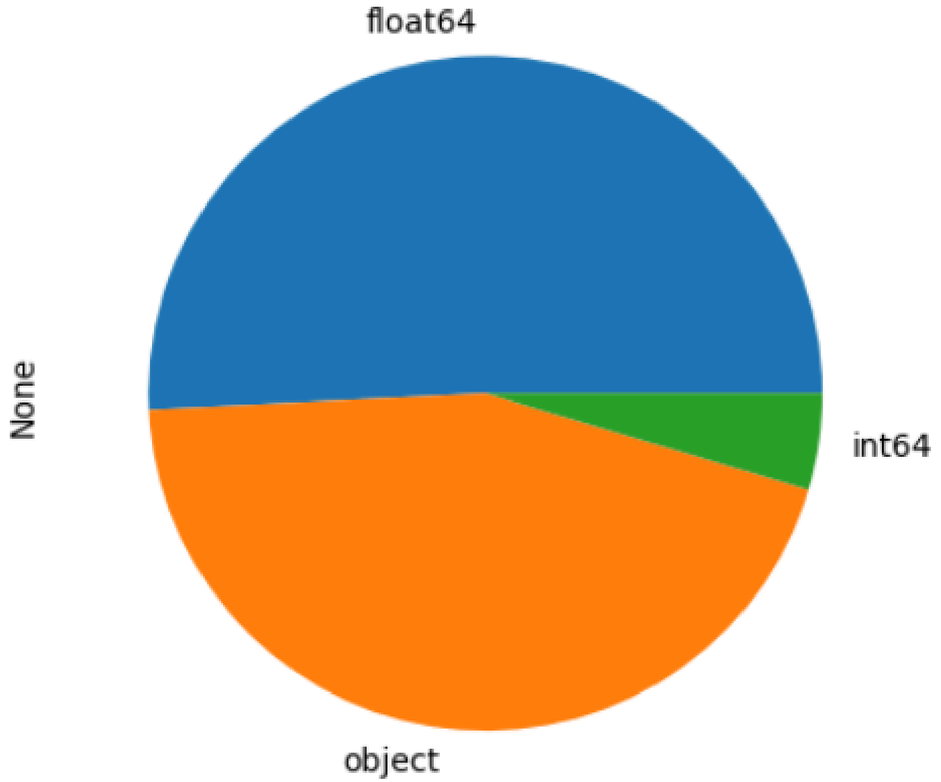
COMPRENDRE NOS DONNÉES



Jeu de donnée

	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	FL_DATE	UNIQUE_CARRIER	AIRLINE_ID	CARRIER	TAIL_NUM	...	DISTANCE_GROUP
0	2016	1	1	6	3	2016-01-06	AA	19805	AA	N4YBAA	...	4.0
1	2016	1	1	7	4	2016-01-07	AA	19805	AA	N434AA	...	4.0
2	2016	1	1	8	5	2016-01-08	AA	19805	AA	N541AA	...	4.0
3	2016	1	1	9	6	2016-01-09	AA	19805	AA	N489AA	...	4.0
4	2016	1	1	10	7	2016-01-10	AA	19805	AA	N439AA	...	4.0
...	...	...	...	...	...	...	...	...	...	...	...	...

+ 5 millions de lignes
65 colonnes



float64(33)

int64(3)

object(29)

IDENTIFICATION TARGET

ARR_DELAY_NEW

ARR_DELAY_NEW
0.0
0.0
7.0
0.0
113.0

Indique si un avion est en retard

> 0 = retard

< 0 = en avance

NETTOYAGE

- **Sélection de colonnes**
  - Suppression des colonnes après atterrissage
  - Suppression des colonnes dont l’utilisateur n’a pas la connaissance

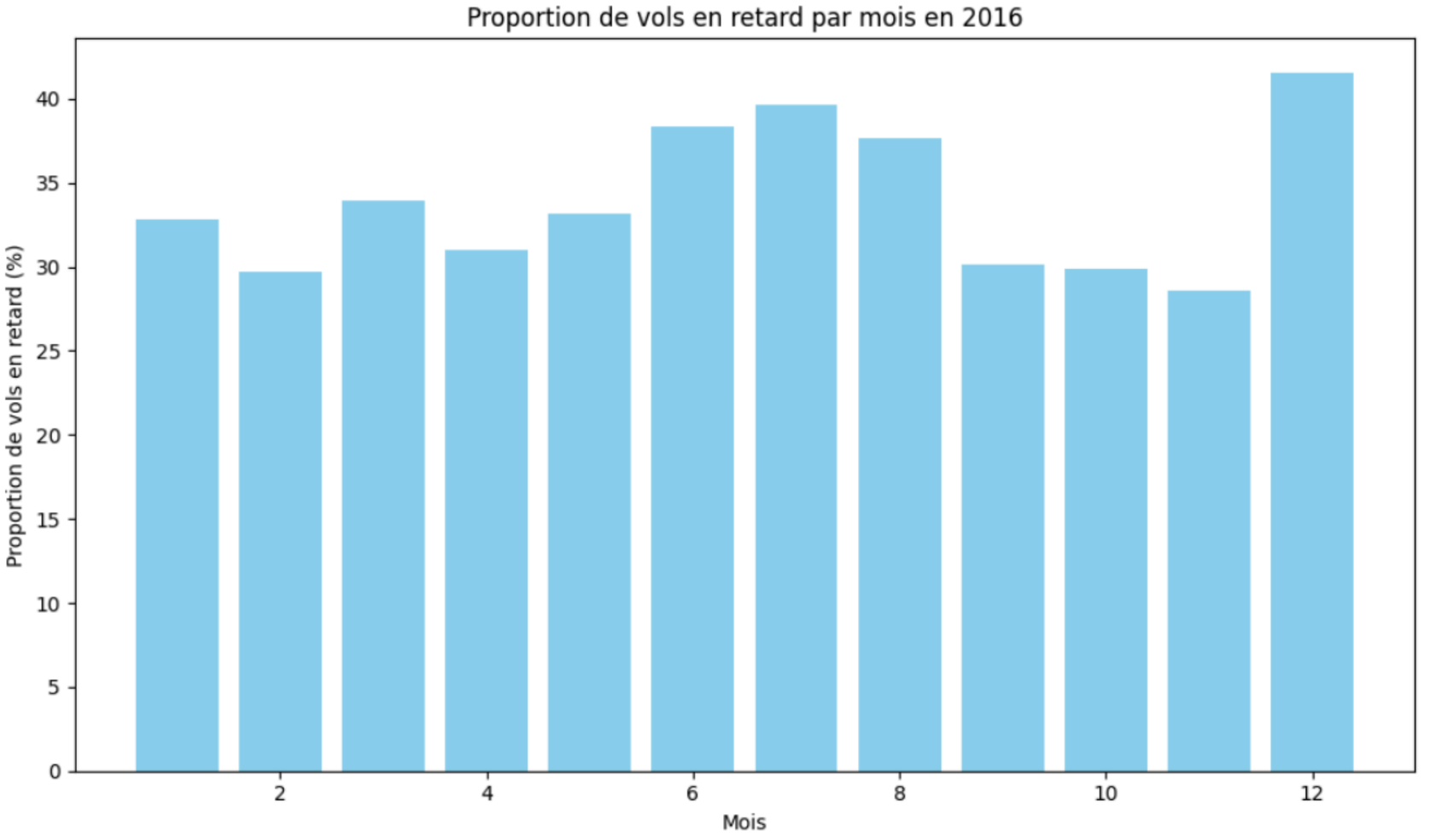
- **Check de toutes les colonnes pour voir les valeurs unique**

- **Suppression des valeur NAN**

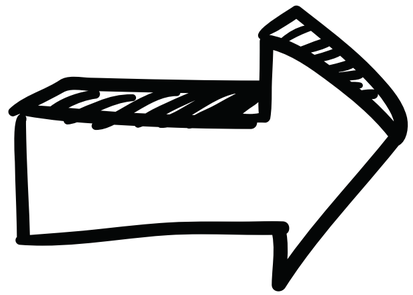
DATA SET NETTOYER

	ARR_DELAY_NEW	YEAR	QUARTER	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	ORIGIN_AIRPORT_ID	ORIGIN_CITY_NAME	ORIGIN_STAT
0	0.0	2016	1	1	6	3	AA	11298.0	Dallas/Fort Worth, TX	
1	0.0	2016	1	1	7	4	AA	11298.0	Dallas/Fort Worth, TX	
2	7.0	2016	1	1	8	5	AA	11298.0	Dallas/Fort Worth, TX	
3	0.0	2016	1	1	9	6	AA	11298.0	Dallas/Fort Worth, TX	
4	113.0	2016	1	1	10	7	AA	11298.0	Dallas/Fort Worth, TX	

16 colonnes


Proportion de vols en retard par mois en 2016
Proportion de vols en retard (%)
0
5
10
15
20
25
30
35
40
2
4
6
8
10
12
Mois

FEATURE SELECTION



IMPORTANCE

indique le poids des features dans le modèle.

Plus le score est haut plus la feature est importante.

```
DAY_OF_MONTH          0.309206
DAY_OF_WEEK           0.137379
CRS_ARR_TIME          0.111819
CRS_DEP_TIME          0.101597
MONTH                 0.068845
DISTANCE              0.050818
DEST_AIRPORT_ID       0.032541
DEST_CITY_NAME        0.031718
ORIGIN_AIRPORT_ID     0.031120
ORIGIN_CITY_NAME      0.030432
UNIQUE_CARRIER        0.026085
DEST_STATE_ABR        0.025639
ORIGIN_STATE_ABR      0.024864
QUARTER               0.017936
YEAR                  0.000000
dtype: float64
```

COLINÉARITÉ

Variance Inflation Factor (VIF):

	Feature	VIF
0	YEAR	NaN
1	QUARTER	17.531539
2	MONTH	17.525829
3	DAY_OF_MONTH	1.000063
4	DAY_OF_WEEK	1.001269
5	UNIQUE_CARRIER	1.034189
6	ORIGIN_AIRPORT_ID	2.527879
7	ORIGIN_CITY_NAME	2.582830
8	ORIGIN_STATE_ABR	1.088012
9	DEST_AIRPORT_ID	2.527203
10	DEST_CITY_NAME	2.586319
11	DEST_STATE_ABR	1.087470
12	CRS_DEP_TIME	1.842900
13	CRS_ARR_TIME	1.836114
14	DISTANCE	1.043277

YEAR apparait en NaN car c’est une constante

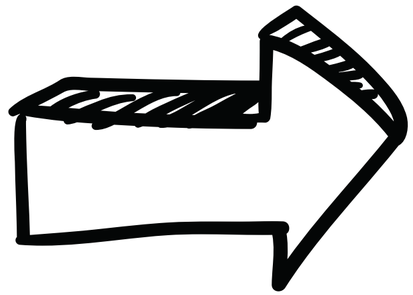
Un Vif Index supérieur à 10 est statistiquement significative.

Nous ne choisirons pas les features QUARTER et MONTH.

LISTE FINAL

DAY_OF_MONTH / jour du mois
DAY_OF_WEEK / 1 = Lundi
CRS_DEP_TIME / heure de départ
ORIGIN_AIRPORT_ID / Aéroport de départ
CRS_ARR_TIME / Heure d'arrivée
DEST_AIRPORT_ID / Aéroport d'arrivée

MODÈLES



PREPROCESS

Encodage de la Target - permet la classification.

Normalisation des données.

Encodage des colonnes catégorielles.

CHOIX DU METRIQUES

- Le recall en classe 1 est supérieur. Cette métrique permet de minimiser le risque de rater un vrai retards.

- Le taux de Faux Positif est inférieur. Cela est important pour le confort des passagers et leurs confiance dans le système.

BASELINE

Utilisation d’un Dummy model : Modèle simple
Sans equilibrage.

- Recall de 0.
- 377 688 FP

	precision	recall	f1-score	support
0	0.66	1.00	0.80	733551
1	0.00	0.00	0.00	377688
accuracy			0.66	1111239
macro avg	0.33	0.50	0.40	1111239
weighted avg	0.44	0.66	0.52	1111239

Confusion Matrix:
[[733551 0]
[377688 0]]
Accuracy: 0.6601199201971854

LOGISTIC REGRESSION

Modèle relativement simple mais adapté à ce genre de problématique :
Sous estimation réalisé de la classe principale

- Recall de 0,57.
- 160 978 FP

L’amélioration du score, incitation à l’essai d’un nouveau modèle

RANDOM FOREST

Modèle plus complexe :
Sous estimation réalisé de la classe principale

- Recall Target de 0.62.
- 143 275 FP

Il nous offre les meilleurs resultats sur les métriques choisit.

BASELINE

RECALL DE 0

377 688 FP

LOGISTIC REGRESSION

RECALL DE 0.57

160 978 FP

RANDOM FOREST

RECALL DE 0.62

143 275 FP

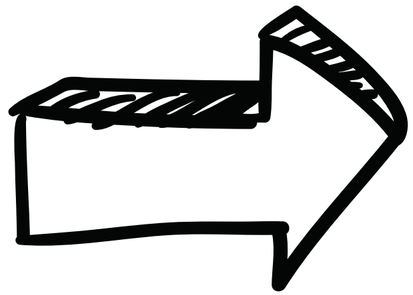
CHOIX MODELES

Nous avons décidé de privilégier Random Forest.

Il est possible d’adapter le choix du modèle à une autre utilisation.

Ex : Gestion de l’espace aérien, et adapter nos observations

DEMO



MERCI POUR VOTRE ATTENTION