CentraleSupélec

OPENCLASSROOMS

Parcours Data Scientist

Projet 4 :

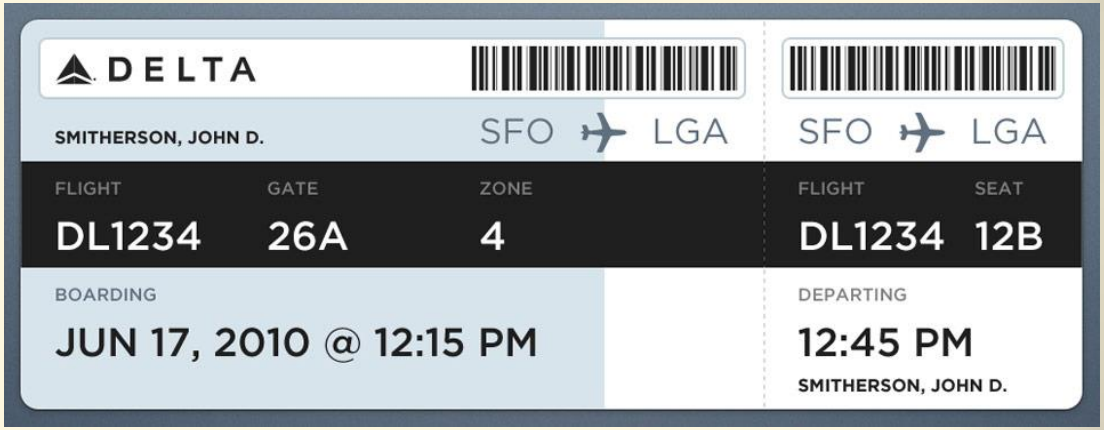
Anticipation du retard de vol des avions

24/01/09 ENREGISTREMENTS 12h00

HEURE	DESTINATION	VOL	HALL	BANQUES	OBSERVATION
06:20	AMSTERDAM	AF 8300	B	35 ▲ 43	
06:20	AMSTERDAM	KL 1314	B	31 ▲ 46	
08:30	LYON	EZY 4102	A	18 ▲ 19	RETARDÉ
10:55	MARRAKECH	8A 159	A	14 ▲ 16	RETARDÉ
10:55	MARRAKECH	AT 9797	A	14 ▲ 16	RETARDÉ
11:25	MARSEILLE	AF 5452	A	22 ▲ 25	ANNULÉ
11:35	MADRID	IB 8551	A	11 ▲ 12	
12:00	AGADIR	8A 2045	A	16 ▲ 18	RETARDÉ
12:40	AMSTERDAM	KL 1318	B	31 ▲ 46	ANNULÉ
12:40	AMSTERDAM	AF 8340	B	35 ▲ 43	ANNULÉ
13:00	PARIS-ORLY	AF 6263	B	35 ▲ 43	ANNULÉ
13:30	PARIS-CDG	AZ 3647	B	35 ▲ 43	ANNULÉ
13:30	PARIS-CDG	DL 8327	B	35 ▲ 43	ANNULÉ
13:30	PARIS-CDG	AF 7627	B	35 ▲ 43	ANNULÉ
13:40	LYON	AF 7805	B	35 ▲ 43	

Sommaire

- Présentation et Objectifs
- Nettoyage
- Exploration
- Modèles
  - Modèle 1
  - Modèle 2
- Interprétation
- API
- Pistes d'évolutions
- Conclusion



Présentation

- Entrée
  - Données concernant les vols (USA)
  - 1 an
  - 5,6 millions de vols
  - 65 features
  - Aucune donnée manquante
- Objectifs
  - Faire un modèle prédictif des retards
- Contrainte
  - « Accessible » à l'utilisateur

Nettoyage

- Nettoyage global
  - Suppression de Features
    - Features redontantes
      - AirlineID, Carrier, dates, airports, …
    - Features inconnus
      - Tail Number
    - Features inutiles
      - Flight number, wheelsON/OFF
    - Features imprévisibles/semi-prévisible
      - Weather Delay, Security Delay
  - Features nécessaires
    - Date et heure
    - Compagnie
    - Aéroport de départ (et d'arrivé)

Nettoyage

Modèle 1

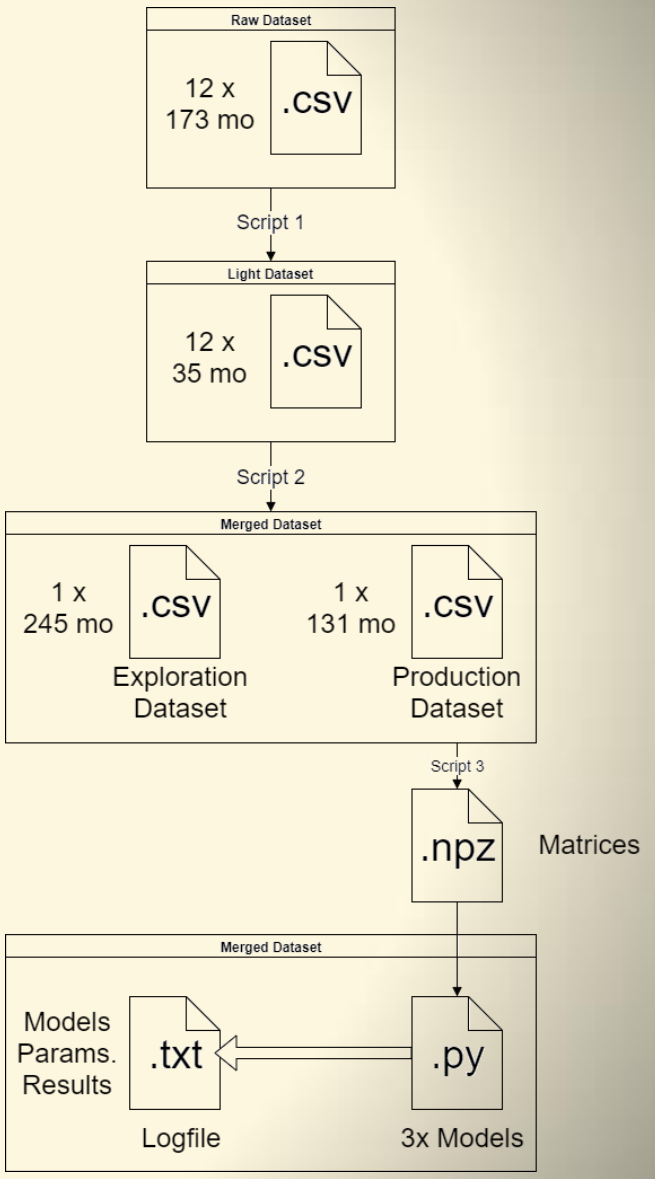
- Mois d'avril – données erronées
- Suppression des vols
  - Retard semi-prévisible > 60 min
    - Weather/NAS/Security delay, Late Aircraft Delay
  - Retard imprévisible > 0 min
    - Cancelled, Diverted

Modèle 2

- Mois d'avril – données erronées
- Agrégation par semaine/jour de la semaine/heure/aéroport de départ
- Pas d'utilisation de l'aéroport d'arrivé
- Ajout du nombre de vols

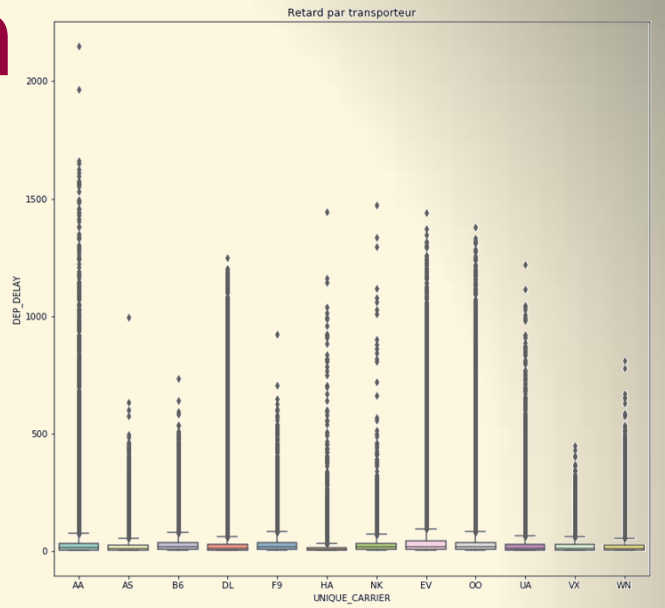
Nettoyage

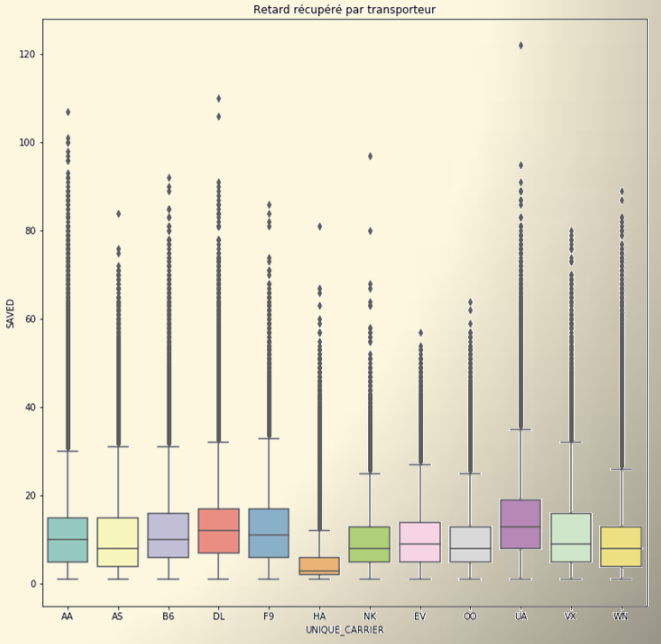
- Procédé **Modèle 1**:
  - Script 1:
    - Allègement par mois
  - Script 2:
    - Regroupement des 12 mois
    - Fin du nettoyage
    - Génération des datasets
      - Exploration
      - Production
  - Script 3 :
    - Scaling
    - Sauvegarde Train/Test sets en matrices
  - Script suivants:
    - 1 modèle par script
    - Ecrit les résultats dans un fichier texte



Exploration

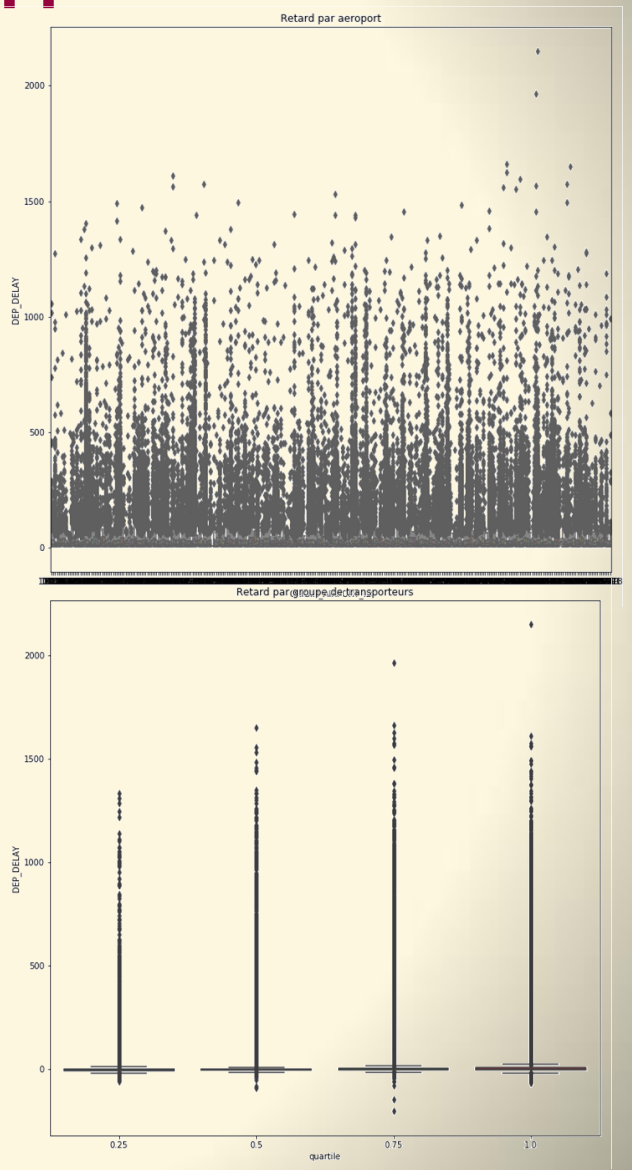
- Retard par Transporteur
  - Très variable
  - Non linéaire
  - Retard récupéré par Transporteur
    - Très variable
    - Non linéaire
- Simplification
  - OHE





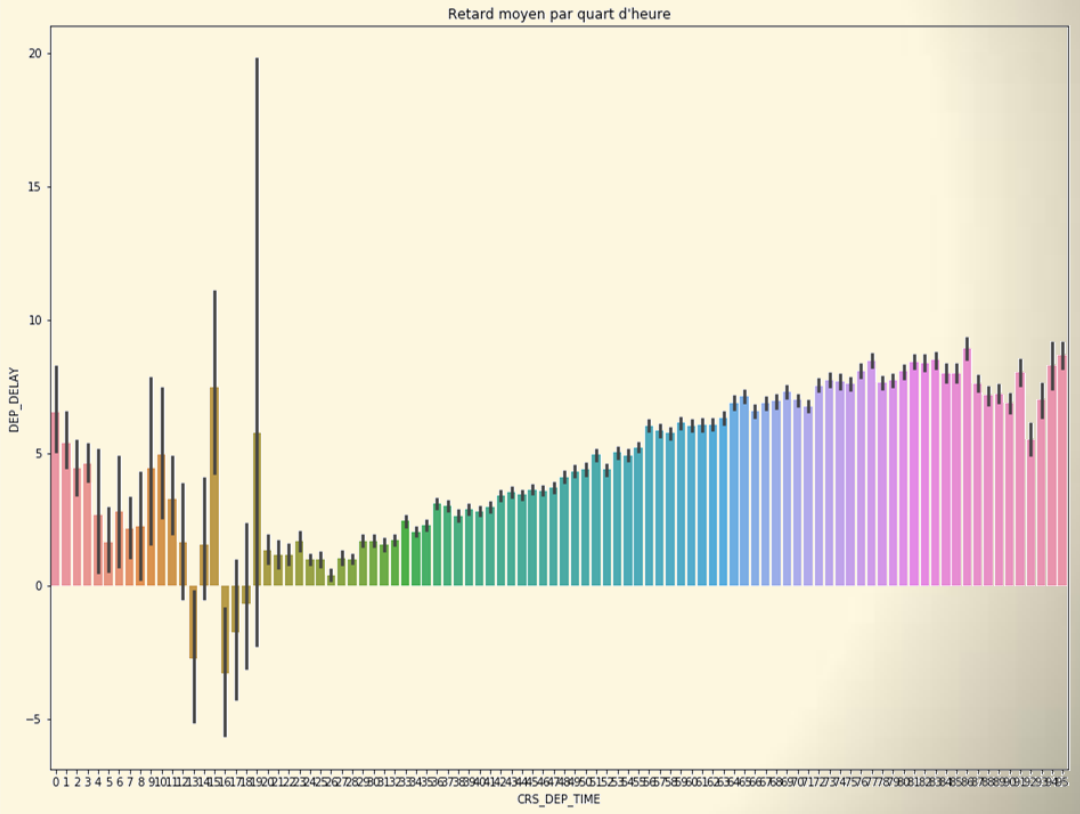
Exploration

- Retard par aéroport
  - Variable
  - Non linéaire
- Regroupement
  - 4 groupes par quartile
  - $retard = \alpha * groupe + b$



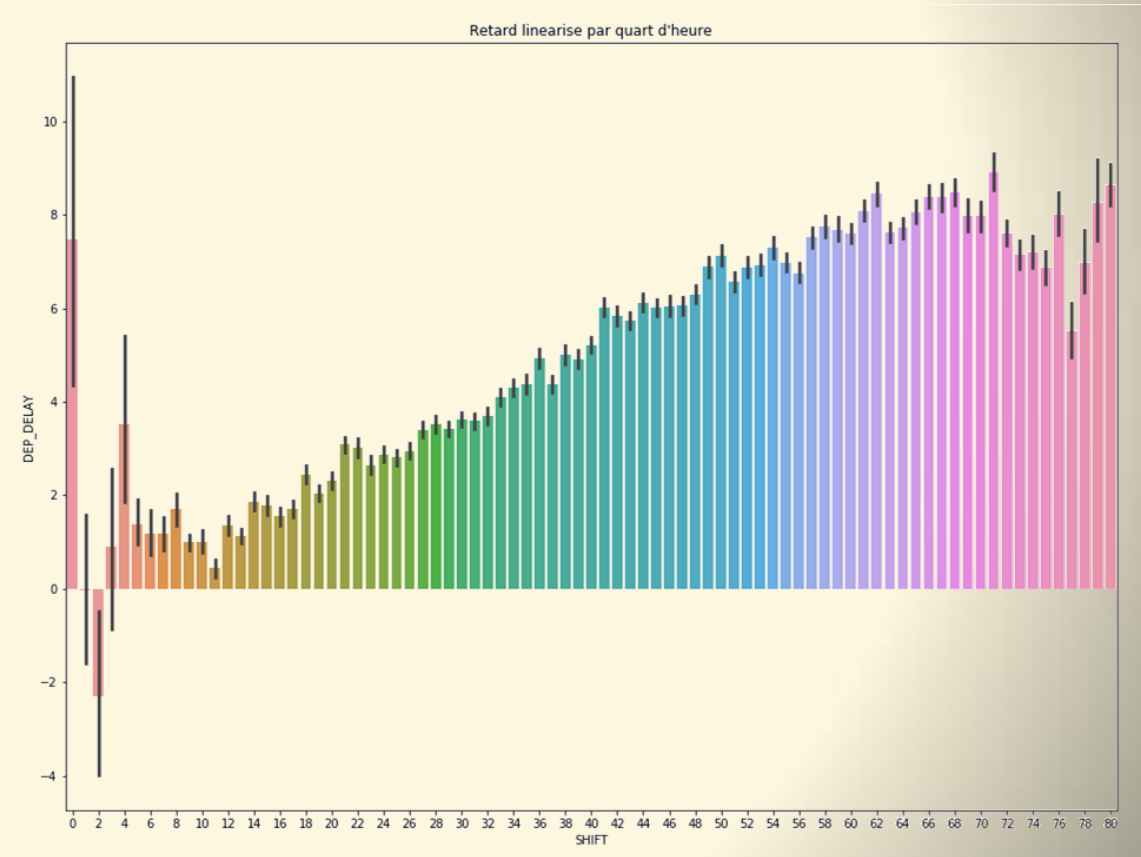
Exploration

- Retard par Date (Modèle 1)
  - Par 15 minutes
    - Assez Linéaire



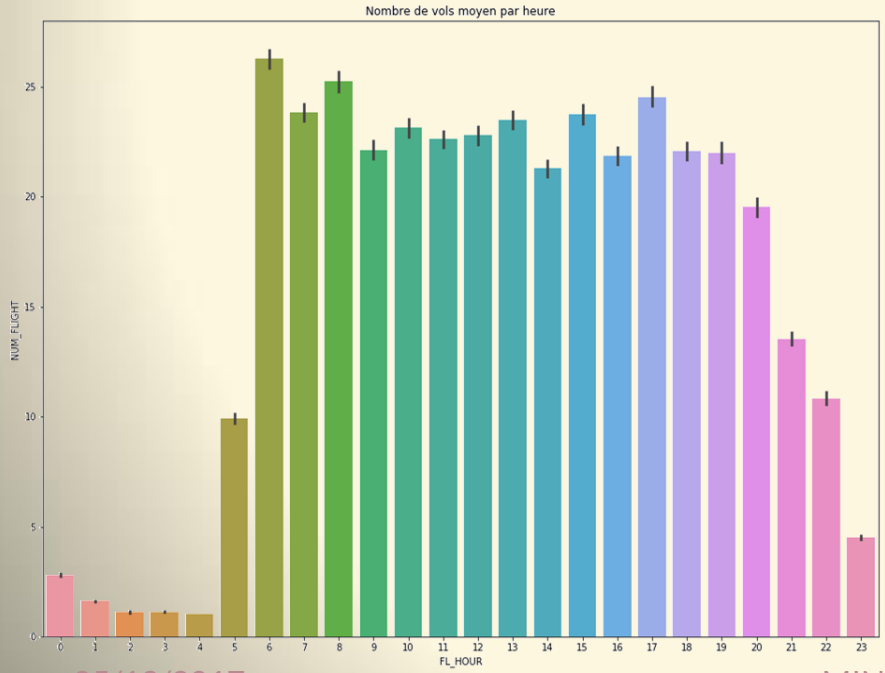
Exploration

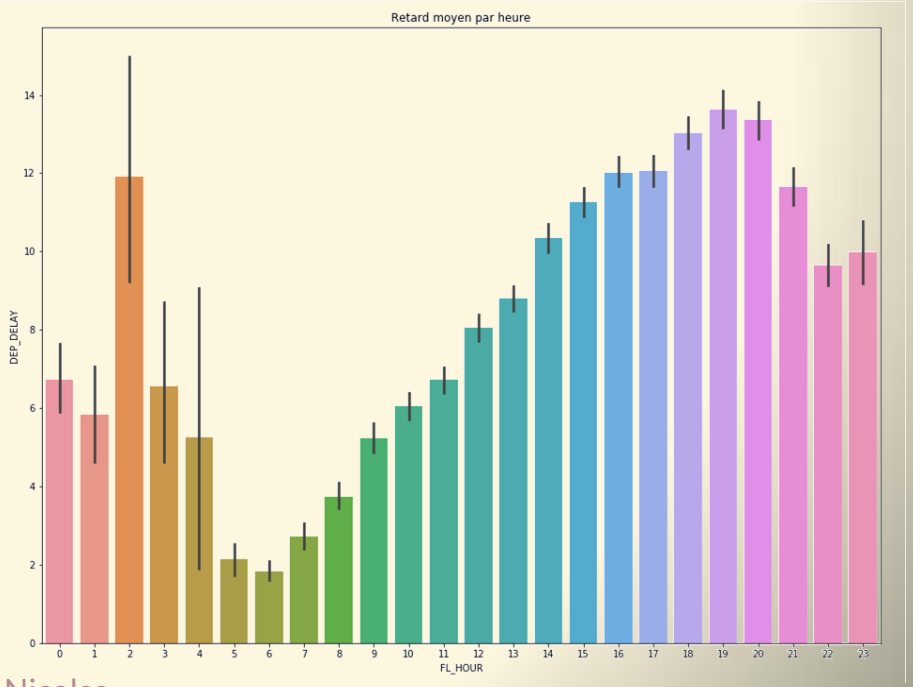
- Retard par Date (Modèle 1)
  - Par 15 minutes
    - Linéarisation
      - X = abs(X-15)
    - Assez stable
    - Pas encodé



Exploration

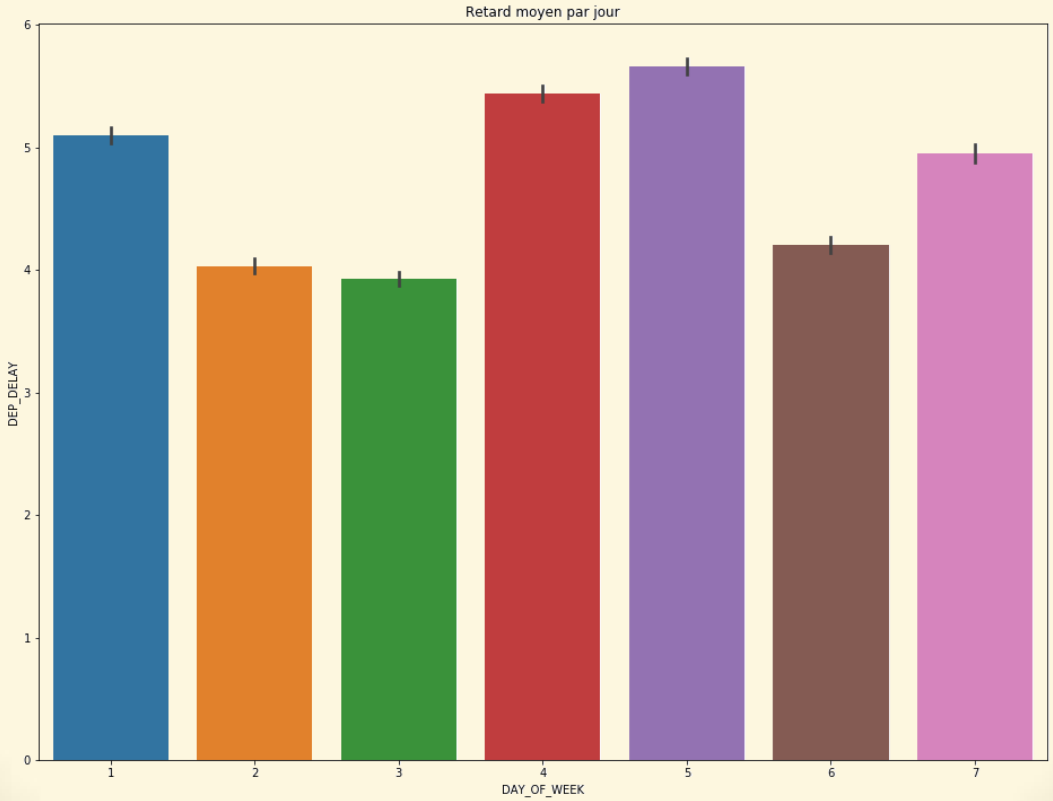
- Retard par Date (Modèle 2)
  - Par heure
    - Pas de linéarisation
    - Pas encodé





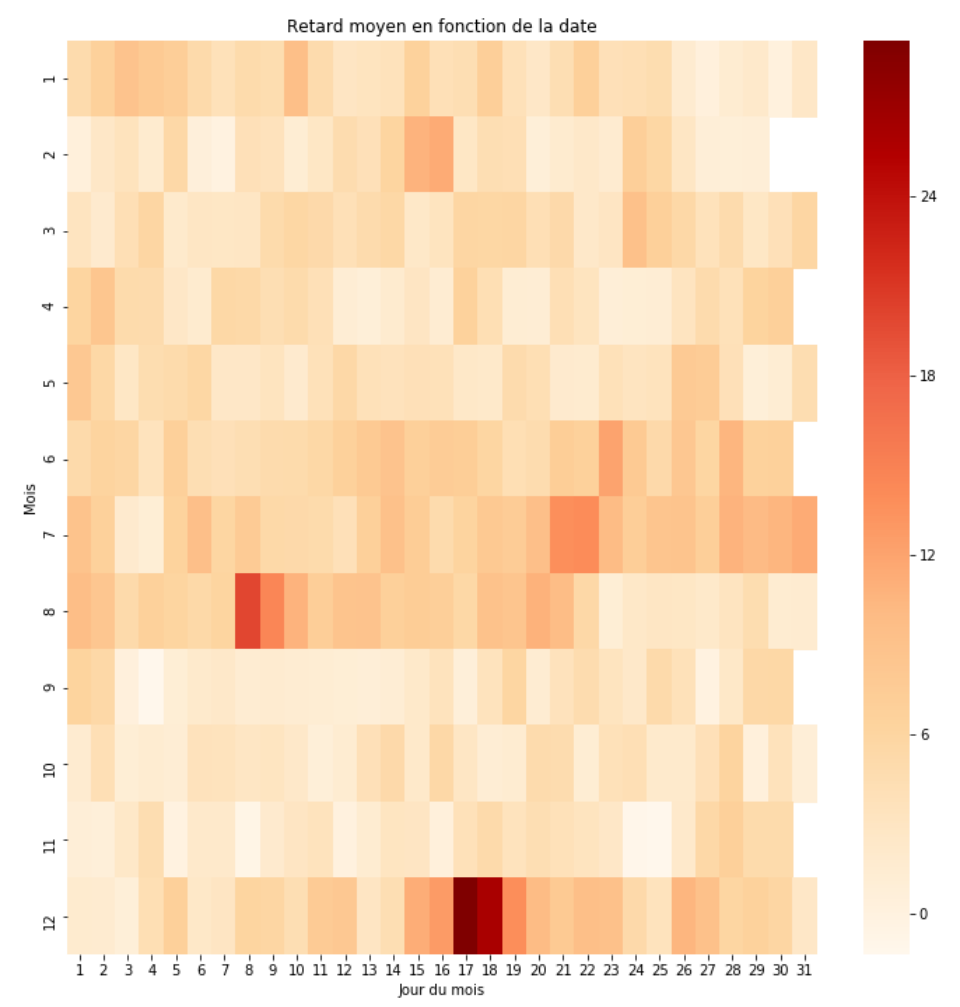
Exploration

- Retard par Date
  - Jour de la semaine



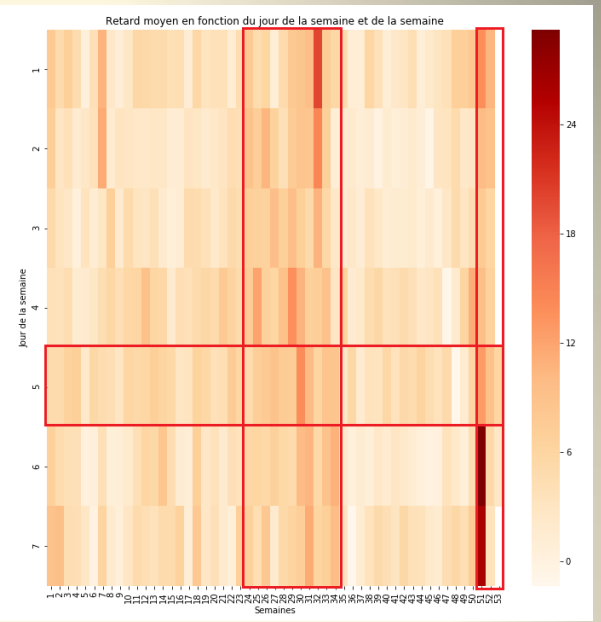
Exploration

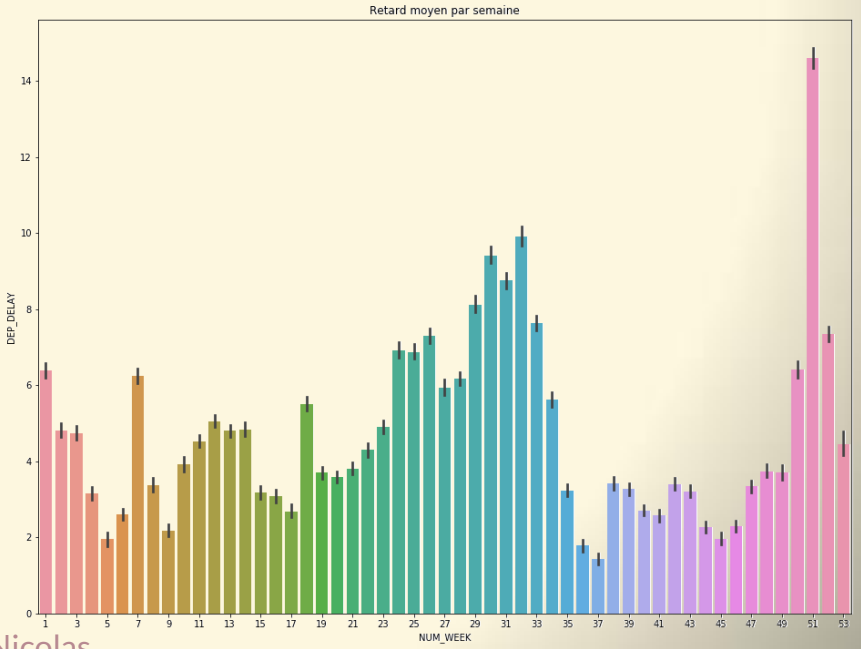
- Retard par Date
  - Jour et Mois de l'année
    - Très irrégulier
    - Pic Juin/Juillet/Aout
    - Pic vers noel
  - Possibilité OHE
    - 31 + 12 dimensions
    - Solution 1



Exploration

- Retard par Date
  - Par semaine / jour de la semaine
    - Irrégulier mais peu variable
    - Pic Juin/Juillet/Aout
    - Pic vers noël
  - Possibilité OHE
    - 53 dimensions
    - Solution 2

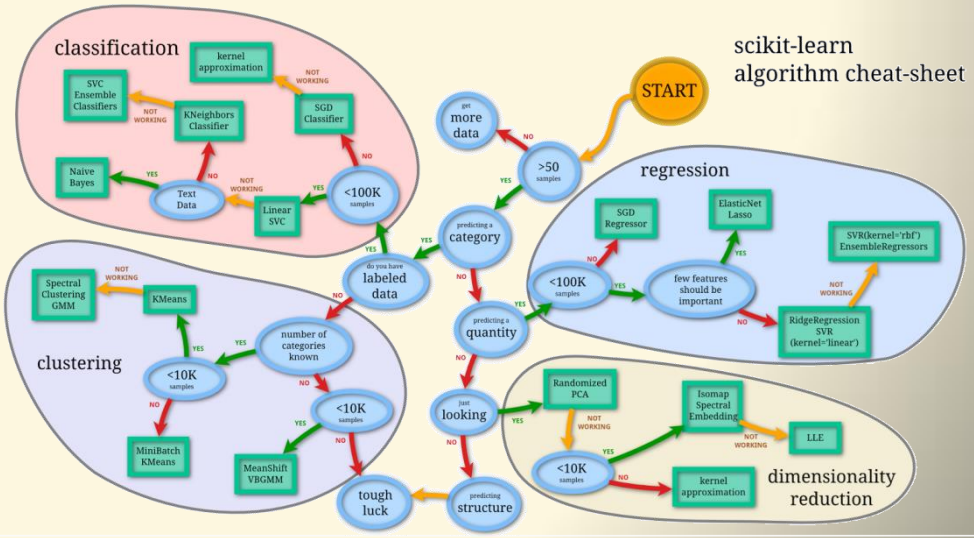




Modélisation

- Choix des modèles
  - Std. Regression Lineaire
  - Batch GD Regressor
  - ✓ Stochastic GD Regressor
  - ✓ Boosting (avec SGD Reg.)
  - Ensembles Learning
  - ✓ Simple ANN
  - KNN Regression
  - SVM Regression

- Choix des métriques
  - ✓ MAE
  - ✓ (R)MSE
  - ❖ RMSLE



Modélisation

Modèle 1

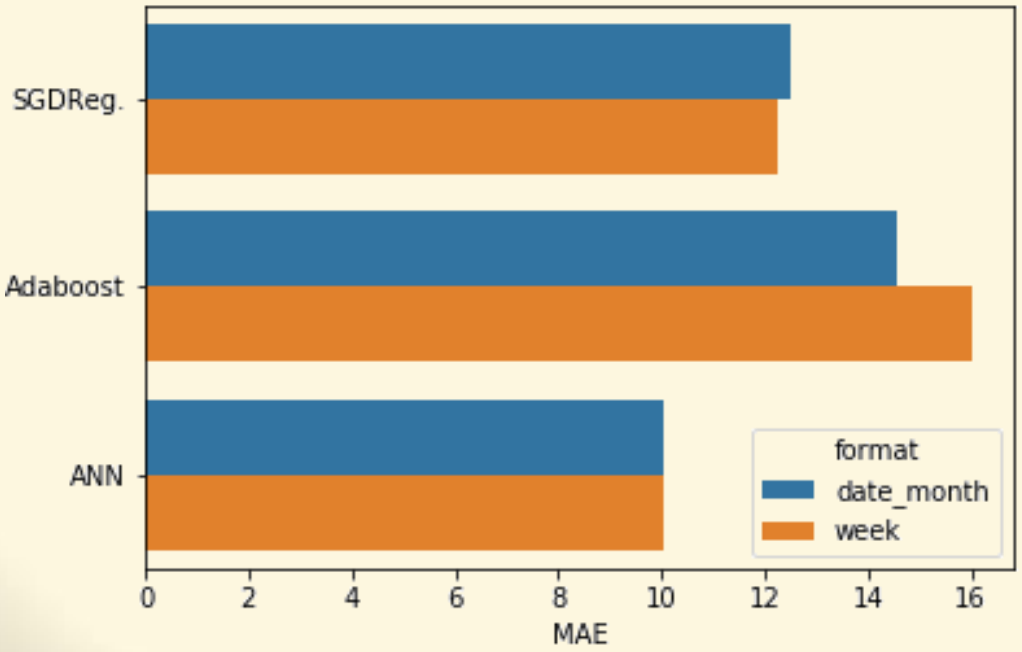
Modélisation

Modèle Jour / mois

- 5,25 millions de lignes
- 55 features (dont 51 OHE)
  - sparse matrices

Modèle Semaines

- 5,25 millions de lignes
- 65 features (dont 61 OHE)
  - sparse matrices



Modélisation

- Modèle choisi – Semaines
  - Plus logique
  - Meilleur ANN/SGD Regressor (Doutes sur le Boosting)
- SGD Regressor & Boosting
  - Optimisation au Grid Search
  - Test avec ou sans OHE Groupe Aéroport

SGDR :

"loss" :	huber
"max_iter":	3, 5, 10
"penalty" :	Aucune, l2, l1, elasticnet
"l1_ratio" :	0.15, 0.50, 0.85

Boosting :

"base_estimator" :	SGDRegressor(best_params_SGDR)
"n_estimators" :	2, 5, 10, 20
"loss" :	linear, square, exponential

Modélisation

- Résultats :
  - AdaBoost
    - 2 estimateurs
    - Loss : exponential
  - SGD
    - L1_ratio : 0,15
    - Iter : 10
    - Penalty : Aucune

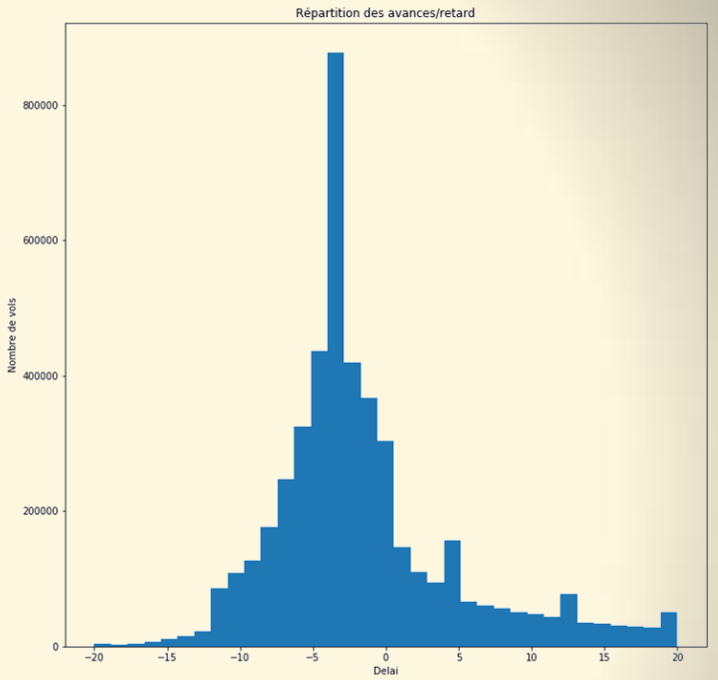
```
##################
Week Format False :
##################

14/11/17 - 18:55:03 => SGD_Regressor ({'loss': 'huber', 'max_iter': 3, 'penalty': None}) : MSE 851.9428, MAE 10.2565
14/11/17 - 18:55:09 => SGD_Regressor ({'loss': 'huber', 'max_iter': 3, 'penalty': 'l2'}) : MSE 852.6893, MAE 10.2602
14/11/17 - 18:55:17 => SGD_Regressor ({'loss': 'huber', 'max_iter': 3, 'penalty': 'l1'}) : MSE 852.6906, MAE 10.2627
14/11/17 - 18:55:25 => SGD_Regressor ({'loss': 'huber', 'max_iter': 5, 'penalty': None}) : MSE 851.4224, MAE 10.2513
14/11/17 - 18:55:34 => SGD_Regressor ({'loss': 'huber', 'max_iter': 5, 'penalty': 'l2'}) : MSE 851.6895, MAE 10.2548
14/11/17 - 18:55:46 => SGD_Regressor ({'loss': 'huber', 'max_iter': 5, 'penalty': 'l1'}) : MSE 851.7479, MAE 10.2568
14/11/17 - 18:56:02 => SGD_Regressor ({'loss': 'huber', 'max_iter': 10, 'penalty': None}) : MSE 850.7918, MAE 10.2490
14/11/17 - 18:56:19 => SGD_Regressor ({'loss': 'huber', 'max_iter': 10, 'penalty': 'l2'}) : MSE 851.5218, MAE 10.2519
14/11/17 - 18:56:41 => SGD_Regressor ({'loss': 'huber', 'max_iter': 10, 'penalty': 'l1'}) : MSE 851.5194, MAE 10.2534
14/11/17 - 18:56:50 => SGD_Regressor ({'l1_ratio': 0.15, 'loss': 'huber', 'max_iter': 3, 'penalty': 'elasticnet'}) : MSE 852.4666, MAE 10.2604
14/11/17 - 18:57:02 => SGD_Regressor ({'l1_ratio': 0.15, 'loss': 'huber', 'max_iter': 5, 'penalty': 'elasticnet'}) : MSE 851.8880, MAE 10.2549
14/11/17 - 18:57:26 => SGD_Regressor ({'l1_ratio': 0.15, 'loss': 'huber', 'max_iter': 10, 'penalty': 'elasticnet'}) : MSE 851.3977, MAE 10.2519
14/11/17 - 18:57:33 => SGD_Regressor ({'l1_ratio': 0.5, 'loss': 'huber', 'max_iter': 3, 'penalty': 'elasticnet'}) : MSE 852.4652, MAE 10.2612
14/11/17 - 18:57:46 => SGD_Regressor ({'l1_ratio': 0.5, 'loss': 'huber', 'max_iter': 5, 'penalty': 'elasticnet'}) : MSE 851.7899, MAE 10.2554
14/11/17 - 18:58:10 => SGD_Regressor ({'l1_ratio': 0.5, 'loss': 'huber', 'max_iter': 10, 'penalty': 'elasticnet'}) : MSE 851.6556, MAE 10.2523
14/11/17 - 18:58:18 => SGD_Regressor ({'l1_ratio': 0.85, 'loss': 'huber', 'max_iter': 3, 'penalty': 'elasticnet'}) : MSE 852.4764, MAE 10.2621
14/11/17 - 18:58:30 => SGD_Regressor ({'l1_ratio': 0.85, 'loss': 'huber', 'max_iter': 5, 'penalty': 'elasticnet'}) : MSE 851.6836, MAE 10.2564
14/11/17 - 18:58:53 => SGD_Regressor ({'l1_ratio': 0.85, 'loss': 'huber', 'max_iter': 10, 'penalty': 'elasticnet'}) : MSE 851.3929, MAE 10.2529

14/11/17 - 19:03:40 => AdaBoost ({'base_estimator': SGDRegressor(alpha=0.0001, average=False, epsilon=0.1, eta0=0.01,
       fit_intercept=True, l1_ratio=0.15, learning_rate='invscaling',
       loss='huber', max_iter=10, n_iter=None, penalty=None, power_t=0.25,
       random_state=None, shuffle=True, tol=None, verbose=0,
       warm_start=False), 'loss': 'linear', 'n_estimators': 2}) : MSE 850.7297, MAE 10.2490
14/11/17 - 19:05:37 => AdaBoost ({'base_estimator': SGDRegressor(alpha=0.0001, average=False, epsilon=0.1, eta0=0.01,
       fit_intercept=True, l1_ratio=0.15, learning_rate='invscaling',
       loss='huber', max_iter=10, n_iter=None, penalty=None, power_t=0.25,
       random_state=None, shuffle=True, tol=None, verbose=0,
       warm_start=False), 'loss': 'linear', 'n_estimators': 5}) : MSE 848.1876, MAE 10.2505
14/11/17 - 19:09:22 => AdaBoost ({'base_estimator': SGDRegressor(alpha=0.0001, average=False, epsilon=0.1, eta0=0.01,
       fit_intercept=True, l1_ratio=0.15, learning_rate='invscaling',
       loss='huber', max_iter=10, n_iter=None, penalty=None, power_t=0.25,
       random_state=None, shuffle=True, tol=None, verbose=0,
       warm_start=False), 'loss': 'linear', 'n_estimators': 10}) : MSE 848.1641, MAE 10.2505
14/11/17 - 19:13:26 => AdaBoost ({'base_estimator': SGDRegressor(alpha=0.0001, average=False, epsilon=0.1, eta0=0.01,
       fit_intercept=True, l1_ratio=0.15, learning_rate='invscaling',
       loss='huber', max_iter=10, n_iter=None, penalty=None, power_t=0.25,
```

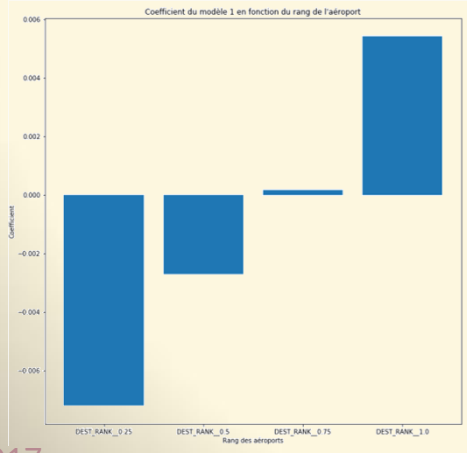
- MAE : 10.2008 (9.9598)
- MSE : 875.1533 (845.0382)
- RMSE : 29.5830 (29.06)

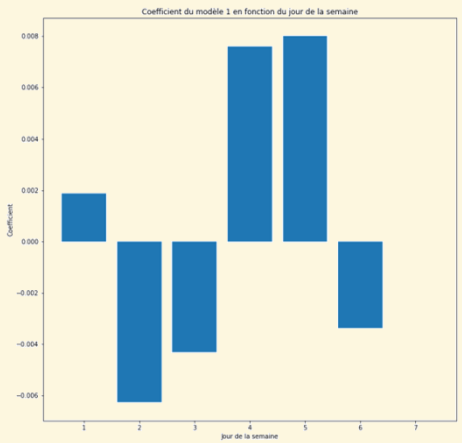
Modèle linéaire similaire
au
Modèle non linéaire

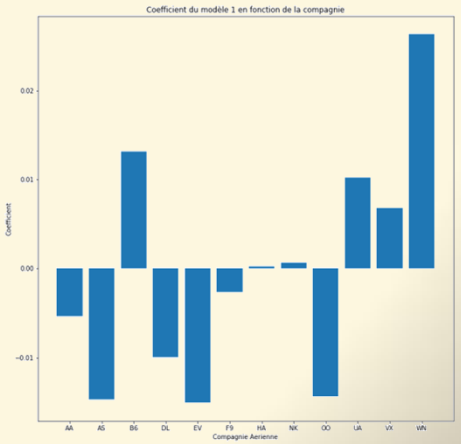
Modélisation

- Problème :
  - Prédiction négatives (-8 à -3 min)
- Solution
  - Retard = max(retard, 0)
- Nouveau problème
  - Prédiction proche de 0 (coefs ~ 0)









Modélisation

Modèle 2

Modélisation

- Agrégation
  - Date & Heure , Aéroport départ, Compagnie
  - ➢ Moyenne retard
  - ➢ Ajout du nombre de vols

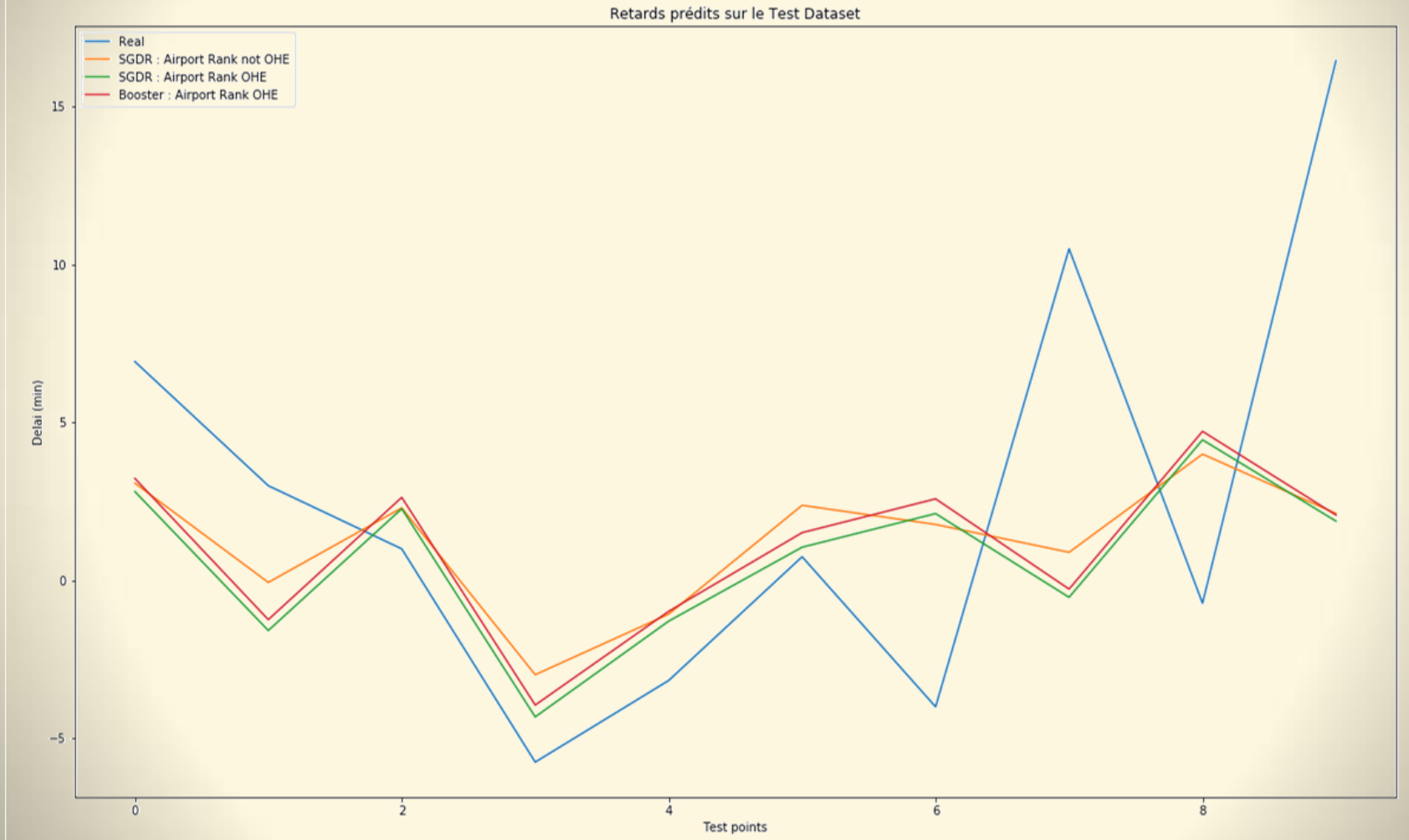
- Optimisation Grid Search
- Multiple modèles:
  - SGDR sans Rang aéroport non OHE
  - SGDR avec Rang aéroport OHE
  - Adaboost avec Rang aéroport OHE

- SGDR avec rang non OHE :
  MSE 652.6672 - MAE 11.3115

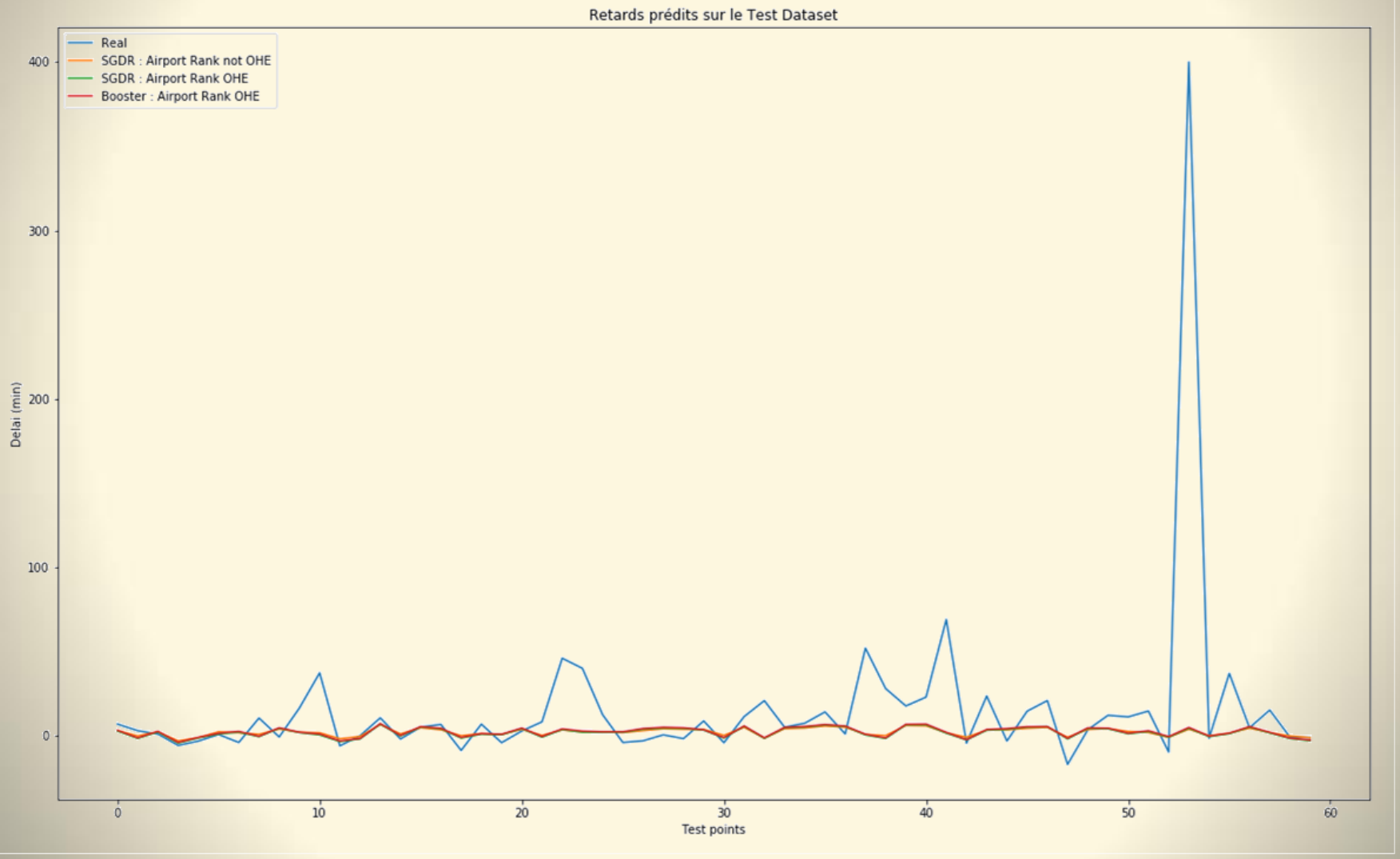
- SGDR avec rang OHE :
  MSE 652.2052 - MAE 11.2355

✓ Boosting avec rang OHE
  MSE 648.2626 - MAE 11.2342

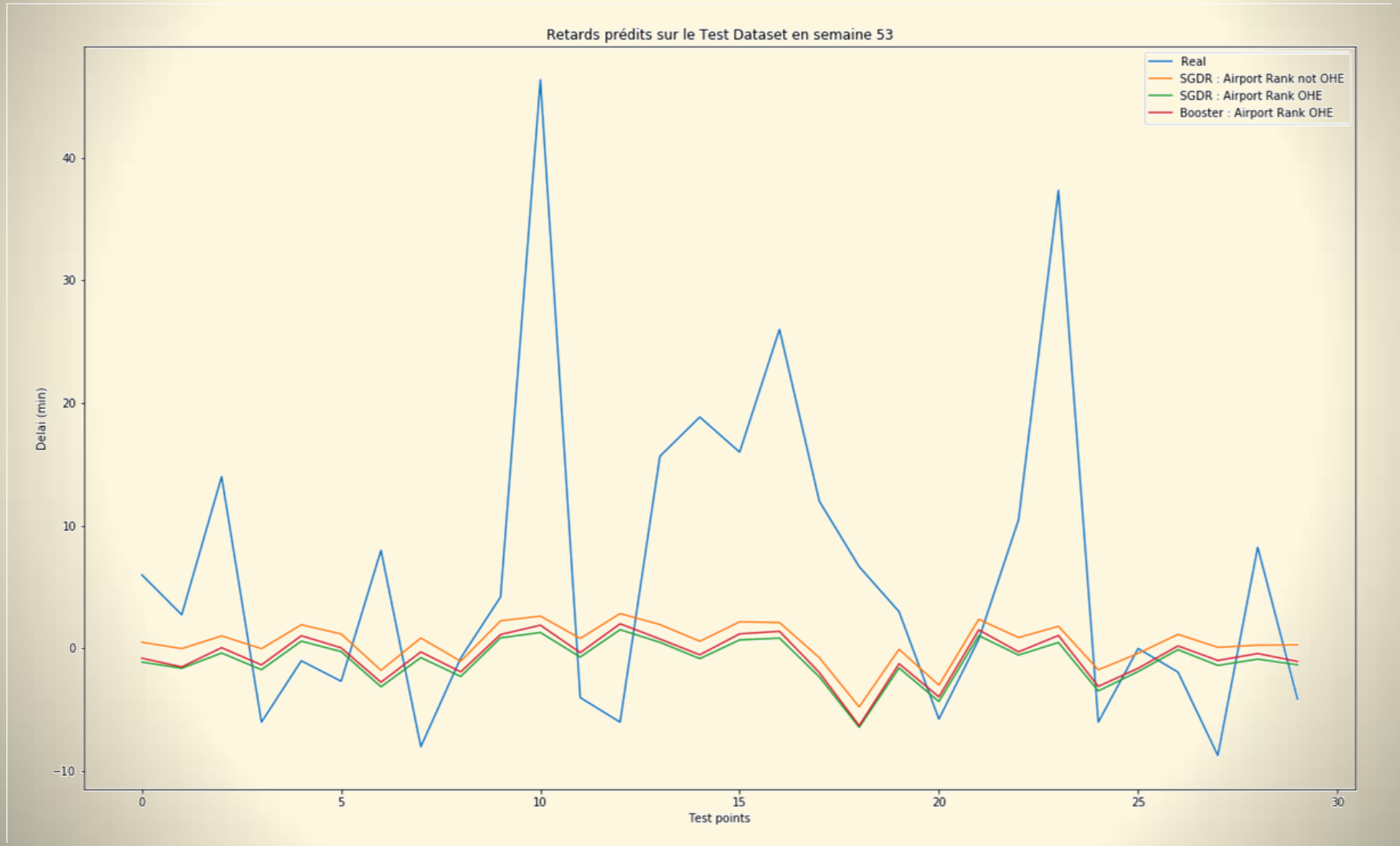
Modélisation



Modélisation

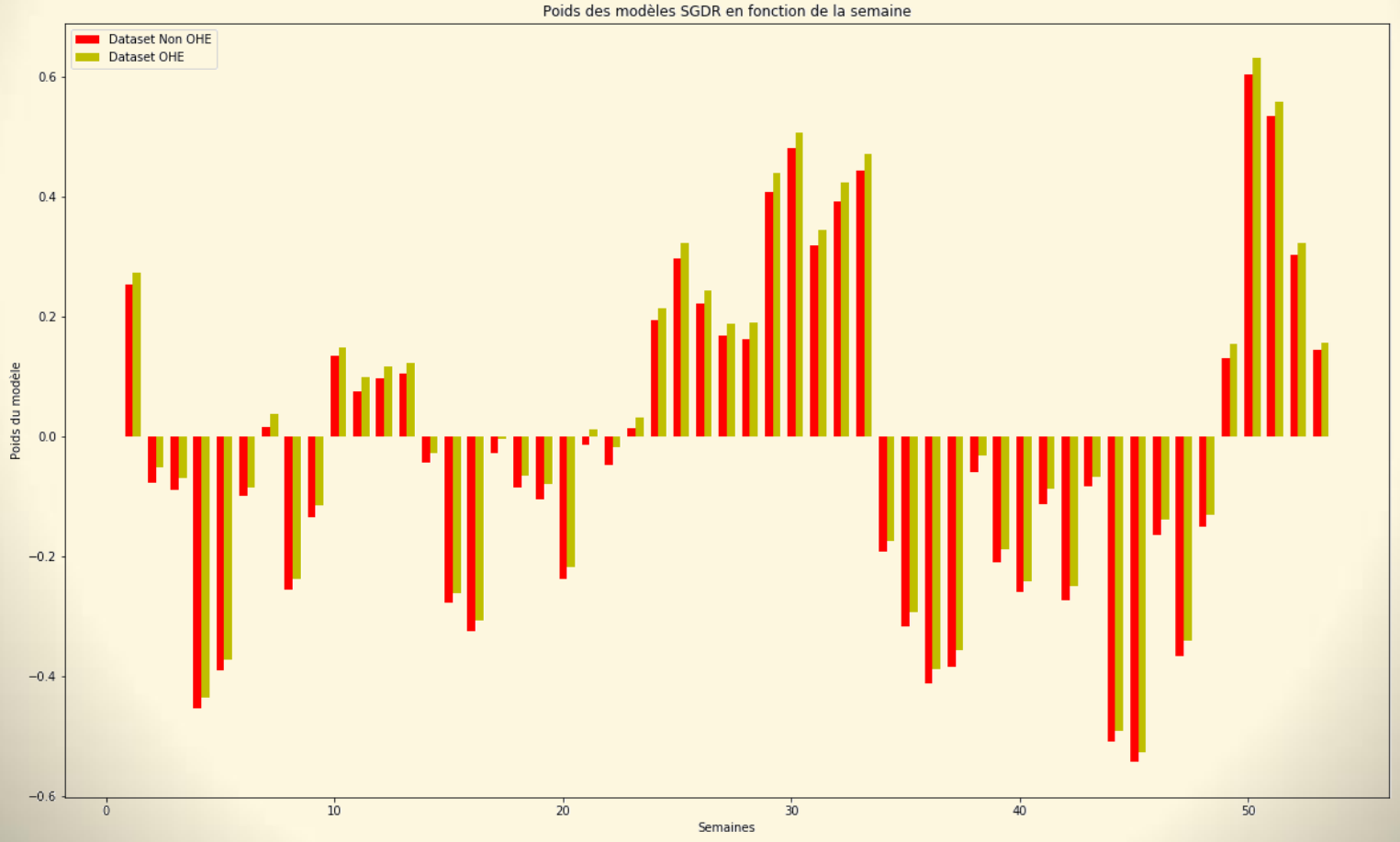


Modélisation



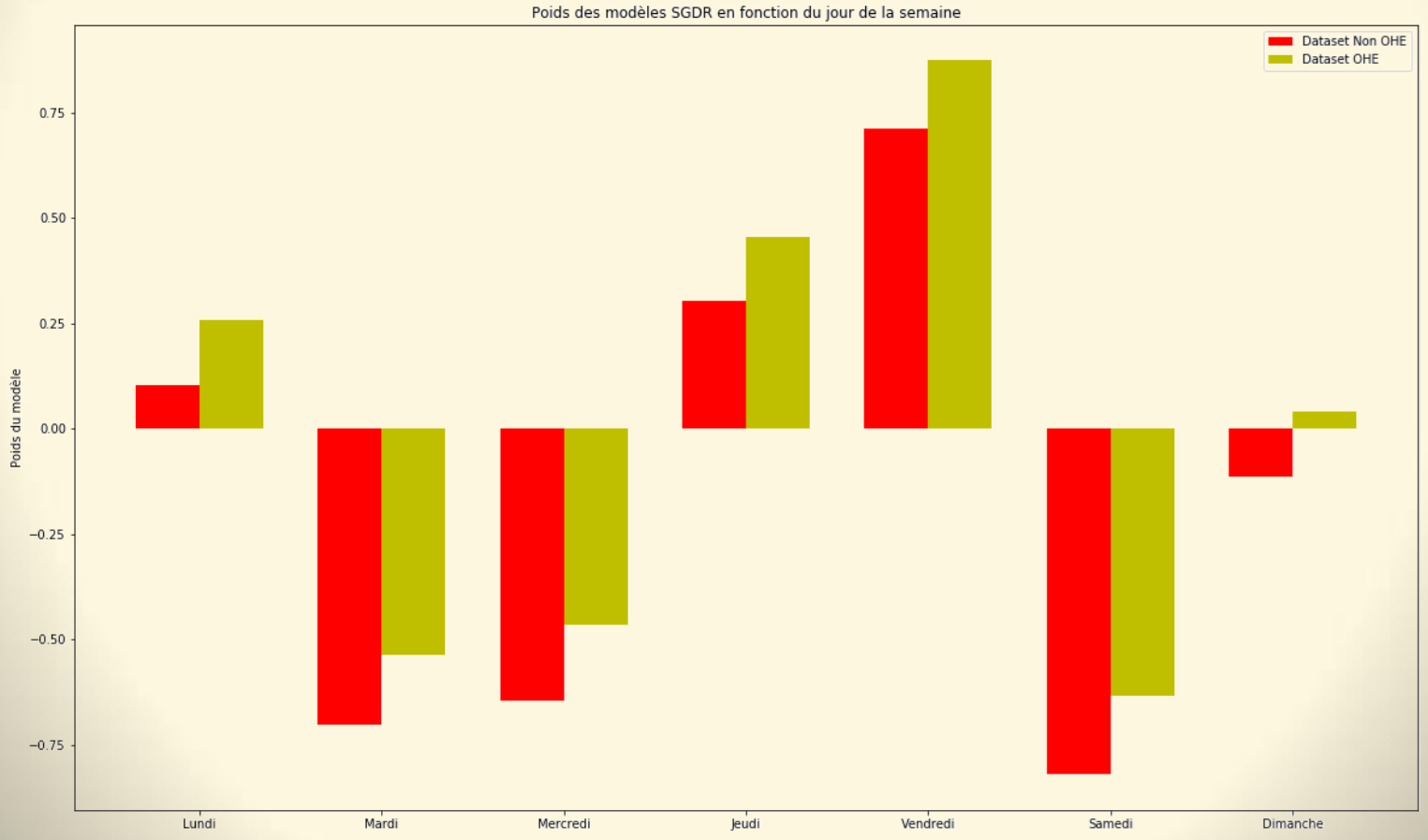
Interprétation

- SGDR – OHE vs non OHE



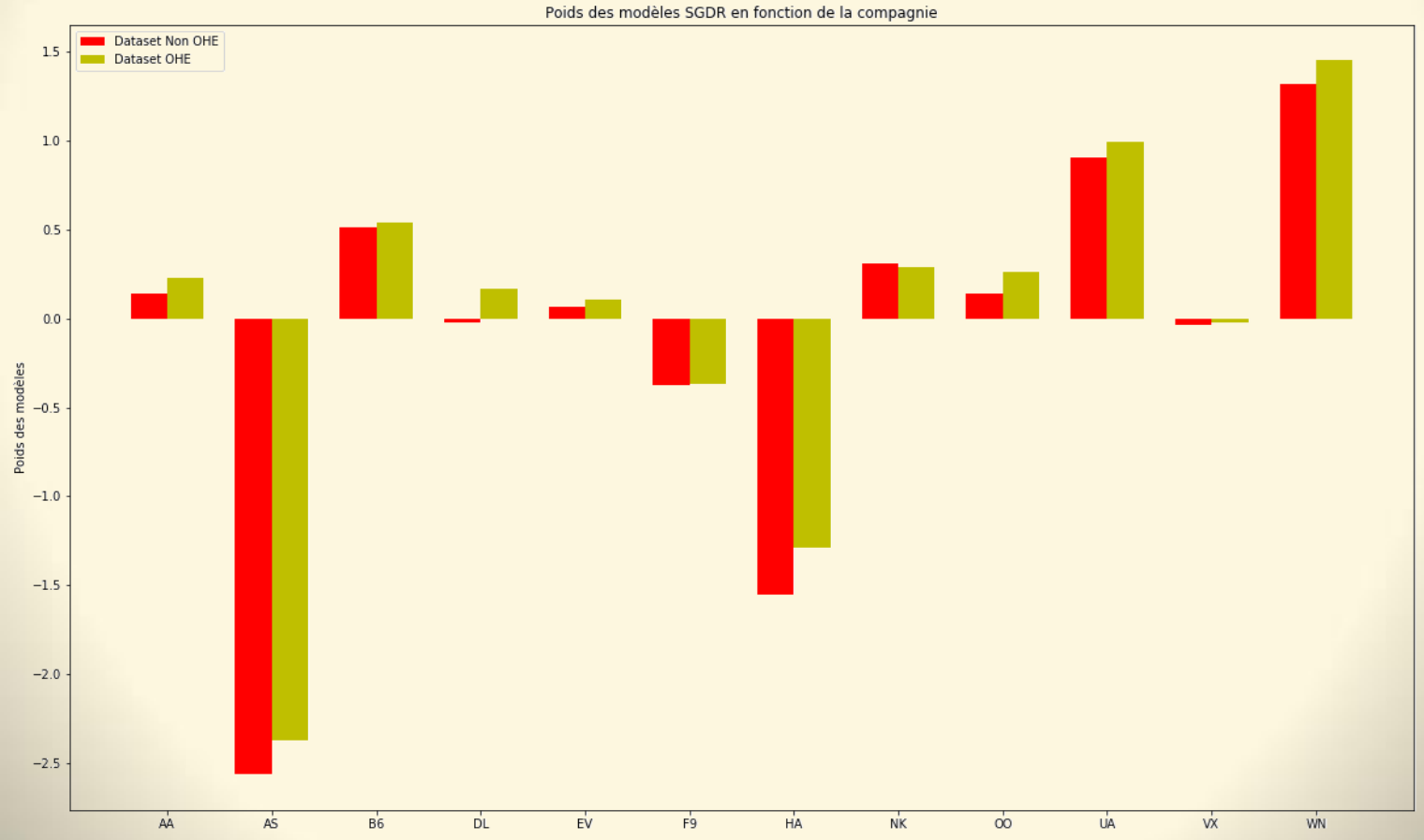
Interprétation

- SGDR – OHE vs non OHE



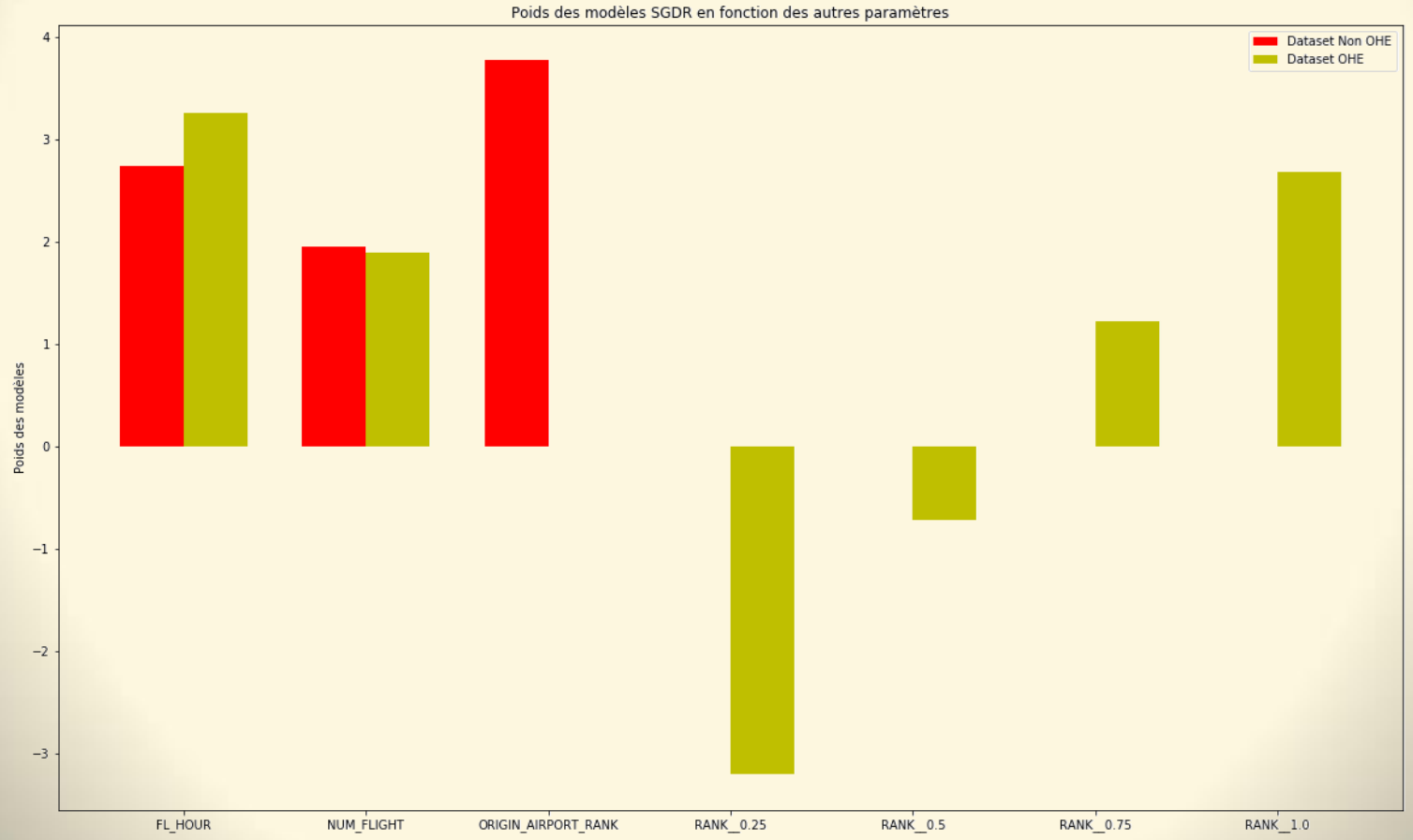
Interprétation

- SGDR – OHE vs non OHE



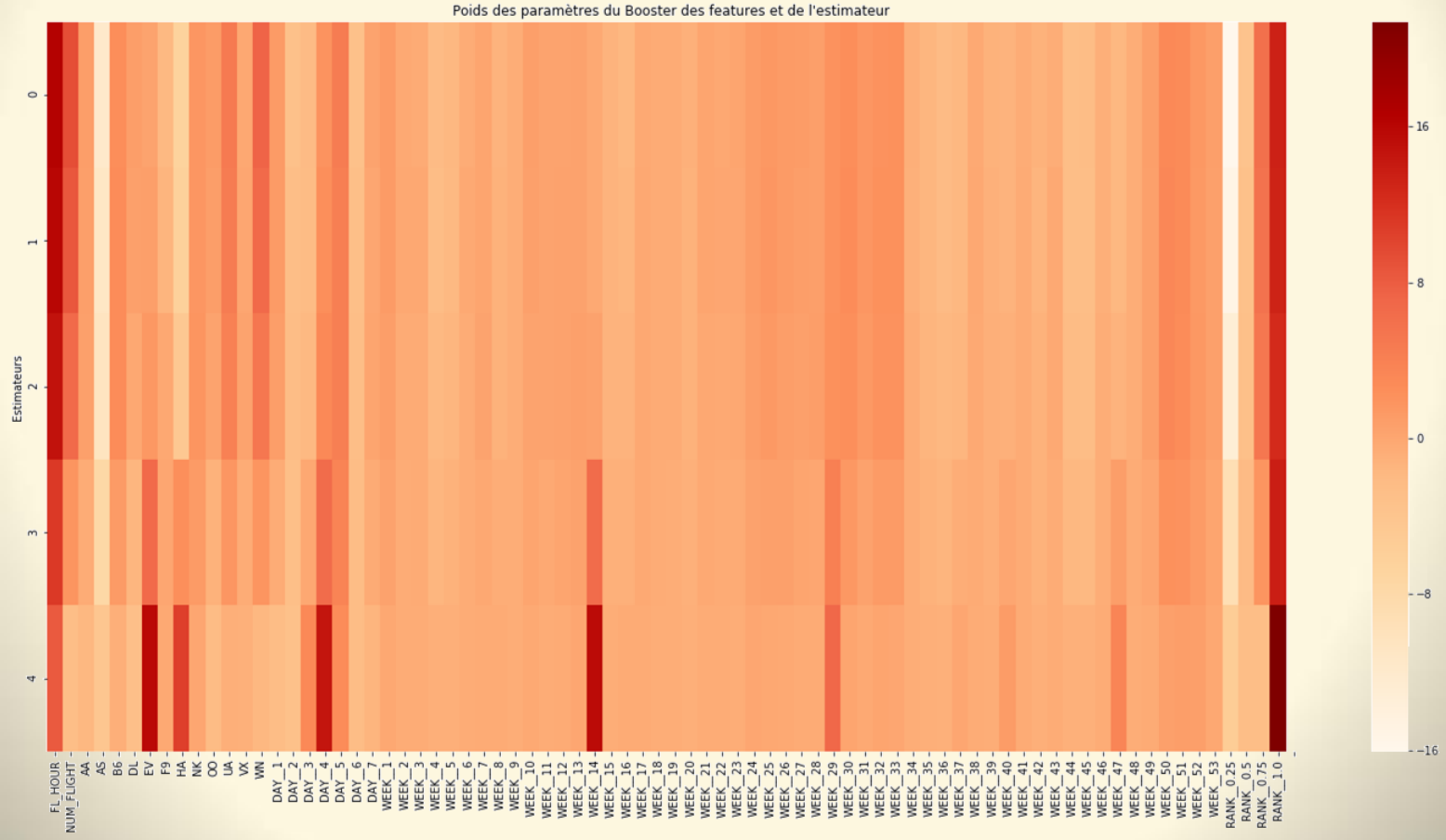
Interprétation

- SGDR – OHE vs non OHE



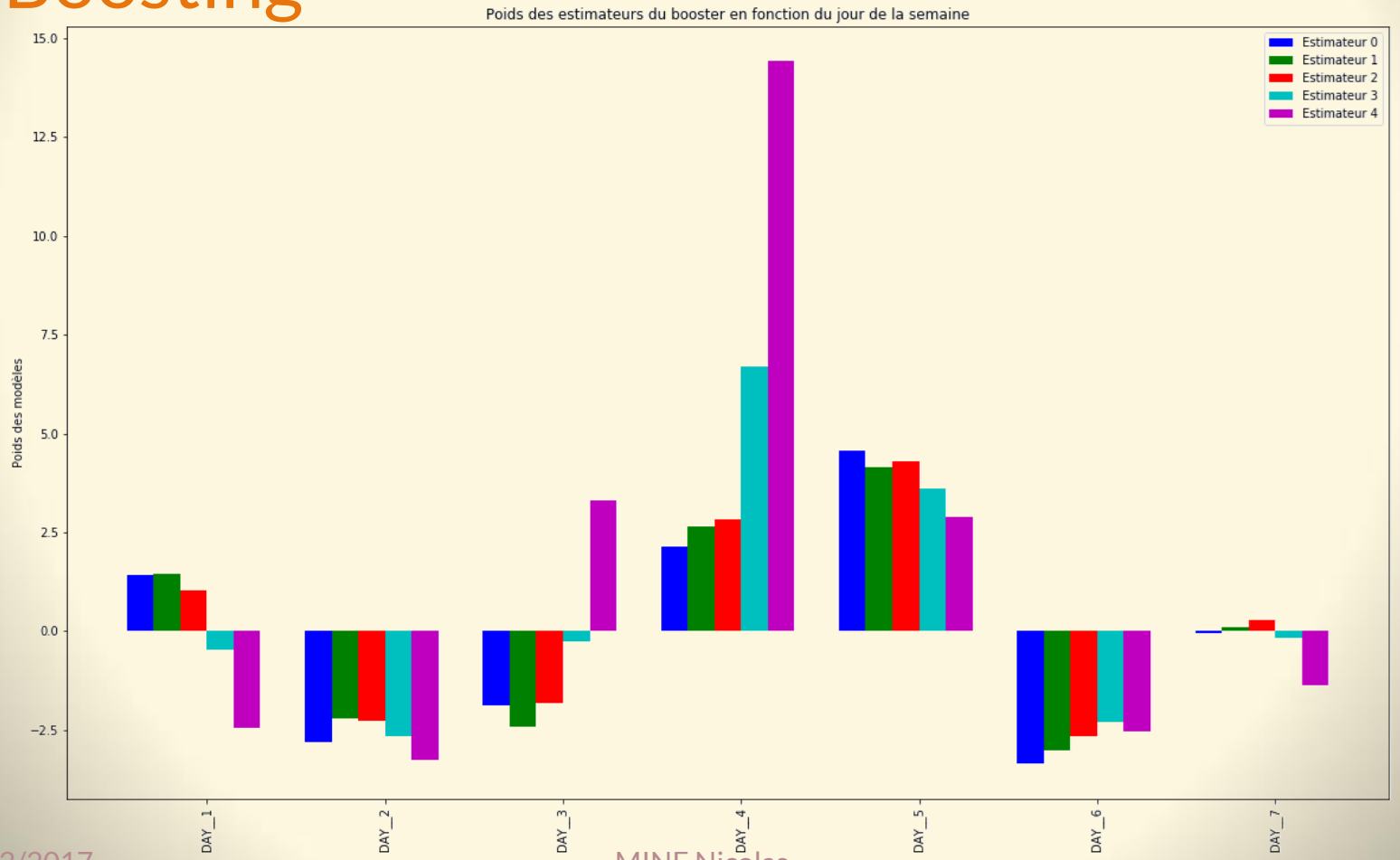
Interprétation

- Boosting



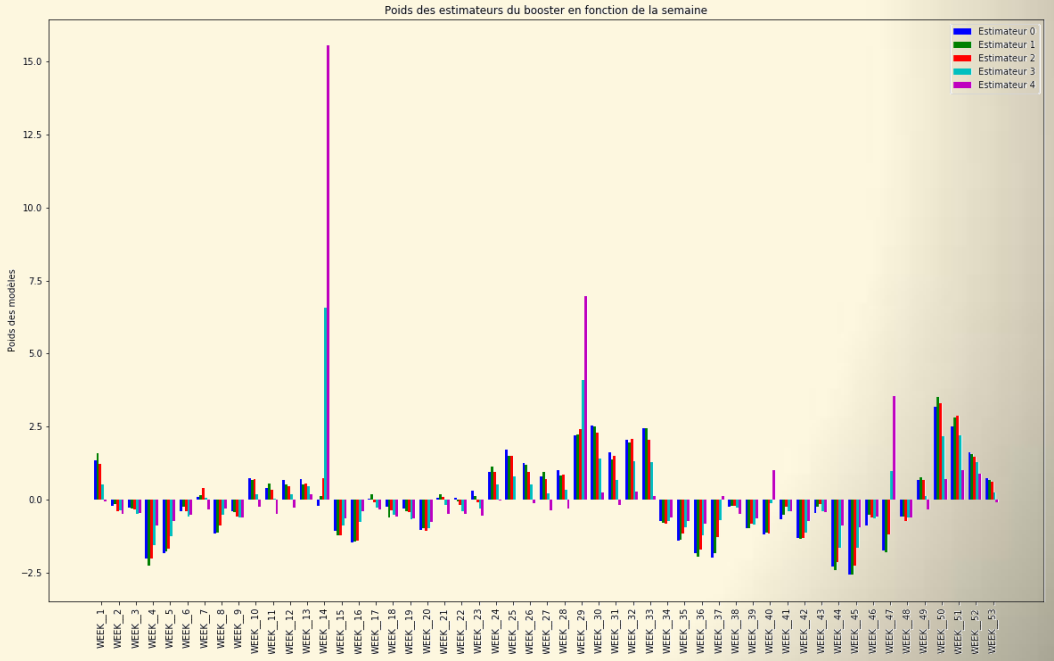
Interprétation

- Boosting



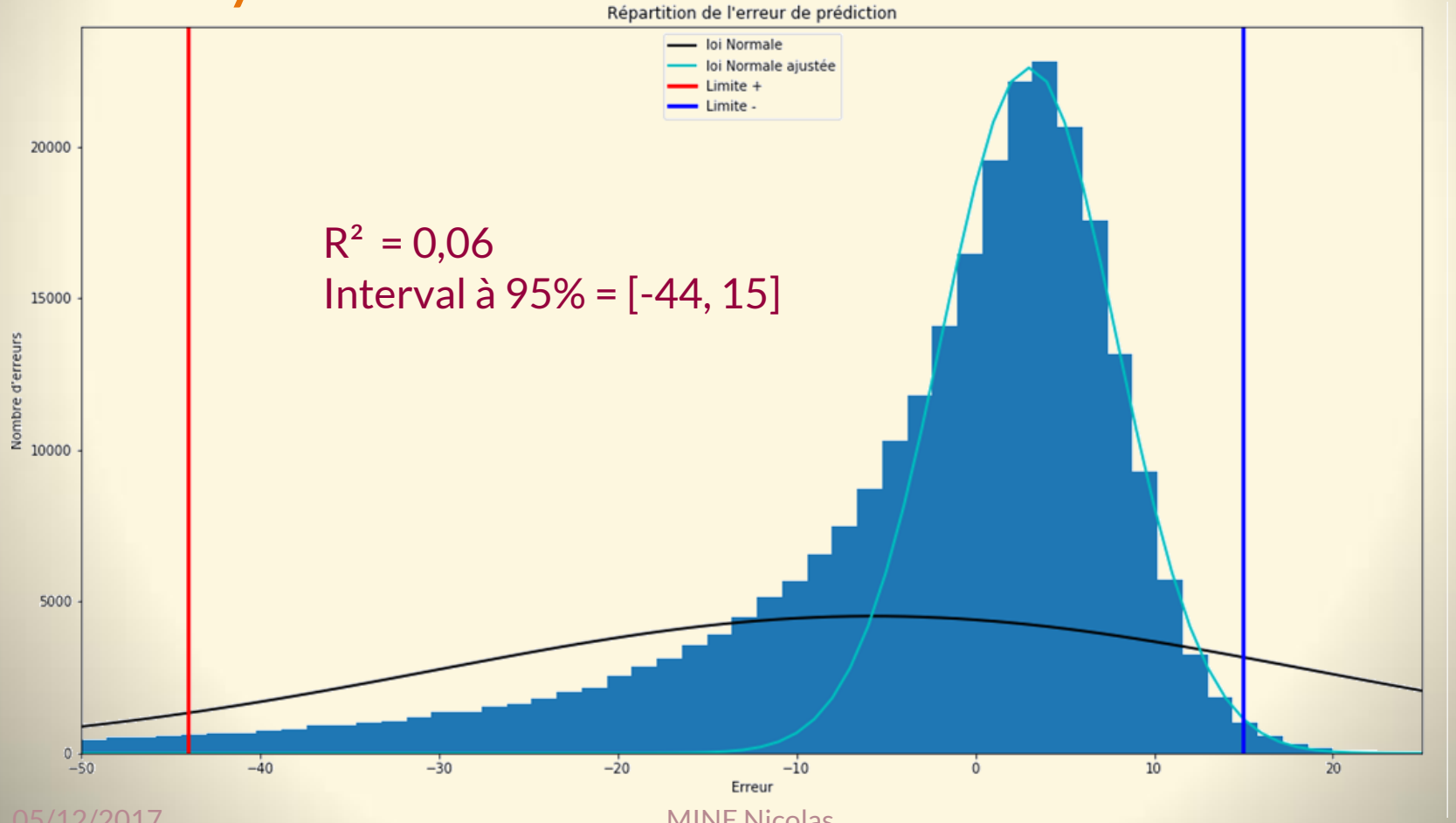
Interprétation

- Prédictions similaires
- Coefficients similaires
  - Offset OHE
  - Bruit Boosting
  - Même tendances



Interprétation

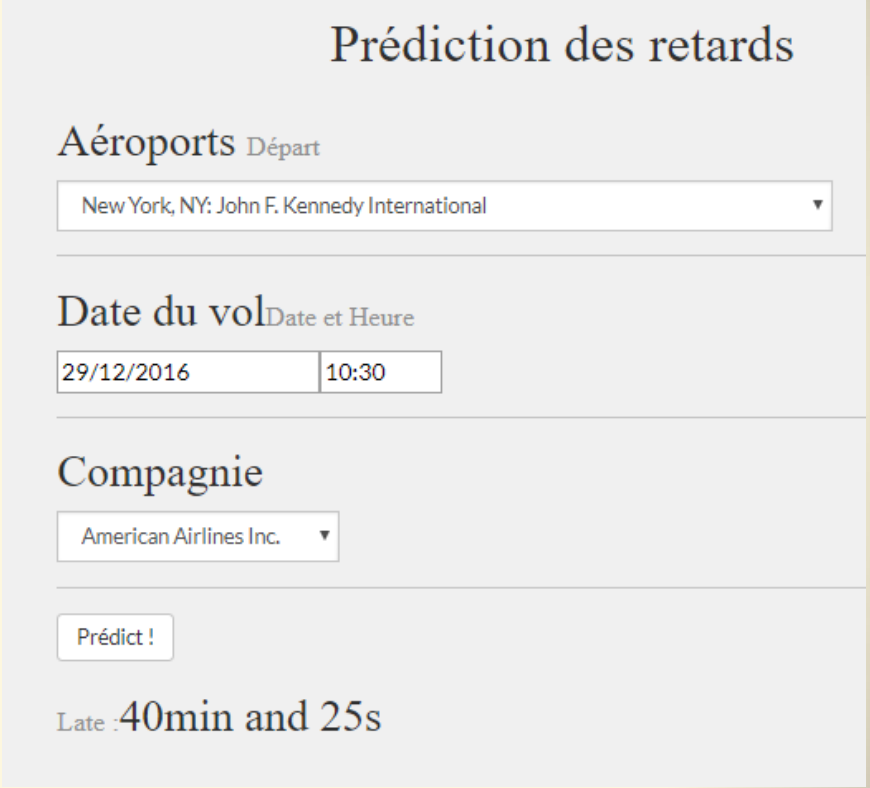
- Analyse des résidus



R^2 = 0,06
Interval à 95% = [-44, 15]

API

- Site Flask
- UI
  - Aéroport de départ
  - Date et Heure
  - Compagnie
- Server
  - Encode/Scale
  - Prédiction
  - Retourne l'info (POST)



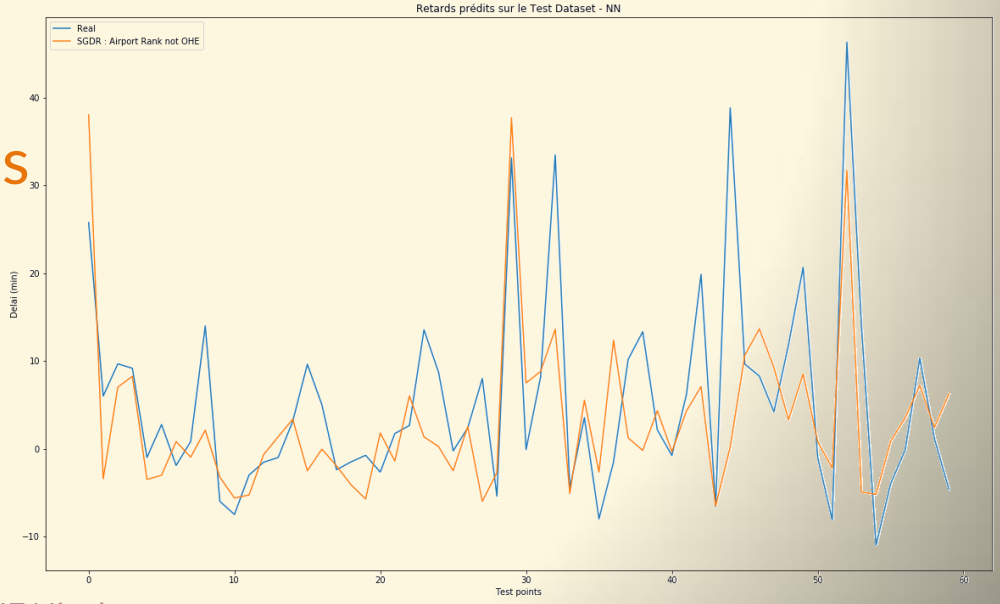
http://coni57.pythonanywhere.com/p4/

Ouverture à l'amélioration

- Modèle non linéaire
  - Actuellement linéaire:

$$Retard = \alpha * heure + \beta_{jour} + \gamma_{semaine} + \delta_{compagnie} + \varepsilon_{aeroport} + \theta * nb_{vols}$$

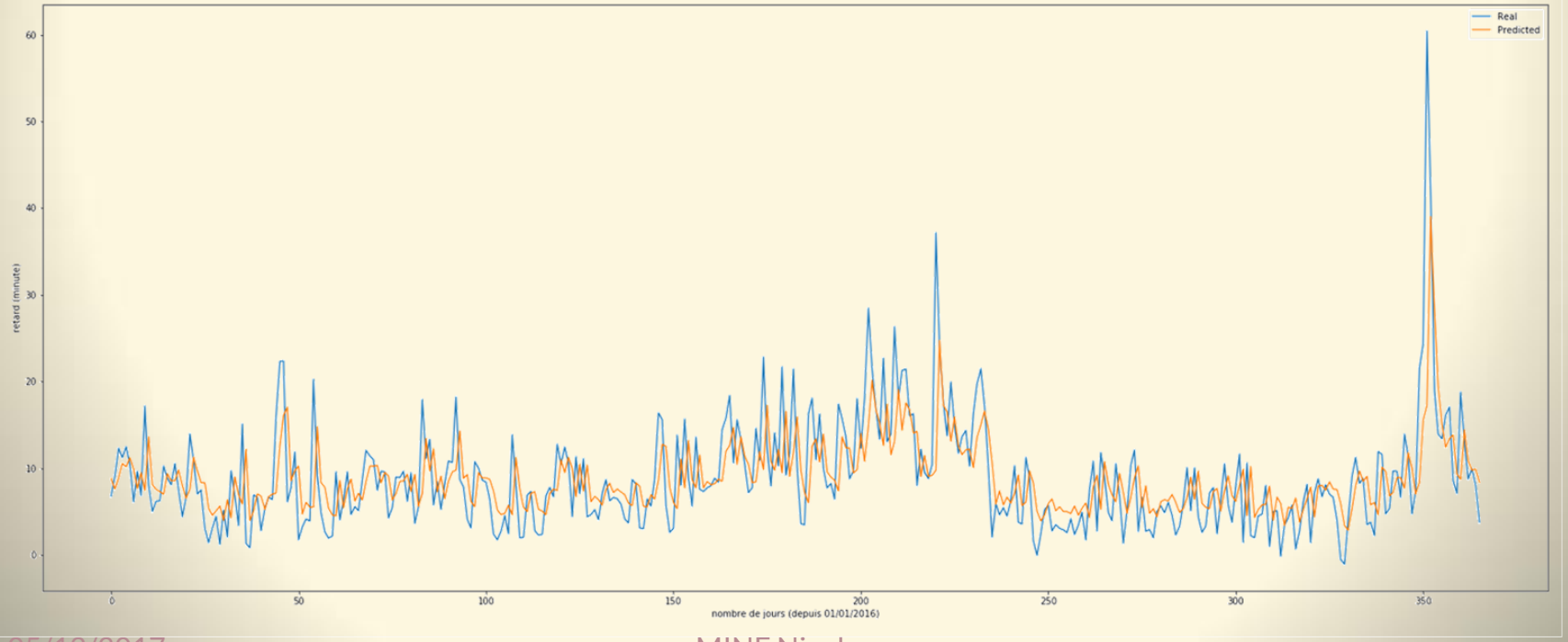
  - Possible :
    - Features interactions
    - Dimensions ?
    - ➢ ANN
      - MAE: 8.2088
      - MSE: 467.2681



Modélisation

- ARIMA (Hors Sujet)
  - Création d'un dataset
    - Date / retard moyen (365 lignes)
  - Recherche de paramètre (p, q, r) par grid search (MAE/MSE)

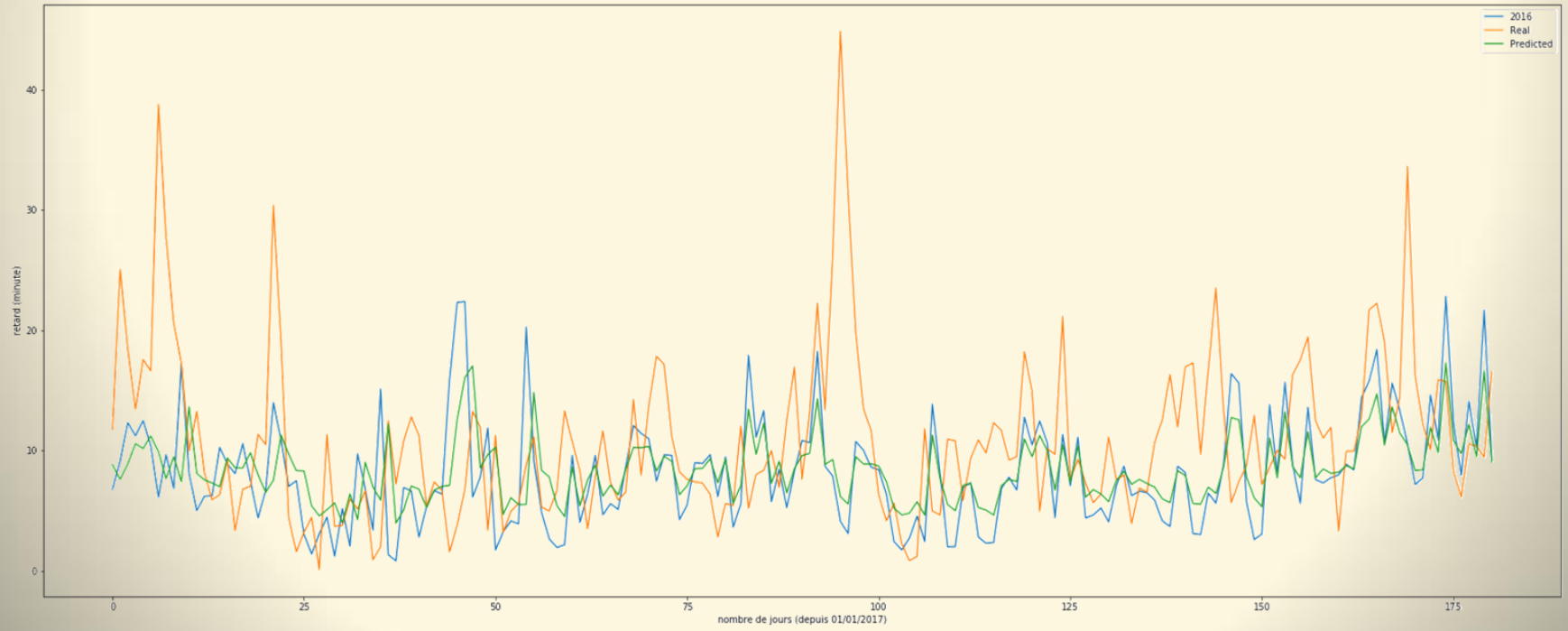
MAE : 5.66
MSE : 65.99



Modélisation

- ARIMA (Hors Sujet)
  - Validation sur 6 mois 2017
  - Prédiction max sur 7j
    - Réutilisation des données 2016

MAE : 4.96
MSE : 52.62



Modélisation

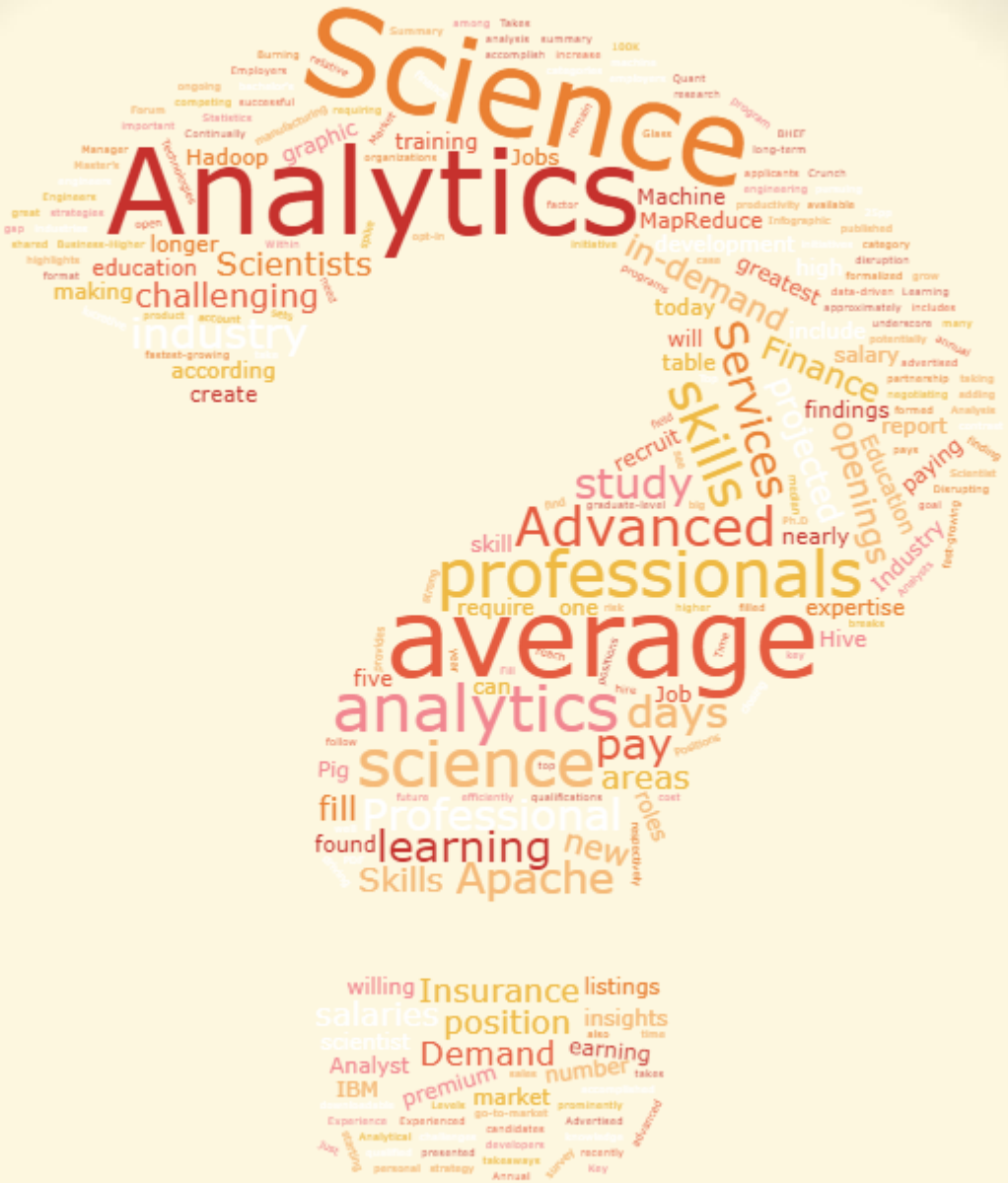
- RNN
  - Type LSTM/GRU
  - Entrée :
    - Nb vol
    - compagnie,
    - aéroport
    - date/heure
    - retard T-1
  - Sortie : retard T
  - Validation sur 6 mois 2017

  - Hors sujet (il faut 1 an minimum)

Conclusion

- Beaucoup de données
- Dataset propre
- Tendances visibles
- Peu de modèles possible
- Problème mémoire
- Underfitting du modèle linéaire
- MAE ~ 10 min (dataset 8,4 +/- 25 min)

Peu d'avantages pour les petits retards


Science
Analytics
training
Hadoop
graphic
Jobs
Machine
MapReduce
longer
education
Scientists
making
challenging
industry
according
create
development
in-demand
greatest
high
today
include
Finance
Services
will
table
salary
skills
findings
report
recruit
study
Advanced
skill
nearly
professionals
require
one
expertise
Hive
average
five
can
Job
analytics
days
pay
Pig
science
areas
fill
Professional
found
learning
new
Skills
Apache
willing
Insurance
listings
salaries
position
insights
scientist
Demand
earning
Analyst
number
IBM
premium
market