Prédiction du retard de vol des avions

Thomas Weber

Introduction

- AirData souhaite évaluer le comportement des différentes compagnies d'aviation et pouvoir anticiper les retards

- Source des données: Transtats (ministère des transports US)

  Tous les vols intérieurs aux US en 2016 (env. 5 500 000 vols)

- Objectif: tester plusieurs modèles supervisés, les optimiser (hyperparamètres), les comparer et implémenter le meilleur dans une API.

Points clés: encoding

- Beaucoup de variables catégorielles et/ou temporelles (jour, mois, etc..)

- Solutions possibles:
    - Label encoding
    - Count encoding
    - One-hot encoding
    - Circular encoding
    - Target (ou mean) encoding

Points clés: encoding

- Circular encoding:
  - Label encoding: variable entre 0 et n-1
  - Transformer cette variable en 2D sur le cercle unité:

$$\sin\frac{2k\pi}{n};\ \cos\frac{2k\pi}{n}$$

- Target encoding:
  - On regroupe les échantillons par valeur de la variable catégorielle
  - Pour chaque groupe, on encode la variable par la moyenne de la variable cible dans le groupe

Points clés: taille du dataset

- + 5 millions de lignes: long à manipuler avec une machine classique

- 2 solutions possibles:
  - Prendre un échantillon représentatif du dataset
  - Séparer les données par compagnie (12 en tout), et entraîner un modèle par compagnie

- 2^{e} solution choisie:
  - Permet de travailler sur tout le dataset
  - Fonctionne bien avec le sujet

Nettoyage

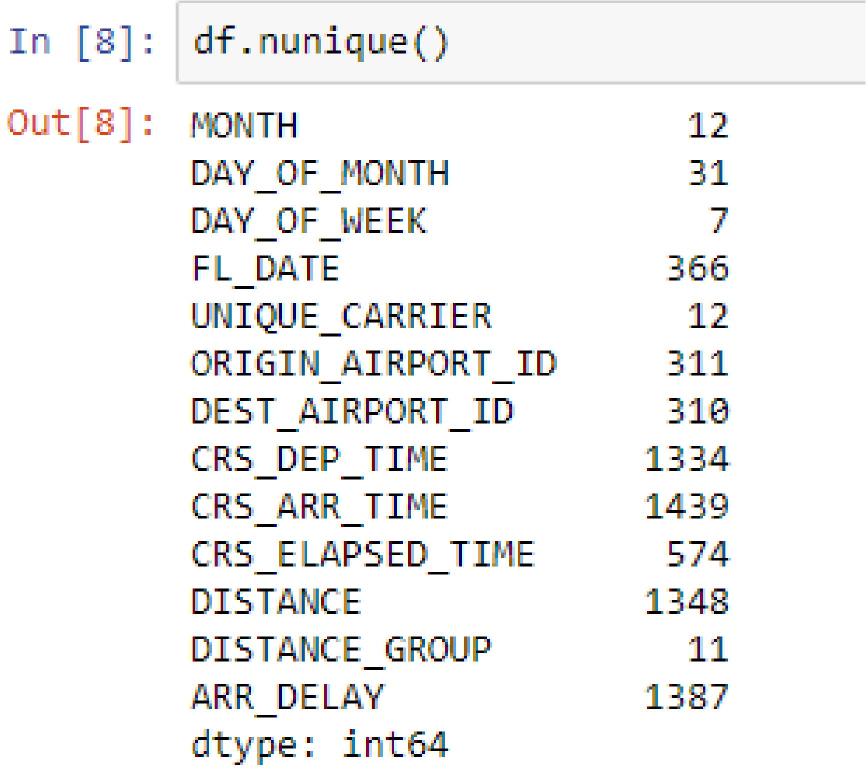
- Choix de la variable cible: ARR_DELAY
  (retard à l'arrivée)

- Suppression des vols annulés ou déroutés

- Informations gardées:
  - Date (jour/mois/jour de la semaine)
  - Compagnie
  - Aéroport de départ/arrivée
  - Heure de départ/arrivée
  - Distance

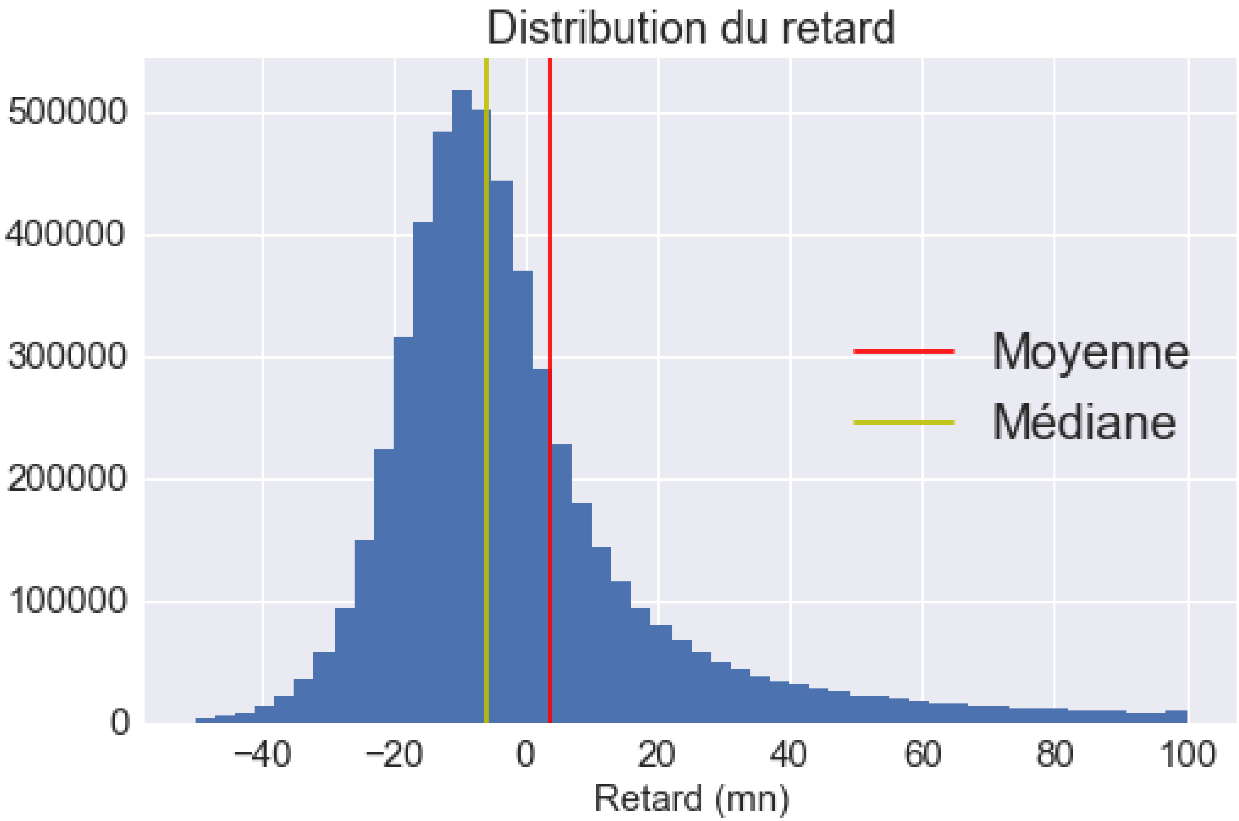
- Pas de valeurs manquantes

- Nouvelle feature: nombre de jours avant/après un jour férié

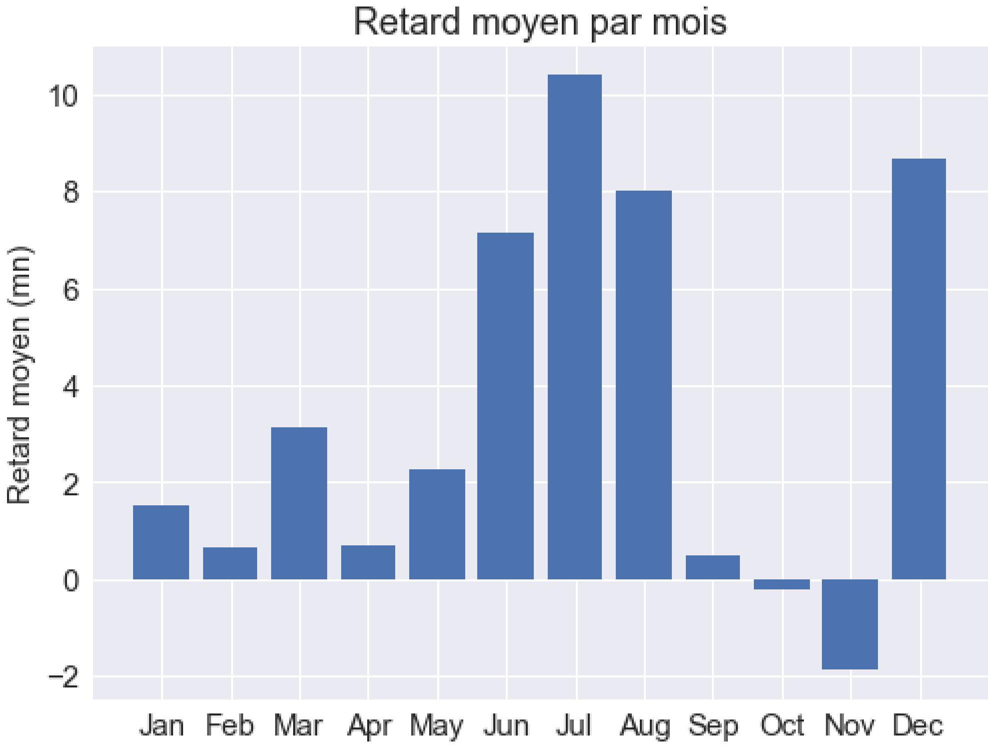


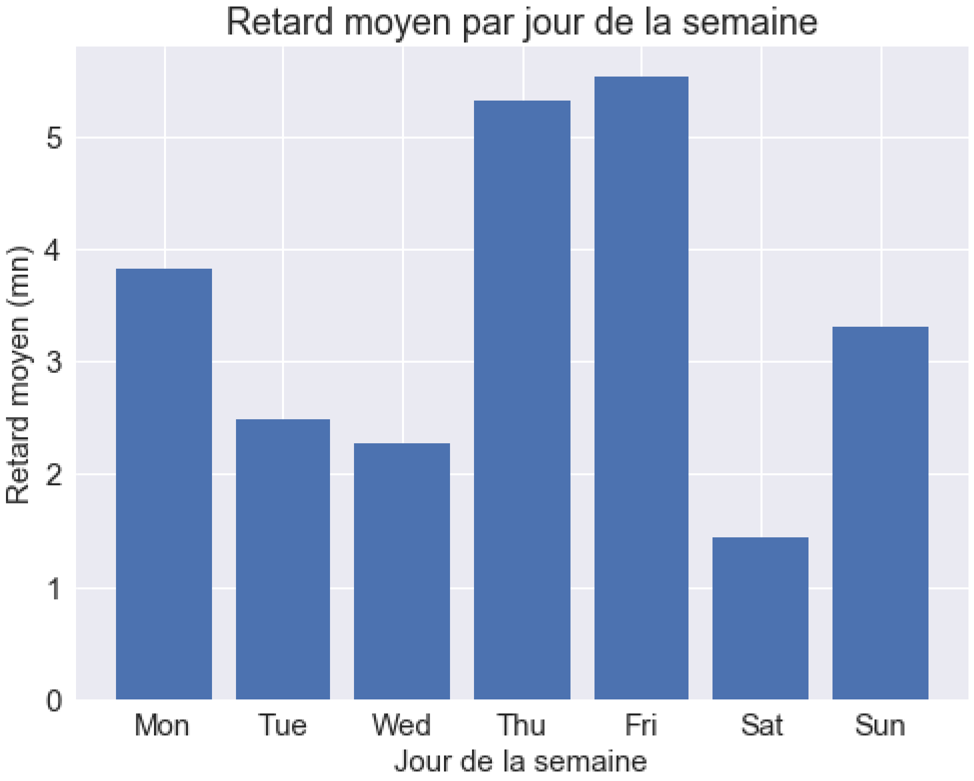
```
In [8]: df.nunique()

Out[8]: MONTH                   12
        DAY_OF_MONTH            31
        DAY_OF_WEEK              7
        FL_DATE                366
        UNIQUE_CARRIER          12
        ORIGIN_AIRPORT_ID      311
        DEST_AIRPORT_ID        310
        CRS_DEP_TIME          1334
        CRS_ARR_TIME          1439
        CRS_ELAPSED_TIME       574
        DISTANCE              1348
        DISTANCE_GROUP          11
        ARR_DELAY             1387
        dtype: int64
```

Exploration – Variable cible

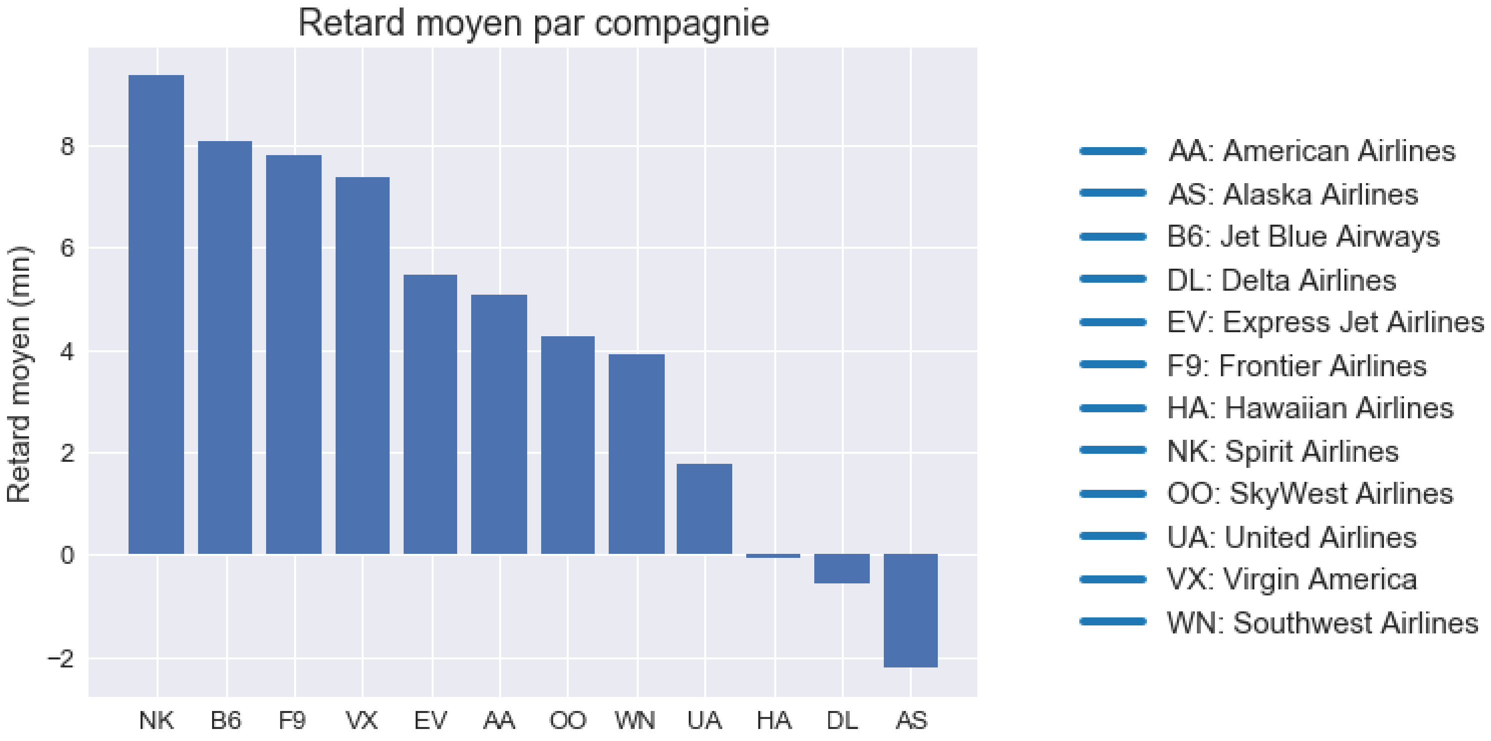


Exploration – Mois et jour de la semaine

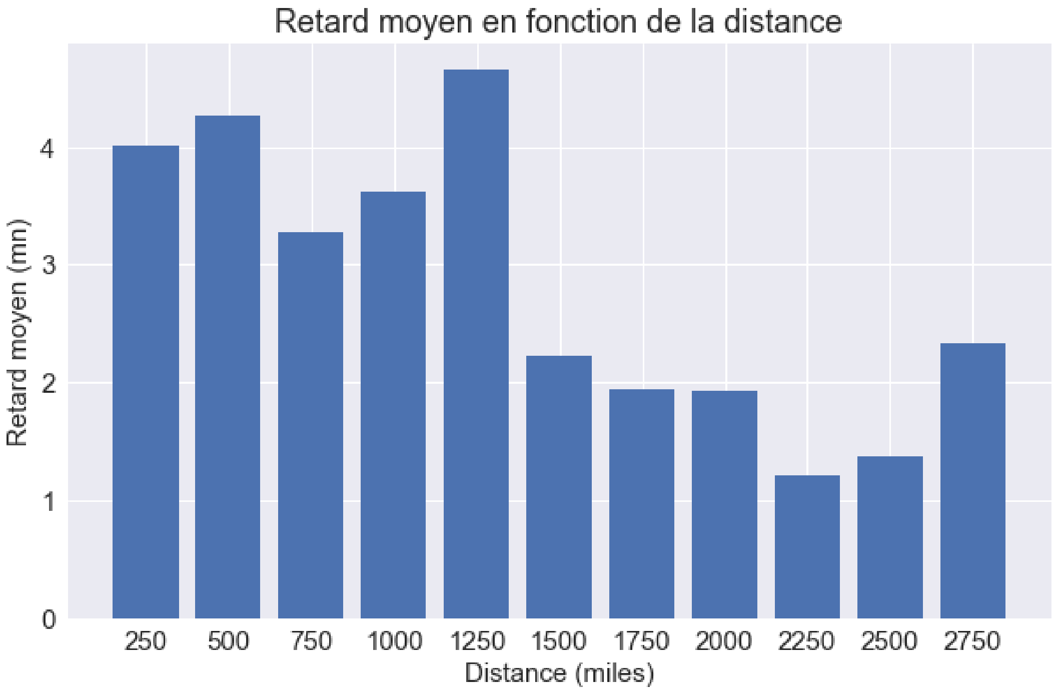


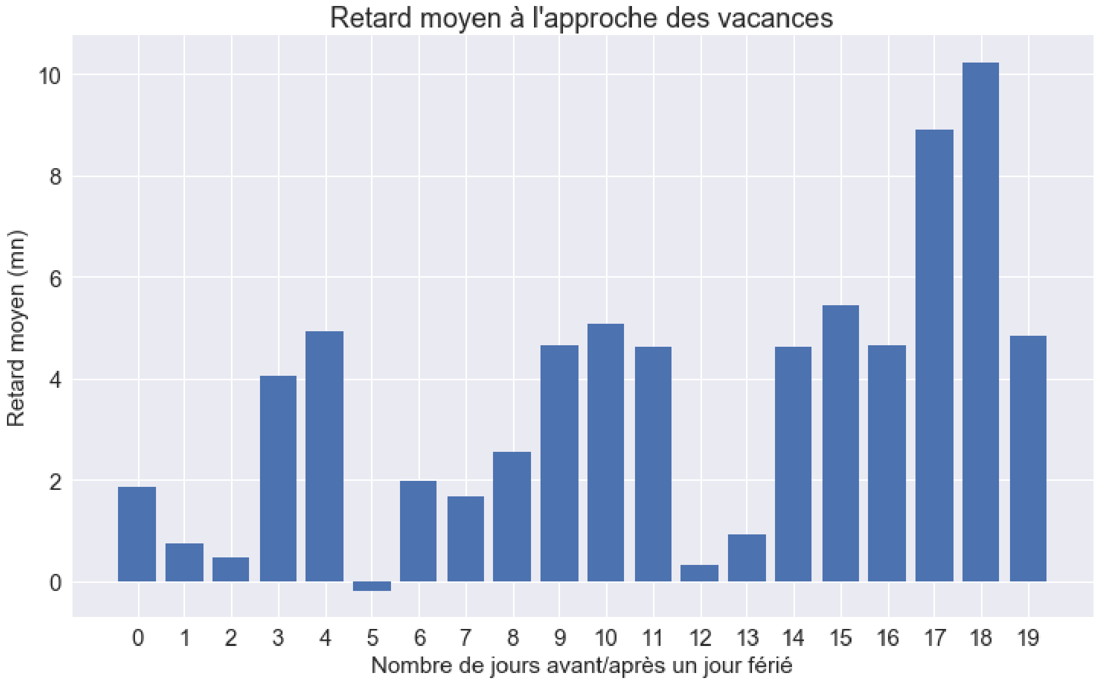


Exploration – Compagnies aériennes

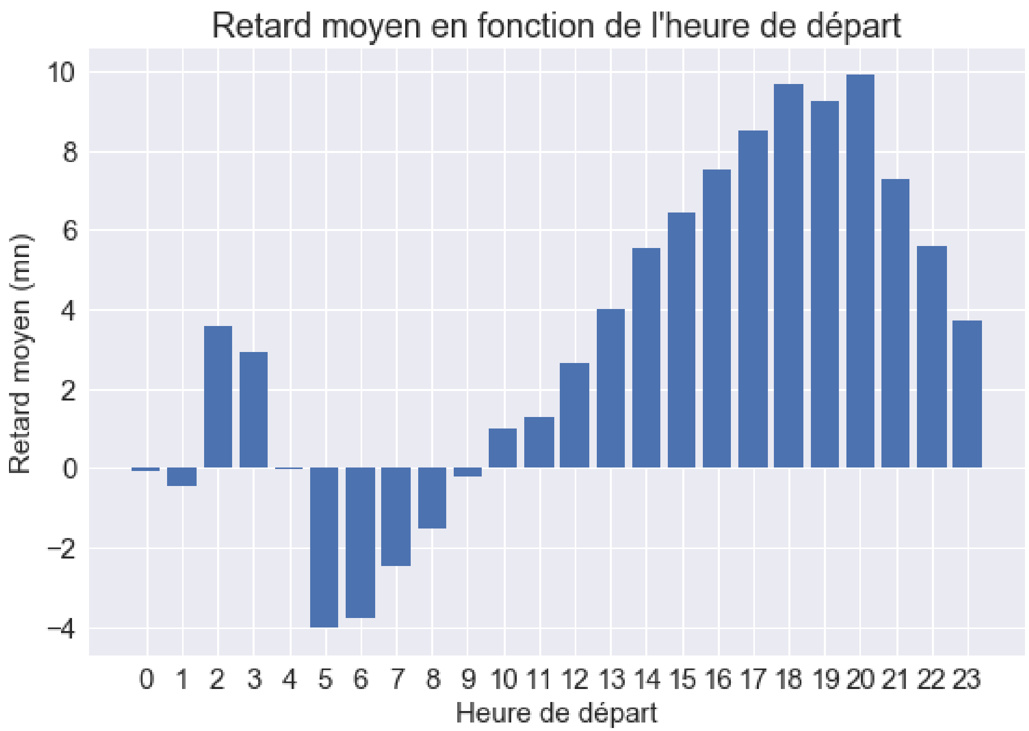


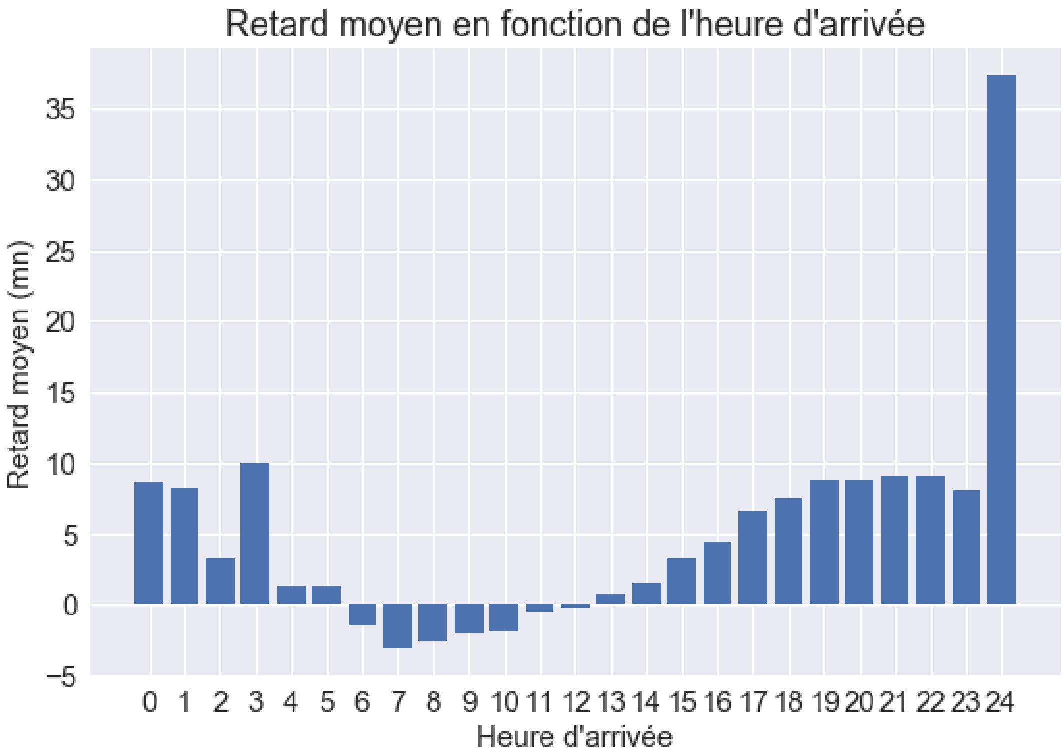
Exploration – Distance et vacances





Exploration – Heure de départ/d’arrivée





Choix de l’encoding

Méthode	MAE (mn)	R2 (%)	RMSE (mn)	Fit time (s)	Pred time (s)	Total time (s)
Label	23.00	4.96	36.62	0.12	0.02	0.14
Count	23.18	2.21	37.14	0.07	0.02	0.08
One-hot	21.12	20.24	33.54	0.18	1.12	1.30
Circular	23.02	4.74	36.66	0.68	0.02	0.71
Target	21.23	19.46	33.71	0.14	0.01	0.15

Outliers

Méthode	MAE (mn)	R2 (%)	RMSE (mn)
Avec outliers	21.23	19.46	33.71
Sans outliers (dans le training set)	20.24	16.67	34.29
Sans outliers	16.87	19.34	26.65

Modèles testés

- Modèles linéaires:
  - Régression linéaire simple
  - Régression ridge
  - Lasso
  - Elastic-Net
  - Régression linéaire après transformation polynomiale des features

- Modèles non linéaires:
  - K-NN
  - Bagging
  - Random Forest
  - Gradient Boosting

Modèles linéaires – Paramètres

- Validation croisée pour le choix optimal des paramètres:
  - Ridge:
    - alpha: $[10^{-5}, 10^{-4}, 10^{-3}, \ldots, 10^{4}, 10^{5}]$
  - Lasso:
    - alpha: $[10^{-5}, 10^{-4}, 10^{-3}, \ldots, 10^{4}, 10^{5}]$
  - Elastic-Net:
    - alpha: $[10^{-2}, 10^{-1}, \ldots, 10^{4}, 10^{5}]$
    - l1_ratio: [0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1]

- Polynomial Features:
  - degree: 3

Modèles non linéaires

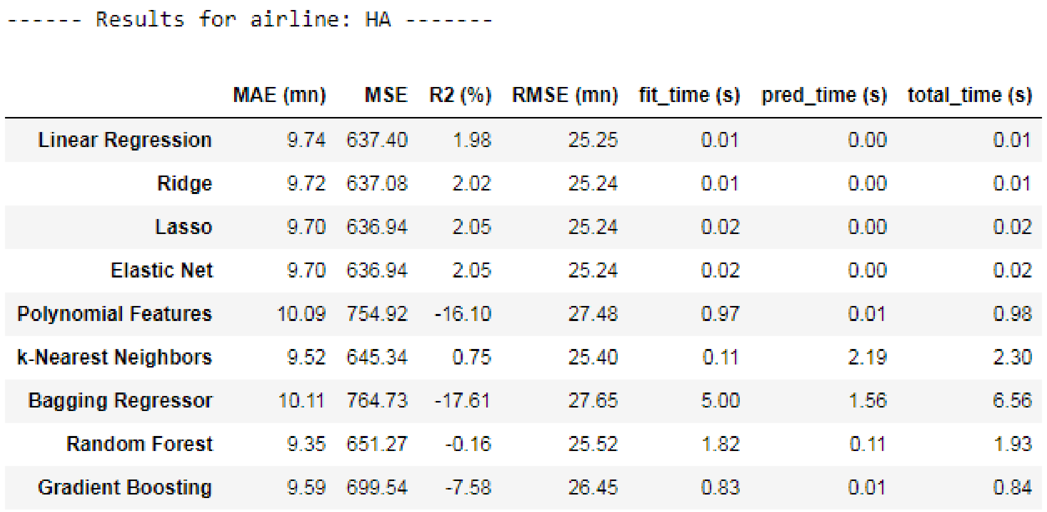
- Validation croisée pour k-NN:
  - n_neighbors: [20, 30, 40, 50]

- Bagging:
  - n_estimators: 50

- Random Forest:
  - n_estimators: 100, max_features: 3, min_samples_leaf: 10

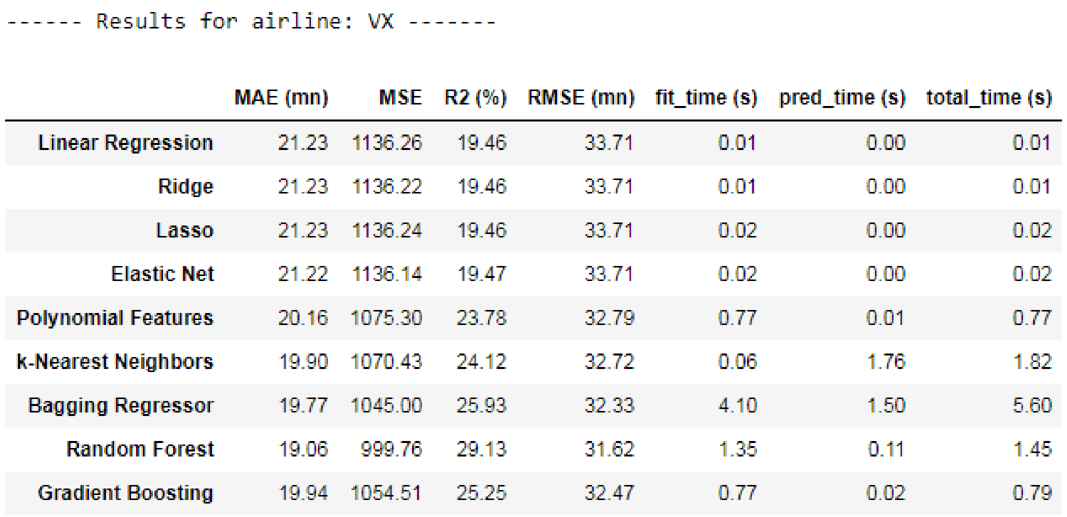
- Gradient Boosting:
  - n_estimators: 100, max_features: 3

Résultats par compagnie



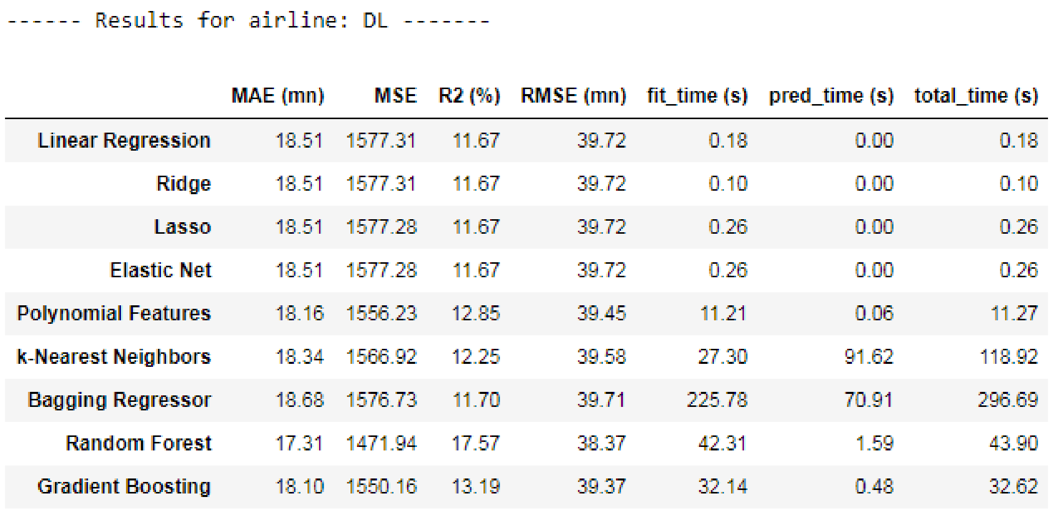
```
------ Results for airline: HA -------
```

	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
Linear Regression	9.74	637.40	1.98	25.25	0.01	0.00	0.01
Ridge	9.72	637.08	2.02	25.24	0.01	0.00	0.01
Lasso	9.70	636.94	2.05	25.24	0.02	0.00	0.02
Elastic Net	9.70	636.94	2.05	25.24	0.02	0.00	0.02
Polynomial Features	10.09	754.92	-16.10	27.48	0.97	0.01	0.98
k-Nearest Neighbors	9.52	645.34	0.75	25.40	0.11	2.19	2.30
Bagging Regressor	10.11	764.73	-17.61	27.65	5.00	1.56	6.56
Random Forest	9.35	651.27	-0.16	25.52	1.82	0.11	1.93
Gradient Boosting	9.59	699.54	-7.58	26.45	0.83	0.01	0.84



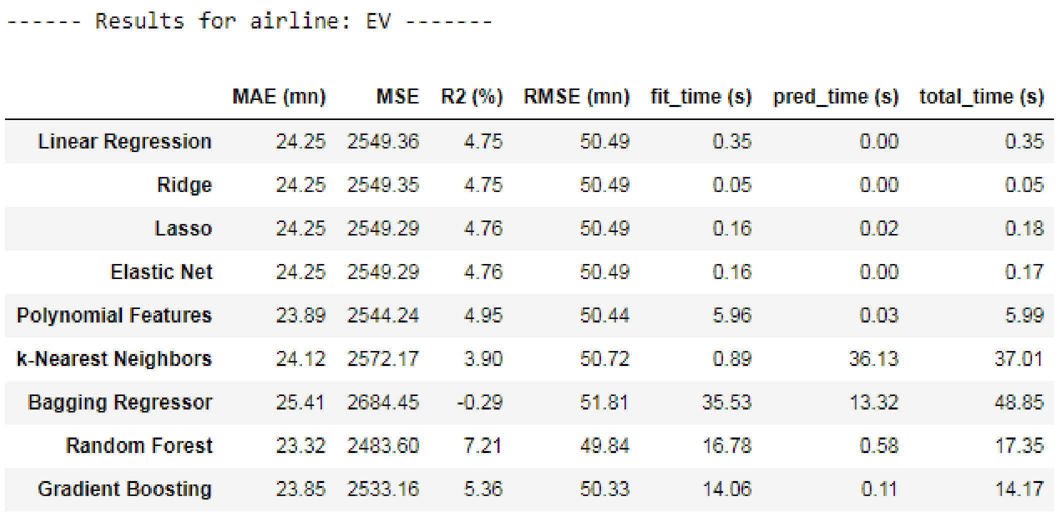
```
------ Results for airline: VX -------
```

	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
Linear Regression	21.23	1136.26	19.46	33.71	0.01	0.00	0.01
Ridge	21.23	1136.22	19.46	33.71	0.01	0.00	0.01
Lasso	21.23	1136.24	19.46	33.71	0.02	0.00	0.02
Elastic Net	21.22	1136.14	19.47	33.71	0.02	0.00	0.02
Polynomial Features	20.16	1075.30	23.78	32.79	0.77	0.01	0.77
k-Nearest Neighbors	19.90	1070.43	24.12	32.72	0.06	1.76	1.82
Bagging Regressor	19.77	1045.00	25.93	32.33	4.10	1.50	5.60
Random Forest	19.06	999.76	29.13	31.62	1.35	0.11	1.45
Gradient Boosting	19.94	1054.51	25.25	32.47	0.77	0.02	0.79



```
------ Results for airline: DL -------
```

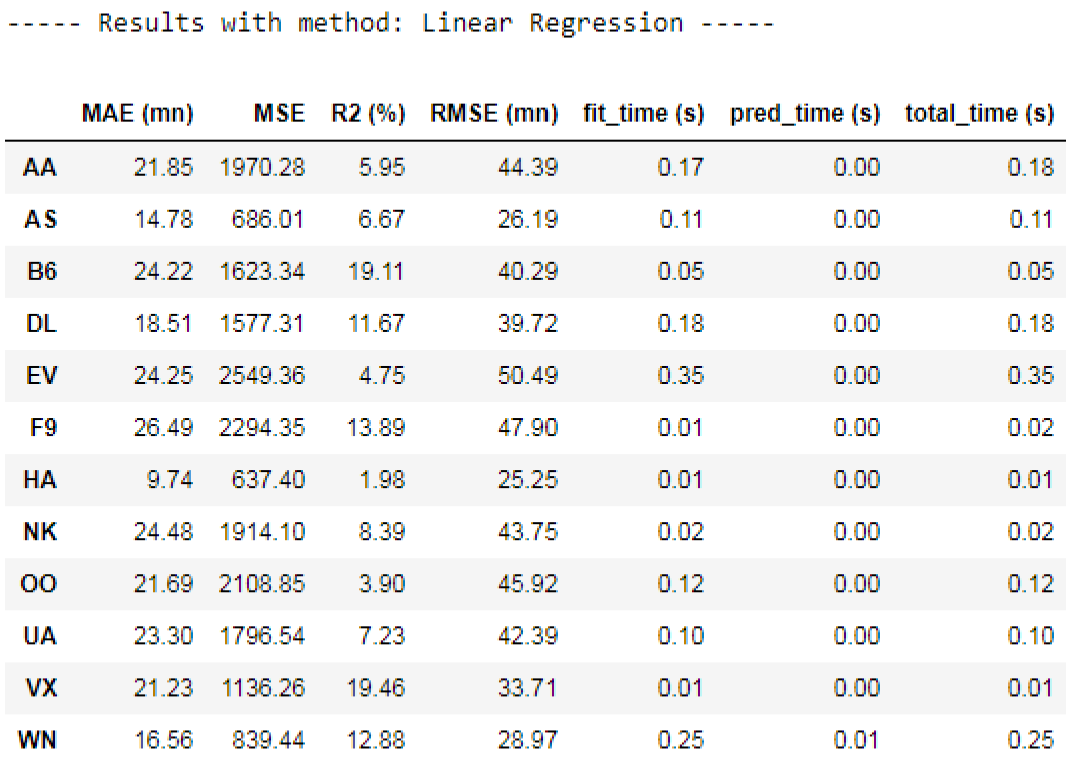
	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
Linear Regression	18.51	1577.31	11.67	39.72	0.18	0.00	0.18
Ridge	18.51	1577.31	11.67	39.72	0.10	0.00	0.10
Lasso	18.51	1577.28	11.67	39.72	0.26	0.00	0.26
Elastic Net	18.51	1577.28	11.67	39.72	0.26	0.00	0.26
Polynomial Features	18.16	1556.23	12.85	39.45	11.21	0.06	11.27
k-Nearest Neighbors	18.34	1566.92	12.25	39.58	27.30	91.62	118.92
Bagging Regressor	18.68	1576.73	11.70	39.71	225.78	70.91	296.69
Random Forest	17.31	1471.94	17.57	38.37	42.31	1.59	43.90
Gradient Boosting	18.10	1550.16	13.19	39.37	32.14	0.48	32.62



```
------ Results for airline: EV -------
```

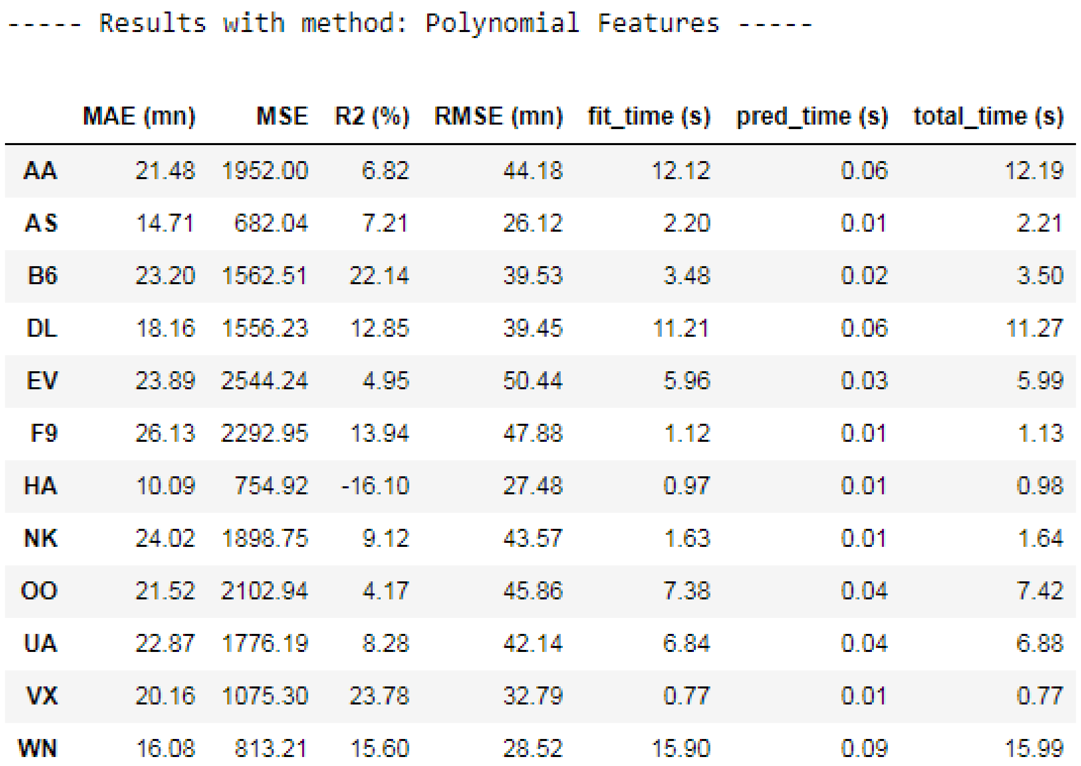
	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
Linear Regression	24.25	2549.36	4.75	50.49	0.35	0.00	0.35
Ridge	24.25	2549.35	4.75	50.49	0.05	0.00	0.05
Lasso	24.25	2549.29	4.76	50.49	0.16	0.02	0.18
Elastic Net	24.25	2549.29	4.76	50.49	0.16	0.00	0.17
Polynomial Features	23.89	2544.24	4.95	50.44	5.96	0.03	5.99
k-Nearest Neighbors	24.12	2572.17	3.90	50.72	0.89	36.13	37.01
Bagging Regressor	25.41	2684.45	-0.29	51.81	35.53	13.32	48.85
Random Forest	23.32	2483.60	7.21	49.84	16.78	0.58	17.35
Gradient Boosting	23.85	2533.16	5.36	50.33	14.06	0.11	14.17

Résultats par méthode



```
----- Results with method: Linear Regression -----
```

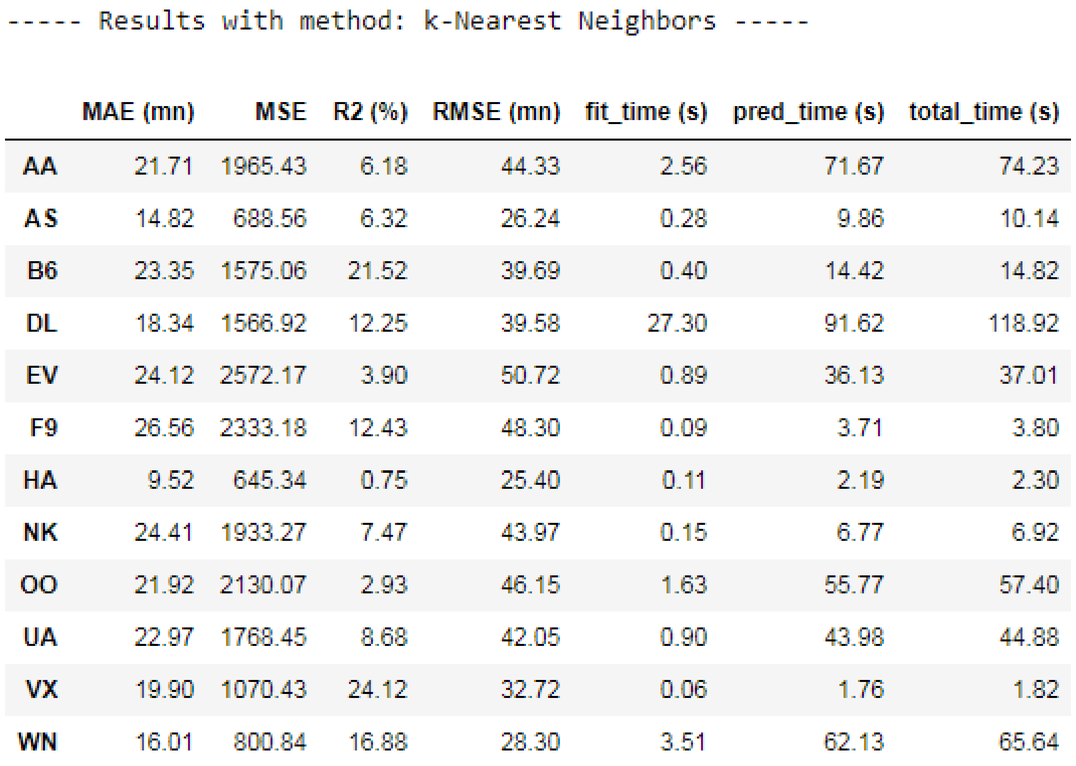
	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
AA	21.85	1970.28	5.95	44.39	0.17	0.00	0.18
AS	14.78	686.01	6.67	26.19	0.11	0.00	0.11
B6	24.22	1623.34	19.11	40.29	0.05	0.00	0.05
DL	18.51	1577.31	11.67	39.72	0.18	0.00	0.18
EV	24.25	2549.36	4.75	50.49	0.35	0.00	0.35
F9	26.49	2294.35	13.89	47.90	0.01	0.00	0.02
HA	9.74	637.40	1.98	25.25	0.01	0.00	0.01
NK	24.48	1914.10	8.39	43.75	0.02	0.00	0.02
OO	21.69	2108.85	3.90	45.92	0.12	0.00	0.12
UA	23.30	1796.54	7.23	42.39	0.10	0.00	0.10
VX	21.23	1136.26	19.46	33.71	0.01	0.00	0.01
WN	16.56	839.44	12.88	28.97	0.25	0.01	0.25



```
----- Results with method: Polynomial Features -----
```

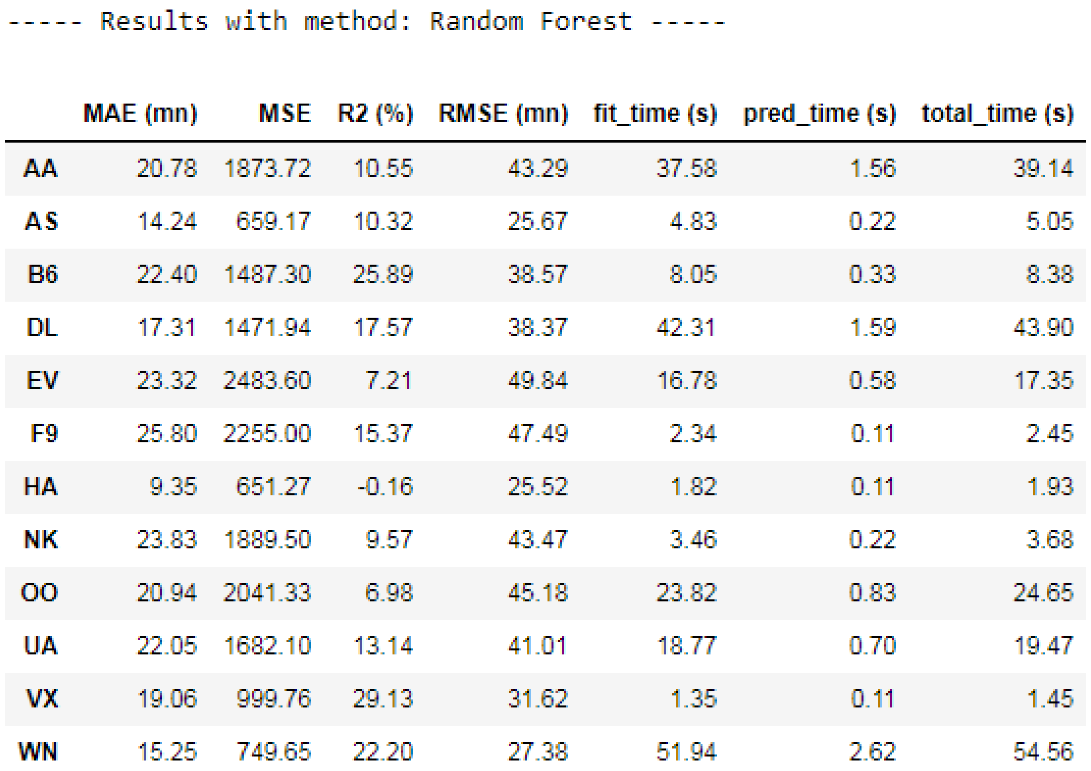
	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
AA	21.48	1952.00	6.82	44.18	12.12	0.06	12.19
AS	14.71	682.04	7.21	26.12	2.20	0.01	2.21
B6	23.20	1562.51	22.14	39.53	3.48	0.02	3.50
DL	18.16	1556.23	12.85	39.45	11.21	0.06	11.27
EV	23.89	2544.24	4.95	50.44	5.96	0.03	5.99
F9	26.13	2292.95	13.94	47.88	1.12	0.01	1.13
HA	10.09	754.92	-16.10	27.48	0.97	0.01	0.98
NK	24.02	1898.75	9.12	43.57	1.63	0.01	1.64
OO	21.52	2102.94	4.17	45.86	7.38	0.04	7.42
UA	22.87	1776.19	8.28	42.14	6.84	0.04	6.88
VX	20.16	1075.30	23.78	32.79	0.77	0.01	0.77
WN	16.08	813.21	15.60	28.52	15.90	0.09	15.99

Résultats par méthode



```
----- Results with method: k-Nearest Neighbors -----
```

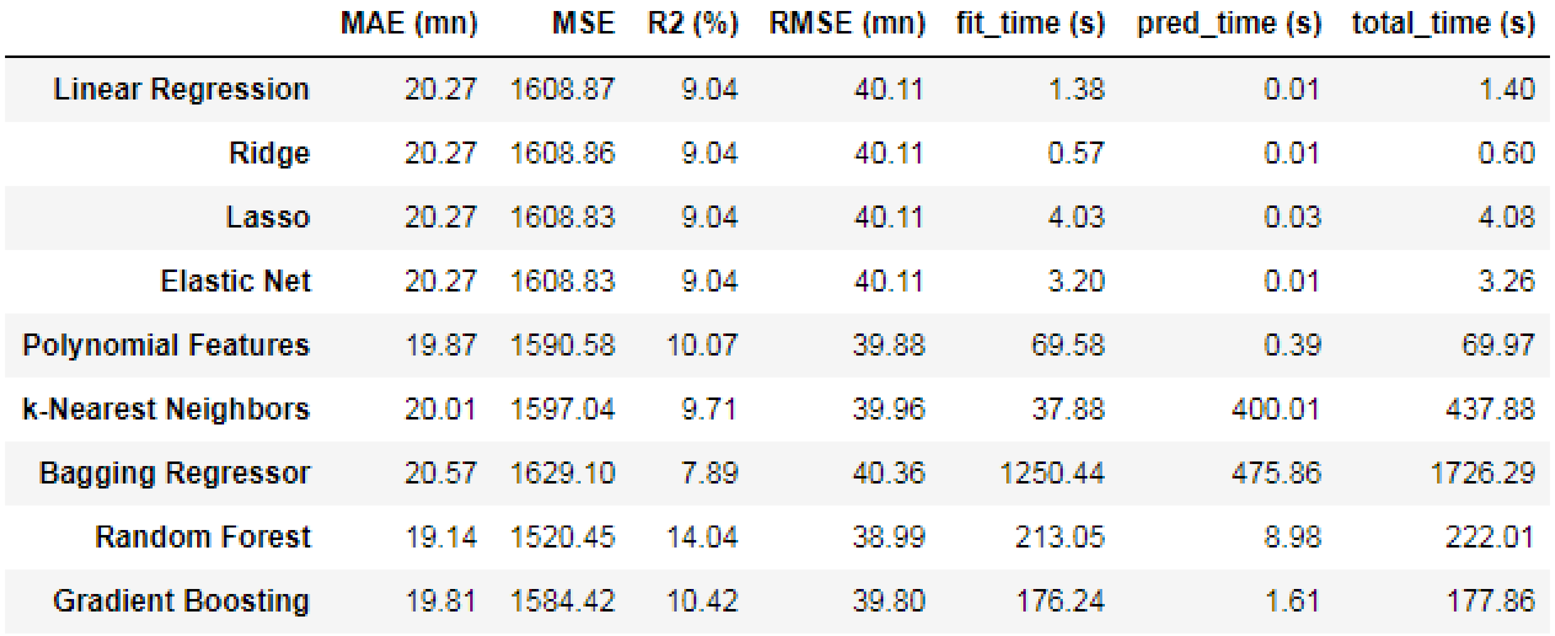
	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
AA	21.71	1965.43	6.18	44.33	2.56	71.67	74.23
AS	14.82	688.56	6.32	26.24	0.28	9.86	10.14
B6	23.35	1575.06	21.52	39.69	0.40	14.42	14.82
DL	18.34	1566.92	12.25	39.58	27.30	91.62	118.92
EV	24.12	2572.17	3.90	50.72	0.89	36.13	37.01
F9	26.56	2333.18	12.43	48.30	0.09	3.71	3.80
HA	9.52	645.34	0.75	25.40	0.11	2.19	2.30
NK	24.41	1933.27	7.47	43.97	0.15	6.77	6.92
OO	21.92	2130.07	2.93	46.15	1.63	55.77	57.40
UA	22.97	1768.45	8.68	42.05	0.90	43.98	44.88
VX	19.90	1070.43	24.12	32.72	0.06	1.76	1.82
WN	16.01	800.84	16.88	28.30	3.51	62.13	65.64



```
----- Results with method: Random Forest -----
```

	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
AA	20.78	1873.72	10.55	43.29	37.58	1.56	39.14
AS	14.24	659.17	10.32	25.67	4.83	0.22	5.05
B6	22.40	1487.30	25.89	38.57	8.05	0.33	8.38
DL	17.31	1471.94	17.57	38.37	42.31	1.59	43.90
EV	23.32	2483.60	7.21	49.84	16.78	0.58	17.35
F9	25.80	2255.00	15.37	47.49	2.34	0.11	2.45
HA	9.35	651.27	-0.16	25.52	1.82	0.11	1.93
NK	23.83	1889.50	9.57	43.47	3.46	0.22	3.68
OO	20.94	2041.33	6.98	45.18	23.82	0.83	24.65
UA	22.05	1682.10	13.14	41.01	18.77	0.70	19.47
VX	19.06	999.76	29.13	31.62	1.35	0.11	1.45
WN	15.25	749.65	22.20	27.38	51.94	2.62	54.56

Résultats – Synthèse



	MAE (mn)	MSE	R2 (%)	RMSE (mn)	fit_time (s)	pred_time (s)	total_time (s)
Linear Regression	20.27	1608.87	9.04	40.11	1.38	0.01	1.40
Ridge	20.27	1608.86	9.04	40.11	0.57	0.01	0.60
Lasso	20.27	1608.83	9.04	40.11	4.03	0.03	4.08
Elastic Net	20.27	1608.83	9.04	40.11	3.20	0.01	3.26
Polynomial Features	19.87	1590.58	10.07	39.88	69.58	0.39	69.97
k-Nearest Neighbors	20.01	1597.04	9.71	39.96	37.88	400.01	437.88
Bagging Regressor	20.57	1629.10	7.89	40.36	1250.44	475.86	1726.29
Random Forest	19.14	1520.45	14.04	38.99	213.05	8.98	222.01
Gradient Boosting	19.81	1584.42	10.42	39.80	176.24	1.61	177.86

API

- Modèle choisi: Régression linéaire avec transformation polynomiale

- Stockage des paramètres dans des dictionnaires pour:
  - Encoding
  - Standardisation
  - Régression

- Pas besoin de stocker les datasets sur le serveur et prédiction rapide