

Projet de Data Science (Openclassrooms PJ4)

Anticiper les retards de vol d'avion

Données :



A photograph showing a man from behind, sitting on a long, curved, modern-style bench in an airport terminal. He is looking out through large windows onto a tarmac where several aircraft are parked. The scene is bathed in warm, golden sunlight, creating strong shadows and highlights on the floor and the man's clothing.

Les enjeux ...

Pour la qualité de service au
client

Les enjeux ...



Pour optimiser
la logistique

Problématique

Prédire numériquement le retard à l'arrivée d'un vol

- ⇒ Problématique 1 : Uniquement à partir des informations connues à la réservation
- ⇒ Problématique 2 : Puis en ajoutant les informations connues le jour du départ

Comment mesurer la performance atteinte ?

- ⇒ Root Mean Square Error
- ⇒ Erreur moyenne 80 % du temps

Fournir une interface web de consultation des prédictions

Démarche suivie

- 1/ Compréhension du cycle de vie d'un vol
- 2/ Feature engineering
- 3/ Choix sur les données manquantes et les outliers
- 4/ Choix d'encodage des features
- 5/ Conception du modèle de prédiction
- 6/ Optimisation des paramètres du random forest par validation croisée
- 7/ Optimisation tenant compte des valeurs élevées de retard plus difficiles à prédire
- 8/ Perspectives pour aller plus loin
- 9/ Interface web



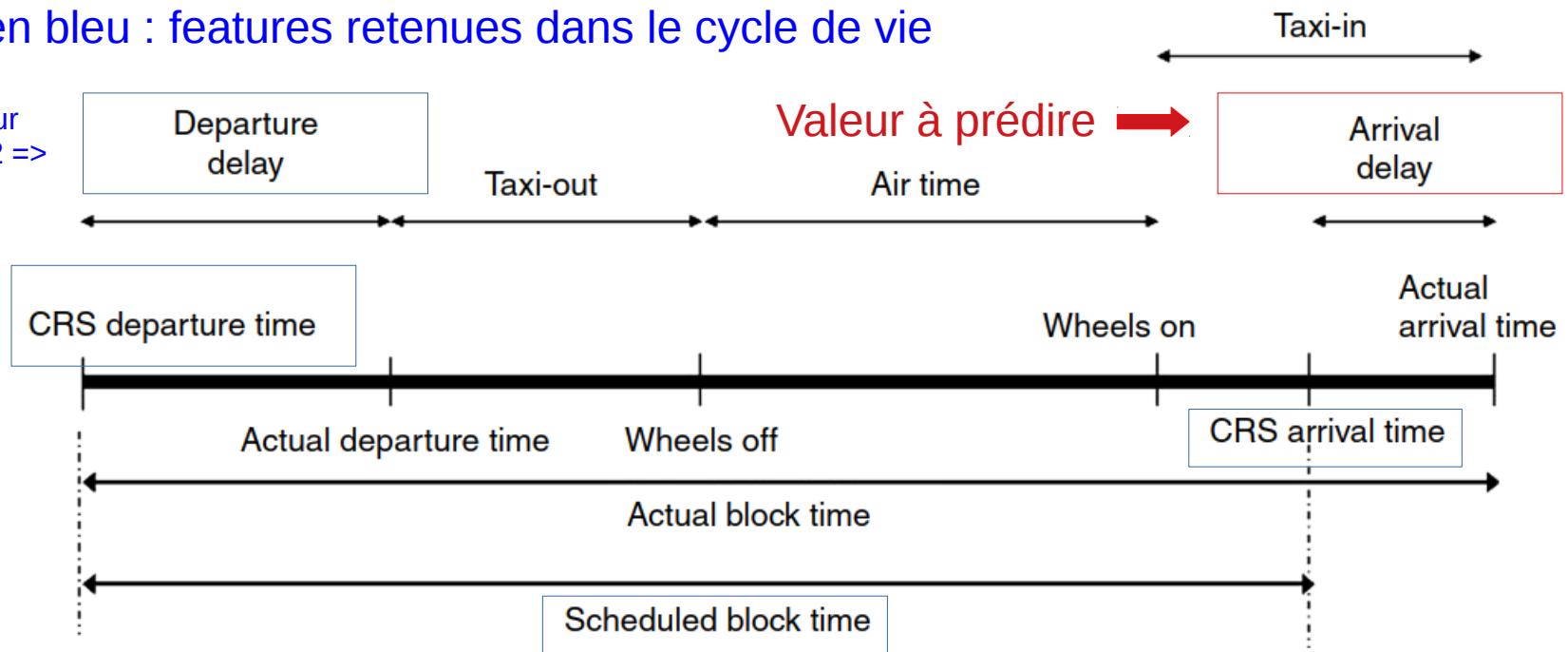
Compréhension du cycle de vie d'un vol

Cycle de vie d'un vol

Figure 2 Main Segments of Air Travel Time

Encadré en bleu : features retenues dans le cycle de vie

Uniquement pour
problématique 2 =>





Feature engineering

Features quantitatives

Nom	Description
CRS_ARR_TIME	Heure d'arrivée prévue => Converti en minutes
CRS_DEP_TIME	Heure de départ prévu => Converti en minutes
CRS_ELAPSED_TIME	Temps de trajet prévu
DAY_OF_MONTH	Jour du mois de départ
DAY_OF_WEEK	Jour de la semaine de départ
DISTANCE	Distance (miles)
MONTH	Mois de départ
DEP_DELAY	Retard au départ (minutes) : uniquement pour le 2ème modèle

Features quantitatives (2/2)

Nom	Description
NBFLIGHTS_FORDAYHOUR_FORAIRPORT	Feature ajoutée : Nombre de vols par heure le jour du départ dans l'aéroport considéré
NBFLIGHTS_FORDAY_FORAIRPORT	Feature ajoutée : Nombre de vols le jour du départ dans l'aéroport considéré

Features qualitatives

Nom	Description
ORIGIN	Aéroport d'origine
DEST	Aéroport de destination
UNIQUE_CARRIER	Compagnie aérienne

Valeur à prédire

Nom	Description
ARR_DELAY	Retard à l'arrivée en minutes Peut être négatif (= avance)



Choix sur les
données manquantes
et les outliers

Choix réalisé sur les dates, les aéroports, les compagnies :

L'année est toujours 2016 dans le jeu de données

=> L'information n'a donc pas été conservée

Plusieurs variables équivalentes sont disponibles pour les aéroports, les compagnies, la date de départ : conservation d'une seule variable pour chaque

La ville n'a pas été conservée (les aéroports d'origine et de destination ont été jugés suffisants)

Choix réalisé sur les retards :

- ⇒ les vols annulés ou déviés ne sont pas considérés comme des retards, et ne sont pas considérés par le modèle
- ⇒ Certains retards sont causés par un vol précédent qui était en dépendance (LATE_AIRCRAFT_DELAY == 1) : ils ne sont pas considérés par le modèle car cela nécessiterait d'avoir a priori l'information sur les vols en dépendance.
- ⇒ Plusieurs variables possibles pour les retards : on a conservé la variable qui apportait le plus d'informations (valeur numérique non catégorisée, pouvant être négative de façon à pouvoir prédire l'avance)

Suppression des outliers :

- ⇒ Certains aéroports qui n'apparaissent qu'une seule fois
- ⇒ Suppression d'une compagnie aérienne
- ⇒ Suppression des 2 % des vols ayant les retards les plus extrêmes



Choix d'encodage des features

Features quantitatives :

- ⇒ Conversion des heures de départ et d'arrivée en minutes
- ⇒ Standard scaling : centrage + écart type de 1

Features qualitatives :

2 types d'encodage ont été testés :

- ⇒ One hot encoding avec conservation de 80 % des valeurs
(20 % des valeurs sont dans une catégorie « OTHERS »)
(Scaling min max des features quantitatives pour une homogénéité avec les features 1hot encodées)
=> RMSE : 27.29 min
- ⇒ Target encoding : encodage numérique en fonction de la moyenne des retards (plus la moyenne est élevée plus le retard est élevé)
=> RMSE : 26.72 min

Conception du modèle de prédiction



Synthèse des traitements du modèle

Data cleaning / split

Airports, companies

Delays

Outliers

Split train / test

Stratified on target delay

Preparation pipeline

Filter percentile

Time conversion

OHE

Target

Feature selection

Scaling

Prediction model

Linear Regression

Random forest

GridSearch on Random Forest

Error optimisation



Les principales étapes de la modélisation

- 1/ Modèle naïf
- 2/ Comparaison target encoding / one hot encoding
 - ⇒ target encoding retenu
- 3/ Comparaison régression linéaire degré 3 / random forest 1000 - 500
 - ⇒ Constat que le random forest fit mieux le training set, et overfit : plus prometteur
- 4/ Recherche de paramètres avec validation croisée sur le random forest
- 5/ Optimisation finale des valeurs prédites du meilleur modèle

Choix réalisés : data preparation

Composante du modèle	Choix testés	Choix retenu
Stratégie de stratification	1/ Répartition uniforme ARR_DELAY => RMSE 26.97	1/
	2/ Répartition aléatoire => RMSE 27.14	
Encodage des features	1/ One hot encoding => RMSE 27.29	2/
	2/ Target encoding => RMSE 26.72	

Modèle naïf « Always mean »

Le modèle retenu au final devra faire mieux que le modèle naïf.

Ce modèle prédit le retard moyen pour chaque instance.

- RMSE : 42.63 min
- Erreur moyenne 90% du temps : **14.07 min**

Comparaison :

> régression linéaire de degré 3
avec

> random forest profondeur max 1000, estimateurs 500

Modèle 1 : régression linéaire

Polynomial features : degré 3

> 12 features degré 1

> 455 features degré 3



RMSE train : 26.64

Mean prediction error 90.0% train : 9.93

RMSE test : 26.97

Modèle 2 : random forest

Profondeur max : 1000

Nb d'estimateurs : 500

Max features : 4



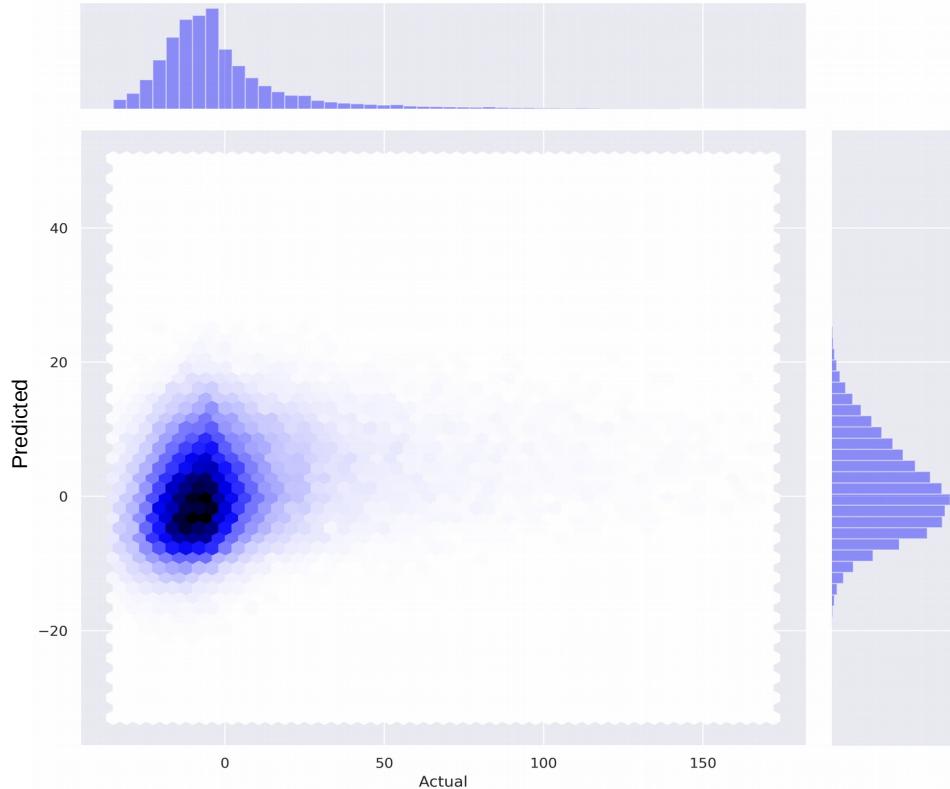
RMSE train : 9.94

Mean prediction error 90.0% train : 3.77

RMSE test : 27.24

Régression linéaire deg 3 / Random forest 1000 - 500: Prédictions sur training set vs valeurs réelles

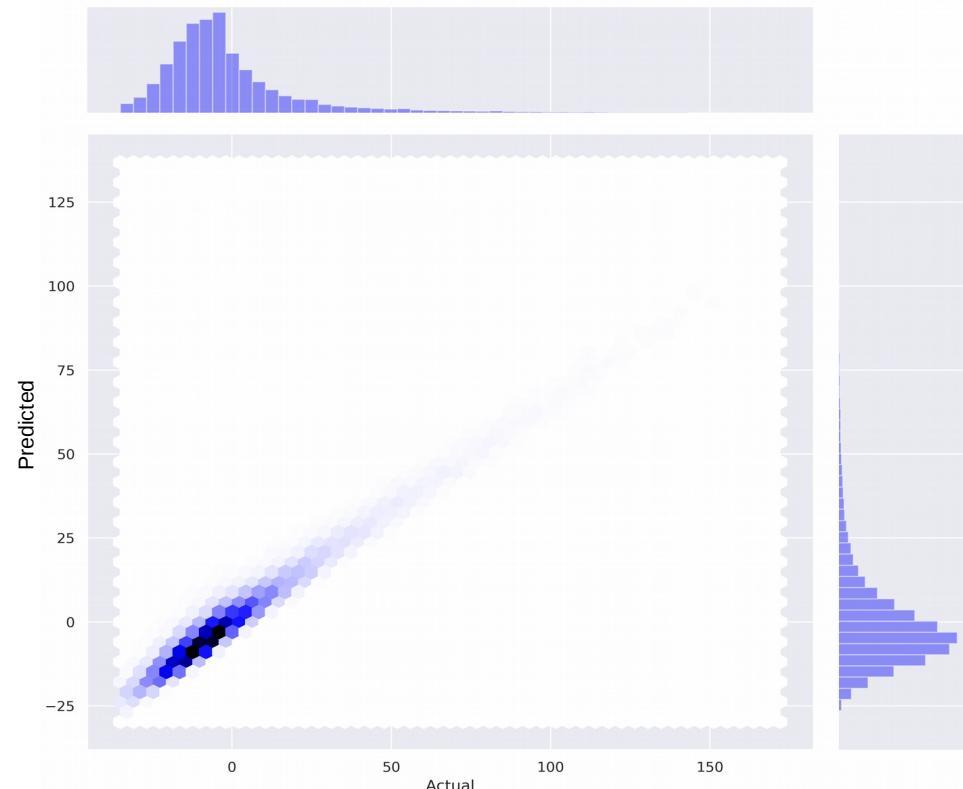
Linear regression : comparison actual values / predict values on training set



RMSE train : 26.64

Mean prediction error 90.0% : 9.93

Random forest : comparison actual values / predict values on training set

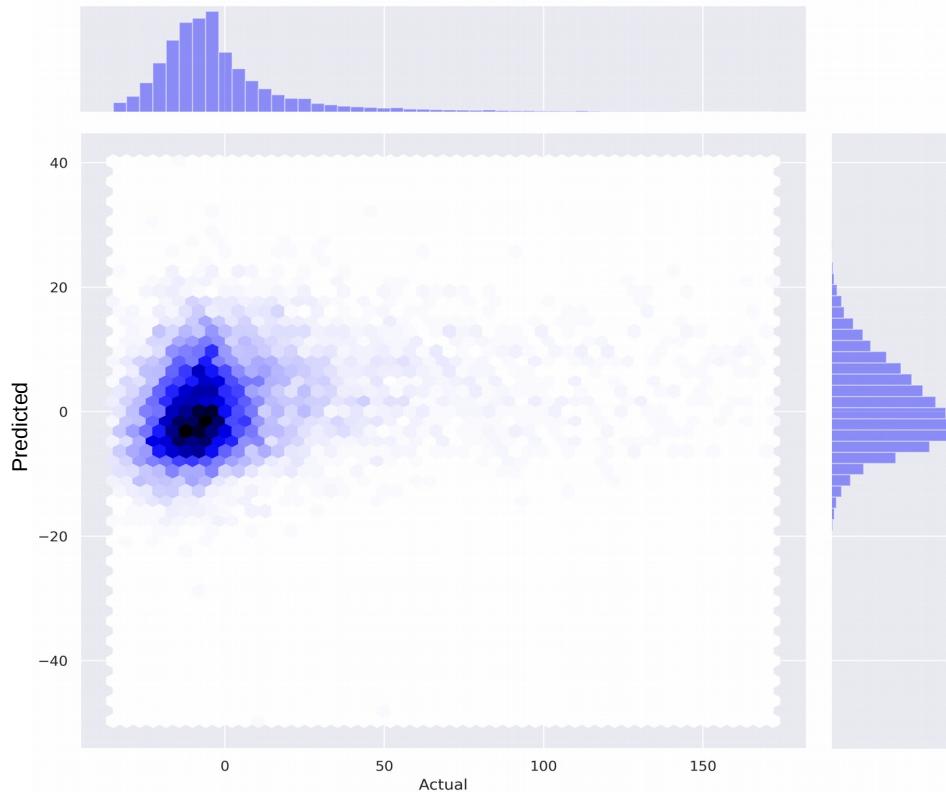


RMSE train : 9.94

Mean prediction error 90.0% : 3.77

Régression linéaire deg 3 / Random forest 1000-500: comparaison des prédictions sur test set vs valeurs réelles

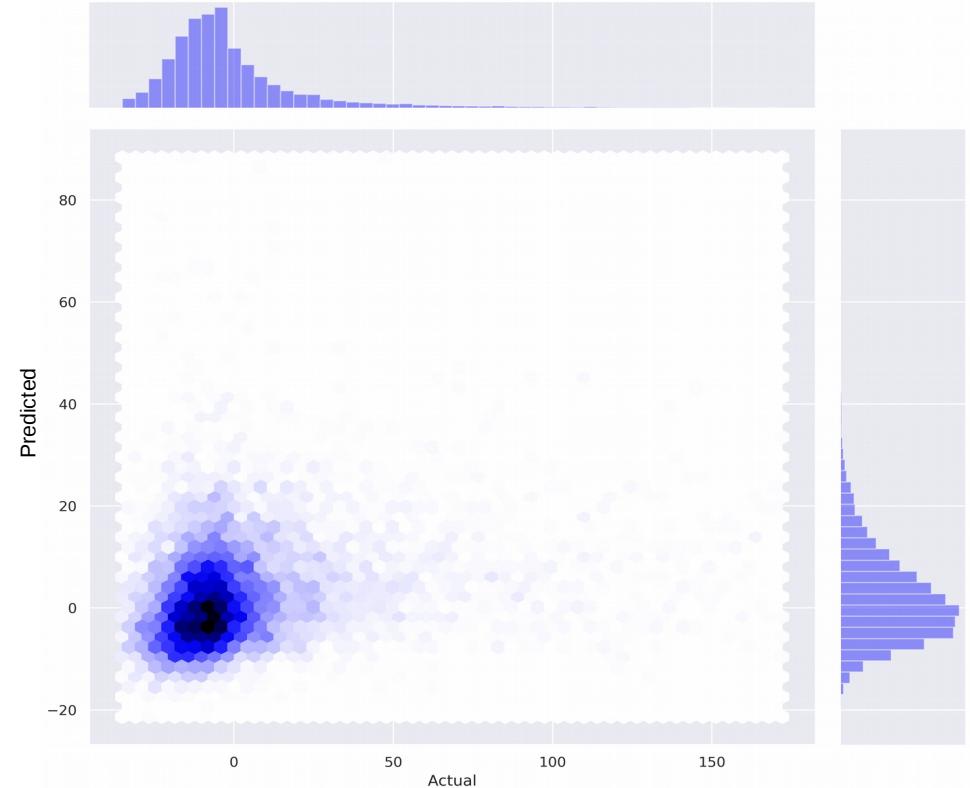
Linear regression : comparison actual values / predict values on test set



RMSE : 26.97

Mean prediction error 90.0% : 9.93

Random forest overfit : comparison actual values / predict values on test set



RMSE : 27.24

Mean prediction error 90.0% : 10,36



Optimisation des
paramètres par
validation croisée

Optimisation du random forest avec recherche de paramètres et validation croisée

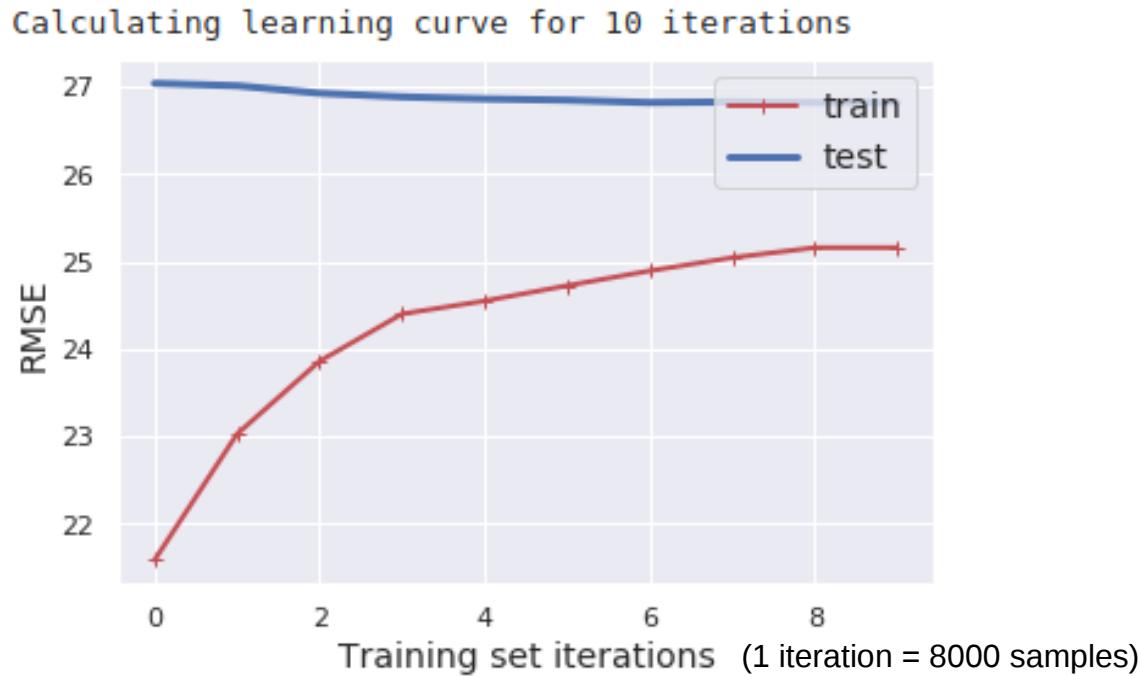
max_depth	max_features	max_leaf_nodes	n_estimators	mean_test_score
10	4	Unlimited	1000	-712,95
10	4	Unlimited	500	-713,07
10	8	Unlimited	1000	-713,52
10	4	Unlimited	200	-713,55
10	8	Unlimited	500	-713,75
10	12	Unlimited	1000	-713,96
10	4	Unlimited	100	-714,08
10	12	Unlimited	500	-714,17
10	2	Unlimited	1000	-714,26

=> Le meilleur modèle retenu est en premier :

RMSE test : 26.81

Mean prediction error 90.0% of the time : 9.92 / 10 % of the time : 45.79

Courbe d'apprentissage du random forest optimisé



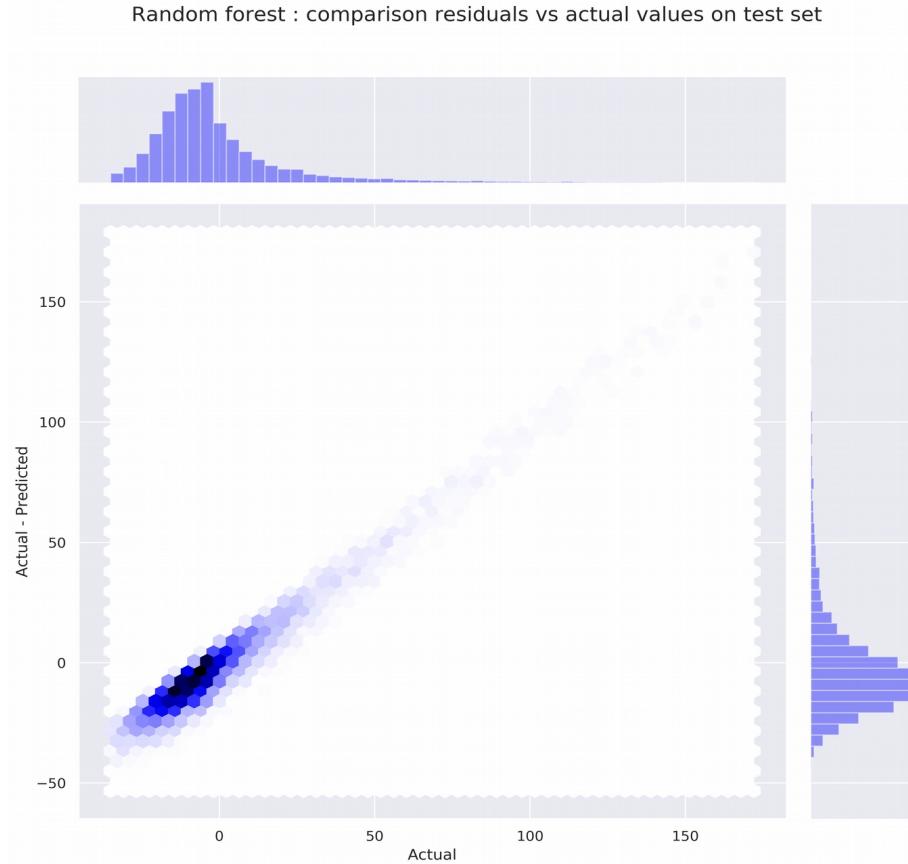
Importance des features pour le modèle random forest optimisé par validation croisée

	Feature name	Feature importance	Cumulated feature importance
1	CRS_DEP_TIME	15.29%	15.29%
0	CRS_ARR_TIME	13.08%	28.38%
5	DEST	11.54%	39.92%
9	NBFLIGHTS_FORDAY_FORAIRPORT	8.46%	48.38%
10	ORIGIN	8.42%	56.80%
7	MONTH	7.53%	64.33%
2	CRS_ELAPSED_TIME	7.36%	71.68%
11	UNIQUE_CARRIER	6.71%	78.39%
3	DAY_OF_MONTH	6.27%	84.66%
6	DISTANCE	6.21%	90.87%
8	NBFLIGHTS_FORDAYHOUR_FORAIRPORT	5.53%	96.41%
4	DAY_OF_WEEK	3.59%	100.00%



Optimisation tenant
compte des valeurs
élevées plus difficiles
à prédire

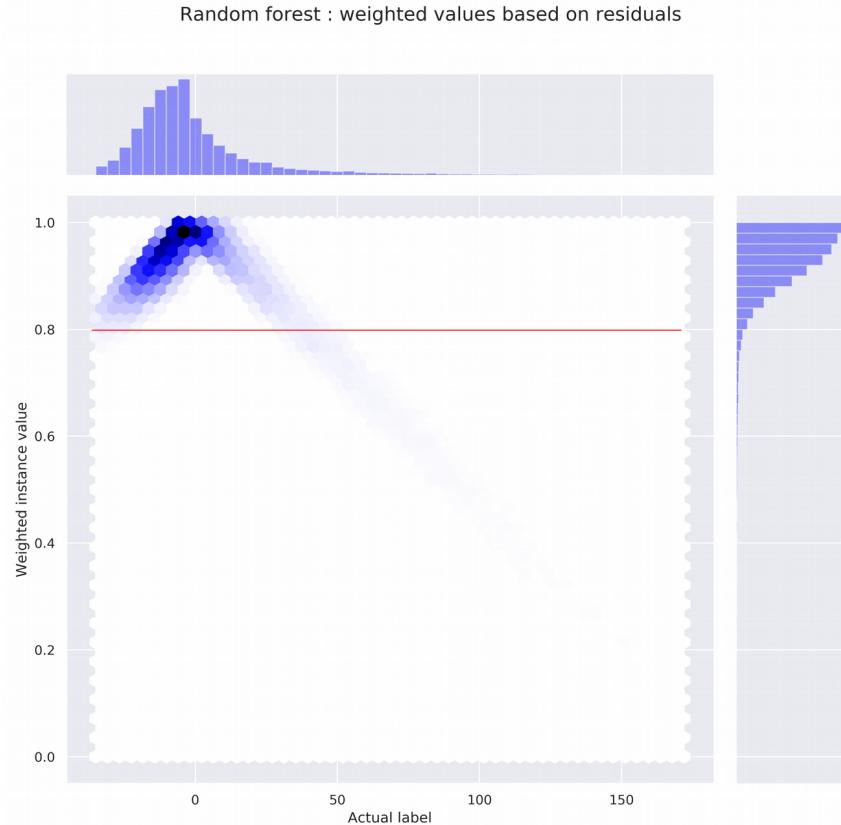
Random forest : comparaison des résidus vs valeurs réelles



=> Constat que plus la valeur à prédire est élevée, plus le modèle fait des erreurs importantes dans ses prédictions

Random forest : optimisation liée aux valeurs élevées de retard plus difficiles à prédire (1/2)

- Affectation d'un poids aux données d'entraînement : moins la valeur a été bien prédite sur le training set, plus on va lui affecter un poids faible



Conservation des instances dans le training set : uniquement celles qui ont un poids > 0.80

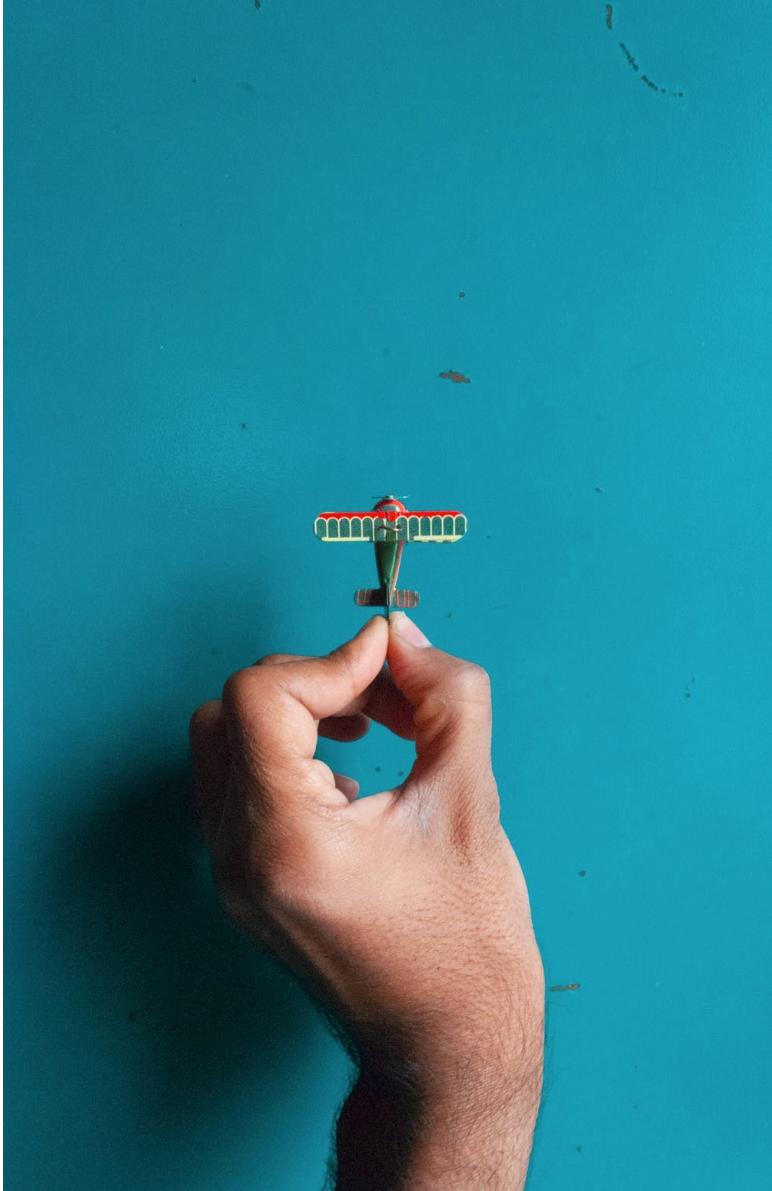
Réentraînement du modèle avec le training set ainsi réduit :

RMSE test set : 27.69 (un peu moins bien),

RMSE training set 12.75

Erreur moyenne 90 % du temps : **7.90** vs **9.92**

Erreur moyenne 10 % du temps : **48.03** vs **45.79**

A photograph of a person's hand holding a small, colorful toy biplane. The toy has red wings with yellow stripes, a green body, and a red tail. It is held between the thumb and forefinger of the hand, which is positioned vertically. The background is a solid teal color.

Difficultés
rencontrées et
perspectives

Difficultés rencontrées

Volumétrie : difficile d'entraîner un arbre de décision en une seule fois sur les 5.5M de lignes

De nombreuses données qualitatives mixées avec données quantitatives (avec relation non linéaire avec la variable à prédire)

- ⇒ Difficile d'obtenir une régression linéaire performante sans monter très fortement le degré des features polynomiales
- ⇒ Explosion du nombre de features avec un nombre d'instances déjà élevé

L'erreur du modèle n'est pas linéaire par rapport aux valeurs de retard à prédire

- ⇒ Plus le retard est élevé plus c'est difficile de le prédire

Perspectives

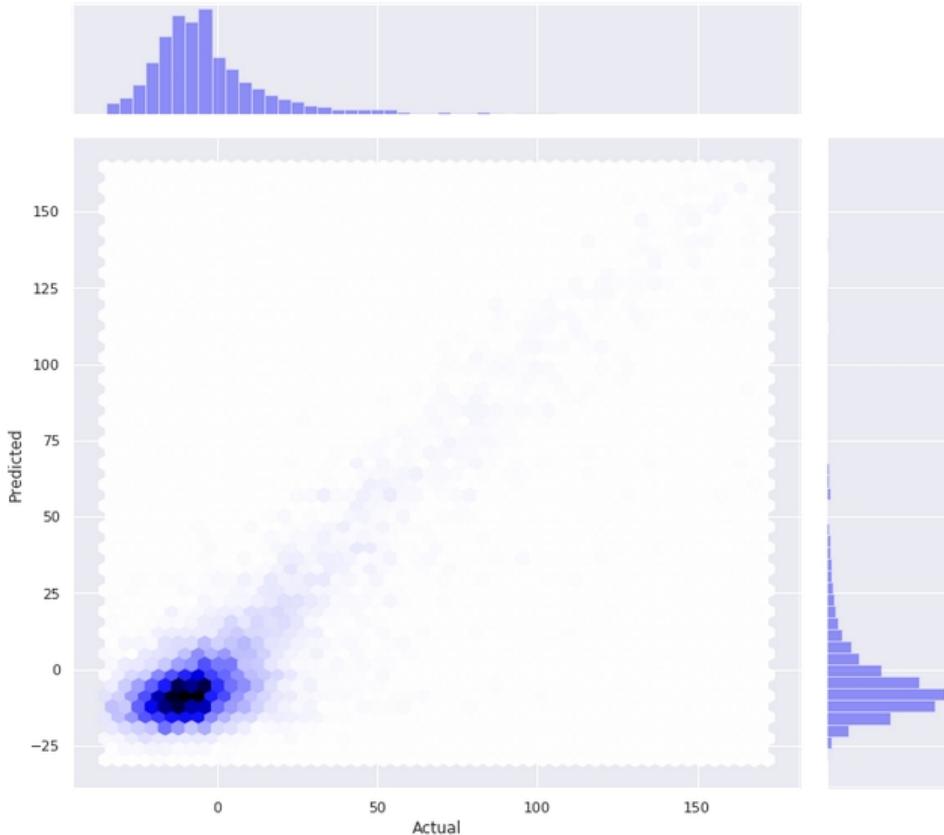
- Améliorer le modèle de random forest en effectuant du bagging sur différentes plages d'instances de données parmi les 5.5M, plutôt que de se limiter aux mêmes 80000 instances pour chaque prédicteur
- Appliquer des techniques de boosting afin de reparamétriser dynamiquement les poids associés à chaque instance (afin de mieux rectifier les erreurs élevées de prédiction pour les instances qui ont une valeur de retard élevée à prédire)
- Améliorer le modèle de régression linéaire en augmentant de façon importante les features polynomiales, en diminuant à 3 ou 4 les features de degré 1 (pour ne pas que le nombre de features polynomiales explosent), jusqu'à obtenir un overfit, puis effectuer une régularisation
- Ajouter une feature qui correspond aux périodes de vacances



Problématique 2 :
Ajouter l'information
de retard connue le
jour du départ

2ème modèle : prévision du retard à l'arrivée à partir du retard au départ, pour aider la logistique

Model 2 with departure delay, random forest : comparison actual values / predict values on test set



Features du modèle :

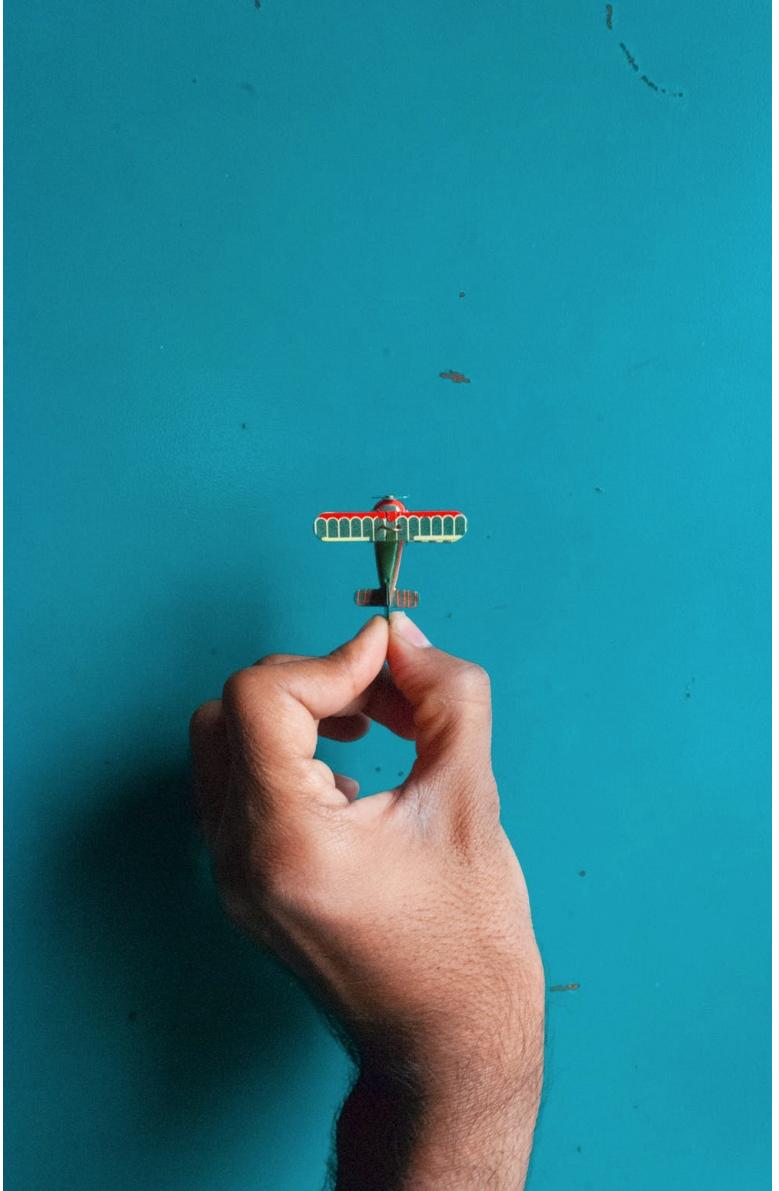
- Distance
- Nombre de vols moyen / heure / jour dans l'aéroport concerné
- Retard au départ

RMSE test : 5.71 min

Erreur moyenne 90 % du temps : 6.25 min

Erreur moyenne 10 % du temps : 23.63

	Feature name	Feature importance	Cumulated feature importance
0	DEP_DELAY	82.67%	82.67%
1	DISTANCE	10.26%	92.94%
2	NBFLIGHTS_FORDAYHOUR_FORAIRPORT	7.06%	100.00%

A photograph of a person's hand holding a small, colorful toy biplane. The toy has red wings with yellow stripes, a red body, and a green tail. It is held between the thumb and forefinger of the hand, which is positioned vertically. The background is a solid teal color.

Interface web
hébergée sur AWS

Flight characteristics

Scheduled Departure time (HHMM)
1000

Scheduled Arrival time (HHMM)
1800

Scheduled Elapsed time (minutes)
60 - +

Day of Month (1-31)
1

Day of week (1-7)
1

Distance (miles)
1000 - +

Month (1-12)
1

Origin airport
ABE

Openclassrooms training projet 4 : predicting flight delays (François BOYER)



Arrival delay prediction
1.61 minutes

Display debug information

Input data

CRS_ARR_TIME	CRS_DEP_TIME	CRS_ELAPSED_TIME	DAY_OF_MONTH	DAY_OF_WEEK	DISTANCE	MONTH	ORIGIN	DEST	UNIQUE_CARRIER	NBFLIGHTS_FORDAYHOUR_FORAIRPORT	NBFLIGHTS_FORDAY_
0	1080	600	60	1	1000	1	279	168	7	200	

Transformed data passed to the model

CRS_ARR_TIME	CRS_DEP_TIME	CRS_ELAPSED_TIME	DAY_OF_MONTH	DAY_OF_WEEK	DEST	DISTANCE	MONTH	NBFLIGHTS_FORDAYHOUR_FORAIRPORT	NBFLIGHTS_FORDAY_FORAIRPORT	ORIGIN	1
0	0.5594	-0.7129	-1.1130	-1.6784	-1.4835	0.0772	0.2584	-1.6192	8.4795	-0.4733	1.4635

URL : <http://3.20.50.249:8501/>

