Novembre 2018

# Retards de Vols

Projet OpenClassrooms

Par Xavier Montamat

# Problématique

Le but du projet est de pouvoir estimer le retard probable d'un trajet en avion.

Nous possédons une base de données sur de nombreux vols américains de l'année 2016.

Il nous faut donc trouver une corrélation entre les données de vols disponibles et les retards constatés

# Axe d’approche envisagé

- Identifier les raisons d’un retard
- Selectionner les plus pertinentes
- Appliquer un algorithme de régression
- Raffiner le modèle en fonction des résultats

## Dangers à éviter

- Le Data Leak
- Overfitting du modèle

## Contexte

- Travail pour une compagnie aérienne
- Accessible sous forme d’API

# Plan de réalisation

## Exploration

- Données disponibles
- Donnée cible
- Exploration des features
- Choix des features

## Modélisation

- Test de modèle
- Transformations
- Amélioration du modèle
- Fitting des resultats

## Modèle final

- Transformations effectuées
- Resultats scores et fitting
- API

# Données

## Données disponibles

- 5.6M de vols
- Année 2016 exclusivement
- Etats-Unis exclusivement
- 55 états & 316 aéroports
- 12 compagnies

## Données manquantes

- Sur 10 features cruciales sélectionnées
- 80K lignes de données manquantes supprimées

  (Soit 1.5% du total)

- 116K duplicates supprimés

  5.44M de vols restants

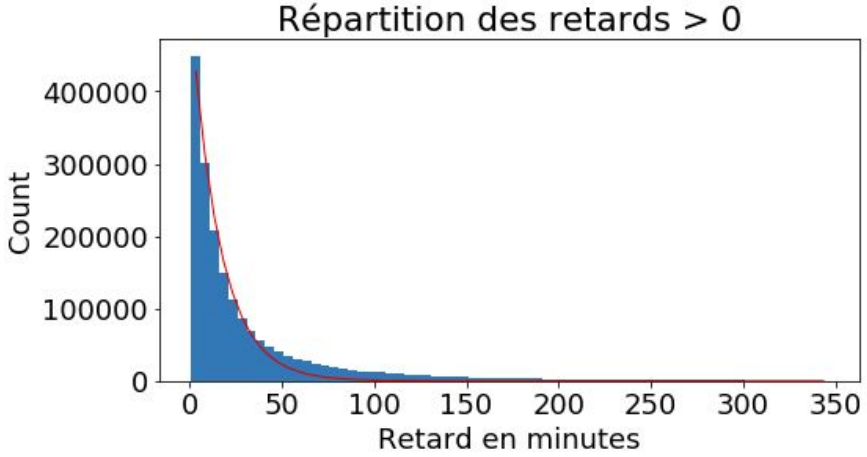
# Variable cible : le retard

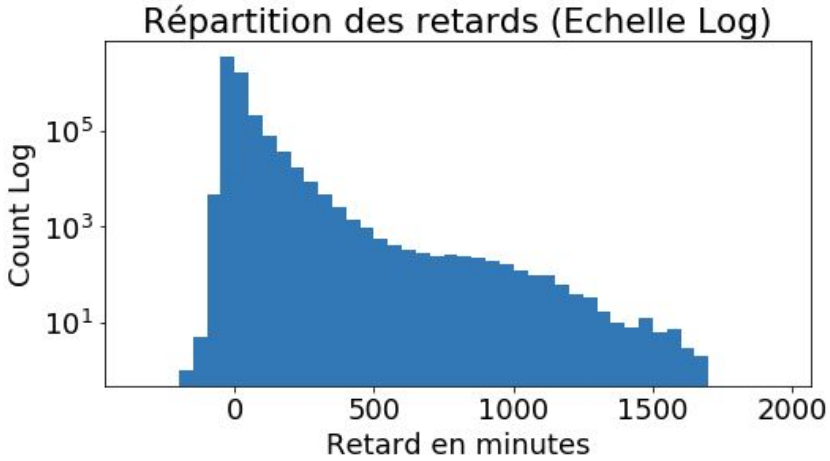
26% des vols ont un retard de plus 5 minutes

Quelques valeurs extrêmes (>20H de retard)

## Retard absolu à l'arrivée

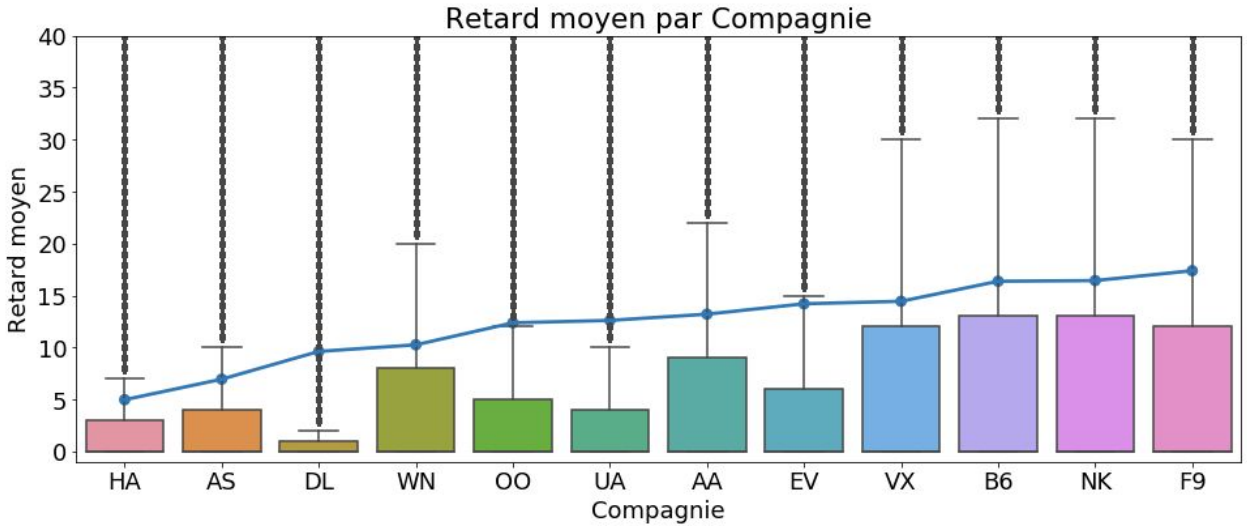
- Retard au départ moins pertinent
- Retard négatif inutiles ( =0 )

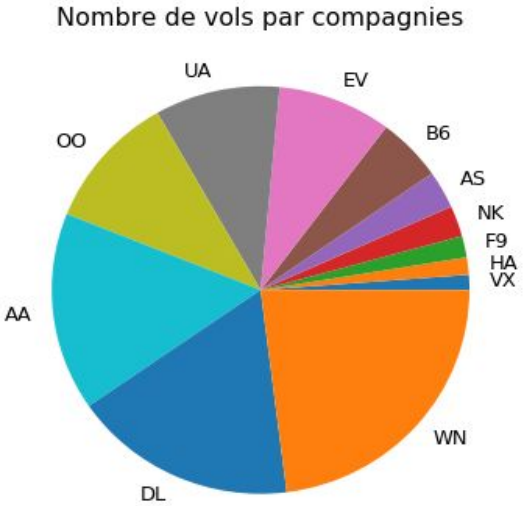


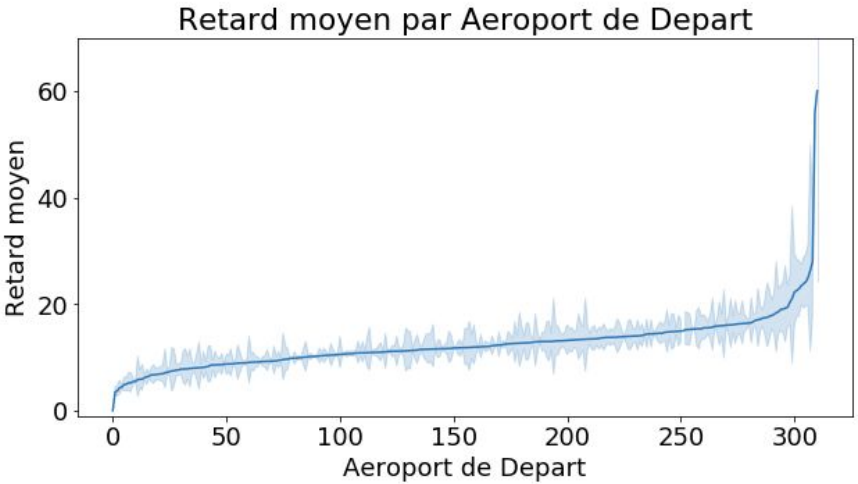
# Exploration - Compagnies aériennes

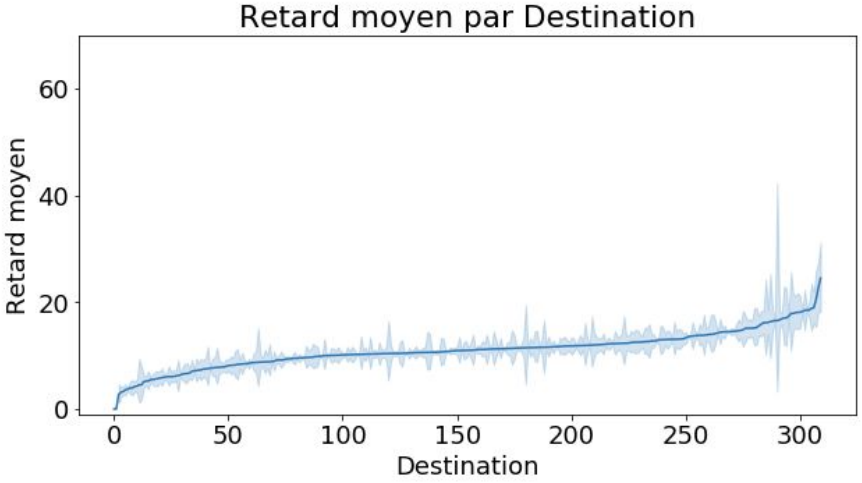
## Distribution du retard par compagnie



# Retards par Aeroports

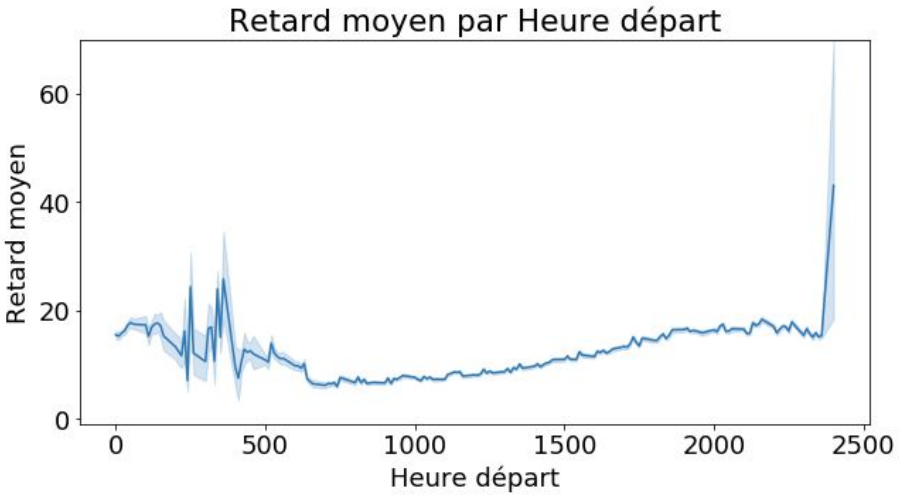


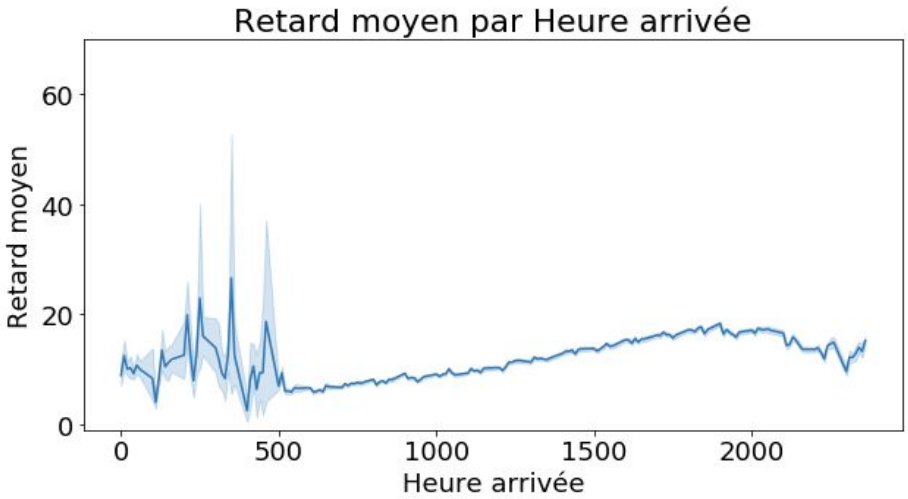




Retard moyen varie en fonction de l'aéroport

Départ / destination

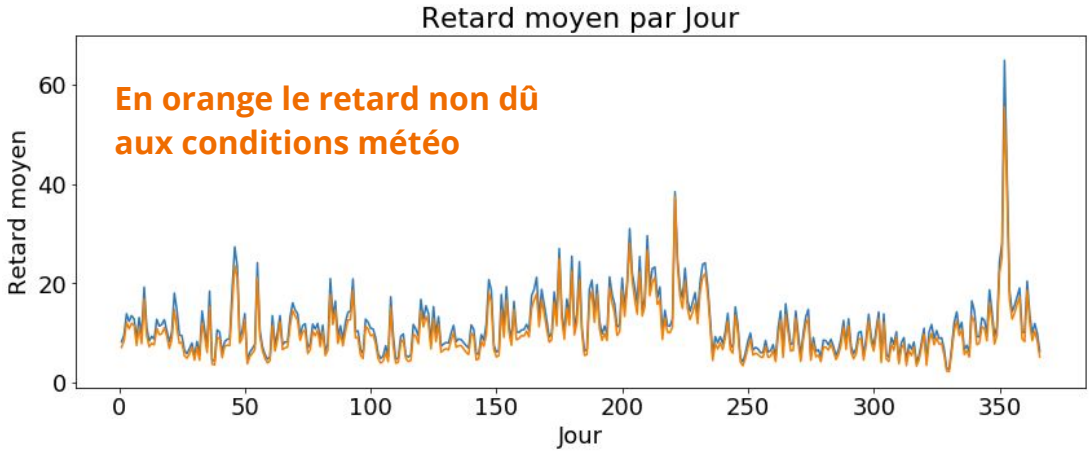
# Retard par heure de la journée

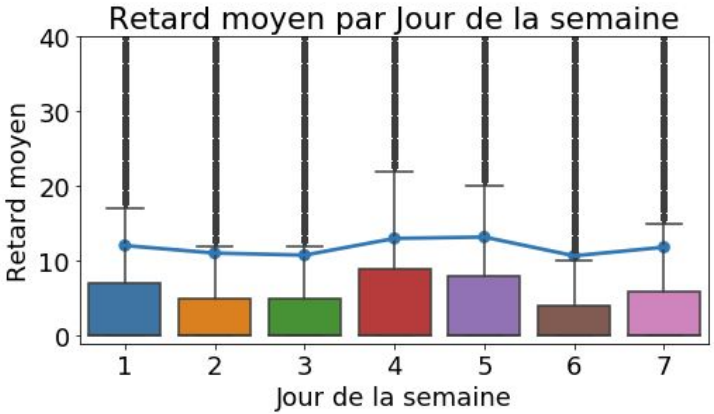



- Plus élevés en fin de journée et durant la nuit

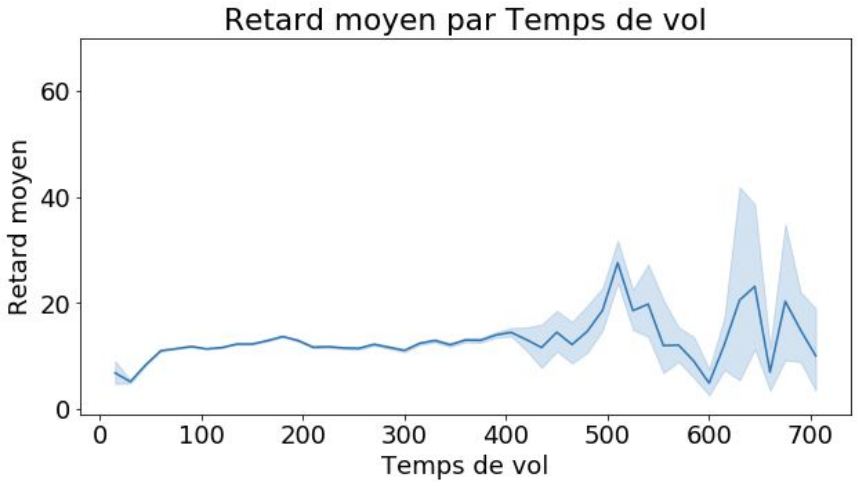
# Retards sur l'année



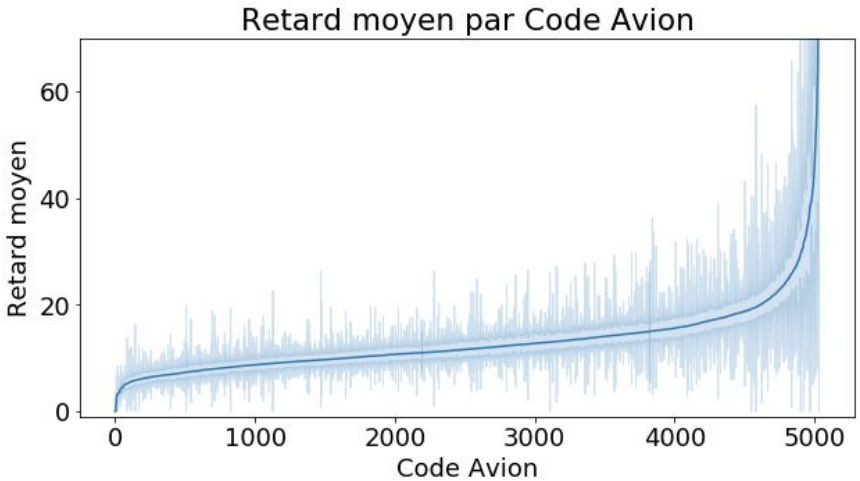




- Retards plus élevés en été
- Fort pic avant  noel
- Variations importantes tout au long de l'année

# Autres




Durée de vol pas vraiment corrélée au retard




Le code d'avion est une donnée intéressante si disponible

Causes possibles : Avion vieillissant / trop gros

# Correlations

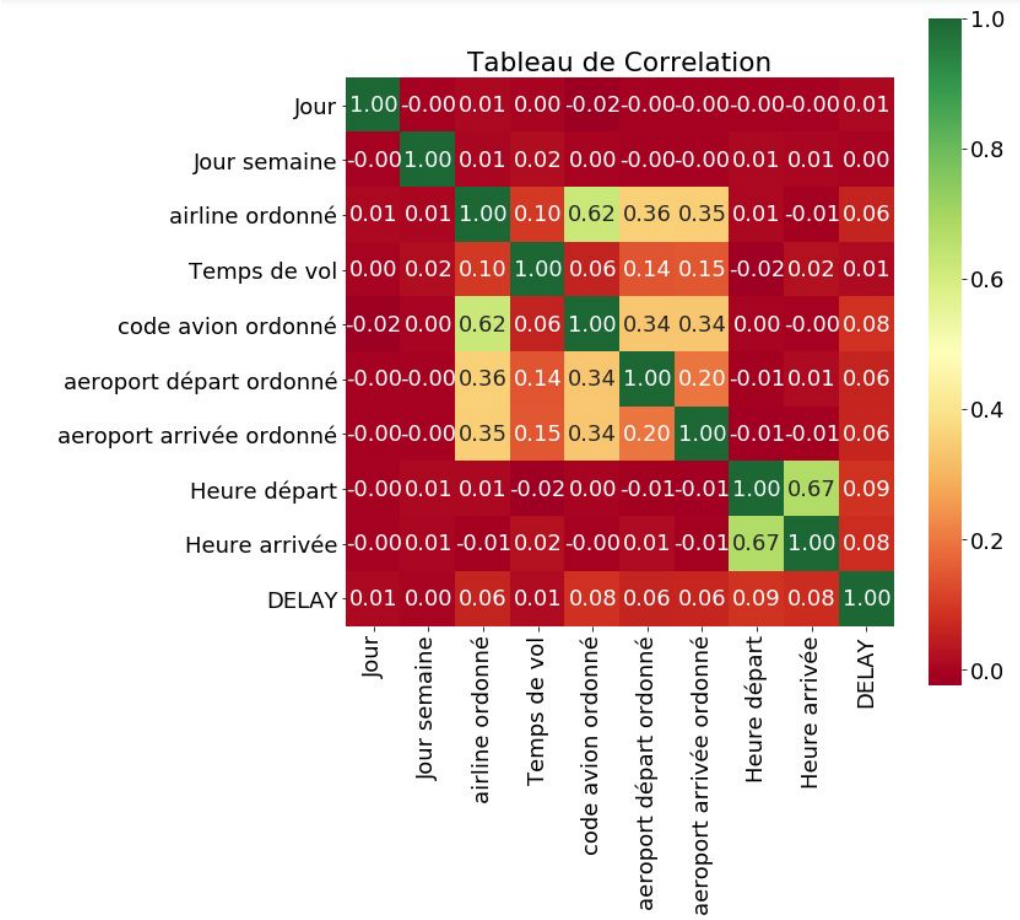
Aperçu rapide des corrélations (pearson) entre nos features

Et avec la variable cible

**Resultats**

Les corrélations au retard sont assez faibles

Entre 0.01 et 0.09



# Modélisation de la régression

Après avoir sélectionné les données les plus pertinentes différentes lors de l'exploration

Je test différents modèles et compare les résultats

Certaines données doivent êtres transformées pour être exploitables par nos modèles

# Méthode de test

## Division du dataset

Séparation

- Données entrainement (75%)
- Données de test (25%)

## Choix de l'hyper paramètre

Meilleur alpha trouvé par GridSearchCV (cv=5)

## Modèles utilisés

Linéaire
LASSO
RIDGE

## Scores

R squared
Valeur moyenne absolue
Erreur carré moyenne

# Premiers tests

## Régression linéaire

5-Folds CV score : 0.011

R squared: 0.0189

Valeur moyenne absolue : 17

RMSE: 38

## LASSO

5-Folds CV score : 0.011

R squared: 0.0189

Valeur moyenne absolue : 17

RMSE: 38

## Ridge

5-Folds CV score : 0.011

R squared: 0.0189

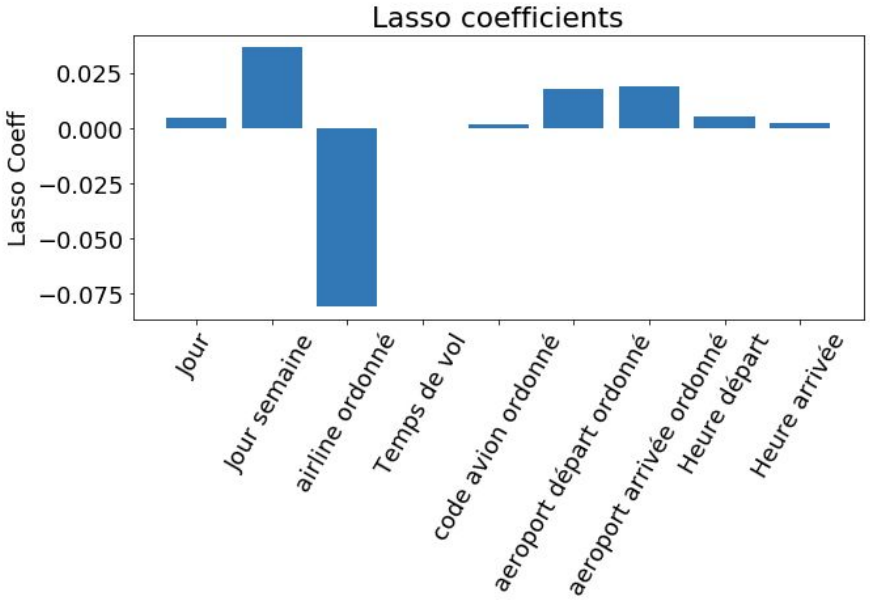
Valeur moyenne absolue : 17

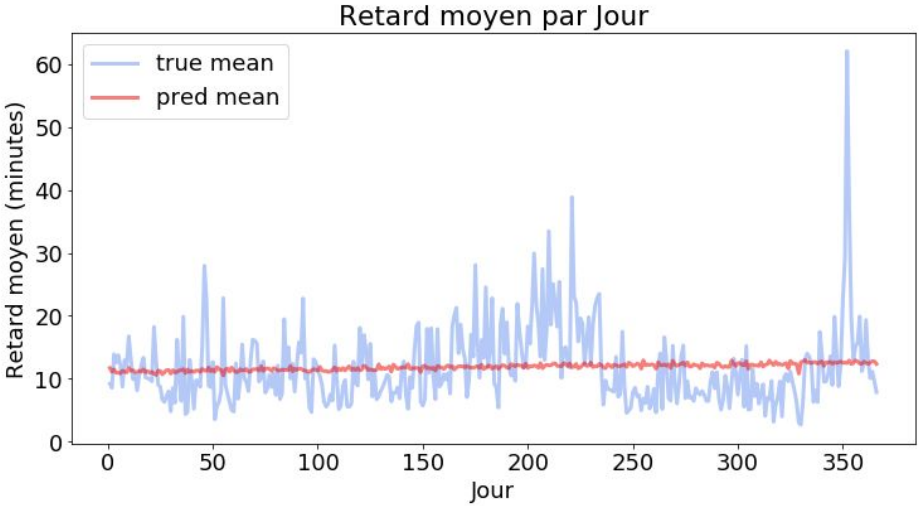
RMSE: 38

## Résultats identiques

## Scores bas

# Détail des résultats




**Temps de vol inutile**

**Fitting non optimal sur certaines features**

# Améliorations

## Filtrage des outliers

Retards extrêmes probablement non prédictibles

Sur 1.44 M de vols avec un retard > 5 min

99% quantile de ces vols = 290 min

Retard max limité à 290 min

## Ordonnancement des jours année et semaine

## Suppression du temps de vols

# Nouveaux scores

## Régression linéaire

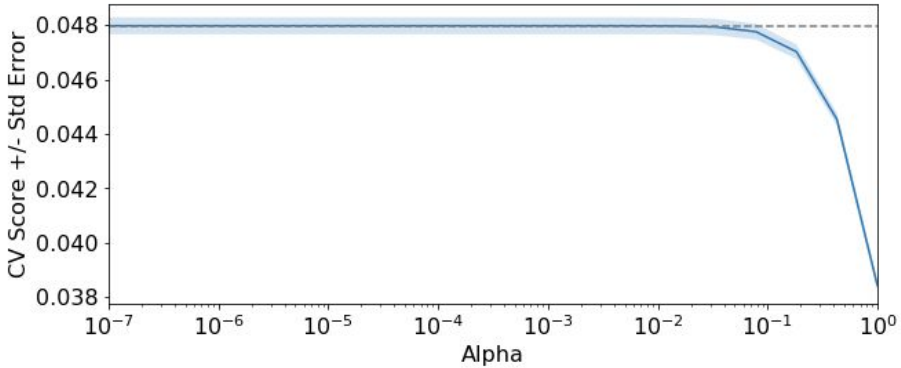
5-Folds CV score : 0.04 (x4)

R squared: 0.047 (x3)

Valeur moyenne absolue : 17 (-0.8)

RMSE :  31.7 (-7.3)

## Meilleur alpha ridge




## Ridge et Lasso toujours identiques

# Ridge

## Recherche du meilleur alpha

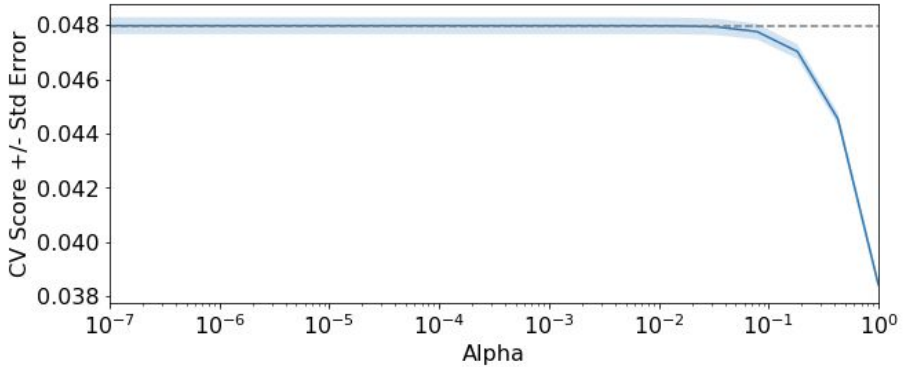
GridSearchCV : Meilleur alpha de 1e-07

## Scores

R squared: 0.0482
Mean absolute value : 16.16
RMSE :

Les scores sont donc les mêmes!

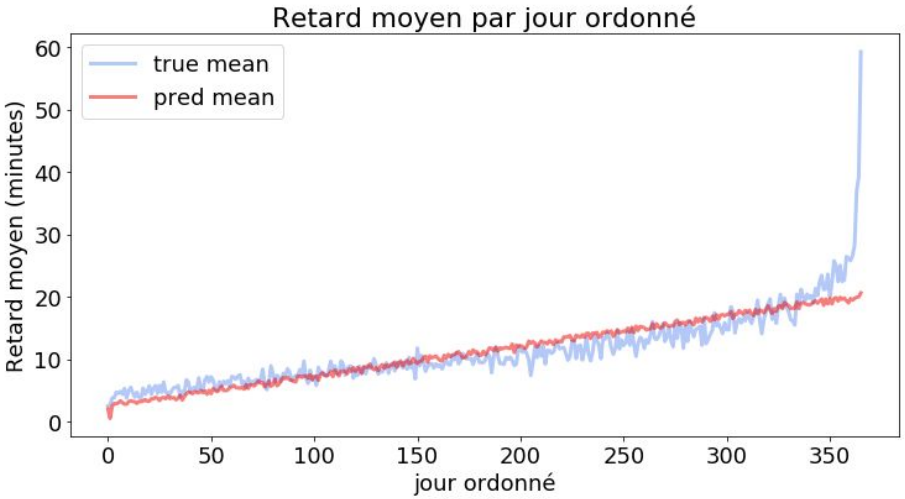
## Recherche du meilleur alpha



# Analyse du fitting

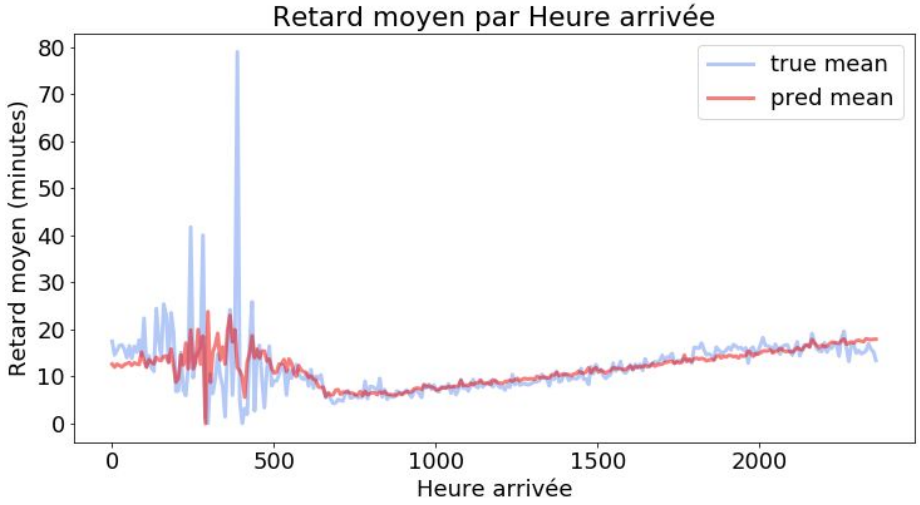
## Jour de l'année

Encore trop linéaire par rapport à la valeur réelle




## Heure d'arrivée

Fit assez bon entre prédiction et réalité



# Modèle Final

Pour le modèle final nous allons à nouveau transformer nos features par degrés polynomiaux

Pour obtenir un meilleur fitting

Nous utiliserons ensuite le modèle de ridge et evaluerons les scores

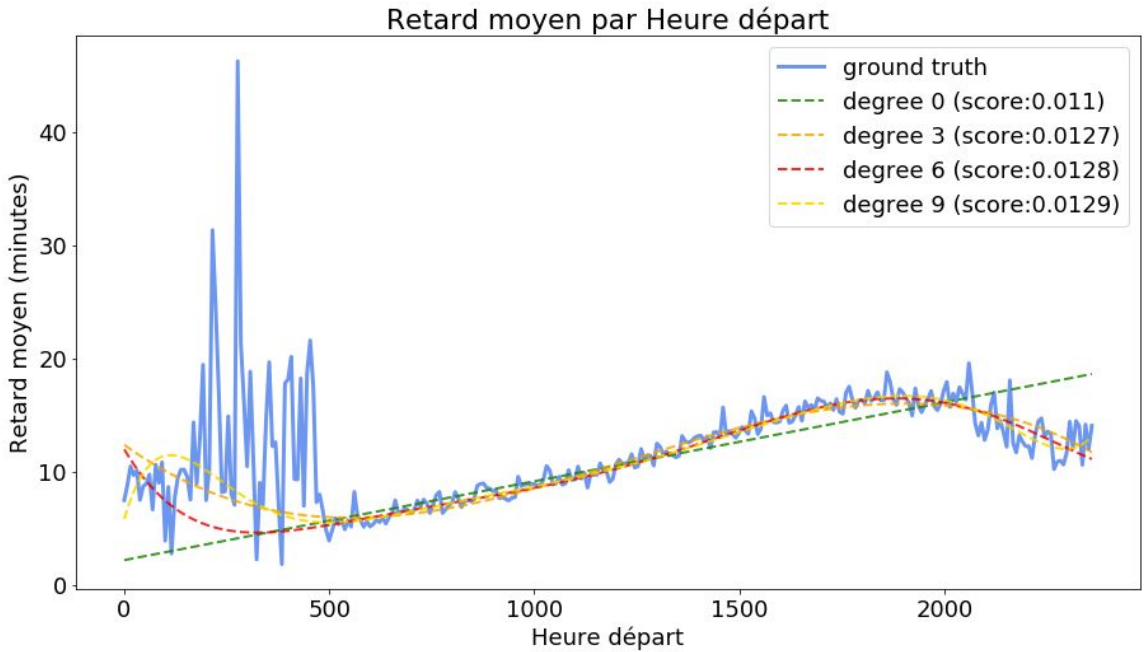
# Transformation des données

## Transformation polynomiale

Analyse du score par degré de transformation

Degrés de transformation

- Heure départ = 3°
- Heure d'arrivée = 3°



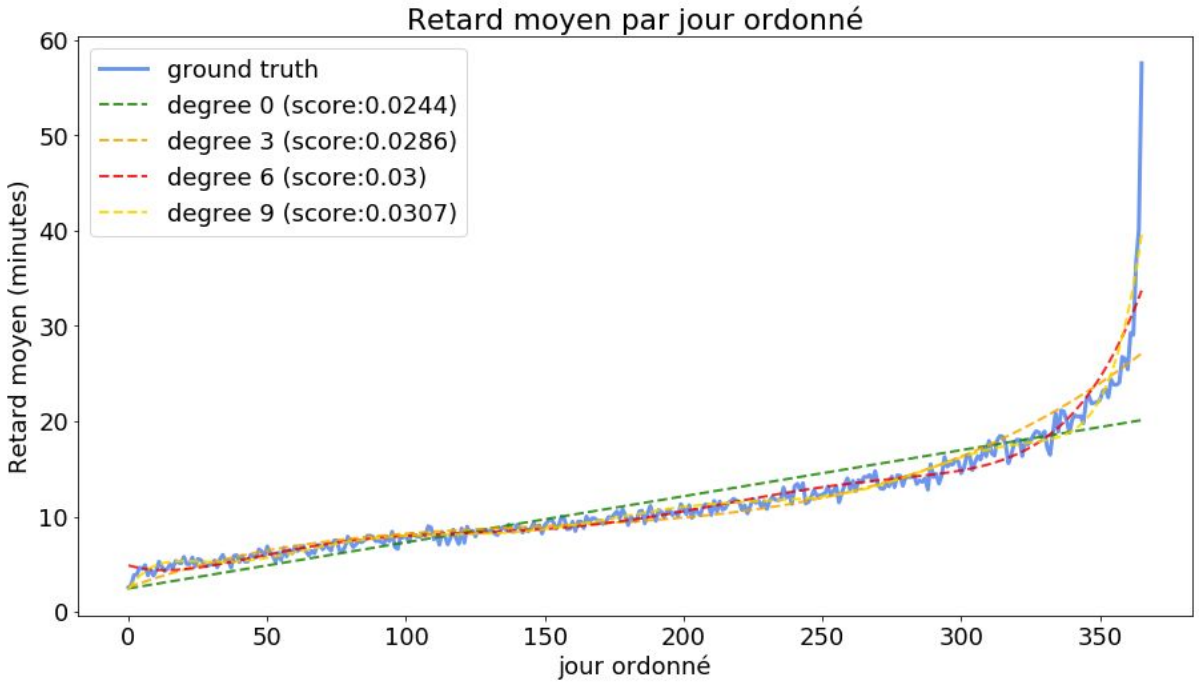
# Transformation des données

## Données de catégories

(Ordonnées par la moyenne sur un train set)

Degrés choisis

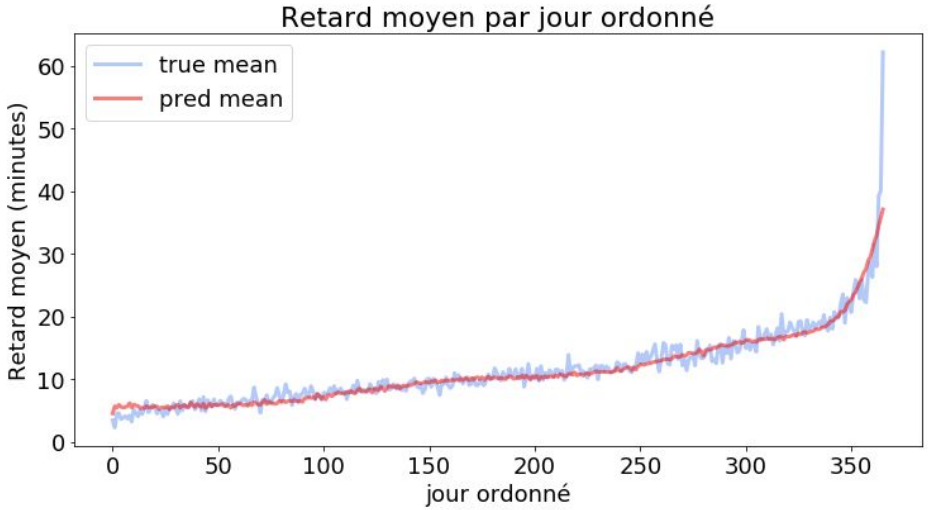
- Code Avion = 3°
- Jour de l'année = 9°
- Jour semaine = 1°
- Aéroport Dest = 3°
- Aéroport Départ = 3°
- Compagnie = 3°



# Fitting final

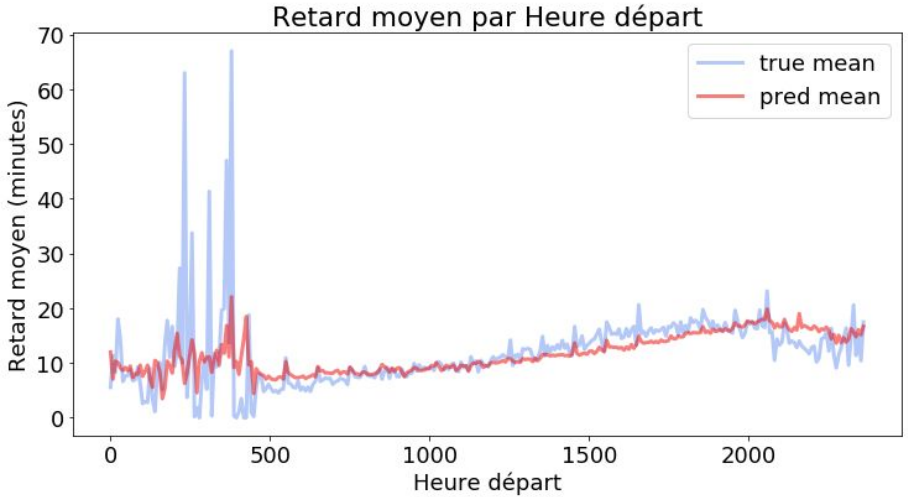
## Jour de l'année

Fit bien meilleur




## Heure d'arrivée

Fit légèrement meilleur



# Performances final

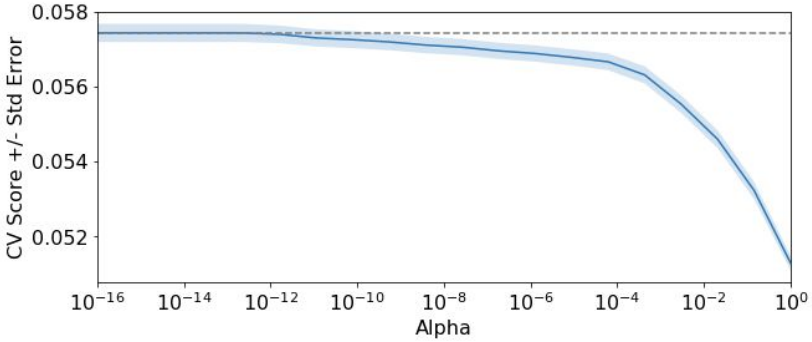
31 features après transformation

## Scores Ridge

5-Folds CV score: 0.49 (x1.2)
R squared: 0.0574 (x1.2)
Mean absolute value : 15.98 (-1.3)
RMSE : 31.5 (-0.2)

## Scores en hausse

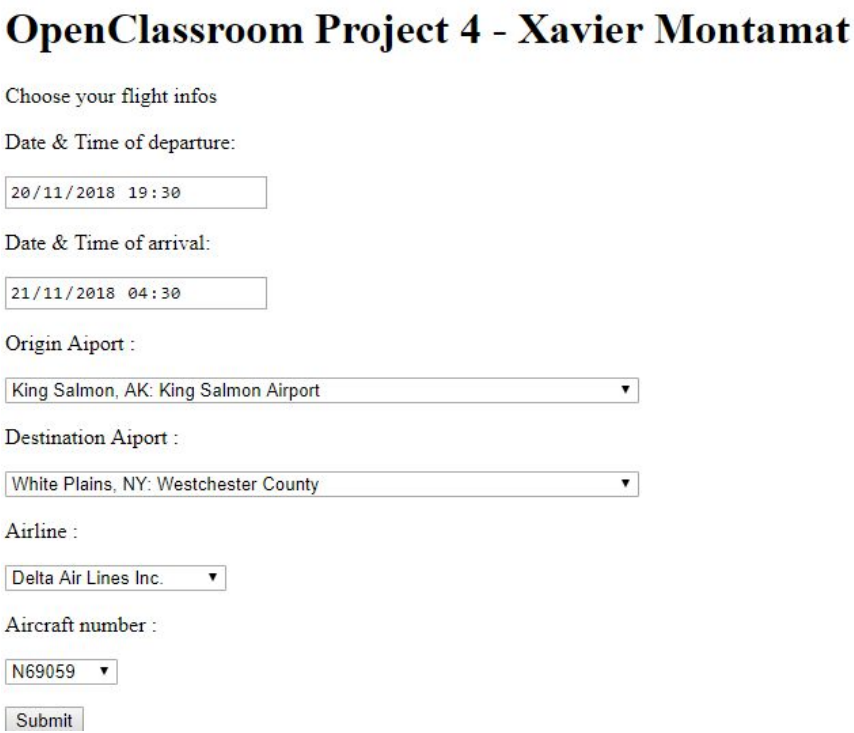
## Recherche du meilleur ridge alpha



# Conclusion

Bon fitting
Encore beaucoup de variations
Retard aléatoire

# API du modèle



# Axes d’amélioration

Récupérer données différentes
Données sur plus d’années
Modèles plus complexes