

Parcours DataScientist Projet 4:

Anticipez le retard de vol d'avions



Jérôme d'Harveng

Mentor: Pierre Comalada

OpenClassrooms avril 2019

Table des matières

PART I: Contexte de l'analyse de données

PART II: Nettoyage des données

PART III: Analyse exploratoire

PART IV : Pistes de modélisation

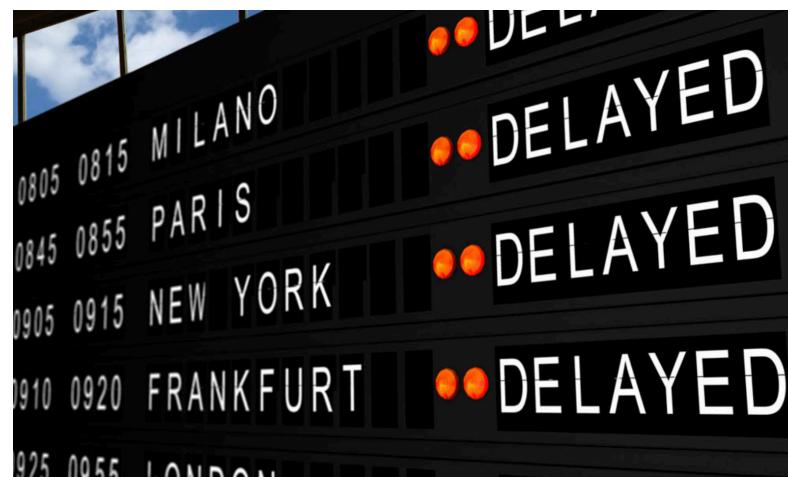
PART V: Modèle final et performances

PART VI: Conclusions



Part I:

Contexte de l'analyse



Problématique

- *AirData:*

- * *But final:* via une API

à partir des paramètres d'un futur vol

=> *prédiction retard*

- * *Informations:*

une **Base de données**

reprenant des vols sur 2016 avec leurs paramètres

- * *Points importants:*

Régression linéaire et évaluation des modèles



Interprétation et déductions



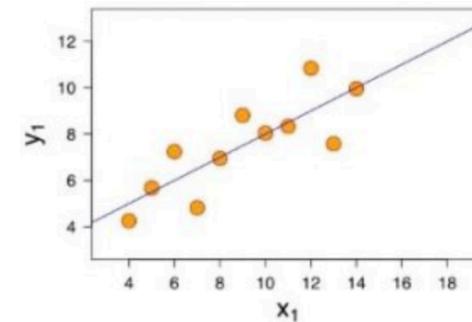
- **Brève analyse de la littérature** (annexes)

- **Etude exploratoire** nécessaire

=> se focaliser les paramètres clé

- **Régression Linéaire:**

- * standard
- * régularisation Ridge
- * LASSO
- * Cross-validation



- **Evaluation du modèle**

- * Entraînement (training) et Test
- * types d'erreur : RMSLE

Part II:

Nettoyage des données



Nettoyage de la Base (CLEANING : 1)

- **Valeurs manquantes** => différentes options

- * **Remplacer certaines valeurs:**

'LATE_AIRCRAFT_DELAY','NAS_DELAY','WEATHER_DELAY','CARRIER_DELAY','SECURITY_DELAY' => 0 min
AIR_TIME','ARR_DELAY' : remplacés par -999 pour les vols non-annulés

- * **Supprimer certaines lignes:**

'ARR_DELAY' for Non-Cancelled flights

- * **Supprimer certaines variables:**

'UNIQUE_CARRIER','AIRLINE_ID','CARRIER' => redondant => GARDER UNIQUE_CARRIER
'ORIGIN_STATE_FIPS', 'ORIGIN_STATE_NM', 'ORIGIN_WAC' / 'DEST_AIRPORT_SEQ_ID', 'DEST_CITY_MARKET_ID'

- **Pas de doublons observés dans la base**



- **Conversion type de données:**

- * **float vers int** (ex. : 'YEAR','DAY_OF_MONTH','DAY_OF_WEEK','CRS_DEP_TIME','DEP_DELAY','CRS_ARR_TIME')
 - * **string vers variable catégorielle** : UNIQUE_CARRIER

- **Valeur erronée:**

Entrée [51]: # first clean line with '16-03-04'
df_flights.YEAR.value_counts()

Out[51]: 2016 5143195
2016 478973
16-03-04 1

Nettoyage de la Base (CLEANING : 2)

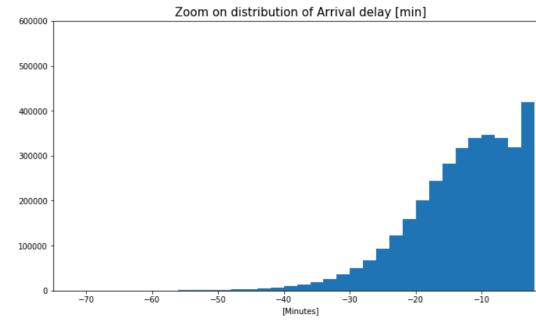
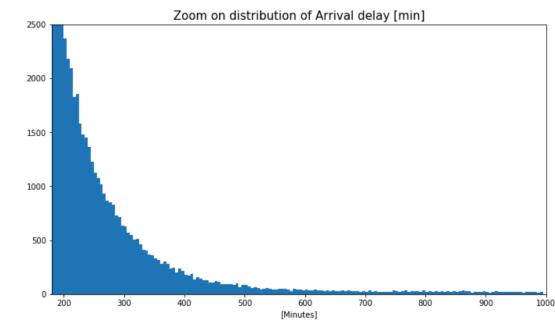
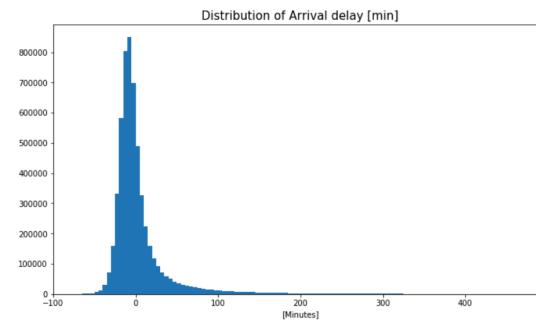
- *Pour la partie REGRESSION LINEAIRE*

- * Suppression de tous les vols annulés (`CANCELLED ==1`)
- * Suppression des outliers pour ARR_DELAY (lignes) :

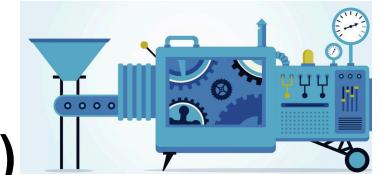
- min = -152 minutes / max = 2142 minutes

- quantile_{1%} = -35 minutes

- quantile_{99%} = 172 minutes



Feature Engineering (pour l'exploration)



- **Variable :** ***DEP_HOUR***

= round(CRS_DEP_TIME/100)



- **Variable :** ***ARR_HOUR***

= round(CRS_ARR_TIME/100)

```
df_flights['DEP_HOUR'] = (df_flights['CRS_DEP_TIME']/100).round(0).astype('int')
df_flights['ARR_HOUR'] = (df_flights['CRS_ARR_TIME']/100).round(0).astype('int')
```

Part III:

Analyse Exploratoire

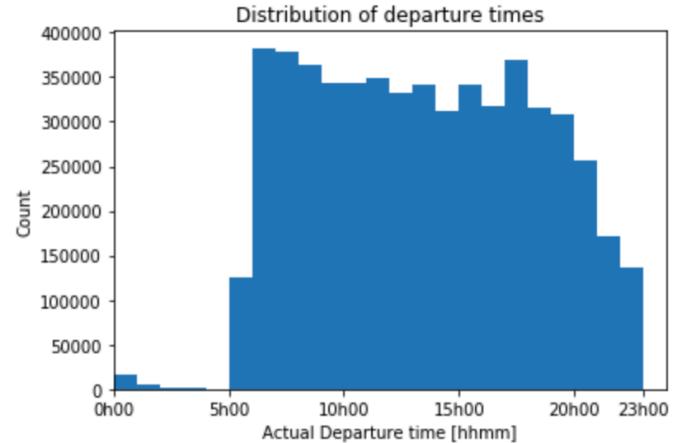


Analyse Univariée (1)

Observations:

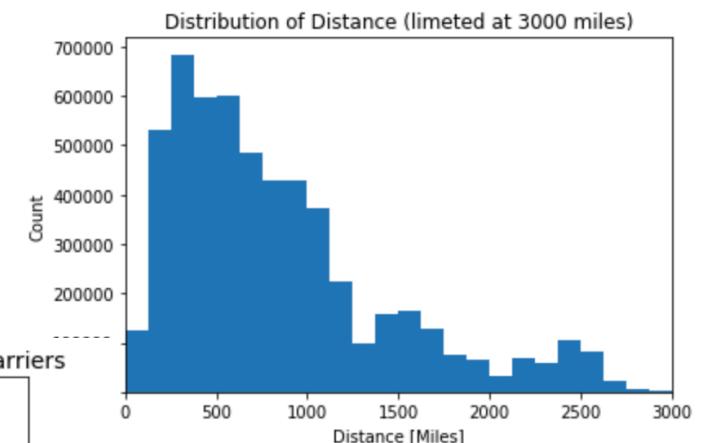
- Departure time:

la majorité des départs sont entre 5h et 23h



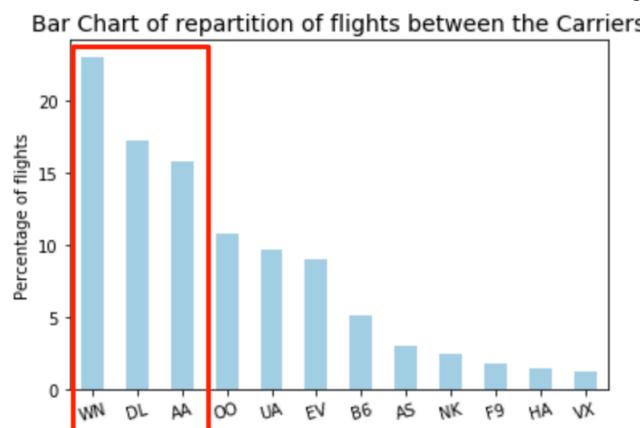
- Distance:

- * 50% des vols moins de 678 miles
- * 75 % des vols moins de 1091 miles
- * min : 28 Miles (même état / AIR_TIME: coherent)
- * max: 4983 miles



- ARR DELAY:

- * 25% des vols moins de 6 min de retard (à l'arrivée)
- * 50% moins 15min
- * 75% moins de 39min



- UNIQUE CARRIER

- * WN: Southwest Airlines
- * DL: Delta Air Lines
- * AA: American Airlines

Analyse Univariée (2)

- Vols annulés :

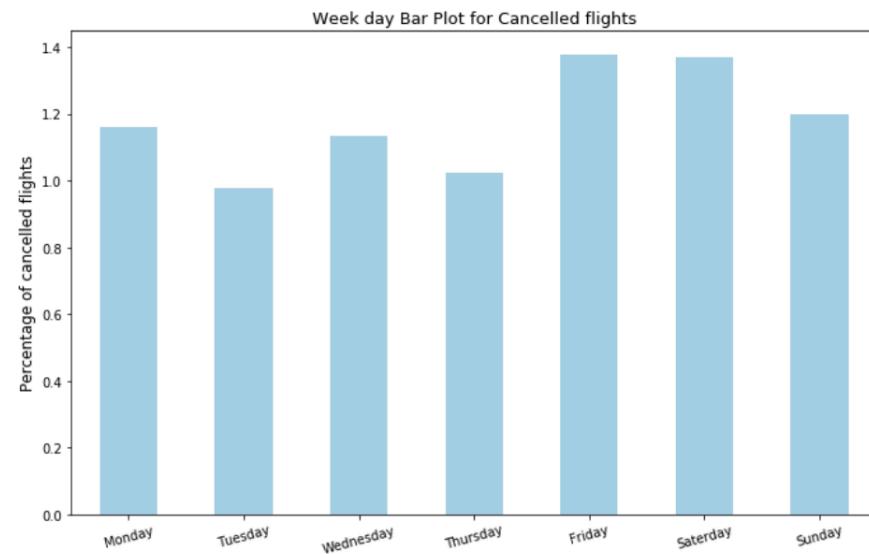
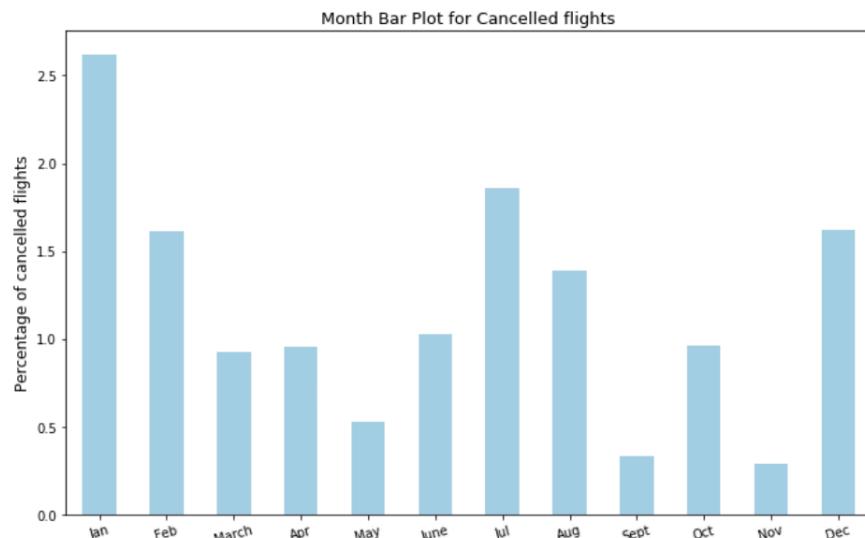
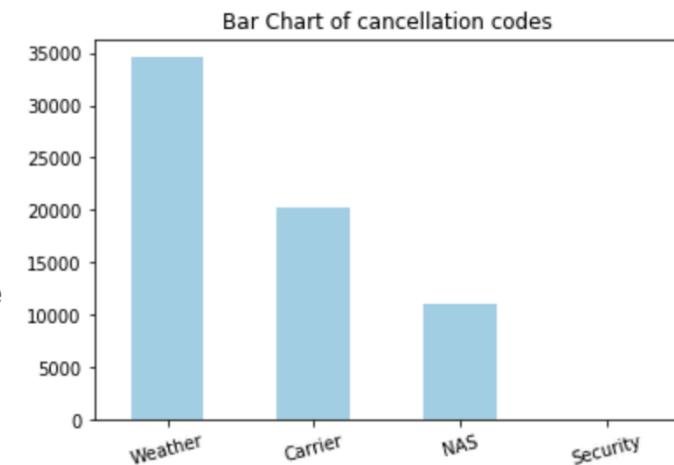
- * 1,17% de vols annulés dans la base de données
- * principalement pour raison de météo et peu en raison sécurité

Sur l'année:

- * Plupart des vols annulé : jan/juill.
- * Moins d'annulation en sept / nov

En semaine:

- * plus d'annulation vendredi et samedi

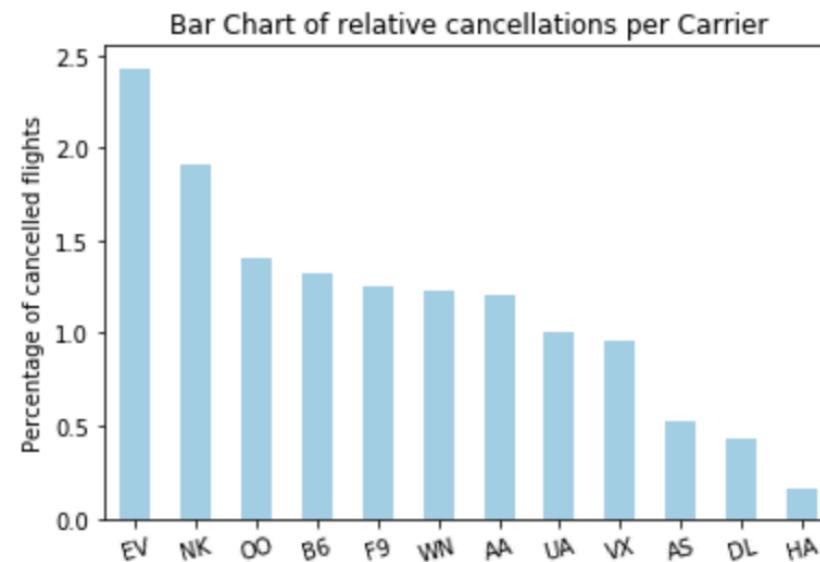


Analyse Univariée (3)

- Vols annulés :

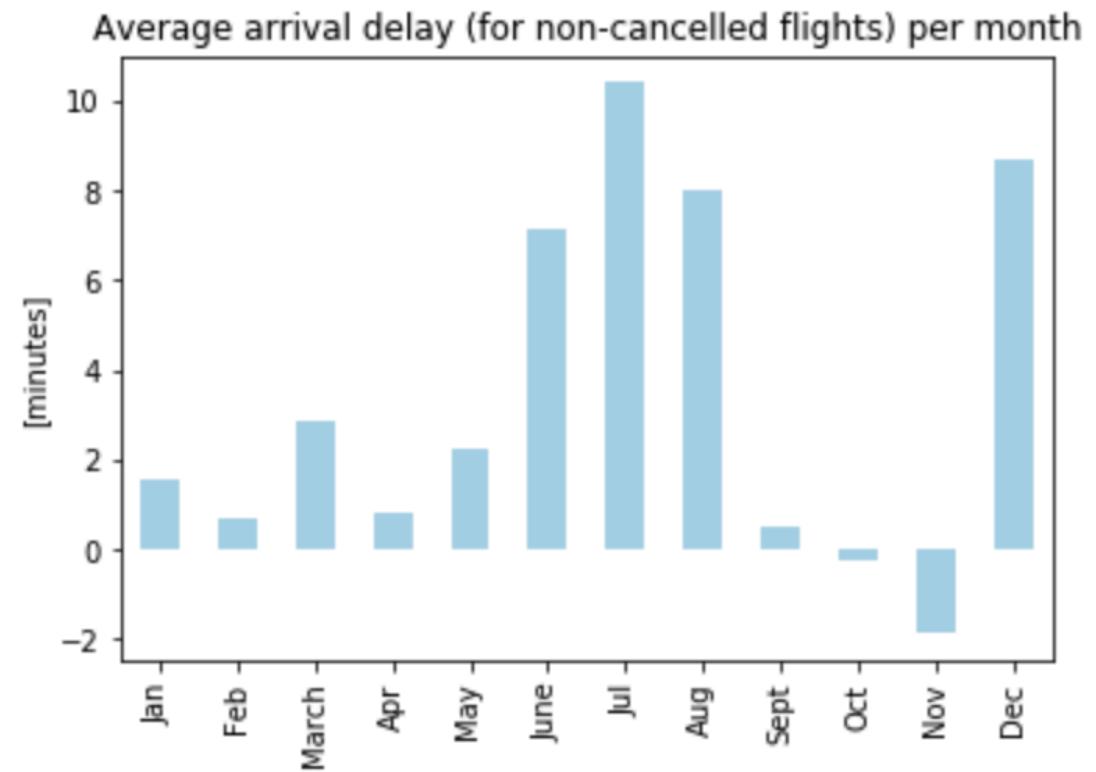
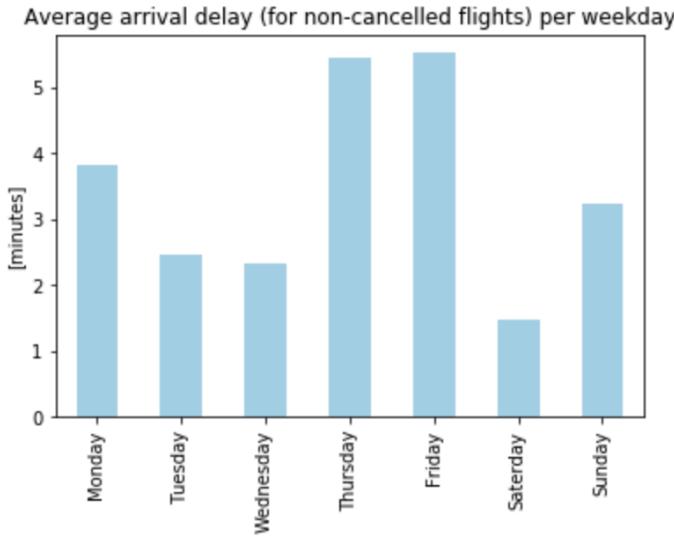
* Les compagnies avec le taux d'annulation le plus élevé :

- EV : ExpressJet Airlines
- NK : Spirit Airlines



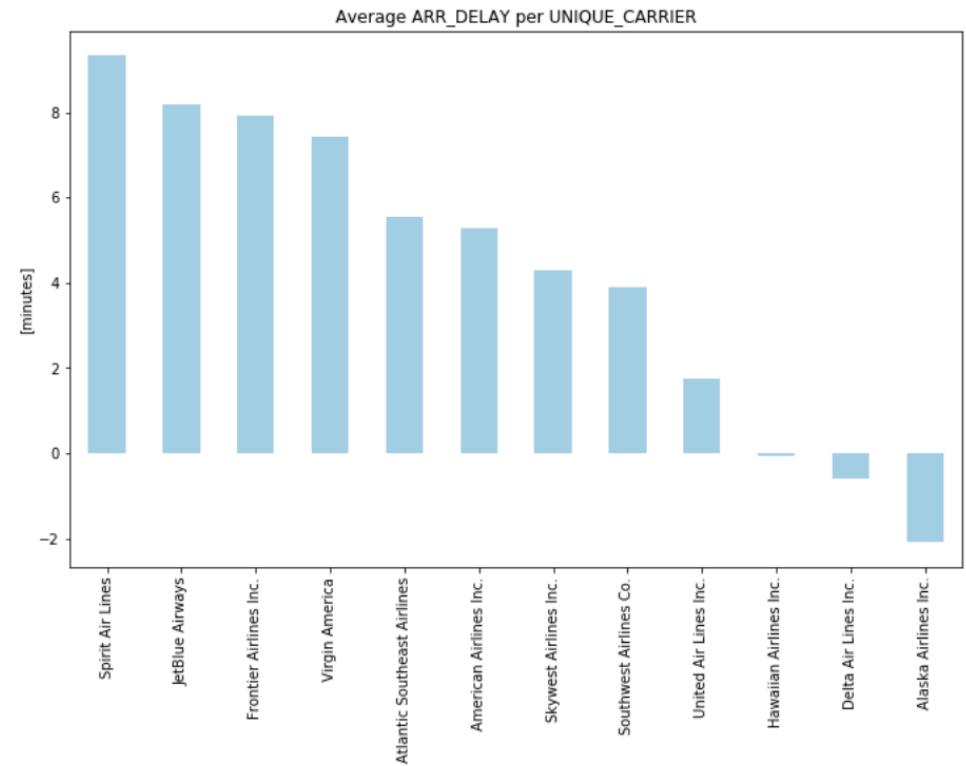
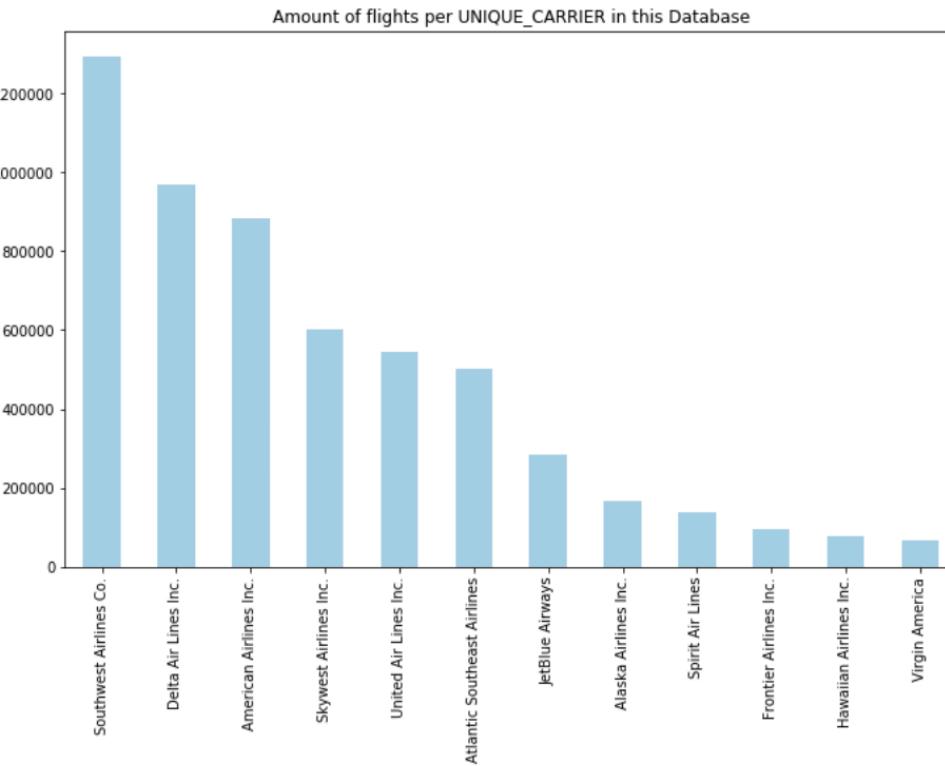
Analyse Multivariée (1)

Délai moyen à l'arrivée:



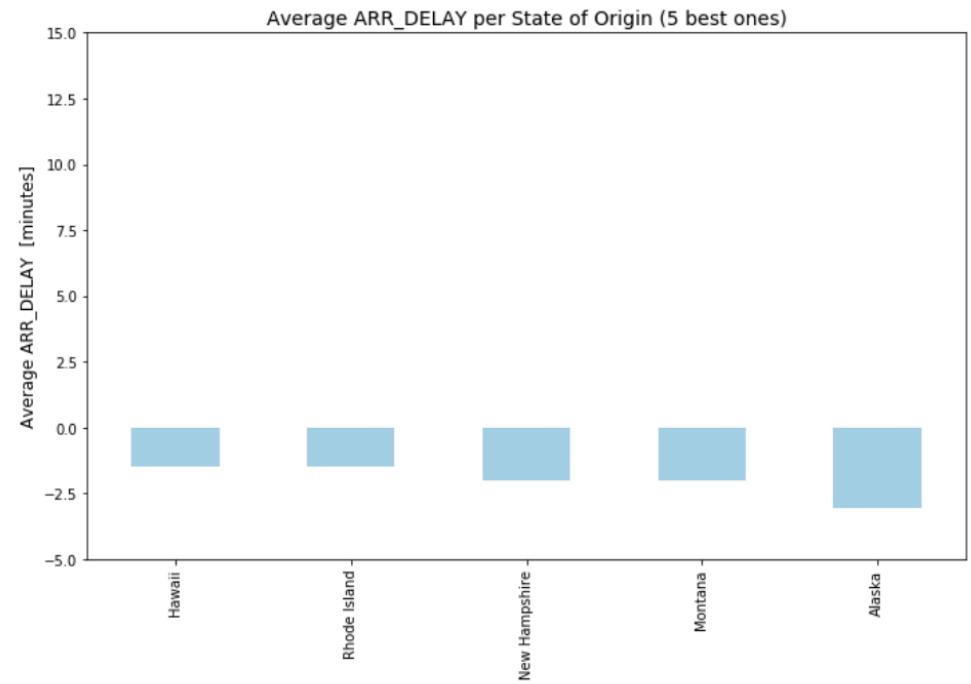
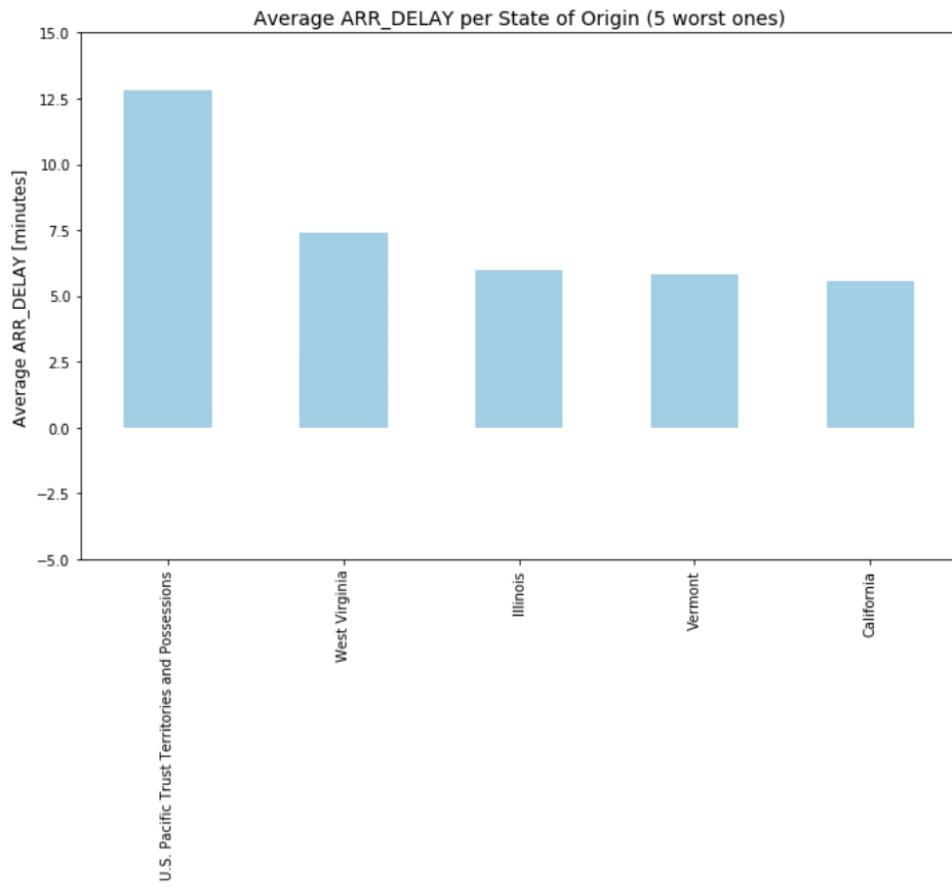
Analyse Multivariée (2)

Délai moyen à l'arrivée en fonction du Carrier



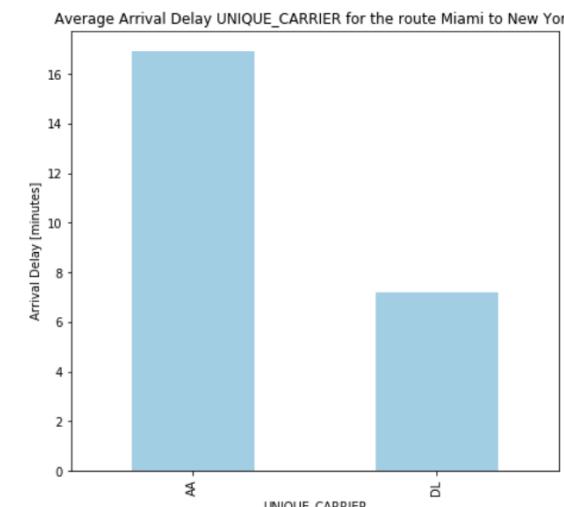
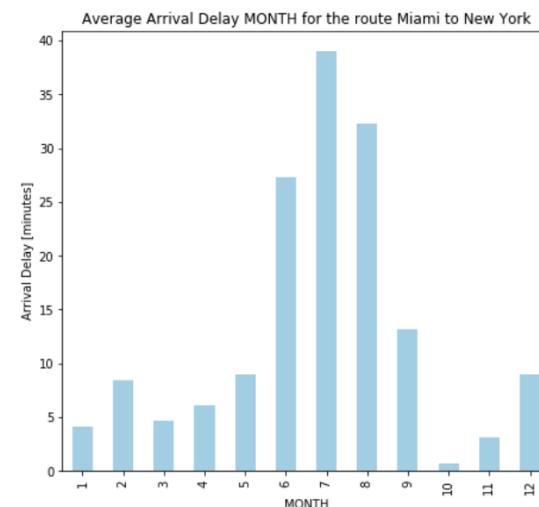
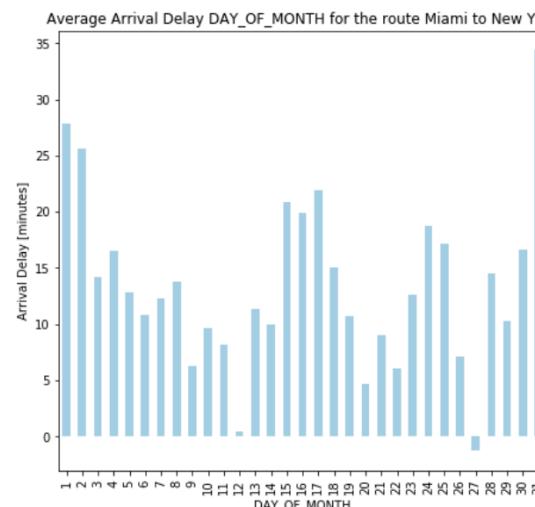
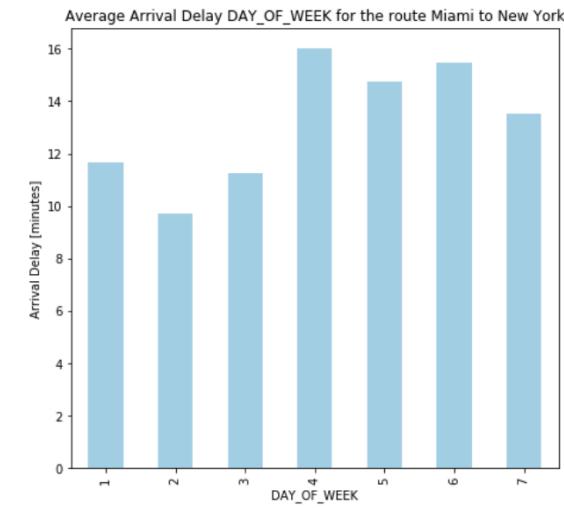
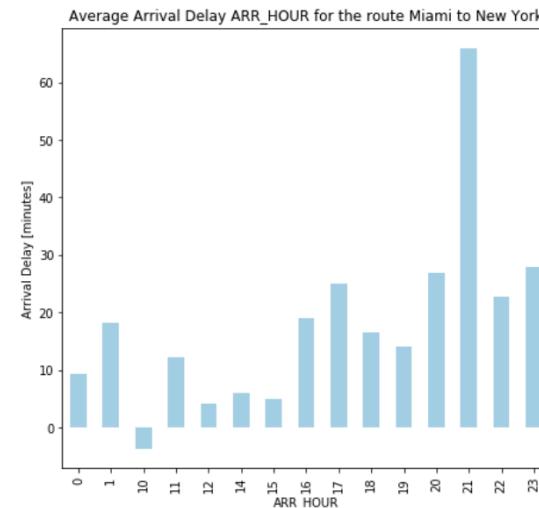
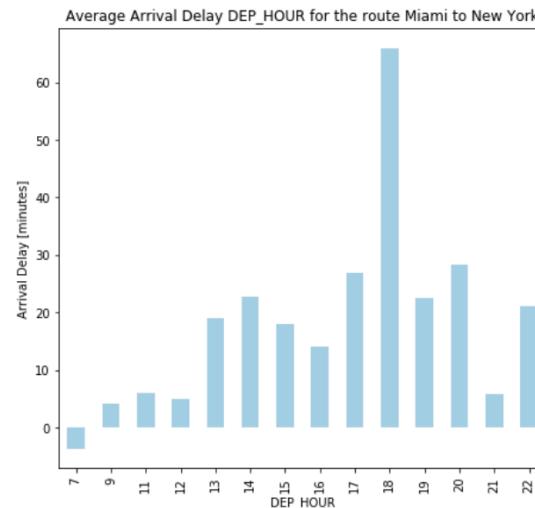
Analyse Multivariée (3)

Délai moyen à l'arrivée en fonction de l'origine



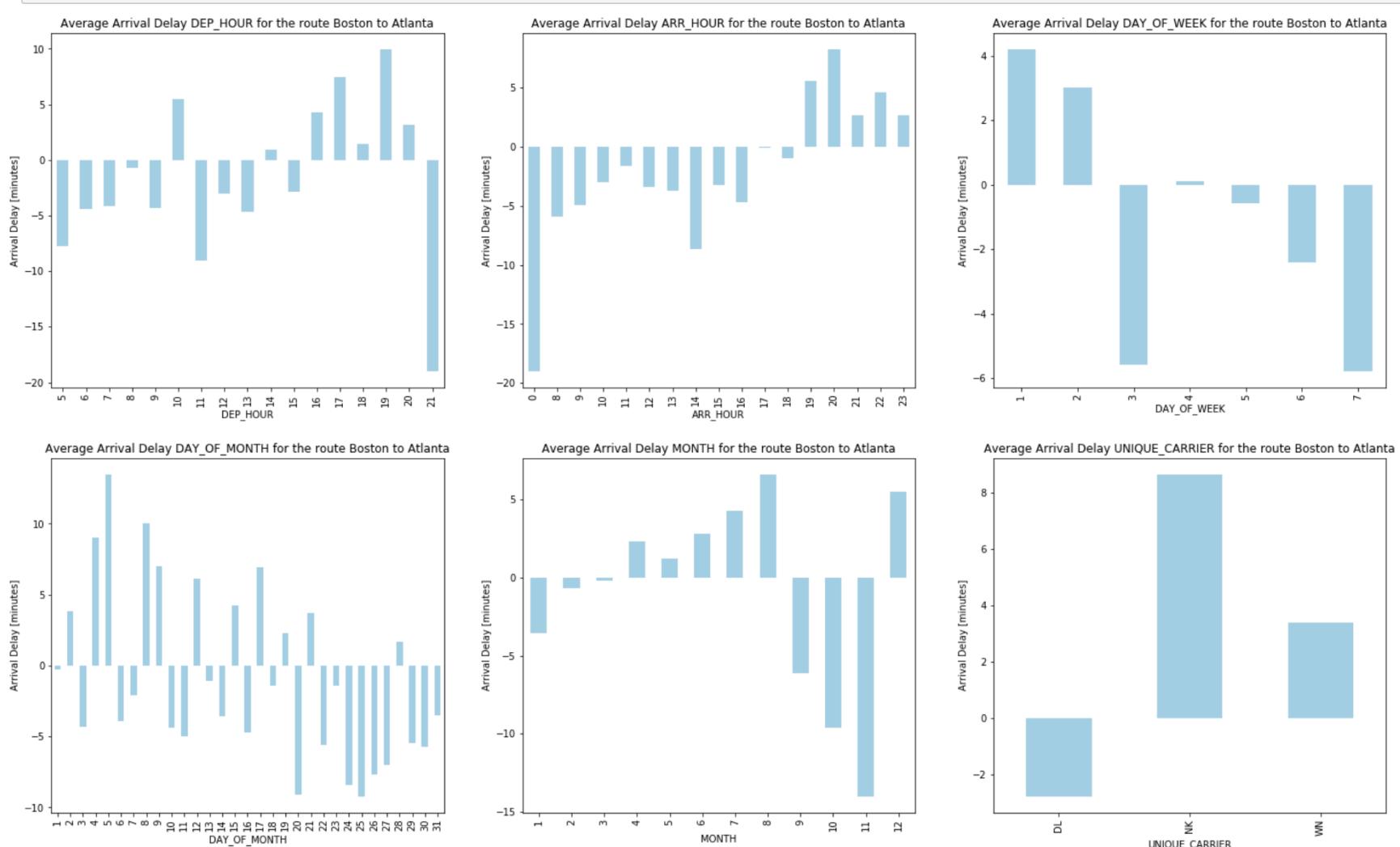
Analyse Multivariée (4)

Délai moyen à l'arrivée: Miami -> New York



Analyse Multivariée (5)

Délai moyen à l'arrivée : Boston -> Atlanta



Part IV:

Pistes modélisation

Difficultés initiales

- Choix des variables pour la régression:

- * celles accessibles bien avant le départ (heure, jour, mois, compagnie,aéroport départ,...)

- Choix de l'encodage:

- * quelles variables en One Hot Encoding
- * quid des variables temporelles

- Problème lié à la taille du dataset près de 5,5 millions de lignes

- * au début 733 features (avec One Hot Encoding)

=> Plantage Notebook



Premières pistes

- Au début avec **One Hot Encoding uniquement** => taille (5.5M; 733)
 - * Séparer dataset selon les 12 compagnies aériennes
 - => idée réduction dimensionnelle mais entraînement de 12 modèles
 - * **Lecture:** Google_colab, Kaggle GPU et AWS EC2
 - * Faire un « sampling » du dataset => mais toujours problématique (temps calcul)
- => **Conclusion :** trop de features
- @ principal cause = OHEncoding des aéroports départ et arrivé
 - => penser à Feature Engineering
 - @ OHEncoding des variables temporelles pas idéal
 - => penser à une alternative



Réduction nombre features

- FEATURE ENGINEERING

* Remplacer

‘ORIGIN_AIRPORT_ID’ par le nombre de vols en partance de l’aéroport

‘DEST_AIRPORT_ID’ par le nombre de vols arrivant à l’aéroport

=> traitement maintenant comme *variable numérique* (non plus catégorielle)

* Rajout variable :

indiquant le nombre de jour par rapport aux congés US les plus proches.

- Variables temporelles (h départ/arrivée, jour du mois/semaine, mois)

Encodage cyclique (cos/sin)

Cyclic variables (month, days, hours)

```
Entrée [18]: df_data['cos_MONTH'] = np.cos(2*np.pi *df_data.MONTH/12)
df_data['sin_MONTH'] = np.sin(2*np.pi *df_data.MONTH/12)
```

- One Hot Encoding

Uniquement pour UNIQUE_CARRIER

Résultat : de **733** à **27** features

Démarches suivies

Features avant encodage:

```
[UNIQUE_CARRIER', 'MONTH', 'DAY_OF_MONTH', 'DAY_OF_WEEK', 'DEP_HOUR', 'ARR_HOUR',  
'ORIGIN_AIRPORT_ID', 'DEST_AIRPORT_ID', 'DISTANCE', 'ARR_DELAY']
```

- Préparation données

- * séparation en X et y
- * ensuite séparation en jeu d'entraînement et test pour X et y
- * mise à l'échelle de variables numériques

- Modèle base line de comparaison

- * Régression linéaire sans régularisation
- * Calcul de l'erreur RMSE = **critère d'évaluation**

RMSE = 27,13

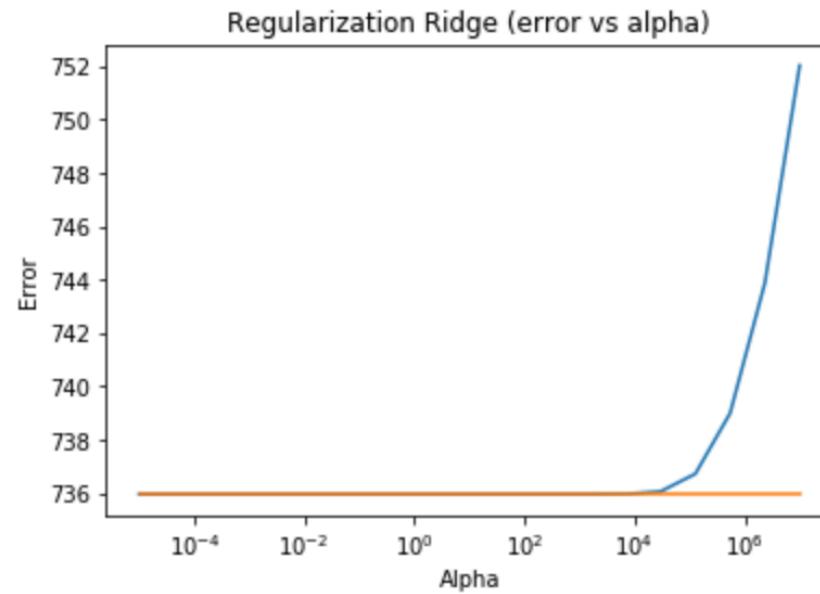
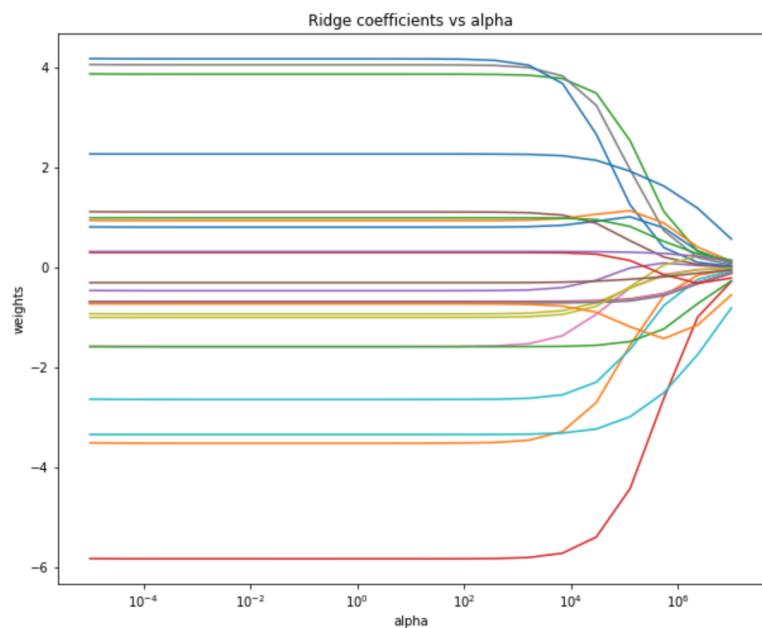
- Entraînement modèle RL avec régularisation (pour différents hyperparamètres alpha)

- * RIDGE
- * LASSO
- * calcul de l'erreur RMSE

- Validation croisée

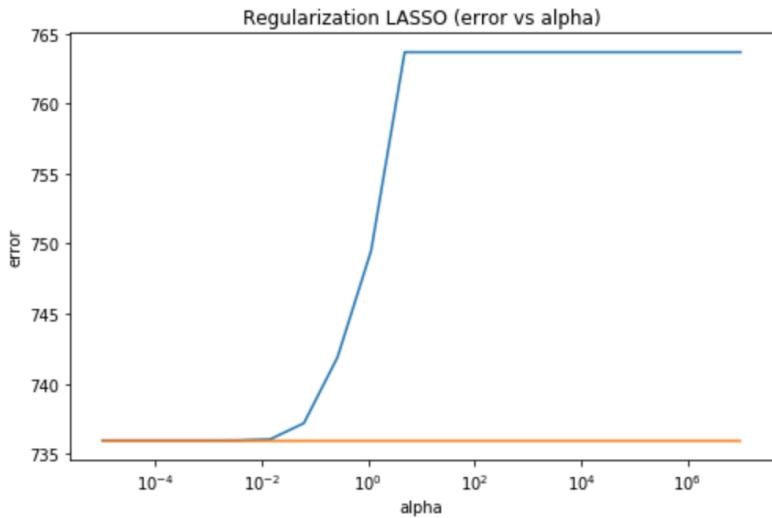
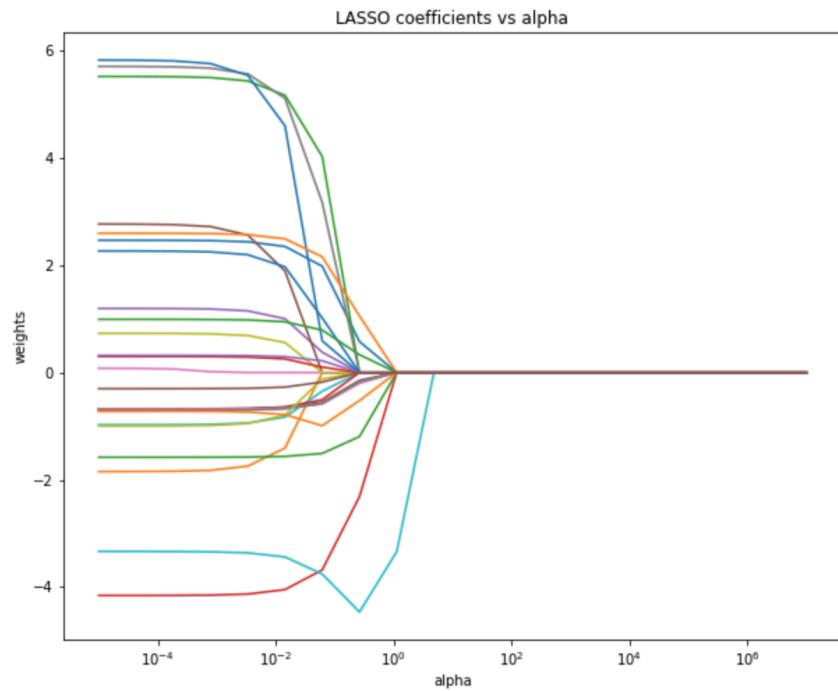
Régularisation Ridge

RMSE = 27,13



Régularisation LASSO

RMSE = 27,13



RMSE la même que pour RIDGE
Donc pas la peine: *Elastic-Net*

Part V:

Modèle final et performances

Prediction retard (minutes) (1)

En entrée

- * **origin** = 'Boston', **destination** = 'Atlanta'
- * **carrier** = 'AA'
- * **dept_time** = 1700, **arr_time** = 1200,
- * **month** = 9, **day_of_month** = 16, **day_of_week** = 3

- 1. **Recherche** de 'ORIGIN_AIRPORT_ID' et 'DEST_AIRPORT_ID'

Ensuite

retrouver la distance

- 2. **Calcul** du nombre de jours par rapport aux congés US les plus proches

- 3. **Encodage:**

- * UNIQUE_CARRIER : OHE
- * 'ORIGIN_AIRPORT_ID' et 'DEST_AIRPORT_ID' : feature engineering
- * variables temporelles : encodage cyclique

- 4. **Mise à l'échelle** des variables numériques

- 5. **Prédiction**

A l'aide du modèle RIDGE

API

http://ocprojet4.pythonanywhere.com/recommend?origin=Boston&dest=Atlanta&carrier=AA&dep_time=1300&arr_time=2000&month=11&day_of_month=21&day_of_week=7

← → C ⓘ Non sécurisé | ocprojet4.pythonanywhere.com/recommend?origin=Boston&dest=Atlanta&carrier=AA&dep_time=1300&arr_time=2000&month=11&day_of_month=21&day_of_week=7 ☆

The expected delay for a flight from Boston to Atlanta with as carrier: AA, the 21th of November (a Sunday) is 8 minutes.

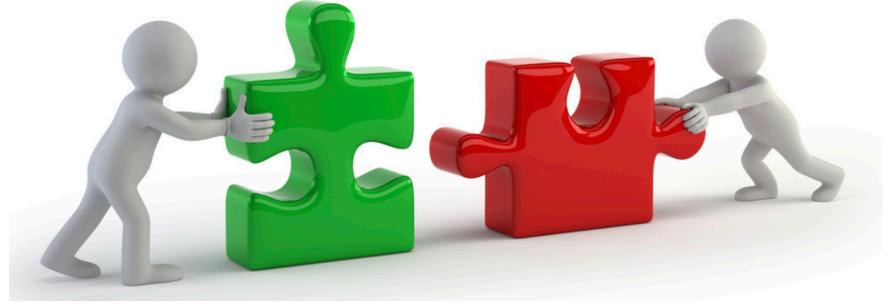
← → C ⓘ Non sécurisé | ocprojet4.pythonanywhere.com/recommend?origin=Miami&dest=Atlanta&carrier=B6&dep_time=1900&arr_time=2000&month=6&day_of_month=15&day_of_week=3

The expected delay for a flight from Miami to Atlanta with as carrier: B6, the 15th of June (a Wednesday) is 18 minutes.

Part VI:

Conclusions

Conclusions



- Un des points clefs dans ce projet a été la dimension des entrées du modèle
 - * le bon choix de variables de départ
 - * le bon choix d'encodage (OHE, cyclique)
 - * le feature engineering
- Ensuite nous avons utilisé une Régression Linéaire
 - Avec et sans régularisation (Ridge / Lasso)
- Comme critères d'évaluation : RMSE

Pistes

Finalement nous avons obtenu une RMSE de 27.13 en testant différentes valeurs d'alphas tant pour la régularisation RIDGE que LASSO.

Un *facteur limitatif* est probablement le type de variables auxquelles on a accès (cause principale météo).

Une possible amélioration serait d'entraîner les modèles par trajet ou par ville de départ.



