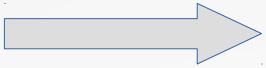
Parcours Datascientist : projet 4

Anticipation des retards de vol

Francois BANGUI

Formulation du problème

Anticipation des retards



Données passées => Prediction future

Classe : Régression

Modèle linéaire

Hypothèses à vérifier :
- Bruit normal
- Observations I.I.D.
- Linéarité

Analyse des variables

DATATION

YEAR
QUARTER
MONTH
DAY_OF_MONTH
DAY_OF_WEEK
FL_DATE
CRS_DEP_TIME
CRS_ARR_TIME
CRS_ELAPSED_TIME

OPERATEUR

UNIQUE_CARRIER
AIRLINE_ID
CARRIER
TAIL_NUM
FL_NUM

LOCALISATION

ORIGIN_AIRPORT_ID
ORIGIN_AIRPORT_SEQ_ID
ORIGIN_CITY_MARKET_ID
ORIGIN ORIGIN_CITY_NAME
ORIGIN_STATE_ABR
ORIGIN_STATE_FIPS
ORIGIN_STATE_NM
ORIGIN_WAC

DEST_AIRPORT_SEQ_ID
DEST_CITY_MARKET_ID
DEST
DEST_CITY_NAME
DEST_STATE_ABR
DEST_STATE_FIPS
DEST_STATE_NM
DEST_WAC
DEST_AIRPORT_ID

DISTANCE
DISTANCE_GROUP

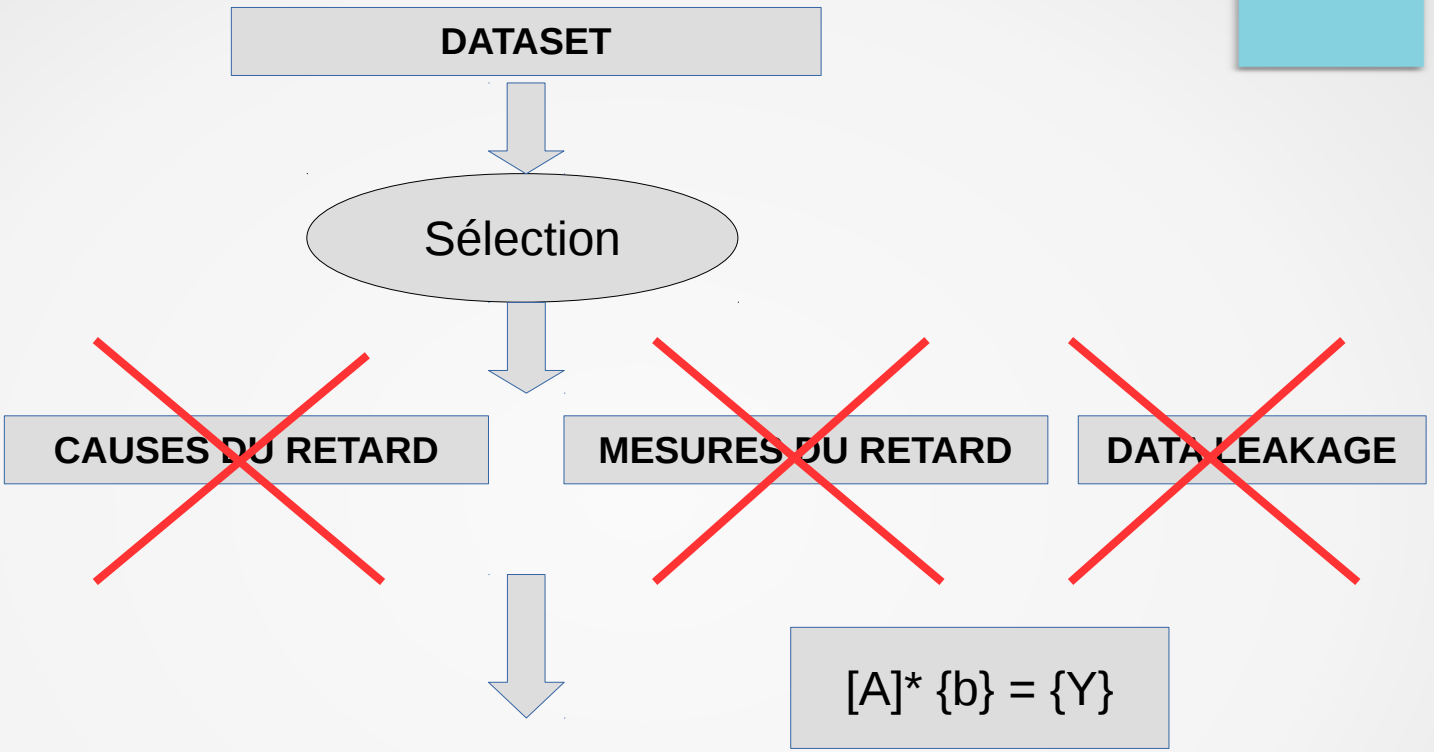
CAUSES DU RETARD

CANCELLED
CANCELLATION_CODE
DIVERTED
TAXI_OUT
TAXI_IN
WHEELS_OFF
WHEELS_ON
AIR_TIME

MESURES DU RETARD

DEP_TIME
DEP_DELAY
DEP_DELAY_NEW
DEP_DEL15
DEP_DELAY_GROUP
ARR_DELAY
DEP_TIME_BLK
ARR_TIME
ARR_DELAY_NEW
ARR_DEL15
ARR_DELAY_GROUP
ARR_TIME_BLK
ACTUAL_ELAPSED_TIME

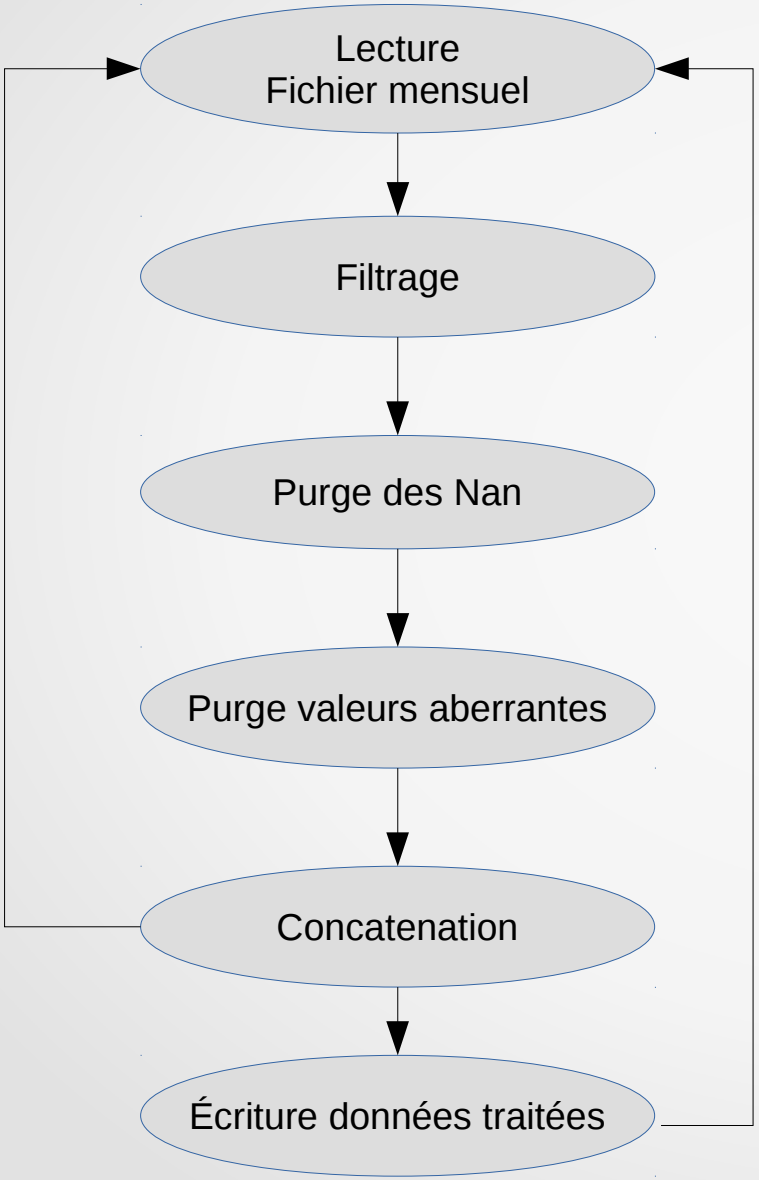
Sélection des variables

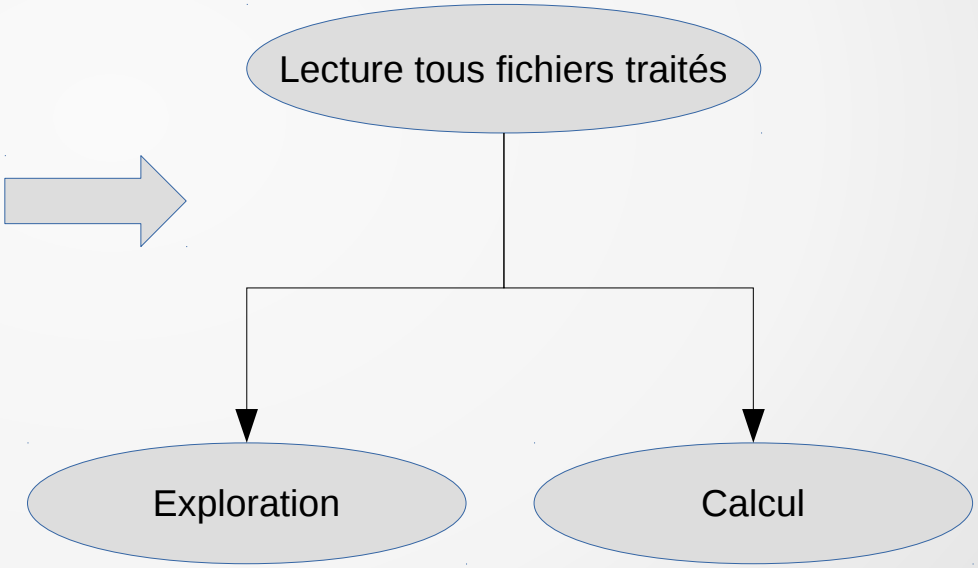


DATATION	LOCALISATION	OPERATEUR
DAY_OF_MONTH DAY_OF_WEEK MONTH **CRS_DEP_TIME** **CRS_ELAPSED_TIME**	ORIGIN_AIRPORT_ID DEST_AIRPORT_ID	AIRLINE_ID

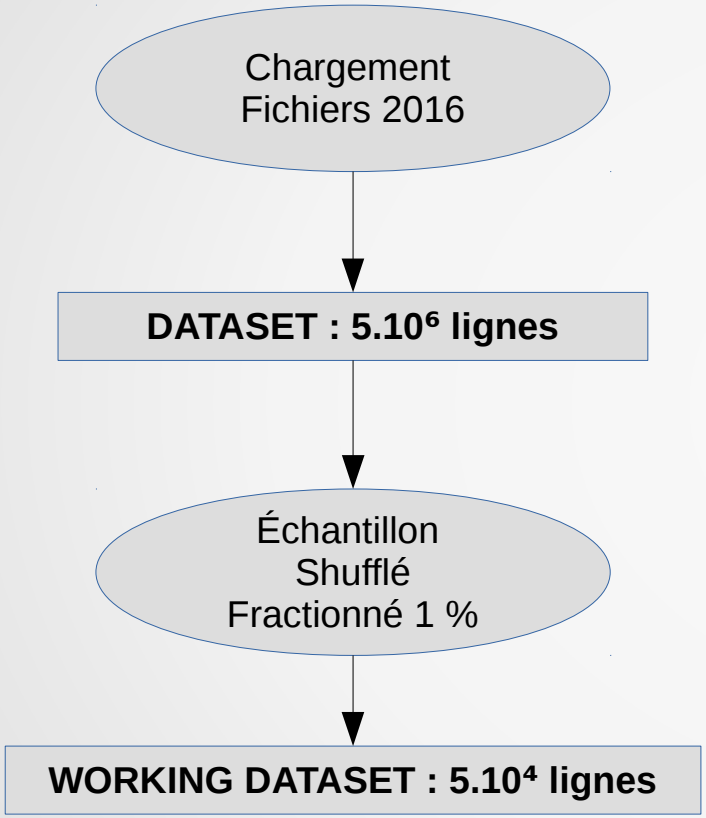
CIBLE
ARR_DELAY

Préparation des données : 1,2GB

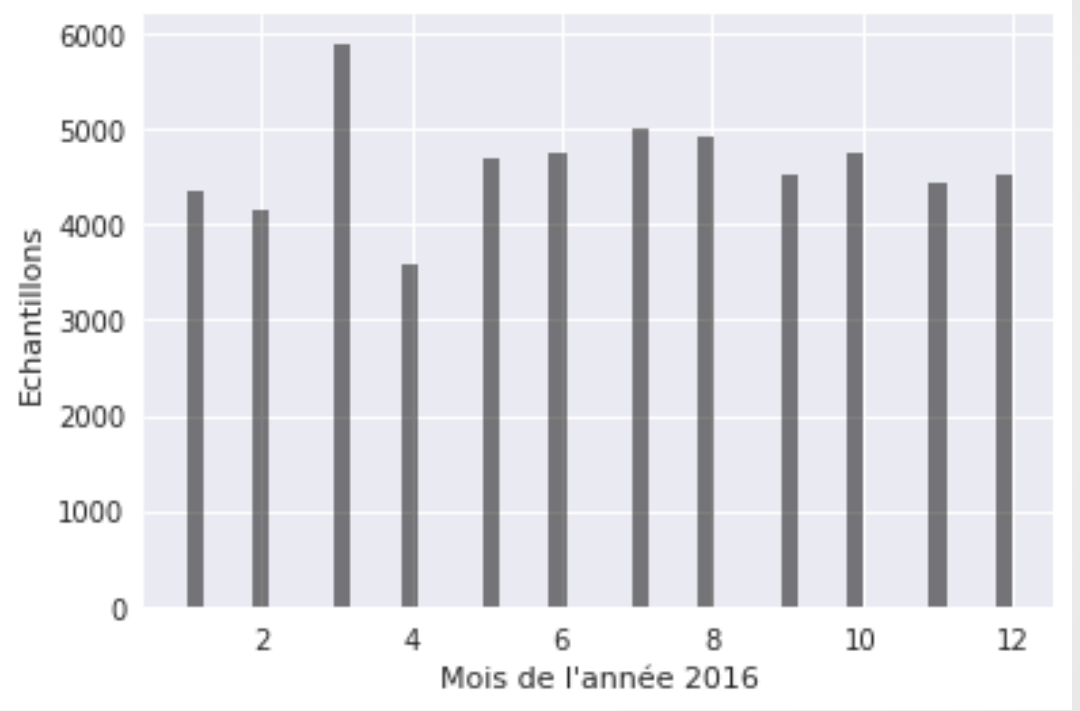




Exploration : Échantillonnage

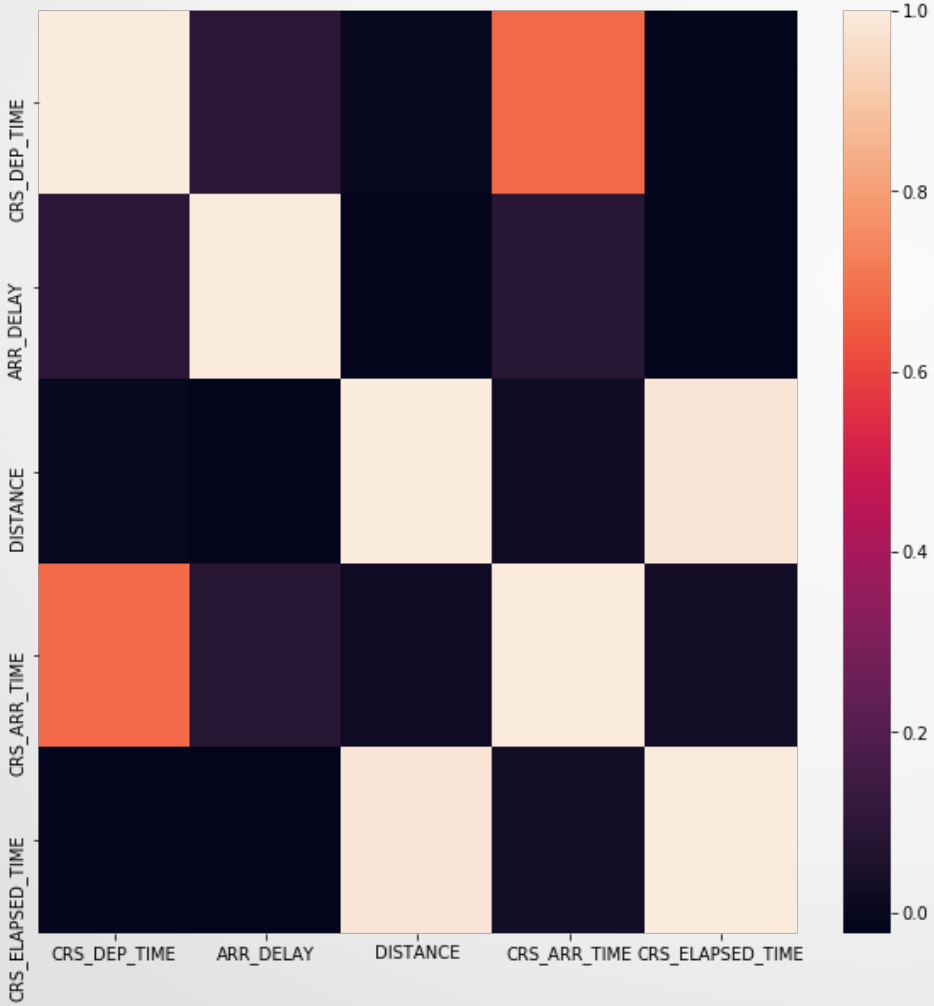


Représentation des mois dans l'échantillon



Corrélation des variables

Modèle : 5 millions de de lignes
Étude : ~ 50 000 lignes « shufflées »



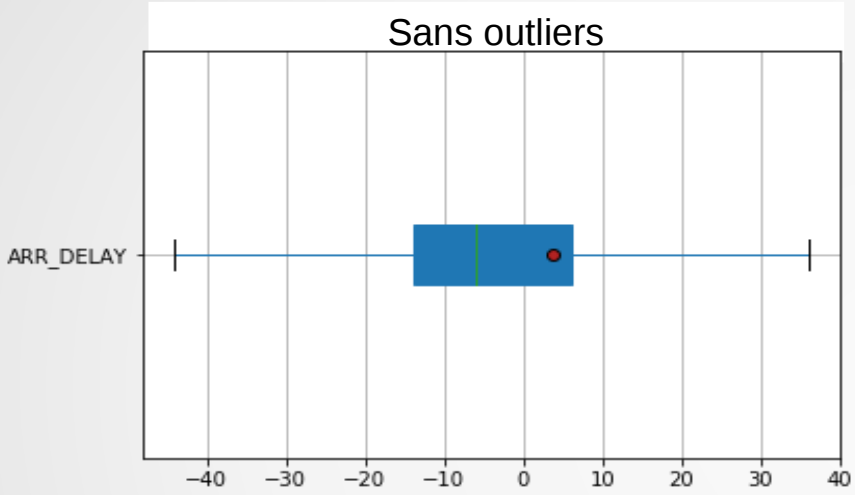
DATA LEAKAGE
- CRS_ELAPSED_TIME.
- CRS_ARR_TIME
- CRS_DEP_TIME.

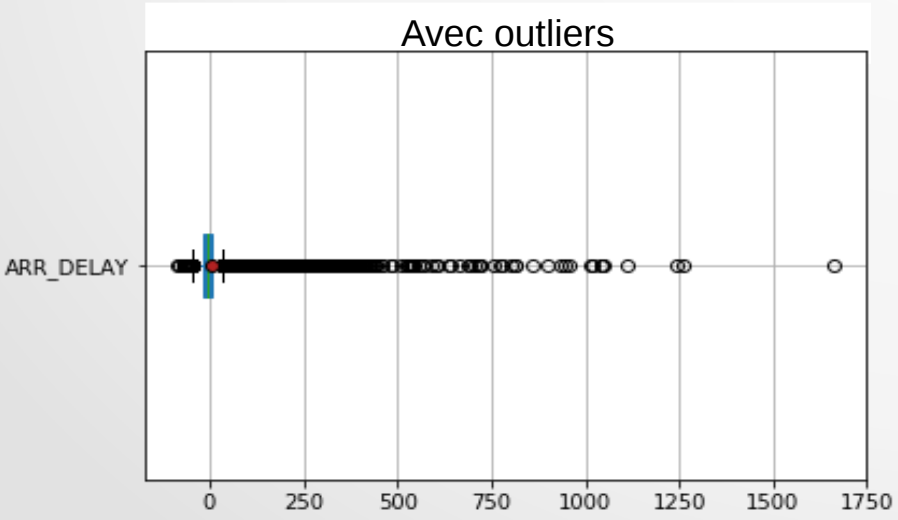
Corrélation entre les variables
- DISTANCE / CRS_ELAPSED_TIME.
- CRS_ARR_TIME / CRS_DEP_TIME.

Il n'est pas nécessaire de garder dans le modèle ces deux variables.

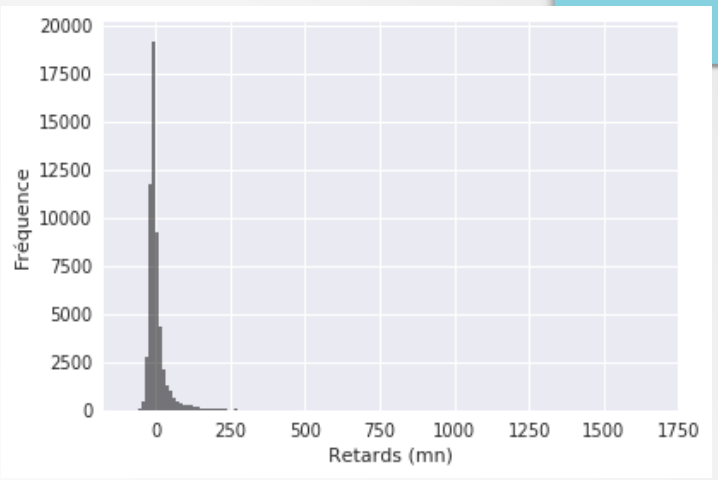
Distribution des retards : ARR_DELAY

Distribution des retards dans l'échantillon

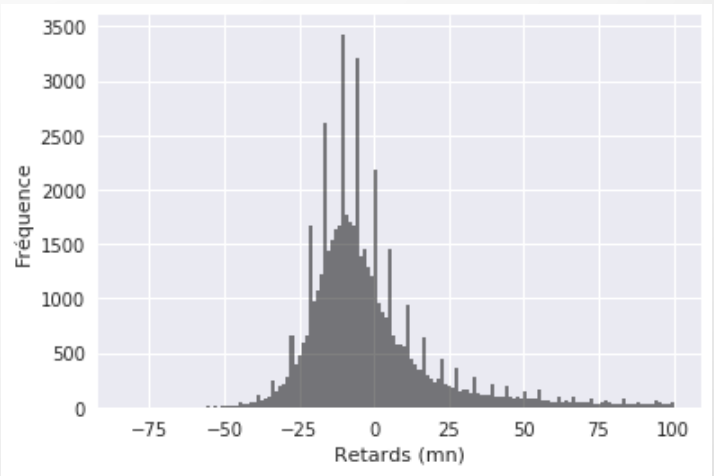




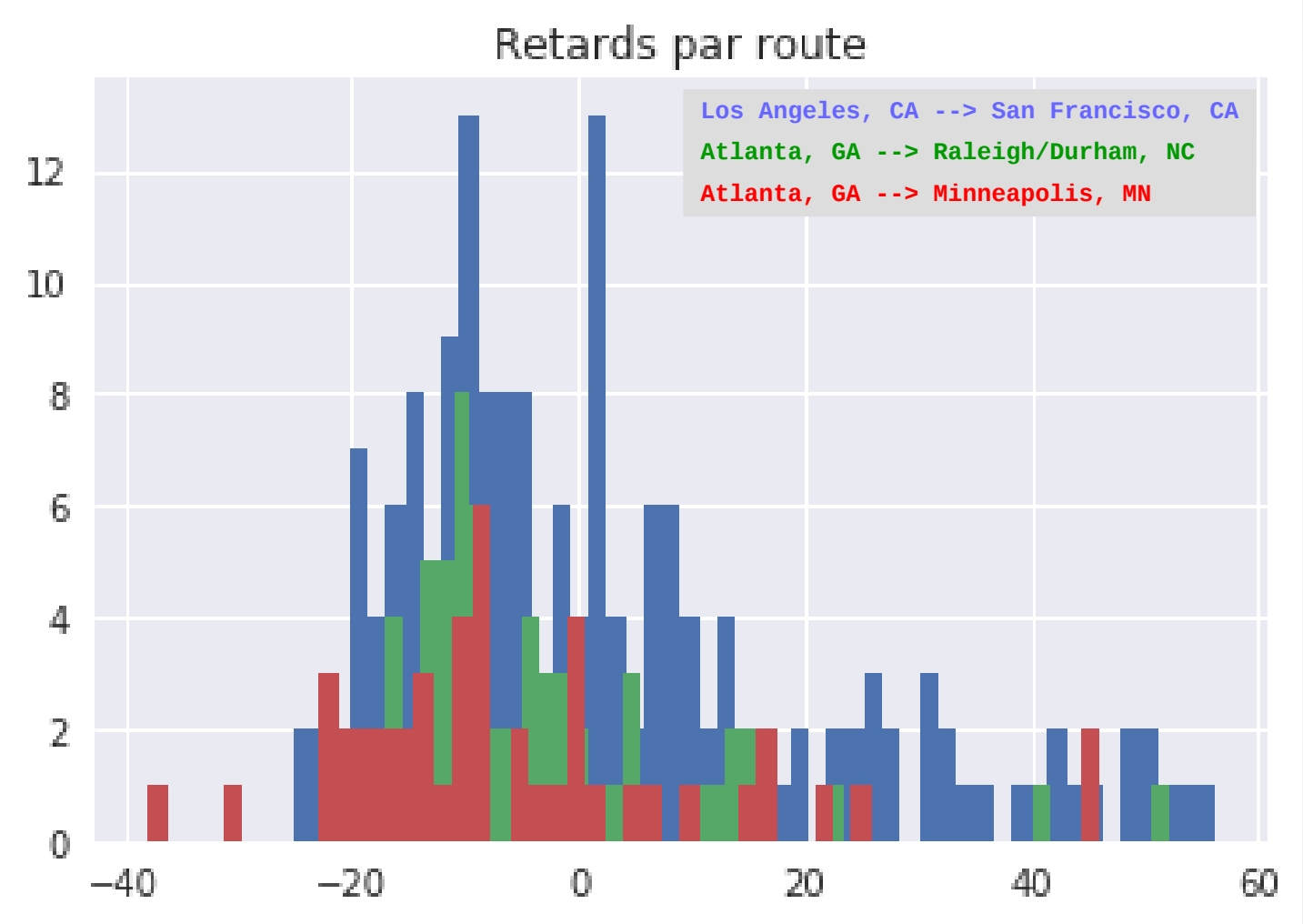
Suppression des outliers : 10 % des données



Moyenne : 4.0 mn
Mediane : -6.0 mn
Variance : 1906 mn
Ecart type :44 mn



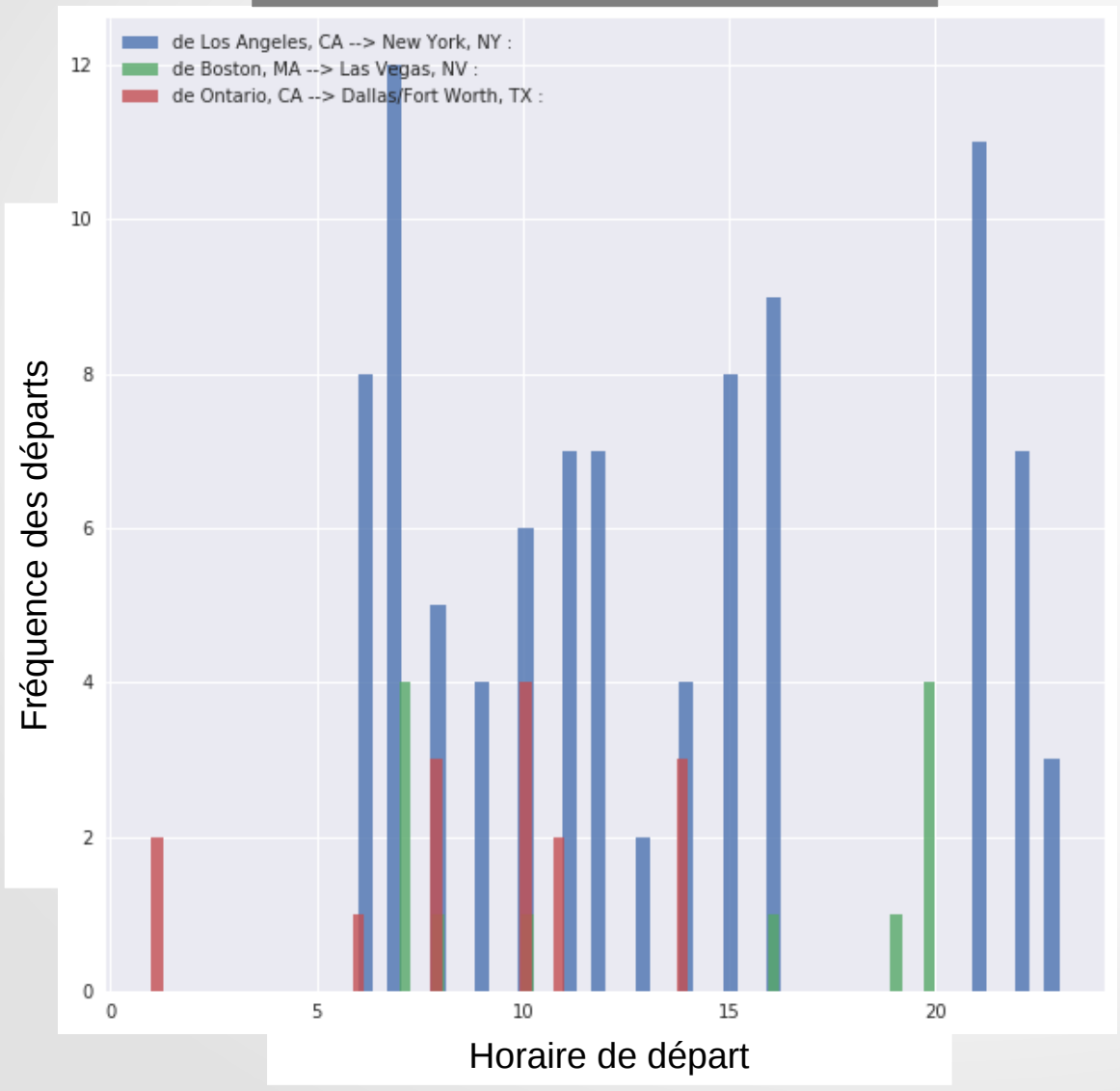
Retards par route



Disparités des distributions des retards

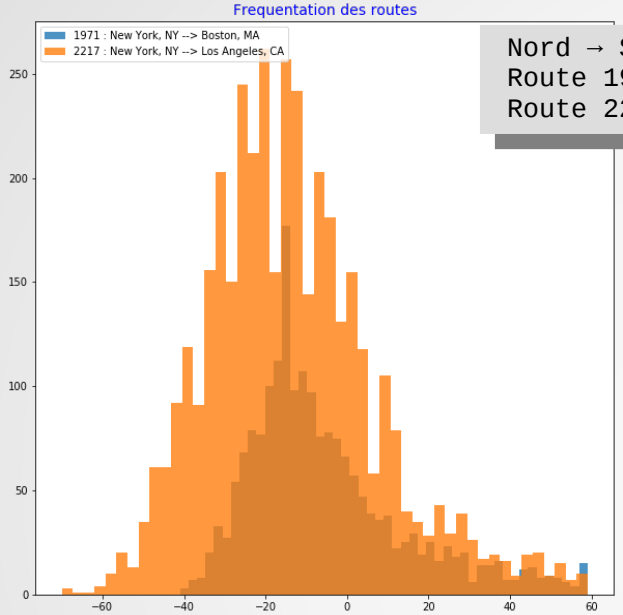
Activité au départ par route

Fréquence des départs par route

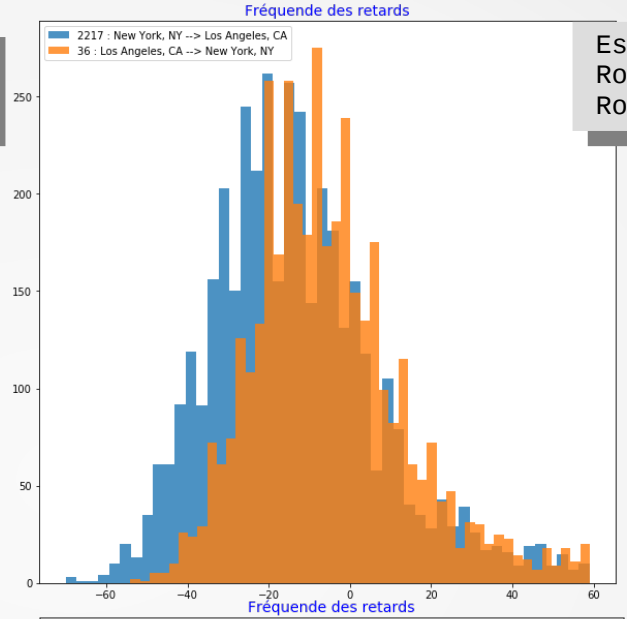


- Variation dans les fréquences
- Variation des amplitudes

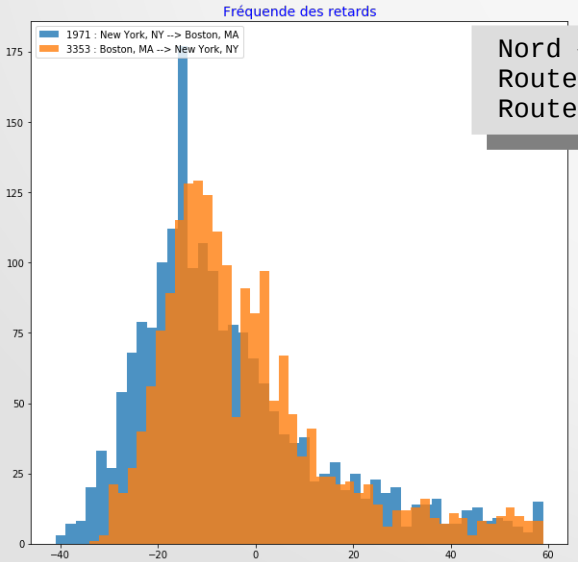
Retards par route



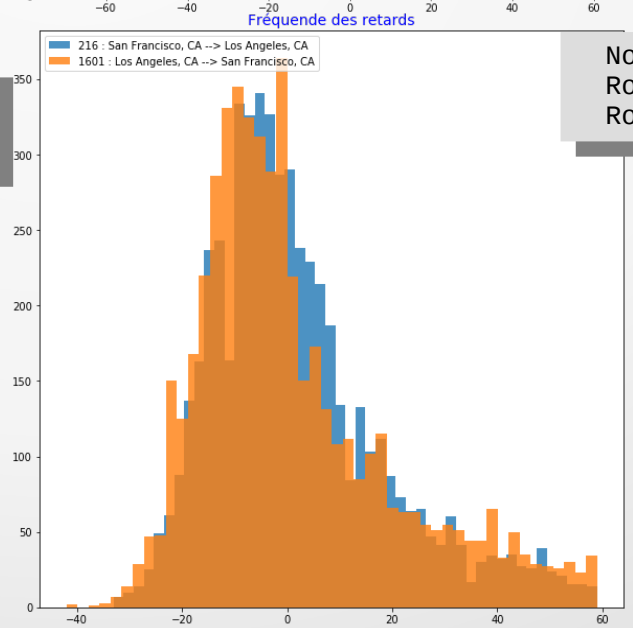

```
Nord → Sud / Est → Ouest
Route 1971 Skew = 1.13
Route 2217 Skew = 0.70
```




```
Est → Ouest / Ouest → Est
Route 2217 Skew = 0.70
Route 36 Skew = 0.77
```




```
Nord → Sud / Sud → Nord
Route 1971 Skew = 1.13
Route 3353 Skew = 1.29
```




```
Nord → Sud / Sud → Nord
Route 216 Skew = 1.03
Route 1601 Skew = 1.05
```

Variance des retards par route

Études des routes sur deux zones géographiques : côtes EST / OUEST :

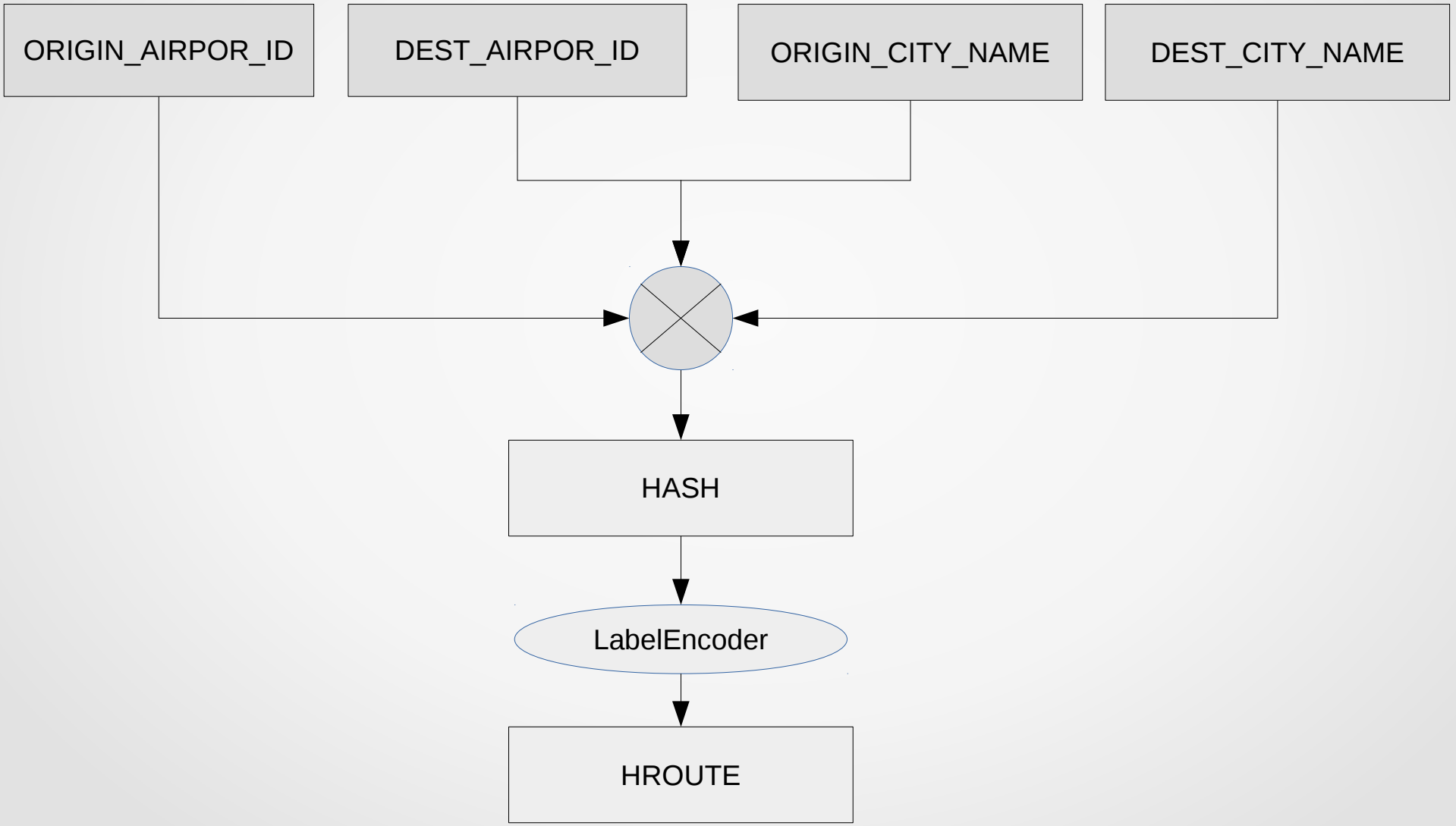
NY / LA : courbes décentrées

Routes NORD / SUD : skewed et variance proches

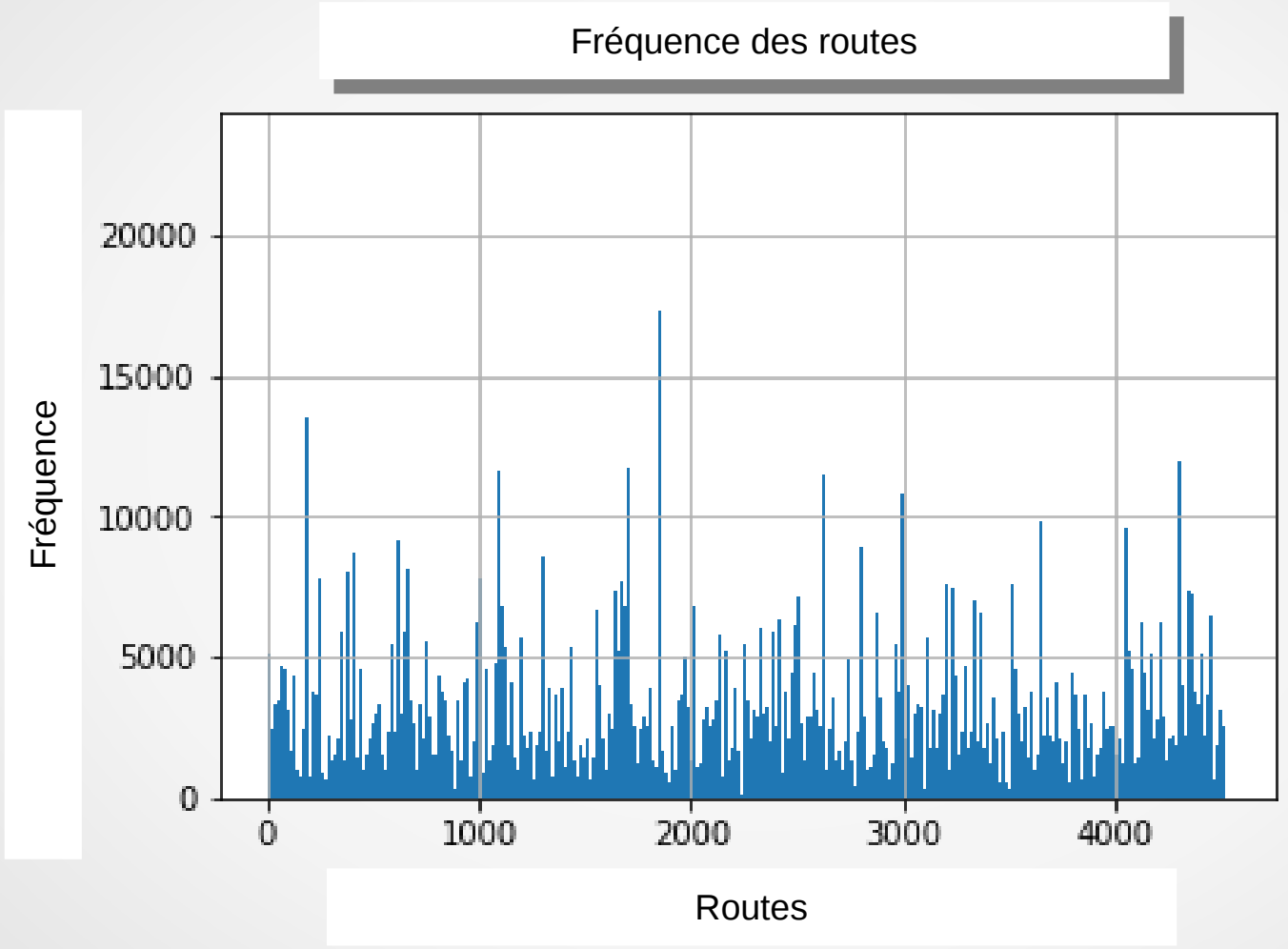
Routes EST/OUEST : skewed et variance proches, décentrage

Incidence des fréquences des départs et arrivées sur les retards

Construction des routes



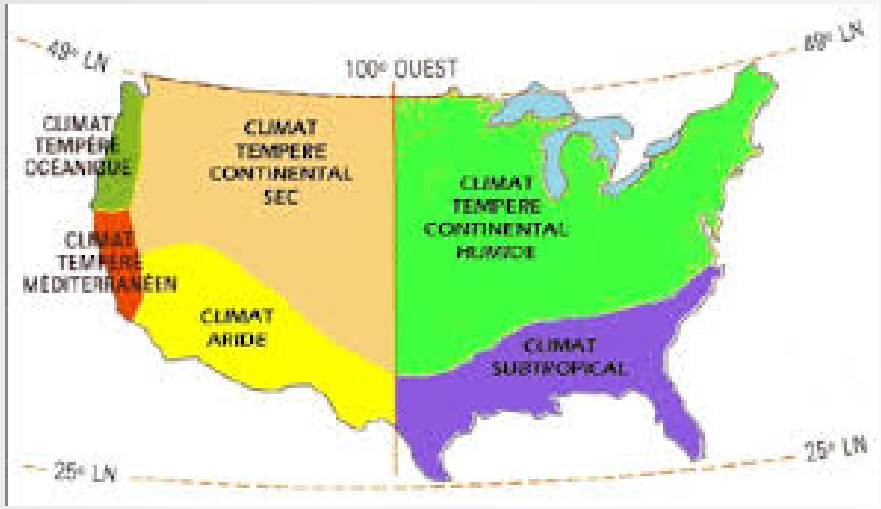
Distribution des routes

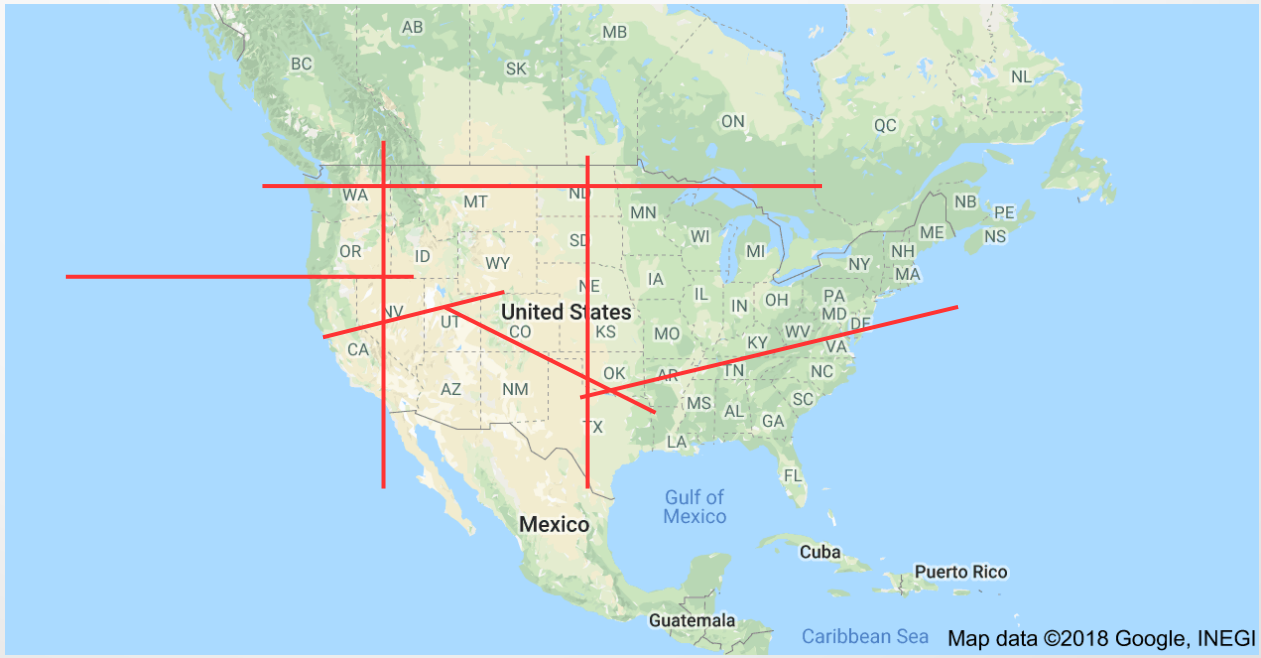


Variables retenues pour le modèle : par route (1)

Quantitatives	Catégorielles	CIBLE
CRS_DEP_TIME	AIRLINE_ID MONTH DAY_OF_MONTH DAY_OF_WEEK	ARR_DELAY

Améliorations : modèle climatique





Variables retenues pour le modèle : par route (2)

Quantitatives	Catégorielles	CIBLE
CRS_DEP_TIME	AIRLINE_ID MONTH DAY_OF_MONTH DAY_OF_WEEK ORIGIN_STATE_CLM DEST_STATE_CLM	ARR_DELAY

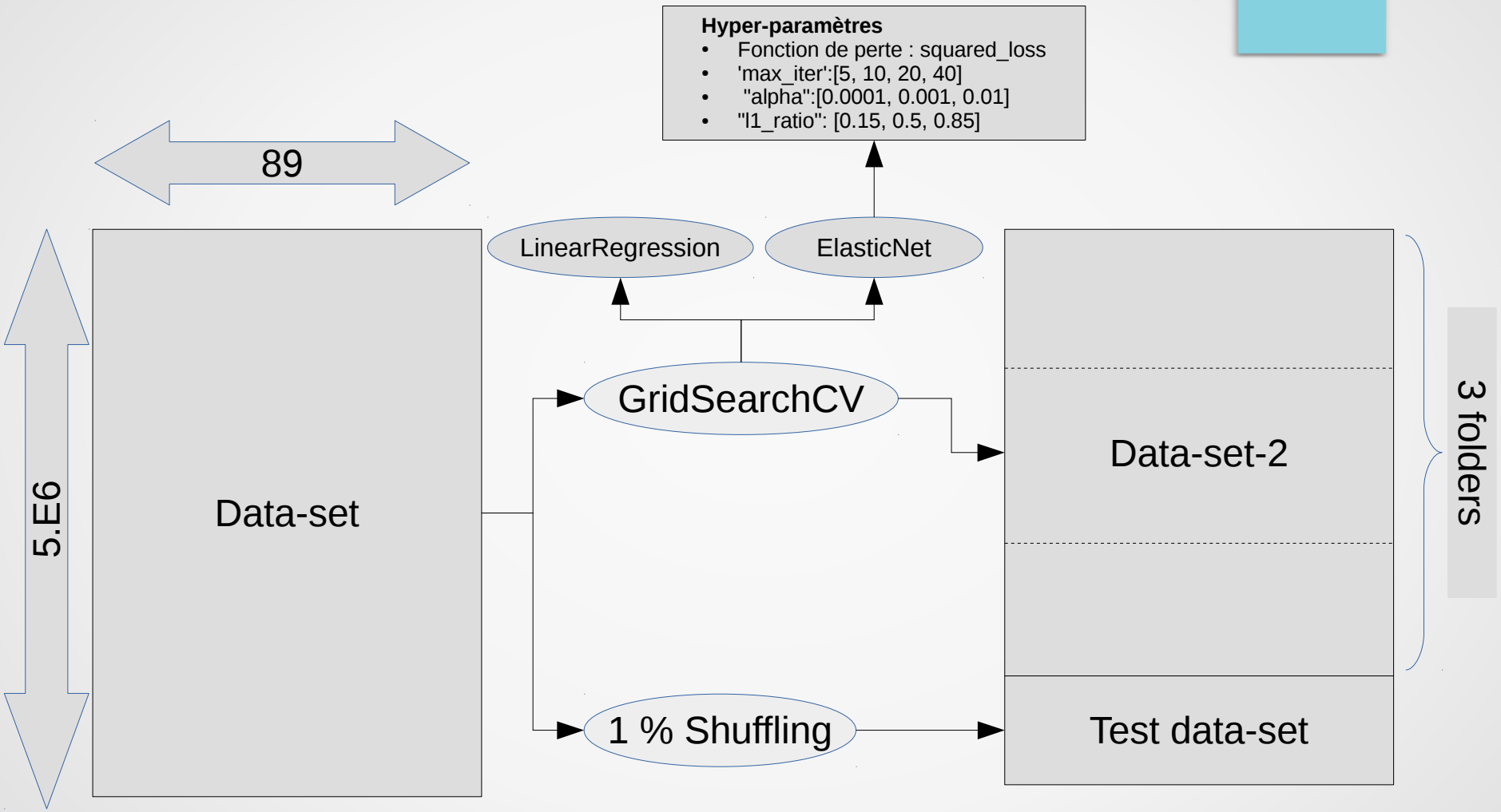
Traitement : one-hot encoding + scaling

	Avant	Augmentation	Après
MONTH	7	+12	28
DAY_OF_MONTH	28	+31	58
DAY_OF_WEEK	58	+7	64
AIRLINE_ID	64	+12	75
ORIGIN_STATE_CLM	75	+7	82
DEST_STATE_CLM	82	+7	88
CRS_DEP_TIME	88	+1	89
Dimension du modèle	**89**		

« One Hot » encoding

Scaling

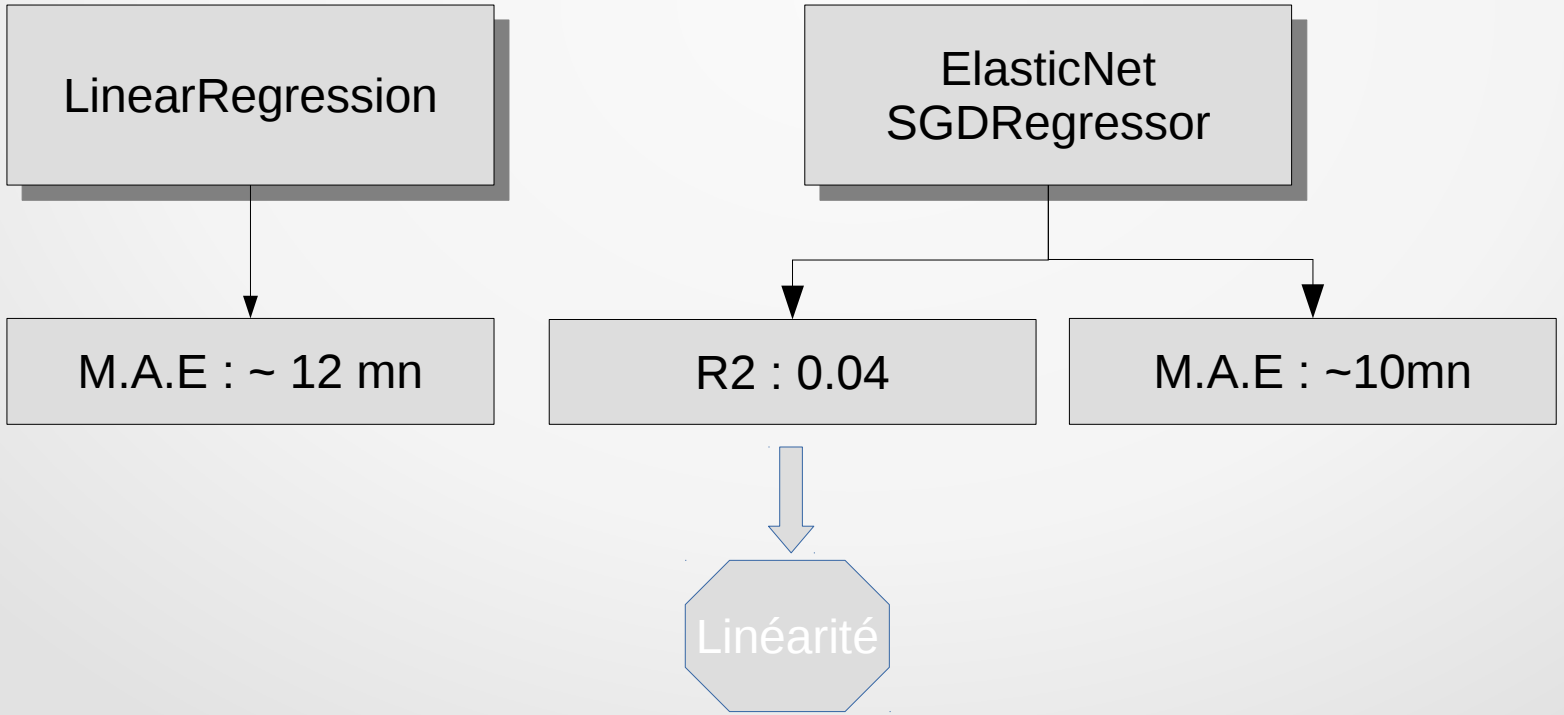
Modèles linéaires : organisation des données



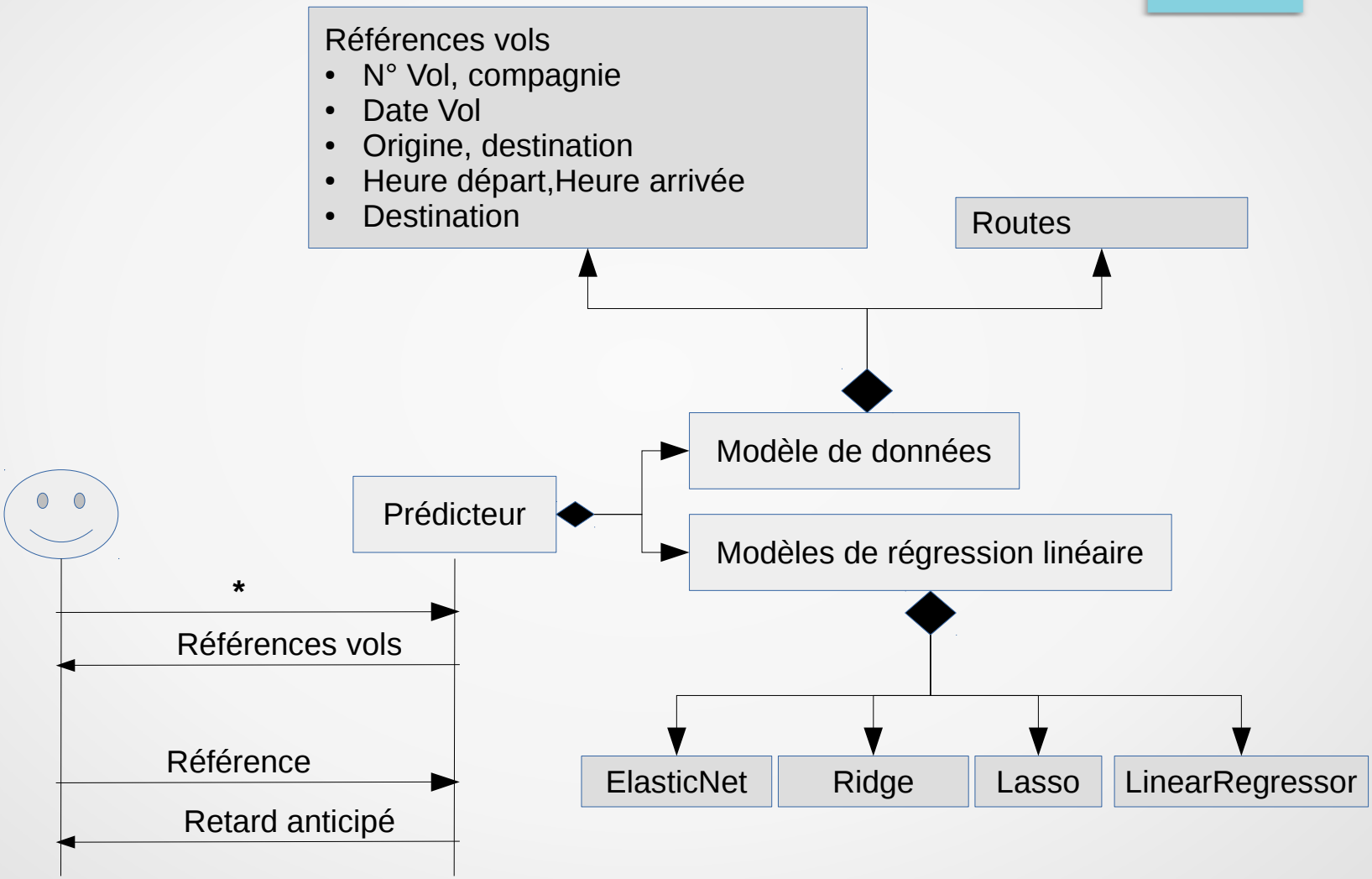
Benchmark des modèles

Modèle régression linéaire : ElasticNet

- SGDRegressor : adapté aux larges problèmes
- Hyper-paramètres :
  - Validation croisée GridSearchCV : 3 folders
- Erreur : MAE et R2 pondérées



Implémentation



Déploiement Heroku & Tests

Application : francoisbangui-flydelaypred.herokuapp.com

Chargement du composant **oLinearDelayPredictor** en RAM

OlinearDelayPredictor.dump 90 MB

Liste des vols disponibles

https://francoisbangui-flydelaypred.herokuapp.com/predictor?'*'

{"_select":
[
{"id":"446011","flight":"3528","company":"WN","origin":"Oakland, CA","destination":"San Diego, CA","departure":"WED 12-28 10:40","arrival":"12:05},
{"id":"348365","flight":"4220","company":"EV","origin":"Houston, TX","destination":"Nashville, TN","departure":"TUE 07-5 15:40","arrival":"17:45},
{"id":"70245","flight":"48","company":"VX","origin":"Kahului, HI","destination":"San Francisco, CA","departure":"THU 07-7 22:15","arrival":"06:25},
{"id":"366211","flight":"3004","company":"WN","origin":"Chicago, IL","destination":"Providence, RI","departure":"TUE 12-6 13:45","arrival":"16:50},
{"id":"150956","flight":"1573","company":"AA","origin":"Chicago, IL","destination":"Kansas City, MO","departure":"SUN 07-17 07:25","arrival":"08:51}]}

Résultat : Régression linéaire vs ElasticNet

{"_result":[{"id":"446011","model":"**SGDRegressor**","evaluated_delay":"0","model":"**LinearRegression**","evaluated_delay":"1","measured_delay":"23"}]}

{"_result":[{"id":"70245","model":"**SGDRegressor**","evaluated_delay":"-23","model":"**LinearRegression**","evaluated_delay":"-8","**measured_delay**":"-23"}]}

{"_result":[{"id":"366211","model":"SGDRegressor","evaluated_delay":"-2","model":"LinearRegression","evaluated_delay":"7","measured_delay":"-5"}]}

{"_result":[{"id":"150956","model":"SGDRegressor","evaluated_delay":"-2","model":"LinearRegression","evaluated_delay":"7","measured_delay":"13"}]}

Conclusions

- Linéarité : pas assez d’information pour une bonne évaluation.
- Avantage : performance
- Inconvénients : Flexibilité : nouvelle route ?

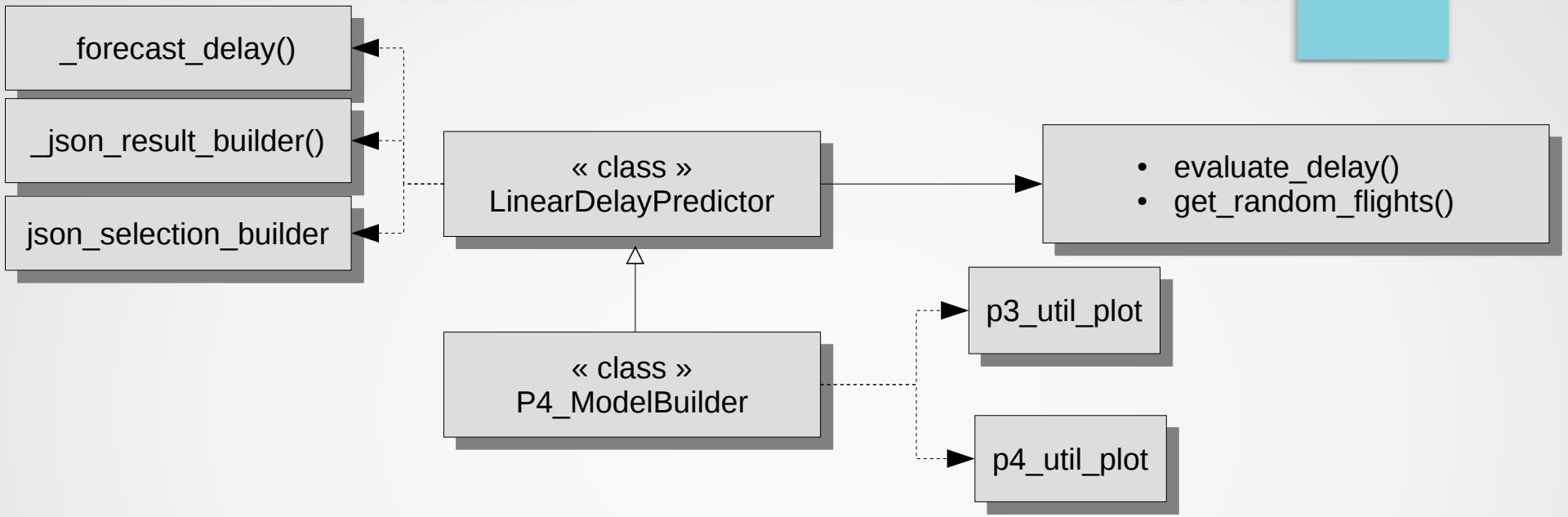
Axes d’améliorations :

- Informations transporteur
- Pré-classification binaire
- Modèle météorologique

Annexe 1 : fichiers du projet

- **Fichiers source python :**
  - heroku/flight_predictor/LinearDelayPredictor.py
  - heroku/flight_predictor/config.py
  - heroku/flight_predictor/views.py
- **Notebook de l'alnalyse exploratoire :**
  - P4.ipynb : Nettoyage / Exploration
- **Notebook des approches de modélisation :**
  - P4_ModelBuilder.ipynb : Évaluation / Tests / Pré-production / Génération d'un objet de type LinearDelayPredictor
- **Rapport sous forme de présentation pdf:**
  - Openclassrooms_ParcoursDatascientist_P4.pdf
- **Points d'entrée de l'API :**
  - Pour récupérer une liste de vols :
    - https://https://francois-bangui-oc-p4.herokuapp.com/predictor/?'*'
  - Pour récupérer l'évaluation du retard d'un vol à partir de son identifiant :
    - https://https://francois-bangui-oc-p4.herokuapp.com/predictor/?flight_id=<ID>

Annexe 2: Ingénierie logicielle



- Linear regression models
  - LinearRegressor
  - Rigde
  - Lasso
  - ElasticNet