# PRÉDICTION DU RETARD D'AVION

IMSD2020

Fatima Zahra MOUHSINI | Fatma BALI

# Plan

1. Contexte du projet
2. Base de données
    a. Présentation du dataset
    b. Data cleaning
    c. La variable à prédire
3. Analyse descriptive
4. Modélisation
    a. Classification
    b. Régression
    c. Evaluation des performances
5. Solution proposé

# Context

*Rien qu'en europe*

**6 MILLIONS**

De passagers ont été victime d'un retard en Juillet/Août 2018

**2,2 MILLIARDS**

D'euro d'indemnisation

**Aéroports et compagnies aériennes:**
Est-il possible de prédire le retard d'un vol ?

**Particuliers:**
Quelle est la probabilité que mon vol soit en retard ?

**Les problèmes liées aux retards**

- Correspondances manquées
- Retard dans les engagements commerciaux
- Frustration des voyageurs

**Bénéfice de la prise de conscience du retard**

- Sélection des vols
- Meilleure relation commerciale
- Rester en bonne santé

# Base de données

## Présentation du dataset

Base des vols de plusieurs compagnie aérienne y compris des données sur les aéroports et les compagnies.

- Dimension : 48  colonnes | 3279290 lignes
- Présentation des variables

| AEROPORT DEPART | AEROPORT ARRIVEE | DEPART PROGRAMME | ARRIVEE PROGRAMMEE | HEURE DE DEPART | HEURE D'ARRIVEE | RETART DE DEPART | RETARD A L'ARRIVEE |
|---|---|---|---|---|---|---|---|
| le code IATA de l'aéroport de départ | le code IATA de l'aéroport d'arrivée | Heure de départ prévue | Heure d'arrivée prévue | Heure de départ réelle | Heure d'arrivée réelle | Le retard effectué au départ | Le retard effectué à l'arrivé |

# Base de données

## Data cleaning

Les valeurs manquantes ont été traitées.
Nous avons supprimé les valeurs manquantes qui étaient désignées par NaN et les colonnes qui représente trop de valeurs manquantes.
Ce qui nous a laissé 2.997.544 vols dans le dataset.

```
IDENTIFIANT                                                  0
VOL                                                          0
CODE AVION                                                   0
AEROPORT DEPART                                              0
AEROPORT ARRIVEE                                             0
DEPART PROGRAMME                                             0
HEURE DE DEPART                                          51414
RETART DE DEPART                                         51414
TEMPS DE DEPLACEMENT A TERRE AU DECOLLAGE                53406
DECOLLAGE                                                53406
TEMPS PROGRAMME                                              6
TEMPS PASSE                                              62529
TEMPS DE VOL                                             62529
DISTANCE                                                     0
ATTERRISSAGE                                             55336
TEMPS DE DEPLACEMENT A TERRE A L'ATTERRISSAGE            55336
ARRIVEE PROGRAMMEE                                           0
HEURE D'ARRIVEE                                          55336
RETARD A L'ARRIVEE                                       62529
DETOURNEMENT                                                 0
ANNULATION                                                   0
RAISON D'ANNULATION                                    3225316
RETARD SYSTEM                                          2673943
RETARD SECURITE                                        2673943
RETARD COMPAGNIE                                       2673943
RETARD AVION                                           2673943
RETARD METEO                                           2673943
DATE                                                         0
NIVEAU DE SECURITE                                           0
COMPAGNIE AERIENNE                                      223233
ORIGIN_AIRPORT                                               0
ORIGIN_LIEU                                                  0
ORIGIN_PAYS                                                  0
ORIGIN_LONGITUDE                                             0
ORIGIN_LATITUDE                                              0
ORIGIN_HAUTEUR                                               0
ORIGIN_PRIX RETARD PREMIERE 10 MINUTES                       0
ORIGIN_PRIS RETARD POUR CHAQUE MINUTE APRES 10 MINUTES       0
DESTINATION_AIRPORT                                          0
DESTINATION_LIEU                                             0
DESTINATION_PAYS                                             0
DESTINATION_LONGITUDE                                        0
DESTINATION_LATITUDE                                         0
DESTINATION_HAUTEUR                                          0
DESTINATION_PRIX RETARD PREMIERE 10 MINUTES                  0
DESTINATION_PRIS RETARD POUR CHAQUE MINUTE APRES 10 MINUTES  0
COMPAGNIE                                                    0
dtype: int64
```

```
AEROPORT ARRIVEE                                             0
AEROPORT DEPART                                              0
ANNULATION                                                   0
ARRIVEE PROGRAMMEE                                           0
ATTERRISSAGE                                                 0
CODE AVION                                                   0
COMPAGNIE                                                    0
COMPAGNIE AERIENNE                                           0
DATE                                                         0
DECOLLAGE                                                    0
DEPART PROGRAMME                                             0
DESTINATION_AIRPORT                                          0
DESTINATION_HAUTEUR                                          0
DESTINATION_LATITUDE                                         0
DESTINATION_LIEU                                             0
DESTINATION_LONGITUDE                                        0
DESTINATION_PAYS                                             0
DESTINATION_PRIS RETARD POUR CHAQUE MINUTE APRES 10 MINUTES  0
DESTINATION_PRIX RETARD PREMIERE 10 MINUTES                  0
DETOURNEMENT                                                 0
DISTANCE                                                     0
HEURE D'ARRIVEE                                              0
HEURE DE DEPART                                              0
IDENTIFIANT                                                  0
NIVEAU DE SECURITE                                           0
ORIGIN_AIRPORT                                               0
ORIGIN_HAUTEUR                                               0
ORIGIN_LATITUDE                                              0
ORIGIN_LIEU                                                  0
ORIGIN_LONGITUDE                                             0
ORIGIN_PAYS                                                  0
ORIGIN_PRIS RETARD POUR CHAQUE MINUTE APRES 10 MINUTES       0
ORIGIN_PRIX RETARD PREMIERE 10 MINUTES                       0
RETARD A L'ARRIVEE                                           0
RETART DE DEPART                                             0
TEMPS DE DEPLACEMENT A TERRE A L'ATTERRISSAGE                0
TEMPS DE DEPLACEMENT A TERRE AU DECOLLAGE                    0
TEMPS DE VOL                                                 0
TEMPS PASSE                                                  0
TEMPS PROGRAMME                                              0
VOL                                                          0
index                                                        0
dtype: int64
```

# Base de données

## Variable à prédire

➔ **Vol retardé au départ:**

Une variable qui permet d'identifier si un vol a été retardée ou pas(>15min).

➔ **Retard de départ :**

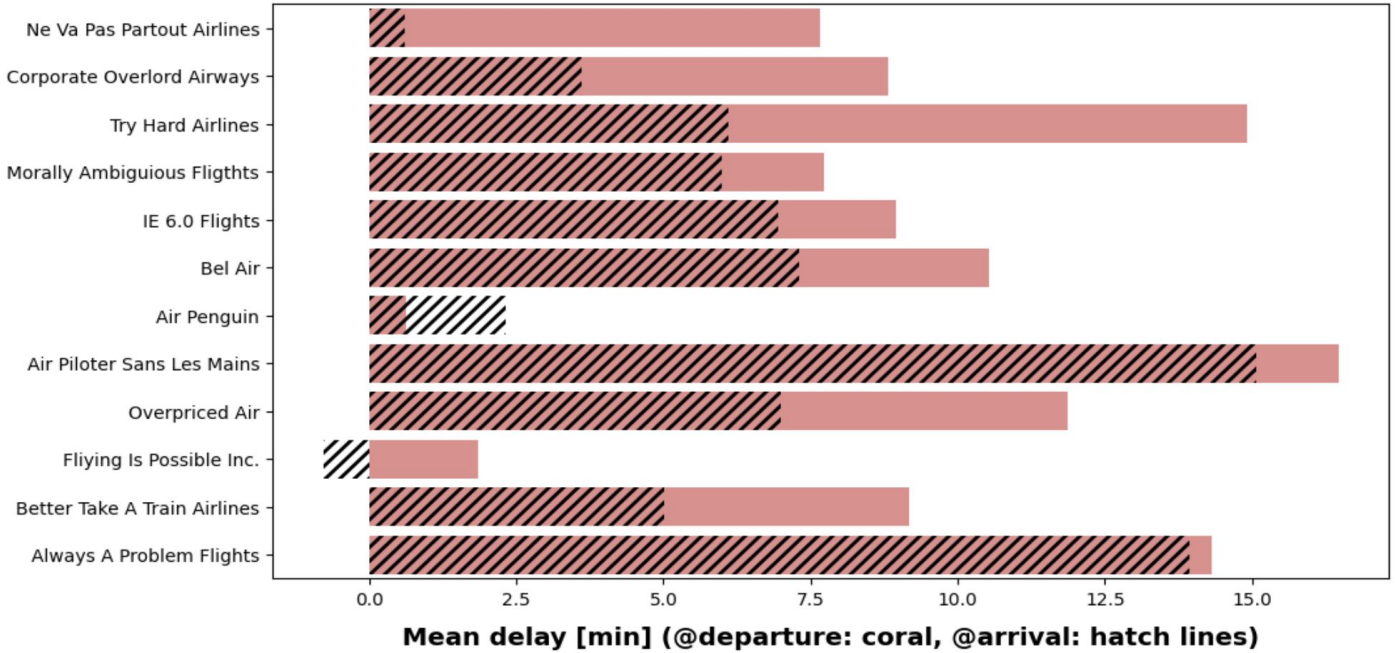
Une variable avec la durée de retard effectuée au départ.

Le rationnel derrière notre choix de la variable cible se justifie par :

- Client : les retards de départs représentent la plus grande frustration chez les clients
- Compagnies aériennes : le retard départ justifié généralement le retard d'arrivée et donc l'entreprise doit surveiller les retards de départ en priorité .

# Analyse descriptive

## Retard : à l'arrivée ou au départ ?

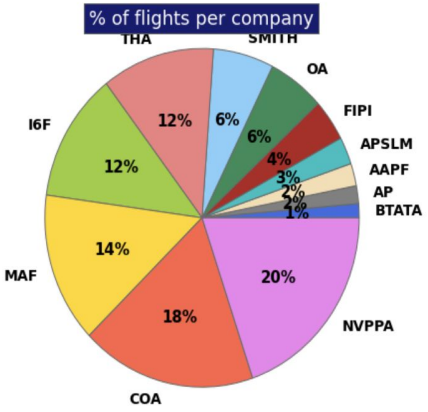


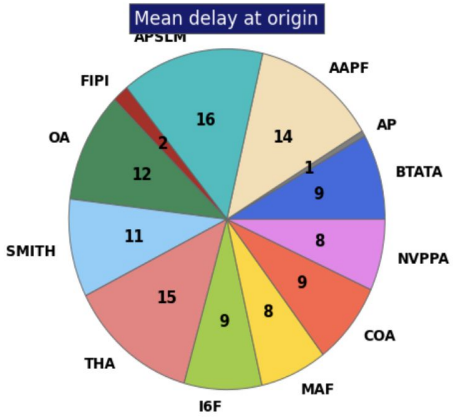

Sur cette figure, on peut constater que les retards à l'arrivée sont généralement plus faibles qu'au départ. Cela indique que les compagnies aériennes adaptent leur vitesse de vol afin de réduire les retards à l'arrivée. Dans ce qui suit, nous nous somme contenter d'examiner les retards au départ, ce qui justifie également notre choix de la variable a prédire.

# Analyse descriptive

## Compagnies aériennes

| COMPAGNIE AERIENNE | min | max | count | mean |
|---|---|---|---|---|
| BTATA | -24.0 | 644.0 | 40139.0 | 9.181370 |
| AP | -27.0 | 1433.0 | 51128.0 | 0.624804 |
| AAPF | -46.0 | 996.0 | 64261.0 | 14.317160 |
| APSLM | -35.0 | 836.0 | 78258.0 | 16.468259 |
| FIPI | -82.0 | 963.0 | 119255.0 | 1.862580 |
| OA | -26.0 | 1006.0 | 174731.0 | 11.860059 |
| SMITH | -36.0 | 1278.0 | 186921.0 | 10.524452 |
| THA | -24.0 | 1314.0 | 347795.0 | 14.929375 |
| I6F | -55.0 | 1236.0 | 371833.0 | 8.955644 |
| MAF | -48.0 | 1215.0 | 431393.0 | 7.726720 |
| COA | -68.0 | 1988.0 | 538897.0 | 8.819524 |
| NVPPA | -61.0 | 1289.0 | 592933.0 | 7.661872 |

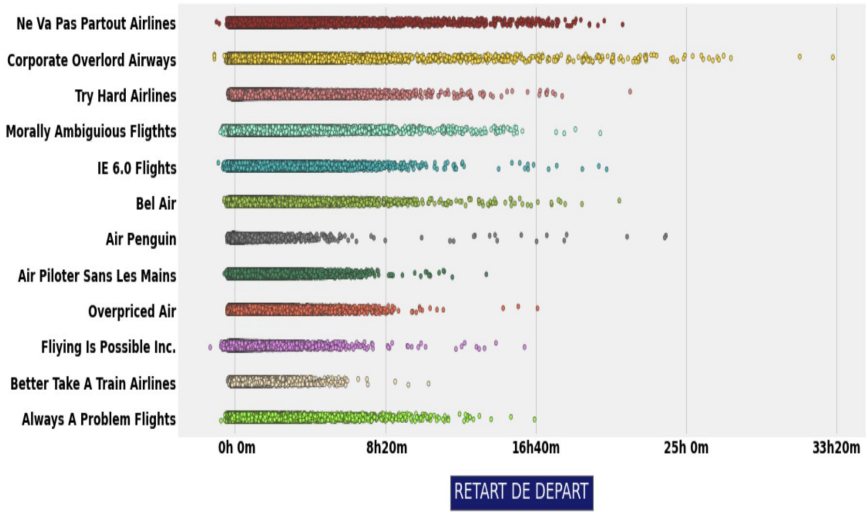



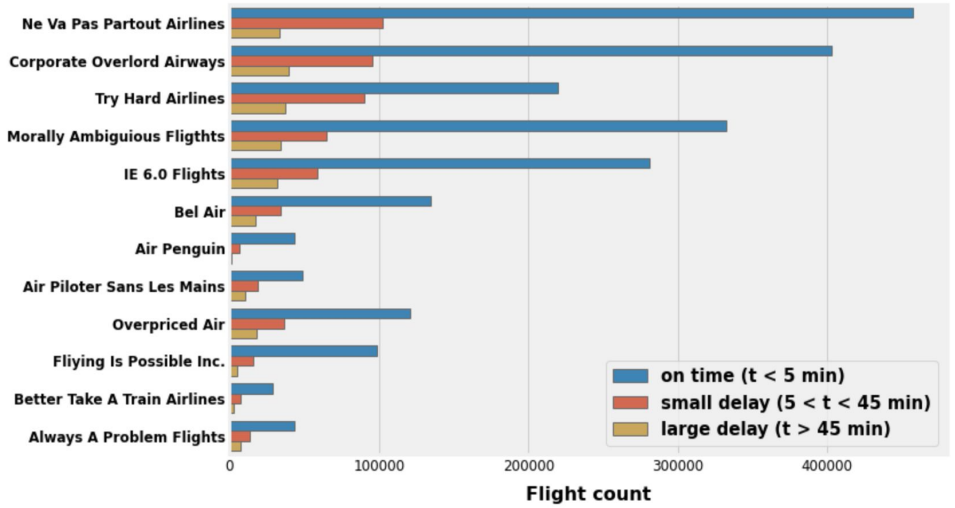

- Pie chart 1: il y a une certaine disparité entre les transporteurs.Par exemple, NVPPA représente ~ 20% des vols, ce qui est similaire au nombre de vols affrétés par les 6 plus petites compagnies aériennes.
- Pie chart 2: les différences entre les compagnies aériennes sont moins prononcées quand il s'agit des retards moyens.

# Analyse descriptive

## Compagnies aériennes




Ce graphique représente toutes les valeurs déclarées pour le retard. nous constatons qu'occasionnellement, nous pouvons être confrontés à des retards vraiment importants qui peuvent atteindre quelques dizaines d'heures !

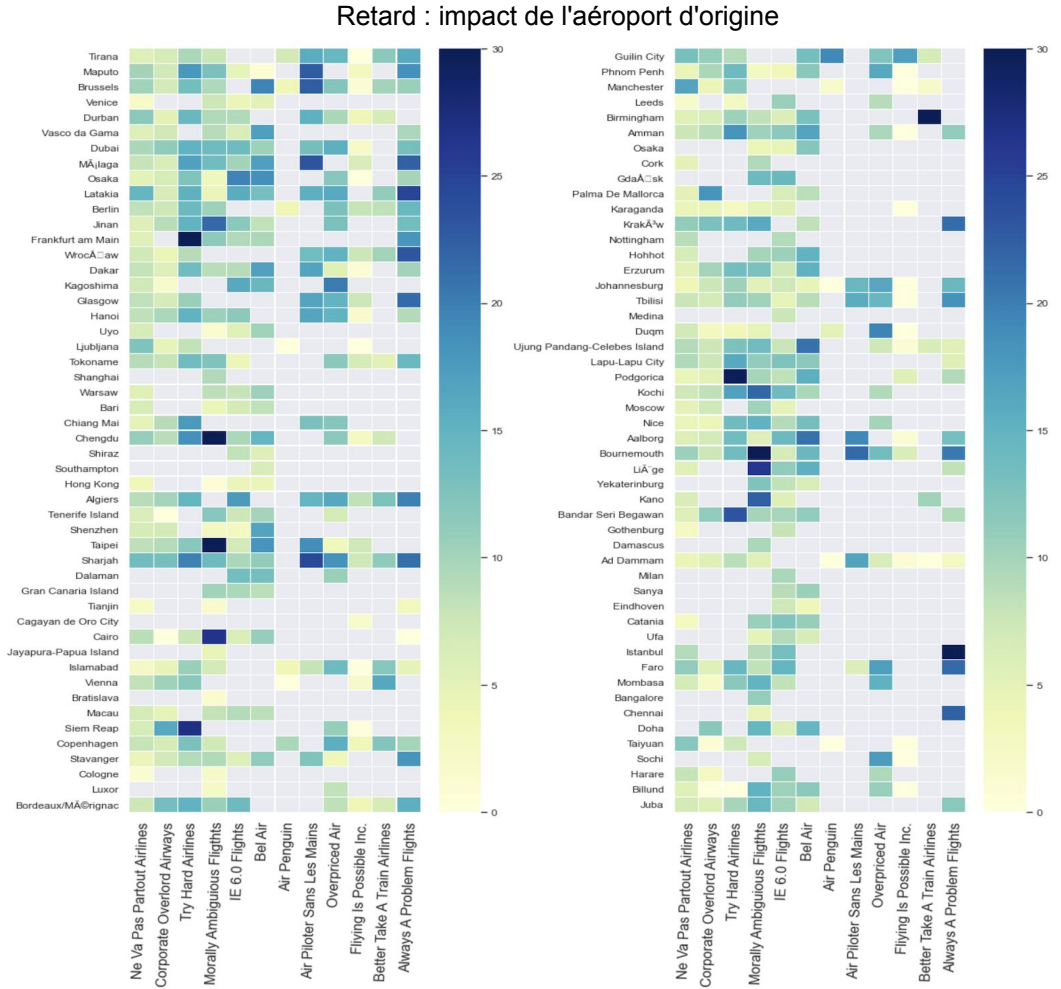



On peut donc dire que indépendamment de la compagnie aérienne, les retards supérieurs à 45 minutes ne représentent que quelques pourcents.Cependant, la proportion des retards dans ces trois groupes dépend de la compagnie aérienne.

# Analyse descriptive

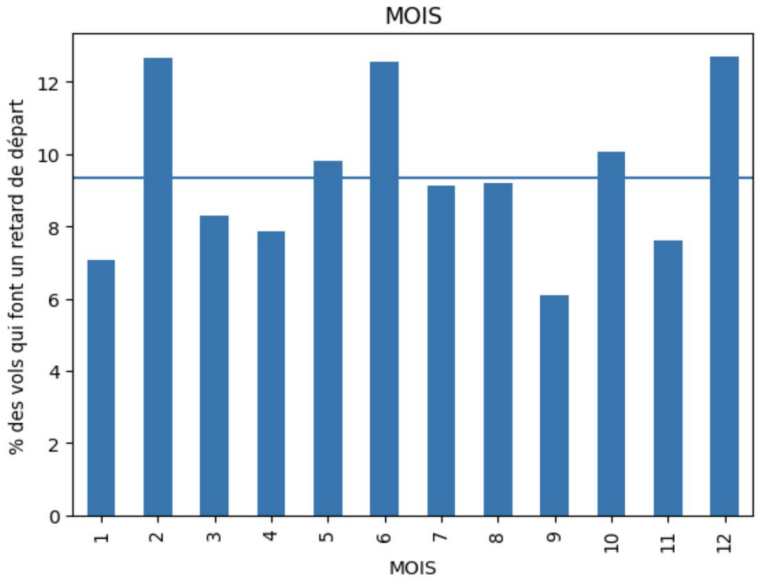
## Relation entre l'aéroport d'origine et les retards

- Ce figure nous permet de tirer quelques conclusions, si nous considérons par exemple le panel à gauche nous verrons que la colonne associée “Always A Problem flights” signale principalement des retards importants. tandis que la colonne associée à “ Ne Va Pas Partout” est principalement associée à des retards de moins de 5 minutes.
- Si nous examinons maintenant les aéroports d'origine, nous verrons que certains aéroports favorisent les départs tardifs : voir par exemple ‘Sharjah’, ‘Bournemouth’ etc. Inversement, d'autres aéroports connaissent surtout des départs à l'heure, comme ‘Southampton’ ou ‘Cologne’.
- Nous  pouvons déduire de ces observations qu'il existe une grande variabilité des retards moyens, à la fois entre les différents aéroports mais aussi entre les différentes compagnies aériennes.

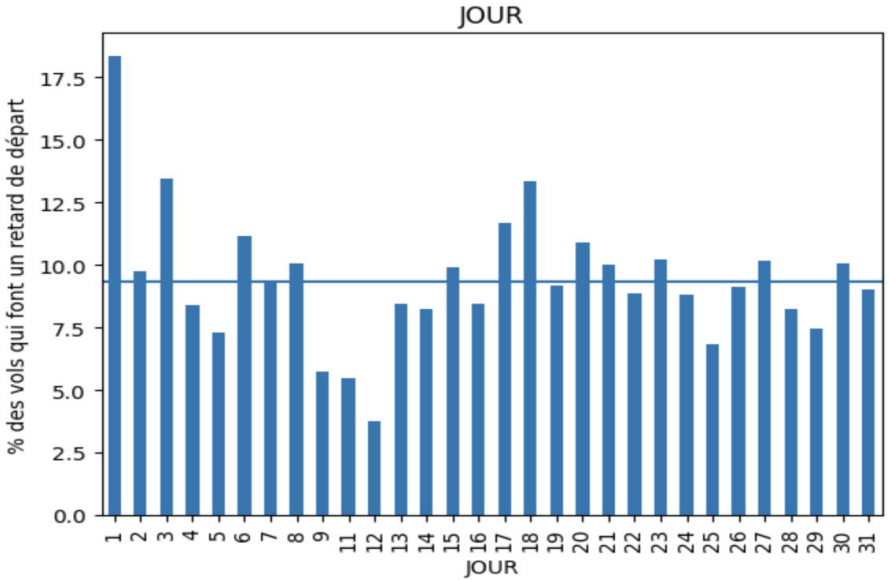


# Analyse descriptive

## Retard par mois et par jour du mois




Les retards varient considérablement en fonction du mois, par exemple les retards sont plus faibles en automne.




Les retards sont plus élevés durant la première semaine et plus au moins faible la dernière semaine du mois

# Modélisation

Hypothése

Les retards des vols sont causés par les facteurs suivants.Les retards des vols peuvent être prédits à l'aide des facteurs suivants

- Météo - Jour du vol - Compagnie - Infrastructures aéroportuaires - heure programmée - saisonnalité - l'état de l'avion..

le modèle suppose que les vols sont sans escale. Nous ne traitons pas les vols de correspondance dans ces données.
le modèle de classification considère qu'un vol est retardé lorsqu'il est 15 minutes plus tard que l'heure prévue.

Critere de selection

- Pertinence des données
- Disponibilité des données
- Précision du modèle et signification statistique

Variables choisies

**Considéré mais abandonné**

- TEMPS DE VOL
- TEMPS PROGRAMME

Raison :

- n'ayant pas d'impact significatif sur la précision du modèle

**Sélectionné**

- ARRIVEE PROGRAMMEE
- ANNEE,MOIS,JOUR,
- DEPART PROGRAMME
- DISTANCE
- AEROPORT ARRIVEE
- AEROPORT DEPART
- COMPAGNIE AÉRIENNE

**Important mais manquant**

- INFORMATIONS MÉTÉOROLOGIQUES HORAIRES
- CONGESTION DES PISTES
- L'ÉTAT DE L'AVION

# Modélisation

## Classification :

- Choisir la variable cible Y : **vol retardée**

```
Entrée [96]: df_vols['Dep_retard15']=df_vols['RETART DE DEPART'].apply(lambda x:1 if x > 15 else 0)
```

- Preprocessing : transformer les variable catégorielle en variable numérique compréhensible par le modèle de classification : en utilisant le concept du One-Hot encoding
- Division des données en base de train et test :

  ➔ Pour ce modèle, nous avons fait une répartition de 70% train, 15% validation, 15% test

- Construire un modèle de **régression logistique :**

  ➔ Car ce modèle est généralement facile à interpréter, et nous permet de savoir quelles sont les plus variables importantes

- Sortir des scores pour évaluer notre modèle

# Modélisation

## Regression :

- Choisir la variable cible Y : **durée de retard du vol**
- Preprocessing : transformer les variable catégorielle en variable numérique compréhensible par le modèle de régression
- Division les donnée en base de train et test
- Construire un modèle de **régression linéaire**
- Sortir des scores pour évaluer notre modèle
- Construire d'autres model : **Random forest** et **XGBRegressor**
- Sortir les scores corresponds à ces models

# Evaluation des performances

## Regression :

**R2_Squared**

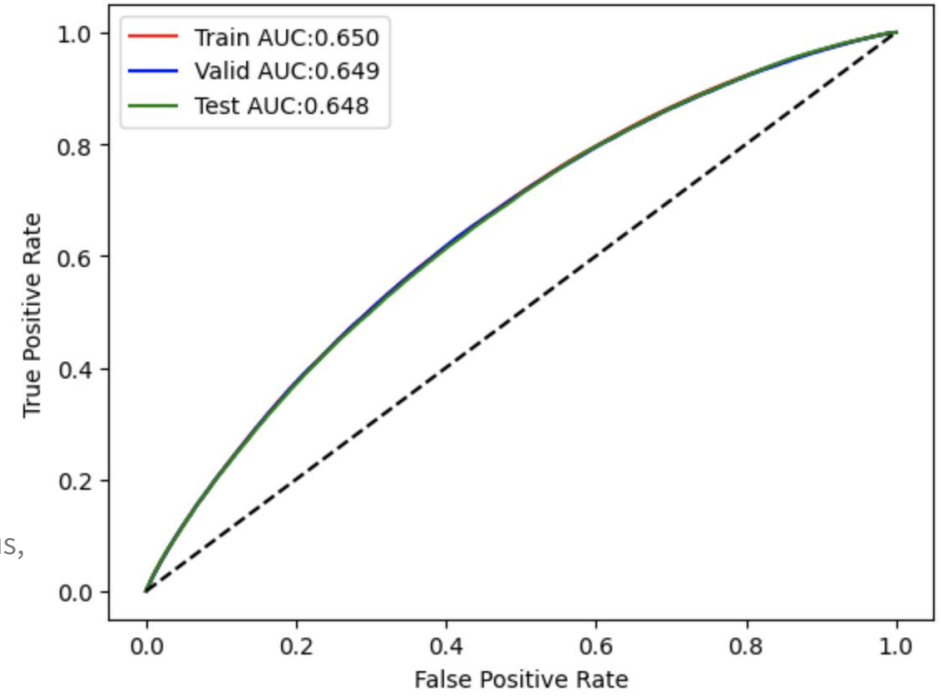
| Linear regression | Random Forest | XGBRegressor |
|---|---|---|
| 0.52 | 0.62 | 0.62 |

- A partir des scores obtenues de la régression linéaire et en faisant la prédiction de la durée de retard on peut conclure que ce modèle a un taux de performance de 52%.
- Nous avons construire d'autres model pour améliorer nos scores .

# Evaluation des performances

## Classification :

| | Training | Validation | Test |
|---|---|---|---|
| AUC | 0.650 | 0.649 | 0.648 |
| Accuracy | 0.608 | 0.601 | 0.600 |
| Recall | 0.620 | 0.621 | 0.618 |
| Precision | 0.606 | 0.246 | 0.244 |
| Specificity | 0.597 | 0.597 | 0.596 |




Ce modèle a permis d'expliquer 62 % des retards au départ. De plus, il a été observé que le retard des vols dépendait fortement des aéroports de départ( Cf. Features Importance ). Cela implique clairement que si un aéroport est occupé et c'est un aéroport important, les chances de retard des vols seront plus élevées par rapport aux autres aéroports.

# Solution Proposé

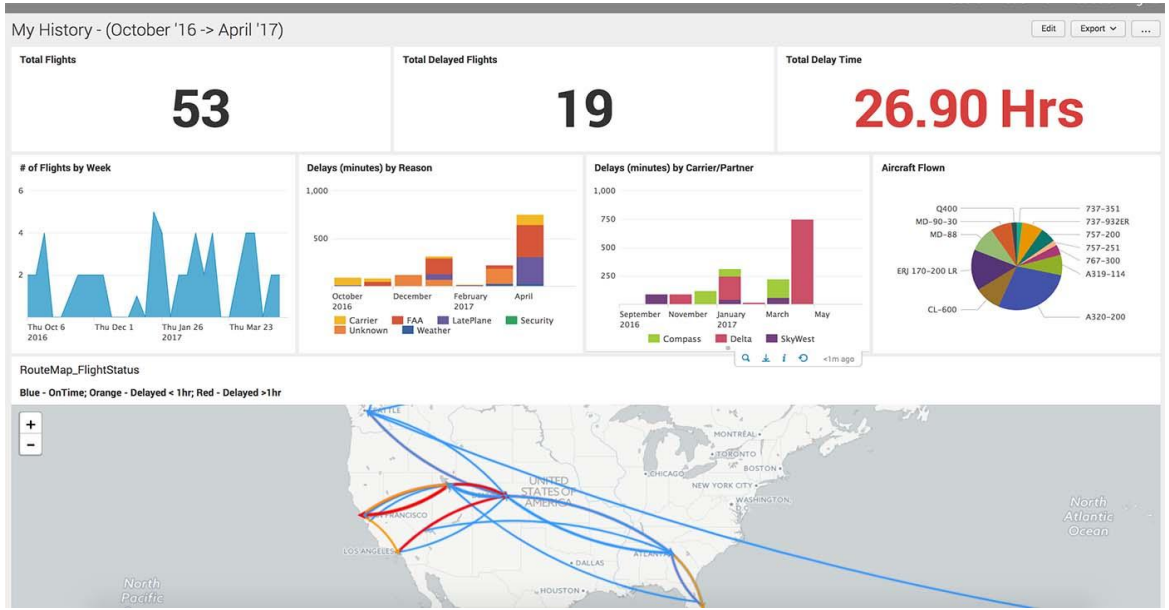
# Solution voyageurs

Nous allons permettre à vos clients de consulter à tout moment une platform en ligne de prédiction du retards des vols.


- Permet aux voyageurs d’une compagnie de voir si leurs vol sera retardé ou non.
- Obtenir une information immédiate et  mesurable afin d'éviter le mécontentement des clients .

# Solution  compagnies

Votre personnels aura un accès plus détaillé avec une vue d'ensemble sur tous les retards de la compagnie et des prediction sur 3 jours.




- Un tableau de bord qui résume l'état des vols retardées .
- Un système d'alerte qui se base sur le modèle de prédiction permet de réagir rapidement et éviter  les retards.

# Conclusion

le CRM est une révolution dans la stratégie d'entreprise, notamment parce qu'il permet une différenciation basée sur des différences autres que le prix.

Par conséquent, de nombreuses entreprises sont entrées dans une stratégie «sur mesure», combinant différents méthode pour fidéliser leurs clients et augmenter leurs satisfaction .