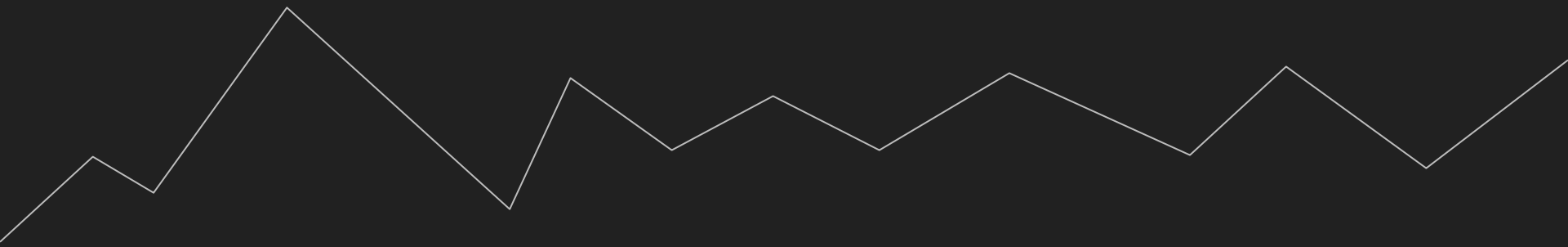
Flight delay consulting

M2 IMSD - Projet de CRM 2020

Jérome O’keeffe
Samir LAZZALI



Objectif de l’étude & contexte

I. Analyser et comprendre les données
II. Mettre à jour le modèle de prédiction des retards
III. Mettre en perspective la rentabilité des compagnies et des aéroports grâce à une classification

Nous sommes spécialiste dans l’analyse des données :

- Visualisation

- Prédiction de retards et d’annulations de vols

Lebonvol.fr

Bla Bla Jet

Creuse Air Line

Périmètre des données

Aéroports :

- 319 aéroports
- 111 pays

Compagnies :

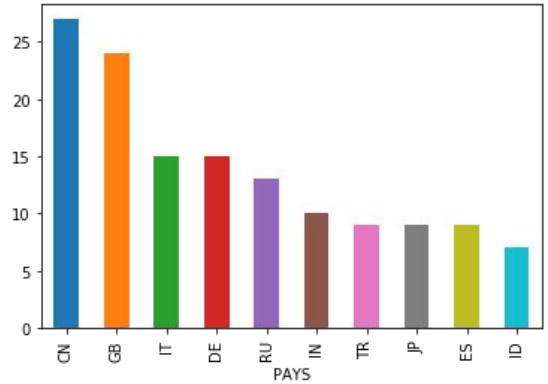
- 13 compagnies

Vols :

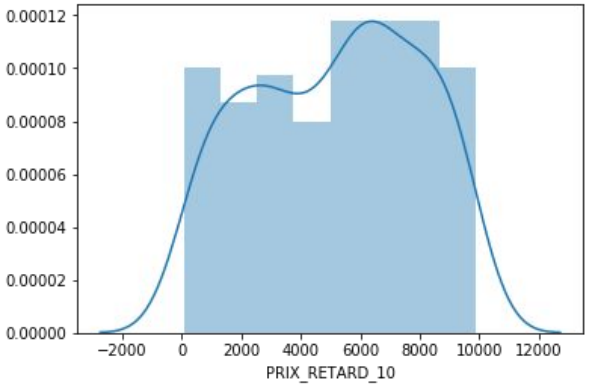
- 3.000.000 vols

Nous avons ensuite fusionné nos données afin de centraliser toutes les informations

Pays ayant le plus d'aéroport :



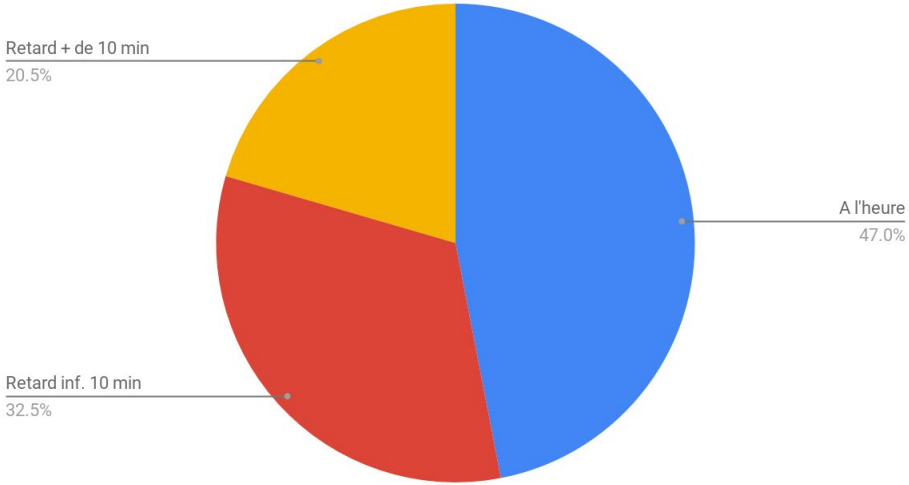
Distribution du prix des 10 premières minutes de retard :



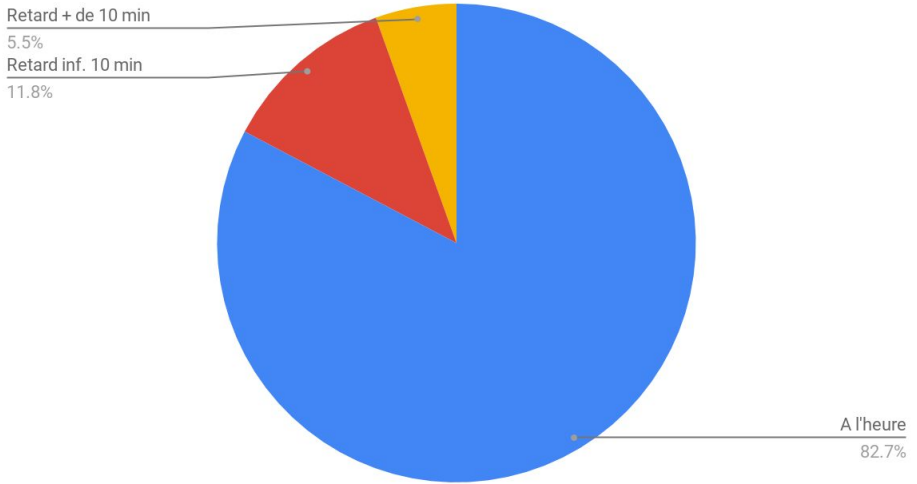
Analyse descriptive

Zoom sur les retards :

Retards à l'arrivé pour les avions partit en retard :



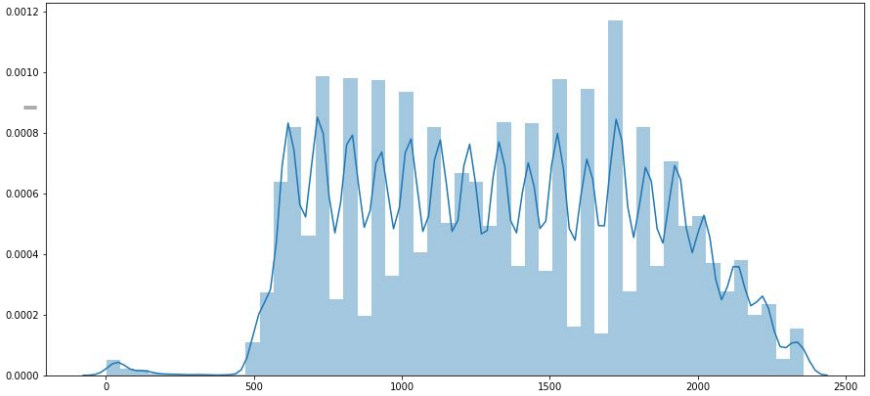
Retards à l'arrivé pour les avions partit à l'heure ou en avance :



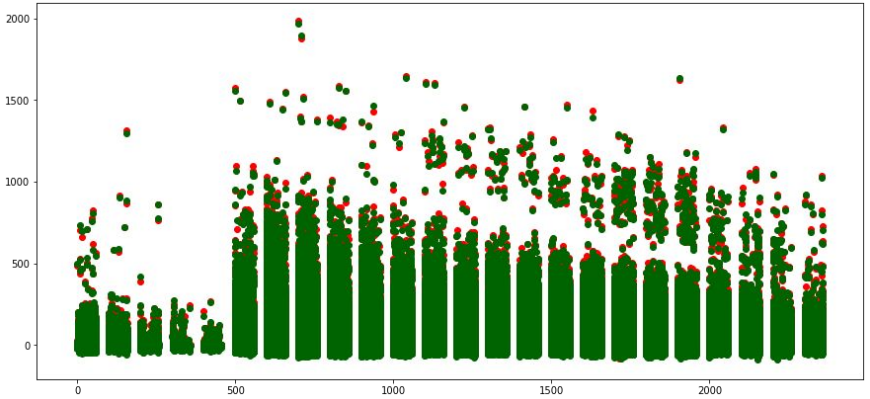
Analyse du dataset de *vols*

- 4185 code_avion différents
- 6380 valeurs différentes dans la colonne vol

Distribution des temps de départ



Valeurs des retard (**depart**, **arrivée**)



Classification des retards

- Les angles d’analyse possibles
- Analyse des retards
- Choix du modèle
- Métriques et résultats

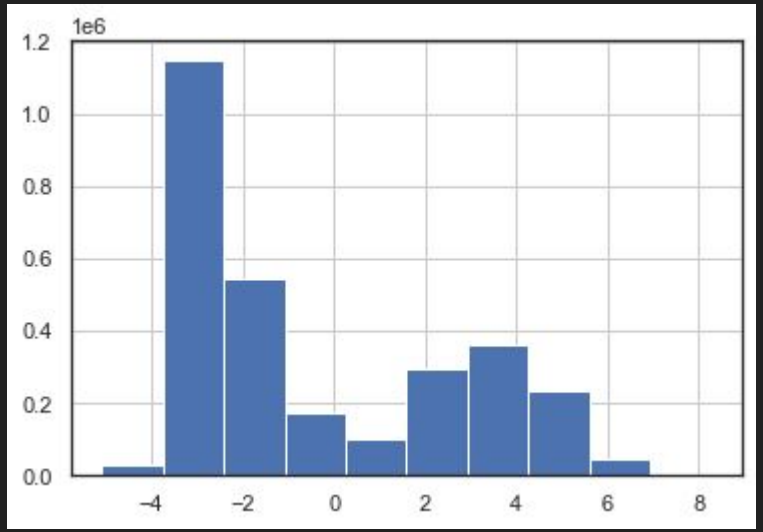
Les angles d’analyse possibles

- Etude de la politique tarifaire des aéroports
- Etude des différents avions et de leur possible impacte sur les retards
- Etude des compagnies aériennes vis-à-vis des amendes

Les objectifs de IMSD2020 :

- Aider les entreprises à modéliser les vols à risque
- Optimiser les frais liés aux amende

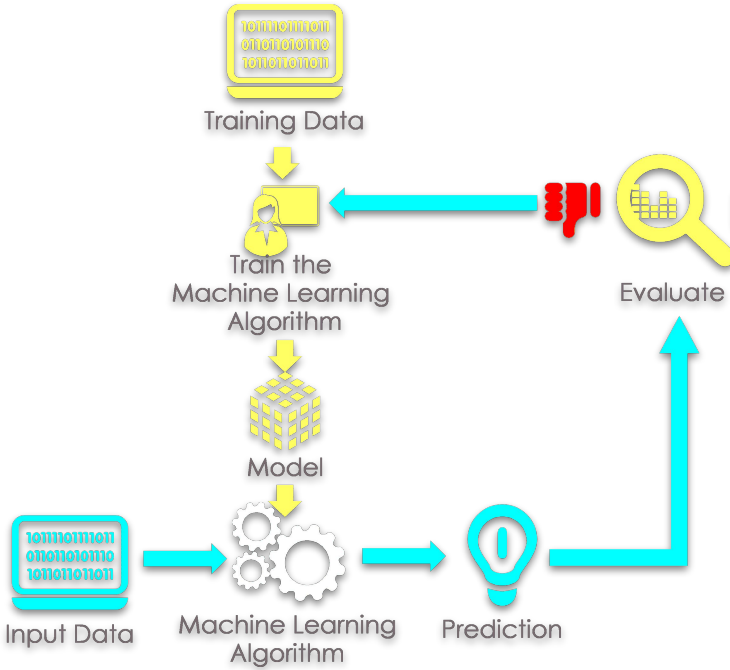
Analyse de la distribution des retards (log)



- On peut voir que la plupart des vols arrivent en avance sur le temps prévu (63%)
- Le temps de retard minimum est - 162 minutes (3 heures d’avance)
- Le retard maximum quant à lui est de 3959 minutes (plus d’un jour)

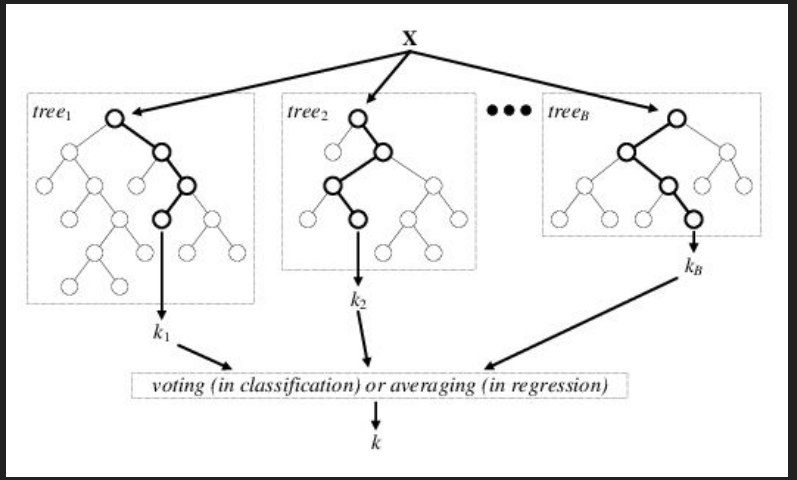
Les différents modèles utiliser

- RandomForestClassifier
- KNeighborsClassifier
- SGDClassifier

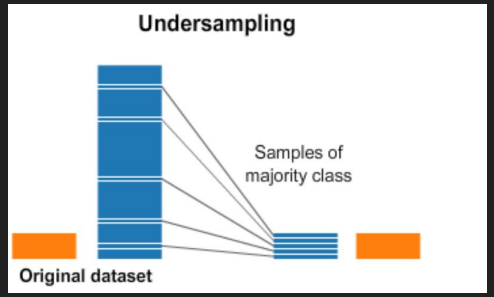


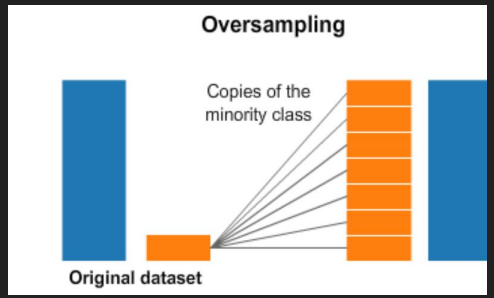
RandomForestClassifier

- Génération de 100 arbres avec des sous-ensembles de la population et des dimensions
- Moyennage des résultats et renvoie de la classe associée



Resampling

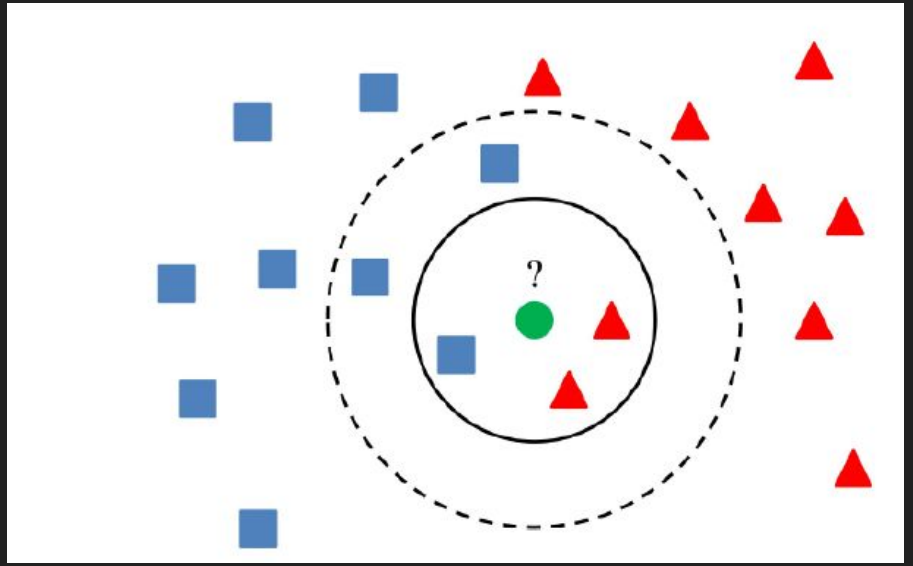




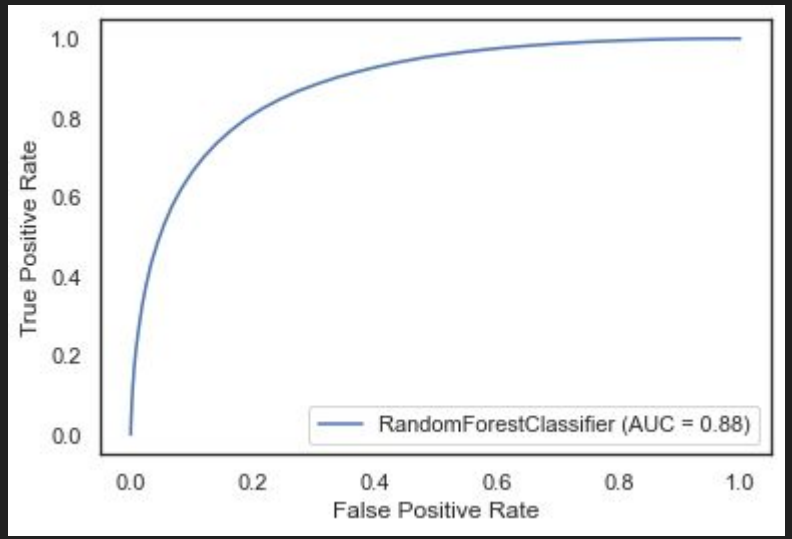
- Lors du premier RandomForestClassifier, on a remarqué qu'il répond beaucoup sourvent plus la classe la plus présente
- Le rééchantillonnage (resampling) permet d'avoir le même nombre de ligne pour chaque modalité en ajoutant ou supprimant des lignes suivant une stratégie

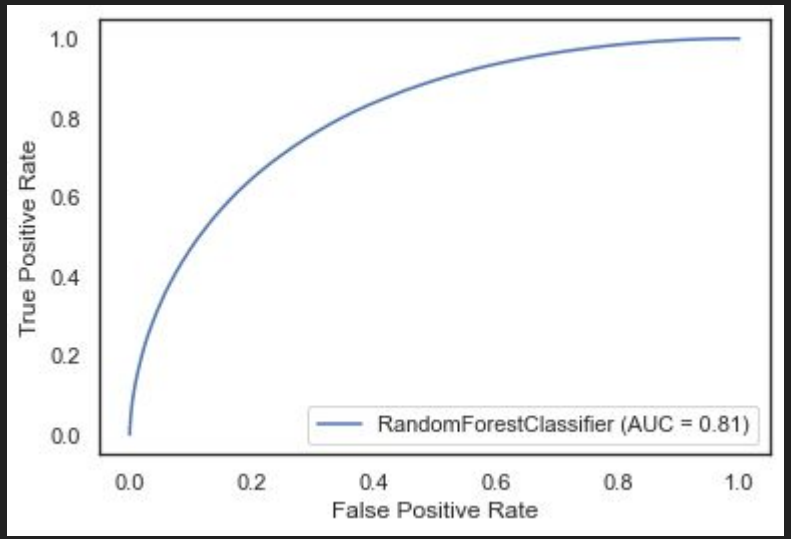
KNeighborsClassifier

- Classification qui utilise les K voisins plus proche et qui effectue un vote
- Problème de performance car il faut interroger la base à chaque prédiction



Courbe ROC sans vs avec resampling





Autres métriques sans resampling

	Précision	Recall	F1
Pas de Retard	0.92	0.81	0.71
Retard	0.63	0.81	0.86
Accuracy			0.82

Autres métriques sans resampling

	Précision	Recall	F1
Pas de Retard	0.50	0.73	0.80
Retard	0.89	0.75	0.62
Accuracy			0.75

Conclusion sur la classification

- L'entraînement sans échantillonnement produit de meilleur résultat
- Les scores sont bon mais étant donné la volumétrie des données, il aurait fallu qu'on utilise un échantillon du dataset pour réduire les temps de développement et permettre l' étude d'autres axes
- On s'est attaqué à un problème qui est la modélisation des retards mais le dataset aurait aussi permis d'avoir beaucoup d'information sur les avions, les pays, le pricing des aéroports et d'autres informations
- Ceci n'est qu'un échantillon des services de l'entreprise IMSD2020, on compte sur vous pour nous permettre de créer de la valeur avec vos données dans le futur.

Clustering des compagnies

- Pre-processing des données
- Analyse des retards
- Choix du modèle
- Analyse et interprétation

Aéroports et des compagnies les moins rentable

Nous avons calculé pour chaque le coût des retards. Ce nuage de mot permet de mettre en avant ceux qui paient le plus :

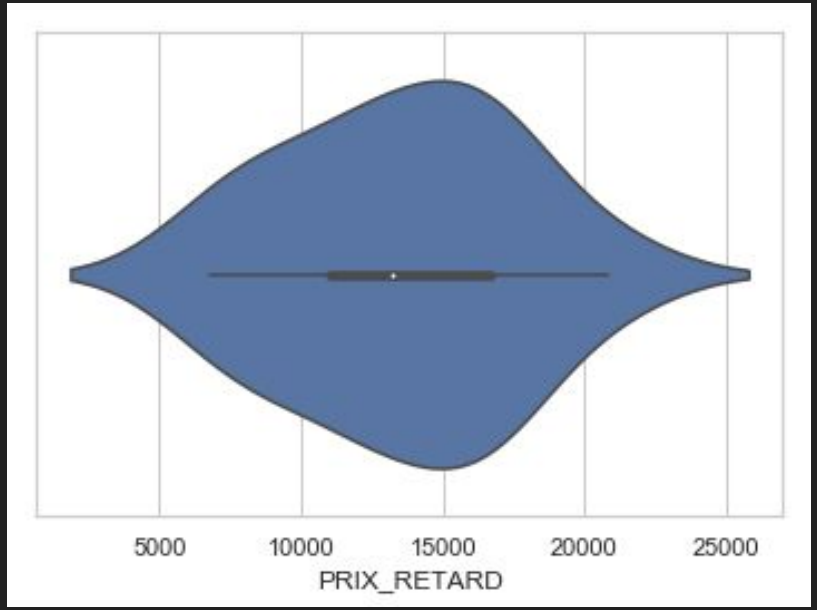
Les aéroports :

Les compagnies

Pre-processing des données

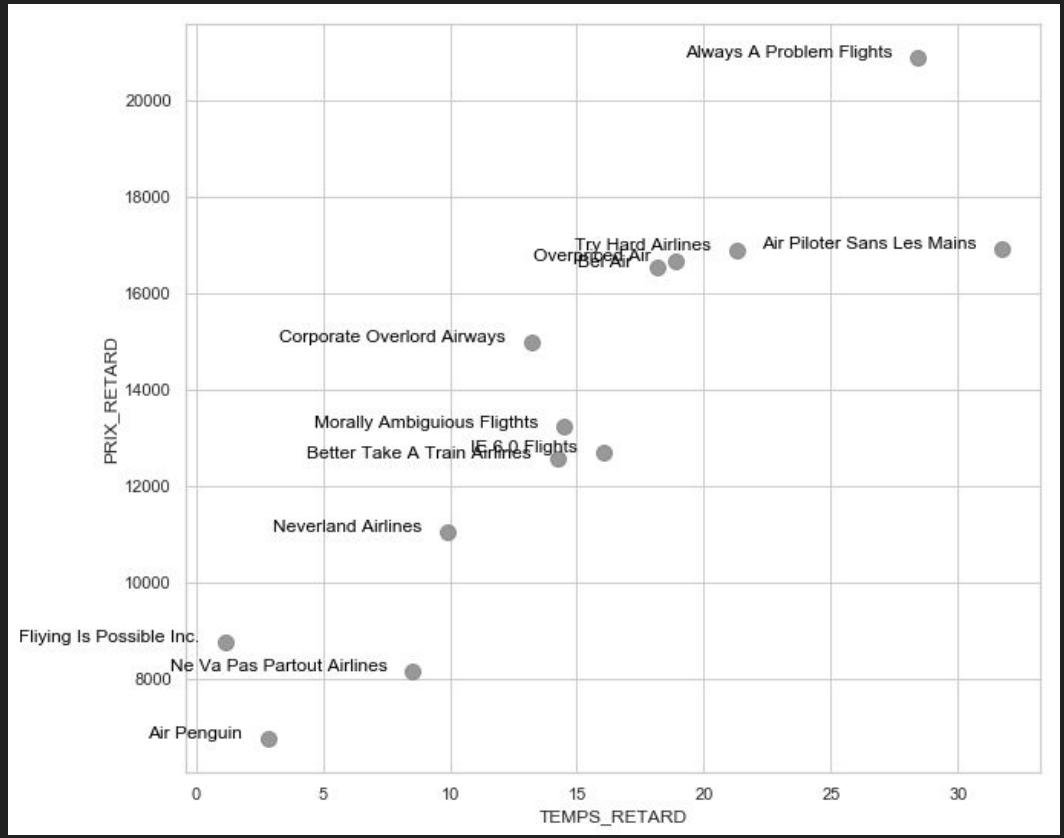
- Calcul du prix des amendes au départ et à l'arrivée, à partir du temps du retard ainsi que du prix avant et après 10 min.
- Agrégation de nos données fusionnées sur les compagnies
- Cumul du temps des informations au départ et à l'arrivée. (La compagnie est responsable de l'ensemble du vol.)

Distribution du prix de retard :



Clustering des compagnies

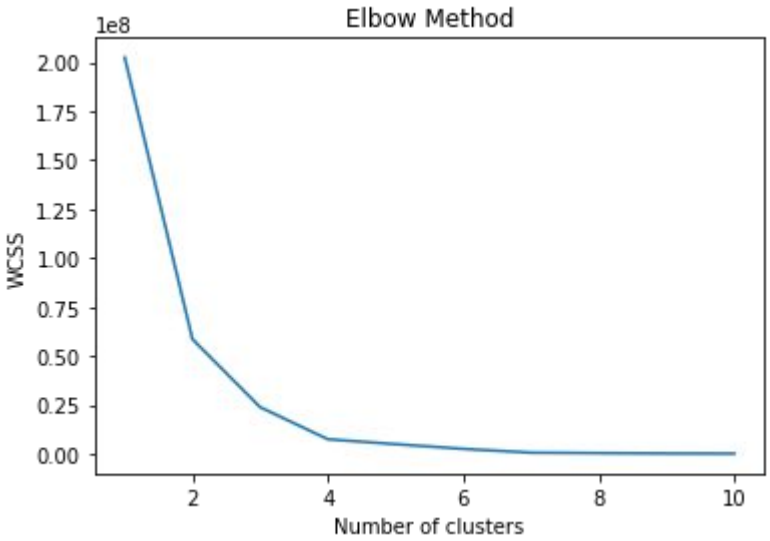
Première itération, cas simple basé sur le **prix des retards** et sur le **temps des retards**.



Processus de clustering

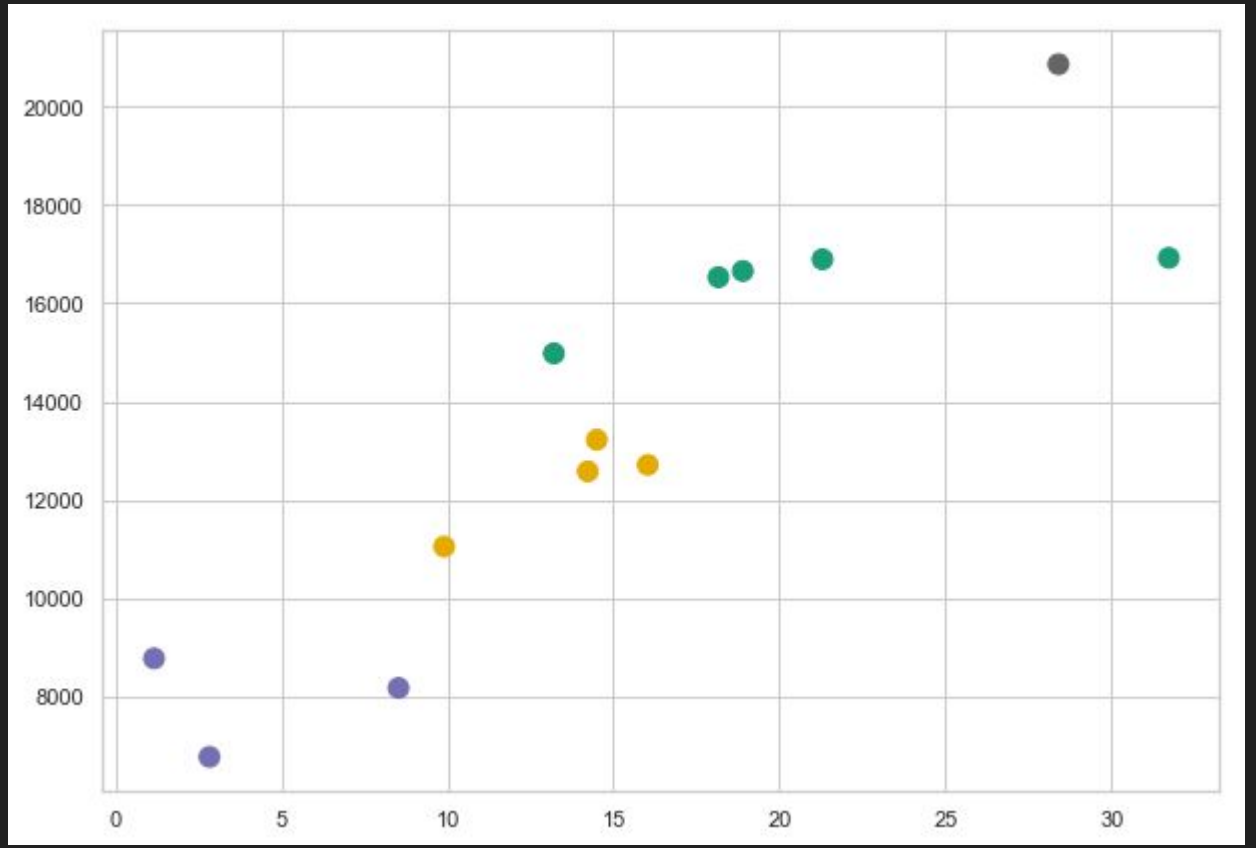
Algorithme : K-means
Nous avons utilisé la technique du coude pour déterminer le nombre de cluster optimal. Nous réaliserons 4 clusters.

Le but est de minimiser la variance au sein des groupes et maximiser celle entre les groupes.



Résulats

Création de cluster qu'il nous faut qualifier avec les experts métiers.



Résulats

Groupe	COMPAGNIE	Temps retard moyen (min)	Prix retard moyen ($)
1	Air Penguin, Fliying Is Possible Inc., Ne Va Pas Partout Airlines	4.1	7904.5
2	Better Take A Train Airlines, IE 6.0 Flights, Morally Ambiguious Fligthts,Neverland Airlines	13.6	12390.6
0	Air Piloter Sans Les Mains, Bel Air,Corporate Overlord Airways, Overpriced Air, Try Hard Airlines	20.6	16397.3
3	Always A Problem Flights	28.4	20864.5

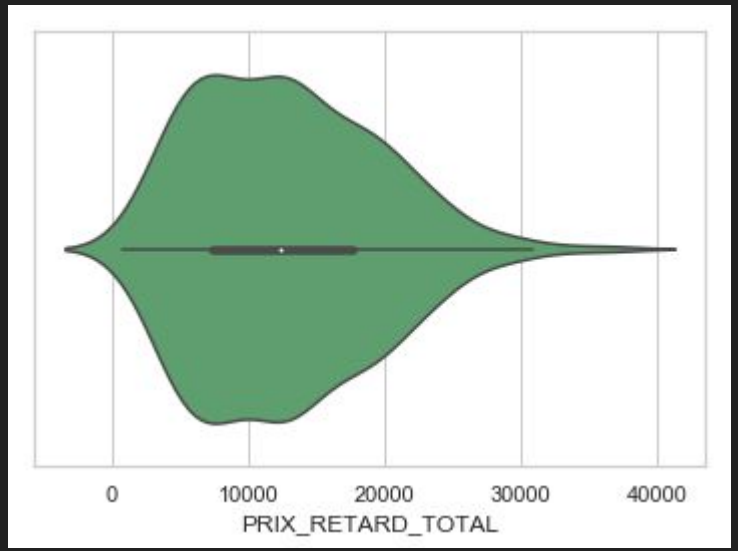
Clustering des aéroports

- Pre-processing des données
- Analyse des retards
- Choix du modèle
- Analyse et interprétation

Pre-processing des données

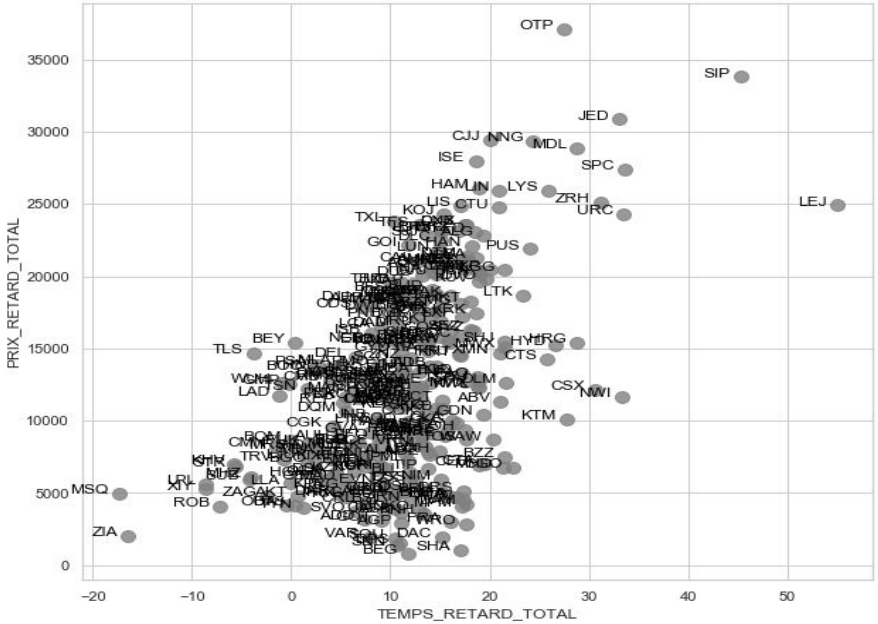
- Séparation des aéroport de départ et d'arrivée pour un même vol (retard à l'arrivée ainsi que son coût associé)
- Agrégation pour obtenir toutes les informations de l'aéroport lorsqu'il était à l'arrivée ou au départ du vol
- Calcul du temps de retard total et du coût total des retards pour chaque aéroport

Distribution du prix des retards par aéroport :

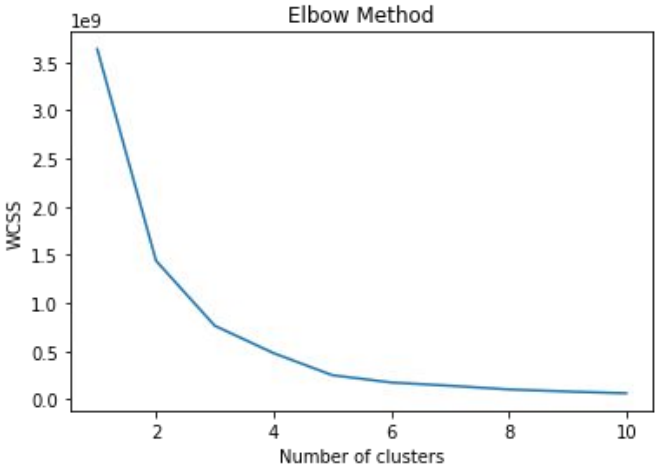


Processus de clustering

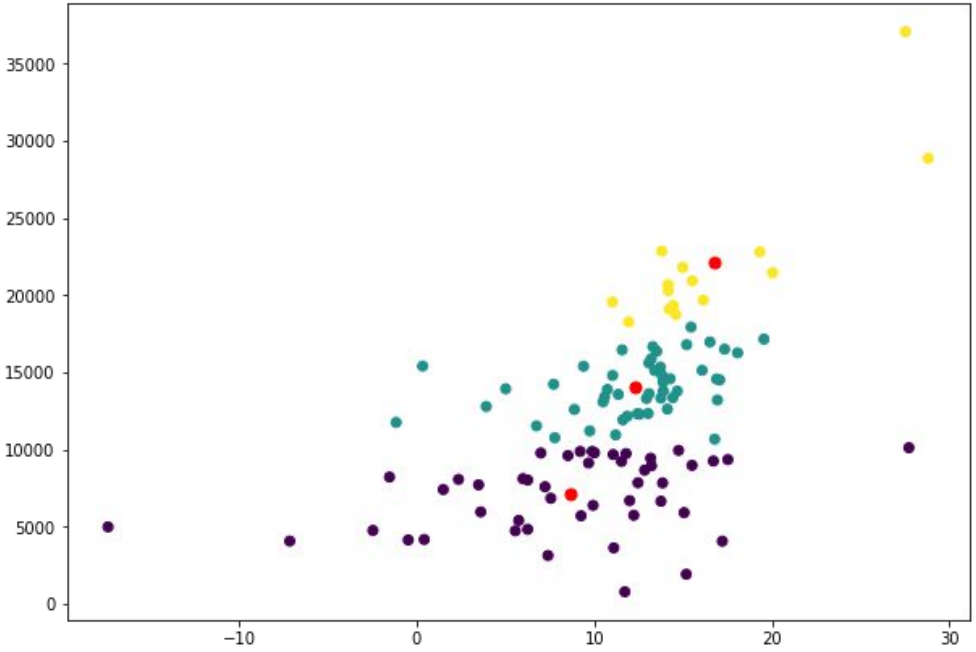
Répartition des aéroports par temps de retard et prix du retard :



Même principe pour la règles du coude. Nous allons créer 3 cluster :



Résultats



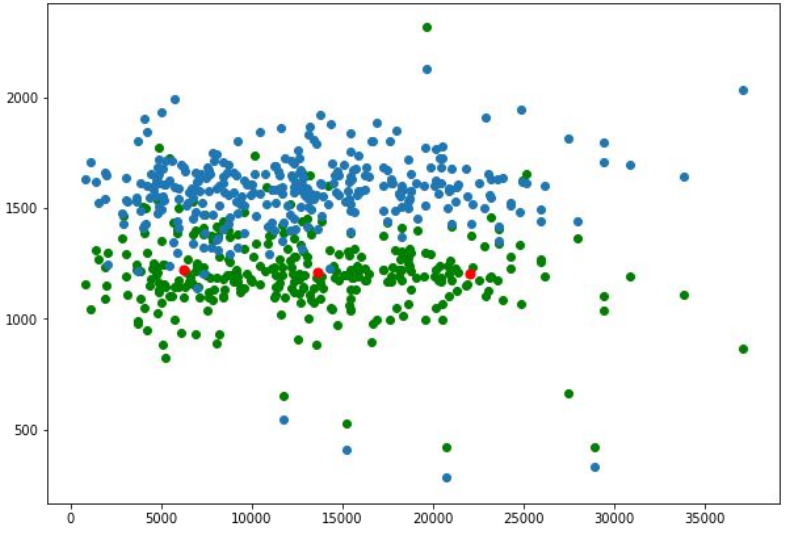
Nous avons donc 3 groupes d'aéroport :

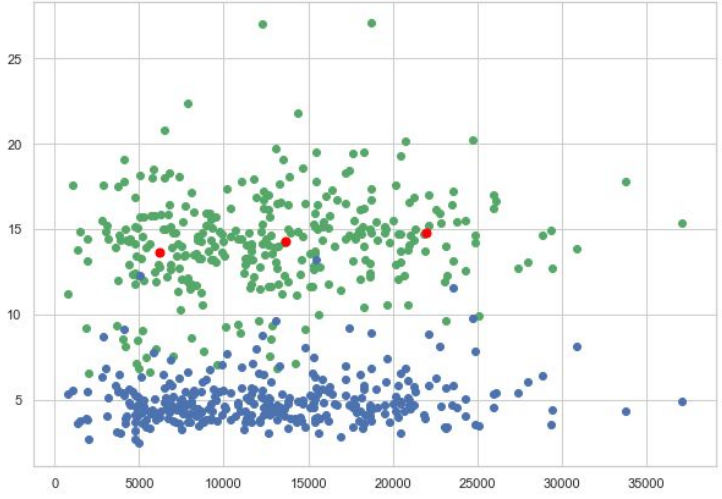
Nom du groupe	Temps de retard moyen (min)	Prix de retard moyen ($)
A (violet)	8,6	6233
B (vert)	12.5	13632
C (jaune)	17.9	21979

Pour aller plus loin

Le clustering n'a ici été réalisé que sur 2 variables, mais il pourrait être intéressant d'en inclure plus comme l'heure de départ et d'arrivée programmé :

Ou encore le temps de déplacement au départ ou à l'arrivée. Ce sont des variables pertinentes pour expliquer les retards :





Conclusion du clustering

L’objectif était de mieux cibler les aéroports et les compagnies défectueuses sur le plan des retards.

Les conseils de IMSD2020 :

- Rediriger des vols pour ne pas encombrer les aéroports les plus en retard
- Privilégier certaines compagnie plutôt que d’autre afin de limiter les coûts liés aux retard comme : *Air Penguin, Fliying Is Possible Inc., Ne Va Pas Partout Airlines* du groupe 1 ou encore celles du groupe 2.