CentraleSupélec

OPENCLASSROOMS

Anticiper le retard des vols

Synthèse

Contexte :

Air Data, une nouvelle compagnie aérienne désire optimiser sa logistique et anticiper les retards possibles de sa flotte. A partir de données existantes sur d'autres compagnies aériennes, celle-ci nous demande de mettre en place un modèle de régression afin de prédire les possibles retards/avances.

Problème :

Ce problème est un problème de régressions. A l'aide de différents modèles, une évaluation de la prédiction sera faite. Par la suite, une optimisation sera faite sur les hyperparamètres des modèles ainsi qu'avec du Boosting. L'objectif majeur étant de prédire les courts retards fréquents.

Données :

Les données fournies sont des données issues d'une base publique gouvernementale (www.transtats.bts.gov). Ce site regroupe les données de chaque vol intérieur aux USA par mois sur les 30 dernières années. Dans notre étude, nous avons à notre disposition les données de l'année 2016.

Approche :

Après un nettoyage des données inutiles dans les différents datasets, différents modèles vont être testés sur différentes configurations des features. Les modèles seront ensuite évalués sur le MAE et MSE en cas d'égalité. Une recherche des meilleurs hyperparamètres sera faite à l'aide de Grid Search pour chacun d'entre eux.

Performances des modèles :

Lors de ce projet, un des problèmes majeurs a été la performance. Beaucoup de modèles ne passent pas en mémoire sur les Notebook et ont donc été découpés dans des scripts. Concernant l'évaluation, le MAE a été le critère principal. L'objectif étant de prévoir majoritairement les petits retards facilement anticipables plutôt que les gros retards potentiellement dus à des problèmes imprévisibles (sécurité, météo, panne, …).

Résultats :

Le 1er modèle souffre fortement du bruit et ne permet pas de prédire une tendance particulière. Quant au modèle 2, l'agrégation par heure du retard moyen permet de prédire tout de même une tendance globale correcte. Cependant les 2 modèles ne sont peut-être pas les plus pertinents pour ce type de prédiction car il y a une souffre d'underfitting. Des pistes d'évolutions, sont donc présentées à la fin du rapport.

MINÉ Nicolas