AIFLUENCE : PRÉDIRE L'AFFLUENCE SUR LE RÉSEAU FRANCILIEN

Julien Roussel | Rima Hajou | Hông-Lan Botterman | Adnene Khalbous

Quantmetry Part of Capgemini Invent · SNCF · école normale supérieure paris—saclay · Région îledeFrance

1 CONTEXTE

- Projet en partenariat avec la SNCF & ENS Paris Saclay suite au gain du challenge « AI for Industry » de la région Ile-de-France.
- L'objectif est de créer des modèles de prédiction de l'affluence dans les trains et dans les gares du réseau francilien à différents horizons temporels.
- L'usage de nos modèles permettrait aux exploitants du réseau de transport d'anticiper les variations importantes de l'affluence et ainsi, de garantir la fluidité du réseau et le confort des voyageurs.

2 OBJECTIFS

- **Développement** de modèles de prédiction de l'affluence performants sur l'ensemble du réseau francilien, indépendamment des sources de données disponibles.
- **Benchmark** State-Of-The-Art des méthodes de détection d'anomalie et d'imputation de données manquantes.
- **Généricité** pour rendre ces modèles complexes applicables à d'autres contextes (ex. Smart Grids, détection de panne…)

3 MÉTHODOLOGIE

CONTEXTE MÉTIER

Dans ce projet, nous avons concentrés nos travaux sur les quatre lignes Transilien du réseau de SNCF, qui représentent 150 arrêts.

Lignes	Nombre de branches	Nombre d'arrêts
J	5	50
H	4	54
L	3	36
K	1	10

Table 1 : Statistiques sur les lignes étudiées

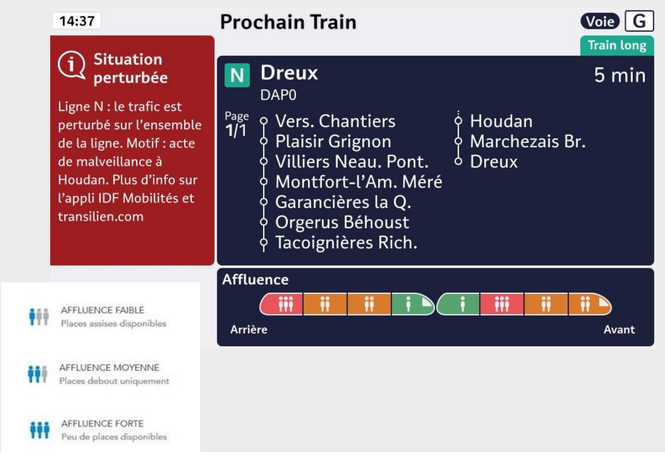
CAS D'USAGE

Trois tâches principales :
- La reconstruction de la charge à bord
- La prévision de la charge à bord par train (à 24h)
- La prévision du nombre de validation billettique à la maille 15 minutes (à 48h)

La pertinence de données exogènes comme les données calendaires, météorologiques et socio-culturelles a été explorée.
La méthodologie est basée sur le développement des modèles de Machine Learning (Light GBM, RandomForest), et une comparaison avec des modèles simples basés sur des règles métiers (baseline).

PRODUIT SOUHAITÉ

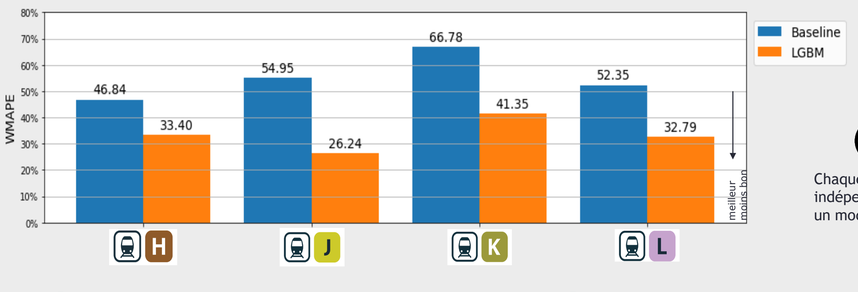
Ce projet entraînera une meilleure visibilité de l'utilisateur sur l'affluence au niveau des trains et des stations.



4 IMPLÉMENTATION ET RÉSULTATS

RECONSTRUCTION DE LA CHARGE À BORD

Pour chaque ligne, 50% des trains sont sélectionnés : toutes leurs mesures sont masqués, puis toutes les données manquantes de la ligne sont reconstruites, en utilisant deux modèles (baseline et LGBM).



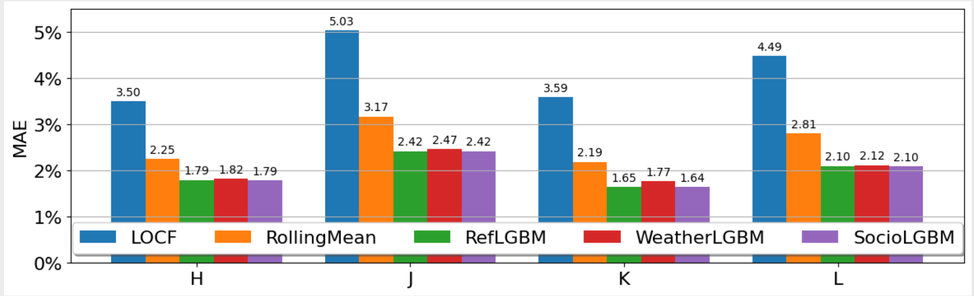
Chaque ligne est traitée indépendamment : un modèle par ligne

Modèle baseline
Modèle prédictif basé sur la moyenne des charges disponibles pour le même type de jour, station et heure.

Modèle LGBM
Modèle ensembliste (Gradient Boosting) utilisant notamment plusieurs variables exogènes (charge à bord des autres trains, destinations, mission, données calendaires ainsi que les données de validation).

PRÉVISION DE LA CHARGE À BORD À 24H

La prévision de la charge sur les différentes lignes par **LGBM** est comparée à des modèles **LOCF (**Last Observation Carried Forward) et **RollingMean** (moyenne mobile). Les LGBM ont globalement de meilleurs performances.



PRÉVISION D'AFFLUENCE DANS LES GARES À 48H:

La prévision des affluences dans les gares avec un modèle LGBM aboutit à une meilleure prévision des affluences pendant le weekend et les jours fériés.

5 CONCLUSION / PERSPECTIVES / USE CASE

CONCLUSION

- Le modèle de Gradient Boosting a montré des meilleurs performances sur les trois tâches. Le feature engineering de l'historique de l'affluence a été le contributeur majeur à ce gain de performance.
- Les features calendaires, météorologiques et spatiales pourraient s'avérer utiles pour des modèles entraînés sur un historique de données plus profond, pour mieux capter l'impact des évènements saisonnier (par ex. vagues de chaleur et des chutes de neige).

FUTURES PISTES

- Développement de modèles « Transformers » pour tester l'intérêt du deep learning pour la prévision des séries temporelles, par rapport au gold standard « modèles ensemblistes ».
- Approfondir les aspects d'IA de confiance, en particulier l'explicabilité et la mesure de l'incertitude de ces modèles.