Prédiction des validations journalières par gare chez SNCF-Transilien

Girondin Audric
Master 2 MIASHS, Université Montpellier III

Contexte et objectif

Participation à un challenge data organisé par SNCF-Transilien. `https://challengedata.ens.fr/participants/challenges/149/`

Contexte : SNCF-Transilien est l'opérateur des trains de banlieue en Île-de-France, faisant circuler plus de 6 200 trains et transportant 3,2 millions de voyageurs quotidiennement. Ces voyageurs valident leurs cartes à puce sur les portiques en moyenne 2,3 millions de fois par jour. Entre 2015 et 2019, le nombre de validations a connu une croissance annuelle de 6%. Anticiper cette évolution permettrait à SNCF-Transilien d'améliorer la performance de son exploitation et d'adapter son offre.

Objectif : Prédire à moyen-long terme le nombre de validations par jour et par gare pour anticiper les volumes de voyageurs, mieux comprendre les dynamiques d'affluence et accompagner les évolutions du réseau.

Métrique d'évaluation : L'erreur est mesurée à l'aide du *Mean Absolute Percentage Error (MAPE)* :

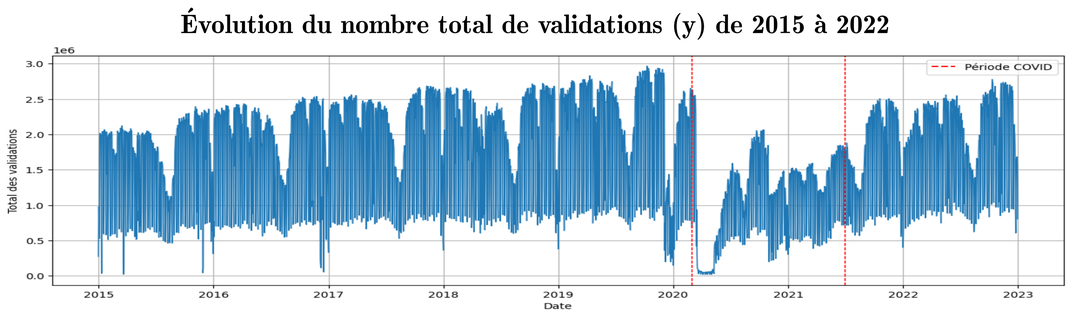
$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Description des données

Origine : Données fournies par Île-de-France Mobilités.
Train.csv : 1 237 971 lignes, 6 colonnes (2015–2022).
Test.csv : 78 652 lignes, 5 colonnes (janvier–juin 2023).
Nombre de stations : 448 gares distinctes.
Variables :
— `date` : jour de la validation
— `station` : identifiant anonymisé de la gare
— `job, ferie, vacances` : indicateurs contextuels
— `y` : nombre de validations
Features engineering :
— Variables temporelles : `weekday`, `weekofyear`, `month`, `quarter`, etc.
— Lags : `lag_1_log`, `lag_7_log`, `lag_30_log`, `lag_365_log`
— Encodage des stations
— Log-transformation de `y`

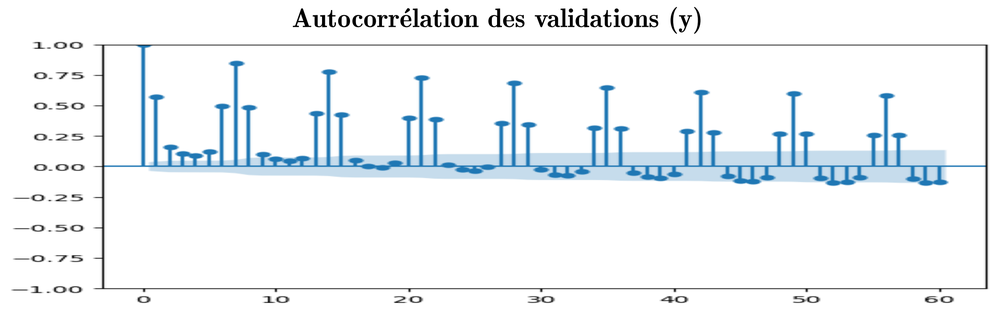
Analyse exploratoire des données

Évolution du nombre total de validations (y) de 2015 à 2022



— Tendance haussière de 2015 à 2019.
— Saisonnalité annuelle très forte.
— Chute brutale en 2020 puis reprise progressive (effet COVID-19).
— Régularité retrouvée à partir de 2022, mais un niveau encore légèrement inférieur à 2019.

Autocorrélation des validations (y)



— Saisonnalité hebdomadaire très marquée : des pics nets tous les 7 jours (trafic récurrent les mêmes jours de semaine).
— Autocorrélation qui décroît lentement indique une tendance à long terme, en plus de la saisonnalité.
— Série très autocorrélée jusqu'à 60 jours (corrélation significative).

Méthodologie

Approches classiques (sans Machine Learning) :

— **Benchmark** : recopie de 2022 sur 2023 en alignant les jours
— Projection de tendance de 2021–2022 sur 2023

Prétraitements et stratégies de modélisation :

— Trois stratégies testées pour traiter la période COVID :

- Ajout d'une variable indicatrice COVID
- Remplacement des données par lissage ou moyenne des y précédents
- Entraînement uniquement sur la période post-COVID (2021+)

— Autres approches explorées :

- SARIMA par station
- XGBoost par station

Modèles Machine Learning et Deep Learning :

— Prophet
— ARIMA
— XGBoost
— LightGBM
— Réseaux de neurones : CNN, LSTM, GRU
— Transformers pour séries temporelles multivariées

Résultats

Sélection de modèles sur jeu de validation (minimisation de la MAPE)

Stratégies	Meilleurs Modèles	MAPE
Variable COVID	LightGBM	0.77
	XGBoost	0.75
Remplacement des données	LSTM	1.52
	CNN	1.00
Apprentissage post-COVID (2021+)	LightGBM	pas assez de données pour évaluer
	XGBoost	pas assez de données pour évaluer
XGBoost par station	**XGBoost**	**0.31**

Classement des soumissions par score public (MAPE)

Rang	Méthode	Features	Score
1	**XGBoost par station post-COVID**	**Features engineering + indicateurs contextuels**	**143,64**
2	XGBoost par station	Features engineering et indicateurs contextuels	150.22
3	XGBoost par station	Indicateurs contextuels	182.66
4	XGBoost post-COVID	Features engineering et indicateurs contextuels	217.43
5	SARIMA	Features engineering et indicateurs contextuels	231.04
6	LightGBM post-COVID	Features engineering et indicateurs contextuels	252.35
7	Projection des tendances 2021-2022	Copie 2022 + Δ(2021 − 2022)	349,52

— **Score du Benchmark officiel du challenge :** 177,0825 (91^{e} place)
— **Mon classement :** 66^{e} sur 200 participants

Analyse des résidus

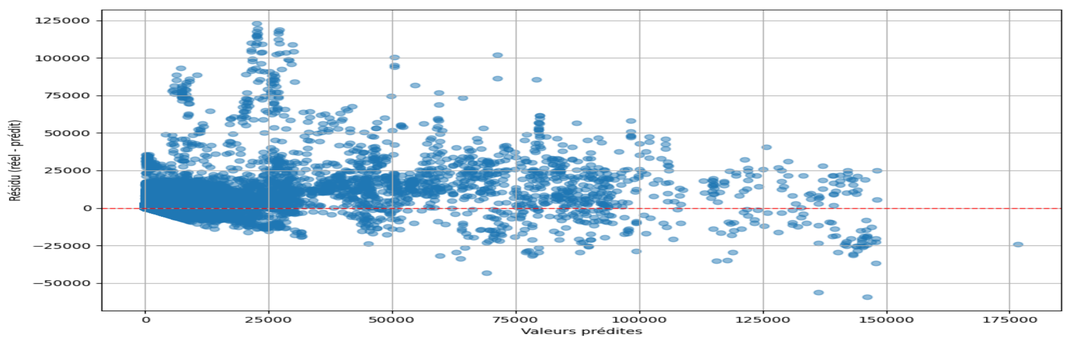
Observations clés :

— Les erreurs sont plus élevées pour les faibles valeurs prédites → instabilité à bas trafic (effet d'hétéroscédasticité).
— Pour les fortes affluences ($\geq 100\,000$), les résidus sont resserrés autour de zéro : meilleure précision.
— Les erreurs sont globalement symétriques autour de 0 mais présence d'outliers.

Perspectives :

— Clustering des gares pour créer des modèles spécialisés
— Ajouter des variables extérieures : événements, météo, etc.
— Analyser les outliers pour mieux comprendre les erreurs
— Tester des modèles hybrides (ex. LSTM + XGBoost)

Graphique des résidus sur le jeu de validation du meilleur modèle (XGBoost par station avec features)



Références

— Olah, C. (2015). *Understanding LSTM Networks*. arXiv :1406.1078v3 [cs.CL].
— Siami-Namini, S., Siami-Namin, A. (2018). *Forecasting Economic and Financial Time Series : ARIMA vs. LSTM*. Texas Tech University.
— Taylor, S.J., Letham, B. (2017). *Forecasting at Scale*. Facebook.
— Vaswani, A., et al. (2017). *Attention Is All You Need*. arXiv :1706.03762v7 [cs.CL].
— Chen, T., Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. arXiv :1603.02754v3 [cs.LG].

poster