Apprentissage et génération par échantillonnage de probabilités

Challenge Data SNCF - Prédiction du nombre de validations par gare

Matteo MARENGO, Teddy ALEXANDRE

ENS Paris Saclay - Master MVA

20 Mars 2024

Président: Hamada SALEH, Institut Louis Bachelier
Provider: Rémi COULAUD, SNCF

Table des matières

Contexte du Challenge & Données

- Grand nombre de voyageurs en IDF, croissance de 6% sur 5 ans. Il faut anticiper et quantifier cette croissance
- 3 fichiers CSV (2 train, 1 test)
- **Variables indicatives :** date, station
- **Variables "features" :** job, ferie, vacances
- **Variable cible :** $y \longrightarrow$ nombre de validations par station chaque jour
- Objectif : prédire y sur le 1er semestre de 2023

Type de problème

Prédiction sur des séries temporelles multivariées

Exploration des données exploration - 1

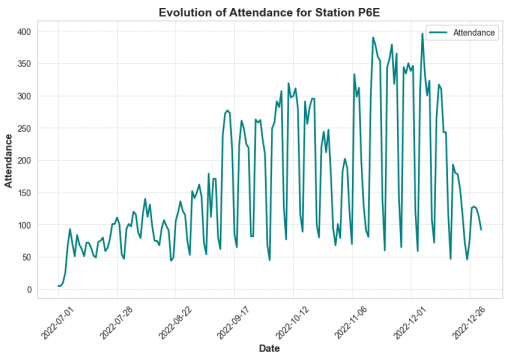

Figure 1: Evolution du nombre de validations à la station 6PE

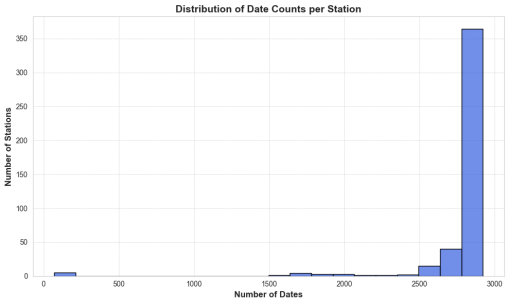

Figure 2: Distribution du nombre de dates disponibles par station

Exploration des données - 2

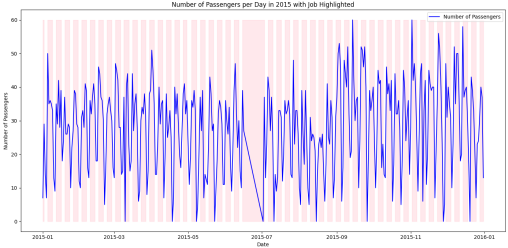

Figure 3: Evolution de l'affluence dans les trains avec les jours ouvrables mis en évidence

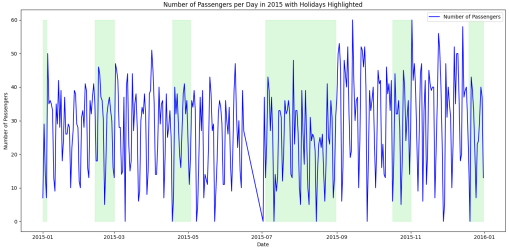

Figure 4: Evolution de l'affluence dans les trains avec les vacances mis en évidence

Techniques de l'état de l'art

Approche Machine Learning

1. Régression linéaire multivariée [Dahui et al., 2020, [3]]
2. Random Forests, Bagging
3. Gradient Boosting [Ding et al., 2016, [4]]

Approche Statistique

- Modèles de prédiction basés sur **ARIMA** (séries temporelles non stationnaires)
- Composante saisonnale + variables exogènes : **SARIMAX** [Milenkovic, 2015, [8]]

Approche Deep Learning

- Multilayer perceptron (MLP)
- Recurrent Neural Networks, LSTMs
- Modèles hybrides : CNN-LSTM [Wang et al., 2019, [13]], RF-LSTM [Toqué et al., 2017, [11]]
- Attention : Transformer-based models [Lim et al., 2020, [7]], [Zhou et al., 2021, [15]]

Nos idées

Hypothèses & Démarche Scientifique

- Traitement indépendant des stations
- Miser sur des modèles avec une certaine robustesse à l'overfitting
- Miser aussi sur des modèles séquentiels (RNNs)
- Deux principaux axes d'exploration : l'**ensemble learning** et les réseaux **LSTMs**

Random Forest and Bagging

- Ensemble learning : fusionner les prédictions de plusieurs modèles
- Bagging : entraîner les modèles sur des sous-ensembles du dataset d'entraînement
- Décorrelation des modèles $\rightarrow$ + robuste

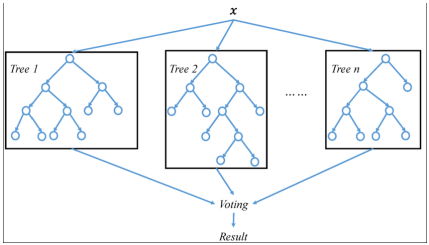

Figure 5: Architecture d'un random forest, extrait de [14]

Gradient Boosting

- Construit des arbres de manière séquentielle, chaque arbre apprenant des erreurs commises avant
- Optimise une loss en ajoutant de manière itérative des arbres à l'ensemble
- Prédiction finale : somme pondérée des prédictions de tous les arbres de l'ensemble

Les LSTMs

- Architectures puissantes pour traiter des données séquentielles.
- RNN : maintien un état interne qui leur permet de traiter des séquences d'entrées, les rendant adaptés aux tâches telles que la prédiction de séries temporelles.
- LSTMs : répond au problème du vanishing gradient, pratique pour garder les informations court et long-terme

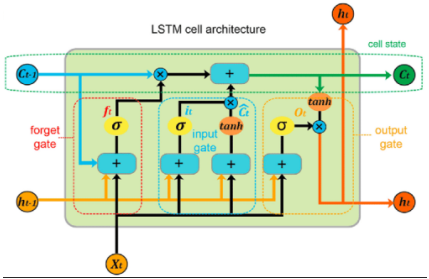

Figure 6: Cellule d'un LSTM

Random Forest Regressor

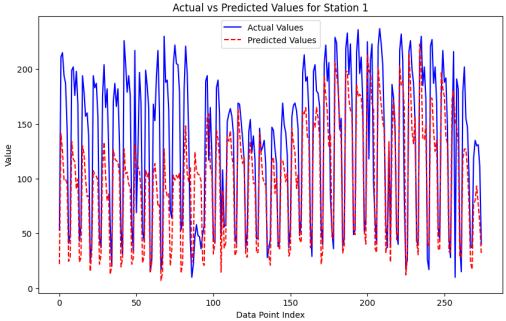

Figure 7: Valeurs actuelles et prédites sur l'ensemble de validation

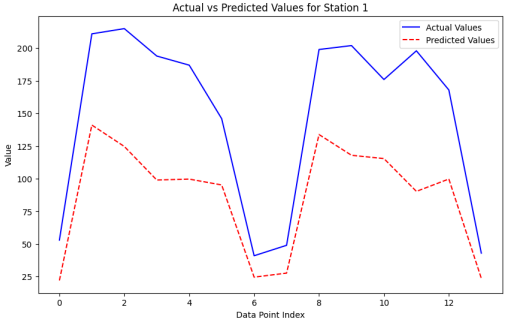

Figure 8: Valeurs actuelles et prédites sur l'ensemble de validation sur deux semaines

Random Forest Regressor

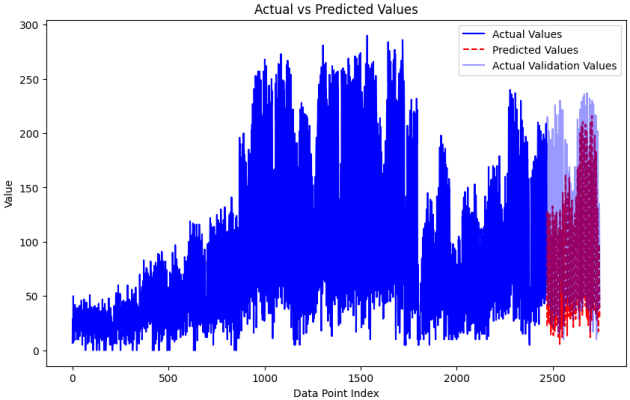

Figure 9: Valeurs actuelles et prédites sur l'ensemble d'entraînement et de validation

LSTM

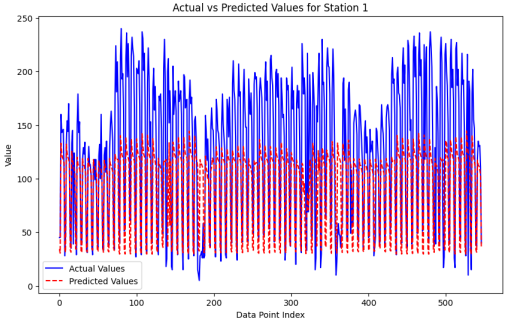

Figure 10: Valeurs actuelles et prédites sur l'ensemble de validation

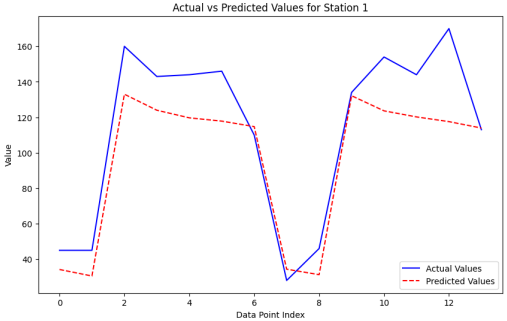

Figure 11: Valeurs actuelles et prédites sur l'ensemble de validation sur deux semaines

LSTM

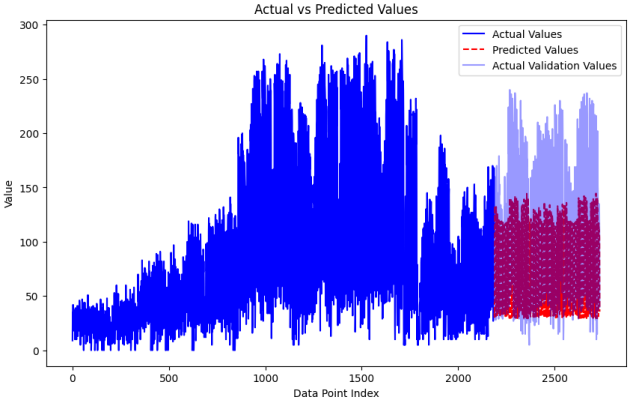

Figure 12: Valeurs actuelles et prédites sur l'ensemble d'entraînement et de validation

Prédictions de XGBOOST pour 2023

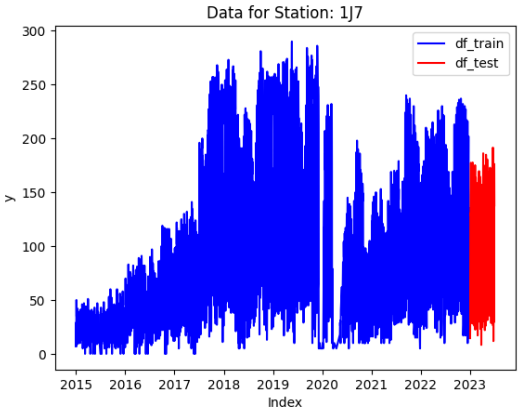

Figure 13: Prédictions de l'année 2023 avec le modèle de XGBOOST

Prédictions de SARIMAX pour 2023

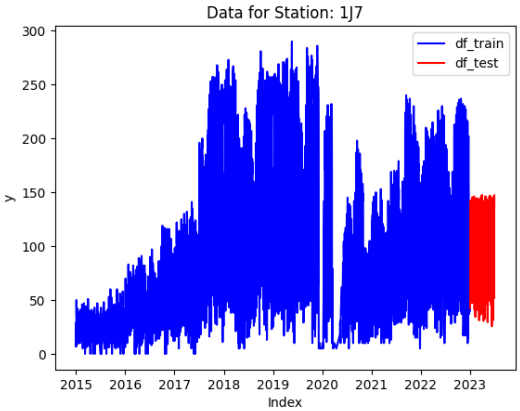

Figure 14: Prédictions de l'année 2023 avec le modèle de XGBOOST

Prédictions de FC pour 2023

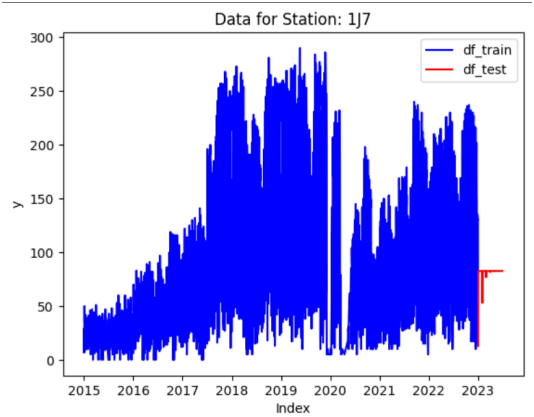

Figure 15: Prédictions de l'année 2023 avec le modèle des couches "fully-connected"

Classement final

6	24 février 2024 17:10	Talexandre150 & Matteo_Marengo	97,0915

(a) Public

4	24 février 2024 17:10	Talexandre150 & Matteo_Marengo	97,3480

(b) Privé

Figure 16: Les classements finaux

MAPE Score, autres métriques ?

Points positifs

- Simple à calculer
- Facilement interprétable

Points négatifs

- Explose aux valeurs cibles y proches de zéro
- Non symmétrique $\rightarrow$ underforecast favorisé sur l'overforecast

D'autres métriques ?

- MAE, (R)MSE
- Symmetric MAPE, MASE.

Le "trou" du COVID

- Observation : chute de la fréquentation voyageur en 2020
- Evénement difficile à anticiper !

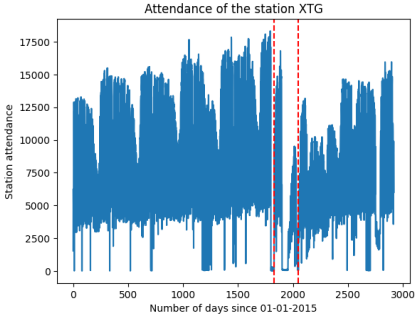

Figure 17: Attendance of the station XTG with the covid gap

Utilisation des transformeurs

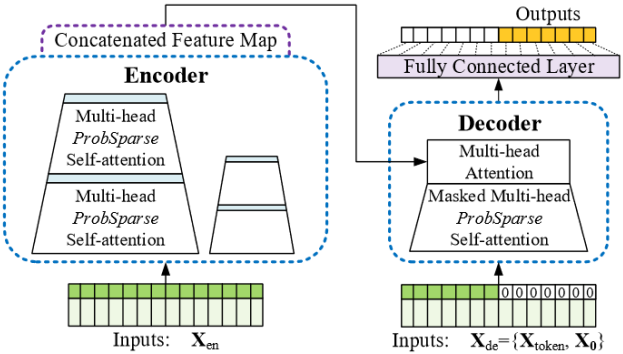

Figure 18: Architecture globale du Informer. Issue de [15].

D'autres techniques

Extraction de features supplémentaires

- Librairie TSFEL
- Lagged Features

D'autres frameworks

- Framework Prophet par Facebook
- Framework Sktime

Conclusion générale

- Challenge éprouvant pendant 2 mois mais également très instructif.
- Découverte de l'univers de la prédiction de séries temporelles.
- Tâche complexe $\rightarrow$ besoin d'aller en profondeur pour comprendre quels étaient les meilleurs algorithmes à utiliser.
- D'autres algorithmes sont à étudier, de nombreuses perspectives dans le futur pour améliorer les prédicitions.

[1] Marília Barandas et al. "TSFEL: Time Series Feature Extraction Library". In: *SoftwareX* 11 (2020), p. 100456.

[2] Maximilian Christ et al. "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)". In: *Neurocomputing* 307 (2018), pp. 72–77. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2018.03.067. URL: https://www.sciencedirect.com/science/article/pii/S0925231218304843.

[3] L.I. Dahui. "Predicting short-term traffic flow in urban based on multivariate linear regression model". In: *Journal of Intelligent & Fuzzy Systems* 39 (June 2020), pp. 1–11. DOI: 10.3233/JIFS-179916.

[4] Chuan Ding et al. "Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees". In: *Sustainability* 8 (Oct. 2016), p. 1100. DOI: 10.3390/su8111100.

[5] Rui Kang et al. "Rapid identification of foodborne bacteria with hyperspectral microscopic imaging and artificial intelligent classification algorithms". In: *Food Control* 130 (June 2021), p. 108379. DOI: 10.1016/j.foodcont.2021.108379.

[6] Bryan Lim and Stefan Zohren. "Time-series forecasting with deep learning: a survey". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194 (Feb. 2021), p. 20200209. ISSN: 1471-2962. DOI: 10.1098/rsta.2020.0209. URL: http://dx.doi.org/10.1098/rsta.2020.0209.

[7] Bryan Lim et al. *Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting*. 2020. arXiv: 1912.09363 [stat.ML].

[8] Milos Milenkovic et al. "SARIMA modelling approach for railway passenger flow forecasting". In: *Transport* 33 (Oct. 2015), pp. 1–8. DOI: 10.3846/16484142.2016.1139623.

[9] Arnaud de Myttenaere et al. "Mean Absolute Percentage Error for regression models". In: *Neurocomputing* 192 (June 2016), pp. 38–48. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2015.12.114. URL: http://dx.doi.org/10.1016/j.neucom.2015.12.114.

[10] Letham B. Taylor SJ. "Forecasting at scale". In: *PeerJ Preprints* (2017). DOI: 10.7287/peerj.preprints.3190v2.

[11] Florian Toqué et al. "Short & long term forecasting of multimodal transport passenger flows with machine learning methods". In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. 2017, pp. 560–566. DOI: 10.1109/ITSC.2017.8317939.

[12] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[13] Yu Wang et al. "Prediction of Passenger Flow Based on CNN-LSTM Hybrid Model". In: *2019 12th International Symposium on Computational Intelligence and Design (ISCID)* (2019), pp. 132–135. URL: https://api.semanticscholar.org/CorpusID:218651276.

[14] Yuanchao Wang et al. "A hybrid ensemble method for pulsar candidate classification". In: *Astrophysics and Space Science* 364 (Aug. 2019). DOI: 10.1007/s10509-019-3602-4.

[15] Haoyi Zhou et al. *Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting*. 2021. arXiv: 2012.07436 [cs.LG].

Thank you !

Random Forest - l'algorithme

- Échantillonner avec remise p éléments des données d'entraînement (X, Y) $(p < n)$: X_p et Y_p.
- Entraîner un arbre de décision ϕ_p sur X_p et $Y_p \rightarrow \hat{y}_p$.
- Répéter ce processus B fois (B est un hyperparamètre, le nombre d'arbres).
- Ensuite, on renvoie la prédiction moyenne de l'ensemble de test $X' = (x'_1, ..., x'_n)$ en calculant leur prédiction moyenne :

$$\hat{y} = \frac{1}{B} \sum_{i=1}^{B} \phi_p(x'_i)$$

Algorithme du Gradient Boosting

- Initialisation: $\hat{f}_0(x) = \text{argmin}_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$
  où L est la fonction de perte choisie, y_i sont les vraies valeurs cibles, et γ est la valeur constante.
- Pour $m = 1$ à M :
  - Calcul des pseudo-résidus : $r_{im} = -\left[\frac{\partial L(y_i, \hat{f}_{m-1}(x_i))}{\partial \hat{f}_{m-1}(x_i)}\right]$
  - Fitter un arbre de décision, aux pseudo-résidus :
    $\phi_m(x) = \text{argmin}_\phi \sum_{i=1}^{n} \Big(L(y_i, \hat{f}_{m-1}(x_i) + \phi(x_i))\Big)$
  - Update le modèle en ajoutant la contribution du nouvel arbre avec un learning rate η : $\hat{f}_m(x) = \hat{f}_{m-1}(x) + \eta\phi_m(x)$
- Prédiction finale : somme pondérée des prédictions de tous les arbres : $\hat{Y} = \sum_{m=1}^{M} \eta\phi_m(x)$

Equations d'une cellule LSTM

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\
C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t \odot \tanh(C_t)
\end{aligned}
$$