Project Architecture - *15/01/2026*

Exploration Notebooks

Exploration.ipynb

This notebook is used for data exploration and provides a detailed implementation of the required transformations and preprocessing steps. This includes removing the COVID period as well as identifying stations with the fewest data points.

OptimizeParameters3.ipynb

This notebook is used for hyperparameter optimization for the model implemented in `LSTMn°3.ipynb`

OptimizeParameters4.ipynb

This notebook is used for hyperparameter optimization for the model implemented in `LSTMn°4.ipynb`

SeePredictionResults.ipynb

This notebook is used to load previously generated predictions in order to visualize and compare them using different types of plots.

PredictionQD6.ipynb

This notebook is used to run the entire pipeline in order to compute prediction for the station QD6.

Modules

utils.py

This module provides essential functions for the end-to-end data pipeline, from raw data cleaning to final submission formatting. It also includes various functions for visualizing predictions and evaluating their results.

modelsV1.py

This module contains the deep learning architectures used for predicting passenger affluence in the Transilien train network.
The models are configured to perform predictions exclusively based on exogenous variables (job, ferie, vacances), meaning that each prediction is independent of previously predicted target values.
The validation data is taken from the last 20% of the period, i.e., the end of 2022.

modelsV2.py

This module is the same as `modelsV1.py`.
In this second version, the validation data is randomly sampled across the entire training period.

modelsQD6.py

This module contains the deep learning architectures used for predicting passenger affluence for station QD6.

AR_models.py

This module contains the 1st autoregressive deep learning architecture used for predicting passenger affluence in the Transilien train network.

Prediction Attempts

LSTMn°3.ipynb *(x2 attempts)*

This notebook was used to produce the 1st and the 2nd valid sets of predictions for 2023:

Non Autoregressive model,
Based on features = ['job', 'ferie', 'vacances'],
After hyperparameters optimization with Optuna,
Random sequences, all possible sequences for train,
Without Validation Data for training.

LSTMn°4.ipynb *(x1 attempt)*

This notebook was used to produce the 3rd valid set of predictions for 2023:

Non Autoregressive model,
Based on features = ['job', 'ferie', 'vacances'],
After hyperparameters optimization with Optuna,
Random sequences, only 25% of all possible sequences for train and validation.
20% Validation Data for training.

LSTMn°5.ipynb *(x0 attempt)*

This notebook will be used to produce the 4th valid set of predictions for 2023:

Autoregressive model,
Based on features = ['y', 'job', 'ferie', 'vacances'],
Random sequences, only 25% of all possible sequences for train and validation.
20% Validation Data for training.

Attempts CSV Files

QD6predictions

This file contains the frozen predictions for station QD6 from `PredictionQD6.ipynb`.

y_test_LSTM_v3.1

2023 predictions generated by the model implemented in `LSTMn°3.ipynb` with randomly chosen hyperparameters.

y_test_LSTM_v3.1_sorted

`y_test_LSTM_v3.1` but correctly sorted and ready for submission.

> The score for this submission is 219.14

y_test_LSTM_v3.2

2023 predictions generated by the model implemented in `LSTMn°3.ipynb` after hyperparameter optimization.

y_test_LSTM_v3.2_sorted

`y_test_LSTM_v3.2` but correctly sorted and ready for submission.

> The score for this submission is 209.61

y_test_LSTM_v4

2023 predictions generated by the model implemented in `LSTMn°4.ipynb` after hyperparameter optimization.

y_test_LSTM_v4

`y_test_LSTM_v4` but correctly sorted and ready for submission.

> The score for this submission is 210.119

Data CSV Files

train_f_x.csv	-	*y_train_sncf.csv*	-	*x_test.csv*
x_train		y_train		x_test