BECOME A KAGGLE MASTER - HW2

Brice Convers
Paul Malet

Sommaire

1. Exploratory data analysis

2. Feature augmentation and selection

3. Models
   a. Transformers?
   b. Base models
   c. Ridge model

4. Hyperparameter tuning

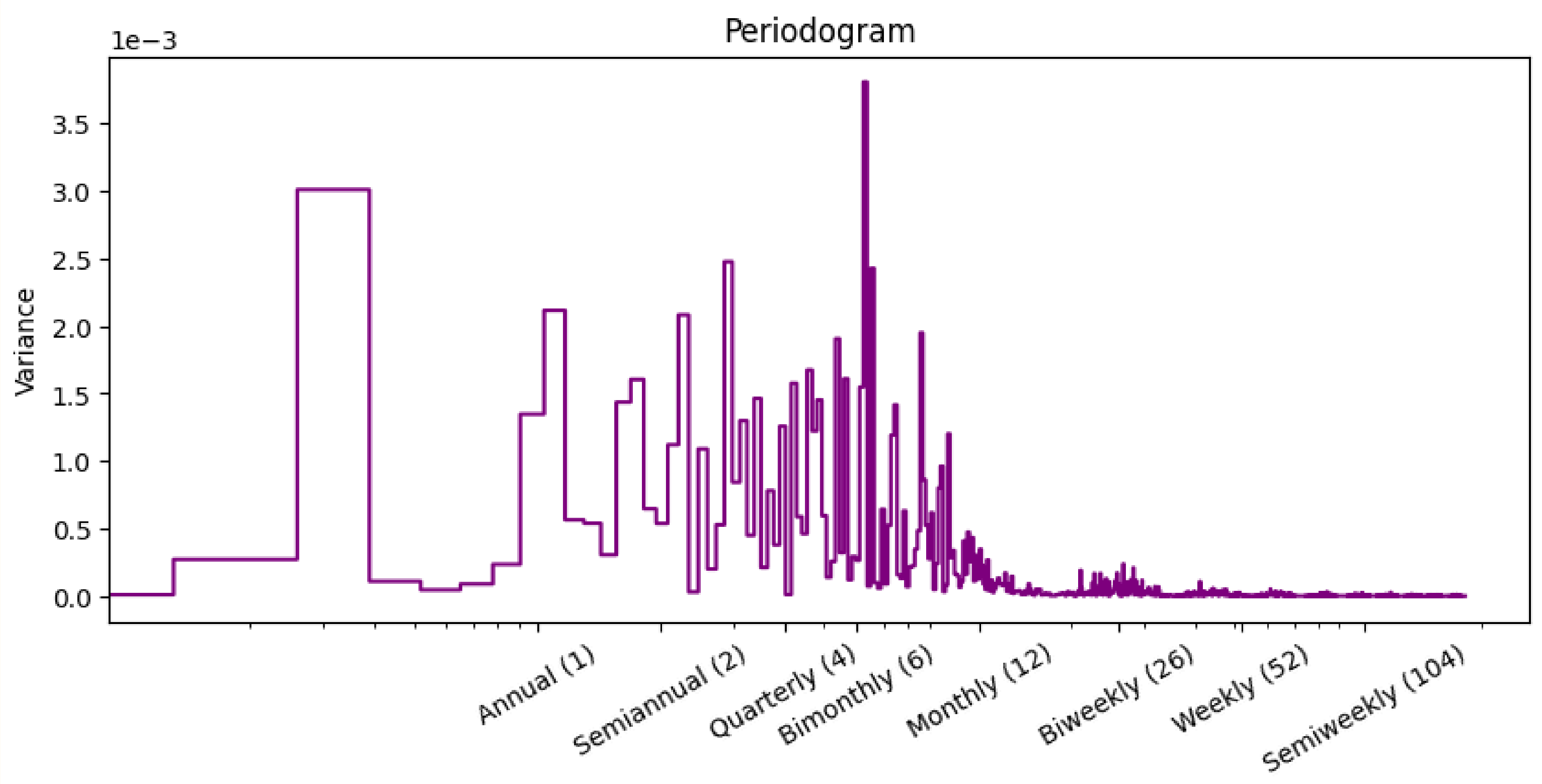
5. Results

1. Exploratory data analysis - 1

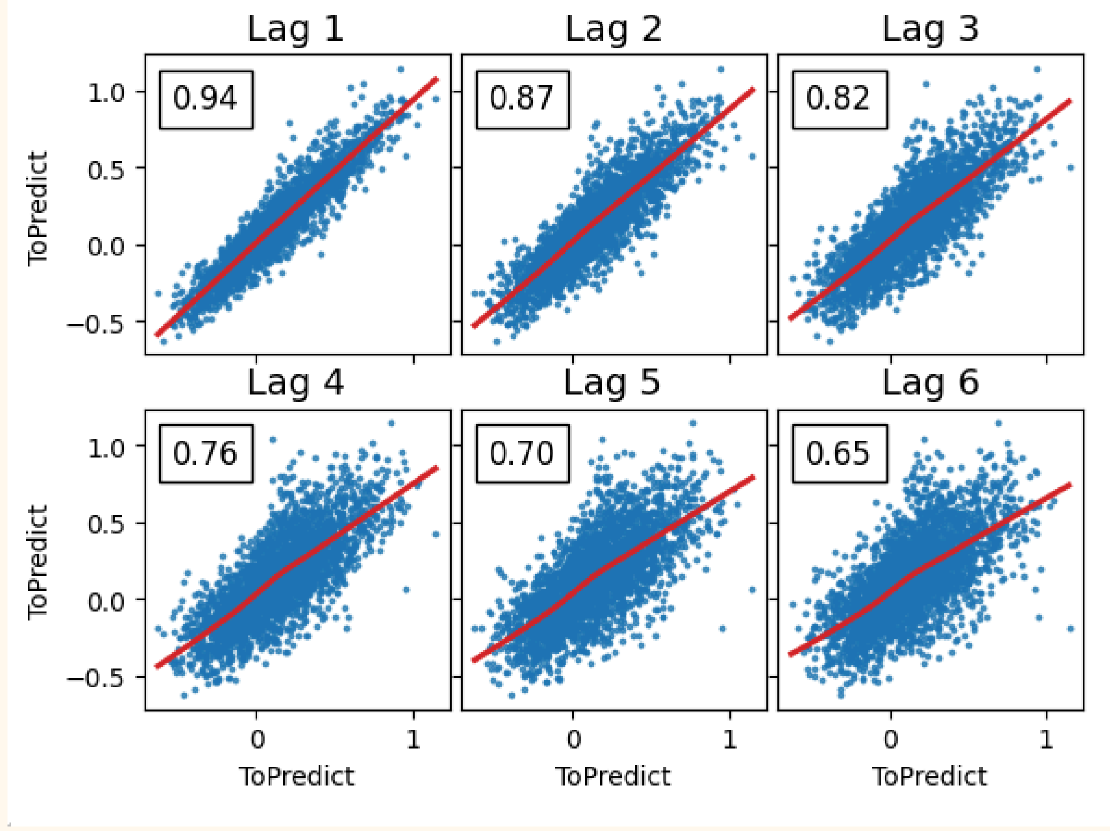
Dataset:

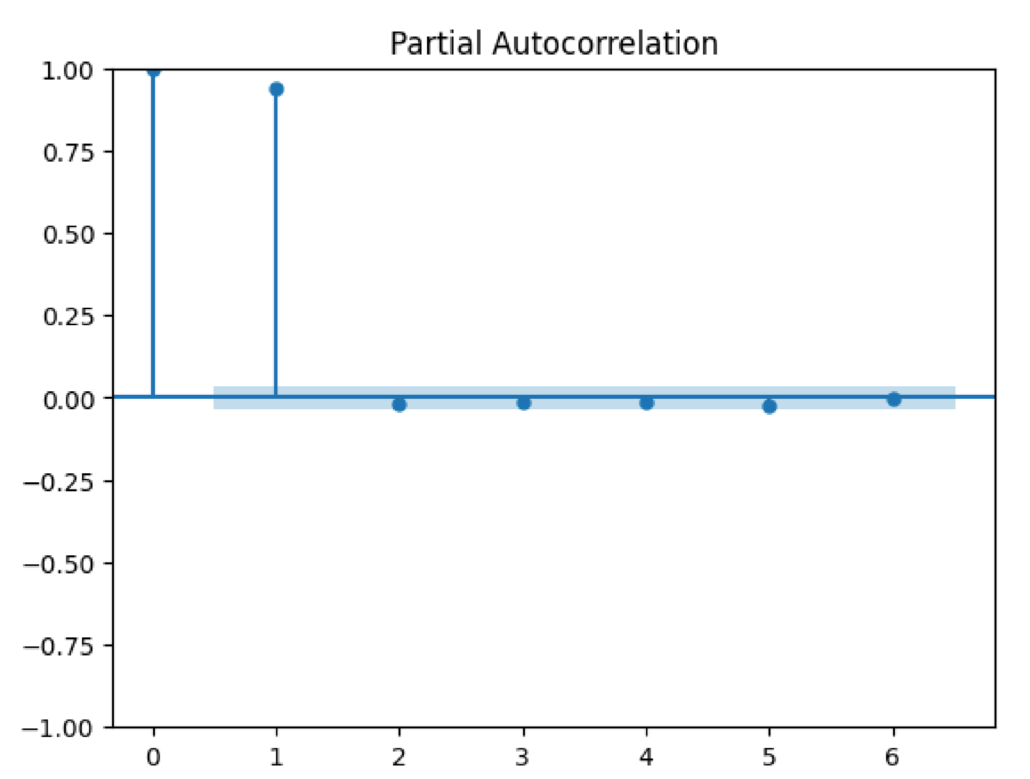
- 133 features
- ToPredict is a float
- Train: 2811 lines
- Test: 1206 lines
- 5 days per week - 260 days per year

1. Exploratory data analysis - 2



Periodogram: strong bimonthly seasonality

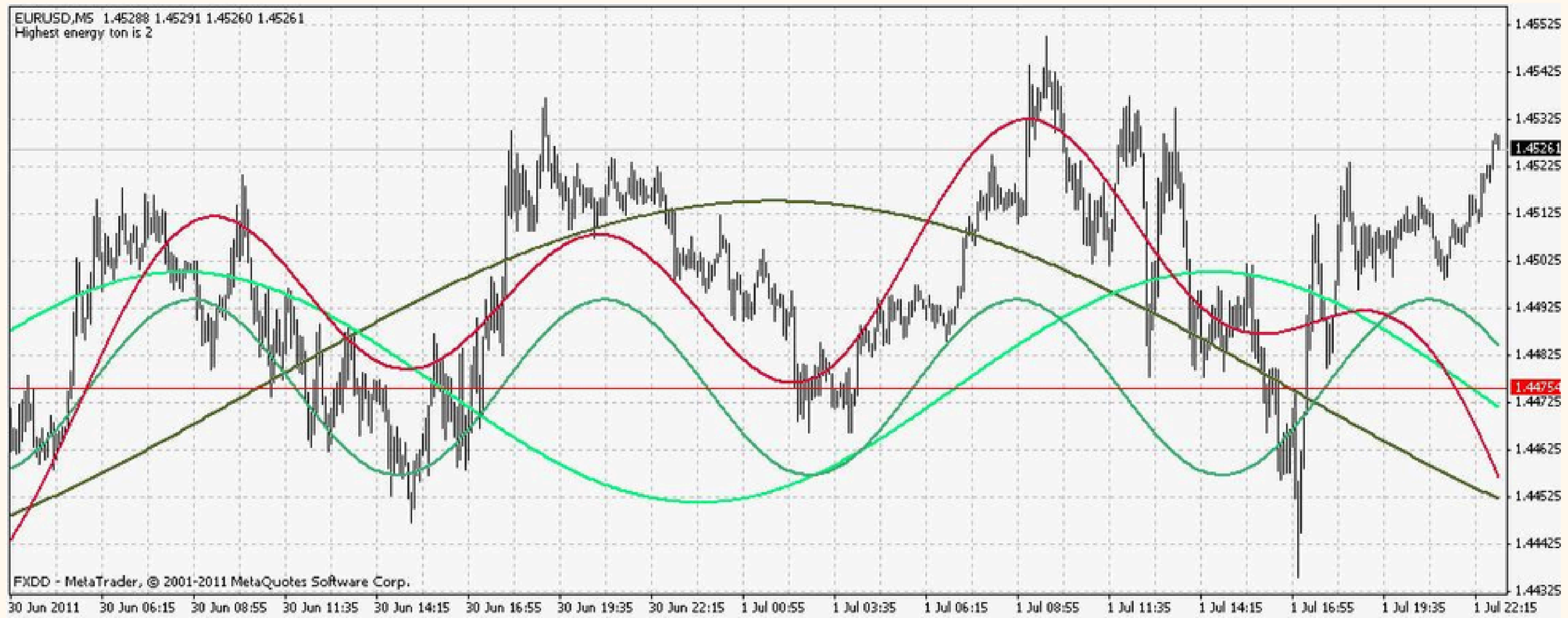


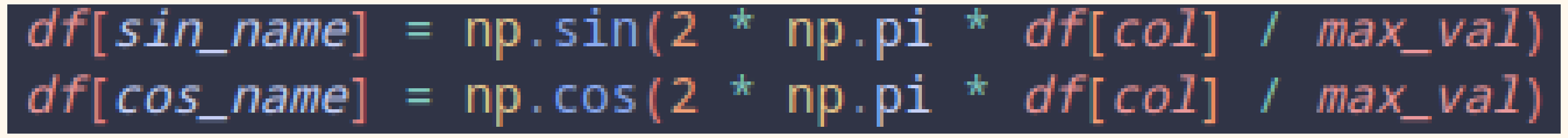


Lag 1 with strong autocorrelation

2. Feature augmentation and selection

- **sinusoidal encoding**
  - weekly: 5 days
  - monthly : 21 days
  - quarterly: 65 days
  - annual: 260 days
  - week of year: 52 weeks
  - month of year: 12 weeks
- **Feature transformations**
  - Log
  - Lag (shift: 1, 3, 7)
  - rolling mean/std (shift: 3, 7)





```
df[sin_name] = np.sin(2 * np.pi * df[col] / max_val)
df[cos_name] = np.cos(2 * np.pi * df[col] / max_val)
```

2. Feature augmentation and selection

Temporal Categorical Encoding

- Day of week effects
  - Monday/ Friday
- Start and End of the month/ year/ Quarter
- Month:
  - January/ December

Technical indicators

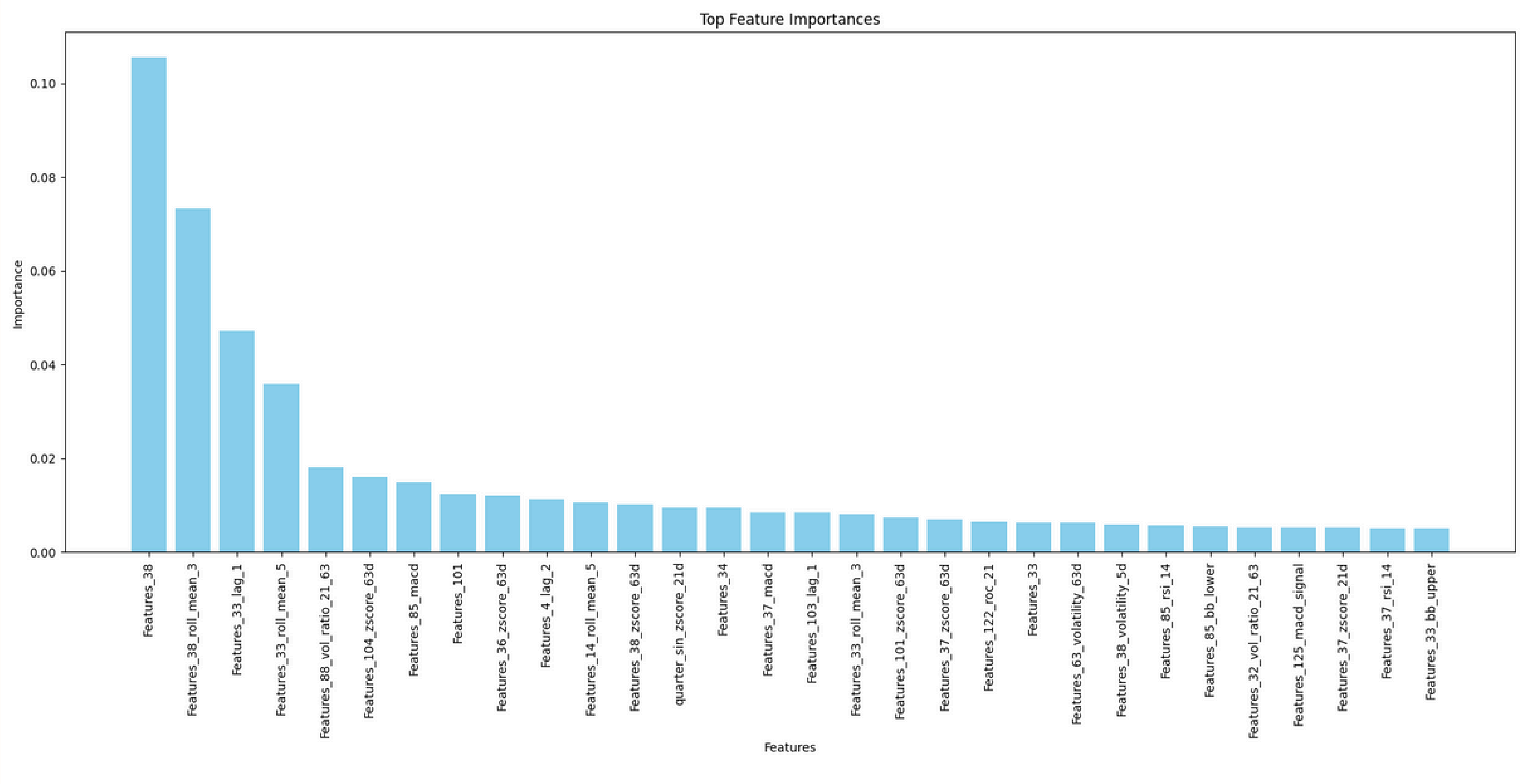
- RSI (Relative Strength Index)
- MACD (Moving Average Convergence Divergence)
- Bollinger Band
- ROC (Rate of change)
- Volatility
- Mean reversion and Momentum indicators

$$RS = \frac{\text{moyenne des gains sur N jours}}{\text{moyenne des pertes sur N jours}}$$

$$\mathrm{ROC}_n = \left(\frac{P_{\text{aujourd'hui}} - P_{t-n}}{P_{t-n}} \right) \times 100$$

2. Feature augmentation and selection

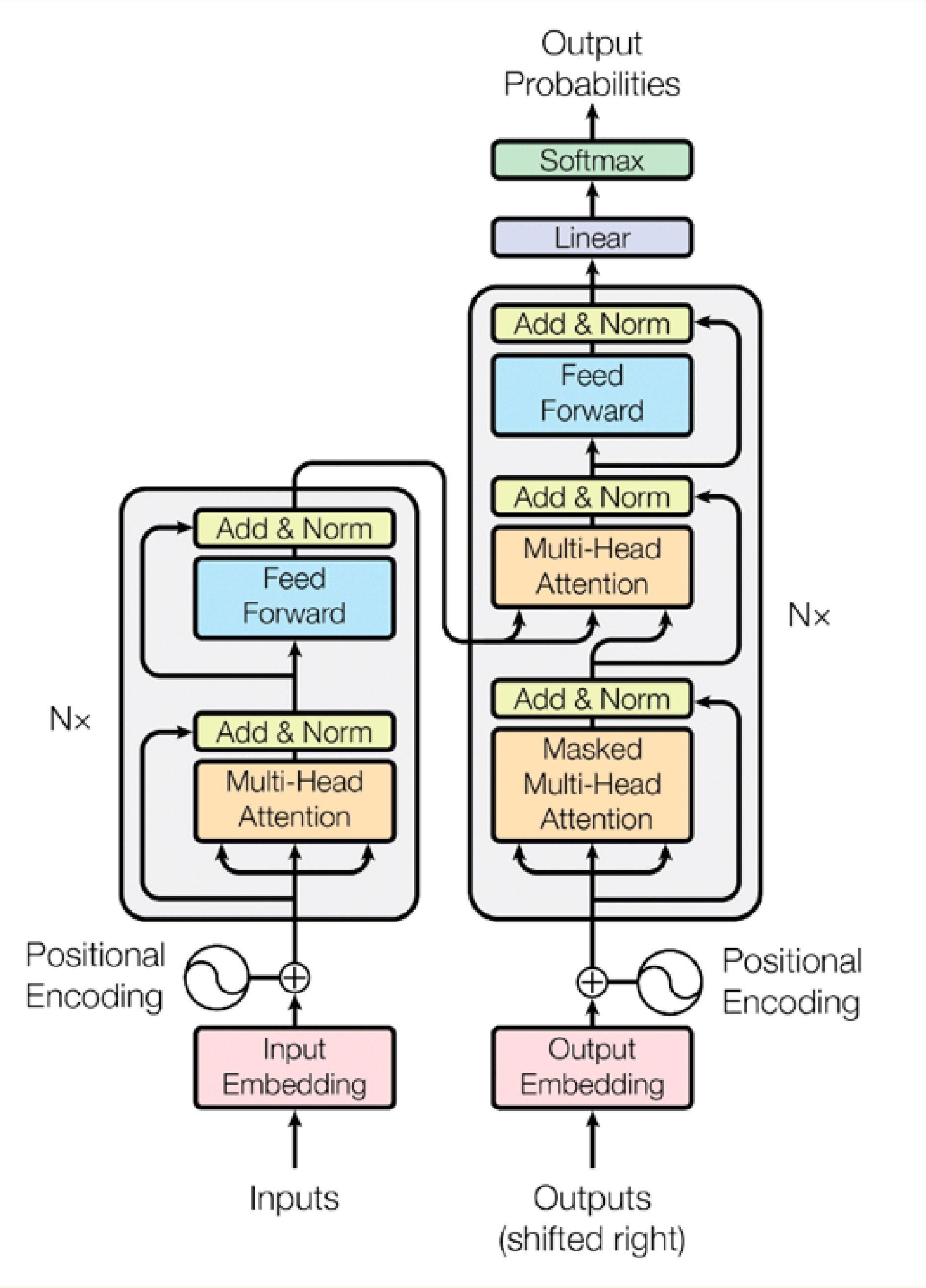
- Feature selection
  - **XGBRegressor** with threshold
  - 4698 columns → 166 columns (Threshold at 0.001)



3. Models

a. Transformers?

- **Differential Former**
  - ○ Noise Cancellation
- Insufficient Data



3. Models
b. Base models

1) CatBoost Regressor:

- Less overfitting
- Performs well with limited training samples

2) XGBoost Regressor:

- Handles high-dimensional data efficiently
- Captures complex non-linear relationships

3) RandomForest Regressor:

- Diversity through ensemble of uncorrelated trees
- Robust to outliers

Objective:

- Balanced Complexity
- Less prone to overfitting
- Diversity

3. Models
c. Ridge model

1) **Regularization**: prevents overfitting when combining the base models

2) **Simplicity**: linear meta-models fast to train

3) **Stability**: less sensitive to correlation between base model predictions

→ **Make the most of our different models**

- Training with **5-fold validation** with **TimeSeriesSplit**
- **One model** at the time, then Ridge model

4. Hyperparameter tuning

Use of **Optuna** library

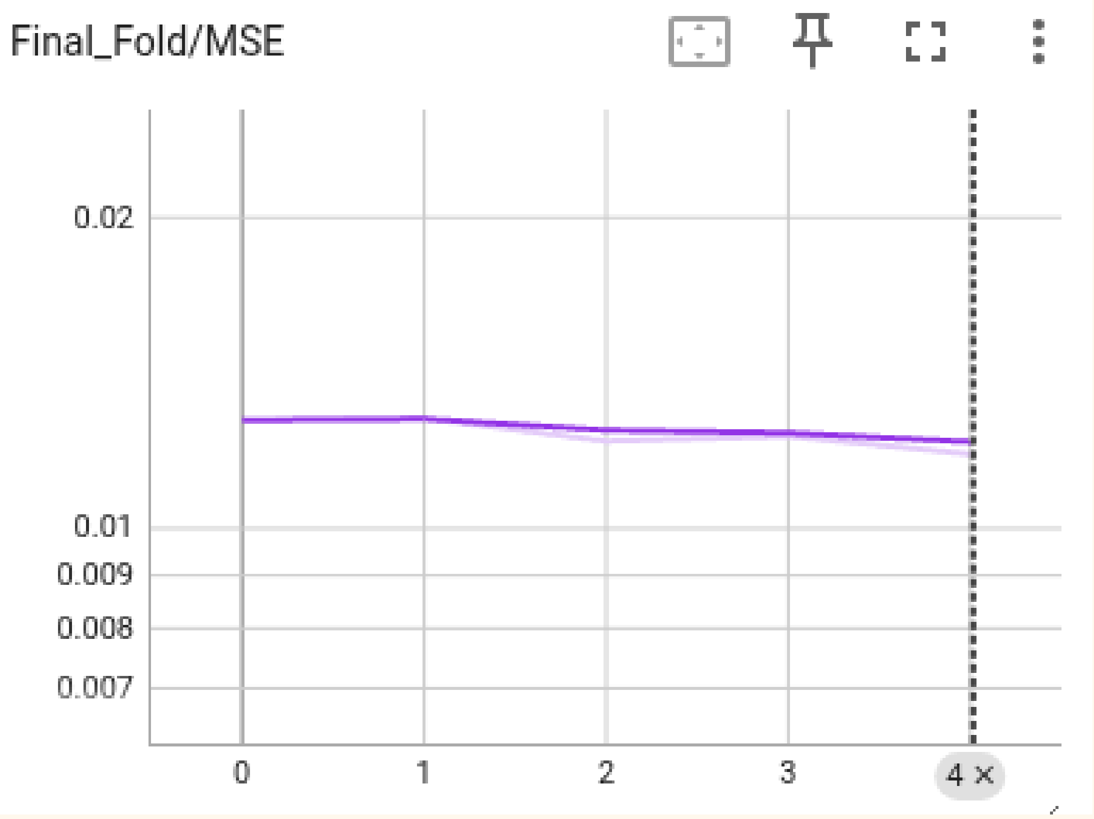
Multi-model approach: separately optimized XGBoost, CatBoost, and RandomForest

→ Optimize Ridge model on the trained models

Optimization method: TimeSeriesSplit (5 folds), 50 jobs, monitored with TensorBoard

Results: combined approach reduced prediction error and maintained computational efficiency

5. Results



- **Stable MSE on train**: from 0.01273 to 0.01213

- MSE on **public** leaderboard: 0.01114
- MSE on **private** leaderboard: 0.01104