Factory Electric Consumption Prediction

A Regression-Based Forecasting Approach

Irene Burri

April 15, 2025

Project Objective

- Develop a regression model to predict future electric consumption in a factory.
- Optimize energy usage, reduce costs, and support sustainability.
- Evaluation metric: **Root Mean Squared Error (RMSE)**.

Motivation

- Rising energy costs and sustainability goals.
- Accurate forecasts enable better planning and optimization.
- Opportunity to identify seasonal trends and usage anomalies.

Dataset Overview

- **Training data:** 13,872 records
- **Test data:** 2,160 records
- **Target: Electric_Consumption**
- Provided by Kaggle: `https://www.kaggle.com/competitions/prediction-of-factory-electric-consumption/`

Tools and Libraries

- Python, Pandas, NumPy
- Scikit-learn, XGBoost
- Matplotlib, Seaborn
- GridSearchCV for hyperparameter tuning

Data Processing Steps

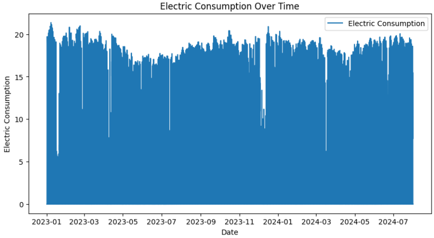
- Date Formatting and Feature Extraction of time-based features (Hour, Day, Month, DayOfWeek)
- Outlier Detection and Clipping
- Handling Missing Values
- Encoding Categorical Features (Cyclical Encoding)

	Factor_A	Factor_B	Factor_C	Factor_D	Factor_E	Factor_F	Hour_sin	Hour_cos	DayOfWeek_sin	DayOfWeek_cos	Month_sin	Month_cos	Day_sin	Day_cos
9205	3.358914	36.536203	12.820157	145.295411	0.0	0.000000e+00	-0.258819	-0.965926	-0.433884	-0.900969	5.000000e-01	0.866025	-0.651372	-0.758758
6779	2.642509	34.672805	6.255283	251.108846	0.0	2.090000e-43	0.258819	-0.965926	0.781831	0.623490	-8.660254e-01	0.500000	0.897805	-0.440394
4148	1.239516	14.581460	41.370703	129.642411	0.0	1.550142e+00	-0.866025	0.500000	0.433884	-0.900969	1.224647e-16	-1.000000	-0.968077	-0.250653
4436	0.724160	7.163648	36.193505	233.031004	0.0	1.812564e+00	-0.866025	0.500000	0.781831	0.623490	-5.000000e-01	-0.866025	0.724793	0.688967
13219	1.676083	11.344591	39.577381	136.578953	0.0	1.976420e+00	-0.965926	0.258819	0.433884	-0.900969	-5.000000e-01	-0.866025	0.724793	0.688967

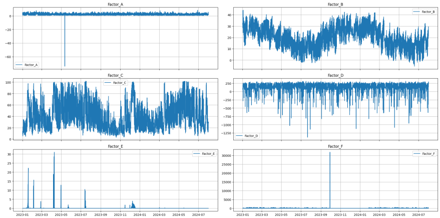
Figure: X_train head after Data preprocessing

Exploratory Data Analysis

- Boxplots and Distribution plots used for outlier analysis
- Correlation analysis to identify key features
- Visualizations of electric consumption trends



(a) Electric_Consumption over Time



(b) Factors trend over Time

Models Used

- Linear Regression
- Polynomial Regression (with Hyperparameters Tuning)
- Random Forest Regressor (with Hyperparameters Tuning)
- XGBoost Regressor (with Hyperparameters Tuning)

Model Evaluation Criteria

- Cross-Validation RMSE
- Validation RMSE
- Hyperparameter tuning using GridSearchCV
- Feature Importance Analysis
- Post-processing to avoid negative predictions

Model: Linear Regression

- Simple baseline model
- Fast to train but limited in performance
- **Validation RMSE:** 4.51

```
############      Training Linear Regression      ############

Linear Regression Cross-validated RMSE: 3.48 ± 0.20
Linear Regression RMSE: 3.192344893933311
```

Figure: Training Linear Regressor

Model: Polynomial Regression

- Captures non-linear relationships
- Tuned degree and bias
  - degree = 3,
  - include_bias = false
- **Validation RMSE:** 2.40

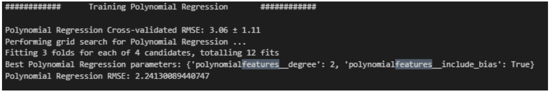
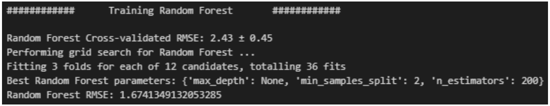


Figure: Training Polynomial Regressor

Model: Random Forest Regressor

- Ensemble of decision trees
- Tuned hyperparameters via GridSearchCV:
  - max_depth = None
  - min_sample = 2
  - n_estimators = 200
- **Validation RMSE:** 1.63



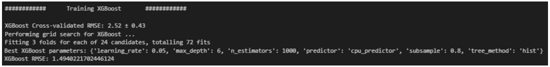
```
############      Training Random Forest       ############

Random Forest Cross-validated RMSE: 2.43 ± 0.45
Performing grid search for Random Forest ...
Fitting 3 folds for each of 12 candidates, totalling 36 fits
Best Random Forest parameters: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
Random Forest RMSE: 1.6741349132053285
```

Figure: Training Random Forest Regressor

Model: XGBoost Regressor

- Gradient Boosting-based model
- Best hyperparameters:
  - learning_rate = 0.01, max_depth = 8
  - n_estimators = 1000, subsample = 0.8
- **Validation RMSE: 1.52** (Best)


```
############      Training XGBoost       ############

XGBoost Cross-validated RMSE: 2.52 ± 0.43
Performing grid search for XGBoost ...
Fitting 3 folds for each of 24 candidates, totalling 72 fits
Best XGBoost parameters: {'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 1000, 'predictor': 'cpu_predictor', 'subsample': 0.8, 'tree_method': 'hist'}
XGBoost RMSE: 1.4940221702446124
```

Figure: Training XGBoost Regressor

Post-processing

- Ensured no negative consumption values.
- Applied post-processing to clip predictions to zero minimum.

Performance Comparison

Model	CV RMSE $\pm$ std	RMSE	Best Params
Linear Reg.	3.48 ± 0.20	3.19	Default
Poly. Reg.	3.06 ± 1.11	2.24	deg=2, include_bias=True
Random Forest	2.43 ± 0.45	1.67	est=200, depth=None, min_sample_split=2
XGBoost	2.52 ± 0.43	**1.49**	lr=0.05, depth=6, est=1000, subsample=0.8

Model Comparison

- XGBoost achieved the best performance but took more time.
- Random Forest also performed well with low variance.
- Polynomial Regression improved over linear baseline.

Feature Importance

- XGBoost and Random Forest revealed most influential features.
- Attempts to remove less important features degraded model performance.
- All available engineered features retained for final model.

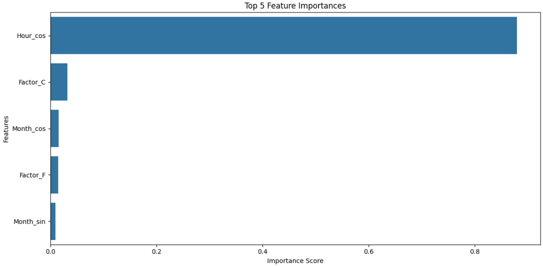


Figure: Top 5 Feature Importances

Final Model

- **Best model: XGBoost Regressor**
- Final RMSE on validation set: **1.52**
- Cleaned and well-engineered features contribute to high performance.
- Further improvements may be possible with deep learning.

Conclusion

- Successfully predicted factory electricity usage using regression models.
- XGBoost yielded best performance (RMSE $= 1.52$).
- Insights can support sustainable and cost-effective energy management.

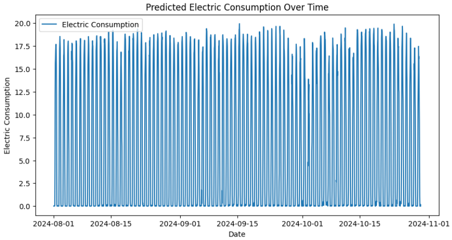


Figure: Predicted Electric Consumption over Time

References

- Kaggle Competition: https://www.kaggle.com/competitions/prediction-of-factory-electric-consumption/
- Scikit-learn, XGBoost Documentation
- GitHub project repository: https://github.com/ireneburri/Burri-PredictionOfFactoryElectricConsumption.git