Ensemble Learning - QRT Challenge

Presented by:
Ayush Tankha
Jatin Singh
Hugo Thevenet
Duoer Gu

WORKFLOW DISTRIBUTION

Hugo	Ayush	Jatin	Duoer
Initial Analysis	**Data Preprocessing + Feature Engineering**	**Modelling + Hypertuning**	**Presentation**
• Conceptualized inital analysis for scope of project. • Performed benchmarking modelling using LR, RF and Grad Boosting.	• Analyzed feature importance and mutual information between features. • Created new features to improve baseline.	• Implemented machine learning models like RF and XGBoost while hypertuning them. • Used ensemble methods like stacking on hypertuned models to improve score.	• Developed the inital and final presentation.

Add Company Name

Challenges and Objectives

Challenges

- **Develop a model** to estimate daily electricity futures price variation in France and Germany.
- Utilize simultaneous weather, energy, and trade data for explanatory variables.
- Aim for a **high Spearman's correlation score** between model predictions and actual price variations.

Objectives

- **Training and testing datasets** with weather, commodity prices, and electricity usage variables.
- Data **includes daily metrics** for two European countries across multiple energy-related dimensions.
- Output model should **predict daily futures price** variation, matched by ID to test dataset.

LETS BEGIN

Baseline Benchmarking

Baseline Models

- The initial baseline model employed was a linear regression, which provided a straightforward, easy-to-interpret model to start the analytical process.
- The linear regression model's performance was evaluated using the Spearman correlation coefficient, which is a non-parametric measure that can capture any monotonic relationship between the features and the target variable.

Objectives

- **Training and testing datasets** with weather, commodity prices, and electricity usage variables.
- Data **includes daily metrics** for two European countries across multiple energy-related dimensions.
- Output model should **predict daily futures price** variation, matched by ID to test dataset.

LETS BEGIN

Simple Model Benchmarking

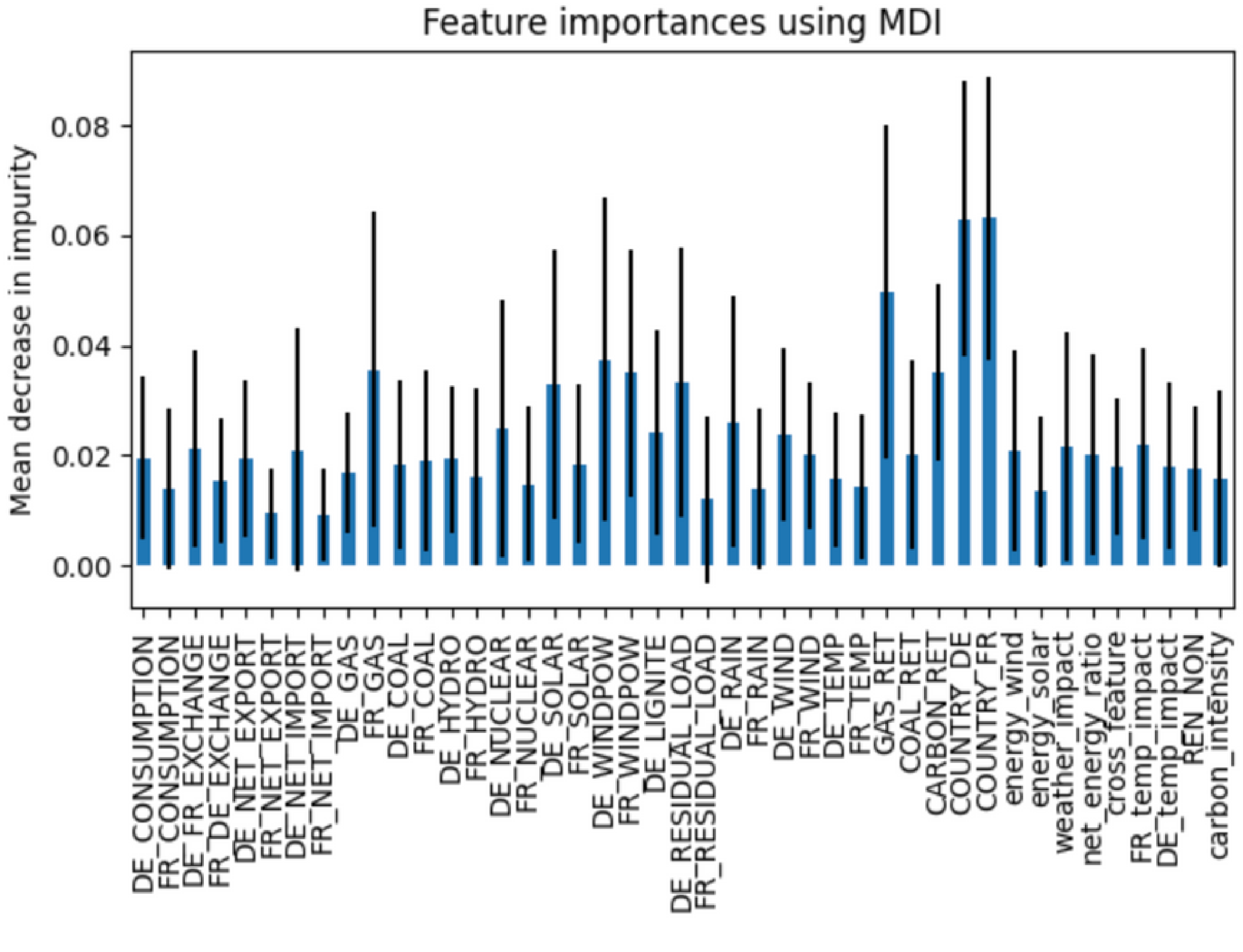
Baseline Models

- The initial baseline model employed was a linear regression, which provided a straightforward, easy-to-interpret model to start the analytical process.
- The linear regression model's performance was evaluated using the Spearman correlation coefficient, which is a non-parametric measure that can capture any monotonic relationship between the features and the target variable.

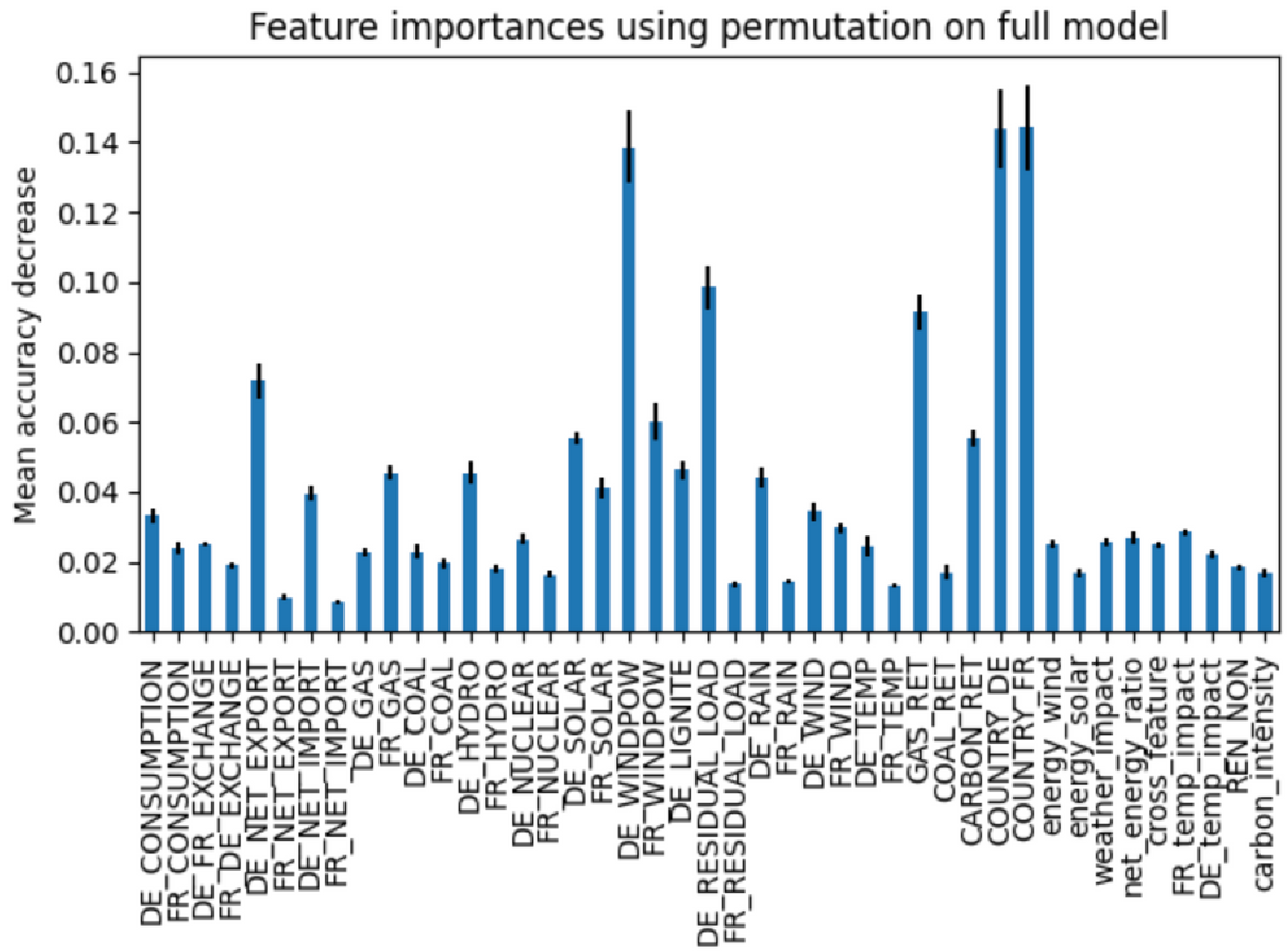
	Model	MSE	MAE	R-squared	Training Time (s)
0	linear_regression	1.005881	0.571140	0.059612	0.010603
1	random_forest	0.172495	0.244220	0.838737	6.371138
2	gradient_boosting	0.563034	0.458647	0.473624	1.277884

LETS BEGIN

Feature Importance



Feature importance based on feature permutation

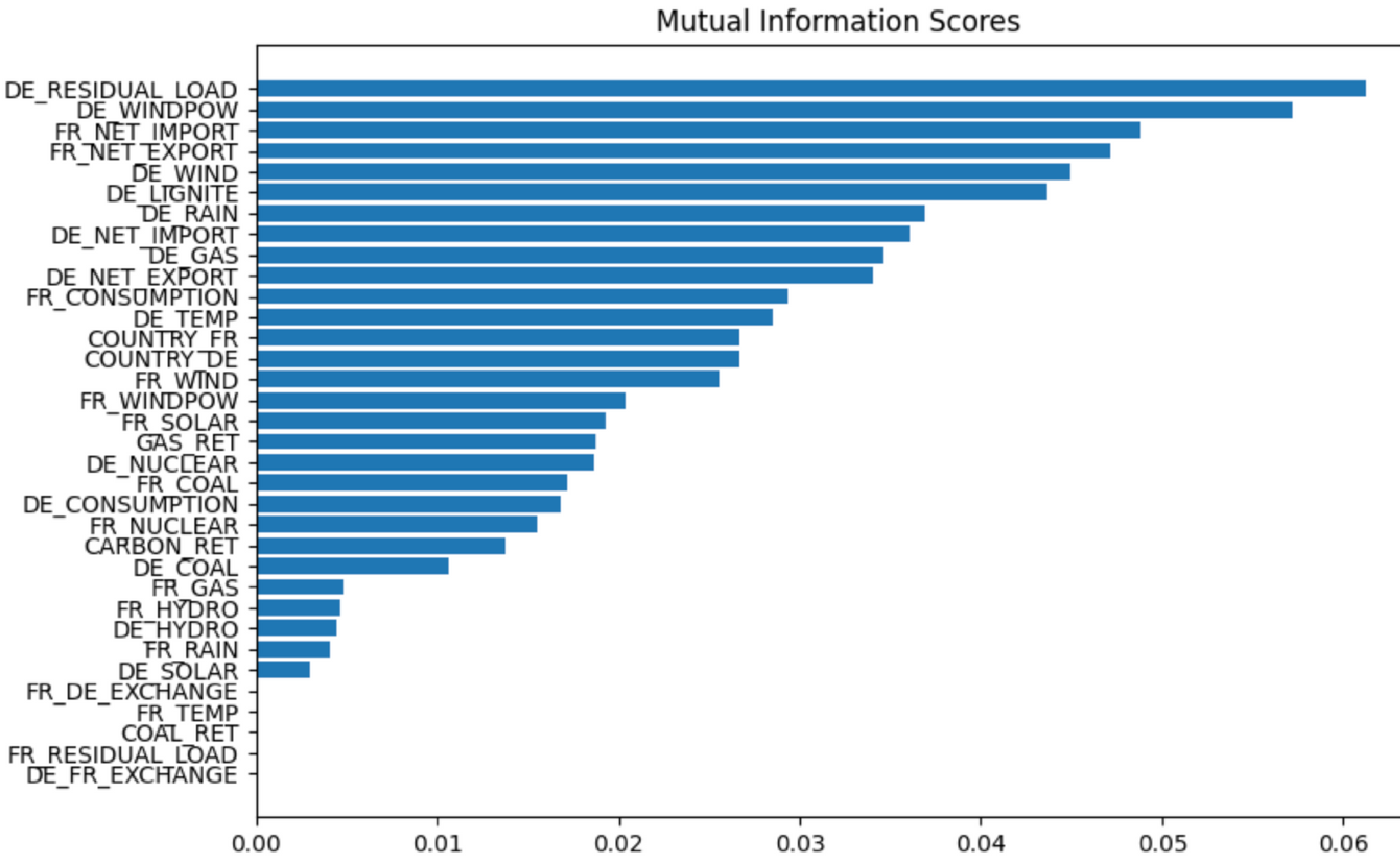


LETS LOOK AT FEATURE IMPORTANCE USING MDI AND PERMUTATION ON FULL MODEL

Data Preprocessing

Categorical Variables

- Since we have a small amount of data we try to convert categorical variables into numerical data for modelling.
- We used **One Hot Encoding** for encoding the countries France and Germany.



Numerical Variables

- We observe for feature importance of variables by finding out mutual information shared between feature variables and target variable

WE CAN SEE THAT MUTUAL INFORMATION SCORE IS VERY LOW FOR ALL FEATURES

SINCE NO FEATURES CAPTURES A MAJORITY OF INFORMATION OF THE PREDICTOR VARIBALE WE ARE FREE TO CHOOSE AMONG FEATURE VARIABLES FOR CREATING NEW FEATURES

FEATURE ENGINEERING	Description	Operation / Formula
Energy Production Efficiency	capture how much of the consumption is covered by renewable sources.	DE_WIND / DE_CONSUMPTION
Weather Impact on Energy	to see if rain has a direct impact on hydroelectric power production	FR_RAIN * FR_HYDRO
Net Exchange Ratios:	contribution of cross-border exchanges	(DE_FR_EXCHANGE - FR_DE_EXCHANGE) / (DE_CONSUMPTION + FR_CONSUMPTION)
Cross-Feature Interactions	capture the proportion of renewable energy in total consumption.	(FR_WIND + FR_SOLAR) / FR_CONSUMPTION
Temperature Effect on Consumption	Capture the effect of changing temperature on consumption	FR_TEMP * FR_CONSUMPTION and DE_TEMP * DE_CONSUMPTION
Renewable vs. Non-renewable Ratios	Calculate the ratio of renewable to non-renewable energy production for each country	(FR_WINDPOW + FR_SOLAR) / (FR_COAL + FR_GAS)
Carbon Intensity	Create a feature representing the carbon intensity of electricity generation	(COAL_RET + GAS_RET) / (DE_CONSUMPTION + FR_CONSUMPTION)

New Baseline Models

(Post Feature Engineering)

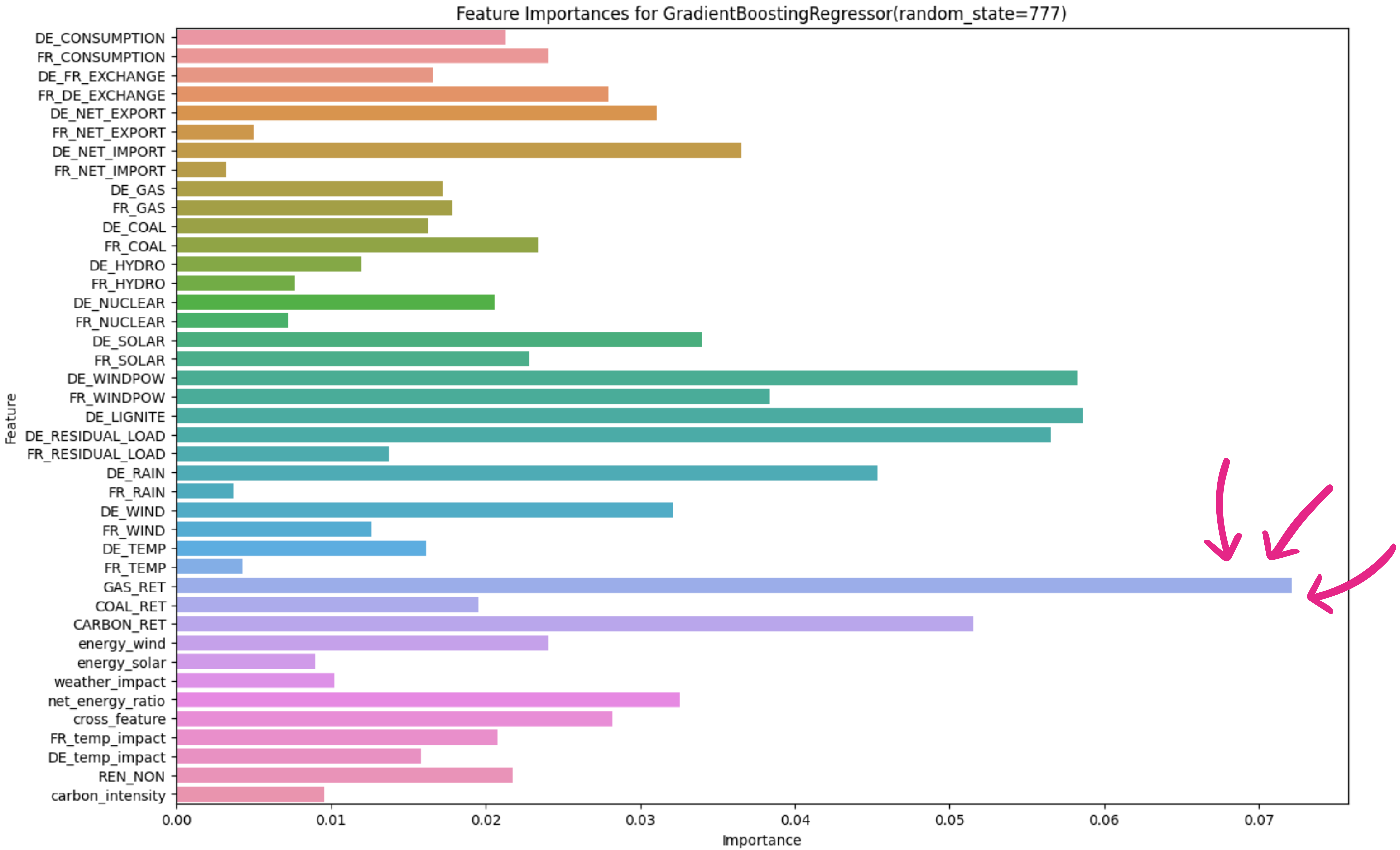
Now that we have a lot new features , it may help us in better capturing information for our baseline models as we can see below -

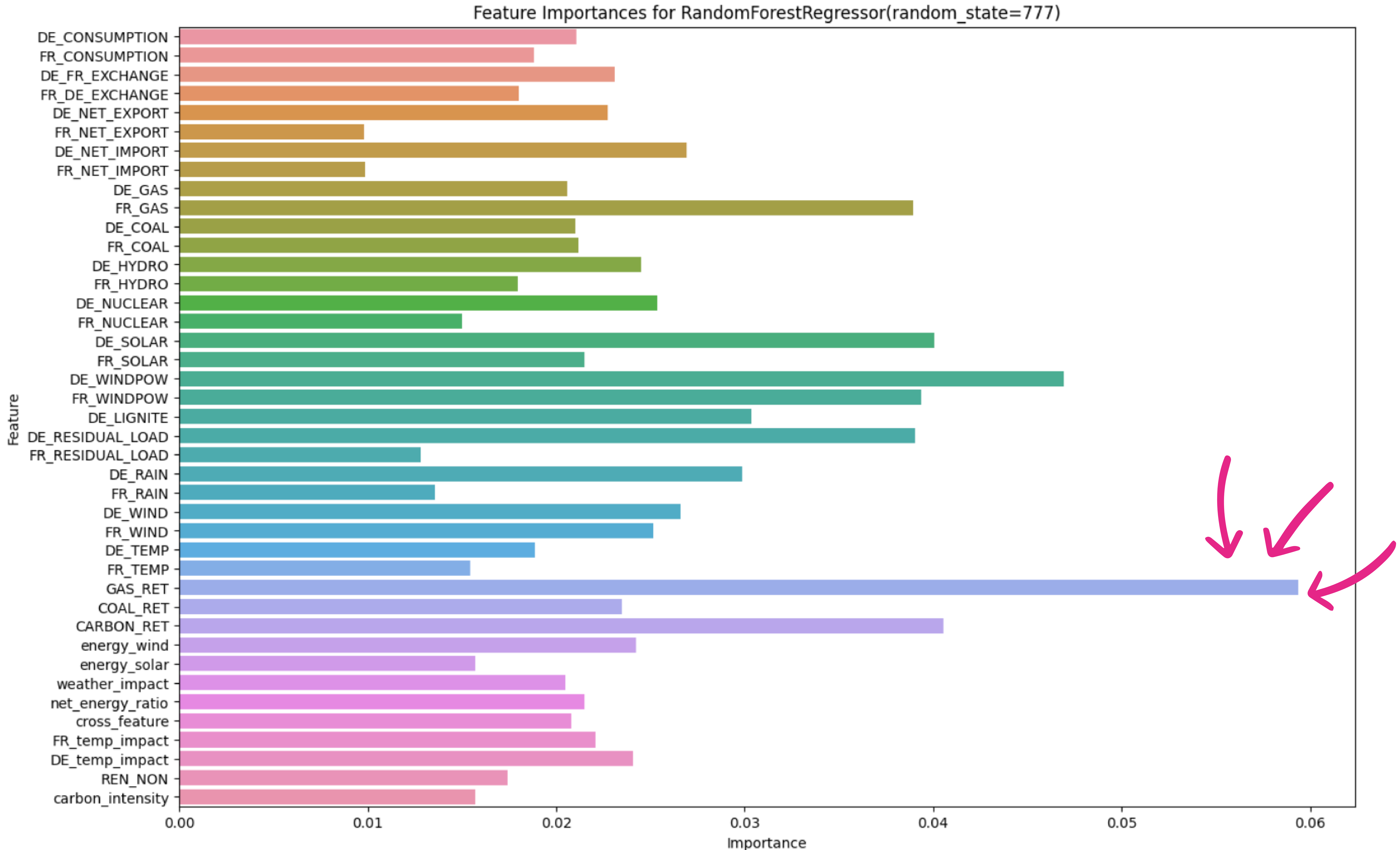
	Model	MSE	MAE	R-squared	Spearman Correlation	Training Time (s)
0	linear_regression	1.007896	0.570350	0.057728	0.282749	0.008636
1	random_forest	0.460711	0.410996	0.569285	0.705836	7.016092
2	gradient_boosting	0.632364	0.485608	0.408809	0.527472	1.703051
3	catboost	0.467540	0.426235	0.562901	0.666339	0.624797
4	adaboost	1.058812	0.767298	0.010127	0.206557	0.468092
5	lightgbm	0.517071	0.430942	0.516595	0.678818	0.500108
6	xgboost	0.448704	0.409082	0.580511	0.697189	0.908484

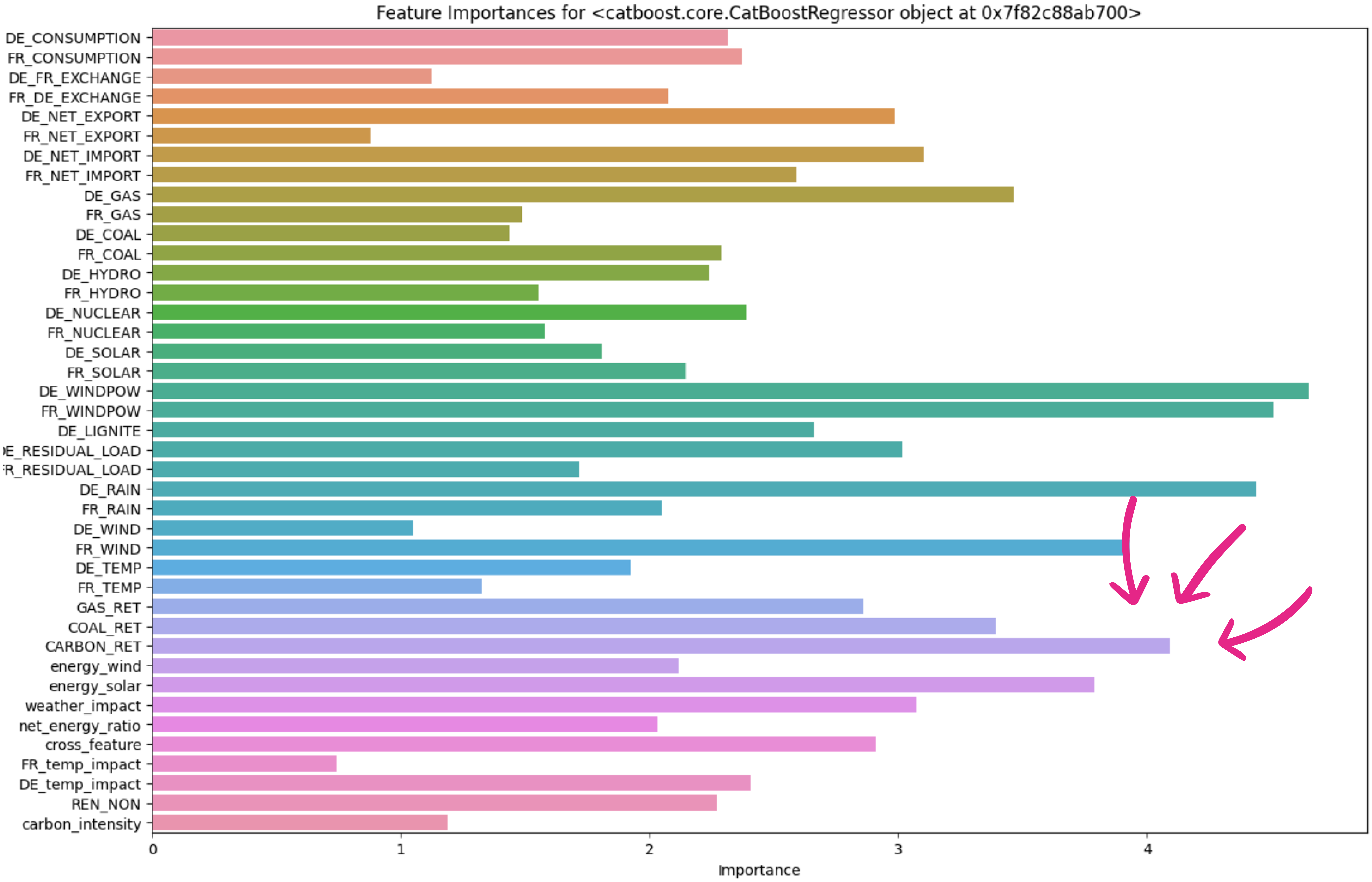
LR SCORE IMPROVED FROM 27.9 TO 28.2

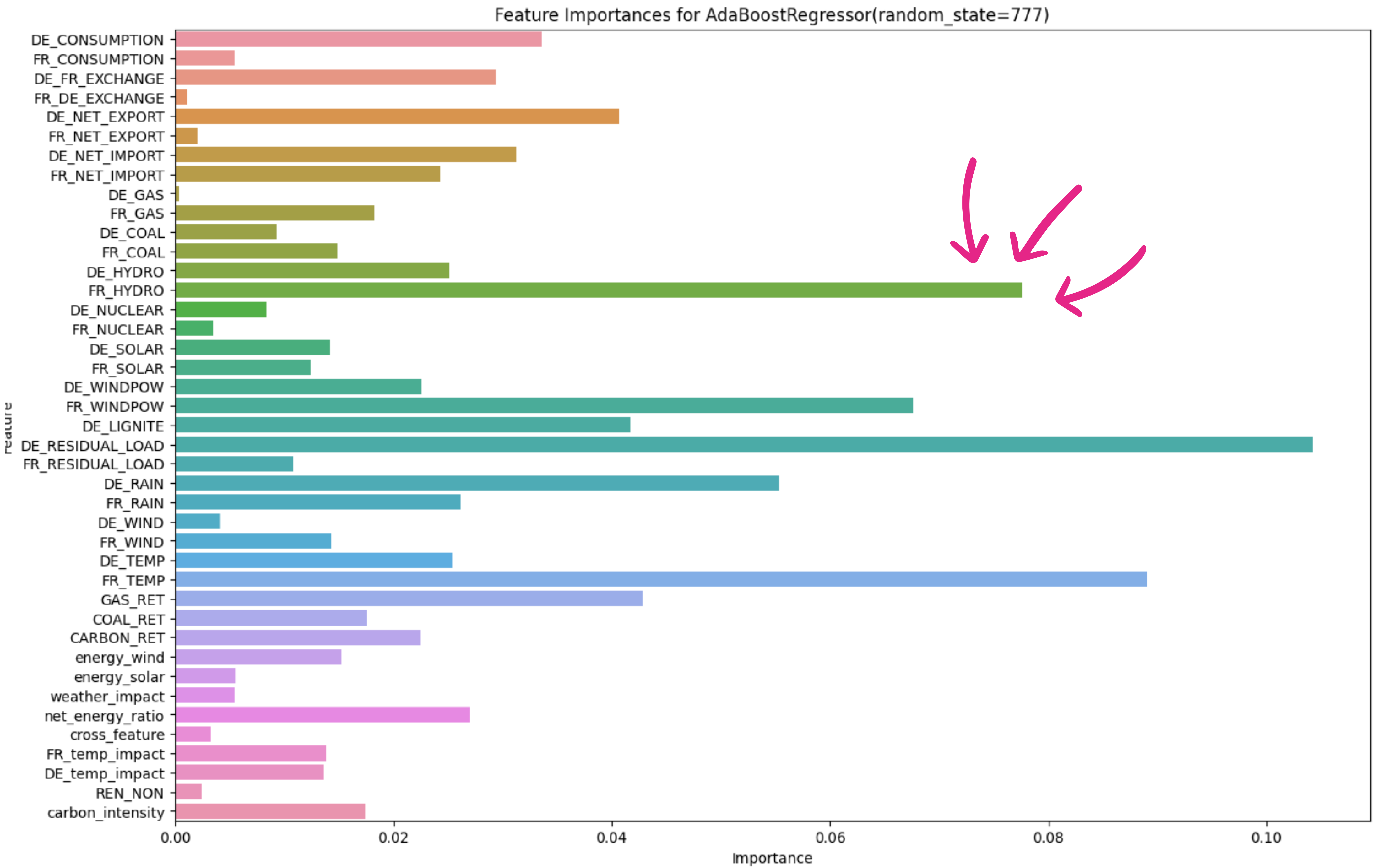
NOW WE WILL LOOK BRIEFLY AT FEATURE IMPORTANCE OF THE NEW BASELINE MODELS POST FEATURE ENGINEERING AND SEE THE IMPACT OF OUR NEW FEATURES

MORE ON THE OTHER SIDE

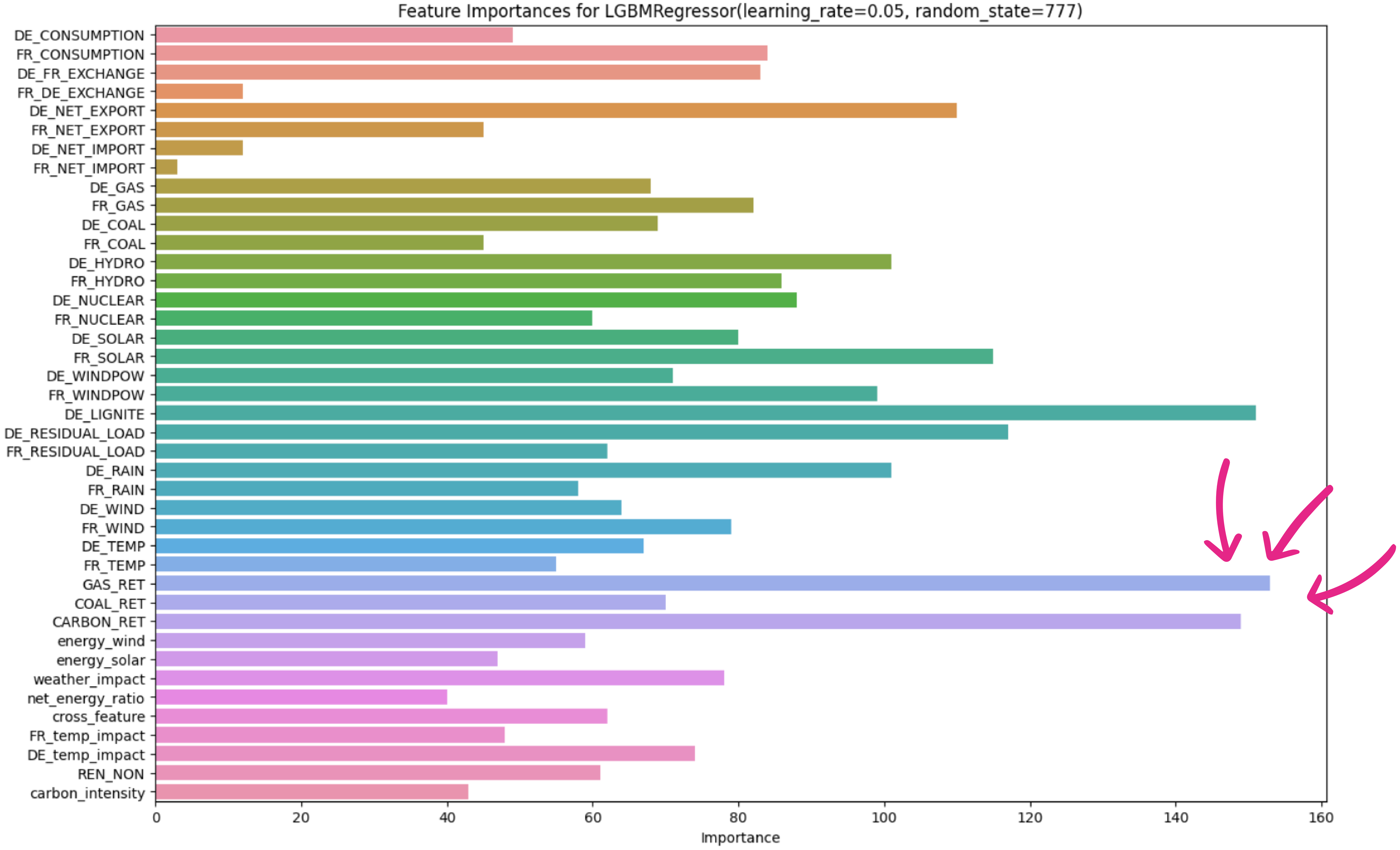

Feature Importances for GradientBoostingRegressor(random_state=777)
DE_CONSUMPTION
FR_CONSUMPTION
DE_FR_EXCHANGE
FR_DE_EXCHANGE
DE_NET_EXPORT
FR_NET_EXPORT
DE_NET_IMPORT
FR_NET_IMPORT
DE_GAS
FR_GAS
DE_COAL
FR_COAL
DE_HYDRO
FR_HYDRO
DE_NUCLEAR
FR_NUCLEAR
DE_SOLAR
FR_SOLAR
DE_WINDPOW
FR_WINDPOW
DE_LIGNITE
DE_RESIDUAL_LOAD
FR_RESIDUAL_LOAD
DE_RAIN
FR_RAIN
DE_WIND
FR_WIND
DE_TEMP
FR_TEMP
GAS_RET
COAL_RET
CARBON_RET
energy_wind
energy_solar
weather_impact
net_energy_ratio
cross_feature
FR_temp_impact
DE_temp_impact
REN_NON
carbon_intensity
Feature
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Importance


Feature Importances for RandomForestRegressor(random_state=777)
DE_CONSUMPTION
FR_CONSUMPTION
DE_FR_EXCHANGE
FR_DE_EXCHANGE
DE_NET_EXPORT
FR_NET_EXPORT
DE_NET_IMPORT
FR_NET_IMPORT
DE_GAS
FR_GAS
DE_COAL
FR_COAL
DE_HYDRO
FR_HYDRO
DE_NUCLEAR
FR_NUCLEAR
DE_SOLAR
FR_SOLAR
DE_WINDPOW
FR_WINDPOW
DE_LIGNITE
DE_RESIDUAL_LOAD
FR_RESIDUAL_LOAD
DE_RAIN
FR_RAIN
DE_WIND
FR_WIND
DE_TEMP
FR_TEMP
GAS_RET
COAL_RET
CARBON_RET
energy_wind
energy_solar
weather_impact
net_energy_ratio
cross_feature
FR_temp_impact
DE_temp_impact
REN_NON
carbon_intensity
Feature
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Importance

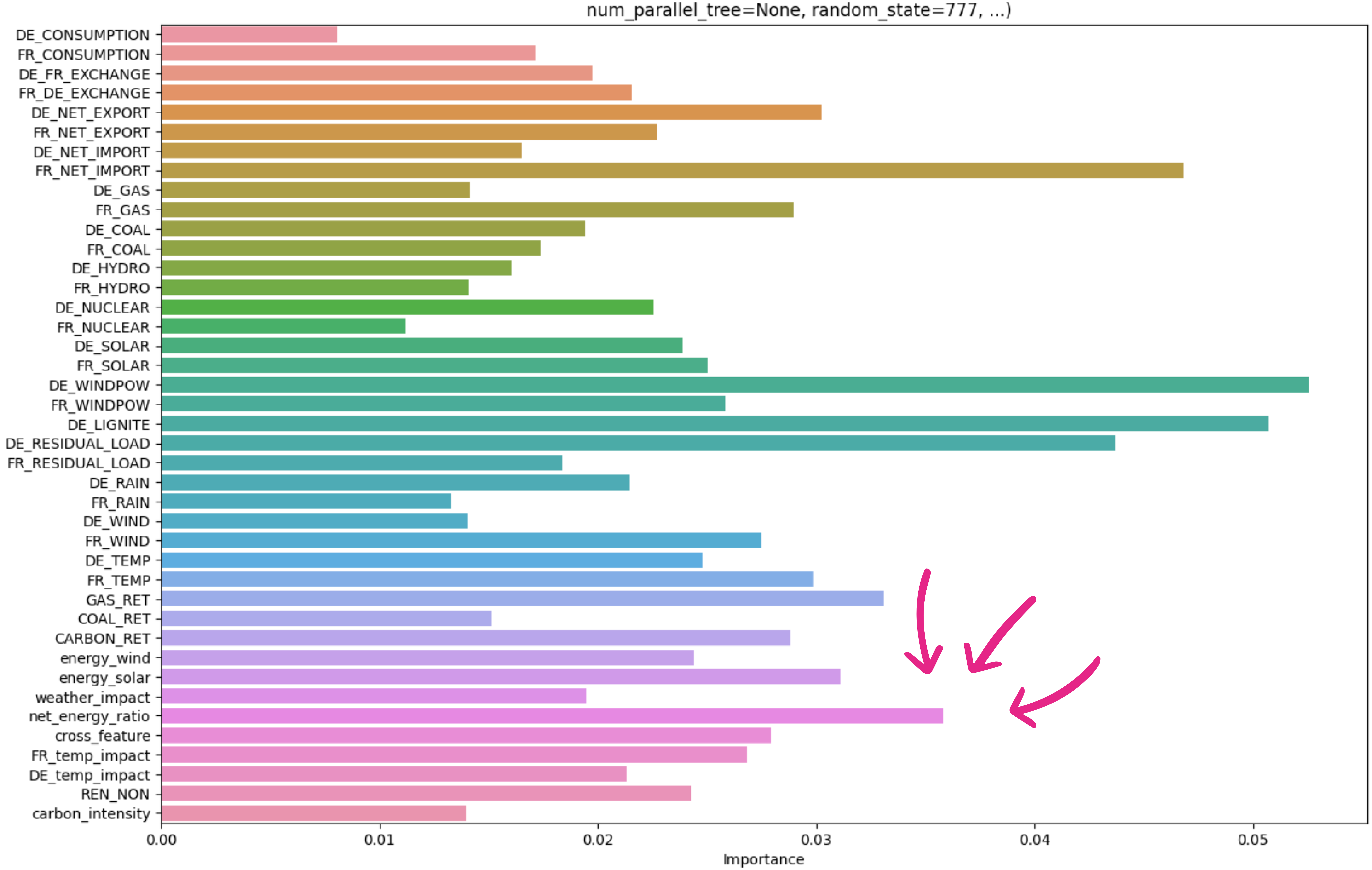

Feature Importances for <catboost.core.CatBoostRegressor object at 0x7f82c88ab700>
DE_CONSUMPTION
FR_CONSUMPTION
DE_FR_EXCHANGE
FR_DE_EXCHANGE
DE_NET_EXPORT
FR_NET_EXPORT
DE_NET_IMPORT
FR_NET_IMPORT
DE_GAS
FR_GAS
DE_COAL
FR_COAL
DE_HYDRO
FR_HYDRO
DE_NUCLEAR
FR_NUCLEAR
DE_SOLAR
FR_SOLAR
DE_WINDPOW
FR_WINDPOW
DE_LIGNITE
E_RESIDUAL_LOAD
R_RESIDUAL_LOAD
DE_RAIN
FR_RAIN
DE_WIND
FR_WIND
DE_TEMP
FR_TEMP
GAS_RET
COAL_RET
CARBON_RET
energy_wind
energy_solar
weather_impact
net_energy_ratio
cross_feature
FR_temp_impact
DE_temp_impact
REN_NON
carbon_intensity
0
1
2
3
4
Importance



Feature Importances for AdaBoostRegressor(random_state=777)
DE_CONSUMPTION
FR_CONSUMPTION
DE_FR_EXCHANGE
FR_DE_EXCHANGE
DE_NET_EXPORT
FR_NET_EXPORT
DE_NET_IMPORT
FR_NET_IMPORT
DE_GAS
FR_GAS
DE_COAL
FR_COAL
DE_HYDRO
FR_HYDRO
DE_NUCLEAR
FR_NUCLEAR
DE_SOLAR
FR_SOLAR
DE_WINDPOW
FR_WINDPOW
DE_LIGNITE
DE_RESIDUAL_LOAD
FR_RESIDUAL_LOAD
DE_RAIN
FR_RAIN
DE_WIND
FR_WIND
DE_TEMP
FR_TEMP
GAS_RET
COAL_RET
CARBON_RET
energy_wind
energy_solar
weather_impact
net_energy_ratio
cross_feature
FR_temp_impact
DE_temp_impact
REN_NON
carbon_intensity
0.00
0.02
0.04
0.06
0.08
0.10
Importance



Feature Importances for LGBMRegressor(learning_rate=0.05, random_state=777)
DE_CONSUMPTION
FR_CONSUMPTION
DE_FR_EXCHANGE
FR_DE_EXCHANGE
DE_NET_EXPORT
FR_NET_EXPORT
DE_NET_IMPORT
FR_NET_IMPORT
DE_GAS
FR_GAS
DE_COAL
FR_COAL
DE_HYDRO
FR_HYDRO
DE_NUCLEAR
FR_NUCLEAR
DE_SOLAR
FR_SOLAR
DE_WINDPOW
FR_WINDPOW
DE_LIGNITE
DE_RESIDUAL_LOAD
FR_RESIDUAL_LOAD
DE_RAIN
FR_RAIN
DE_WIND
FR_WIND
DE_TEMP
FR_TEMP
GAS_RET
COAL_RET
CARBON_RET
energy_wind
energy_solar
weather_impact
net_energy_ratio
cross_feature
FR_temp_impact
DE_temp_impact
REN_NON
carbon_intensity
0
20
40
60
80
100
120
140
160
Importance


num_parallel_tree=None, random_state=777, ...)
DE_CONSUMPTION
FR_CONSUMPTION
DE_FR_EXCHANGE
FR_DE_EXCHANGE
DE_NET_EXPORT
FR_NET_EXPORT
DE_NET_IMPORT
FR_NET_IMPORT
DE_GAS
FR_GAS
DE_COAL
FR_COAL
DE_HYDRO
FR_HYDRO
DE_NUCLEAR
FR_NUCLEAR
DE_SOLAR
FR_SOLAR
DE_WINDPOW
FR_WINDPOW
DE_LIGNITE
DE_RESIDUAL_LOAD
FR_RESIDUAL_LOAD
DE_RAIN
FR_RAIN
DE_WIND
FR_WIND
DE_TEMP
FR_TEMP
GAS_RET
COAL_RET
CARBON_RET
energy_wind
energy_solar
weather_impact
net_energy_ratio
cross_feature
FR_temp_impact
DE_temp_impact
REN_NON
carbon_intensity
0.00
0.01
0.02
0.03
0.04
0.05
Importance

NOW THAT WE WILL PICK PROCEED WITH A SIMPLE WEIGHTED AVERAGE TECHNIQUE OF OUR MOST EFFECTIVE BASELINE MODELS THAT INCLUDE RANDOM FOREST , XGBOOST AND CATBOOST

MORE ON THE OTHER SIDE

Weighted Average

```
weights = {
    "random_forest": 0.4,
    "xgboost": 0.35,
    "catboost": 0.25
}
```

WE GET AN
IMPORVED SCORE

Ranking	Date	Method	Parameters	Public score	Selection
1	Feb. 24, 2024, 9:46 p.m.	weighted avg	rf++cat+XGB	0.20025497739154208	Select

WEIGHTS TAKEN IN ACCORDANCE TO THE BASELINE SCORE

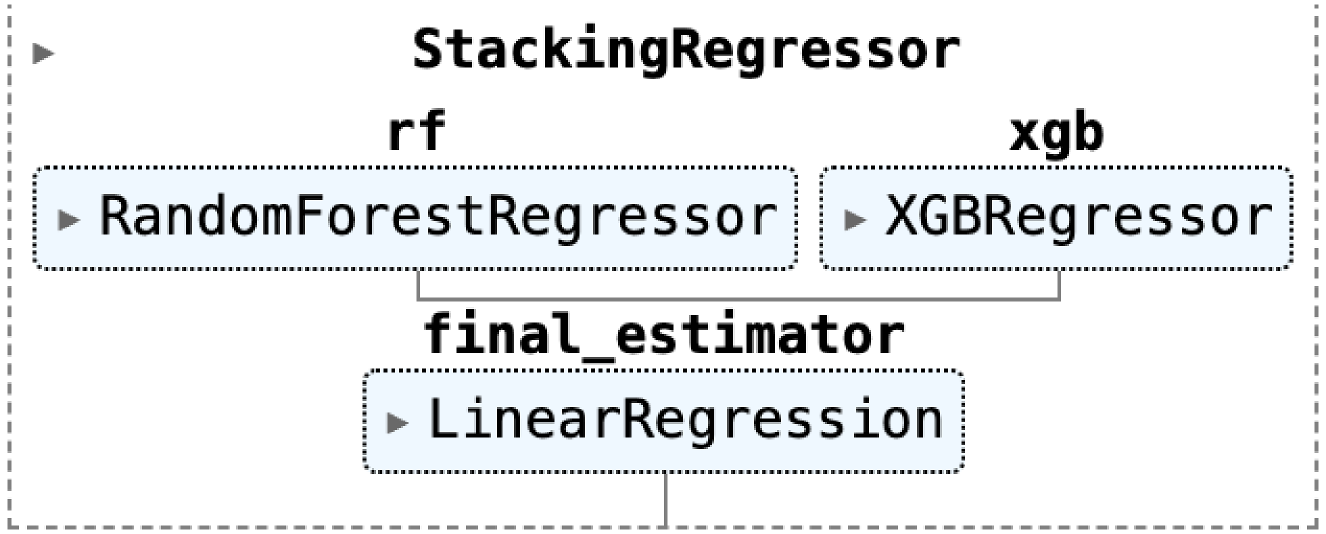
IMPLEMENTING NEW STATEGY USING ENSEMBLE METHODS AND HYPERTUNING

MORE ON THE OTHER SIDE

Hypertuned Models + Stacking

Tuned XGBoost Parameters

```
Best hyperparameters: {'n_estimators': 958, 'max_depth': 3, 'learning_rate': 0.012983502840524876, 'subsample': 0.
8613472551443894, 'colsample_bytree': 0.8592122590269452, 'gamma': 6.265636328745365e-08, 'reg_lambda': 32.1947181
6344407, 'reg_alpha': 42.70703483364794}
```



Tuned Random Forest Parameters

```
hyperparameters: {'n_estimators': 352, 'max_depth': 14, 'min_samples_split': 5, 'min_samples_leaf': 13
```

WE HYPERTUNED MODELS USING OPTUNA AS IT PROVIEDS MORE ACCURATE RESULTS THAN GRIDCV

Best Score !

Ranking	Date	Method	Parameters	Public score	Selection
1	Feb. 15, 2024, 7:56 p.m.	stacking	xgb+rf	0.20284439043303762	Select

All submission results

Ranking	Date	Method	Parameters	Public score	Selection
1	Feb. 15, 2024, 6:48 p.m.	rf+hypertuning	rf	0.19228844668280182	Select
2	Feb. 12, 2024, 9:57 p.m.	Baseline	LR	0.18594542447369478	Select
3	Feb. 15, 2024, 7:17 p.m.	XGB	hypertuned	0.1808433026736426	Select
4	Feb. 12, 2024, 10:04 p.m.	New_Baseline_28.5	LR	0.15908193724817526	Select

Merci !