PREDICTION OF THE NET HOURLY ELECTRICAL ENERGY OUTPUT OF A COMBINED CYCLE POWER PLANT

MACHINE LEARNING FOUNDATION FOR PRODUCT MANAGERS

Carlo De Rossi, October 2024

ASSIGNMENT

- In this project we will build a model to predict the electrical energy output of a Combined Cycle Power Plant, which uses a combination of gas turbines, steam turbines, and heat recovery steam generators to generate power.
- We have a set of 9568 hourly average ambient environmental readings from sensors at the power plant which we will use in our model.
- The columns in the dataset consist of hourly average ambient variables:

Ambient Temperature (AT) in the range 1.81°C to 37.11°C
Ambient Pressure (AP) in the range 992.89-1033.30 millibar
Relative Humidity (RH) in the range 25.56% to 100.16%
Exhaust Vacuum (V) in the range 25.36-81.56 cm Hg
Net hourly electrical energy output (PE) 420.26-495.76 MW (Target we are trying to predict)

DETERMINE APPROACH AND OUTPUT METRIC

- For predicting the electrical energy output, we need a REGRESSION approach, as the target variable (PE) is continuous (numeric).

- Output Metric to evaluate performance:
  Mean Squared Error (MSE)
  Root Mean Squared Error (RMSE)
  R^2 (R-squared) Score (aka Coefficient of determination)

- Algorithms:
  Linear Regression
  Random Forest Regressor
  [Opt] Gradient Boosting Regressor
  [Opt]Support Vector Regression (SVR)

DATA ANALYSIS

- Dataset is clean: No missing Data, no invalid data
- As suggested by National Oceanic and Atmospheric Administration (NOAA) we add the Heat Index (newly derived feature combining temperature and humidity). Reasons for this are:
- Combining Influences: By integrating these factors, the Heat Index offers a more comprehensive representation of the actual conditions at the power plant.
- Capturing Non-linear Relationships: The relationship between temperature, humidity, and energy output isn't straightforward. The Heat Index captures the compounded effect of these variables in a non-linear way, which might be challenging for a model to pick up if treated separately.
- Enhancing Model Performance: Adding this derived feature can help the model grasp the true impact of weather conditions on the power plant's performance, potentially improving prediction accuracy.
- Relevance to the Context: In scenarios like power generation, the combined effect of temperature and humidity (as captured by the Heat Index) can have a significant impact on operations, making it a valuable feature for more precise forecasting.

DATA ANALYSIS

Given the data ranges, we need also to normalize the data. Here is why:

Consistency Across Features and Uniform Scale

Different features might have varying scales (e.g., age in years, income in dollars). Normalization brings all features to a similar scale, making it easier for algorithms to process.

Improved Performance: Algorithms converge faster on normalized data, boosting computational efficien cy.

Equal Weight and Better Accuracy

Prevents features with larger ranges from dominating the model, ensuring each feature contributes prop ortionately to the learning process.

Enhances the model's accuracy by reducing bias toward features with inherently larger values.

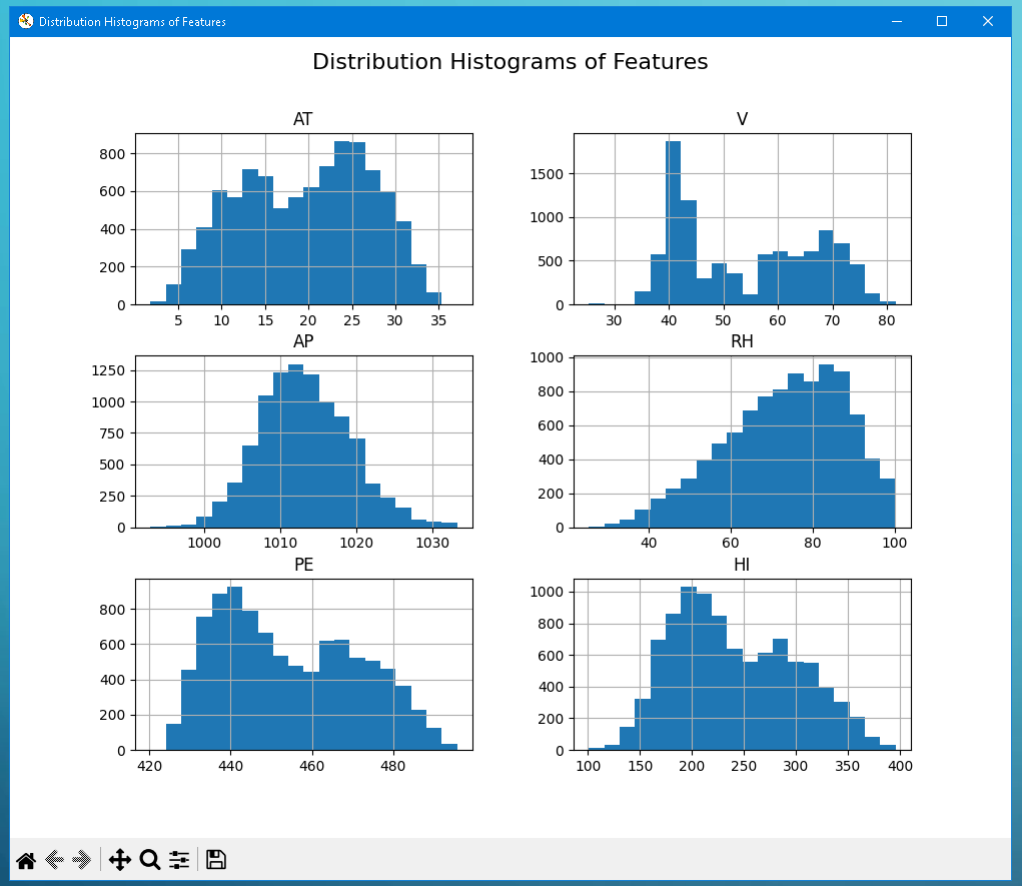
Distribution of dataset:

	AT	V	AP	RH	PE	HI
count	9568.000000	9568.000000	9568.000000	9568.000000	9568.000000	9568.000000
mean	19.651231	54.305804	1013.259078	73.308978	454.365009	240.823205
std	7.452473	12.707893	5.938784	14.600269	17.066995	58.379786
min	1.810000	25.360000	992.890000	25.560000	420.260000	100.886634
25%	13.510000	41.740000	1009.100000	63.327500	439.750000	194.329472
50%	20.345000	52.080000	1012.940000	74.975000	451.550000	230.695371
75%	25.720000	66.540000	1017.260000	84.830000	468.430000	286.932419
max	37.110000	81.560000	1033.300000	100.160000	495.760000	395.716914

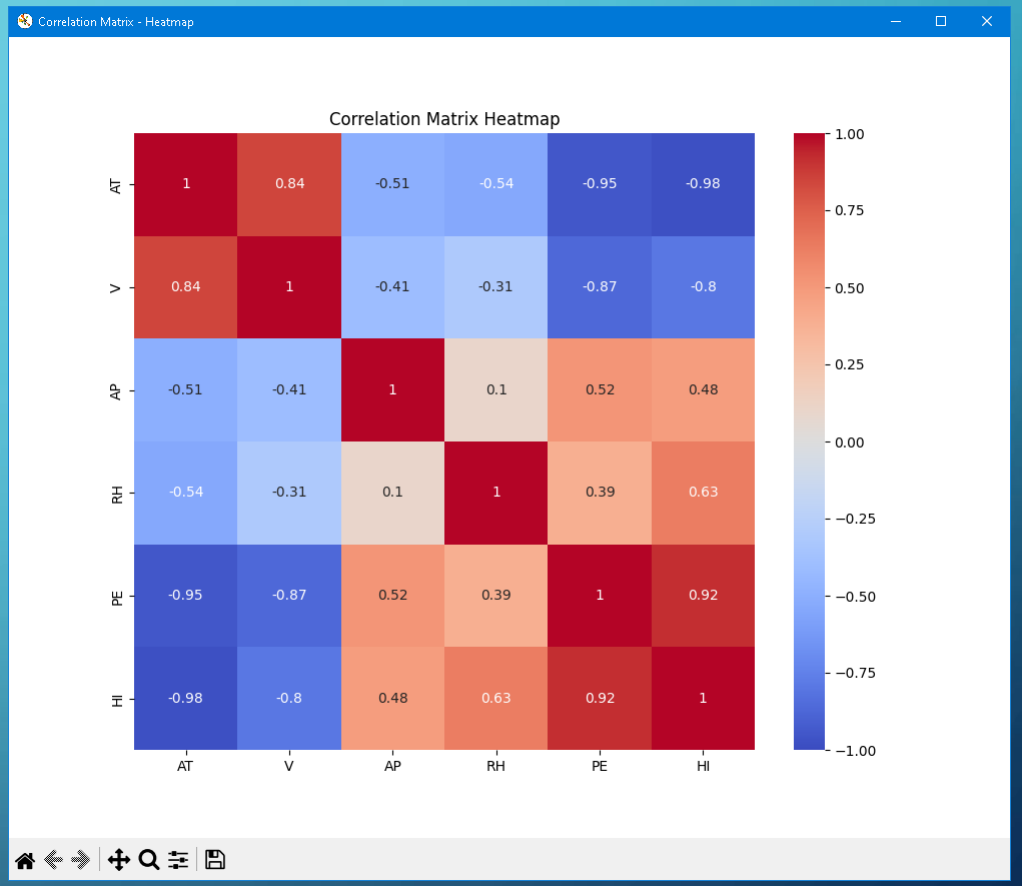
Normalized data:

	AT	V	AP	RH	PE	HI
0	0.372521	0.291815	0.771591	0.638204	463.26	0.592143
1	0.662040	0.669039	0.671863	0.449330	444.37	0.319329
2	0.093484	0.249822	0.476862	0.892493	488.56	0.915568
3	0.539660	0.568683	0.429349	0.684718	446.48	0.451011
4	0.255241	0.216014	0.404355	0.952547	473.90	0.709942
...	...	...	...	...	...	...
9563	0.420397	0.432918	0.522643	0.877212	460.03	0.547888
9564	0.322380	0.245907	0.761693	0.552547	469.62	0.611902
9565	0.835977	0.871352	0.495669	0.146381	429.57	0.106514
9566	0.642210	0.784520	0.518931	0.493700	435.74	0.344452
9567	0.560623	0.661210	0.602326	0.567158	453.28	0.419629

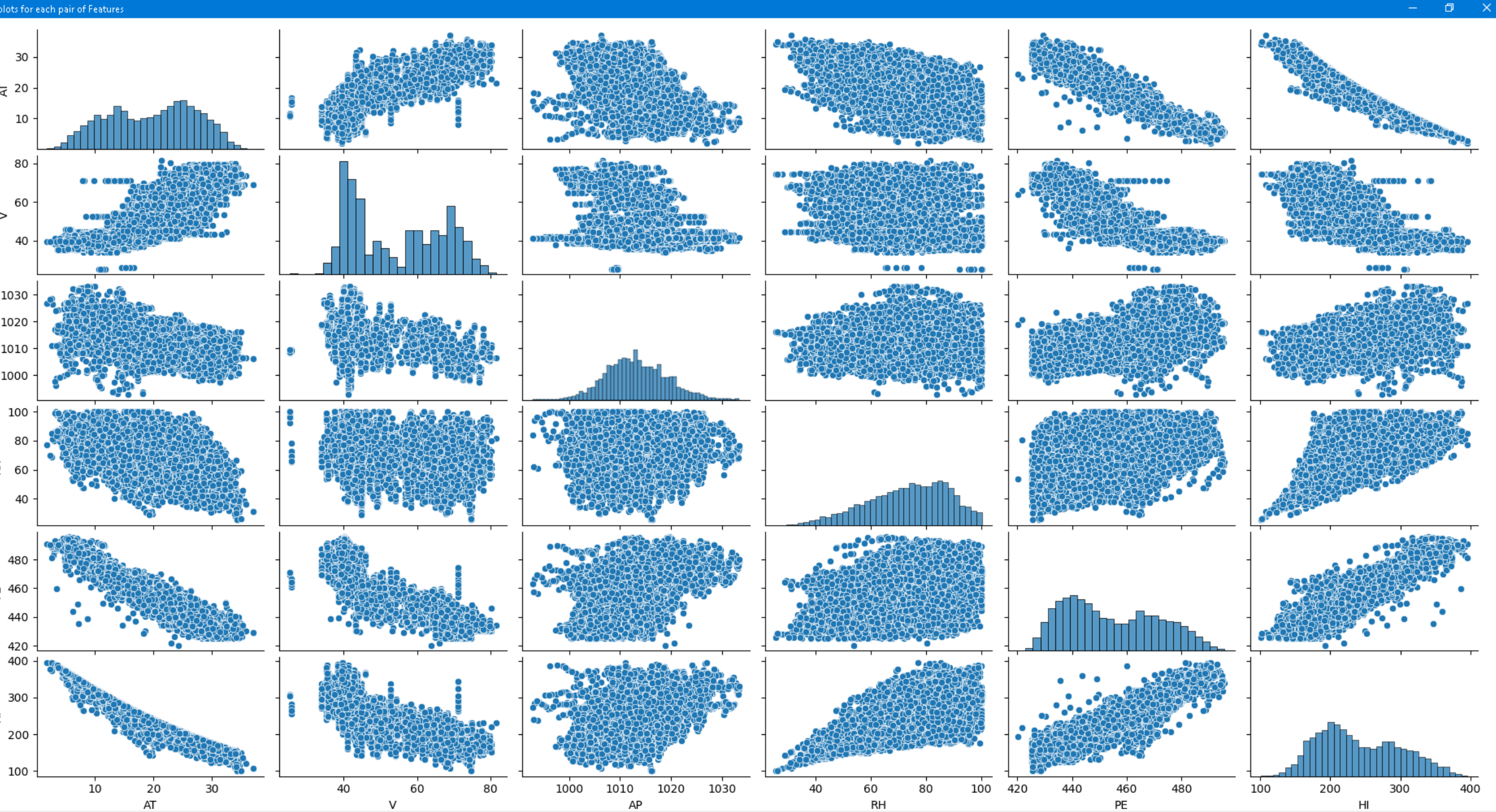
DATA ANALYSIS - DIAGRAMS



Distribution Histograms of Features



Correlation Matrix HeatMap


plots for each pair of Features
AT
V
AP
RH
PE
HI
(x, y) = (485.9, 1024.7)

LINEAR REGRESSION

Starting with a simple model like Linear Regression has its perks:

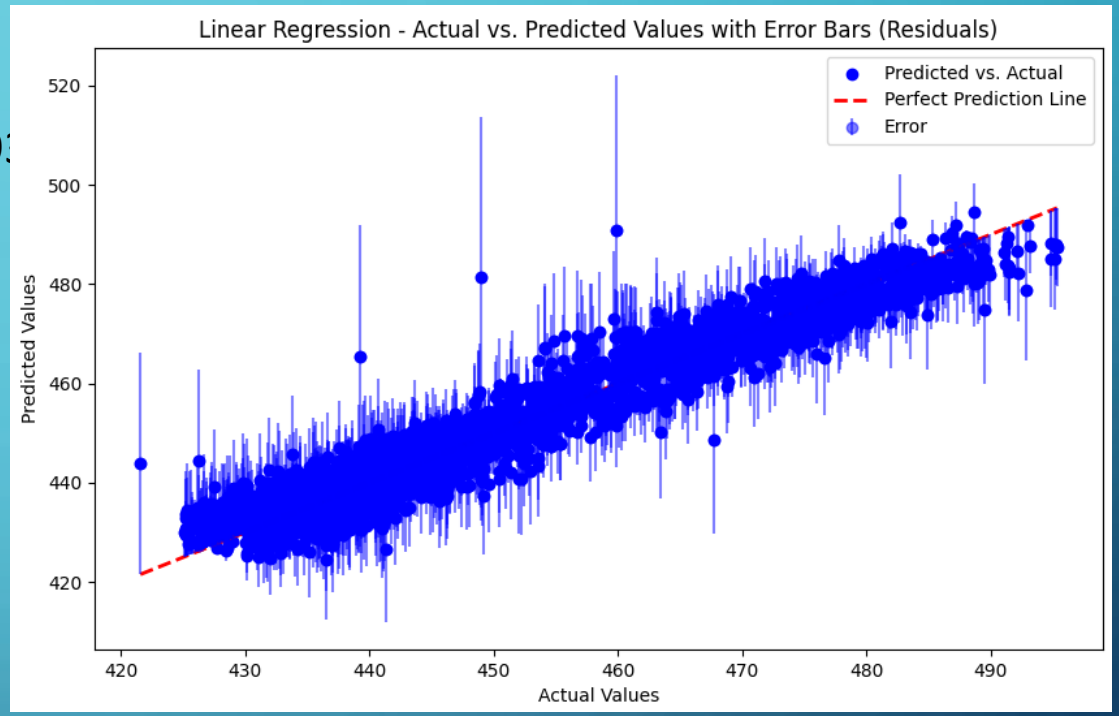
- Clarity and Interpretability: Linear models are straightforward and their results are easy to interpret, making it clear how features impact the target variable.
- Baseline Performance: It provides a benchmark. If a more complex model doesn't significantly out-perform Linear Regression, it might not be worth the extra complexity.
- Quick and Efficient: Training a simple model is fast, allowing you to quickly get a sense of how well your features can predict the target.
- Diagnosing Issues: Insight into Relationships: It helps in understanding the basic relationships in your data. If Linear Regression performs poorly, it can highlight issues like nonlinearity, multi-collinearity, or the need for feature transformation.

Given the data diagrams, we expect the Linear Regression model to perform well (~80%)

Data are split in a train (~70-80%) and a evaluate and test datasets, with random assignment

LINEAR REGRESSION - EVALUATION

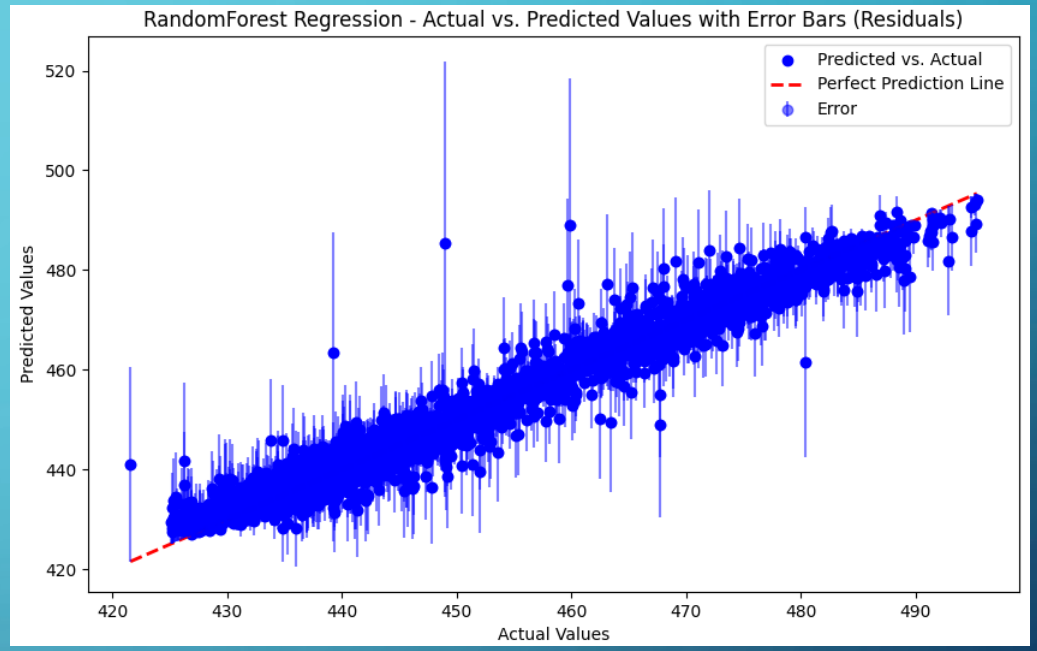
- Elapsed Training Time: 0.0039746761322021484
- Measured Prediction Time: 0.0010280609130859
- Linear Regression Cross-Validation MSE: 19.51
- Linear Regression Cross-Validation RMSE: 4.42
- Linear Regression Cross-Validation R^2: 0.93



As expected (less than 10K rows) the model is efficient in training and makes predictions quickly. The low values errors in the model's predictions indicate good accuracy. An RMSE of 4.42 suggests that the predictions are reasonably close to the actual values (see also residual graph) Cross-Validation R^2 0.93 is an excellent score, as it means that 93% of the variance in the data is explained by the model. In other words, the independent variables do a good job of predicting the target

RANDOM FOREST - EVALUATION

- Elapsed Training Time: 4.314298868179321
- Measured Prediction Time: 0.05500030517578125
- RandomForest Regression Cross-Validation MSE: 11.61
- RandomForest Regression Cross-Validation RMSE: 3.41
- RandomForest Regression Cross-Validation R^2: 0.96



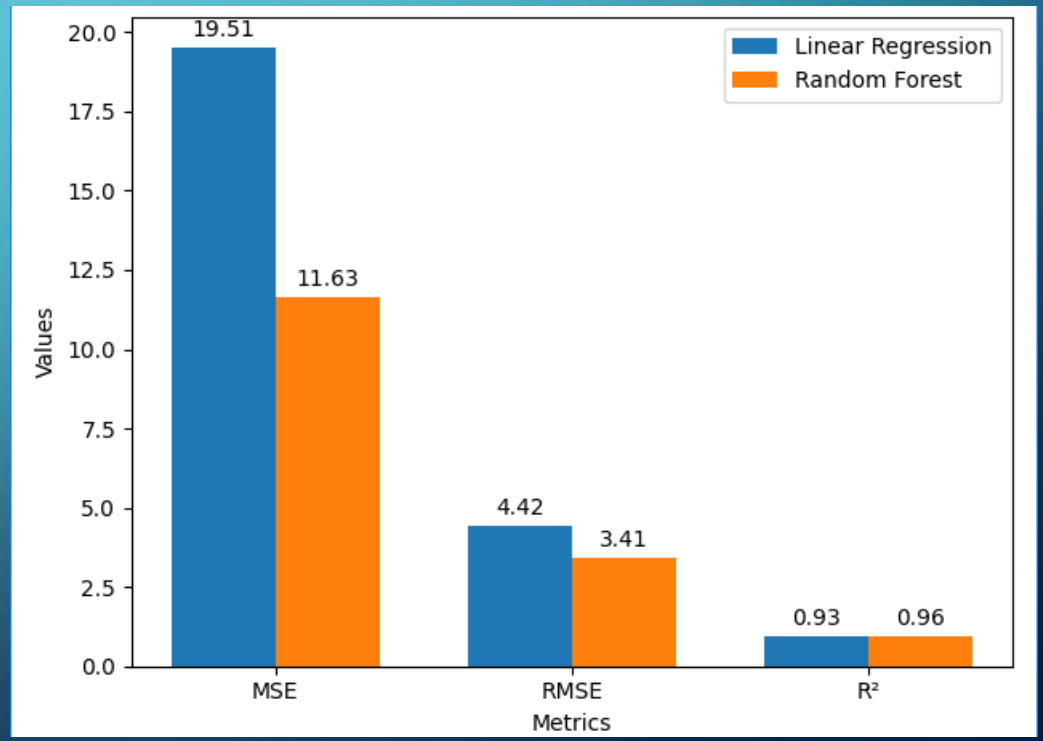
As expected, the model is significantly slower in training and in making predictions than the previous one (computationally more onerous despite small dataset). The lower values errors in the model's predictions indicate very good accuracy. An RMSE of 3.41 suggests that the predictions are reasonably close to the actual values (see also residual graph) Cross-Validation R^2 0.96 is an excellent score, as it means that 96% of the variance in the data is explained by the model.

MODELS COMPARISON

The two models show a significant difference in the training and prediction time and a comparable accuracy and RMSE values.

Only in the case when the small increase in accuracy would justify the significant training and prediction cost it would make sense not to use the Linear Regression model.
Given that for a Power Plant also a small increase in accuracy can produce significant monetary benefits, it needs to be evaluated within a specific Business Case.

For the context of this exercise, we conclude that the Linear Regression model is accurate enough and continue its evaluation.



ORDINARY LEAST SQUARES (OLS)

OLS is a method used in linear regression to estimate the parameters of the model: it aims to find the best-fitting line by minimizing the sum of the squared differences between the observed values and the values predicted by the linear model. OLS is widely used due to its simplicity and interpretability.

R-squared: 0.933: The model explains 93.3% of the variance in the dependent variable.
Adj. R-squared: 0.933: The Adjusted R-squared accounts for the number of predictors in the model.
A high Adjusted R-squared value close to the R-squared value signifies that the model is a good fit and isn't unnecessarily complex.
F-statistic: 2.663e+04: The F-statistic measures the overall significance of the model. A large F-statistic indicates that the independent variables collectively have a significant impact on the dependent variable.
Prob (F-statistic): 0.00: This is the p-value associated with the F-statistic. A value of 0.00 means that there is virtually no chance that the observed F-statistic is due to random chance, hence the model is statistically significant.

OLS Regression Results

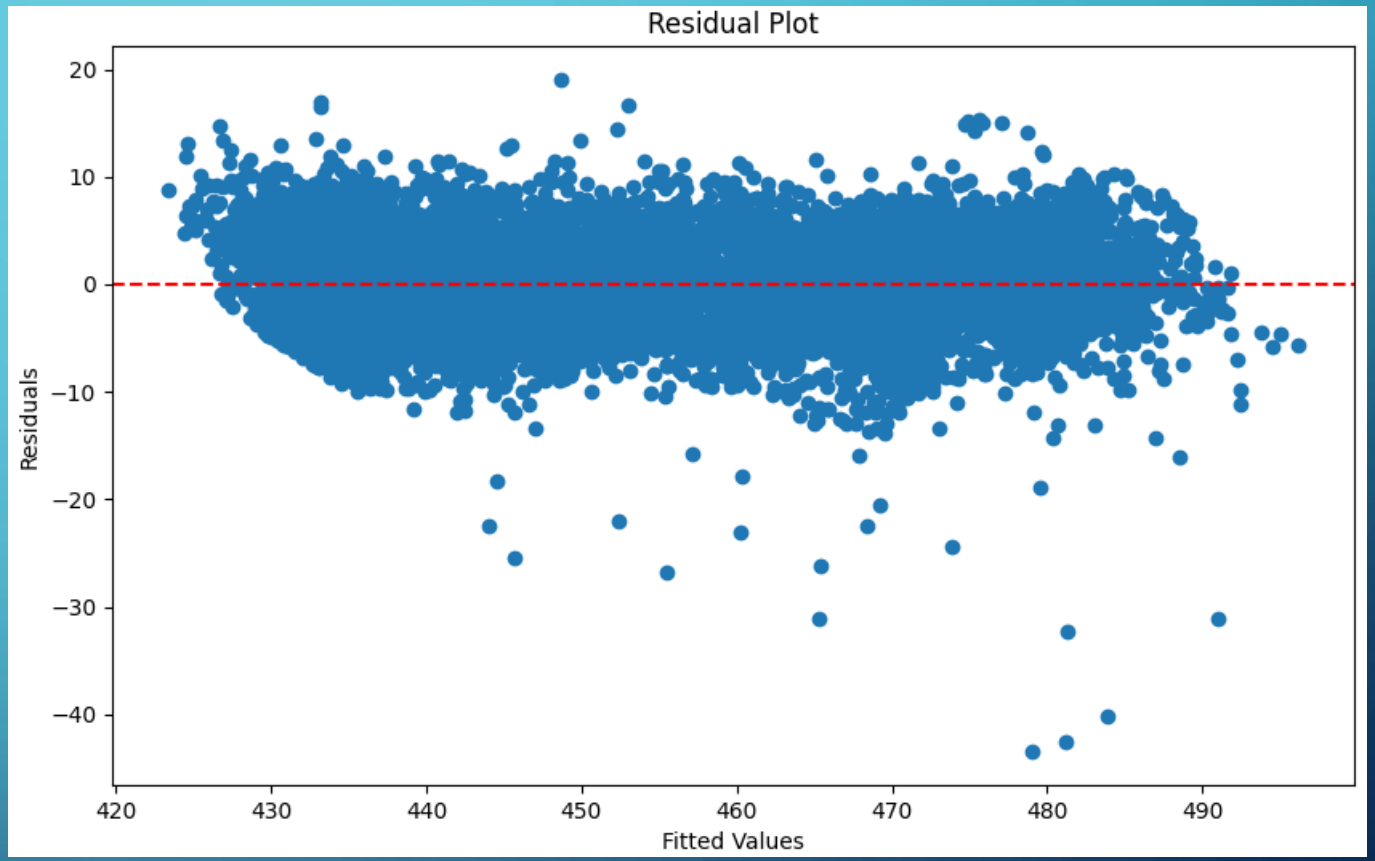
Dep. Variable:	PE	R-squared:	0.933
Model:	OLS	Adj. R-squared:	0.933
Method:	Least Squares	F-statistic:	2.663e+04
Date:	Tue, 15 Oct 2024	Prob (F-statistic):	0.00
Time:	15:46:13	Log-Likelihood:	-27790.
No. Observations:	9568	AIC:	5.559e+04
Df Residuals:	9562	BIC:	5.564e+04
Df Model:	5		
Covariance Type:	nonrobust		

OLS - RESIDUAL PLOT

This plot shows the residuals (errors) on the y-axis and the fitted values on the x-axis.
It helps to check for any patterns in the residuals (like a curve), which can indicate issues with the model.

The residuals in the picture are randomly scattered around the horizontal axis (y=0):

This suggests that the model's predictions are unbiased across all levels of the fitted values.



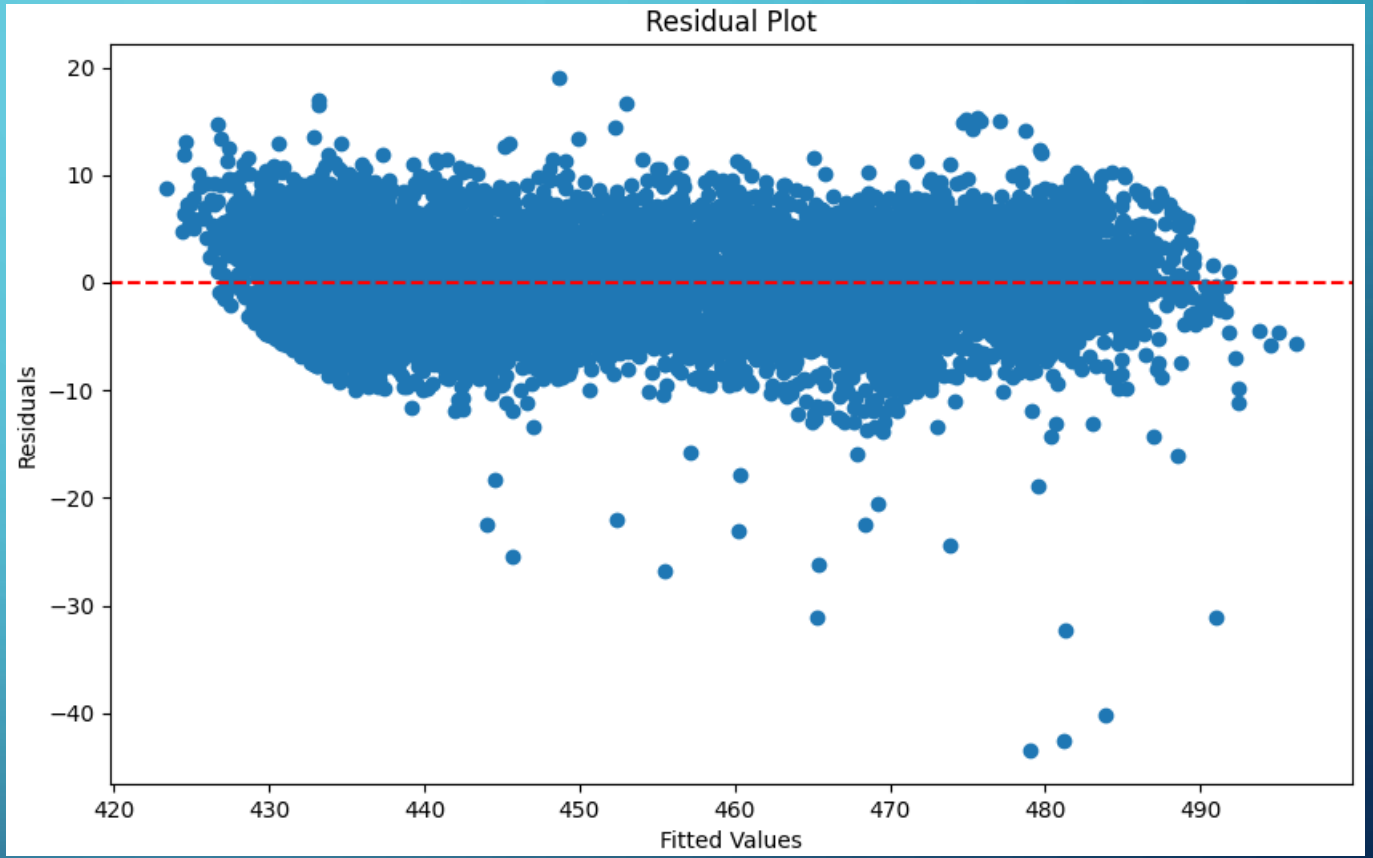
OLS - FITTED VS. ACTUAL VALUES PLOT

This plot compares the predicted values with the actual values..

Points far from the 45-degree reference line (where fitted values equal actual values) suggest larger errors in prediction.

The picture shows that the points lie close to the reference line :

This suggests that the model's predictions are close to the actual values.



REFERENCES

- National Oceanic and Atmospheric Administration (NOAA) and the Weather Prediction Center (www.wpc.ncep.noaa.gov)
- Heat Index Equation: Heat Index Equation (noaa.gov)
- GitHub Public Repo: https://github.com/carloderossi/
- Jupyter Notebook: https://github.com/carloderossi/CCP_NOTEBOOK.jpynb