Linear Regression Analysis

CS4372

Roman Hauksson-Neill

Ivan Masyuk

Introduction

We chose to analyze a dataset called “Combined Cycle Power Plant”, which contains readings of temperature, pressure, relative humidity, and exhaust vacuum of a power plant, which can be used to predict its net hourly electrical energy output.

Pre-Processing

Normality Analysis

Running the Shapiro-Wilk test on the features, we found that all features are non-normal.

Feature	p-value	Normality
ambient_temp	2.10e-30	Non-normal
vacuum	6.40e-48	Non-normal
ambient_pressure	5.96e-12	Non-normal
relative_humidity	1.47e-28	Non-normal
power_output	6.50e-36	Non-normal

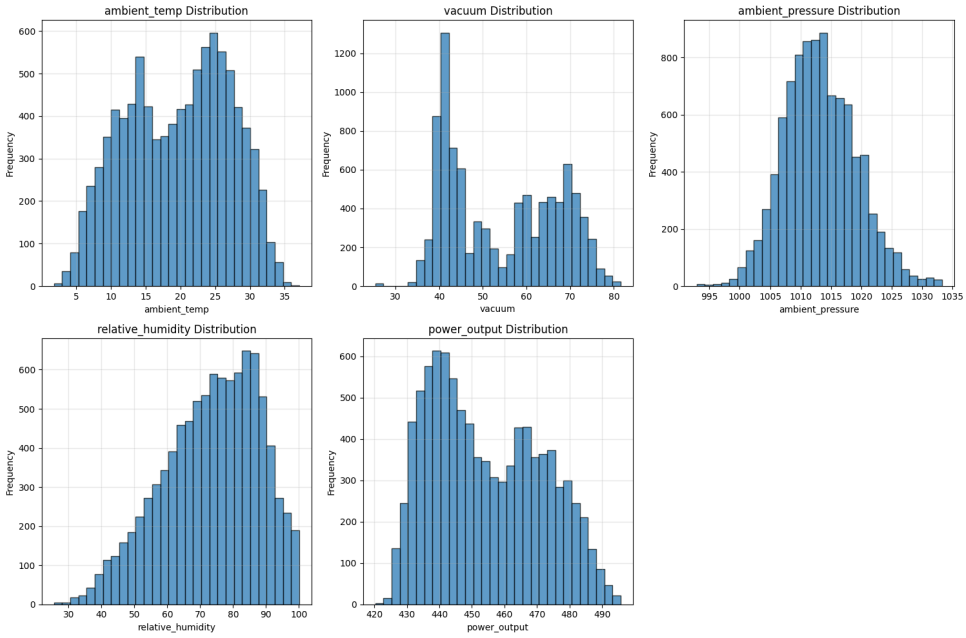

Figure 1: Feature distributions.

Each of the features has a totally different unit and scale. We standardized them so that they all have a mean of 0 and a standard deviation of 1.

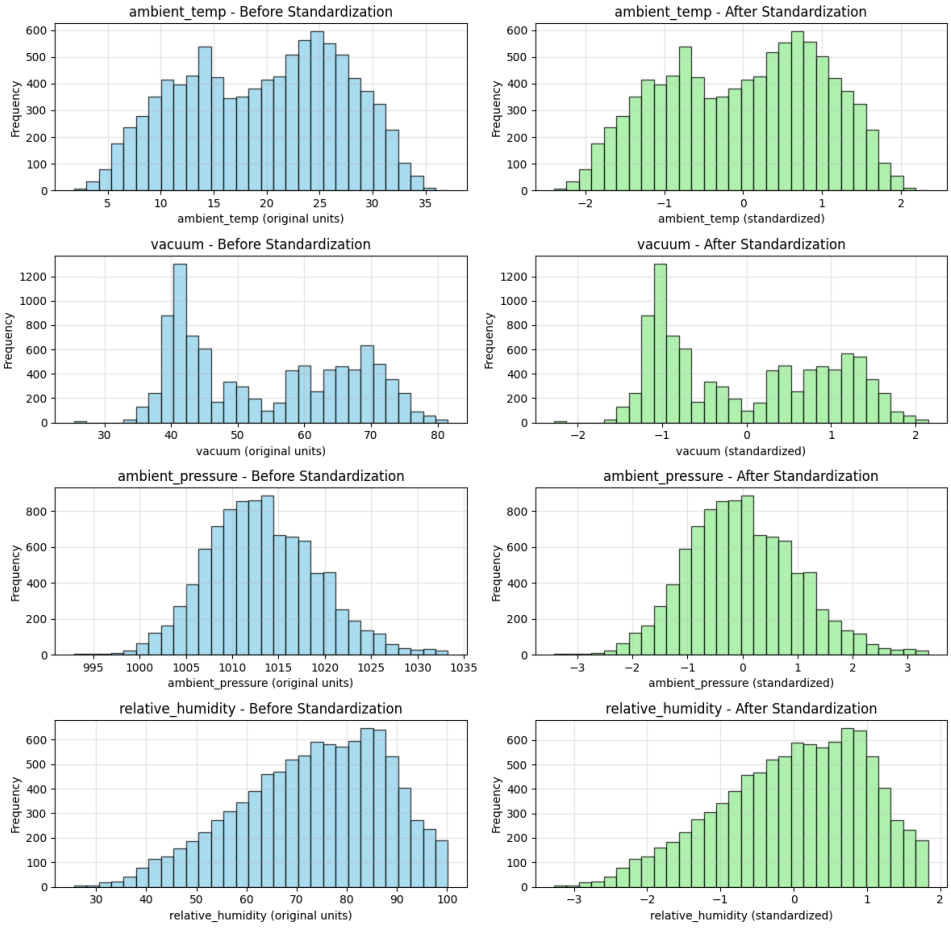

Figure 2: Feature distributions after standardization.

Correlation Analysis

Correlation analysis showed that the target variable (power output) is strongly correlated with the ambient temperature and vacuum features, and the ambient temperature and vacuum features are strongly correlated with each other.

Feature	Correlation	Direction
ambient_temp	−0.948	Negative
vacuum	−0.870	Negative
ambient_pressure	+0.519	Positive
relative_humidity	+0.391	Positive

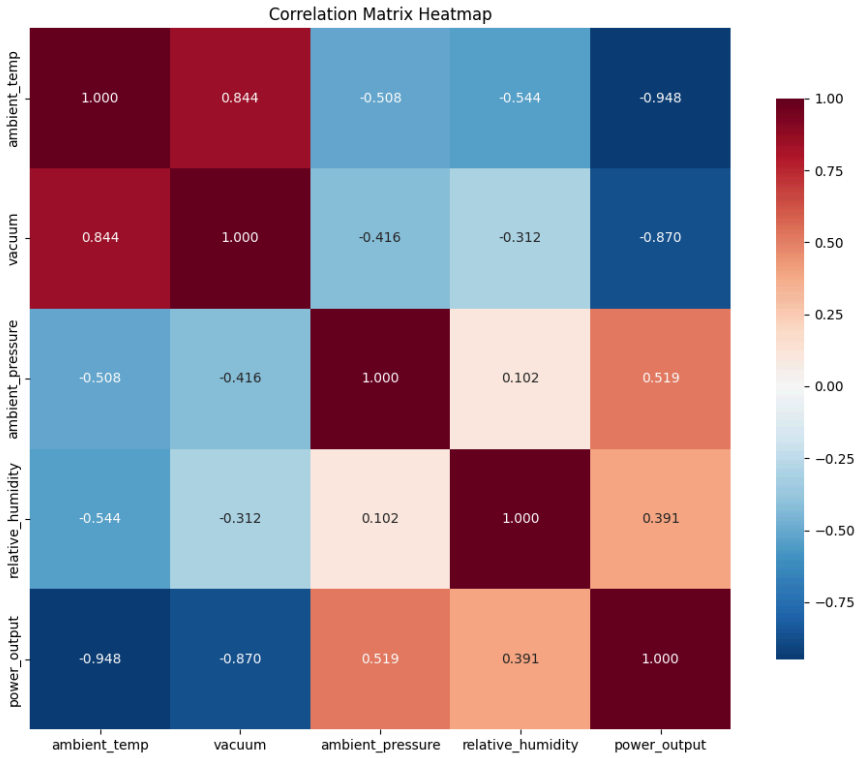

Figure 3: Correlation matrix.

Feature Selection

We tested every possible combination of features to use in an OLS model and found that including every feature in our model achieved the best test R^2 value of 0.9284. However, its maximium variance inflation factor was 5.89, which is above the threshold of 5, indicating that it's an untrustworthy combination of features.

Using only two features – ambient temperature and relative humidity – achieves an R^2 of 0.9204 and a maximum variance inflation factor of 1.41.

Number of Features	Features	R^2	VIF	Condition Number	Trustworthy?	All features significant?
4	ambient_temp, vacuum, ambient_pressure, relative_humidity	0.9284	5.89	4.8	False	True
3	ambient_temp, vacuum, relative_humidity	0.9281	4.88	4.3	True	True
3	ambient_temp, ambient_pressure, relative_humidity	0.9205	2.01	2.4	True	True
2	ambient_temp, relative_humidity	0.9204	1.41	1.8	True	True
3	ambient_temp, vacuum, ambient_pressure	0.9177	3.81	3.8	True	True

2	ambient_temp, vacuum	0.9154	3.41	3.4	True	True
2	ambient_temp, ambient_pressure	0.9001	1.35	1.7	True	True
1	ambient_temp	0.8981	1.00	1.0	True	True
3	vacuum, ambient_pressure, relative_humidity	0.8021	1.32	1.7	True	True
2	vacuum, ambient_pressure	0.7841	1.21	1.6	True	True
2	vacuum, relative_humidity	0.7692	1.10	1.4	True	True
1	vacuum	0.7530	1.00	1.0	True	True
2	ambient_pressure, relative_humidity	0.3848	1.01	1.1	True	True
1	ambient_pressure	0.2698	1.00	1.0	True	True
1	relative_humidity	0.1493	1.00	1.0	True	True