Duke University Machine Learning Foundations for Product Managers

Ruken Zilan

January 20, 2024

Project

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.

Data

The columns in the data consist of hourly average ambient variables:

- **Temperature** (T) in the range 1.81°C to 37.11°C,

- Ambient Pressure (AP) in the range 992.89-1033.30 milibar,

- **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)

AT	V	AP	RH	PE
1496	4176	102407	7317	46326
2518	6296	102004	5908	44437
511	394	101216	9214	48856
2086	5732	101024	7664	44648
1082	375	100923	9662	4739
1799	4372	100864	7504	45302
2014	4693	101466	6422	45399

Project:
Power Predictor

Goals

Priorities

- Based on the project topic, we are trying **to predict a continuous output variable (PE) from a set of input variables (AT, AP, RH, V).**
- This is a typical **regression problem**, which requires a machine learning approach that can **learn the relationship between the input and output variables and minimize the prediction error.**

Followed Steps

- Identify the Machine Learning Approach and Output Metric
- Feature Selection and Algorithm Identification
- Data Splitting
- Validation Strategy
- Model Building
- Model Evaluation
- Final Model Evaluation

Identify the Machine Learning Approach and Output Metric

- This is a **regression problem** since we're predicting the electrical energy output, which is a **continuous variable.**
- The **output metric** for regression tasks could be Mean Absolute Error (MAE), Mean Squared Error (MSE), or **Root Mean Squared Error (RMSE)**. I will use **RMSE.**

Feature Selection and Algorithm Identification

- Identified potential features: **Ambient Pressure (AP), Relative Humidity, Temperature (T), Exhaust Vacuum (V).**
- Considered Algorithm: **Supervised Learning & Multiple Linear Regression.**

Data Splitting

- I splited the dataset into **training** and **test** sets.
- I used **80-20** split considering the size of the data dataset as **training and test data sets.**

Validation Strategy

I compared the R2 and RMSE values of these data sets for both Models.

R2 indicates how well the independent variables explain the variability of the dependent variable. The R-squared value ranges from 0 to 1, where:

- **0 indicates** that the model does not explain any of the variability in the dependent variable.
- **1 indicates** that the model explains all of the variability in the dependent variable

RMSE shows the average difference between predicted value by the model and the real value.

Model Building

Trained **two different models** using my training set via **Excel Data Analysis Pack**.

This involves trying **different features** for the same algorithm.

- **Model 1** uses two features: **Used Temperature (AT) & Exhaust Vacuum (V).**
- **Model 2** uses four features: **Ambient Pressure (AP), Relative Humidity, Temperature (T), Exhaust Vacuum (V)**

Model Evaluation

- Evaluated the models on the **training and test** sets using the chosen metric.
- Selected the model that performs best on the validation set. This could be the model with the lowest MAE, MSE, or RMSE. I choose **RMSE**.

Model 1: Regression Results for Training Data (%80) for using Temperature and Vacume Data

AT	V	AP	RH	PE	*Predicted PE*
14,96	41,76	1024,07	73,17	463,26	466,4415371
25,18	62,96	1020,04	59,08	444,37	442,1307838
5,11	39,4	1012,16	92,14	488,56	484,0774479
20,86	57,32	1010,24	76,64	446,48	451,342037
10,82	37,5	1009,23	96,62	473,9	474,9018588
26,27	59,44	1012,23	58,77	443,67	441,3915747
15,89	43,96	1014,02	75,24	467,35	464,1424644
9,48	44,71	1019,12	66,43	478,42	474,8863236
14,64	45	1021,78	41,25	475,98	465,9510333
11,74	43,56	1015,14	70,72	477,5	471,3822634
17,99	43,72	1008,64	75,04	453,02	460,6208045
20,14	46,93	1014,66	64,22	453,99	455,9072657
24,34	73,5	1011,31	84,15	440,29	440,1907148
25,71	58,59	1012,77	61,83	451,28	442,6237462
26,19	69,34	1009,48	87,59	433,99	438,3543583
21,42	43,79	1015,76	43,08	462,19	454,7206257
18,21	45	1022,86	48,84	467,54	459,8333916
11,04	41,74	1022,6	77,51	477,2	473,1653603
14,45	52,75	1023,97	63,59	459,85	463,7916852
13,97	38,47	1015,15	55,28	464,3	469,1929218
17,76	42,42	1009,09	66,26	468,27	461,4317664

Model 1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,957222917
R Square	**0,916275713**
Adjusted R Square	0,916253827
Standard Error	4,944116676
Observations	7654

RMSE	**4,942825**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2046776,128	1023388,064	41866,14	0
Residual	7651	187023,2605	24,44428971		
Total	7653	2233799,389			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*
Intercept	505,4671705	0,267615825	1888,779073	0	504,9425702	505,9917709	504,9426	505,9918
AT	-1,713625117	0,014114884	-121,4055362	0	-1,74129416	-1,685956075	-1,74129	-1,68596
V	-0,320637013	0,008273164	-38,75627638	3,5E-300	-0,336854681	-0,304419345	-0,33685	-0,30442

Model 1: Regression Results for Test Data (%20) for using Temperature and Vacume Data

AP	RH	PE	*Predicted PE*
1016,88	71,44	444,38	440,2782844
1009,17	45,79	442,85	444,1329145
1008,98	44,32	432,33	424,3300137
1017,47	90,47	477,91	474,6268421
1011,68	70,33	434,99	439,4178367
1018	68,99	469,8	473,8837199
1006,26	59,15	426,66	434,7603379
1012,65	80,25	448,71	449,2233013
1013,65	41,54	463,35	458,529571
1016,6	97,09	435,58	479,6042135
1011,56	48,03	432,72	434,3689768
1012,2	47,06	441,32	442,5686895
1017,62	44,31	442,12	435,1457382
1014,13	69,67	445,16	449,1099711
1020,68	72,07	462,22	460,7773453
1010,86	54,03	445,03	434,0920428
1016,42	67,42	469,69	466,890761
1024,34	79,61	465,14	464,7438044
1011,63	65,97	435,98	431,2486454
1013,11	61,97	437,11	439,0786507

Model 1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,95575982
R Square	**0,913476834**
Adjusted R Square	0,913386233
Standard Error	4,999829368
Observations	1913

RMSE	**4,994591**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	504090,9	252045,4556	10082,51	0
Residual	1910	47746,74	24,99829371		
Total	1912	551837,7			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*
Intercept	505,5214059	0,548522	921,6065985	0	504,4456411	506,5971708	504,4456	506,5972
AT	-1,666290399	0,028849	-57,75951224	0	-1,72286879	-1,609712012	-1,72287	-1,60971
V	-0,340123217	0,016953	-20,06258653	2,13E-81	-0,37337177	-0,306874665	-0,37337	-0,30687

Model 2: Regression Results for Training Data (%80) for using all four parameters

AT	V	AP	RH	PE	*Predicted PE*
14,96	41,76	1024,07	73,17	463,26	467,2608222
25,18	62,96	1020,04	59,08	444,37	444,0988612
5,11	39,4	1012,16	92,14	488,56	483,6359385
20,86	57,32	1010,24	76,64	446,48	450,5600592
10,82	37,5	1009,23	96,62	473,9	471,792436
26,27	59,44	1012,23	58,77	443,67	442,2937569
15,89	43,96	1014,02	75,24	467,35	463,9542681
9,48	44,71	1019,12	66,43	478,42	478,2829264
14,64	45	1021,78	41,25	475,98	472,1454924
11,74	43,56	1015,14	70,72	477,5	473,1071284
17,99	43,72	1008,64	75,04	453,02	459,5247427
20,14	46,93	1014,66	64,22	453,99	456,6222959
24,34	73,5	1011,31	84,15	440,29	438,8124293
25,71	58,59	1012,77	61,83	451,28	443,1445106
26,19	69,34	1009,48	87,59	433,99	435,4066726
21,42	43,79	1015,76	43,08	462,19	458,2480421
18,21	45	1022,86	48,84	467,54	463,8820978
11,04	41,74	1022,6	77,51	477,2	474,2864076
14,45	52,75	1023,97	63,59	459,85	467,3120176
13,97	38,47	1015,15	55,28	464,3	472,30082
17,76	42,42	1009,09	66,26	468,27	461,7157718
5,41	40,07	1019,16	64,77	495,24	487,7136466

Model 2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,964198326
R Square	0,929678412
Adjusted R Square	0,929641642
Standard Error	4,532268542
Observations	7655

RMSE	4,530659

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	4	2077479,659	519369,9	25283,98	0
Residual	7650	157142,1548	20,54146		
Total	7654	2234621,813			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*
Intercept	454,9357467	10,81009621	42,08434	0	433,745	476,1265	433,745	476,1265
AT	-1,992042842	0,016966154	-117,413	0	-2,0253	-1,95878	-2,0253	-1,95878
V	-0,227178302	0,008086631	-28,0931	2,4E-165	-0,24303	-0,21133	-0,24303	-0,21133
AP	0,061871663	0,010487818	5,899384	3,8E-09	0,041313	0,082431	0,041313	0,082431
RH	-0,160556401	0,004622445	-34,7341	1,3E-245	-0,16962	-0,1515	-0,16962	-0,1515

Model 2: Regression Results for Test Data (%20) for using all four parameters

AT	V	AP	RH	PE	*Predicted PE*
23,86	74,93	1016,88	71,44	444,38	441,3758904
27,16	47,43	1009,17	45,79	442,85	445,5304381
33,97	72,29	1008,98	44,32	432,33	426,1833569
10,37	40,03	1017,47	90,47	477,91	473,5865489
25,15	71,14	1011,68	70,33	434,99	439,7273777
10,52	41,48	1018	68,99	469,8	476,1308656
28,88	66,56	1006,26	59,15	426,66	435,0794615
19,89	68,08	1012,65	80,25	448,71	449,2110455
19,31	43,56	1013,65	41,54	463,35	462,5110054
7,14	41,22	1016,6	97,09	435,58	478,4390046
28,18	71,14	1011,56	48,03	432,72	437,205147
26,08	57,32	1012,2	47,06	441,32	445,0241777
29,01	64,79	1017,62	44,31	442,12	438,2009395
21,84	58,86	1014,13	69,67	445,16	449,5338625
16,82	49,15	1020,68	72,07	462,22	461,7554454
29,03	67,79	1010,86	54,03	445,03	435,5171259
14,78	41,17	1016,42	67,42	469,69	468,1717398
15,55	43,71	1024,34	79,61	465,14	464,7263451
30,22	70,32	1011,63	65,97	435,98	430,8561656
25,97	68,12	1013,11	61,97	437,11	440,2686972
19,21	53,16	1013,18	81,64	454,01	454,236825
18,94	48,7	1007,82	92,88	453,36	453,9192798

Model 2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,96169601
R Square	**0,924859216**
Adjusted R Square	0,924701688
Standard Error	4,661807552
Observations	1913

RMSE	**4,6549092**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	510372,1383	127593	5871,083869	0
Residual	1908	41465,51393	21,73245		
Total	1912	551837,6523			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*Ipper 95,0%*
Intercept	453,3625423	22,5194027	20,13209	6,80893E-82	409,1973076	497,5277771	409,1973	497,5278
AT	-1,918531534	0,035213803	-54,4824	0	-1,987593129	-1,849469938	-1,98759	-1,84947
V	-0,26095625	0,0167297	-15,5984	1,01462E-51	-0,293766674	-0,228145827	-0,29377	-0,22815
AP	0,062848982	0,021849344	2,87647	0,004066167	0,019997871	0,105700093	0,019998	0,1057
RH	-0,147913081	0,009631665	-15,357	2,84476E-50	-0,166802781	-0,129023382	-0,1668	-0,12902

Final Model Evaluation

Used Set	R2 Value	RMSE Value
Model 1 Training	0,916	4,94
Model 1 Test	0,913	4,99
Model 2 Training	0,929	4,53
Model 2 Test	0,925	4,65

1. R2 Value:

- **Model 2 performs slightly better than Model 1** on both the training and test sets based on R2 values.
- **This suggests that Model 2 explains a larger proportion of the variance in the data compared to Model 1**.

2. RMSE Value:

- **Model 2 also has a lower RMSE value on both the training and test sets,** indicating **better predictive** accuracy compared to Model 1.
- Lower RMSE values generally signify **better model** performance.

Summary

Results

Model 1: R2 value: 0,916 , RMSE value: 4,94

Model 2: R2 value: 0,929 , RMSE value: 4,53

Insights Gained

Based on the provided metrics, **Model 2 appears to be a better-performing model** compared to Model 1.

It exhibits **higher R2 values and lower RMSE values on both the training and test sets**, suggesting **better overall predictive performance** and a better fit to the data.

Furher Improvements

Enhancing the model's robustness can be achieved through additional data cleaning and testing.