Machine Learning and Big Data Analytics

Project for Taiwan’s ETC (Electronic Toll Collection)

Energy Consumption and Electricity Cost Estimation

Project Overview

This project aims to build a predictive model to **estimate daily electricity consumption and the cost of Taiwan's ETC (Electronic Toll Collection) system,** using temperature and infrastructure data. The model is trained on real-world datasets provided by the instructor, including two-month electricity billing records, minute-level voltage and current sensor data, and daily temperature readings from 19 meteorological stations across Taiwan.

Dataset

- Data provided by instructor

	A	B	C	Q	R
1	點位編號	緯度	經度	計費期間(起)	計費期間(訖)
2	Location ID	Latitude	Longitude	Billing Period (Start)	Billing Period (End)
3	01F0233N	25.07301944	121.5307028	1130429	1130626
4	01F0233N	25.07301944	121.5307028	1130627	1130828
5	01F0233N	25.07301944	121.5307028	1130829	1131029
6	01F0233N	25.07301944	121.5307028	1131030	1131226
7	01F0256N	25.07806667	121.5091056	1130227	1130502
8	01F0256N	25.07806667	121.5091056	1130503	1130701
9	01F0256N	25.07806667	121.5091056	1130702	1130902
10	01F0256N	25.07806667	121.5091056	1130903	1131103
11	01F0256N	25.07806667	121.5091056	1131104	1131229
12	01H0206S	25.07275	121.557739	1130425	1130624
13	01H0206S	25.07275	121.557739	1130625	1130826
14	01H0206S	25.07275	121.557739	1130827	1131027
15	01H0206S	25.07275	121.557739	1131028	1131224
16	01F0213N	25.073053	121.550358	1130425	1130624
17	01F0213N	25.073053	121.550358	1130625	1130826
18	01F0213N	25.073053	121.550358	1130827	1131027
19	01F0213N	25.073053	121.550358	1131028	1131224
20	01H0200N	25.070994	121.562519	1130418	1130618
21	01H0200N	25.070994	121.562519	1130619	1130819
22	01H0200N	25.070994	121.562519	1130820	1131020
23	01H0200N	25.070994	121.562519	1131021	1131218
24	01H0208N	25.072261	121.555014	1130418	1130618
25	01H0208N	25.072261	121.555014	1130619	1130819
26	01H0208N	25.072261	121.555014	1130820	1131020
27	01H0208N	25.072261	121.555014	1131021	1131218

myebill-phase 1

項目：平均氣溫(°c)Temperature
時間：2024年
緯度：22°39'09.20" N經度：120°31'08.90" E

Day\Mon	1	2	3	4	5	6	7	8	9	10	11	12	Day\Month
1	21.8	24.4	21	28.9	28.2	31.4	31.9	30.8	29.5	27.4	27.9	22.1	1
2	21.4	24.4	18.5	28.8	29	28.2	32.1	31.6	31	25.3	28.1	22.9	2
3	20.3	24.2	21.3	29.5	29	27.8	32.1	30.2	30.9	24.8	28.4	23.7	3
4	18.9	25	23.7	28.9	29.5	27.3	31.7	31.3	30.8	27.9	28.1	24.7	4
5	20	25.7	27.6	29	30	27.6	31.2	29.8	29.8	28.3	27.7	24.1	5
6	20.4	23.6	26.8	28.6	30.2	28.6	30.6	27.9	29.2	27.7	26.2	24.7	6
7	20.9	23.4	21.9	28.9	30.3	28.8	29.2	29.6	29.1	28.3	25.7	24.3	7
8	20.7	21.9	20.6	29.3	29.3	30	28.2	29.9	27.8	28.6	25.8	22.1	8
9	22	19.4	21.1	27.7	28.4	30.4	29.7	28.4	28.3	27.6	27.7	22.2	9
10	20.2	19	21	27.3	28.3	30	31.5	27.2	28.6	27.1	27.8	22.7	10
11	18	20.2	22.7	27.8	29.2	29.3	31.9	29.8	29.3	26.7	27.7	23.8	11
12	19	19	22.3	28.6	30.1	28.9	32.5	29.6	29.8	27.5	26.4	22.9	12
13	20	19.8	22.8	28.6	29.4	28.6	31.6	30.3	28.6	28.3	28	22.4	13
14	20.5	22.8	22.3	29.3	26.9	30.3	31.5	29.2	27.2	28.9	27.6	19.8	14
15	20.7	23.8	24.2	29.3	28.5	31	29.6	28	29.5	28.9	26.1	18.9	15
16	20.2	23.1	25.3	28.8	28.4	31	31.3	28.3	29.9	29.1	25.2	19.2	16
17	20.7	22.5	26.1	29	28.3	31.3	31.5	27.9	31.9	28.8	28	21.1	17
18	21.9	24.3	25	29.2	29	31.7	31.3	28.5	30.1	29.7	25.6	20.7	18
19	23.8	25.6	24.3	29.6	29.3	31.7	32	27.1	30.7	29.3	25.4	19.1	19
20	23.5	26.8	23.1	29.5	28.6	31.4	31.7	29.6	29.3	29.6	25.5	19.2	20
21	21.9	26.4	23.1	29.8	29.4	32	29.6	30.5	28.8	29.2	25.8	21.5	21
22	18.9	25.8	24	29.9	30.1	32.3	31.2	31.1	26.3	29.3	25.2	19	22
23	13.5	25.2	26.9	29.6	30.4	32.4	30.4	31.5	29.3	28.2	25.6	17.5	23
24	15.2	24.5	28	27.8	29.1	31.7	28.2	29.7	30.8	27.5	25.8	18.8	24
25	17.2	23.8	28.1	24.9	30.3	32.4	27.8	29.1	30.5	27.6	24	22.3	25
26	17.8	21.6	28.1	25.3	28.5	32.5	29.3	29	30.8	28.6	24.9	23.3	26
27	18.8	21.1	27.5	26.1	28.6	31.2	29.6	30.8	30.7	28	21.7	20.8	27
28	19.3	23.5	27.2	27.4	26.9	32.1	30.2	32	30.2	26.4	21.3	18.9	28
29	20.4	24.4	27.2	29.7	29.4	30.2	28.1	31.8	30.1	27.3	20.9	18.3	29
30	22.2		27.7	28.9	30	30.9	29.8	31.2	28.1	27.1	21.3	21.9	30
31	23.5		28.9		31.2		29.8	29.6		25.8		21.7	31
平均	20.1	23.3	24.5	28.5	29.2	30.4	30.6	29.7	29.6	27.9	25.9	21.4	平均

1	toll_booth	gantry	data_time	ac_a_curre	ac_b_curre	rs485_tem	mains_pov	mains_pov	mains_pov	smr_dc_cu	smr_dc_voltage
2	泰山	01F-000.5N	6/1/2024 0:00	0.1	0.2	16.4	5.6	79.3	155.3	36	26.7
3	泰山	01F-000.5S	6/1/2024 0:00	0.1	2	22.1	7.1	115	122.4	35.3	26.7
4	泰山	01F-001.7S	6/1/2024 0:00	0.1	1	19.6	8.8	115.6	115.7	55.5	26.8

Data Preprocessing

- We transformed the data from each meteorological weather station from a matrix format (with days and months on both the x and y axes, respectively) into a columnar format containing the date, temperature, station name, station ID, latitude, and longitude.
- The raw data from ETC station (server's shelter + gantry structure) was merged with the billing dataset by taking a unique daily consumption for each gantry (initially stated in an hourly manner). Finally, the cleaned temperature dataset containing all 19 stations was also introduced by matching each gantry (with now all its essential information gathered) to the nearest meteorological station by using the latitude and longitude. Our final training dataset ended up as the following:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	gantry_nc	date	ac_a_curr	ac_b_curr	rs485_ten	mains_po	mains_po	mains_po	smr_dc_c	smr_dc_v	latitude	longitude	equipmer	lane coun	stationid	stationna	station_la	station_lc	temperati	distance_	avg_ac_vc	ac_power	dc_power	total_ene	ac_energy	dc_energy	total_energy_kwh	
2	01F0005N	6/1/2024	0.1	1.666667	16.07083	6.666667	79.975	157.0667	35.34167	26.75833	25.11831	121.7316	3	2	466940	keelung	25.13513	121.7323	26.1	1.871	237.0417	418.7736	945.6841	1580.278	0.418774	0.945684	1.580278	
3	01F0005N	6/1/2024	0.1	1.666667	16.07083	6.666667	79.975	157.0667	35.34167	26.75833	25.11831	121.7316	3	2	466940	keelung	25.13513	121.7323	26.1	1.871	237.0417	418.7736	945.6841	1580.278	0.418774	0.945684	1.580278	
4	01F0005N	6/1/2024	0.1	1.666667	16.07083	6.666667	79.975	157.0667	35.34167	26.75833	25.11831	121.7316	3	2	466940	keelung	25.13513	121.7323	26.1	1.871	237.0417	418.7736	945.6841	1580.278	0.418774	0.945684	1.580278	
5	01F0005N	6/1/2024	0.1	1.666667	16.07083	6.666667	79.975	157.0667	35.34167	26.75833	25.11831	121.7316	3	2	466940	keelung	25.13513	121.7323	26.1	1.871	237.0417	418.7736	945.6841	1580.278	0.418774	0.945684	1.580278	
6	01F0005S	6/1/2024	0.1	1.766667	20.80833	6.641667	116.0083	123.4083	34.49583	26.74583	25.11879	121.7318	3	2	466940	keelung	25.13513	121.7323	26.1	1.818	239.4167	446.9111	922.6198	1590.126	0.446911	0.92262	1.590126	
7	01F0005S	6/1/2024	0.1	1.766667	20.80833	6.641667	116.0083	123.4083	34.49583	26.74583	25.11879	121.7318	3	2	466940	keelung	25.13513	121.7323	26.1	1.818	239.4167	446.9111	922.6198	1590.126	0.446911	0.92262	1.590126	
8	01F0005S	6/1/2024	0.1	1.766667	20.80833	6.641667	116.0083	123.4083	34.49583	26.74583	25.11879	121.7318	3	2	466940	keelung	25.13513	121.7323	26.1	1.818	239.4167	446.9111	922.6198	1590.126	0.446911	0.92262	1.590126	
9	01F0005S	6/1/2024	0.1	1.766667	20.80833	6.641667	116.0083	123.4083	34.49583	26.74583	25.11879	121.7318	3	2	466940	keelung	25.13513	121.7323	26.1	1.818	239.4167	446.9111	922.6198	1590.126	0.446911	0.92262	1.590126	
10	01F0017S	6/1/2024	0.1	1.529167	22.34167	9.304167	116.7667	117.225	55.42917	26.77083	25.10957	121.7259	3	2	466940	keelung	25.13513	121.7323	26.1	2.914	233.9917	381.2114	1483.885	2177.097	0.381211	1.483885	2.177097	

Model (1-2)

• We used CatBoost Regressor, a decision tree-based machine learning model. It handles both numeric and categorical features well, making it suitable for our dataset. It also performs strongly in non-linear and complex real-world scenarios.

- **<u>Key model settings:</u>**
- Tuned for non-linear daily patterns
- Regularization to prevent overfitting
- Learning optimized progressively over 1000 iterations
- Only 2024 data used to preserve future test integrity

Model (2-2)

How It Works

- Learns from daily energy usage patterns.
- Builds many small trees, each improving the last.
- Focuses on reducing error step by step.
- Produces a final prediction by combining trees.

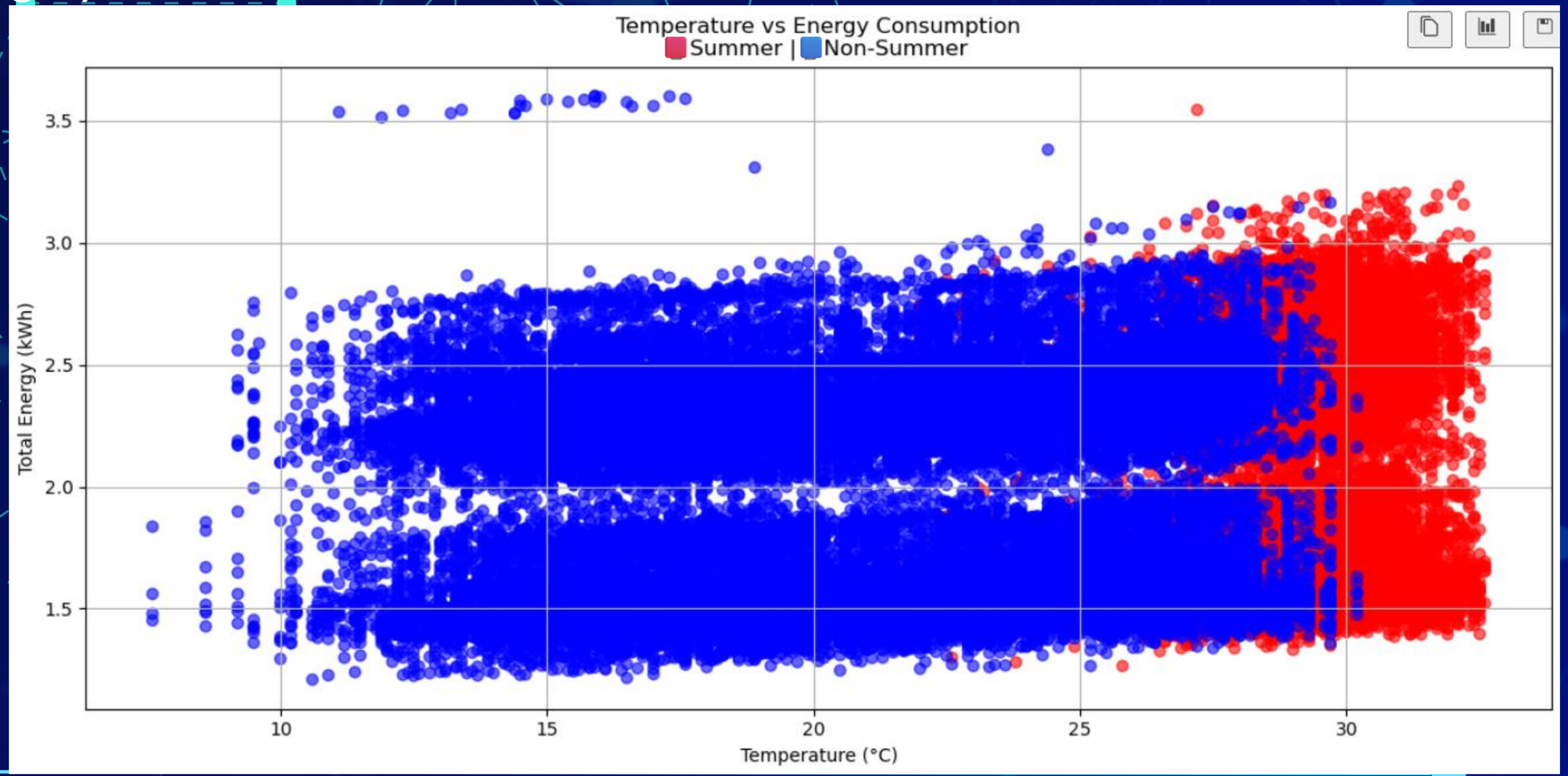
Features

Feature	Why
Temperature	Energy
Lane Count	More la
Station ID	Each st
is_summer	Season
Temperature × Lane Count	Interac

What the model shows (1-3)

- **Temperature Signal**

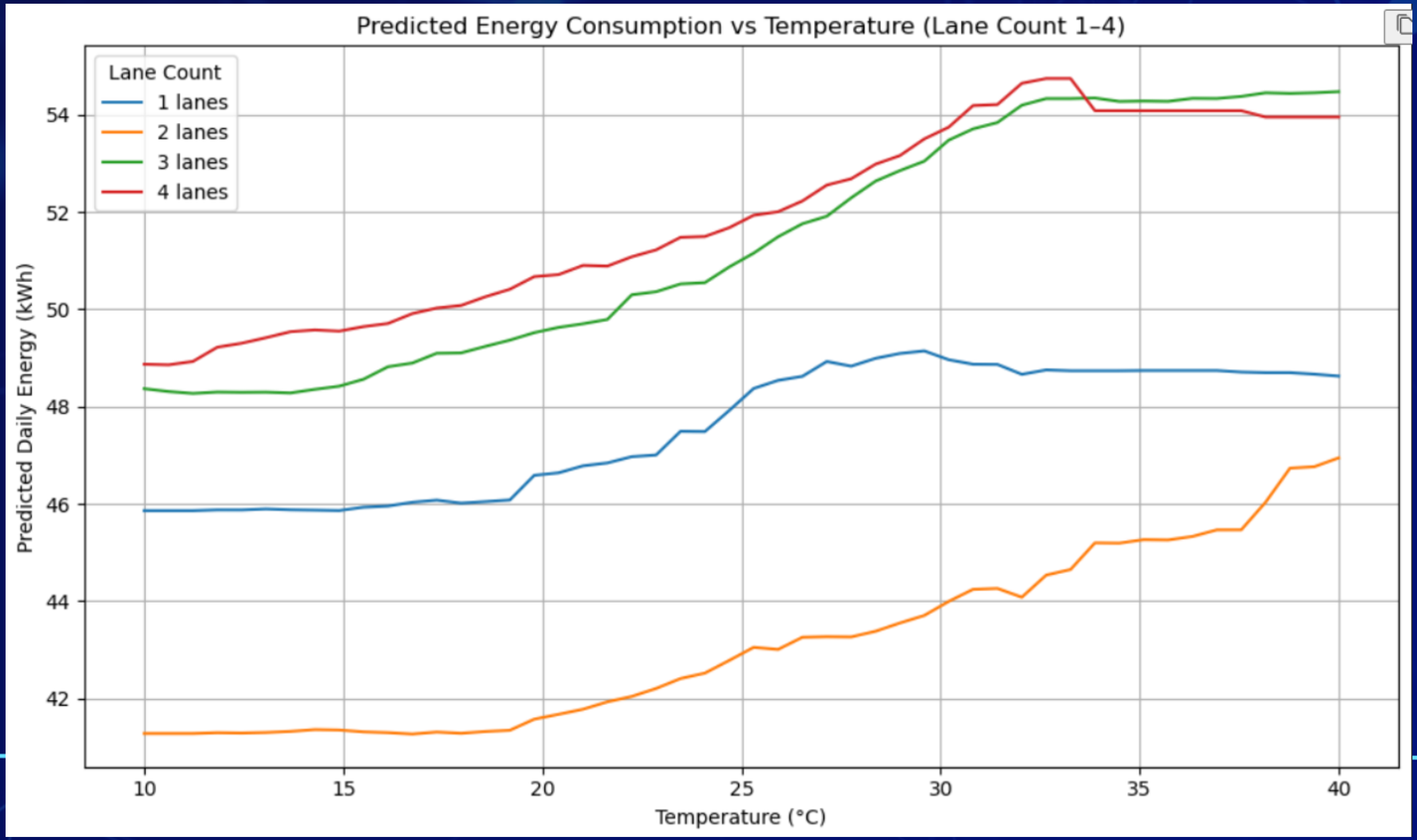
As the temperature rises, energy consumption increases, especially during summer. This aligns with our expectations, since hotter weather drives up demand for cooling systems at toll stations.



What the model shows (2-3)

- **Lane Count Effect**

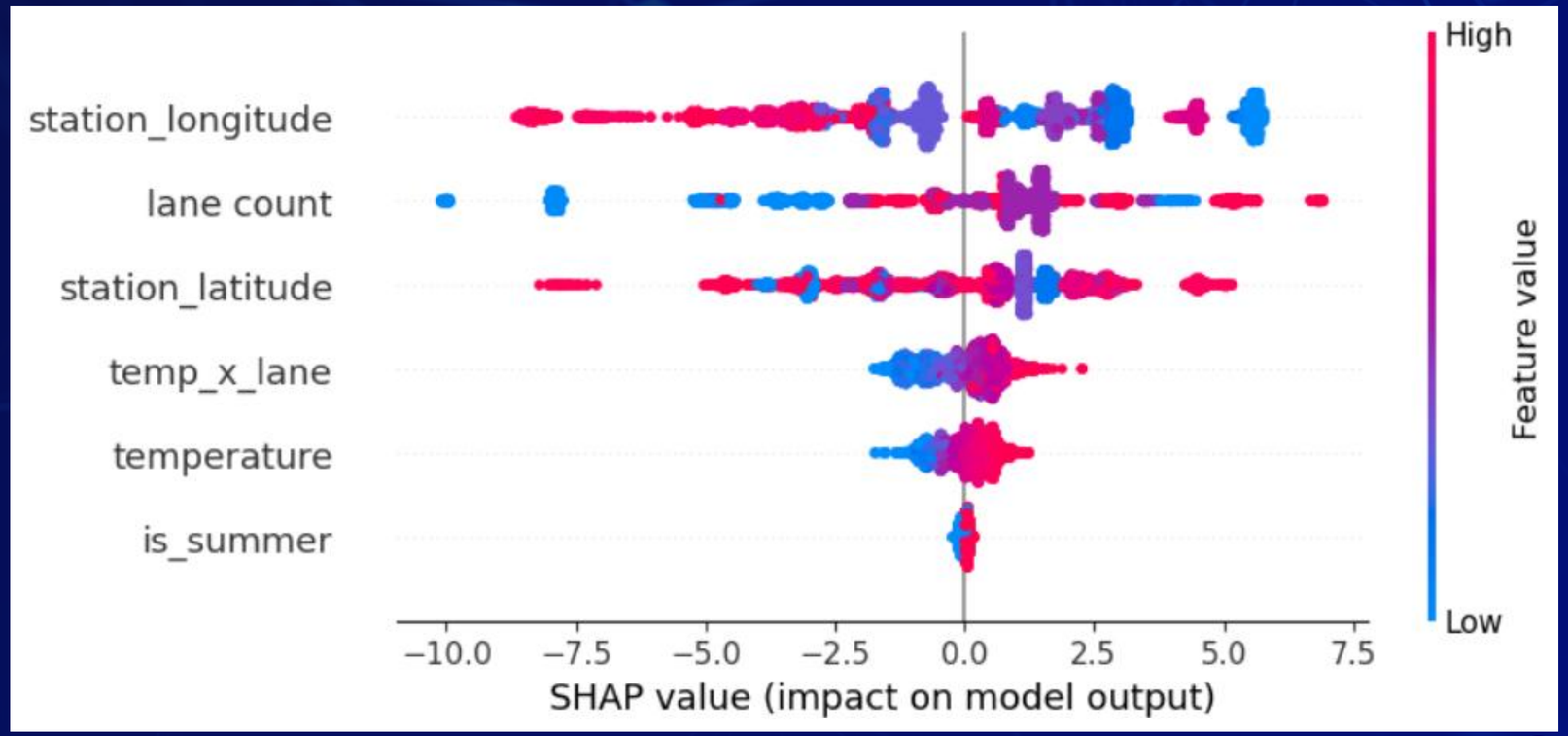
This graph illustrates the average relationship between lane count and energy consumption across all gantries. It reveals that energy consumption tends to increase as the number of lanes increases.



What the model shows (3-3)

- **Combined Effect**

When both temperature and lane count are high, energy tend to increase. We can observe the importance of every feature for the model by the SHAP plotting



Evaluation – Metrics Used

We used four key metrics to assess forecasting performance:

- **R^2 (Explained Variance)**

Indicates how well the model captures overall consumption patterns.
Higher is better; a value near 1 indicates an excellent fit.

- **RMSE (Root Mean Squared Error)**

Measures prediction error magnitude, penalizing significant mistakes.
Essential for planning, especially with high-usage spikes.

- **MAE (Mean Absolute Error)**

Reports average prediction error in kWh.
Straightforward and easy to communicate.

- **Accuracy% (vs. Real Billing)**

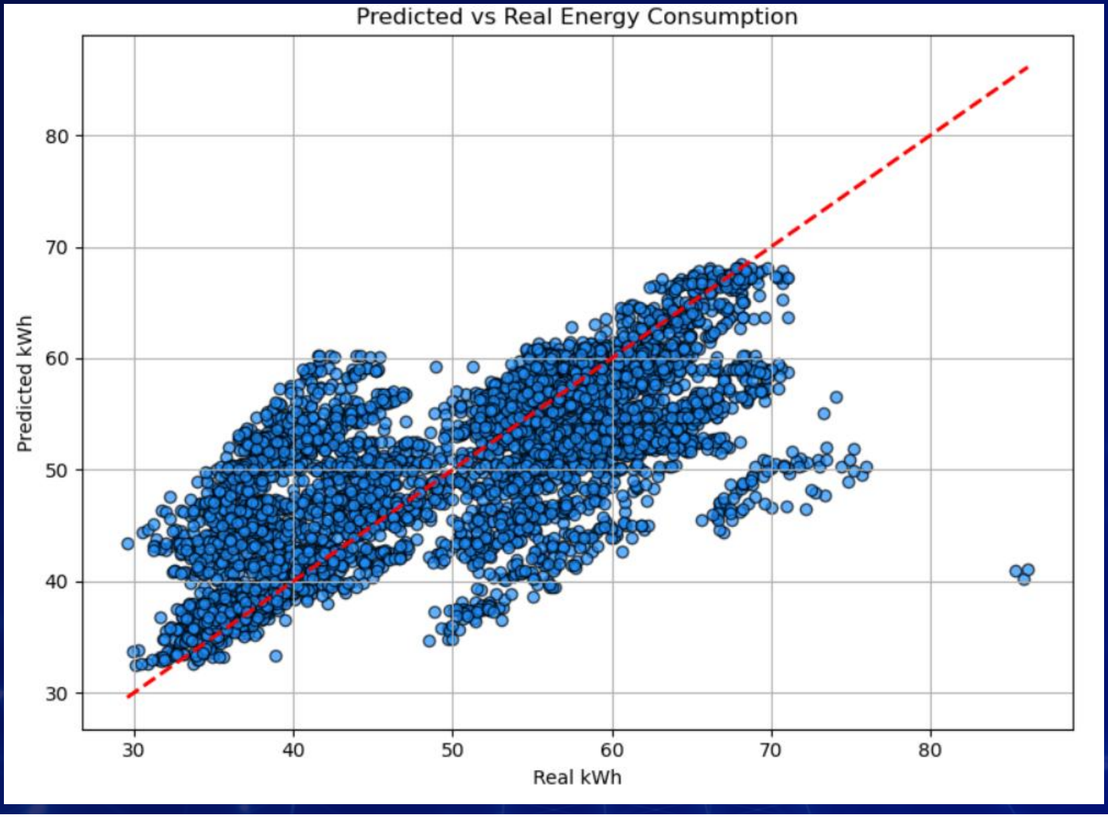
Shows how close predictions are to actual billed values.
Useful for business interpretation

```
Model Evaluation on Test Set:
- R² Score: 0.5841
- RMSE: 6.59 kWh
- MAE: 4.61 kWh
```

Predictions

We tested the model on real billing periods from late 2024 to early 2025. It achieves high accuracy, often above 80%, with small prediction errors, showing its reliability for energy forecasting.

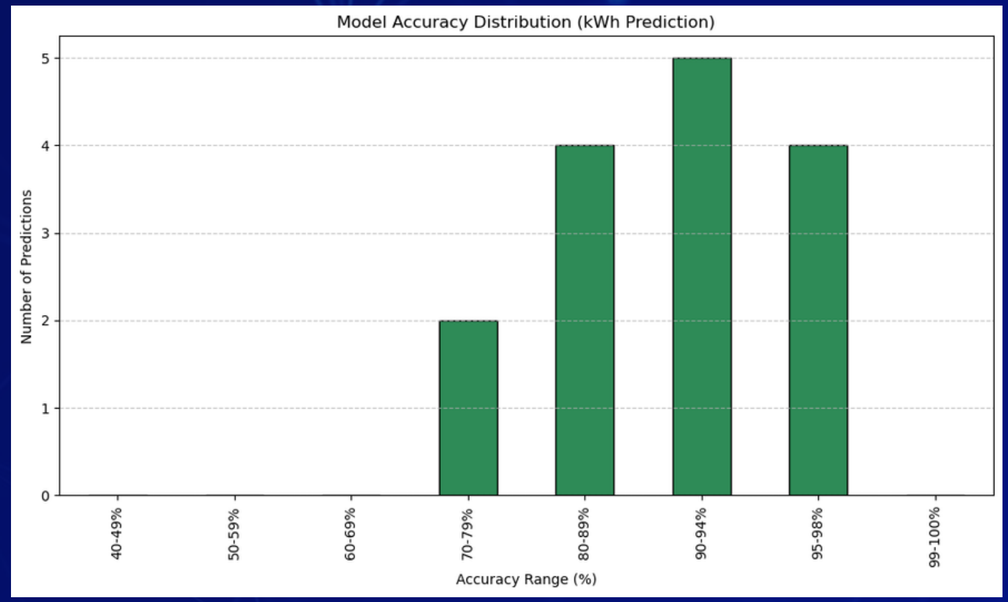
Scatter plot




```
Predictions for toll stations:
- 01F0233N: 2024-04-29 → 2024-06-26 | Pred: 2888.18 kWh | Real: 2453 kWh | Accuracy: 82.26% | Pred NT$: 9794.90 vs Real NT$: 7795.00 | Cost Accuracy: 74.34%
- 01F0233N: 2024-06-27 → 2024-08-28 | Pred: 3154.33 kWh | Real: 2672 kWh | Accuracy: 81.95% | Pred NT$: 11988.39 vs Real NT$: 9611.00 | Cost Accuracy: 75.26%
- 01F0233N: 2024-08-29 → 2024-10-29 | Pred: 3056.95 kWh | Real: 2605 kWh | Accuracy: 82.65% | Pred NT$: 10585.76 vs Real NT$: 8527.00 | Cost Accuracy: 75.86%
- 01F0233N: 2024-10-30 → 2024-12-26 | Pred: 2739.59 kWh | Real: 2294 kWh | Accuracy: 80.58% | Pred NT$: 8356.98 vs Real NT$: 6549.00 | Cost Accuracy: 72.39%
- 01F0256N: 2024-02-27 → 2024-05-02 | Pred: 3164.98 kWh | Real: 2645 kWh | Accuracy: 80.34% | Pred NT$: 9687.49 vs Real NT$: 7600.00 | Cost Accuracy: 72.53%
- 01F0256N: 2024-05-03 → 2024-07-01 | Pred: 2946.91 kWh | Real: 2486 kWh | Accuracy: 81.46% | Pred NT$: 10282.34 vs Real NT$: 8043.00 | Cost Accuracy: 72.16%
- 01F0256N: 2024-07-02 → 2024-09-02 | Pred: 3155.81 kWh | Real: 2651 kWh | Accuracy: 80.96% | Pred NT$: 12245.99 vs Real NT$: 9524.00 | Cost Accuracy: 71.42%
- 01F0256N: 2024-09-03 → 2024-11-03 | Pred: 3042.63 kWh | Real: 2558 kWh | Accuracy: 81.05% | Pred NT$: 10374.61 vs Real NT$: 8211.00 | Cost Accuracy: 73.65%
- 01F0256N: 2024-11-04 → 2024-12-29 | Pred: 2643.73 kWh | Real: 2215 kWh | Accuracy: 80.64% | Pred NT$: 8148.68 vs Real NT$: 6277.00 | Cost Accuracy: 70.18%
```

Results to validate our model
Predictions from Nov-2024 to April-2025

Gantry	Start Date	End Date	Real kWh	Predicted kWh	Error kWh	Accuracy (%)	Real Payable	Predicted Payable	Cost Accuracy (%)
01F0532S	12/13/2024	2/11/2025	3598	3960.17	362.17	89.93	12143	13260.15	90.8
01F0699S	12/11/2024	2/9/2025	2795	3090.35	295.35	89.43	8271	9422.87	86.07
01F0339S	11/26/2024	1/20/2025	2400	2613.41	213.41	91.11	6913	7704.89	88.54
01F1699S	11/5/2024	1/1/2025	3181	2922.4	-258.6	91.87	9934	9167.46	92.28
01F1699S	1/2/2025	3/3/2025	3470	3201.52	-268.48	92.26	11465	10082.45	87.94
01F2603N	11/11/2024	1/7/2025	3048	3141.79	93.79	96.92	9230	9746.91	94.4
01F2603N	1/8/2025	3/9/2025	2752	3405.92	653.92	76.24	8124	10446.63	71.41
01F3286N	11/7/2024	1/5/2025	3318	3426.32	108.32	96.74	15015	11041.16	73.53
01F3286N	1/6/2025	3/5/2025	3420	3462.82	42.82	98.75	15660	11016.17	70.35
01F0339S	1/21/2025	3/24/2025	2705	2997.37	292.37	89.19	7962	8835.48	89.03
01F1699S	1/2/2025	3/3/2025	3470	3201.52	-268.48	92.26	11465	10082.45	87.94
01F2603N	1/8/2025	3/9/2025	2752	3405.92	653.92	76.24	8124	10446.63	71.41
01F3286N	1/6/2025	3/5/2025	3420	3462.82	42.82	98.75	15660	11016.17	70.35
01F0532S	2/12/2025	4/13/2025	3644	4017.96	373.96	89.74	12386	13468.77	91.26
01F0699S	2/10/2025	4/9/2025	2823	3067.56	244.56	91.34	8368	9340.59	88.38



Conclusion

Although we are proud of what we've achieved, we also recognize that there is always room for improvement. We welcome any suggestions or comments that will help us refine our approach. Thank you for your attention, and we look forward to answering your questions.