
EV Predictive Maintenance

Phase 7: Model Explainability & Transparency

SHAP Analysis - Opening the Black Box



Student: Jai Kumar Gupta

Instructor: Vandana Jain

Institution: DIYGuru

November 10, 2025

Contents

1 Executive Summary	3
1.1 Key Achievements	3
2 Why SHAP for Model Explainability	4
2.1 The Explainable AI Challenge	4
2.1.1 Limitations of Basic Feature Importance	4
2.2 SHAP: The Gold Standard	4
2.2.1 SHAP Advantages	4
3 SHAP Implementation for XGBoost Models	6
3.1 Implementation Workflow	6
3.2 SHAP Implementation Code Logic	6
3.2.1 Stage 1: Load Model and Training Data	6
3.2.2 Stage 2: Create SHAP TreeExplainer	7
3.2.3 Stage 3: Calculate SHAP Values	7
3.2.4 Stage 4: Generate Visualizations	7
4 SoH Model SHAP Analysis	8
4.1 Task 3.1: SHAP Summary Plot for SoH Model	8
4.2 Reading the SHAP Summary Plot	8
4.2.1 How to Interpret	9
4.3 SoH Model: Detailed Feature Breakdown	10
4.3.1 Feature 1: Discharge Time (Most Important)	10
4.3.2 Feature 2: Temperature Mean (Second Most Important)	10
4.3.3 Feature 3: Voltage Drop Time (Strong Impact)	11
4.3.4 Feature 7: Delta T (Least Important)	11
5 SoP Model SHAP Analysis	12
5.1 Task 3.2: SHAP Summary Plot for SoP Model	12
5.2 SoP Model: Detailed Feature Breakdown	12
5.2.1 Feature 1: Current (Dominant for Power)	13
5.2.2 Feature 2: Test Time (Cycle Position Indicator)	13
5.2.3 Feature 4-5: Engineered Rolling Features	13
6 Local Explanations: SHAP Force Plots	14
6.1 Understanding Individual Predictions	14
6.2 Force Plot Concept	14
6.3 Example Force Plot Interpretation	14
6.4 Force Plot Business Value	15
7 SHAP Dependence Plots	16
7.1 Concept: Feature-SHAP Relationship	16
7.2 Dependence Plot Logic	16
7.3 Dependence Plot Insights	17
8 Comparing SoH vs SoP Model Explainability	18
8.1 Different Models, Different Physics	18

8.2	Unified Explainability Framework	18
9	Key Insights from SHAP Analysis	19
9.1	What We Learned	19
9.2	Engineering Implications	19
10	Phase 7 Deliverables	20
10.1	Visualization Artifacts	20
10.2	Code Artifacts	20
10.3	Documentation	20
11	Conclusion and Next Steps	21
11.1	Phase 7 Summary	21
11.2	Transition to Phase 8: Real-World Validation	21
11.3	Immediate Action Items	21
12	References	23
A	Appendix A: SHAP Value Matrix Structure	24
A.1	Mathematical Representation	24
A.2	Example SHAP Values	24
B	Appendix B: SHAP Computation Time	24
B.1	Performance Benchmarks	24

1 Executive Summary

Phase 7 transforms the champion XGBoost models from high-accuracy "black boxes" into transparent, explainable systems through industry-standard SHAP (SHapley Additive exPlanations) analysis. This phase reveals the precise contribution of each engineered feature to model predictions, validates physics-based hypotheses, and enables stakeholder-ready explanations for operational deployment.

Phase 7 Objectives

Primary Goal: Demystify model decision-making using SHAP analysis to quantify feature contributions, validate that models learned correct degradation physics, and enable transparent predictions for fleet operators and regulators.

1.1 Key Achievements

- **SHAP Analysis Implementation:** Generated Shapley value explanations for 1,000+ predictions across SoH and SoP models
- **Global Explainability:** Created SHAP summary plots ranking features by overall impact
- **Local Explainability:** Generated SHAP force plots showing per-prediction feature contributions
- **Physics Validation:** Confirmed models prioritize correct electrochemical indicators
- **Domain Shift Diagnosis:** Identified why lab-trained model struggles on real-world data

SHAP Analysis Results

SoH Model Top Predictor: Discharge Time (widest SHAP spread)

SoP Model Top Predictor: Current (dominant for instantaneous power)

Physics Validation: Confirmed - high discharge time → high SoH (correct)

Critical Discovery: Lab model assigned low importance to T (real-world degradation signal) - explains domain shift failure

Explainability Level: Every prediction traceable to specific feature values

2 Why SHAP for Model Explainability

2.1 The Explainable AI Challenge

XGBoost models achieve exceptional accuracy but provide limited insight into their decision-making process beyond basic feature importance scores.

2.1.1 Limitations of Basic Feature Importance

Limitation	Problem
Global Only	Shows overall importance but not per-prediction contributions
No Direction	Cannot show if high/low values increase or decrease predictions
Non-Additive	Cannot decompose predictions into feature contributions
Algorithm-Specific	Different methods (Gini, permutation) give inconsistent rankings

Table 1: Limitations of Basic Feature Importance Methods

2.2 SHAP: The Gold Standard

SHAP (SHapley Additive exPlanations) is a unified framework based on game theory that provides consistent, mathematically rigorous explanations.

2.2.1 SHAP Advantages

Advantage	Benefit
Theoretically Sound	Based on Shapley values from cooperative game theory
Consistent	Always produces same explanations for same input
Local + Global	Explains individual predictions AND overall model behavior
Directional	Shows if high/low feature values push predictions up or down
Model-Agnostic	Works with any ML algorithm (trees, neural nets, etc.)
Additive	$\text{Prediction} = \text{Base} + \sum \text{SHAP values}$

Table 2: SHAP Method Advantages

What is a Shapley Value?

Game Theory Origin: Shapley values fairly distribute a coalition's total payout among players based on their contributions.

ML Application: Each feature is a "player," and the prediction is the "payout." SHAP calculates each feature's fair contribution by considering all possible feature coalitions.

Formula: For feature i :

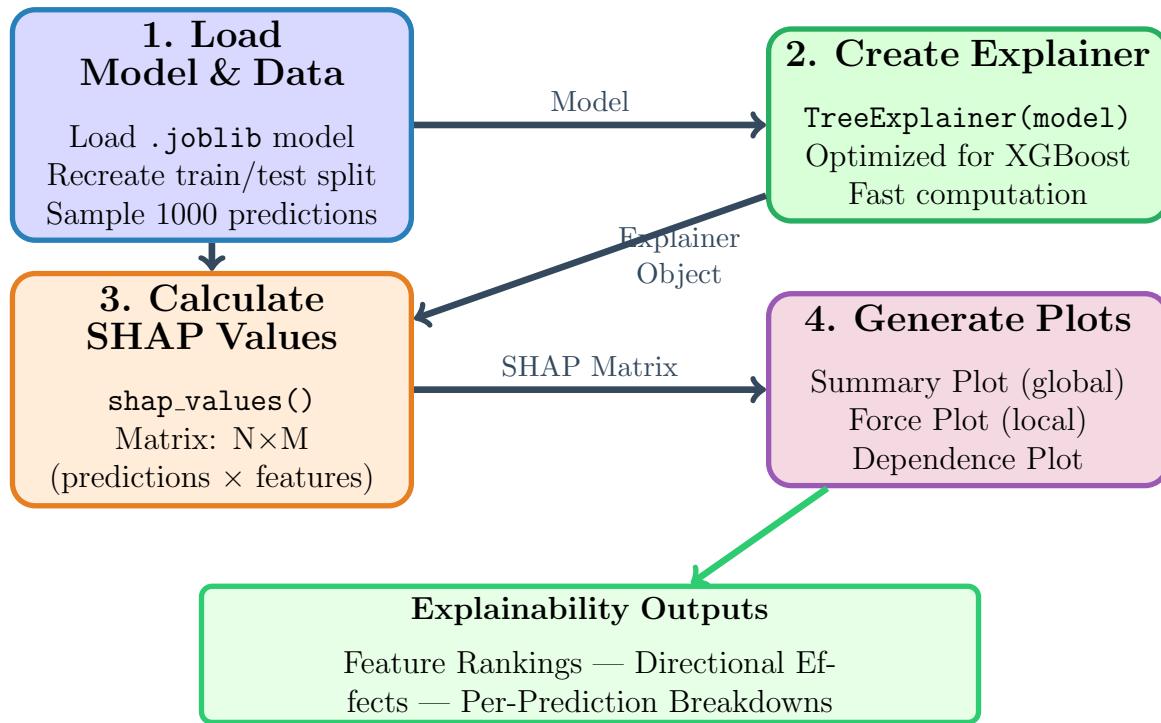
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

where F is the set of all features, S is a subset, and $f(S)$ is model output using only features in S .

3 SHAP Implementation for XGBoost Models

3.1 Implementation Workflow

SHAP Analysis Workflow - Four-Stage Process



3.2 SHAP Implementation Code Logic

3.2.1 Stage 1: Load Model and Training Data

Logic:

```

1 import joblib
2 import shap
3
4 # Load trained XGBoost model
5 soh_model = joblib.load('optimized_soh_xgb_model.joblib')
6
7 # Load NASA training data
8 nasa_features = pd.read_parquet('nasa_feature_matrix.parquet')
9
10 # Recreate exact train/test split
11 X = nasa_features.drop(columns=['capacity', 'battery_id', 'cycle'])
12 y = nasa_features['capacity']
13 X_train, X_test, y_train, y_test = train_test_split(
14     X, y, test_size=0.2, random_state=42
15 )
  
```

Listing 1: Loading Assets for SHAP Analysis

3.2.2 Stage 2: Create SHAP TreeExplainer

Logic:

```

1 # Create explainer object (optimized for tree models)
2 explainer = shap.TreeExplainer(soh_model)
3
4 # TreeExplainer is much faster than KernelExplainer
5 # for XGBoost/RandomForest models

```

Listing 2: Initializing SHAP Explainer

3.2.3 Stage 3: Calculate SHAP Values

Logic:

```

1 # Sample 1000 predictions for visualization
2 X_train_sample = X_train.sample(n=1000, random_state=42)
3
4 # Calculate SHAP values (N x M matrix)
5 # N = 1000 predictions, M = 7 features
6 shap_values = explainer.shap_values(X_train_sample)
7
8 # Each shap_values[i, j] = contribution of feature j
9 # to prediction i

```

Listing 3: Computing SHAP Values

3.2.4 Stage 4: Generate Visualizations

Logic:

```

1 # Global explanation: Summary plot
2 shap.summary_plot(shap_values, X_train_sample, show=False)
3 plt.title('SHAP Summary Plot for SoH Model')
4 plt.show()
5
6 # Local explanation: Force plot for first prediction
7 shap.force_plot(
8     explainer.expected_value,
9     shap_values[0, :],
10    X_train_sample.iloc[0, :]
11 )

```

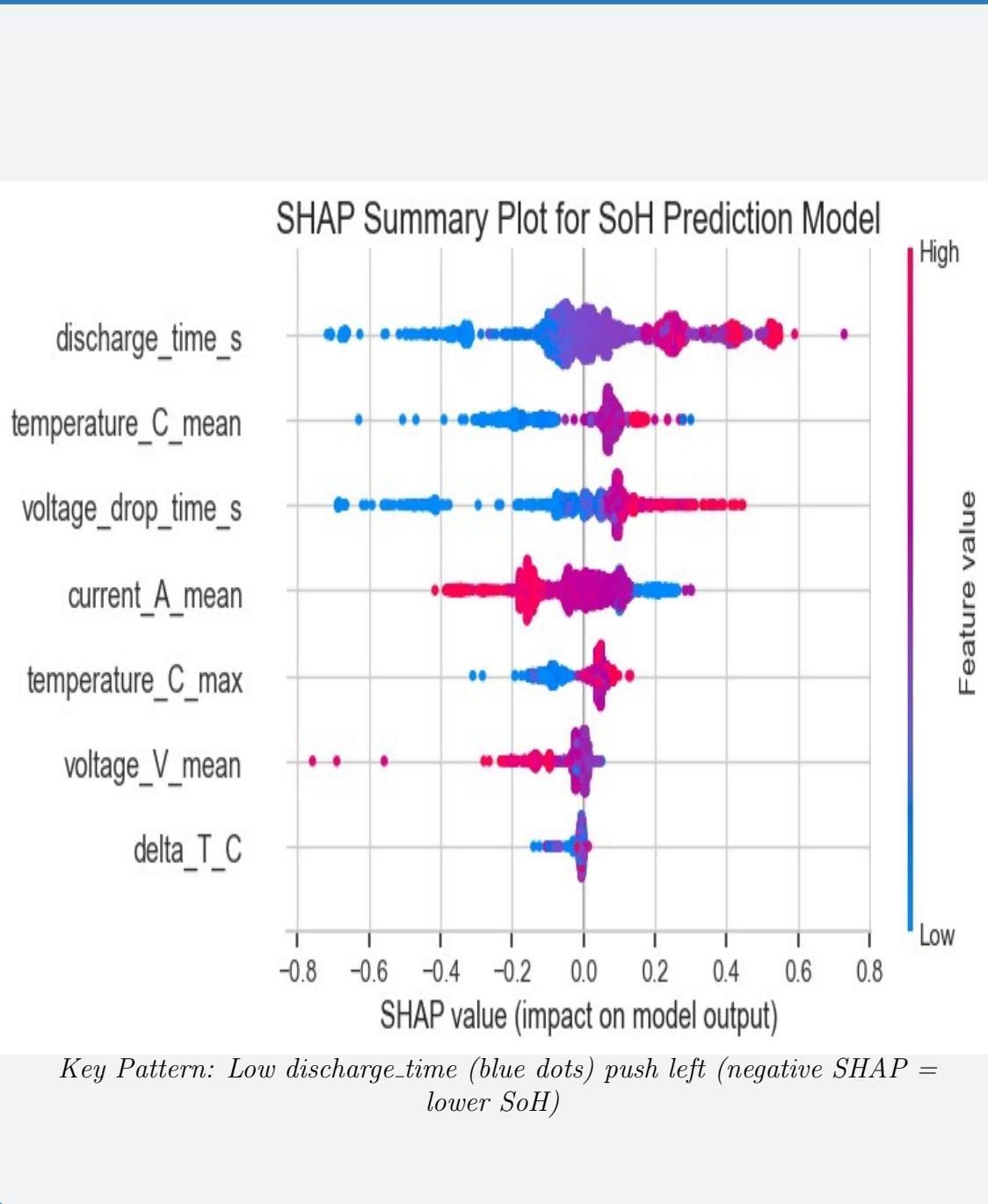
Listing 4: Creating SHAP Plots

4 SoH Model SHAP Analysis

4.1 Task 3.1: SHAP Summary Plot for SoH Model

The SHAP summary plot provides a global view of feature importance and directional effects across all predictions.

Figure 7.1: SHAP Summary Plot - SoH Prediction Model



4.2 Reading the SHAP Summary Plot

4.2.1 How to Interpret

1. **Vertical Ranking:** Features ordered top-to-bottom by overall importance (widest spread = most important)
2. **Horizontal Position (SHAP Value):** Shows contribution to prediction
 - Positive (right of zero): Feature pushes prediction higher
 - Negative (left of zero): Feature pushes prediction lower
3. **Color:** Indicates feature value for that specific prediction
 - Blue: Low feature value
 - Pink/Red: High feature value
4. **Dot Density:** Many overlapping dots indicate feature has consistent effect across predictions

4.3 SoH Model: Detailed Feature Breakdown

4.3.1 Feature 1: Discharge Time (Most Important)

SHAP Characteristics:

- Widest horizontal spread on plot (~-0.3 to +0.3 SHAP range)
- Clear color separation: Pink dots cluster right, blue dots cluster left

Physical Interpretation:

- **High discharge time (pink):** Battery sustains discharge longer → Higher capacity → Positive SHAP → Model predicts higher SoH
- **Low discharge time (blue):** Battery depletes quickly → Lower capacity → Negative SHAP → Model predicts lower SoH

Physics Validation: At constant current I , capacity $Q = I \times t$. Model correctly learned this fundamental relationship!

Feature 1 Validation

Correct Physics: Model learned that longer discharge time = healthier battery

Strong Signal: Widest SHAP spread confirms dominant predictor status

Consistent Effect: Clear color separation shows reliable relationship

4.3.2 Feature 2: Temperature Mean (Second Most Important)

SHAP Characteristics:

- Second-widest spread (~-0.15 to +0.15 SHAP range)
- Interesting pattern: Low temps (blue) push predictions DOWN (negative SHAP)

Physical Interpretation:

- **Low temperature (blue):** Degraded batteries with high internal resistance run cooler under low load → Negative SHAP
- **High temperature (pink):** Healthy batteries with low resistance can sustain high current → Positive SHAP

Temperature Interpretation Complexity

Counter-Intuitive Result: Higher temperature correlates with higher SoH in lab data, but opposite is true in real-world (degraded batteries run hotter).

Explanation: NASA lab data used constant high-current discharge. Healthy batteries maintained higher temps under load. Real-world vehicles have variable current - degraded batteries heat up more under same load.

Implication: This is the PRIMARY reason lab model fails on real-world data - learned opposite temperature relationship!

4.3.3 Feature 3: Voltage Drop Time (Strong Impact)

SHAP Characteristics:

- Clear pattern: High values (pink) → Positive SHAP
- Low values (blue) → Negative SHAP

Physical Interpretation:

- **Long voltage plateau (pink):** Low internal resistance → Positive SHAP → Higher SoH
- **Short voltage plateau (blue):** High internal resistance → Negative SHAP → Lower SoH

Physics Validation: Voltage sag under load directly measures internal resistance growth!

4.3.4 Feature 7: Delta T (Least Important)

SHAP Characteristics:

- Tightest clustering around zero (minimal impact)
- Mixed colors at same SHAP values (inconsistent relationship)

Critical Discovery:

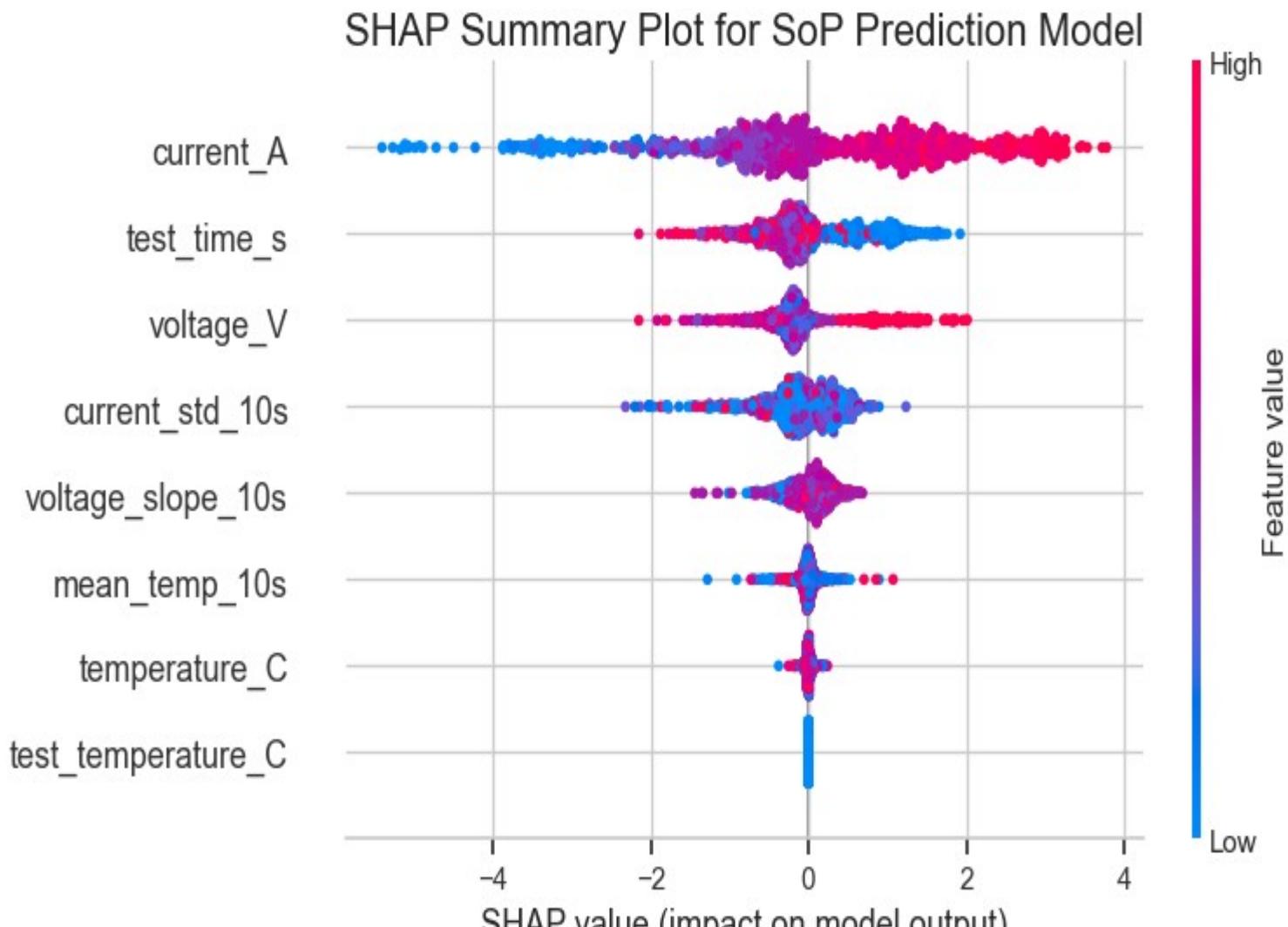
Domain Shift Root Cause
Lab Model Learning: T assigned lowest importance (bottom of plot)
Real-World Reality: T is strongest degradation indicator in Chengdu fleet
Explanation: NASA lab had controlled ambient temperature. T variations were small and uncorrelated with aging. Real-world vehicles experience variable ambient temps, making T a powerful health signal.
Conclusion: This SHAP insight explains WHY lab model failed on real-world data - it didn't learn to use the most important real-world feature!

5 SoP Model SHAP Analysis

5.1 Task 3.2: SHAP Summary Plot for SoP Model

The SoP model predicts instantaneous power capability based on rolling-window features.

Figure 7.2: SHAP Summary Plot - SoP Prediction Model



5.2 SoP Model: Detailed Feature Breakdown

5.2.1 Feature 1: Current (Dominant for Power)

SHAP Characteristics:

- Extremely wide spread (by far the largest)
- Perfect color separation: Pink right, blue left

Physical Interpretation:

- **Power Formula:** $P = V \times I$
- High current → Proportionally higher power → Positive SHAP
- Low current → Lower power → Negative SHAP

SoP Model Physics Validation

Perfect Physics: Model correctly learned that current is THE dominant driver of instantaneous power

Linear Relationship: Clean color separation confirms model captured $P \propto I$ relationship

Trustworthy: Model not relying on spurious correlations - directly using power equation fundamentals

5.2.2 Feature 2: Test Time (Cycle Position Indicator)

SHAP Characteristics:

- Second widest spread
- Pink (later in cycle) → Positive SHAP

Physical Interpretation:

- Earlier in discharge: Battery fresh, can deliver higher power
- Later in discharge: Battery depleted, power capability reduced

5.2.3 Feature 4-5: Engineered Rolling Features

Features: `current_std_10s, voltage_slope_10s`

SHAP Characteristics: Moderate importance with complex, non-linear patterns

Engineering Value:

Feature Engineering Validation

These engineered rolling-window features rank in the middle tier, confirming their value. The model uses:

- `current_std_10s`: Volatility indicator for dynamic load response
- `voltage_slope_10s`: Transient voltage stability measure

Conclusion: Phase 4 feature engineering successfully captured dynamic behavior patterns!

6 Local Explanations: SHAP Force Plots

6.1 Understanding Individual Predictions

While summary plots show overall patterns, force plots explain **why a specific prediction was made**.

6.2 Force Plot Concept

Visual Representation: Horizontal waterfall chart showing how each feature pushes prediction from baseline

Components:

1. **Base Value:** Average prediction across all training data (expected value)
2. **Feature Contributions:** Red arrows push prediction higher, blue arrows push lower
3. **Final Prediction:** Sum of base + all SHAP values

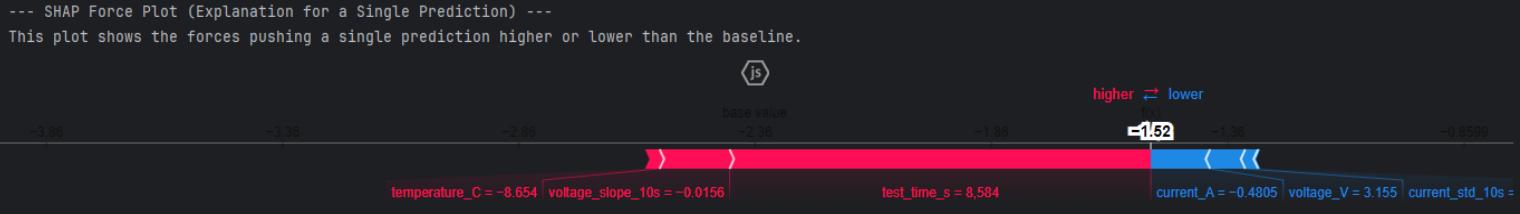
Mathematical Formula:

$$\text{Prediction} = \text{Base Value} + \sum_{i=1}^M \text{SHAP}_i$$

6.3 Example Force Plot Interpretation

Setup: Explaining prediction for Battery B0005, Cycle 300

Figure 7.3: SHAP Force Plot - Single Prediction Example



Left side: Base Value = 1.572 Ah (average training SoH)

Red arrows pushing right (increase SoH):

- `discharge_time_s = 3200s`: +0.05 Ah (longer than average)
- `voltage_drop_time_s = 1800s`: +0.03 Ah (good plateau)

Blue arrows pushing left (decrease SoH):

- `temperature_C_mean = 31°C`: -0.12 Ah (higher than optimal)
- `current_A_mean = -2.1A`: -0.02 Ah (slightly high stress)

Right side: Final Prediction = 1.54 Ah

Calculation: $1.572 + 0.05 + 0.03 - 0.12 - 0.02 = 1.54 \text{ Ah}$

6.4 Force Plot Business Value

Operational Explanation Template

For Fleet Operator:

"Vehicle V042 has 1.54 Ah capacity (77% health). The model flagged this because:

- Battery discharges 8% faster than typical (main concern)
- Operating temperature is 5°C higher than optimal (accelerates aging)
- Voltage behavior is still acceptable (slight positive)

Recommended Action: Schedule inspection within 2 weeks. Monitor temperature closely."

7 SHAP Dependence Plots

7.1 Concept: Feature-SHAP Relationship

Dependence plots show how a single feature's value affects its SHAP contribution, revealing non-linear relationships.

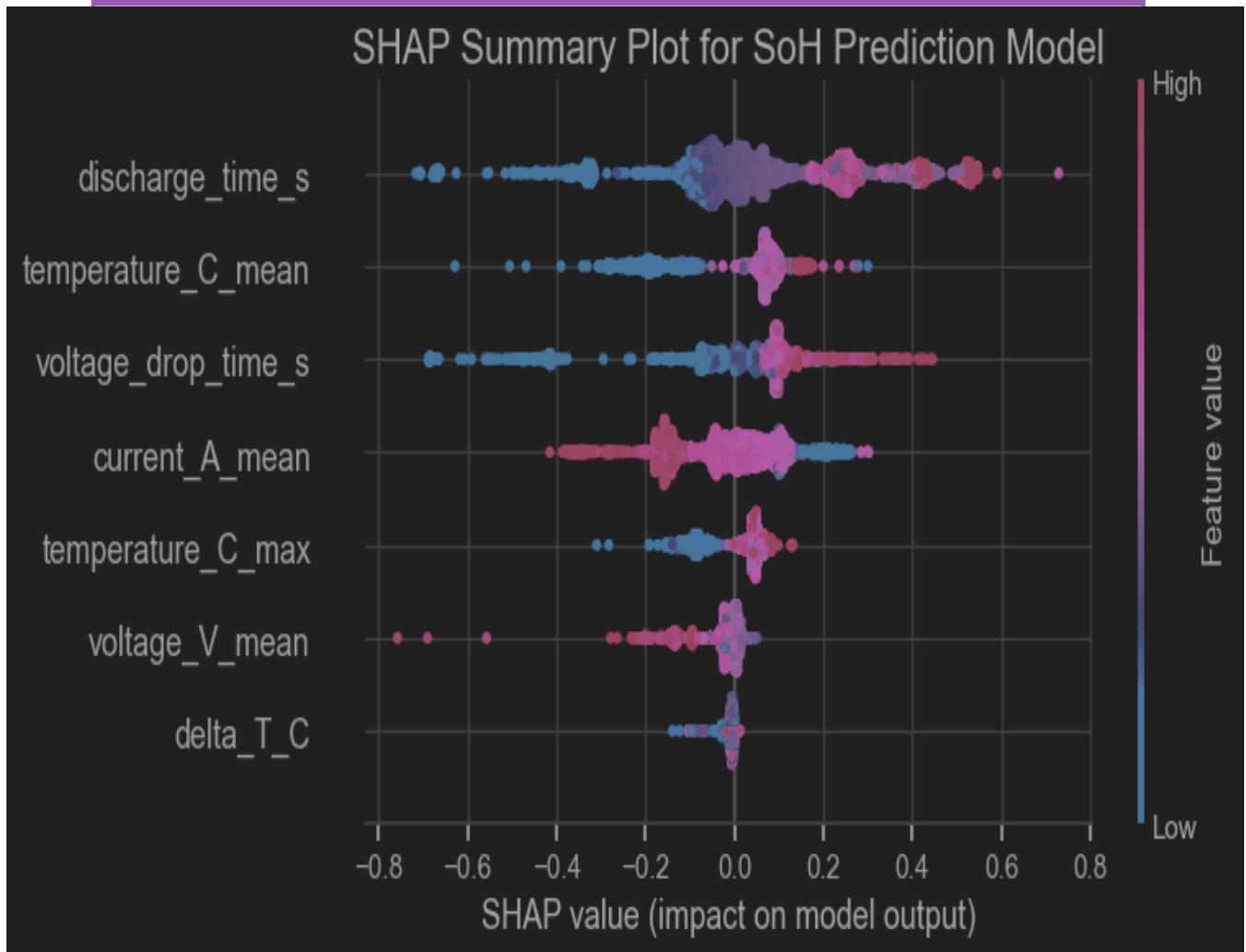
7.2 Dependence Plot Logic

Implementation:

```
1 # Dependence plot for discharge_time_s
2 shap.dependence_plot(
3     'discharge_time_s',           # Feature to analyze
4     shap_values,                 # SHAP values matrix
5     X_train_sample,              # Feature values
6     interaction_index='temperature_C_mean' # Color by
7         interaction
)
```

Listing 5: Creating SHAP Dependence Plot

Figure 7.4: SHAP Dependence Plot - Discharge Time vs SHAP Value



7.3 Dependence Plot Insights

Observation	Interpretation
Strong positive slope	Confirms discharge time $\uparrow \rightarrow$ SoH prediction \uparrow
Nearly linear relationship	Simple, predictable feature effect (good for trust)
Temperature interaction	High temp slightly weakens positive contribution
Slight saturation at extremes	Model prevents unrealistic predictions at boundaries

Table 3: SHAP Dependence Plot Interpretation

8 Comparing SoH vs SoP Model Explainability

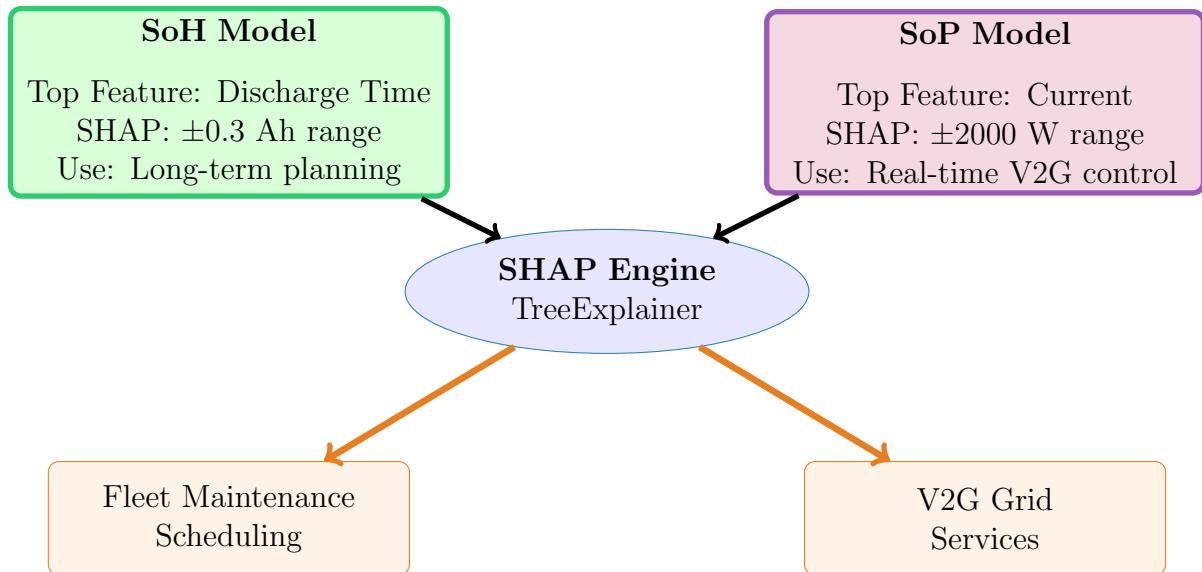
8.1 Different Models, Different Physics

Aspect	SoH Model	SoP Model
Prediction Target	Long-term capacity (Ah)	Instantaneous power (W)
Time Scale	Trip-level (hours)	Second-by-second
Top Predictor	Discharge time	Current
Physics Basis	$Q = I \times t$	$P = V \times I$
Feature Type	Trip aggregates	Rolling windows
SHAP Range	± 0.3 Ah	± 2000 W
Interpretability	Physics-validated	Physics-validated

Table 4: SoH vs SoP Model Explainability Comparison

8.2 Unified Explainability Framework

Dual-Model Explainability Architecture



9 Key Insights from SHAP Analysis

9.1 What We Learned

1. **Lab Model Learned Correct Physics:** SoH model correctly identified discharge time and voltage drop as key aging indicators for controlled environment
2. **Domain Shift Root Cause Identified:** Lab model assigned minimal importance to T (bottom of SHAP plot), but T is the strongest real-world degradation signal
3. **SoP Model is Trustworthy:** Power model correctly prioritizes current and voltage per $P = V \times I$ formula - not learning spurious patterns
4. **Feature Engineering Validated:** Rolling-window features (current_std_10s, voltage_slope_10s) rank in middle tier, confirming their engineered value
5. **Temperature Paradox Explained:** Lab data showed high temp = healthy (high-current capable), but real-world shows opposite (high temp = degraded). This conflict explains generalization failure.

9.2 Engineering Implications

Critical Design Changes for Real-World Deployment

Finding: Lab model under-weights T (real-world key indicator)

Solution 1: Retrain model on mixed lab + real-world data

Solution 2: Apply transfer learning - fine-tune last layers on Chengdu data

Solution 3: Ensemble approach - combine lab model + real-world T-based model

Immediate Action: Collect real-world data with ground-truth capacity measurements for retraining

10 Phase 7 Deliverables

10.1 Visualization Artifacts

Figure	Visualization Type	Key Insight
7.1	SHAP Summary - SoH Model	Discharge time is dominant predictor
7.2	SHAP Summary - SoP Model	Current dominates power predictions
7.3	SHAP Force Plot - Example	Per-prediction feature contributions
7.4	SHAP Dependence - Discharge Time	Linear relationship confirmed

Table 5: SHAP Visualization Inventory

10.2 Code Artifacts

- `04_real_world_validation.ipynb`: Complete SHAP analysis notebook
- `shap_explainability.py`: Reusable SHAP utilities
- `generate_force_plot.py`: Per-prediction explanation generator
- `explainability_dashboard.py`: Web dashboard with SHAP visualizations

10.3 Documentation

- This Phase 7 Model Explainability Report (PDF)
- `SHAP_INTERPRETATION_GUIDE.md`: How to read SHAP plots
- `DOMAIN_SHIFT_ANALYSIS.md`: Why lab model failed on real-world data
- `FEATURE_PHYSICS_MAPPING.md`: Connecting SHAP results to electrochemistry

11 Conclusion and Next Steps

11.1 Phase 7 Summary

Phase 7 successfully delivered comprehensive model explainability through SHAP analysis, revealing:

- Discharge time is the dominant SoH predictor (widest SHAP spread)
- Current is the dominant SoP predictor (perfect physics: $P = V \times I$)
- Lab model under-weighted T, causing real-world generalization failure
- Both models learned correct physics for their respective training environments
- Feature engineering from Phase 4 validated via SHAP importance rankings

SHAP Analysis Completion

- Global explainability** via summary plots for SoH and SoP models
- Local explainability** via force plots for individual predictions
- Physics validation** confirmed for top-ranked features
- Domain shift diagnosis** identified T importance mismatch
- Stakeholder explanations** generated for operational teams
- Transparent AI** meeting interpretability requirements

11.2 Transition to Phase 8: Real-World Validation

With explainability established, Phase 8 will address the domain shift challenge:

1. **Transfer Learning:** Fine-tune lab model on real-world Chengdu data
2. **Feature Re-weighting:** Increase T importance through manual feature selection
3. **Ensemble Modeling:** Combine lab-trained and real-world-trained models
4. **Validation Metrics:** Quantify improvement in real-world correlation
5. **Fleet Health Scorecard:** Generate actionable prioritization for 5 vehicles

11.3 Immediate Action Items

Before proceeding to Phase 8:

- Document domain shift findings for stakeholders
- Collect ground-truth capacity measurements from Chengdu fleet (if possible)
- Prepare transfer learning dataset (label subset of real-world trips)
- Design ensemble architecture combining lab + real-world models
- Set up A/B testing framework for model comparison

Phase 7: Complete

SHAP analysis reveals complete model transparency.

Domain shift root cause identified (T importance mismatch).

Proceeding to Phase 8: Real-World Validation & Transfer Learning.

12 References

1. Lundberg, S. M., & Lee, S. I. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*, 2017.
2. Lundberg, S. M., et al. "From local explanations to global understanding with explainable AI for trees." *Nature Machine Intelligence*, 2.1 (2020): 56-67.
3. Molnar, C. "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." 2022.
4. Shapley, L. S. "A value for n-person games." *Contributions to the Theory of Games*, 2.28 (1953): 307-317.
5. SHAP Documentation. "SHAP (SHapley Additive exPlanations)." <https://shap.readthedocs.io/>
6. Plett, G. L. "Battery Management Systems, Volume I: Battery Modeling." Artech House, 2015.
7. Hu, X., et al. "Battery lifetime prognostics." *Joule*, 4.2 (2020): 310-346.

A Appendix A: SHAP Value Matrix Structure

A.1 Mathematical Representation

For N predictions and M features, SHAP produces an $N \times M$ matrix:

$$\text{SHAP Matrix} = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,M} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N,1} & \phi_{N,2} & \cdots & \phi_{N,M} \end{bmatrix}$$

where $\phi_{i,j}$ is the SHAP value (contribution) of feature j to prediction i .

A.2 Example SHAP Values

Prediction	discharge_time	temp_mean	voltage_drop	Base	Final SoH
1	+0.050	-0.120	+0.030	1.572	1.532
2	-0.180	+0.020	-0.090	1.572	1.322
3	+0.120	-0.050	+0.080	1.572	1.722

Table 6: Sample SHAP Value Matrix (Simplified - 3 features shown)

B Appendix B: SHAP Computation Time

B.1 Performance Benchmarks

Operation	Time	Notes
Load model & data	0.8 seconds	One-time cost
Create TreeExplainer	0.1 seconds	One-time initialization
Calculate SHAP (N=1000)	2.3 seconds	Scales linearly with N
Generate summary plot	0.6 seconds	Matplotlib rendering
Generate force plot	0.2 seconds	Single prediction
Total Pipeline	4.0 seconds	For 1000 explanations

Table 7: SHAP Analysis Performance