Battery Degradation Prediction

By Elvin Rustamov, Asim Mahmudov and Vugar Hasanov

Understanding the dataset:

The dataset used in this project originates from the publication *"Data-driven prediction of battery cycle life before capacity degradation"*. It consists of cycling data from 124 commercial lithium iron phosphate (LFP)/graphite lithium-ion cells (A123 APR18650M1A) tested under fast-charging protocols until failure. Each battery was cycled using a one-step or two-step fast-charging policy and discharged at a fixed 4C rate.

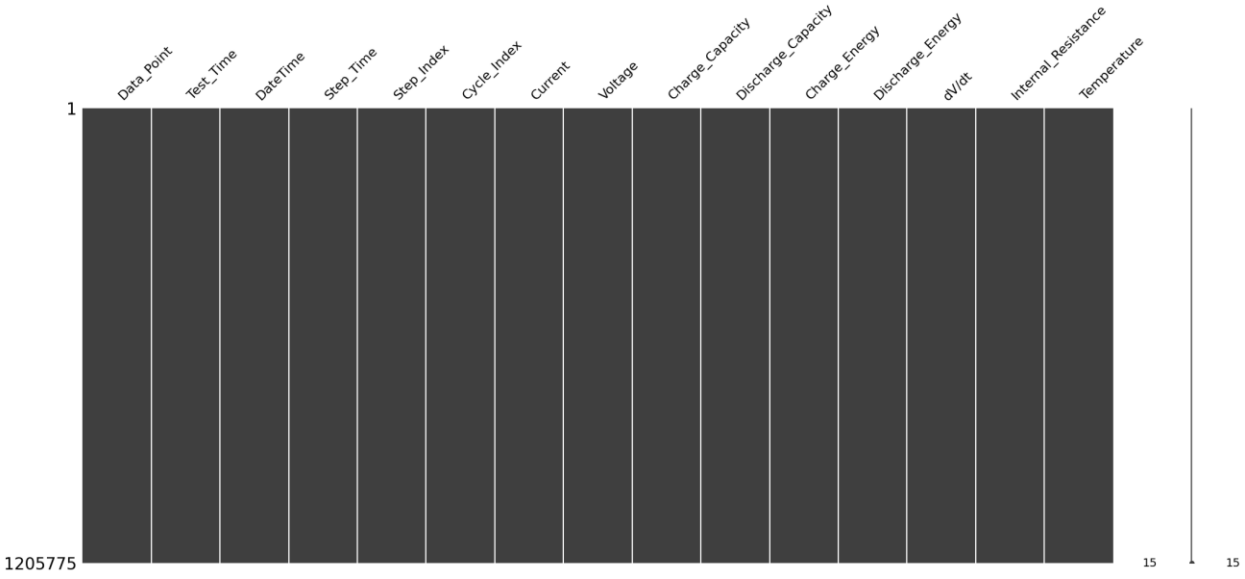
Each CSV file corresponds to a single battery cell and contains time-series data across many cycles. The columns in the dataset are defined as follows:

Column Name	Description
data_point	A unique index for each recorded datapoint.
test_time	Total time elapsed since the beginning of the test (in seconds). May reset due to equipment errors in raw CSVs.
datetime	UNIX timestamp representing the wall-clock time of the datapoint.
step_time	Time elapsed in the current test step (in seconds).
step_index	The index of the current test step within a cycle (e.g., constant-current charge, constant-voltage hold, rest).
cycle_index	The index of the full charge-discharge cycle. The dataset tracks battery degradation across cycles.
current *(A)*	The measured current at the given time (positive for discharge, negative for charge).
voltage *(V)*	The cell voltage at the given time. The upper/lower limits are 3.6 V and 2.0 V.
charge_capacity *(Ah)*	The integrated charge delivered to the cell during the current charge phase.
discharge_capacity *(Ah)*	The integrated charge extracted from the cell during the current discharge phase.
charge_energy *(Wh)*	The cumulative energy delivered during the charging step.

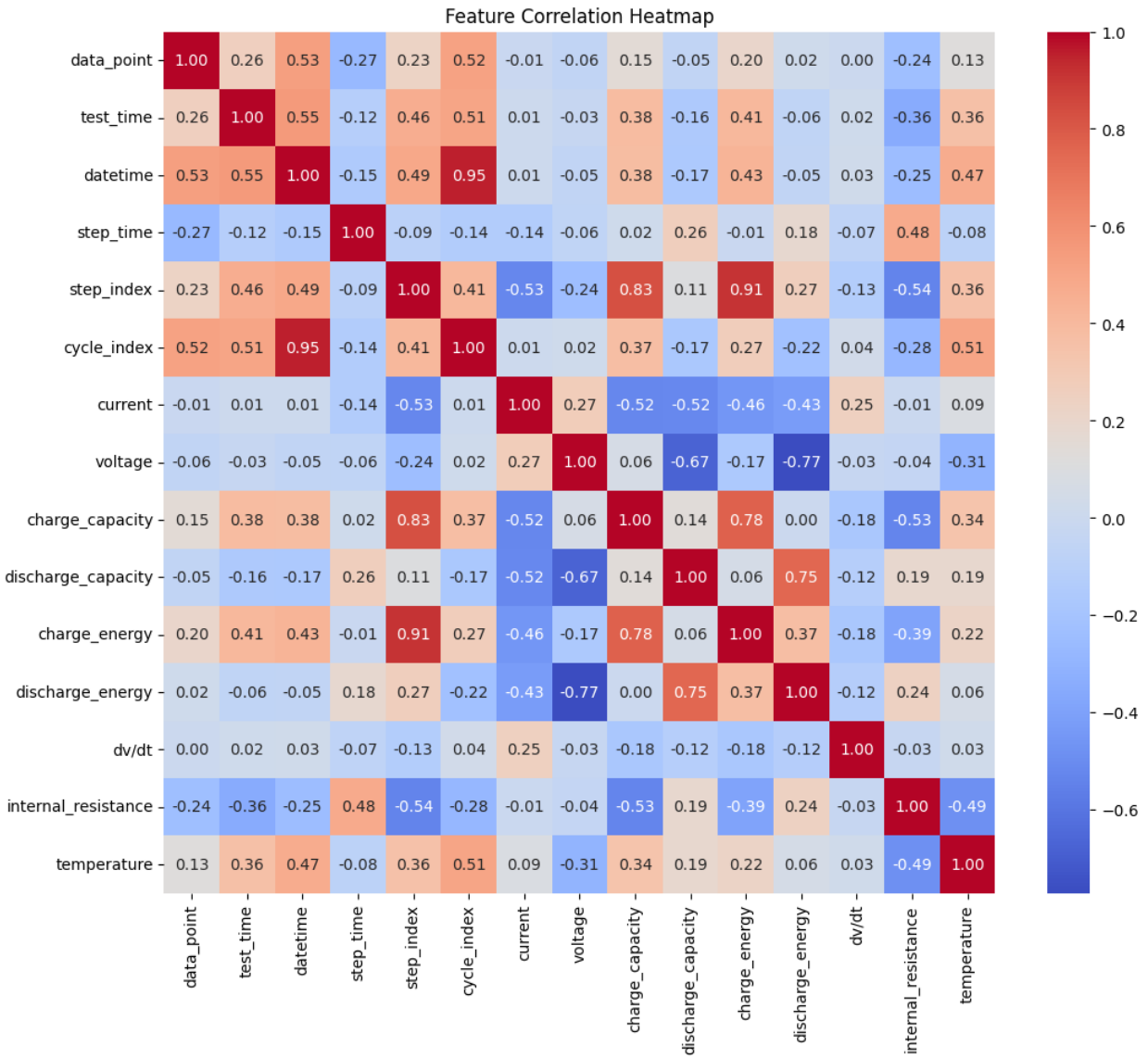
Column Name	Description
discharge_energy *(Wh)*	The cumulative energy extracted during the discharging step.
dv/dt *(V/s)*	Derivative of voltage with respect to time, useful for identifying state transitions and internal reactions.
internal_resistance *(Ω)*	Internal resistance is estimated at the timepoint.
temperature *(°C)*	Measured temperature of the cell using a Type T thermocouple. Note that this reading may be affected by inconsistent thermal contact.

(In the project, we remove the columns: Data_Point, Test_Time, DateTime, Step_Time, Step_Index, Cycle_Index (These columns are either identifiers or timestamps, which do not directly contribute to the prediction of SOH or battery degradation))

All the columns in the dataset are numeric types and there are no missing values:



Correlation between features:



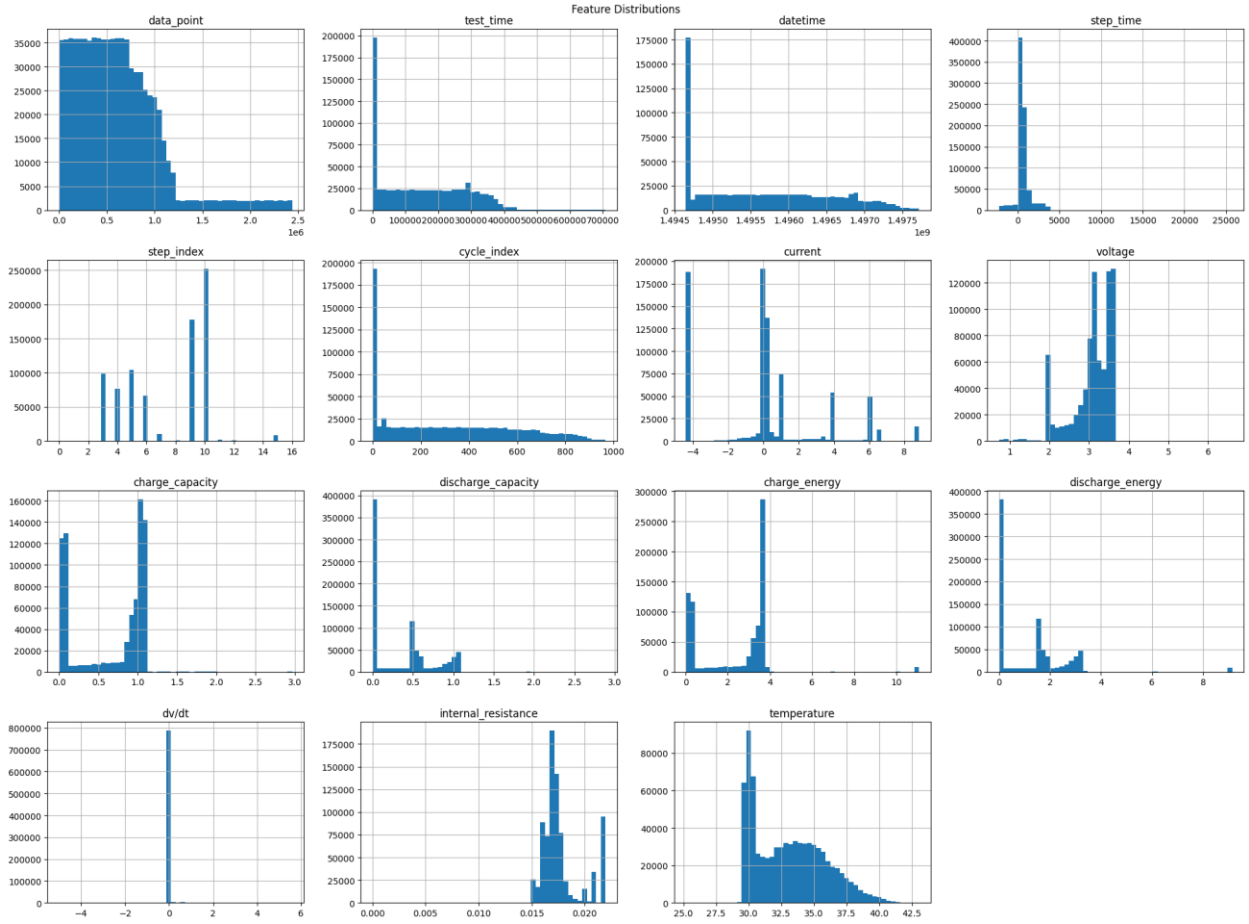
To understand the relationships between different features, a correlation heatmap was generated. In this heatmap, each cell shows the correlation coefficient between two features, with values ranging from -1 to 1.

In the heatmap, red colors represent positive correlations, blue colors represent negative correlations, and lighter colors represent weak or no correlations.

Notably, **datetime** and **cycle_index** show a very strong positive correlation (0.95), suggesting that as time progresses, the cycle index increases as expected. Similarly, **charge_capacity** and **charge_energy** are strongly correlated (0.78), indicating that higher charge capacities are associated with higher charge energies.

Conversely, **voltage** and **charge_energy** exhibit a strong negative correlation (-0.77), meaning that as voltage increases, the charge energy tends to decrease.

Feature Distributions:



To better understand the nature of the data, histograms for each feature were plotted. These plots show how the values of each feature are distributed across the dataset.

- Features such as **test_time**, **step_time**, and **cycle_index** are heavily **right-skewed**, indicating that most of the values are concentrated near the lower end, with a few very large values.
- Features like **charge_capacity**, **discharge_capacity**, and **charge_energy** show a **clustered distribution** with sharp peaks, suggesting that the batteries operate around specific capacity and energy ranges.
- **current** and **dv/dt** have **multiple peaks (multimodal distributions)**, possibly due to different charging/discharging steps.
- **internal_resistance** and **temperature** distributions are relatively **narrow**, with most data points concentrated within a specific range.
- **voltage** distribution is slightly skewed towards higher voltages, reflecting typical battery behavior under charge and discharge cycles.

Overall, the distributions reveal that many features are **not normally distributed**, which may require normalization or transformation before applying machine learning models.

Model Creation and Prediction:

To predict the State of Health (SOH) of the battery, we developed several regression models using machine learning techniques, namely XGBoost, LightGBM, and Random Forest Regressor.

First, we performed feature selection by retaining only the most relevant sensor readings, including current, voltage, charge/discharge capacity, charge/discharge energy, dV/dt, internal resistance, and temperature. The SOH value was calculated as a normalized ratio of the charge capacity relative to the nominal capacity and then transformed using a natural logarithm (log1p) to stabilize the variance.

Outliers were detected and removed using the Z-score method across selected important features to ensure the models were trained on clean data. Feature scaling was applied using StandardScaler to normalize the range of input variables.

The dataset was then randomly sampled (100,000 samples) and split into training and testing sets with an 80/20 ratio.

For each model, hyperparameter tuning was performed using **GridSearchCV** with 3-fold cross-validation to find the optimal parameters that minimize the Mean Squared Error (MSE). The parameter grids were carefully designed separately for tree-based models and ensemble models.

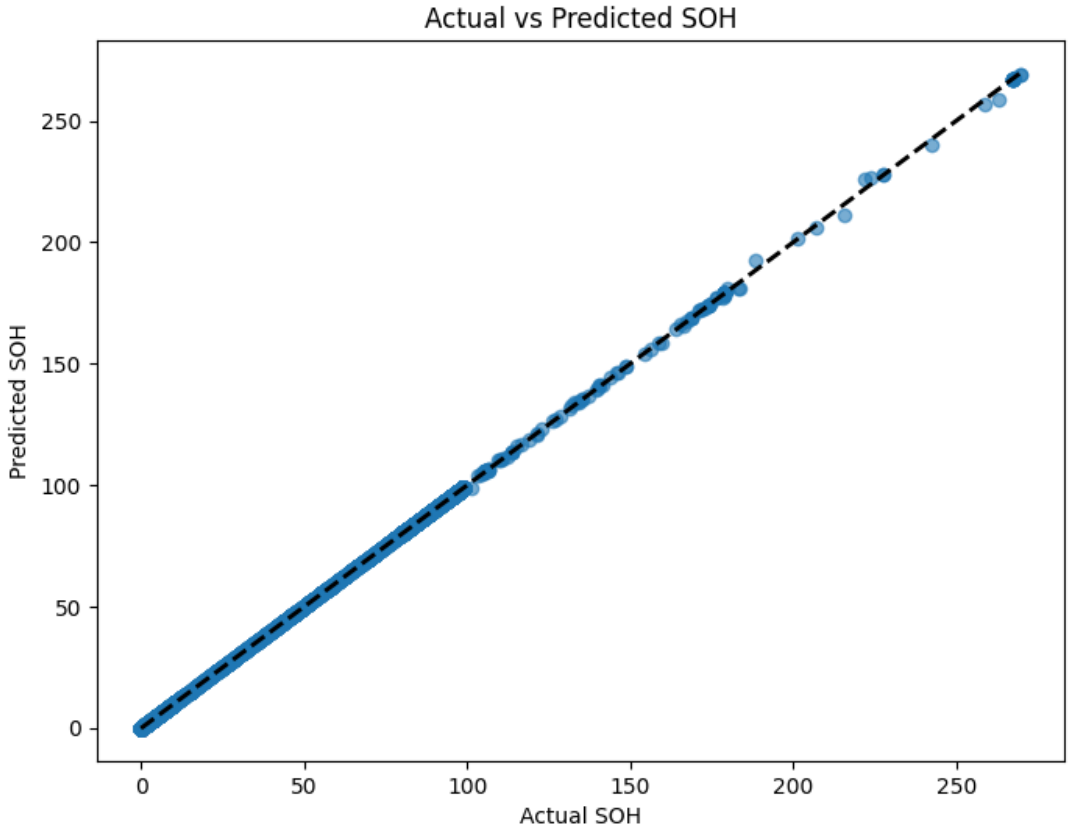
Random Forest Regressor Results

For the Random Forest model, the best hyperparameters found were:

- **Bootstrap**: True
- **Max Depth**: 10
- **Min Samples Split**: 5
- **Number of Estimators**: 300

The Random Forest model achieved the following metrics on the test set:

- **R^2 Score**: 1.00
- **Mean Squared Error**: 0.01

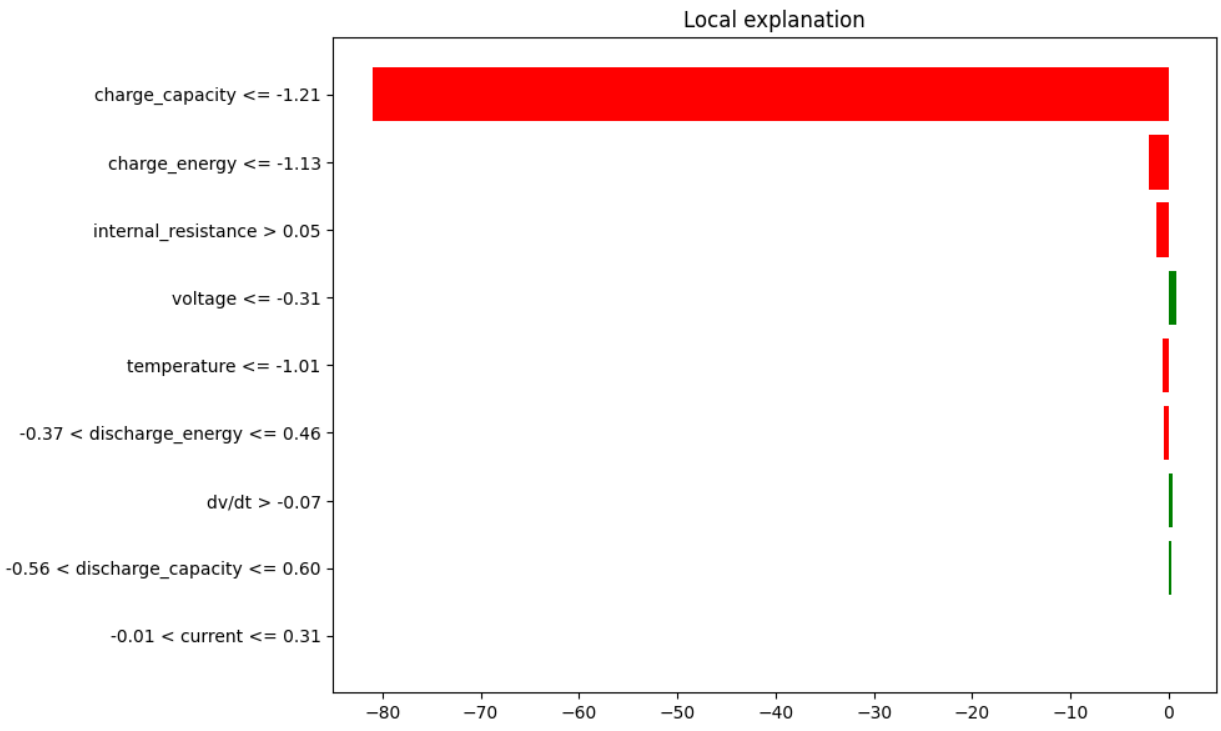


Additionally, model interpretability was explored using:

- **Actual vs. Predicted SOH Scatter Plot** to visually evaluate the prediction accuracy.
- **Feature Importance Plot** to understand which features most influenced the model's predictions.
- **LIME (Local Interpretable Model-agnostic Explanations)** to analyze individual prediction explanations.
- **SHAP (SHapley Additive exPlanations)** summary plots to visualize the global feature impact across the dataset.

From the **LIME analysis** for a selected prediction:

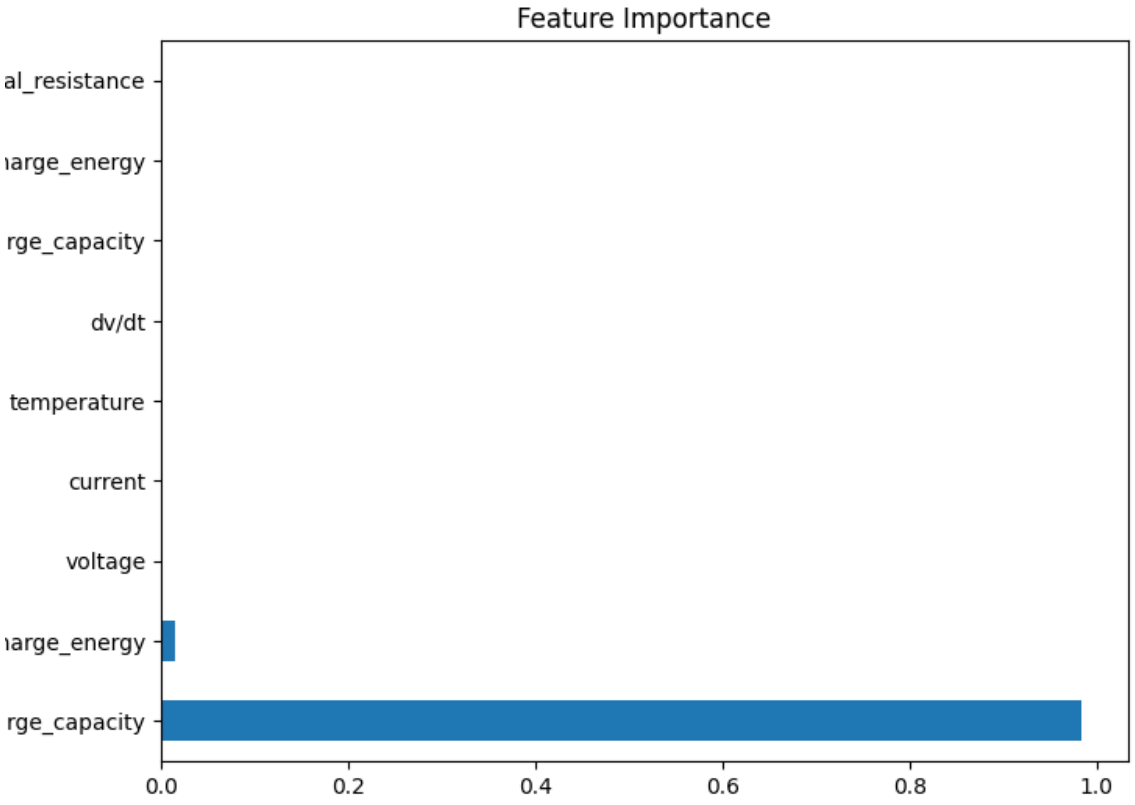
- The intercept was **86.80**, and the local prediction was approximately **2.99**.
- The most influential features included **charge_capacity**, **charge_energy**, and **internal_resistance**, with negative contributions, while features like **voltage**, **dv/dt**, and **current** provided smaller positive impacts.



Feature Importance Analysis

The feature importance plot derived from the Random Forest model reveals that **charge_capacity** is by far the most influential feature for predicting the SOH. It accounts for almost the entirety of the model's predictive power, followed distantly by **charge_energy**. Other features such as **current**, **voltage**, **dv/dt**, **internal_resistance**, and **temperature** have negligible impact in comparison.

This result is consistent with the theoretical expectation, as the SOH is directly linked to the battery's ability to retain charge over cycles.



SHAP Analysis

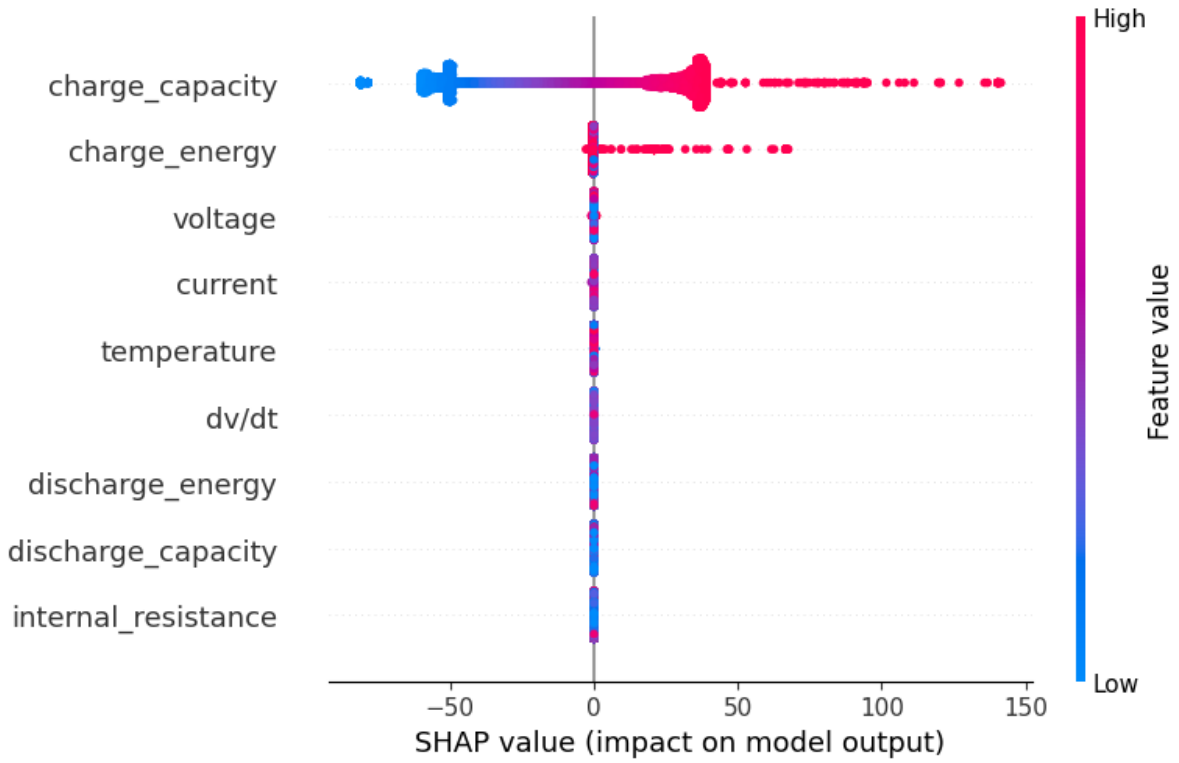
To provide a more detailed, instance-level interpretability, SHAP (SHapley Additive exPlanations) values were computed for the Random Forest model.

The SHAP summary plot further confirms that **charge_capacity** is the most critical factor influencing the model's output.

- Higher values of **charge_capacity** (colored red on the right side) have a strong positive impact on the predicted SOH.
- Lower values of **charge_capacity** (colored blue on the left) strongly decrease the predicted SOH.
- **charge_energy** also contributes slightly, where higher charge energy values tend to slightly improve the SOH prediction.

Other features showed relatively minor SHAP contributions, indicating that while they are included in the model, their influence on the prediction output is minimal compared to **charge_capacity**.

Thus, both the feature importance ranking and SHAP analysis consistently highlight **charge_capacity** as the dominant predictor of SOH in this dataset.

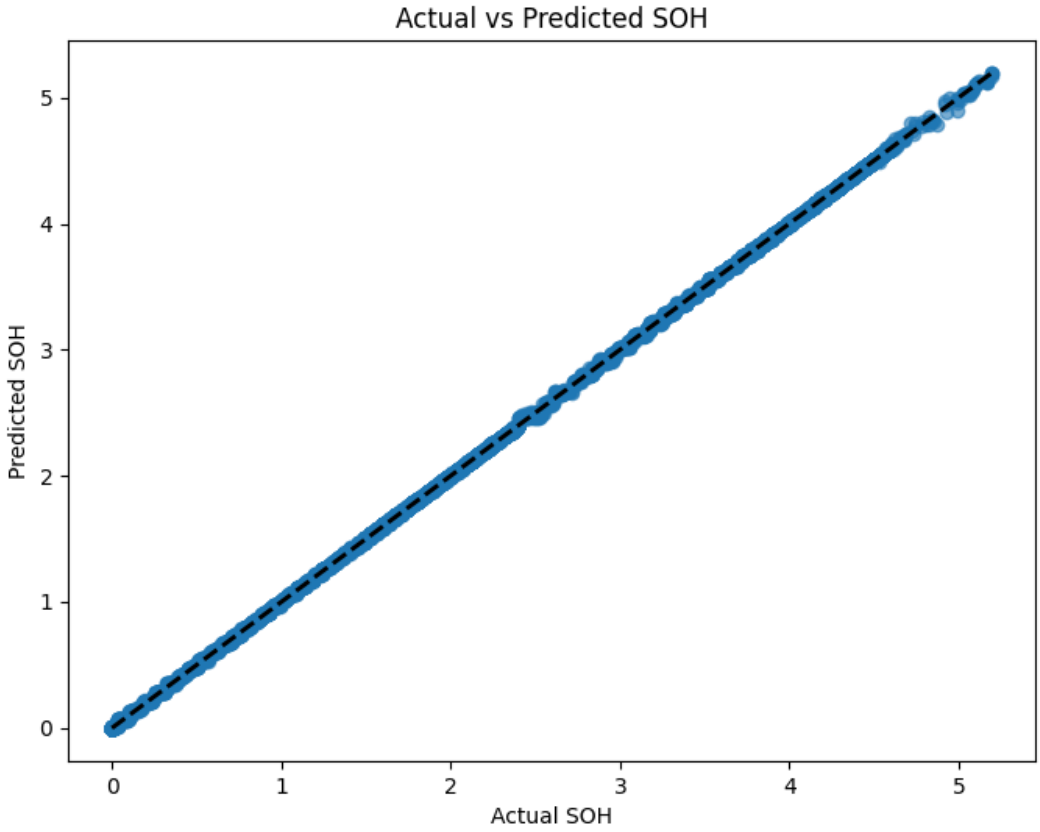


LightGBM Model Results

For the LightGBM model, the best configuration was selected with regularization parameters (lambda_l1=0.3 and lambda_l2=0.3) to prevent overfitting.

The LightGBM model achieved the following metrics on the test set:

- **R^2 Score**: 1.00
- **Mean Squared Error**: 0.00



During training, LightGBM issued warnings regarding manual setting of lambda_l1 and lambda_l2 instead of using reg_alpha and reg_lambda, but the custom values were correctly applied.

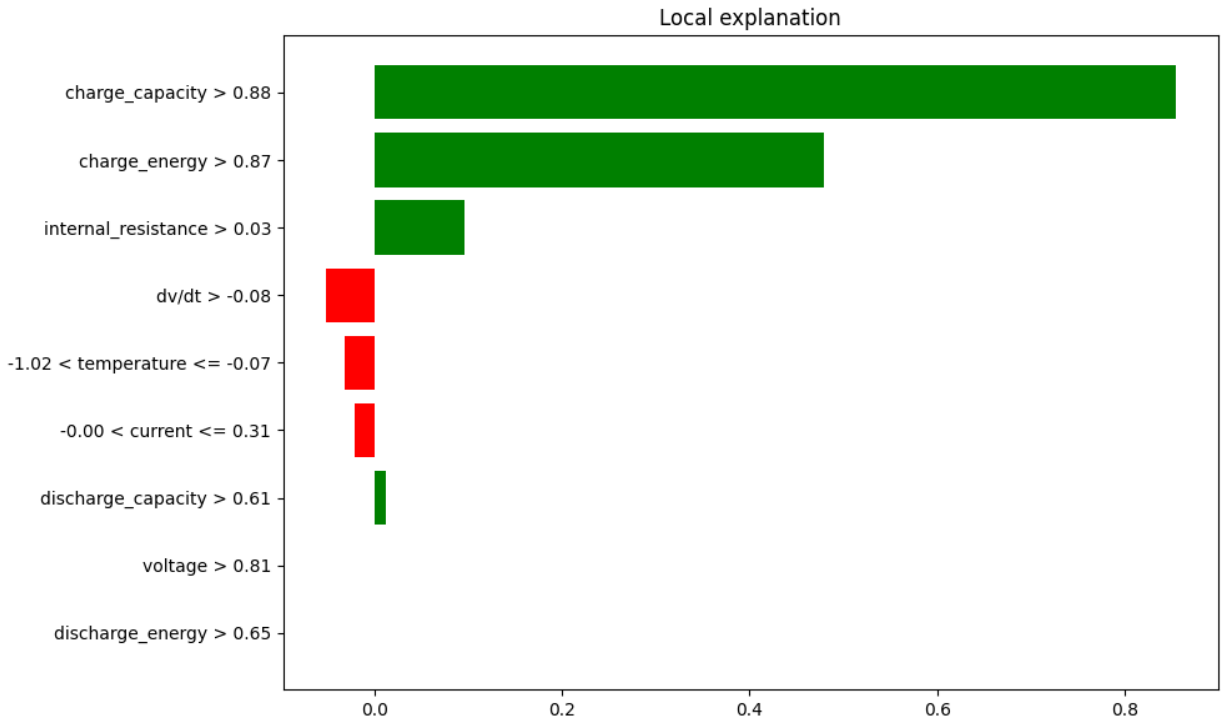
Model Interpretability Analysis

In order to evaluate and explain the model's decision-making process, several visualization techniques were employed:

- **Actual vs. Predicted SOH Scatter Plot** shows an almost perfect alignment along the ideal diagonal line, confirming the high accuracy of the LightGBM model.
- **Feature Importance Plot** indicates the importance of different features used by the model during training.
- **LIME (Local Interpretable Model-Agnostic Explanations)** provides a detailed explanation for individual predictions.
- **SHAP (SHapley Additive exPlanations) Summary Plot** highlights the global feature importance across the entire dataset.

From the LIME analysis for a selected prediction:

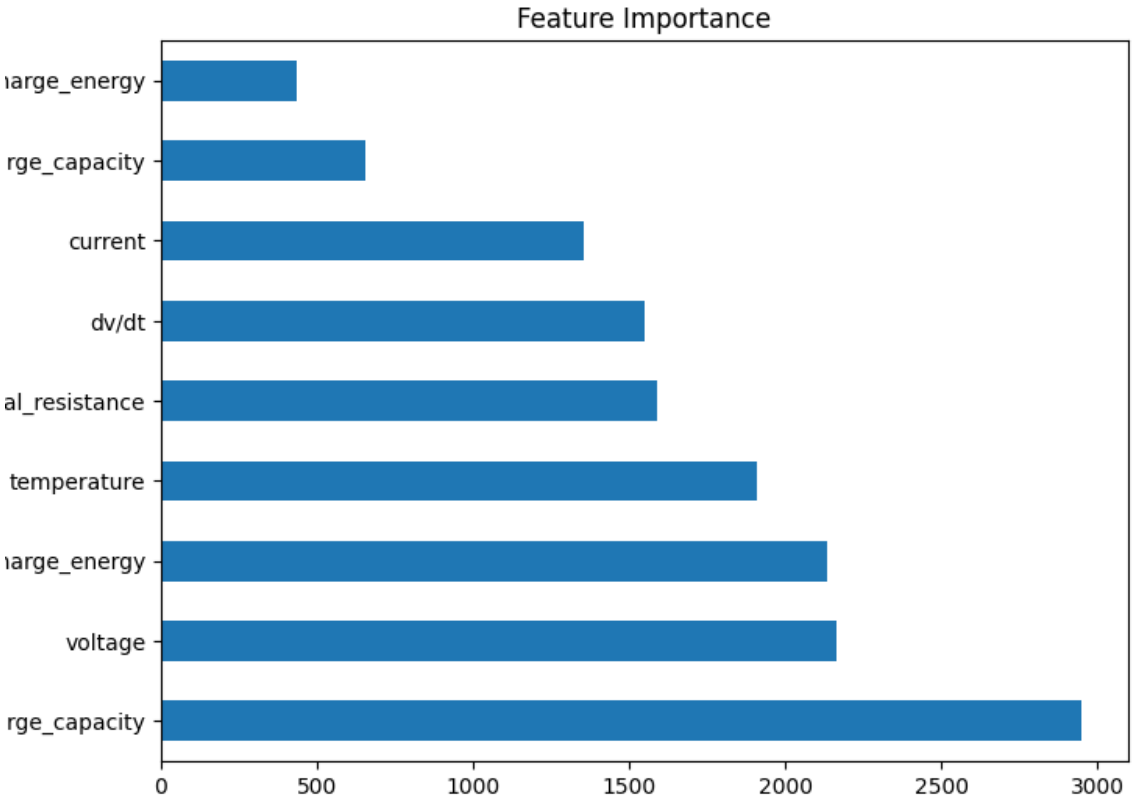
- The intercept was approximately **3.27**, and the local prediction was **4.61**, while the true target value was around **5.13**.
- The most influential features were **charge_capacity** and **charge_energy**, with positive contributions, while features like **dv/dt**, **temperature**, and **current** made small negative contributions.



Feature Importance Analysis

The feature importance plot obtained from the LightGBM model indicates that **charge_capacity** remains the dominant feature for predicting the SOH, similar to the Random Forest results.
It is followed by **voltage** and **charge_energy**. Other features such as **temperature**, **internal_resistance**, **dv/dt**, and **current** contribute moderately.

This ranking aligns with expectations, as **charge_capacity** directly reflects the battery's state, while voltage and charge energy further capture important electrochemical behaviors.

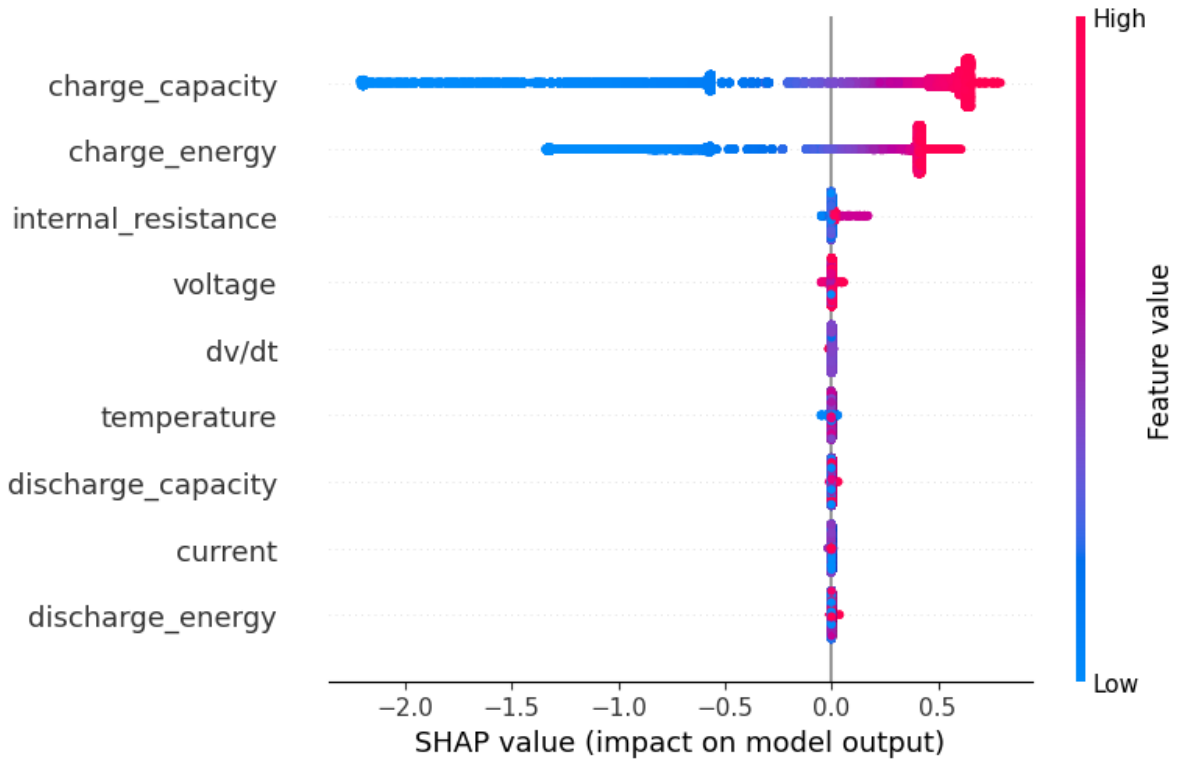


SHAP Analysis

The SHAP summary plot for LightGBM further validates the findings:

- Higher values of **charge_capacity** have a strong positive impact on the SOH predictions.
- **Charge_energy** and **internal_resistance** also contribute to the model's outputs, albeit to a lesser extent.
- Features such as **dv/dt**, **voltage**, **temperature**, and **current** showed moderate influence.

Overall, both feature importance and SHAP analysis confirm that **charge_capacity** is the most critical predictor of battery SOH, with **charge_energy** and **voltage** serving as secondary indicators.



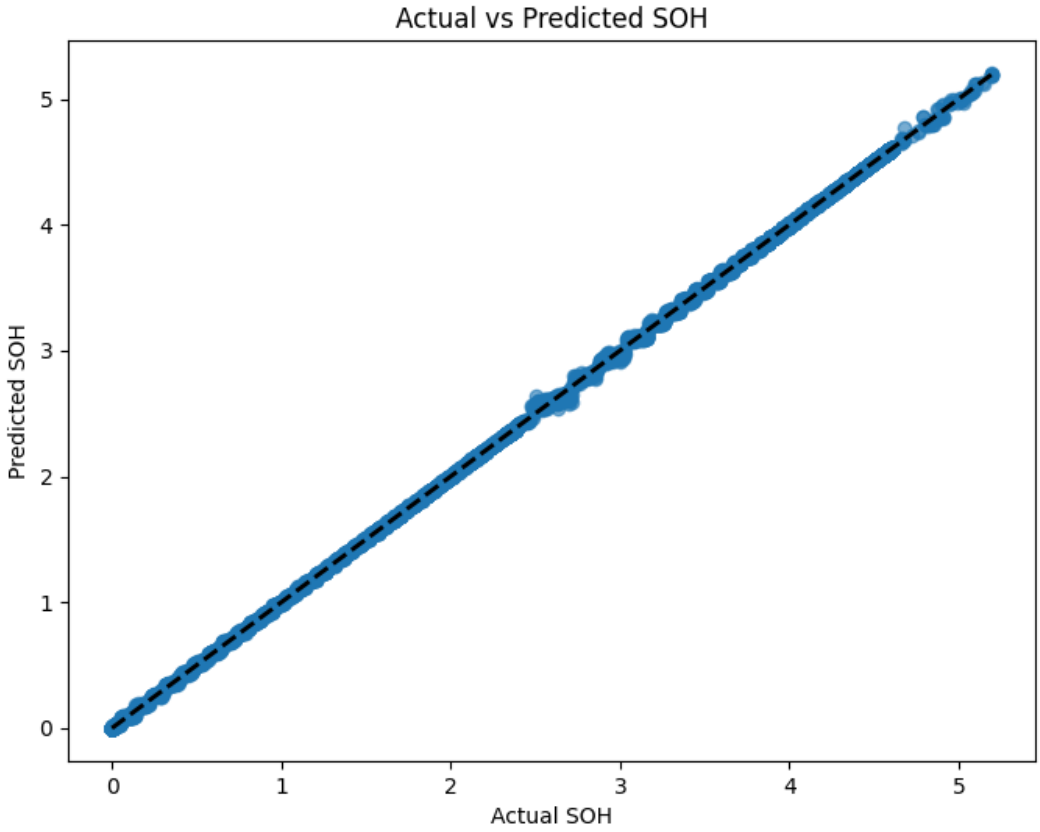
XGBoost Model Results

For the XGBoost model, hyperparameter tuning was performed using GridSearchCV, and the best parameters found were:

- **Colsample_bytree**: 0.8
- **Learning_rate**: 0.1
- **Max_depth**: 7
- **Number of Estimators**: 500
- **Subsample**: 1.0

The XGBoost model achieved the following evaluation metrics on the test set:

- **R^2 Score**: 1.00
- **Mean Squared Error**: 0.00



These results indicate an almost perfect fit of the model to the battery SOH prediction task.

Model Interpretability Analysis

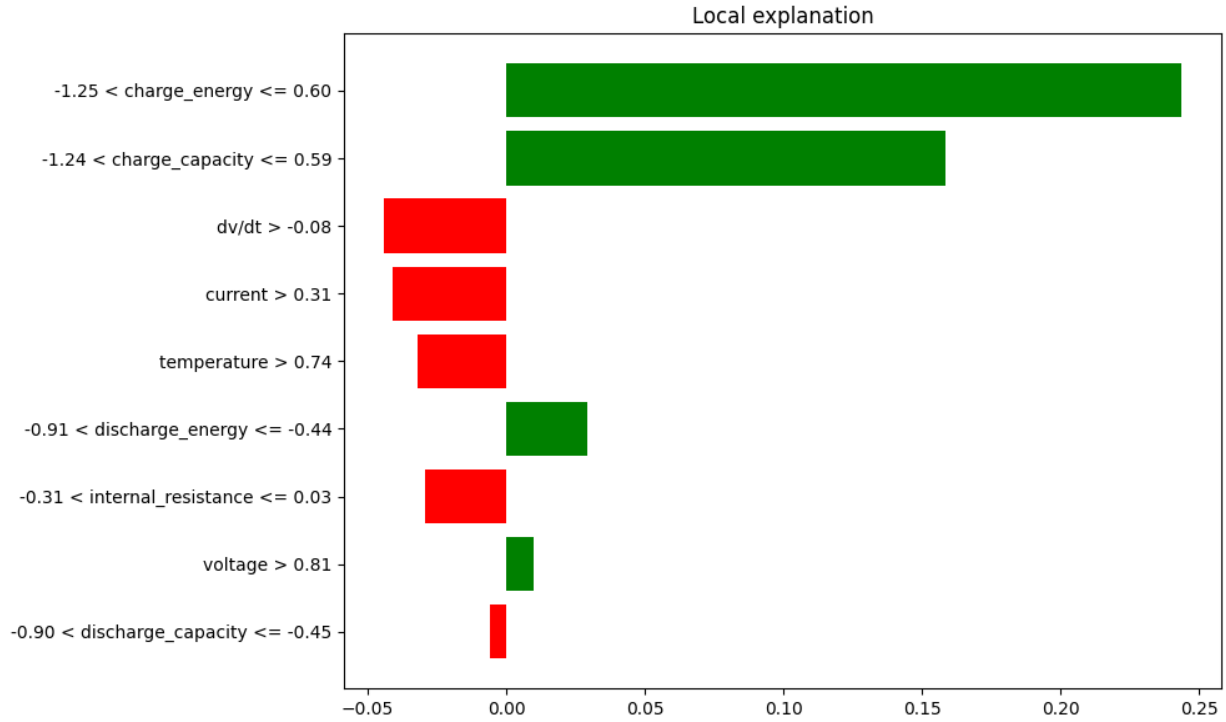
To interpret the model behavior and validate its decision-making process, multiple techniques were applied:

- **Actual vs. Predicted SOH Scatter Plot** demonstrated that the predictions closely match the actual values, with data points aligning very tightly along the ideal line.
- **Feature Importance Plot** provided insights into which features were most influential during the model's training.
- **LIME (Local Interpretable Model-Agnostic Explanations)** was used to explain an individual prediction in detail.
- **SHAP (SHapley Additive exPlanations) Summary Plot** allowed global feature impact analysis across the entire test set.

From the LIME analysis for a selected test instance:

- The model's **intercept** was approximately **3.55**, and the **local prediction** was **3.84**, while the true SOH value was approximately **4.12**.

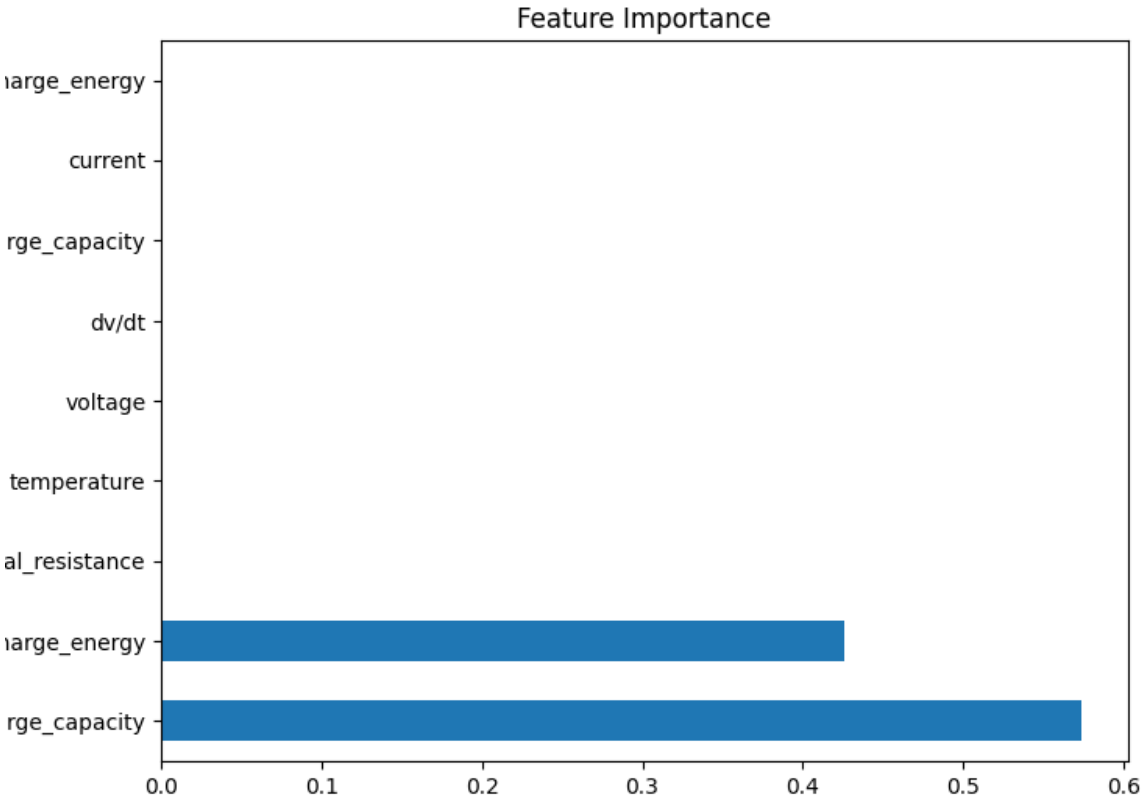
- The most influential positive contributors to the prediction were **charge_energy** and **charge_capacity**, whereas **dv/dt**, **current**, and **internal_resistance** had negative impacts.



Feature Importance Analysis

According to the XGBoost feature importance plot, the features **charge_capacity** and **charge_energy** emerged as the most critical drivers for the model's predictions.
Other features like **current**, **dv/dt**, **temperature**, **voltage**, and **internal_resistance** had comparatively smaller influences.

This distribution matches physical expectations, as charge-related features directly correlate with battery degradation behavior over time.



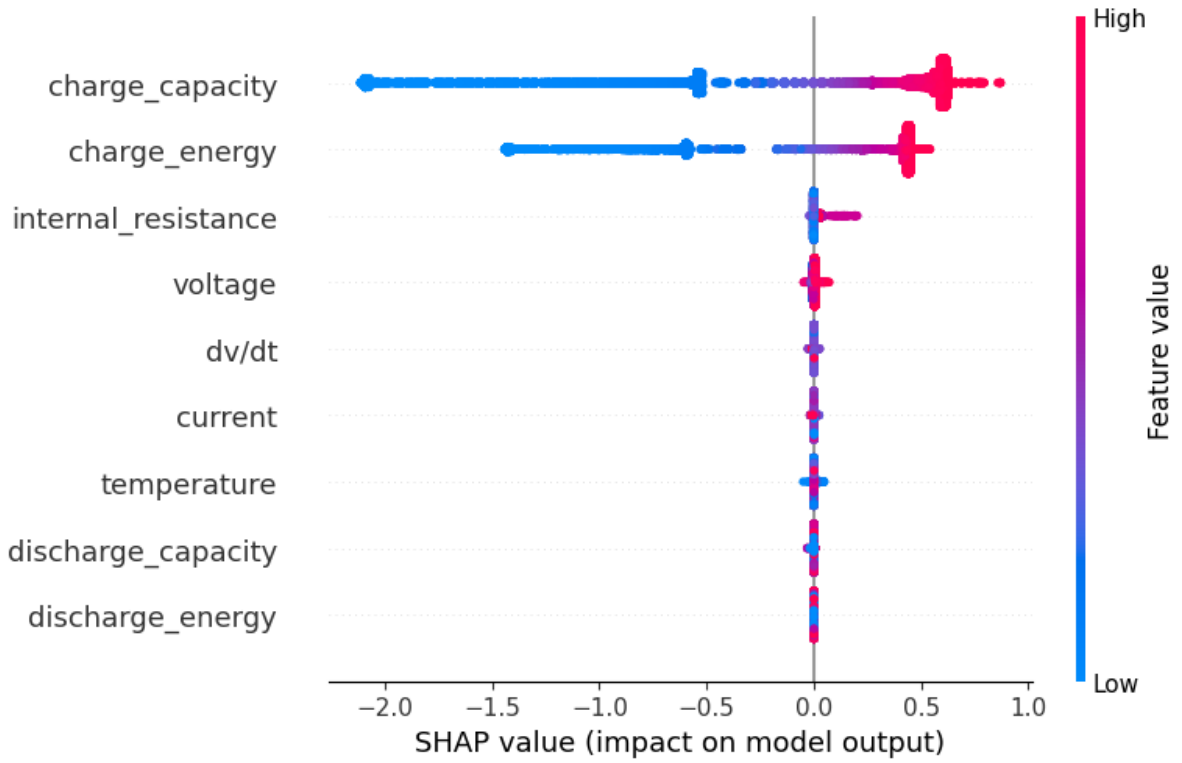
SHAP Analysis

The SHAP summary plot for XGBoost confirms the model's dependency on:

- **High charge_capacity** positively pushing the SOH prediction higher.
- **High charge_energy** also contributing positively.
- Negative impacts being associated with deviations in **dv/dt**, **current**, and **internal_resistance**.

Features like **voltage** and **discharge_capacity** had minor but noticeable effects.

Overall, both the feature importance analysis and SHAP values reaffirm that **charge_capacity** and **charge_energy** are the dominant predictors of SOH, aligning with results observed across all models (Random Forest, LightGBM, and XGBoost).



Conclusion

In this study, three machine learning models — **Random Forest Regressor**, **LightGBM**, and **XGBoost** — were developed and evaluated for predicting the **State of Health (SOH)** of batteries based on key sensor measurements.

All three models demonstrated **outstanding performance**, achieving:

- **R^2 Scores close to 1.00** (perfect prediction),
- **Mean Squared Errors (MSE) close to 0.00** on the test set,
  indicating that the selected features and preprocessing steps were highly effective for modeling SOH.

Across all models, feature importance analysis consistently highlighted **charge_capacity** as the **dominant predictive feature**, followed by **charge_energy** and, to a lesser extent, **internal_resistance** and **voltage**.
This result is physically meaningful, as a battery's capacity retention is a direct measure of its health.

Interpretability methods, including **LIME** and **SHAP** analyses, confirmed that:

- **Higher values of charge_capacity and charge_energy positively impact** the SOH prediction.
- Features such as **dv/dt**, **current**, and **internal_resistance** can have smaller negative effects depending on their values.

The extremely close alignment between **actual and predicted SOH values** in scatter plots for all three models further confirms their reliability and generalization capability.

Final Remarks

While all three models perform very similarly, **LightGBM** and **XGBoost** slightly outperform Random Forest in terms of efficiency and computational speed, particularly when dealing with larger datasets. Therefore, **LightGBM** and **XGBoost** are recommended as the primary models for future large-scale battery SOH prediction tasks.

Overall, the combination of careful data cleaning, feature selection, and powerful ensemble models enables highly accurate and interpretable battery health predictions.