Project title: Predicting Battery Lifecycle using Charging-Discharging Characteristics

Team members: Ritoban Ghosh, Barath Tirumuruhan, Ashutosh Nehete

Project executive summary:

This report presents a machine learning approach for predicting the lifecycle of lithium-ion batteries, which is essential in their development and maintenance for efficient energy storage solutions. The report uses a dataset from the Battery Archive Repository, which comprises experiments conducted at Sandia National Labs. The dataset includes 86 commercial ‘18650’ NCA, NMC, and LFP cells cycled at different charging-discharging rates, and the experiments examine the influence of temperature, depth of discharge, and discharge current on the long-term degradation of commercial cells. Here, we develop a model that predicts the lifecycle of a battery using only the initial charging-discharging characteristics. The report explains the data structure and visualization of raw data, feature engineering, and calculating the target value (lifetime of a cell). The predictive model uses five statistical parameters, four features from the capacity fade curves and the data provided covers six experimental conditions. Here, the target is the lifetime of a cell, which is defined as the number of cycles after which the maximum discharge capacity drops by 80% of its initial value. Seven files with ambiguous raw data were excluded from the analysis. The classification machine learning models such as Support Vector Classification, Logistic Regression, and Voting Classifier, all indicate that the most important feature is the slope of discharge capacity fade curve for the initial 100 cycles. Further, there is a commonality between the top features for these models. Random Forest and Decision Tree Classifier models seem to be overfit. Overall, insights into predicting the lifecycle of lithium-ion batteries is provided, which is an essential step in designing and maintaining efficient energy storage systems.

Background

Efficient energy storage solutions are key for renewable energy to succeed. One of the most critical components of an energy storage system is the battery, and predicting the lifecycle of a battery is a crucial factor in its development and maintenance. The lifecycle of a battery refers to the time it takes for a battery to degrade to a level where it can no longer provide sufficient power for its intended use. Battery degradation can occur due to a variety of factors such as temperature, depth of discharge, and number of charge-discharge cycles.

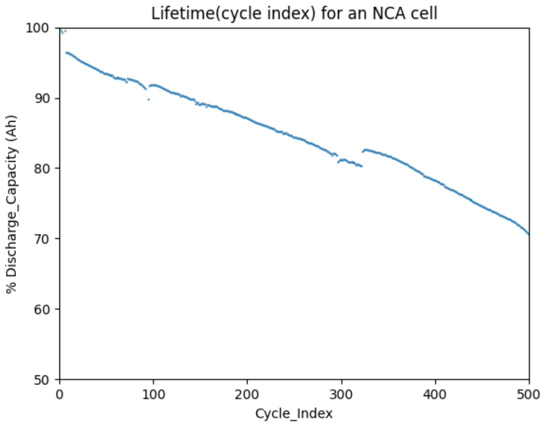
Predicting battery lifecycle is an essential step in the design process. Recent advancements in machine learning techniques have led to the development of accurate predictive models [1]. It takes a significant amount of time to test the charging-discharging cycle of batteries. Here, we develop a model that attempts to predict the lifecycle of a battery only using the initial charging-discharging characteristics. For example, even taking the lowest capacity battery in our dataset, it takes about a year to complete the charging-discharging cycle. We also aim to predict whether a given battery shall have a lifetime above a certain expected threshold.

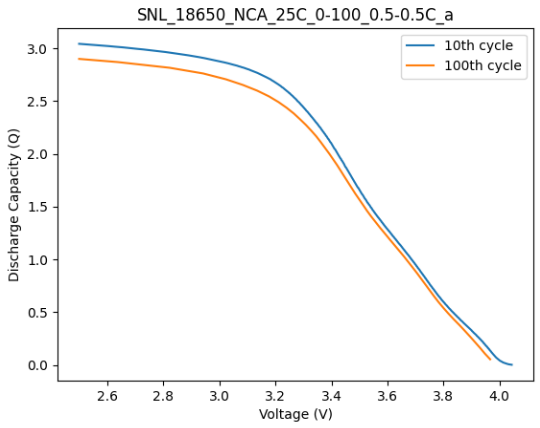
Data description

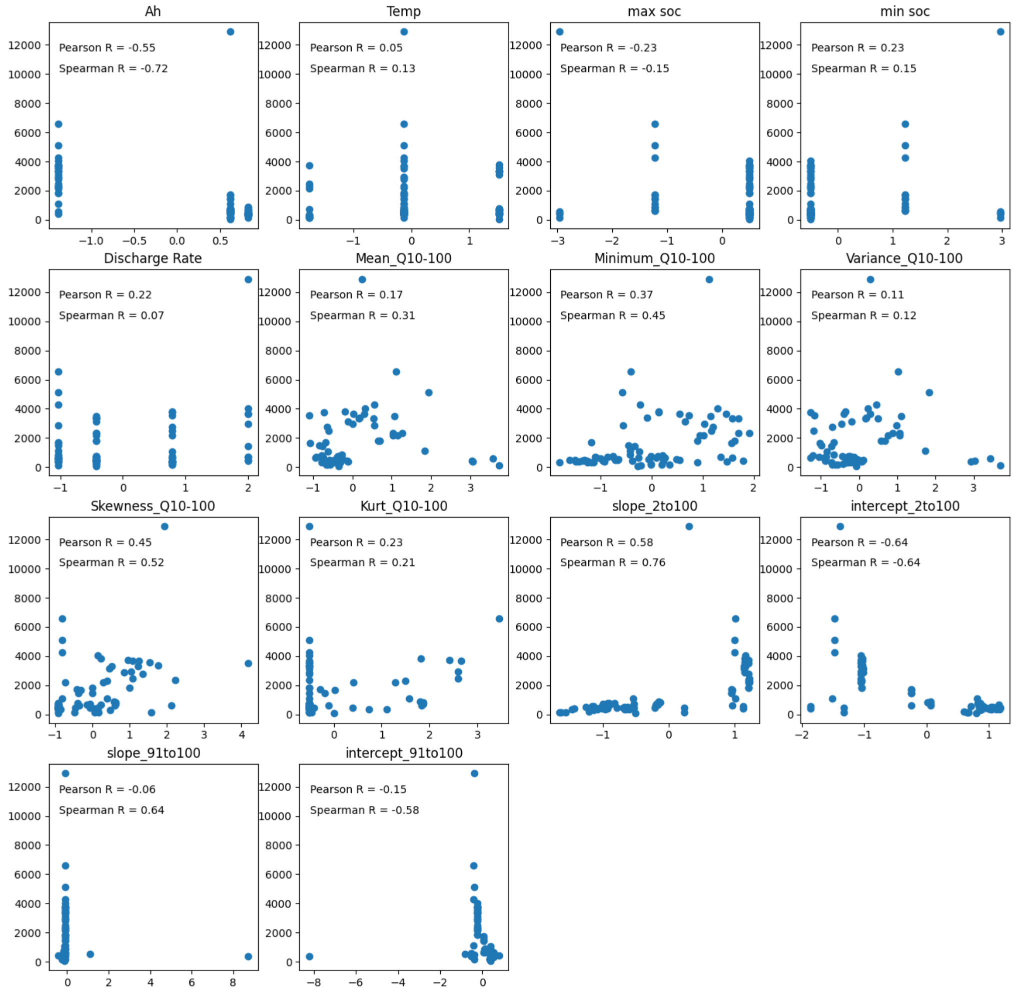
The source of the dataset is Battery Archive Repository [1]. The experiments were conducted at Sandia National Labs, and the data is used in the publication "Degradation of Commercial Lithium-ion Cells as a Function of Chemistry and Cycling Conditions" [2]. In this study, commercial '18650' NCA (Lithium Nickel-Cobalt-Aluminum Oxide), NMC (Nickel Manganese Cobalt), and LFP (Lithium Iron Phosphate) cells are cycled to 80% capacity. In the experiments, each cell is cycled (charged to its 100% capacity, followed by completely discharging it) multiple times until the maximum charging capacity of the cell is dropped to 80% of its initial maximum capacity. This study examines the influence of temperature, depth of discharge, and discharge current on the long-term degradation of commercial cells.
Structure of Raw Data: Data from the cycling tests of a total of 86 cells are analyzed in our machine learning project (30 LFP, 24 NCA, and 32 NMC). A summary 'ba_cell_list_v2' XLSX file reports the experimental conditions such as cathode material, Ah, temperature, maximum SOC, minimum SOC, Charge rate, and Discharge Rate for all 86 cells. This file would create the skeleton of our feature matrix. The structure of raw data consisted of a 'timeseries_data' CSV file and a 'cycle_data' CSV file corresponding to each cell (thus, a total of 172 CSV files). The nomenclature of these files for each cell is as follows:
'SNL_18650_<chemistry>_<temp>_<minSOC>-<maxSOC>_<charge capacity>-<discharge capacity>_b_timeseries_data.csv'
'SNL_18650_<chemistry>_<temp>_<minSOC>-<maxSOC>_<charge capacity>-<discharge capacity>_a_cycle_data.csv'
The 'cycle_data' file has data on variables like minimum current, maximum current, minimum voltage, maximum voltage, charge capacity, discharge capacity, charge energy, and discharge energy recorded during each cycle for the particular cell. The number of rows (cycles) depends on the type of chemistry (LFP/NCA/NMC), and they are rough of the order of 3000, 600, and

400, respectively. The ‘timeseries_data’ file has temporal data of variables such as Current, Voltage, Charge Capacity, Discharge Capacity, Charge Energy, Discharge Energy, Environment Temperature, and Cell Temperature across all the cycles for that particular cell.

Raw Data Visualization:







Methods

Creation of feature matrix:

The 'timeseries_data' file is used to calculate the statistical parameters of the $\Delta Q_{100-10}(V)$ data. For each cell, the discharge capacity (Q) data as a function of Voltage (V) is extracted for the 10th and the 100th cycle (during the discharging portion of the cycle). A spline function fitting procedure is used to generate fit $Q_{100}(V)$ and $Q_{10}(V)$ curves. Further, $\Delta Q_{100-10}(V)$ data array is calculated by subtracting $Q_{10}(V)$ from $Q_{100}(V)$. Five statistical parameters (features), such as minimum, mean, variance, skewness, and kurtosis, are calculated from this $\Delta Q_{100-10}(V)$ data array corresponding to each cell.

The 'cycle_data' file is used to calculate certain features using the capacity fade curves (Discharge capacity as a function of cycle index data). Two linear fits are made on this discharge capacity curve between cycles 2-100 and 91-100, and the respective slopes and intercepts for both fits are used as four features for that particular cell. Three additional columns are added as features to implement one-hot encoding to indicate if the cell has LFP, NCA, or NMC chemistry.

Finally, as the summary file ba_cell_list forms the skeleton of our feature matrix, six additional features (experimental conditions), such as Battery Rated Capacity (Ah), temperature, maximum SOC, minimum SOC, Charge rate, and Discharge Rate, are included.

As discharge capacity decreases as the cells are cycled multiple times, the target value (lifetime of a cell) is also calculated from the cycle_data file. Lifetime is the number of cycles after which the max discharge capacity drops by 80% of its initial value.

There were 7 files (cells) that had ambiguous raw data (3 LFP and 4 NMC cells). The value of discharge capacity did not drop across by any amount/by the sufficient amount for the measured cycle indices. The possible reason for this could be the fact that the experiments are still running presently. These 7 cases were discarded for our analysis. Thus, the feature has a dimensionality of 79 rows (cells) and 18 columns (features). Data cleaning was attempted to remove some outliers but was not executed efficiently (hence included in future work).

Choice of the features:

Cathode and Anode material: These features could be important in understanding the chemical composition and behavior of the battery during cycling, which could impact its lifetime.

Capacity: By including the capacity of a battery as a feature in a machine learning model, the model can take into account the initial capacity of the battery and how it changes over time. For example, a battery with a higher initial capacity may be expected to have a longer lifetime than a battery with a lower initial capacity, all other things being equal.

Temperature: Temperature is a critical factor that affects battery lifetime, so including this feature could help improve the accuracy of the model.

Max SOC and Min SOC: The battery's level of charge, represented as a percentage of its maximum charge, is indicated by the SOC. The maximum and minimum SOC numbers can shed

light on how the battery is being used and how it is performing. The SOC might fluctuate during cycling when the battery is charged and depleted.
Charge rate and Discharge rate: The rate at which the battery is charged or discharged could impact its lifetime. For example, high discharge rates can cause batteries to lose capacity more quickly and have shorter lifespans than batteries that are discharged more slowly.
Mean_Q10-100, Minimum_Q10-100, Variance_Q10-100, Skewness_Q10-100, and Kurt_Q10-100: These features are summary statistical results of the cycle data. It could provide insights into the distribution of the battery's behavior during cycling.
Index_lifetime: This is the cycle number at which the capacity of the battery has dropped to 80% of the initial capacity.
Slope_2to100, Intercept_2to100, Slope_91to100, and Intercept_91to100: These features could represent various linear or nonlinear trends in the battery's behavior during cycling at the beginning (2 to 100 trend) and at later stages (91 to 100 trend).

Implementation of regression models:
Regression machine learning models are implemented to predict the lifetimes of batteries. A preliminary Linear L1 LASSO model was used that yielded poor prediction and accuracy. Thus, non-linear models such as Decision Tree regressor (DT), Gradient Boosting DT, Adaboosting DT, Support Vector, Random Forest, and Artificial Neural Network are used. A 5-fold cross-validation methodology is implemented to calculate the mean and standard deviation of the absolute percentage error in the lifetime of a cell. A stratified 5-fold method is used to represent each LFP, NMC, and NCA chemistry equally when doing cross-validation. Grid search is performed to obtain optimal values of the relevant hyperparameters of each model.

Model	**Hyperparameters tuned: Optimal Value**
Linear L1 LASSO	{'alpha': 100, 'fit_intercept': True}
Decision Tree Regressor	{'criterion': 'absolute_error', 'max_depth': 2, 'max_features': 'sqrt', 'min_samples_split': 2}
Gradient Boosting DT regressor	{'learning_rate': 0.1, 'max_depth': 2, 'n_estimators': 50}
Adaboosting DT regressor	{'learning_rate': 0.1, 'n_estimators': 20}
Support Vector Regressor	{'C': 1, 'epsilon': 0.01, 'gamma': 1, 'kernel': 'poly'}
Random Forest Regressor	{'criterion': 'absolute_error', 'max_depth': 40, 'min_samples_split': 4, 'n_estimators': 2}
Artificial Neural Network	{'activation': 'identity', 'alpha': 1e-05, 'hidden_layer_sizes': 4}

Implementation of classification models:
Standard LFP, NMC, and NCA batteries have expected lifetimes of 3000, 600, and 400 cycles, respectively. We have used classification models to predict whether a battery has a lifetime above

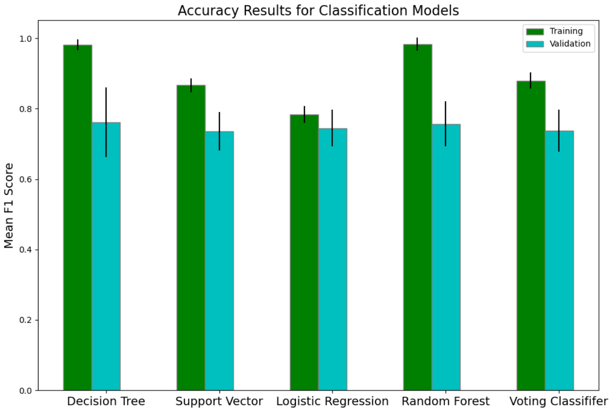
or below the listed threshold cycles. Similar to the regression models, a 5-fold cross-validation methodology is implemented to calculate the mean and standard deviation of the absolute percentage error in the lifetime of a cell. Grid search is performed to obtain optimal values of the relevant hyperparameters of each model.

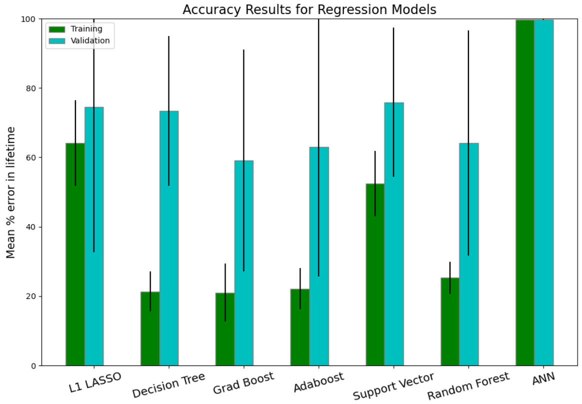
Model	Hyperparameters tuned: Optimal Parameter
Decision Tree Classifier	{'criterion': 'gini', 'max_depth': 2, 'max_features': 'sqrt'}
Support Vector Classifier	{'C': 0.1, 'gamma': 0.1, 'kernel': 'sigmoid'}
Logistic Regression	{'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}
Random Forest Classifier	{'criterion': 'entropy', 'max_depth': 6, 'min_samples_split': 2, 'n_estimators': 20}

Results

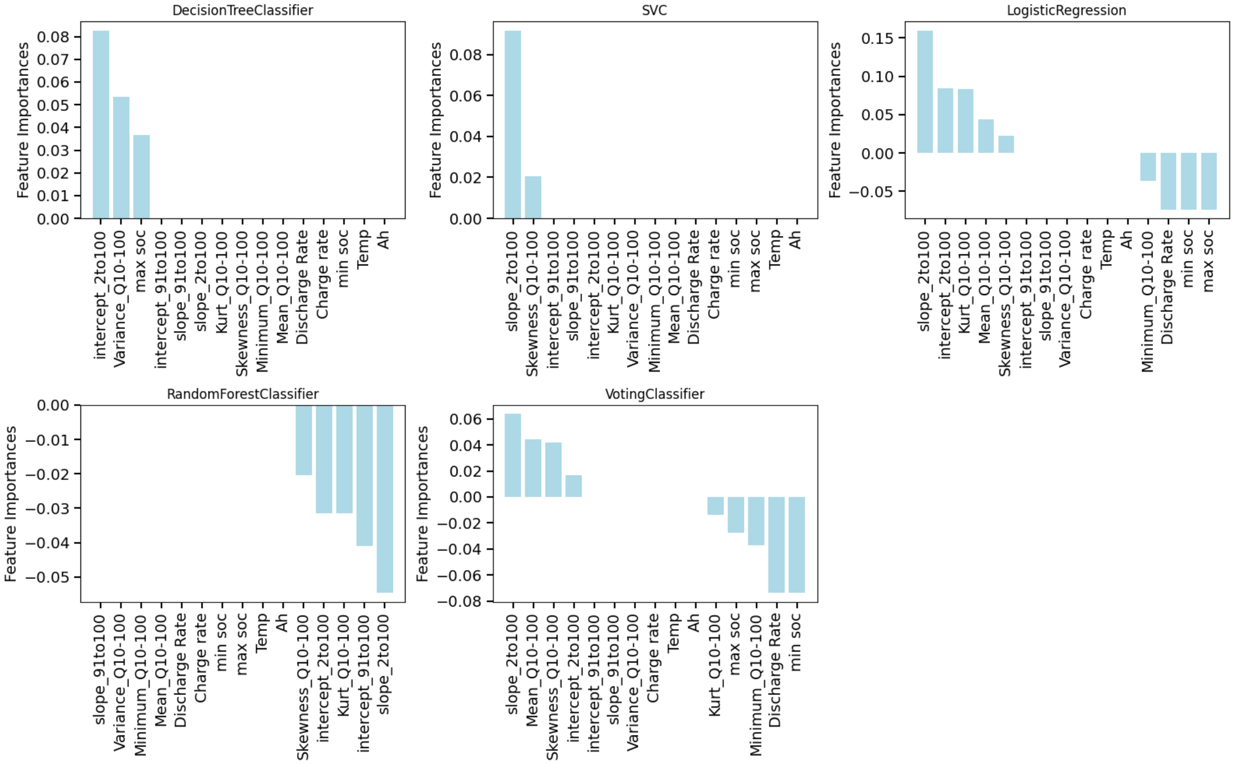
Regression Model	% error (training)	% error (validation)
Linear L1 LASSO	64.169 +/- 12.3399	74.531 +/- 41.9188
Decision Tree Regressor	21.301 +/- 5.7639	73.415 +/- 21.5766
Gradient Boosting DT regressor	20.975 +/- 8.3715	59.103 +/- 31.9192
Adaboosting DT regressor	22.103 +/- 5.9497	63.030 +/- 37.4022
Support Vector Regressor	52.407 +/- 9.4600	75.870 +/- 21.4563
Random Forest Regressor (n_est = 8)	25.260 +/- 4.6401	64.154 +/- 32.5178
Artificial Neural Network (n_layers = 6)	99.713 +/- 0.0176	99.749 +/- 0.0468

Nonlinear Classifier Model	Mean f1 score (training)	Mean f1 score (validation)
Decision Tree Classifier	0.982 +/- 0.0157	0.761 +/- 0.0991
Support Vector Classifier	0.867 +/- 0.0198	0.736 +/- 0.0544
Logistic Regression	0.784 +/- 0.0243	0.745 +/- 0.0524
Random Forest Classifier (n_est = 20)	0.983 +/- 0.0186	0.757 +/- 0.0642
Voting Classifier (log_reg + svm + dtc)	0.880 +/- 0.0223	0.737 +/- 0.0600





Model	**F1 score obtained on the reserved 20% Test data**
Decision Tree Classifier	0.7059
Logistic Regression	0.8235
Support Vector Classifier	0.7778
Random Forest Classifier	0.8421
Voting Classifier	0.7059



Discussion

For the regression models, the mean % error is very high for linear models, which indicates the possibility of data being non-linear. The mean % training error decreases by a significant amount as we switch to non-linear models like Decision Tree, Gradient Boosting DT, Adaboosting DT, and Random Forest regressor, but the validation error still remains high. The mean % error values for Artificial Neural Network is excessively high, and this could be attributed to a large number of hyperparameters in addition to the limited dataset. For the classification models, we observe f1 scores on the test data between 0.73-0.76 for all the models. The classification models were trained and tested on battery life cycle data, and the results show that the Nonlinear Classifier Model and Random Forest Classifier achieved the highest mean f1 score (training). The Support Vector Classifier and Logistic Regression models had lower mean f1 scores (training and validation) compared to the Nonlinear Classifier and Random Forest models. The Voting Classifier, which combined the Logistic Regression, Support Vector, and Decision Tree models, shows moderate performance compared to the other models. Feature importance indicates that the slope_2-100 is the most important feature for all models except Decision Tree Classifier and Random Forest Classifier. There are commonalities between the top features identified by SVC, Logistic Regression, and Voting Classifier models. It appears that Decision Tree and Random Forest models are possibly overfitting, as indicated by a very high mean training F1 score.

Future work:
After every 3% capacity loss, Electrochemical impedance Spectroscopy (EIS) test is performed. Thus, this leads to a few data points which shoot the discharge capacity by a large value before a cycle just after the EIS test starts. The data can be cleaned by excluding the outlier data points. This process could improve the feature values obtained from slopes and intercepts from linear fits. The ΔQ(V) between the 10th and 100th cycle may not capture the non-linear behavior of discharging for LFP cells as the ΔQ100-10(V) values are very low, as observed in data visualization. ΔQ(V) for different combinations of cycles (other than the 100th and 10th cycle) can be taken to calculate new statistical features. ΔQ(V) can be calculated for the charging part of the cycle as well, in addition to the discharge portion. The cathode and anode materials have different chemistries (LFP/NCA/NMC), and these can featurized based on the physical and chemical properties of the material.

References

1) Severson, K.A., Attia, P.M., Jin, N. *et al.* Data-driven prediction of battery cycle life before capacity degradation. *Nat Energy* 4, 383–391 (2019). DOI

2) Mayilvahanan, K. S., Takeuchi, K. J., Takeuchi, E. S., Marschilok, A. C., West, A. C. *et al.* Supervised Learning of Synthetic Big Data for Li-Ion Battery Degradation Diagnosis. *Batteries and Supercaps*, Vol 5, Issue 1 (2021) DOI

3) Yuliya Preger *et al.* Degradation of Commercial Lithium-Ion Cells as a Function of Chemistry and Cycling Conditions. *J. Electrochem. Soc.* 167 120532 (2020). DOI

4) Jie Xiong, Tong-Xing Lei, Da-Meng Fu, Jun-Wei Wu, Tong-Yi Zhang. Data driven discovery of an analytic formula for the life prediction of Lithium-ion batteries. *Progress in Natural Science: Materials International*, Volume 32, Issue 6, 793-799, (2022). DOI

5) Source of dataset: Battery Archive: https://www.batteryarchive.org/