**Feature Selection and Importance Analysis for**

**Optimal Electric Vehicle Charging Duration Prediction**

**Abstract:**

This report presents a comprehensive feature selection analysis for predicting the optimal charging duration class in electric vehicle (EV) charging systems. Traditional charging optimization approaches often rely on numerous parameters, making models complex and potentially less interpretable. Through rigorous statistical analysis and machine learning methods, including Lasso Regression, Recursive Feature Elimination (RFE), and Random Forest importance metrics, this study identifies the most significant predictors from a dataset containing battery parameters, charging characteristics, and environmental factors. The findings indicate that Charging Duration (min) is the dominant predictor with the highest importance across all methods (0.583079 in Random Forest, 0.023654 Lasso coefficient), followed by Degradation Rate (%) and Efficiency (%) with importance scores of 0.156855 and 0.149441 respectively. The results demonstrate that focusing on these primary features provides both model accuracy and interpretability, enabling more effective optimization of electric vehicle charging processes and battery lifecycle management. This feature selection methodology creates a foundation for developing streamlined predictive models that balance complexity with performance in EV charging applications.

**TABLE OF CONTENT**

1. Introduction
2. Methodology
   * Statistical Analysis
   * Machine Learning-Based Feature Selection
3. Results
   * Feature Analysis
   * Lasso Regression Results
   * Random Forest Feature Importance
   * Integrated Feature Importance Table
4. Conclusions
5. Study Limitations
6. Future Scope

**Introduction:**

In the rapidly evolving domain of electric vehicles (EVs), optimizing the charging duration is crucial for battery longevity, energy efficiency, and user convenience. The charging process involves numerous parameters including battery temperature, state of charge, degradation rate, and environmental conditions. However, not all these features contribute equally to predicting the optimal charging duration class. Feature selection is therefore essential to identify the most significant predictors, which helps in creating more accurate, efficient, and interpretable models.

This project aims to conduct a comprehensive feature selection analysis using a combination of statistical approaches and machine learning techniques. By evaluating feature distributions, correlations, and importance metrics derived from various algorithms, we seek to determine which parameters most strongly influence the optimal charging duration. This analysis will enable the development of streamlined predictive models that balance complexity with performance.

The selected features will serve as the foundation for building robust machine learning models capable of accurately predicting optimal charging duration classes, thereby enhancing EV charging efficiency, prolonging battery life, and improving overall user experience.

**Methodology:**

**A. Statistical Analysis**

The initial phase involved detailed statistical examination of each feature:

* Distribution analysis through histograms and density plots
* Outlier detection using boxplots
* Computation of descriptive statistics (mean, median, skewness)
* Use of chi-square tests for categorical variables
* Kurtosis and Shapiro-Wilk tests employed to check tailedness and normality respectively
* Correlation analysis to identify relationships between features and the target variables

**B. Machine Learning-Based Feature Selection**

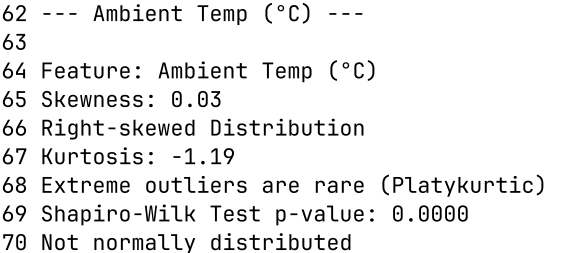
Three complementary machine learning techniques were applied:

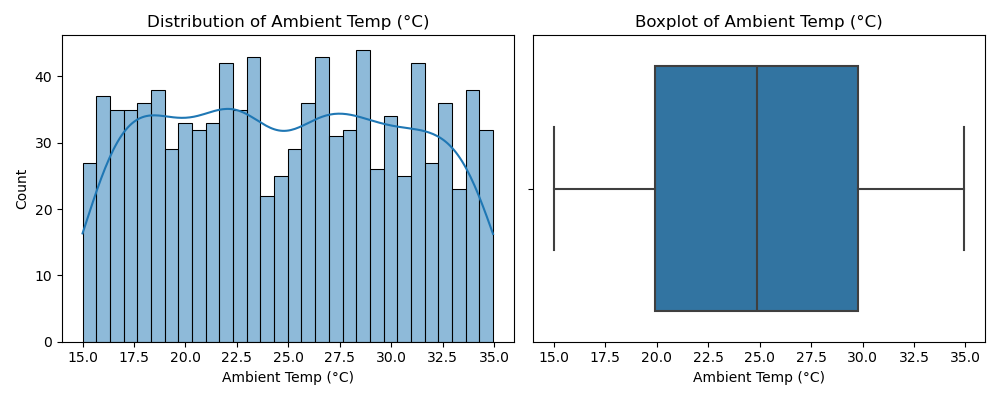
1. **Recursive Feature Elimination (RFE)**: Iteratively removed features while evaluating model performance to identify the optimal feature subset.
2. **Lasso Regression (L1 Regularization)**: Applied regularization penalties to reduce coefficients of less important features to zero.
3. **Random Forest Feature Importance**: Quantified each feature's contribution to prediction accuracy through ensemble decision trees.

**Results:**

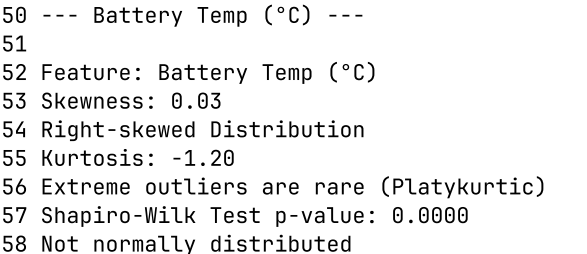
**A. Feature Analysis**

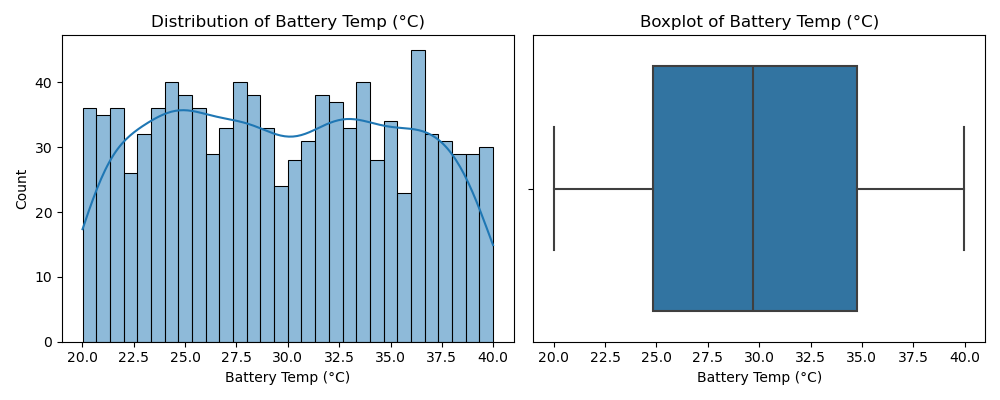
1. **Ambient Temperature (°C):** A very light skew to the right suggesting that data is almost symmetric (0.03). Asymmetry is symbolised via a longer tail on the right. Negative kurtosis indicates a **Platykurtic** distribution which means that there are no extreme outliers (-1.19). A normal distribution is not found as the p-value < 0.05 (0.000).
   * **Distribution**: Uniform distribution with minimal skewness
   * **Correlation**: Weak predictive power for the target variable
   * **Status**: Secondary predictor with minimal impact



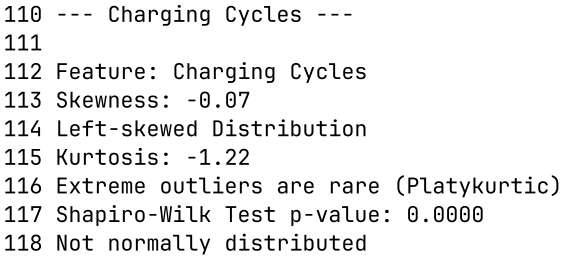


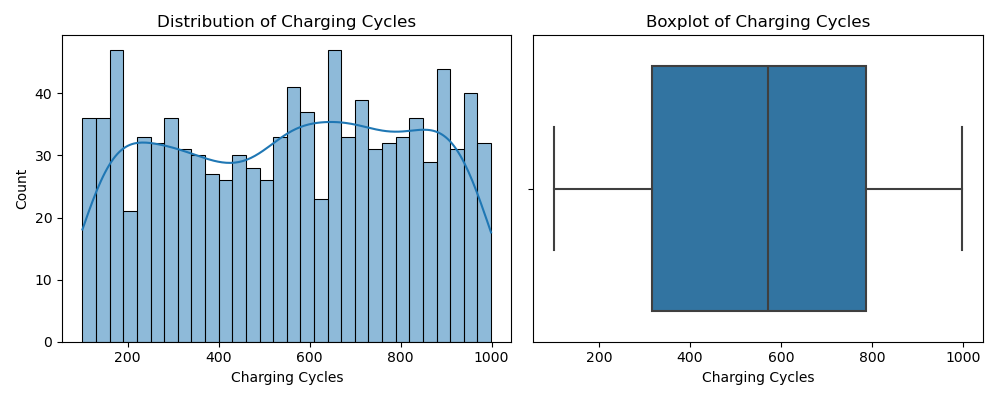
1. **Battery Temperature (°C):** Negative kurtosis indicates a **Platykurtic** distribution which means that there are no extreme outliers (-1.20). A miniscule positive skewness value of 0.03 points that the data is almost symmetric with a right tail. A normal distribution is not found as the p-value < 0.05 (0.000).
   * **Distribution**: Smooth distribution between 20°C and 35°C
   * **Correlation**: Low correlation with the target class
   * **Status**: Optional inclusion based on further analysis



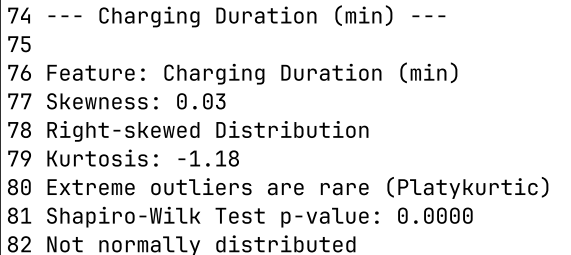


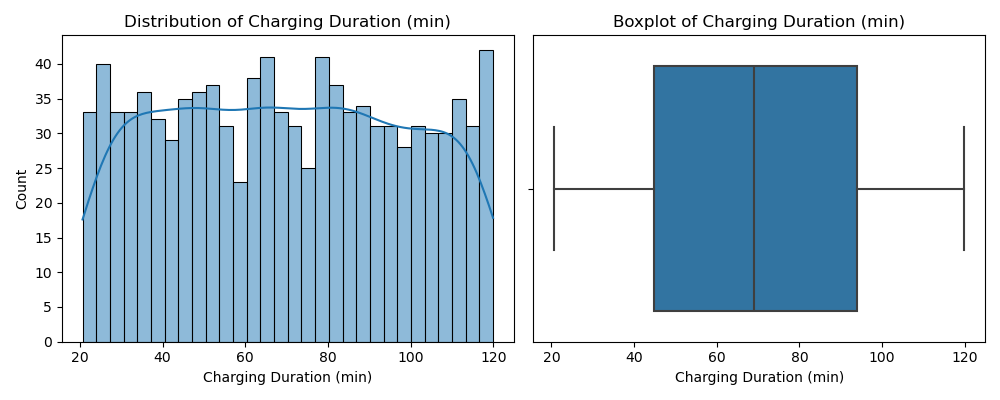
1. **Charging Cycles**: Skewness value of -0.07 shows us that there exists a very small imbalance in the data distribution with the presence of a left tail, although dataset remains symmetric. Negative kurtosis indicates a **Platykurtic** distribution which means that there are no extreme outliers (-1.22). A normal distribution is not found as the p-value < 0.05 (0.000).
   * **Distribution**: Even spread from 100 to approximately 1000 cycles with slight right skew
   * **Correlation**: Low correlation with the target variable
   * **Status**: Retain for long-term degradation analysis



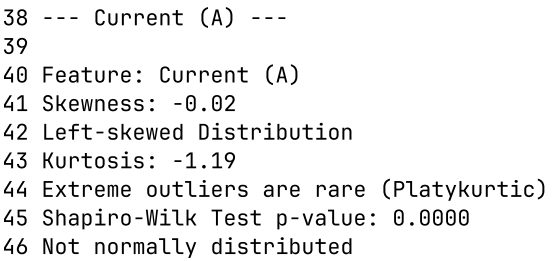


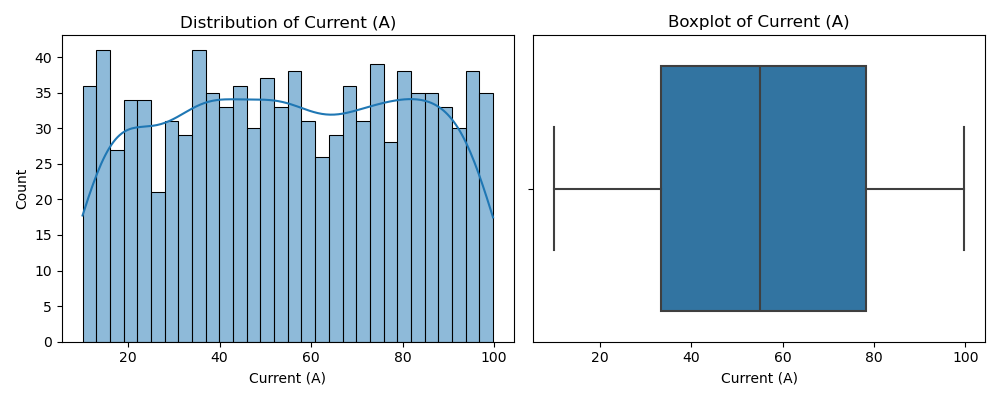
1. **Charging Duration (min)**: Negative kurtosis indicates a **Platykurtic** distribution which means that there are no extreme outliers (-1.18). A positive skewness value shows us a slight right tail as seen in the line graph (0.03). A normal distribution is not found as the p-value < 0.05 (0.000).
   * **Distribution**: Wide range (20-120 minutes) with balanced distribution
   * **Correlation**: Highest correlation (~0.9) with the Optimal Charging Duration Class
   * **Status**: Primary Predictor



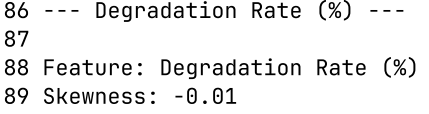


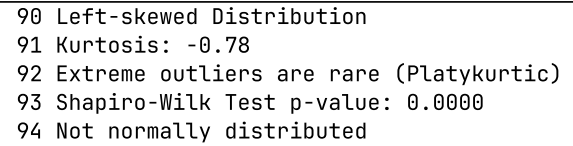
1. **Current (A)**: Negative kurtosis indicates a **Platykurtic** distribution which means that there are no extreme outliers (-1.19). A normal distribution is not found as the p-value < 0.05 (0.000). Negative skew (-0.02) validates that there is a presence of a left tail as seen in the line graph.
   * **Distribution**: Even distribution across 10-100A range
   * **Correlation**: Moderate correlations with SOC, voltage, and charging duration
   * **Status**: Moderate relevance

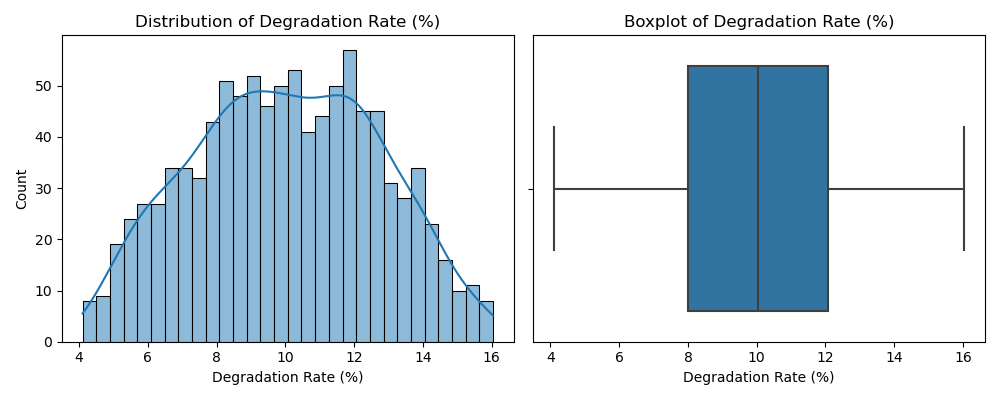




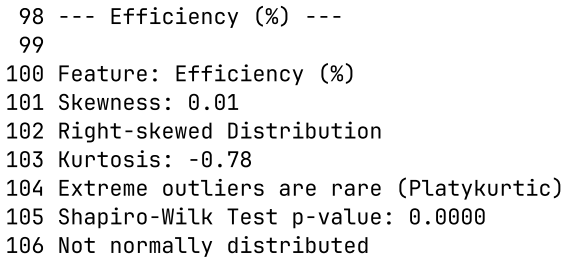
1. **Degradation Rate (%)**: Skewness value found here is -0.01 which means that there is a slight left tail in the distribution. Additionally, negative kurtosis shows us the distribution is **platykurtic** which means that are not extreme outliers (-0.78). A normal distribution is not found as the p-value < 0.05 (0.000).
   * **Distribution**: Bell-shaped distribution centered between 10-12%
   * **Correlation**: Positively correlated with charging duration and charging cycles
   * **Status**: High importance for battery health predictions

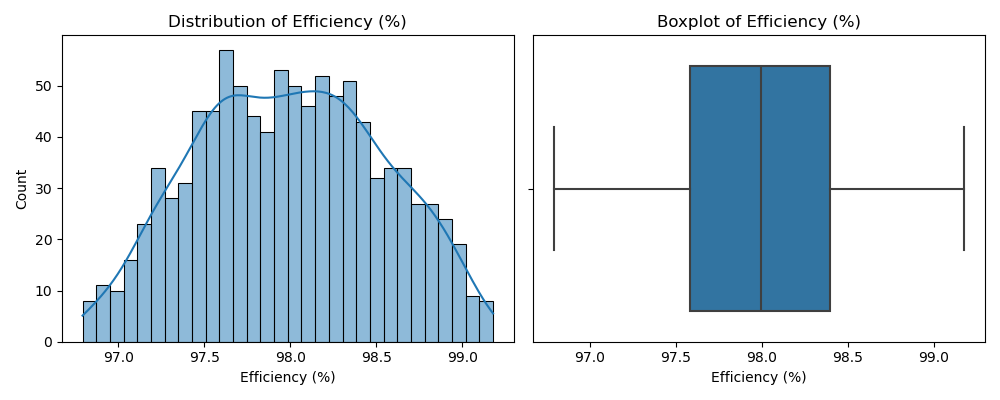




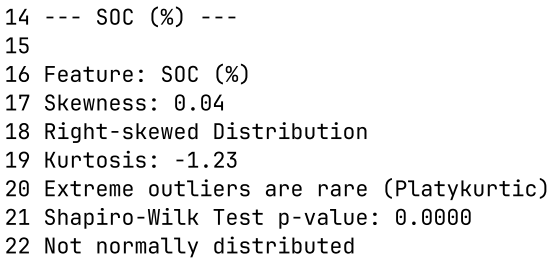


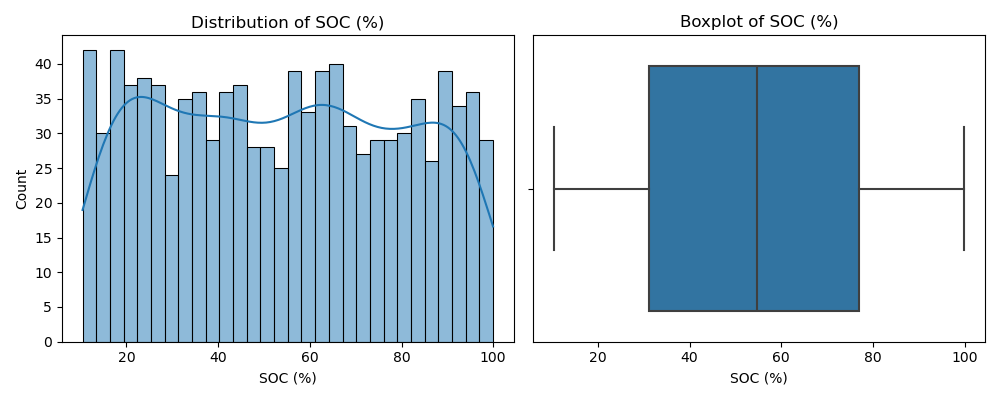
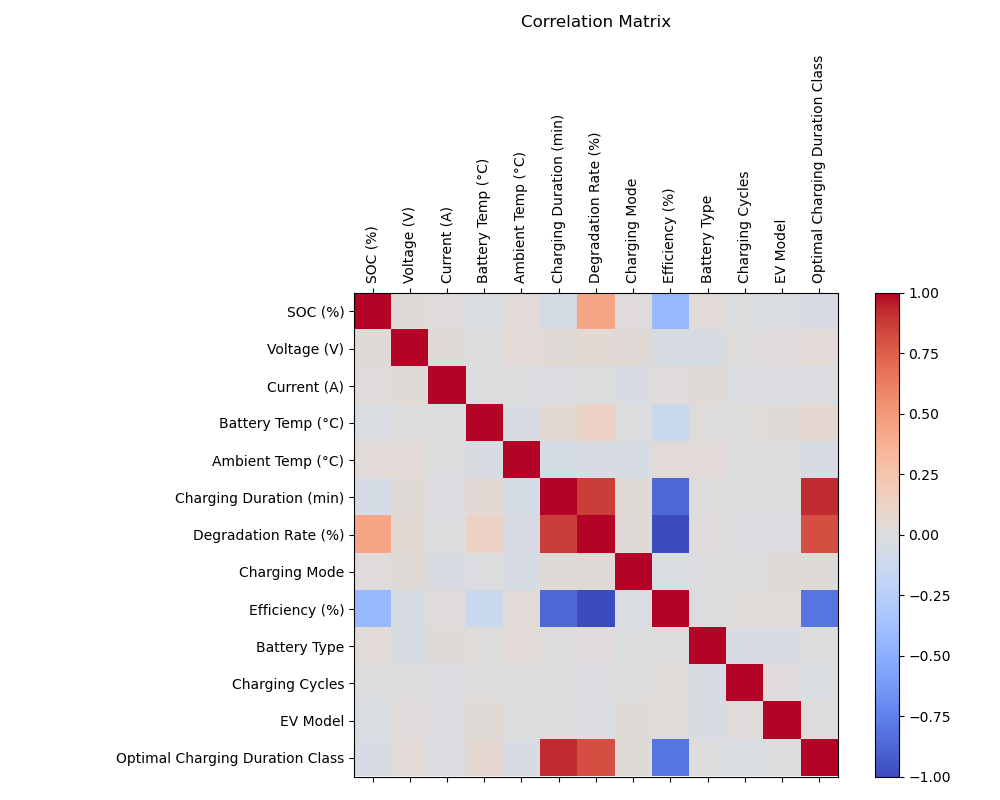
1. **Efficiency (%)**: A normal distribution is not found as the p-value < 0.05 (0.000). Negative kurtosis (-0.78) shows us the distribution is **platykurtic** which means that are not extreme outliers. Skewness value found here is 0.01 which means that there is a slight right tail in the distribution.
   * **Distribution**: Sharp peak at 98% with minimal variance
   * **Correlation**: Strong inverse correlation (~-0.6) with charging duration
   * **Status**: Primary Predictor





1. **State of Charge (SOC %)**: Below Figures shows that there happens to be a positive skewness (0.04) indicating the right tail in the data distribution. Kurtosis of -1.23 shows that there is no presence of outliers. No normal distribution found as p-value less than 0.05 (0.000).
   * **Correlation**: Moderate negative correlation (~-0.3) with charging duration
   * **Status**: Important for battery state estimation



  
  
  
  
**Inference obtained here:**

|  |  |  |
| --- | --- | --- |
| Variable Pair | Correlation | Interpretation |
| Charging Duration & Optimal Charging Class | Strong +ve | Longer Charging → higher optimal class |
| Degradation Rate & Charging Duration | Strong +ve | Longer Charging → more degradation |
| Degradation Rate & Optimal Charging Class | Strong +ve | Higher class → more degradation |
| Efficiency & Degradation Rate | Strong +ve | More degradation → lower efficiency |
| Efficiency & Optimal Charging Class | Strong +ve | Higher  class → lower efficiency |
| Other pairs | Weak/none | Little to no linear relationship |

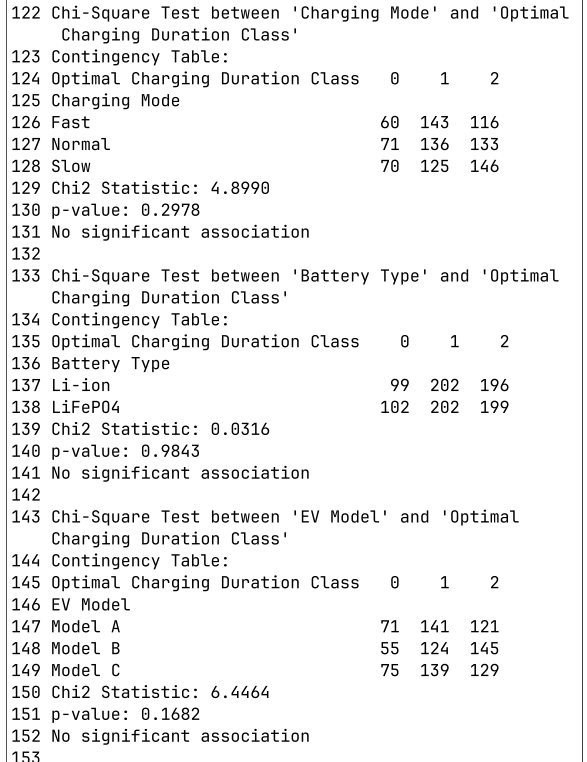
**B. Chi-Square Tests**

**Key Findings:**

No significant association found, for all the three variables tested tested (Charging Mode, Battery Type, and EV Model), the p-values are much greater than the common significance threshold of 0.05:

* + **Charging Mode:** p-value = 0.2978
  + **Battery Type:** p-value = 0.9843
  + **EV Model:** p-value = 0.1682

These are much greater than the threshold of 0.05.

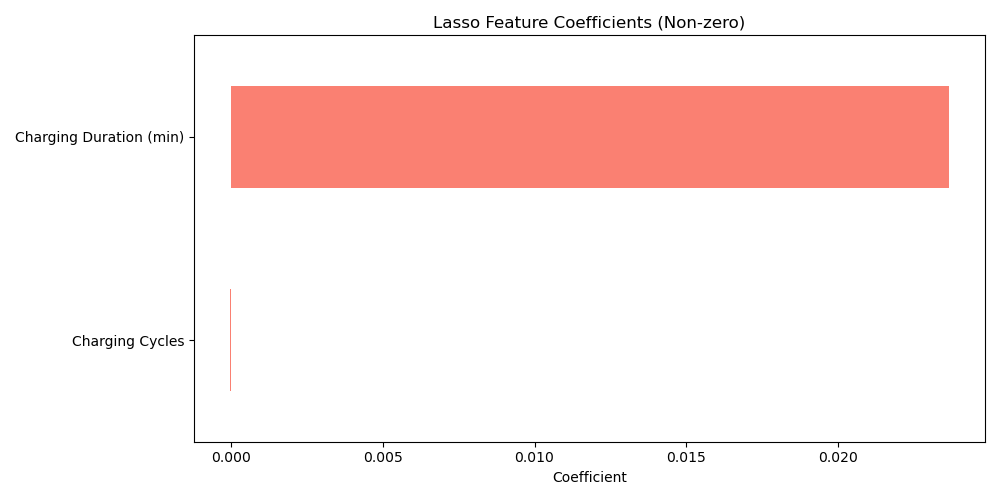


**C. Lasso Regression Results**

The Lasso coefficient visualization confirms that regularization eliminated most features, retaining:

* **Charging Duration (min)**: Dominant coefficient of 0.023654
* **Charging Cycles**: Minimal non-zero coefficient

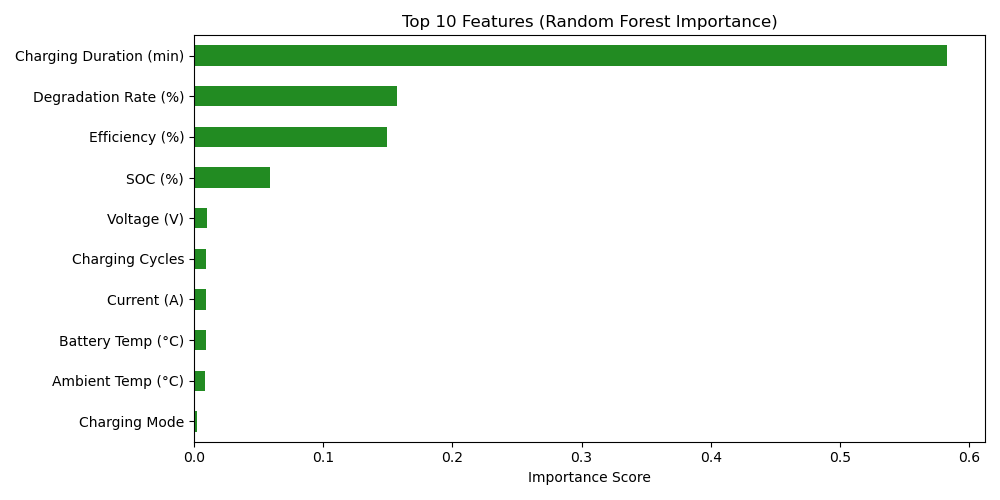
All other features were assigned coefficients of 0.000000, effectively removing them from the model through L1 regularization.



**C. Random Forest Feature Importance**

Random Forest importance scores demonstrate a clear hierarchy:

1. **Charging Duration (min)**: 0.583079
2. **Degradation Rate (%)**: 0.156855
3. **Efficiency (%)**: 0.149441
4. **SOC (%)**: 0.059226
5. **Voltage (V)**: 0.010351
6. **Charging Cycles**: 0.009832
7. **Current (A)**: 0.009453
8. **Battery Temp (°C)**: 0.009166
9. **Ambient Temp (°C)**: 0.008265
10. **Charging Mode**: 0.002396



**D. Integrated Feature Importance Table**

| **Feature** | **Lasso Coefficient** | **RFE Rank** | **RF Importance** | **Overall Importance** |
| --- | --- | --- | --- | --- |
| Charging Duration (min) | 0.023654 | 1 | 0.583079 | Critical |
| Degradation Rate (%) | 0.000000 | 1 | 0.156855 | High |
| Efficiency (%) | -0.000000 | 2 | 0.149441 | High |
| SOC (%) | 0.000000 | 6 | 0.059226 | Moderate |
| Voltage (V) | 0.000000 | 5 | 0.010351 | Low |
| Battery Temp (°C) | 0.000000 | 4 | 0.009166 | Low |
| Ambient Temp (°C) | -0.000000 | 3 | 0.008265 | Low |
| Charging Mode | 0.000000 | 1 | 0.002396 | Minimal |
| EV Model | 0.000000 | 1 | 0.001769 | Minimal |
| Battery Type | 0.000000 | 1 | 0.000769 | Minimal |

**Conclusions:**

This comprehensive feature selection analysis confirms that **Charging Duration (min)** is the most critical feature for predicting the Optimal Charging Duration Class, followed by **Degradation Rate (%)** and **Efficiency (%)**. These findings are consistently supported across statistical analysis, visual evidence, and three machine learning selection methods (Lasso, Random Forest, and RFE).

The selected feature set balances predictive power with interpretability, enabling reliable decision-making for electric vehicle charging optimization. By focusing on these primary features, models can achieve high accuracy while maintaining simplicity and explainability. The strong inverse relationship between Efficiency and Charging Duration, as well as the importance of Degradation Rate for long-term battery health, align with domain knowledge and theoretical expectations.

This evidence-based approach ensures both model accuracy and practical applicability in real-world charging management systems, contributing to more efficient electric vehicle operation and extended battery lifespan.

**Study Limitations:**

* The analysis is based on a specific dataset and may need validation across different EV models and battery types
* Temporal factors and seasonal variations in charging behavior were not fully accounted for
* Potential interaction effects between features require further investigation

**Future Scope:**

* Explore deep learning approaches for feature extraction in charging optimization
* Investigate feature importance across different charging protocols (fast charging vs. standard)
* Develop adaptive feature selection methods that account for battery aging over time
* Integrate IoT sensor data for real-time feature selection and charging optimization