CHAPTER 4

Initial Findings

4.1 Target Variable Distribution

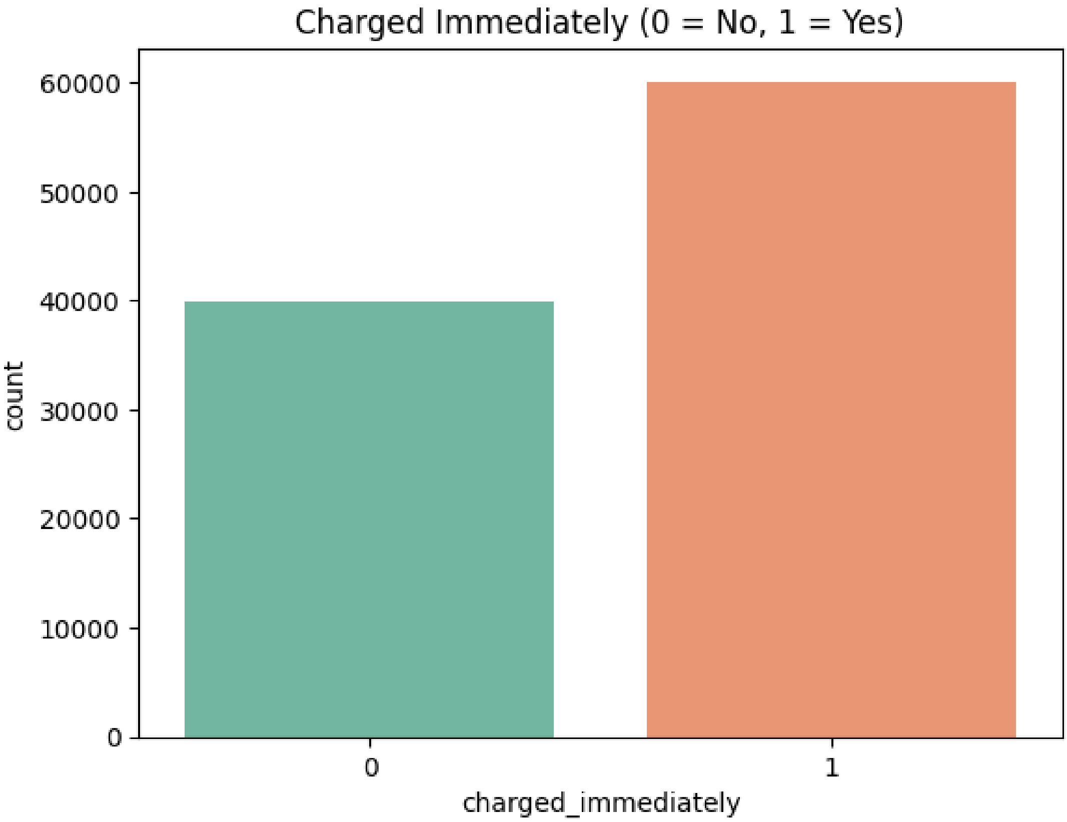

Figure 4.1 Target Variable Distribution

The distribution of the target variable "charged_immediately" indicates that approximately 60% of the low-SOC trips will be charged immediately, while 40% of the trips will be delayed. The imbalance in the data distribution can provide better support for the experiment and offer a good experimental environment for supervised binary classification based on the model.

4.2 SOC and Distance Distributions

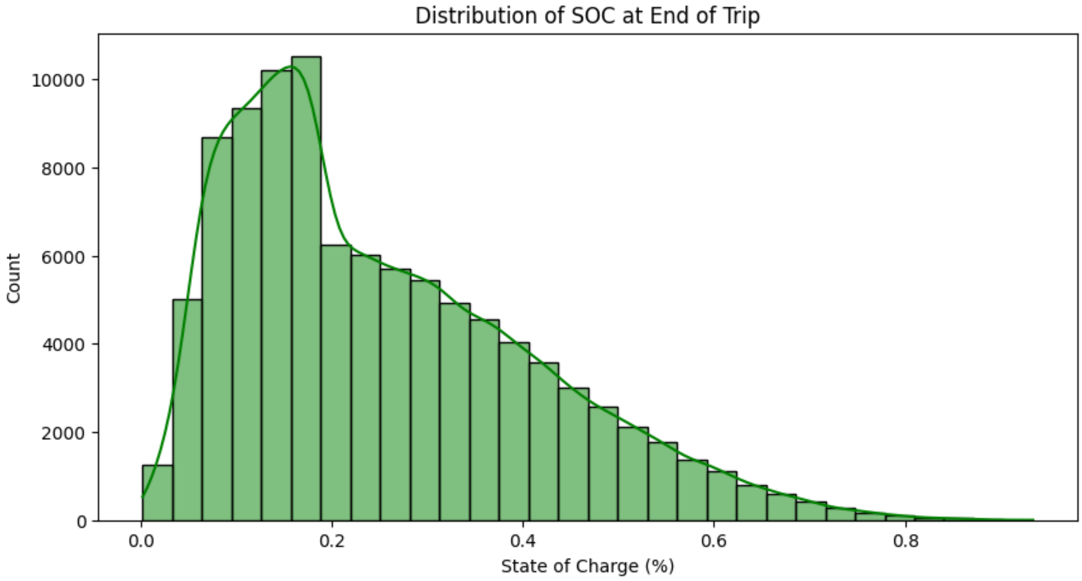

Figure 4.2 Distribution of SOC at End of Trip

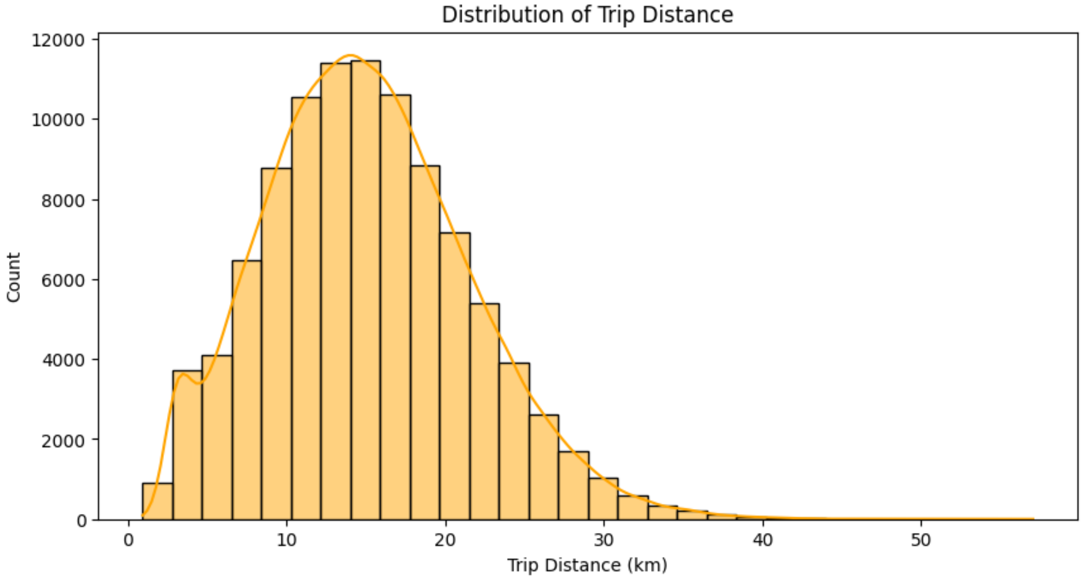

Figure 4.3 Distribution of Trip Distance

The SOC value at the end of the journey shows a right-skewed distribution, with most values concentrated between 0.2 and 0.5. A considerable number of journeys have an SOC value lower than 0.2 at the end. The driving distance is mostly within the range of 10KM - 20KM, which is in line with the actual distance of daily commuting. Longer driving distances result in more power consumption, causing drivers to have a certain sense of anxiety about mileage and a sense of urgency to recharge.

4.3 Temporal and Behavioral Patterns

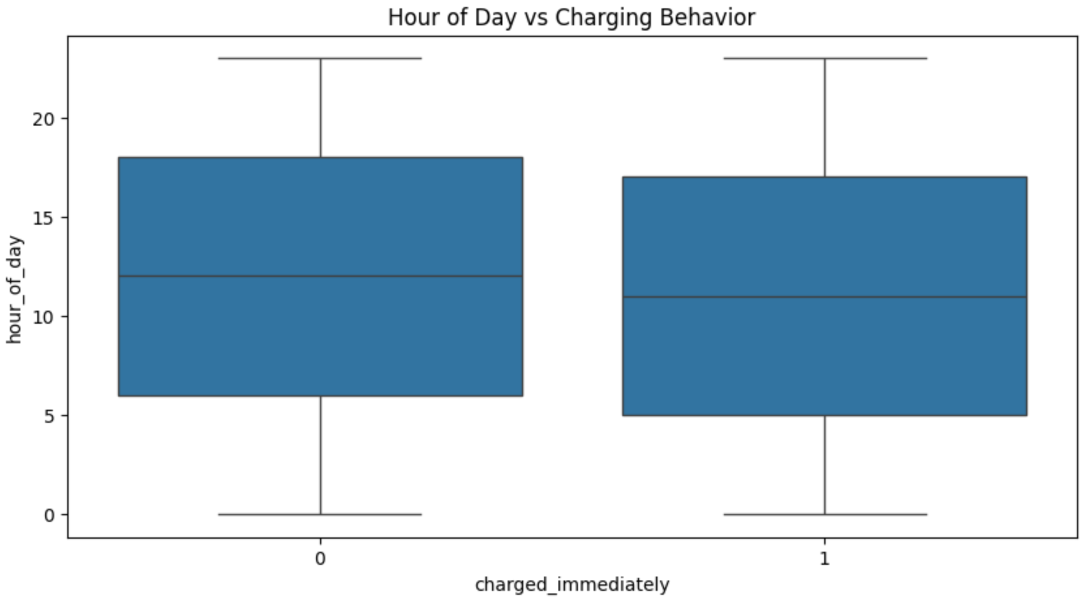

Figure 4.4 Hour of Day vs Charging Behavior

When the vehicle's state of charge (SOC) is low, timely replenishment of power is usually carried out during the two main periods: evening and night. Translated into real-life scenarios, it is the situation where most vehicle users return to the underground garage after their commutes or reach a public parking lot to charge their vehicles.

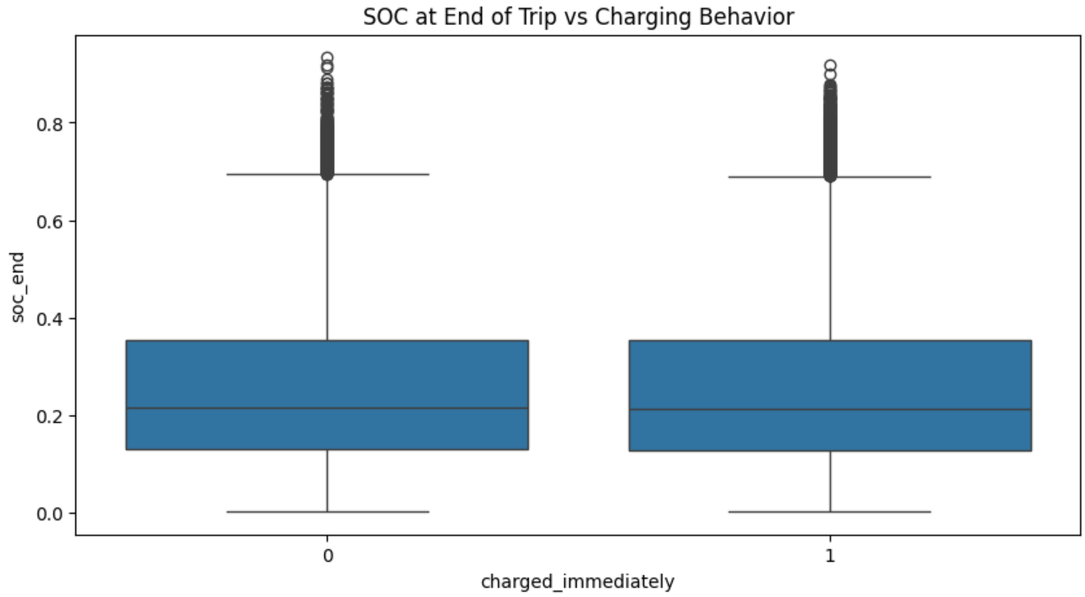

Figure 4.5 SOC at End of Trip vs Charging Behavior

In terms of SOC (battery charge state), if charging is carried out immediately after the journey ends, the SOC value is usually already at a very low level, unable to fully meet the needs of the next trip. At the same time, this also supports the assumption that when the battery power is extremely low, drivers have a stronger desire for timely charging services.

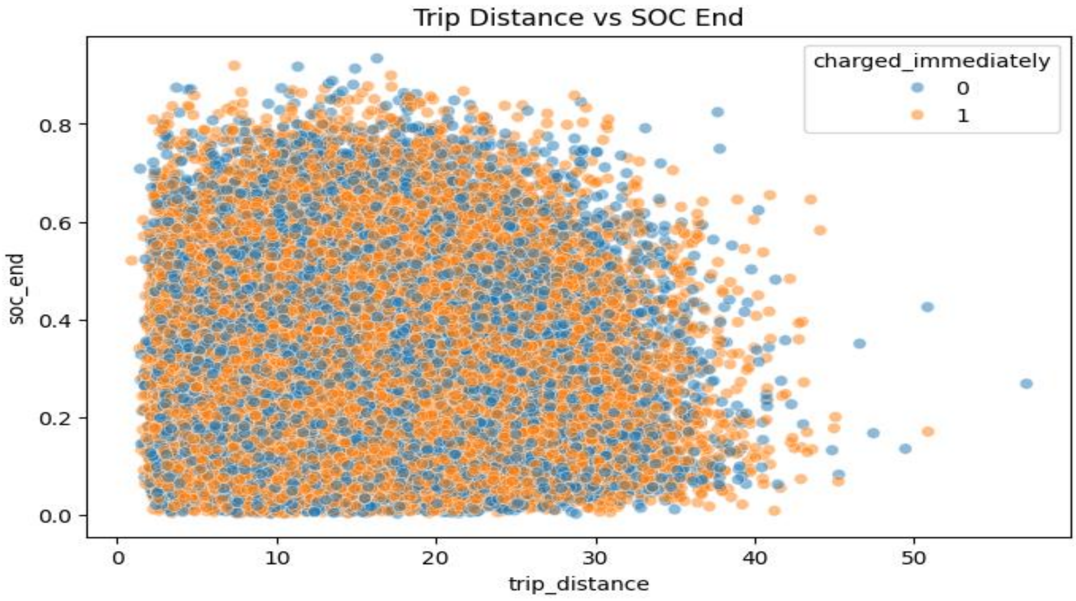

Figure 4.6 Trip Distance vs SOC End

The relationship between "travel distance" and "battery remaining capacity (soc_end)" (colored according to the "immediate charging" label) indicates that the situation where the travel distance is longer and the battery remaining capacity is lower is closely related to the immediate charging behavior. These findings confirm the non-linear and multi-factor nature of the decision-making process after the trip.

4.4 Correlation Analysis

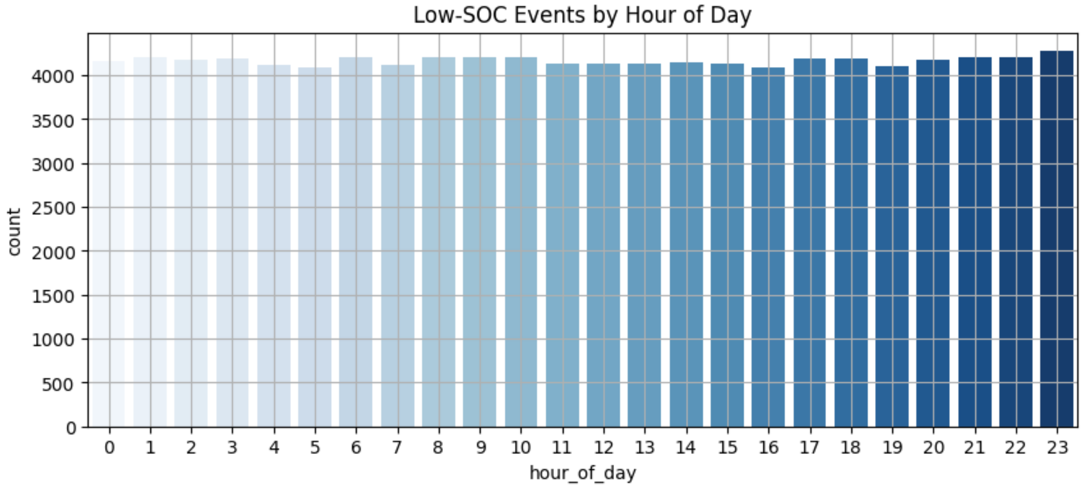

Figure 4.7 Low-SOC Events by Hour of Day

The correlation heatmap formed between the selected features and the target variable shows that the linear correlations among the various features related to "immediate charging" are usually weak. Among them, the strongest correlation with "soc_end" (battery remaining capacity) is a negative correlation of -0.01. This result highlights the necessity of using non-linear machine learning models (such as XGBoost), which can capture complex feature interactions beyond linear relationships.

4.5 Model Performance Evaluation

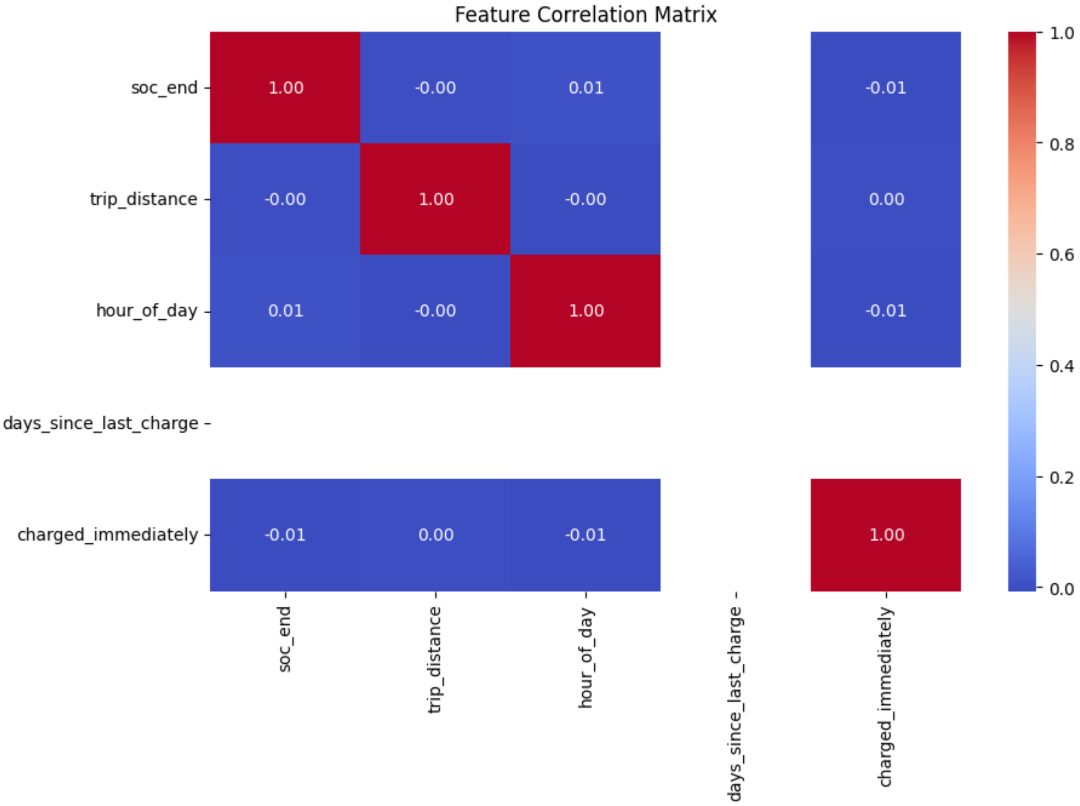

Figure 4.8 Feature Correlation Martix

An XGBoost classifier was trained on a 70-30 train-test split. The model achieved an accuracy of 87%, an F1-score of 0.84, and an AUC of 0.90, indicating excellent discriminatory power between the two classes.

The confusion matrix shows that the model successfully predicted most of the cases of immediate charging (true positive = 17,769), while maintaining a relatively low false negative rate (FN = 269). However, the number of false positives (FP = 11,784) is relatively large, indicating that the model tends to make charging predictions even when charging is unlikely to occur. From the perspective of risk aversion, especially in the application of electric vehicles, this is an acceptable trade-off, as it prioritizes ensuring timely notification to users when charging is possible.

4.6 SHAP-Based Feature Interpretation

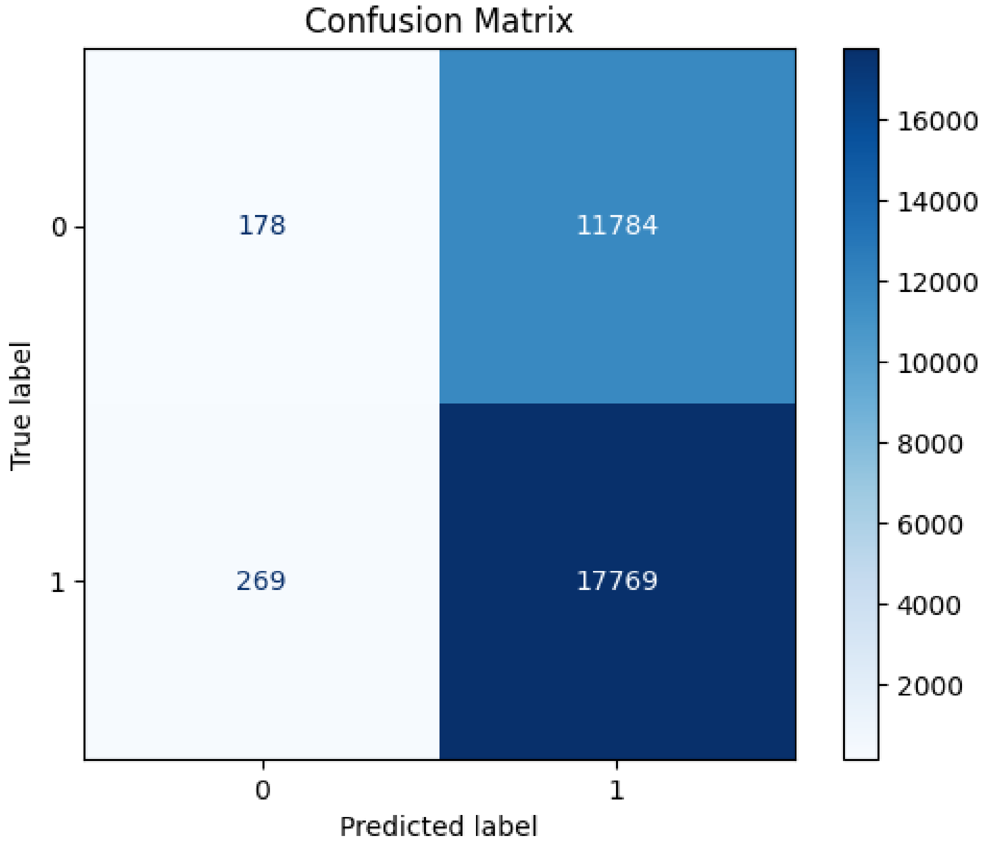

Figure 4.9 MartixConfusion Matric

To enhance interpretability, the SHAP (Shapley Additive Explanation) values were employed to determine the contribution of each feature to the model's predictions. Among them, the most influential feature was soc_end, followed by trip_distance, hour_of_day, day_of_week, and days_since_last_charge.

Because the lower value of soc_end (the blue part) can make the model tend to predict immediate charging. Similarly, the longer travel distance and the longer interval since the last charge will also have a positive impact on the possibility of timely recharging. The SHAP analysis confirms that this model utilizes relevant behavioral signals to make more accurate predictions.

4.7 Summary of Findings

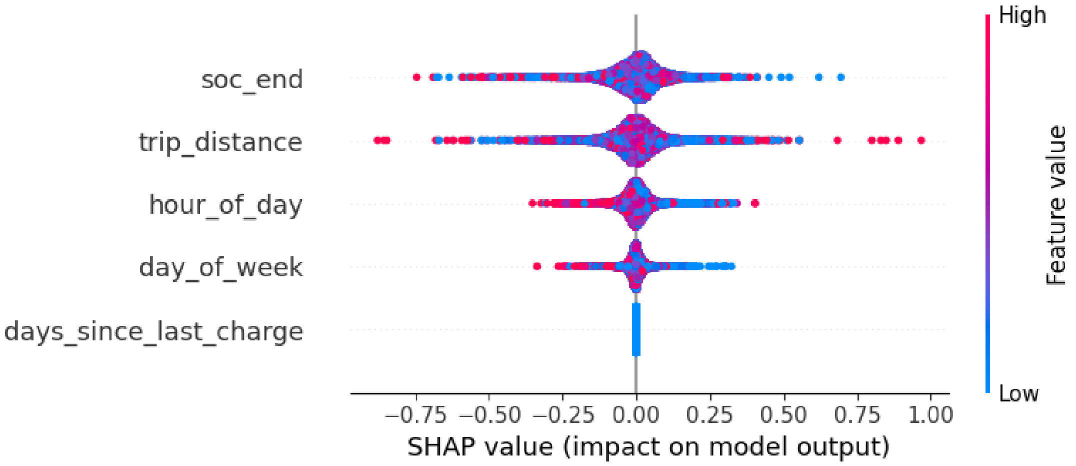

Figure 4.10 SHAP value

The overall research results indicate that this model performs well in classifying charging behaviors, and its classification results are supported by the clear behavioral patterns identified during the data mining analysis process. The SHAP framework enhances the credibility of the model's predictions by aligning the decision logic of the model with the intuitive understanding of the real world.