CS 562 Course Project Work

BU ID's: 362971, 364394

Electric Vehicle Charging Duration Prediction Project Report

1. Project Summary

This project focuses on predicting the charging duration of Electric Vehicles (EVs) using a dataset containing 1,320 samples of EV charging session data. The primary goal is to develop a model that can accurately forecast the time it takes to charge an EV based on various features such as energy consumption, charging rate, vehicle details, and environmental factors like temperature. We explore two machine learning algorithms: **Linear Regression, MLP Regressor, K Neighbors Regressor** and **Random Forest Regressor** to identify the best model for this prediction task.

2. Feature Engineering

Feature engineering plays a crucial role in improving model performance. The following steps were taken during the feature engineering process:

1. **Handling Missing Values**: Missing values were imputed using mean imputation. Specifically, columns like 'Energy Consumed (kWh)', 'Charging Rate (kW)', and 'Distance Driven (since last charge) (km)' were imputed by replacing missing entries with their respective column, means.
2. **Normalization**: Several numerical columns such as 'Energy Consumed (kWh)', 'Charging Rate (kW)', and others were standardized using Z-scores and then normalized using a StandardScaler. This ensures that the features are on a similar scale, preventing models from being biased toward higher-value features.
3. **Time-Based Feature Extraction**: The dataset contained timestamps for charging start and end times. We extracted new features such as 'Charging Start Hour', 'Charging Start Day', and 'Charging Start Month' to capture time-based patterns in charging behavior.
4. **Additional Derived Features**:
   - **Charging Efficiency**: Calculated as the ratio of energy consumed to charging duration.
   - **Energy per Charge %**: The energy consumed divided by the difference in State of Charge percentages.
   - **Distance per kWh**: Distance driven per unit of energy consumed.
   - **Temperature Adjusted Consumption**: Adjusted energy consumption considering temperature deviations from 20°C, providing insights into the impact of environmental factors on charging behavior.
5. **Categorical Feature Encoding**: Categorical features like 'Time of Day', Charger Type', and 'Vehicle Model' were encoded using one-hot encoding, transforming them into numerical values for use in machine learning models.
6. **Irrelevant Feature Removal**: Non-predictive features like 'User ID' and 'Charging Station ID' were dropped to reduce complexity and improve model accuracy.

3. ML Algorithm

For this project, we employed two machine learning algorithms:

- **Linear Regression**: A simple yet powerful model for regression tasks that assumes a linear relationship between the independent variables and the target variable (charging duration). Linear Regression was used as a baseline model.
- **Random Forest Regressor**: A more complex ensemble learning model that builds multiple decision trees and averages their outputs to improve prediction accuracy. Random Forest is less sensitive to outliers and can capture non-linear relationships in the data.
- **MLP Regressor (Multi-layer Perceptron Regressor):** MLP Regressor is a type of artificial neural network model that consists of multiple layers of interconnected neurons
- **K-Nearest Neighbors Regressor (KNN):** KNN is a non-parametric, instance-based learning algorithm used for regression tasks. It works by predicting the target variable for a new data point based on the average of the target values of the 'K' nearest data points in the feature space.

Both models were trained on the same set of features, and their performance was evaluated using metrics such as **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **R² Score**.

4. Data Set

The dataset used in this project is the **EV Charging Patterns dataset** which contains 1,320 samples with features like 'Energy Consumed (kWh)', 'Charging Rate (kW)', 'Vehicle Model', and 'Charging Duration (hours)' (target variable). The dataset includes both numerical and categorical features.

- **Dependent Variable**: 'Charging Duration (hours)' is the target variable we aim to predict.
- **Independent Variables**: These include features such as 'Energy Consumed (kWh)', 'Charging Rate (kW)', 'Temperature (°C)', 'State of Charge (Start %)', 'Distance Driven (since last charge) (km)', 'Vehicle Age (years)', and various time-based and engineered features.

5. Results and Discussion

5.1 Model Performance

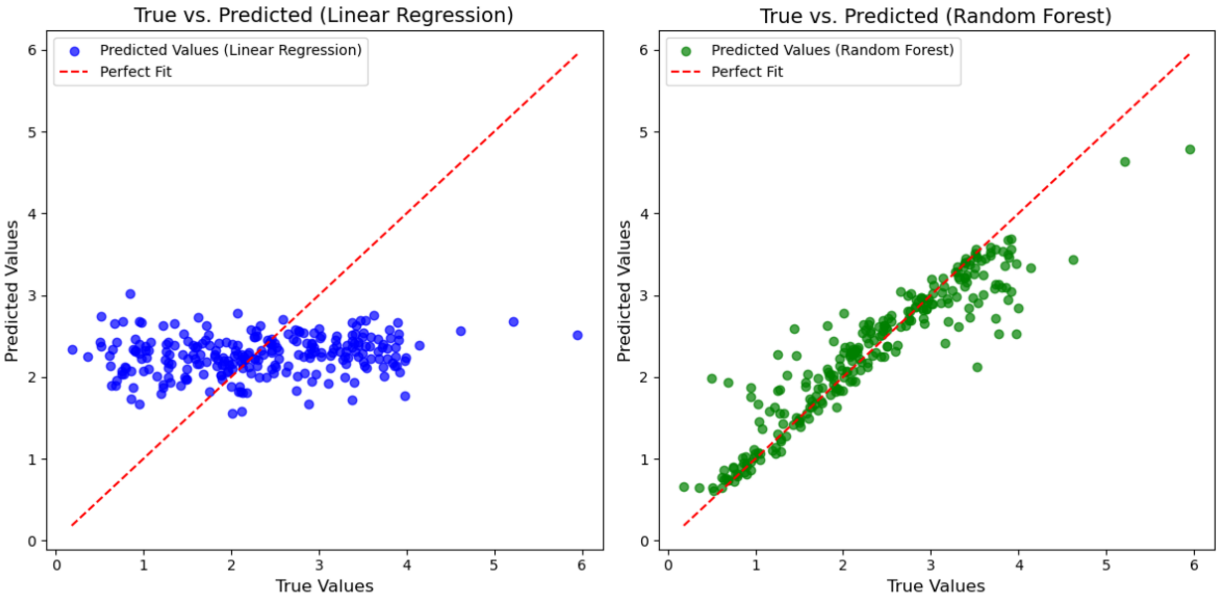
The performance of the models was evaluated using the following metrics:

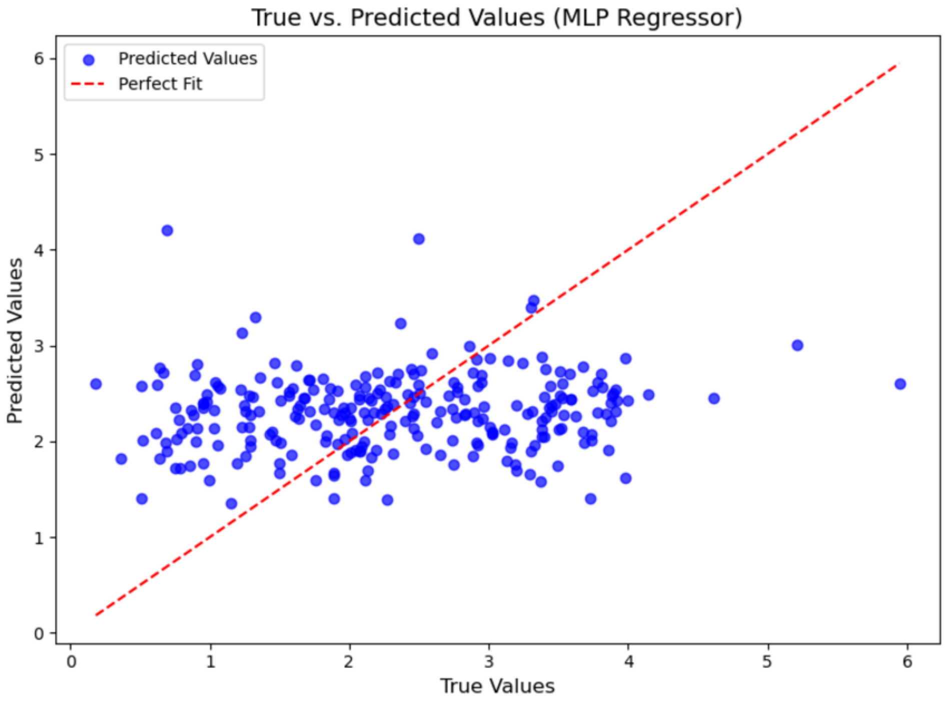
	Linear Regression	**Random Forest Regressor**	**MLP Regressor**	**K-Nearest Neighbors Regressor**
MAE	0.83	0.25	0.86	0.87
MSE	1.03	0.14	1.12	1.20
RMSE	1.01	0.38	1.05	1.09
R² Score	0.02	0.86	0.06	0.14
CV MSE	1.17	0.18	2.08	1.30

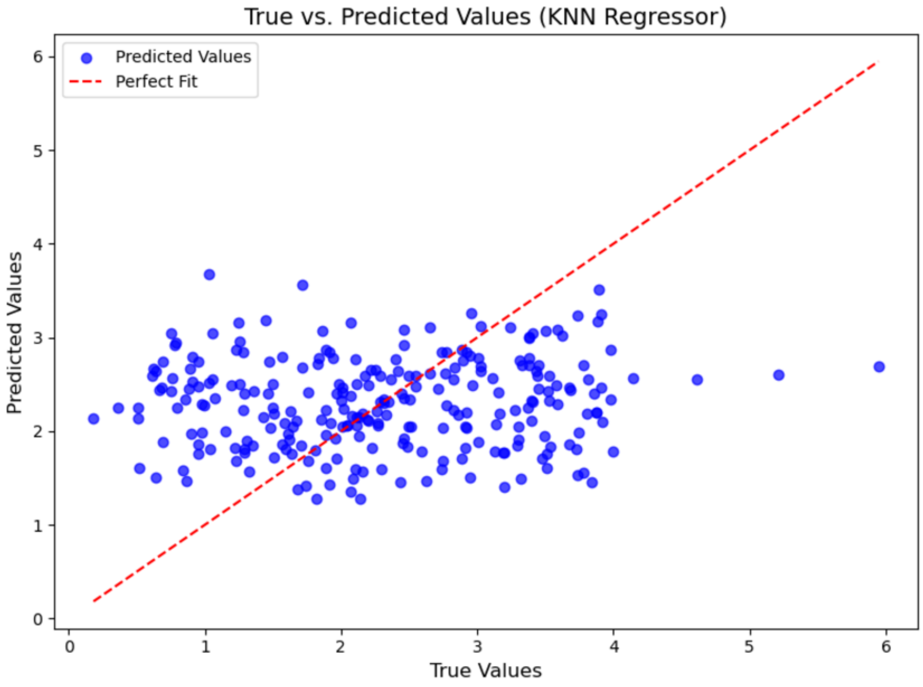
As evident from the metrics, the **Random Forest Regressor** outperforms the **Linear Regression** model and others by a significant margin. The **R² Score** for Random Forest is much higher, indicating that it explains 86% of the variance in the data, whereas Linear Regression explains only 2%. The Random Forest model also produces lower MAE, MSE, and RMSE values, indicating better prediction accuracy.

5.2 Visual Analysis

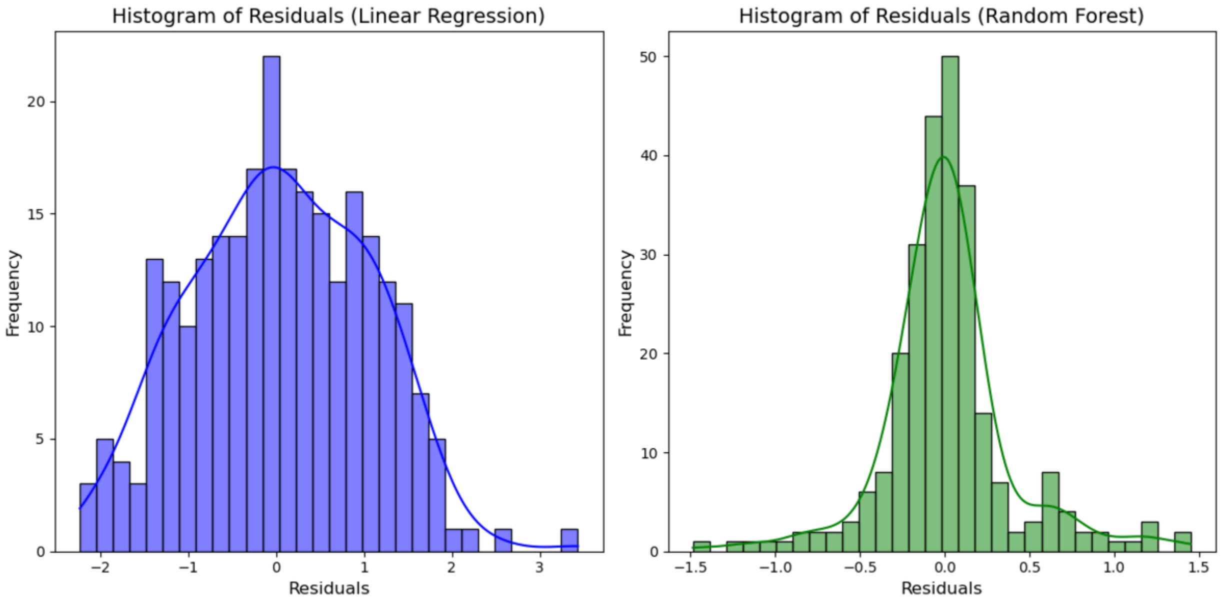
Scatter Plot for True vs. Predicted Values: The scatter plot for the Random Forest model shows that the predicted values are much closer to the true values, indicating better accuracy. In contrast, the Linear Regression model's predictions are more scattered.




True vs. Predicted Values (MLP Regressor)
Predicted Values
Perfect Fit
Predicted Values
True Values
0
1
2
3
4
5
6
0
1
2
3
4
5
6


True vs. Predicted Values (KNN Regressor)
Predicted Values
Perfect Fit
Predicted Values
True Values
0
1
2
3
4
5
6
0
1
2
3
4
5
6

Histograms of Residuals: The residual histogram for the Random Forest model is more symmetrical and centered, suggesting fewer biases and better overall accuracy. The Linear Regression model's residuals, while roughly normal, show a slight right skew, indicating overestimation of some charging durations.



5.3 Conclusion

The **Random Forest Regressor** was the best performing model for predicting EV charging duration. With a high R² score of 0.86, it demonstrated the ability to capture complex relationships between the features and the target variable. In comparison, **Linear Regression, MLP and KNN** struggled with this task, yielding a low R² scores. Based on the evaluation metrics and visual results, we conclude that Random Forest is the most suitable model for this dataset and task.