Final Project Report: Predicting Electric Vehicle User Types from Charging Data

---

1. Introduction

In this project, we aim to predict the type of electric vehicle (EV) user based on charging session data. The target variable is User Type, a multiclass categorical column with categories such as *Commuter*, *Casual Driver*, and *Long-Distance Traveler*. This is a multiclass classification task.

We selected this topic because of growing interest in EV infrastructure, sustainability, and smart energy systems. By understanding how different types of users behave at charging stations, we can support better planning for locations, pricing, and load distribution.

The dataset was obtained from Kaggle and contains 1,320 rows and 20 columns, each row representing one EV charging session. It includes numeric, categorical, and temporal features such as:

Battery capacity, charging duration, and energy consumed

Charger type, time of day, and temperature

Battery start/end percentages, driving distance, and more

The dataset reflects commerce and environmental sectors, particularly sustainable transportation systems.

---

2. Exploratory Data Analysis

The dataset had:

- 1320 rows, each a single EV charging event
- 10 numerical and 10 categorical features

Some missing values were present in 3 columns:

- Energy Consumed (kWh)
- Charging Rate (kW)
- Distance Driven (km)

These were handled using mean imputation.

Key Findings:

- A correlation heatmap revealed no strong correlation ($|r| > 0.3$) with charging duration.
- Most users were labeled as Commuter, indicating class imbalance (but not severe).
- Histograms and boxplots showed that Long-Distance Travelers tend to charge longer and consume more energy.

Categorical columns like Charger Type, Time of Day, and Day of Week also showed variation by user type. This suggested that combining time-of-day behavior with charging metrics could be predictive.

---

3. Data Preparation and Preprocessing

A single, consistent preprocessing pipeline was applied using ColumnTransformer.

- Numerical features: Imputed with mean and standardized
- Categorical features: Imputed with mode and one-hot encoded
- Dropped columns: User ID, timestamps, and location info (to prevent data leakage)
- No text or cyclic features were used
- No class balancing (like SMOTE) was applied because the distribution wasn't highly skewed

The data was split into:

- 60% training, 20% validation, and 20% test
  This split was reused for every model, ensuring consistency.

---

4. Model Training and Results

We trained six classification models, following course guidelines:

Model	Accuracy	F1 Score	Best Parameters
Logistic Regression	0.758	0.751	C=0.1
Decision Tree	0.708	0.704	max_depth=10
Random Forest	0.788	0.781	n_estimators=100, max_depth=20
Gradient Boosting	0.792	0.787	Early stopping (no grid search used)
SVM (SVC)	0.774	0.765	C=1.0, gamma='scale'
Neural Network (MLP)	0.770	0.760	hidden_layer_sizes=(100, 50)

Key Points:

- Gradient Boosting performed best overall
- Random Forest was close and gave useful feature importance
- SVM performed better than Decision Tree and Logistic Regression
- Neural Network (MLPClassifier) performed decently with just two hidden layers

---

5. Visualizations and Insights

- Decision Tree: We visualized a tree of depth 3 to show key decision splits
- Random Forest: Feature importances showed Charging Duration, State of Charge, and Energy Consumed as top predictors
- Confusion Matrix (SVM): Revealed confusion between Casual Driver and Commuter labels, hinting at overlapping behavior

These insights can help optimize charging policies, like dynamic pricing based on predicted user type.

---

6. Conclusions and Next Steps

This project demonstrated that it's feasible to accurately classify EV user types using structured session data.

Best Model:

Gradient Boosting (Highest accuracy and F1, simple config with early stopping)

Next Steps:

- Add cyclic encoding for Time of Day
- Test class balancing using SMOTE
- Use grid search on neural networks (tune learning rate, dropout)
- Explore ensemble voting or model stacking
- Consider using geolocation or external weather data

This model could help EV infrastructure companies to anticipate user needs and allocate resources more effectively.