ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN MỨC NĂNG LƯỢNG KHI SẠC CỦA MỘT SỐ DÒNG XE ĐIỆN PHỔ BIẾN TẠI MỸ

Nhóm 21			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Nguyễn Cao Thắng	22521329	CNTT
2	Võ Phi Thân	22521323	CNTT
3	Võ Minh Quyền	22521227	CNTT
4	Võ Anh Quân	22521192	CNTT

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Đề tài này thực hiện phân tích các yếu tố ảnh hưởng đến mức độ tiêu thụ năng lượng khi sạc của một số dòng xe điện phổ biến tại Mỹ. Trong đề tài này, chúng tôi đã sử dụng các thư viện và các module trong các thư viện như pandas, numpy, matplotlib [4], seaborn [5], sklearn [2], scipy, geopy, plotpy [6] để phân tích và trực quan dữ liệu; mô hình hồi quy đa thức và các thư viện hỗ trợ tương ứng. Thông qua việc sử dụng các công cụ hỗ trợ, nhóm chúng tôi đã thực hiện thành công những nhiệm vụ gồm: tiền xử lý dữ liệu, phân tích thăm dò và trực quan hóa, đưa ra một mô hình phù hợp để dự đoán cho bộ dữ liệu.

Bộ dữ liệu phân tích được tham khảo tại Electric Vehicle Charging Patterns [1]. Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế và không dựa trên đề tài nào khác.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu này cung cấp các mẫu dữ liệu về các thông số đáng lưu ý trong việc sạc xe điện (EV) cũng như thói quen di chuyển của chủ phương tiện. Bộ bao gồm 1.320 mẫu dữ liệu về các phiên sạc, với các thông số như năng lượng tiêu thụ, thời gian sạc và thông tin chi tiết về xe, ... Mỗi mục dữ liệu ghi lại nhiều khía cạnh khác nhau của việc sử dụng EV, cho phép thực hiện phân tích chuyên sâu và xây dựng các mô hình dự đoán.

Giải thích các thuộc tính trong bộ dữ liệu:

User ID	Mã định danh duy nhất cho mỗi người dùng
Vehicle Model	Mẫu xe điện đang được sạc (ví dụ: Tesla Model 3, Nissan Leaf).
Battery Capacity (kWh)	Tổng dung lượng pin của xe tính bằng kilowatt-giờ.
Charging Station ID	Mã định danh duy nhất cho trạm sạc được sử dụng
ChargingStation Location	Vị trí địa lý của trạm sạc (ví dụ: New York, Los Angeles).
Charging Start Time	Thời điểm phiên sạc bắt đầu
Charging End Time	Thời điểm phiên sạc kết thúc
Energy Consumed (kWh)	Tổng năng lượng tiêu thụ trong quá trình sạc, được đo bằng kilowatt-giờ.
Charging Duration (hours)	Tổng thời gian sạc, tính bằng giờ
Charging Rate (kW)	Tốc độ cung cấp điện trung bình trong quá trình sạc, được đo bằng kilowatt.
Charging Cost (USD)	Tổng chi phí phát sinh cho phiên sạc, tính bằng đô la Mỹ.
Time of Day	Khoảng thời gian xảy ra việc sạc (ví dụ: Buổi sáng, Buổi chiều).
Day of Week	Ngày trong tuần xảy ra tình trạng tính phí

State of Charge (Start %)	Phần trăm sạc pin khi bắt đầu phiên sạc
State of Charge (End %)	Phần trăm pin được sạc vào cuối phiên sạc
Distance Driven	Quãng đường đã đi được kể từ lần sạc cuối cùng, tính bằng kilômét.
Temperature (°C)	Nhiệt độ môi trường trong quá trình sạc (℃)
Vehicle Age	Tuổi của xe điện, tính bằng năm
Charger Type	Loại bộ sạc được sử dụng (ví dụ: Bộ sạc nhanh DC cấp độ 1, cấp độ 2).
User Type	Phân loại người dùng dựa trên thói quen lái xe (ví dụ: Người đi làm, Người đi du lịch đường dài).

Thống kê ban đầu:

- Số lượng cột: 20
- Số lượng dòng: 1320
- Số lượng biến phân loại: 10
- Số lượng biến số: 10

Bảng mô tả kiểu dữ liệu, loại biến và giá trị khuyết:

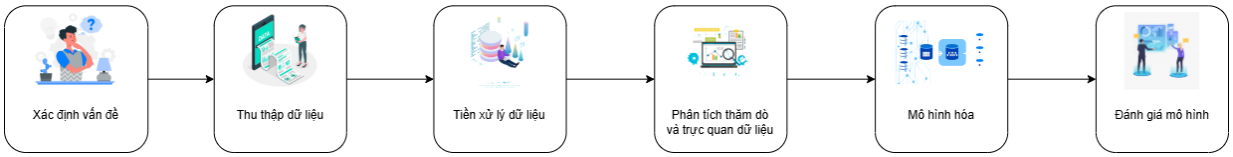
Yếu tố / Tên thuộc tính	Kiểu dữ liệu	Loại biến	Giá trị khuyết
User ID	Object	Biến phân loại	0
Vehicle Model	Object	Biến phân loại	0
Battery Capacity (kWh)	float64	Biến số	0
Charging Station ID	object	Biến phân loại	0
Charging Station Location	object	Biến phân loại	0
Charging Start Time	object	Biến phân loại	0
Charging End Time	object	Biến phân loại	0
Energy Consumed (kWh)	float64	Biến số	66
Charging Duration (hours)	float64	Biến số	0
Charging Rate (kW)	float64	Biến số	66
Charging Cost (USD)	float64	Biến số	0

Time of Day	object	Biến phân loại	0
Day of Week	object	Biến phân loại	0
State of Charge (Start %)	float64	Biến số	0
State of Charge (End %)	float64	Biến số	0
Distance Driven (since last charge)	float64	Biến số	66
Temperature (°C)	float64	Biến số	0
Vehicle Age (years)	float64	Biến số	0
Charger Type	object	Biến phân loại	0
User Type	object	Biến phân loại	0

3. PHƯƠNG PHÁP PHÂN TÍCH

Sau khi cân nhắc về nội dung đề tài và tham khảo một số quy trình phân tích, nhóm chúng tôi đã đưa ra quy trình phân tích dữ liệu tổng quát như sau:

–



Hình 1. Quy trình phân tích dữ liệu tổng quát.

Bước 1 – Xác định vấn đề: cần phân tích dữ liệu về lượng năng lượng các dòng xe điện tiêu thụ.

Bước 2 – Thu thập dữ liệu: ở đây nhóm chọn bộ dữ liệu [1], sau đó đưa bộ dữ liệu vào bằng pandas để bắt đầu phân tích.

Bước 3 – Tiền xử lý dữ liệu:

- Kiểm tra trùng lặp dữ liệu (kết quả không có).
- Loại bỏ một số cột có ít giá trị sử dụng cho phân tích (User ID do cột này có quá nhiều giá trị độc lập với mục đích đánh số thứ tự).
- Xử lý dữ liệu bị khuyết:
  + Kiểm tra số giá trị rỗng trong các thuộc tính (3 thuộc tính số Energy Consumed (kWh), Distance Driven (since last charge) (km), Charging Rate (kW)).

    + Điền các giá trị khuyết bằng giá trị trung bình của cột (Energy Consumed (kWh), Distance Driven (since last charge) (km)).
    + Riêng với cột Energy Consumed (kWh) là cột mục tiêu, thực hiện điền giá trị theo trung bình nhóm một số thuộc tính được cho là có liên quan bao gồm Vehicle Model, Battery Capacity (kWh), và Charger Type. Những nhóm không có giá trị trung bình, điền bằng trung bình toàn thể.
    + Kiểm tra số giá trị khuyết còn lại một lần nữa.
- Chuyển đổi các cột có giá trị thời gian sang định dạng datetime và trích xuất các thông tin thời gian.
- Lọc lại các giá trị nhiễu trong bộ dữ liệu: Tìm và loại bỏ các giá trị nằm ngoài khoảng $\pm$ 1.5 IQR, sau đó tìm chỉ số của các hàng nhiễu và xóa các hàng đó.

Bước 4 – Phân tích thăm đò và trực quan dữ liệu:

- Thống kê mô tả
    + Các thông số hướng trung tâm và phân tán của các cột số: Mean, Mode, Median, Range, Quartiles, IQR, Variance, Standard Deviation, CV.
    + Hình dạng của dữ liệu: Skewness và Kurtosis của các cột số.
- Xác định tương quan:
    + Chuyển biến phân loại thành biến số bằng Label Encoding.
    + Tính toán ma trận tương quan và trực quan bằng Heatmap.
- Phân tích ANOVA:
    + Trực quan phân phối dữ liệu của Energy Consumed theo từng giá trị của biến phân loại.
    + Thực hiện kiểm định ANOVA qua từng biến phân loại để xác định F-statistic và P-value của các biến phân loại với Energy Consumed và đưa ra kết luận.
- Trực quan hóa dữ liệu bằng các biểu đồ.

Bước 5 – Mô hình hóa:

- Tiền xử lý dữ liệu:

+ KNN Imputer: Điền giá trị khuyết bằng phương pháp KNNImputer.
+ Label Encoding: Mã hóa các cột dạng chuỗi (categorical features).
+ Chuẩn hóa dữ liệu (StandardScaler): Chuẩn hóa các đặc trưng số.
+ PolynomialFeatures: Tạo các đặc trưng mới bằng cách kết hợp các cột số hiện tại thành các bậc đa thức.

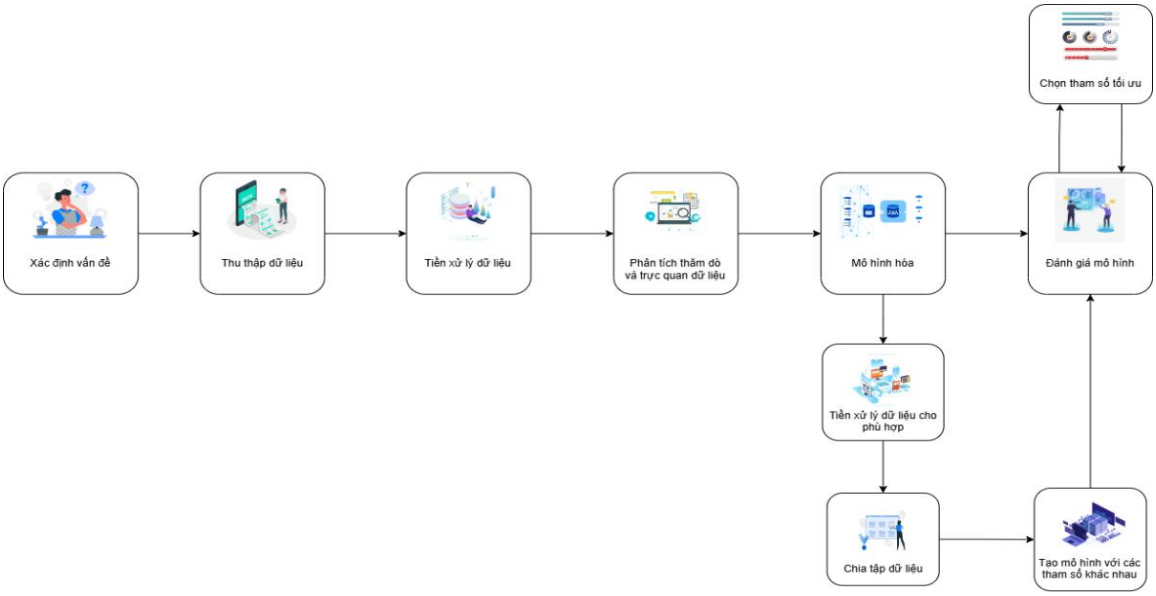
- Chia tập dữ liệu thành tập test và tập train.
- Mô hình chính: Sử dụng ColumnTransformer để chuẩn hóa các cột số và mã hóa các cột phân loại. Đồng thời, tạo các đặc trưng đa thức từ các cột số. Sử dụng Pipeline để kết hợp các bước tiền xử lý và mô hình hồi quy Ridge kết hợp Ploynominal Features để tạo mô hình hồi quy đa thức.
- Tìm kiếm tham số tối ưu:

+ Sử dụng Ridge Regression với các Polynominal Features và One Hot Encoder cho các biến phân loại.
+ Sử dụng cross-validation để tìm ra bậc đa thức (từ 1 đến 5) và giá trị alpha (từ 10^{-3} đến 10^{3}) tốt nhất bằng Cross-Validation (CV) với chỉ số đánh giá là Root Mean Squared Error (RMSE).
+ Huấn luyện mô hình với các giá trị tốt nhất.

Bước 6 – Đánh giá mô hình

- Tính toán RMSE và R-squared trên tập kiểm tra.
- Sử dụng cross-validation để tính toán MSE trung bình và độ lệch chuẩn.
- Vẽ biểu đồ so sánh giữa giá trị thực và giá trị dự đoán.

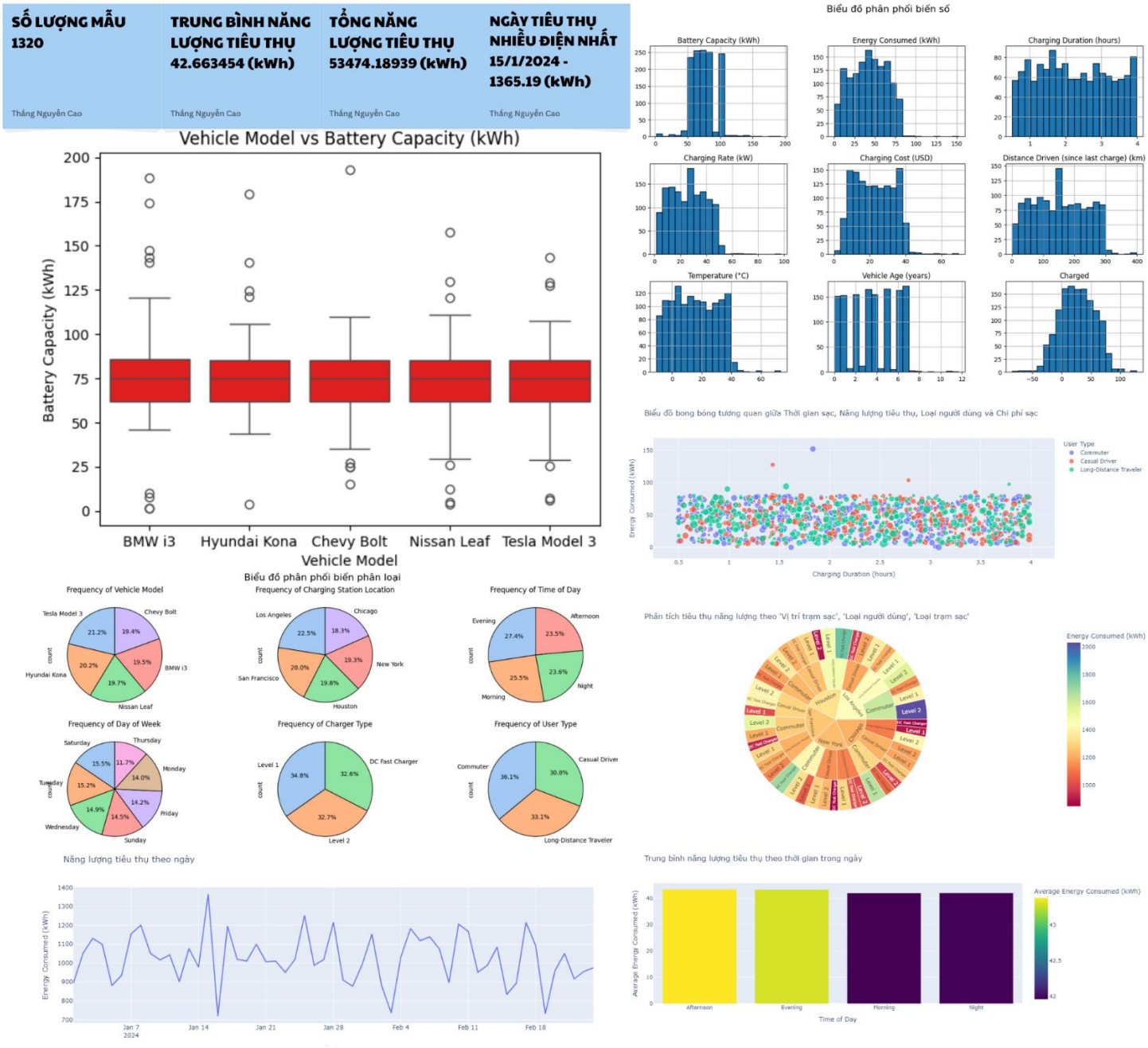
Sau khi hoàn thành đánh giá mô hình, chúng tôi đã xây dựng được một quy trình phân tích dữ liệu tổng thể như sau:



Hình 2. Quy trình phân tích dữ liệu chi tiết.

4. PHÂN TÍCH THĂM DÒ

Sau khi thực hiện thành công phân tích thăm dò và trực quan hóa, một phần kết quả phân tích thu được của nhóm được trình bày trong dashboard sau (phần còn lại trong file notebook TXL - PTTD - Trực quan.ipynb):



Hình 3. Dashboard dữ liệu

Các thống kê, biểu đồ được hiển thị trong dashboard (thứ tự liệt kê từ trái sang phải, trên xuống dưới):

- Các thống kê về số mẫu, trung bình năng lượng tiêu thụ, tổng năng lượng tiêu thụ, thời điểm tiêu thụ năng lượng nhiều nhất.
- Các biểu đồ cột thể hiện phân phối của các biến số gồm Battery Capacity (kWh), Energy Consumed (kWh), Charging Duration (hours), Charging Rate (kW), Charging Cost (USD), State of Charge (Start %), State of Charge (End

%), Distance Driven (since last charge) (km), Temperature (°C), Vehicle Age (years).

- Boxplot thể hiện phân phối Battery Capacity (kWh) đối với các mẫu xe.
- Biểu đồ bong bóng thể hiện tương quan giữa Charging Duratuon và Energy Consumed, phân biệt bởi User Type và Charging Cost.
- Các biểu đồ tròn thể hiện phân phối của các biến phân loại gồm Vehicle Model, Charging Station Location, Time of Day, Day of Week, Charger Type, User Type.
- Biểu đồ Sunburst phân tích Energy Consumed (kWh) theo Charging Station Location, User Type và Charger Type.
- Biểu đồ đường Energy Consumed (kWh) theo từng ngày được ghi nhận trong bộ đữ liệu.
- Biểu đồ cột – nhiệt thể hiện Energy Consumed (kWh) theo các khoảng thời gian trong ngày.

5. KẾT QUẢ PHÂN TÍCH

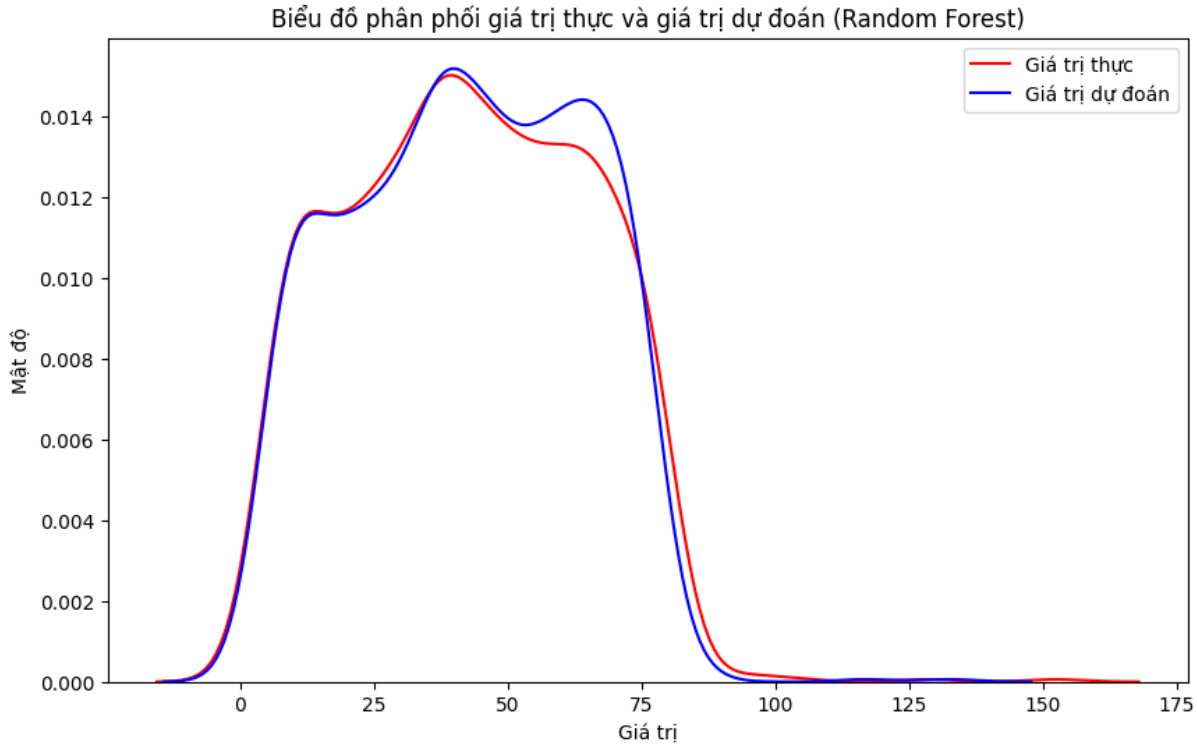
Phân tích kết quả thu được từ dashboard:

- Từ các biểu đồ cột thể hiện phân phối của biến số, có thể rút ra các nhận xét:
  + Pin có dung lượng tập trung nhiều ở khoảng 50-90 kWh, phản ánh phổ biến của các mẫu xe với pin dung lượng trung bình.
  + Hầu hết năng lượng tiêu thụ rơi vào khoảng 20-60 kWh, một số ít mẫu có mức năng lượng tiêu thụ cao hơn 70 kWh, đây có thể là những xe với hành trình dài hoặc điều kiện bất thường. Có thể nhận xét phân bố mức năng lượng tiêu thụ phù hợp với phân phối của dung lượng pin khi hầu hết các xe có dung lượng pin trung bình.
  + Phần lớn các chi phí sạc nằm trong khoảng 10-35 USD, một số trường hợp vượt mức 40 USD nhưng hiếm, tương ứng với mức năng lượng tiêu thụ.
- Từ boxplot thể hiện phân phối Battery Capacity (kWh) đối với các mẫu xe: Tesla Model 3 có dung lượng pin cao nhất trung bình so với các mẫu xe khác, phù hợp cho người dùng có nhu cầu đi đường dài. Có một số ngoại lệ dưới tập trung ở các mẫu xe BMW i3 Nissan Leaf và Chevy Bolt.

- Biểu đồ bong bóng cho thấy mối quan hệ giữa thời gian sạc, năng lượng tiêu thụ, loại người dùng và chi phí sạc. Thời gian sạc của các kiểu tài xế phân bố khá đều trong khoảng 0.5 đến 4 giờ, tuy nhiên Comuter và Casual Driver thường sử đụng ít năng lượng, cũng như kích thước bong bóng (chi phí) ít hơn so với Long Distance Traveller.
- Các biểu đồ tròn thể hiện phân phối của biến phân loại:
  + Tần suất theo loại xe: Tesla Model 3 chiếm tỷ lệ cao nhất (21.2%), tiếp theo là Hyundai Kona và Nissan Leaf.
  + Thời gian trong ngày: Phần lớn các phiên sạc diễn ra vào buổi tối(27.4%) và buổi sáng (25.5%).
  + Tần suất sử dụng loại trạm sạc: Phân bố khá đều, tỉ lệ gần như bằng nhau.
  + Tần suất theo ngày trong tuần: Số lần sạc cao nhất vào Thứ Bảy (15.5%) và thấp nhất vào Thứ Năm (11.7%).
- Biểu đồ Sunburst thể hiện sự phân bổ năng lượng tiêu thụ theo vị trí trạm sạc (Location), loại người dùng (User Type), và loại trạm sạc (Charger Type): Người dùng loại sạc DC Fast Charger ở các thành phố đều tiêu thụ khá ít năng lượng. Lượng năng lượng trung bình tiêu thụ cho việc sạc ở Los Angeles là lớn nhất, trong đó của các tài xế kiểu Comuter dùng sạc Level 2 là lớn nhất.
- Biểu đồ đường thể hiện xu hướng năng lượng tiêu thụ theo từng ngày: Có sự dao động rõ rệt với mức tiêu thụ theo từng ngày, thường là tăng vào cuối tuần, sau đó giảm vào hai ngày đầu tuần, lại tăng vào giữa tuần, giảm vào thứ năm, thứ sáu và tăng lại vào cuối tuần.
- Biểu đồ cột – nhiệt thể hiện Energy Consumed (kWh) theo các khoảng thời gian trong ngày: Lượng năng lượng tiêu thụ cao vào buổi chiều và tối, thấp hơn vào buổi đêm và sáng, nhưng chênh lệch cũng không quá lớn.

Phân tích kết quả thu thập từ các mô hình dự đoán: Trong bài toán này, nhóm chúng tôi đã thử sử dụng 4 mô hình dự đoán, bao gồm 2 mô hình hồi quy đa thức (sử dụng Linear Regression và Ridge Regression tích hợp tiền xử lý Polynominal Features), Decision Tree và Random Forrest. Trong cả 4 mô hình, Random Forrest cho ra được hiệu suất dự đoán tốt nhất.

- Overall RMSE (Random Forest) = 2.2193: RMSE thấp chứng tỏ rằng mô hình dự đoán khá chính xác và sai số giữa giá trị thực và dự đoán là không quá lớn.
- Overall R-squared (Random Forest) = 0.9898: Mô hình có thể giải thích khoảng 98.98% sự biến thiên trong dữ liệu, chứng tỏ đây là một mô hình rất mạnh và chính xác.
- MSE = 34.9494: Cho thấy rằng mô hình có sai số nhỏ, với giá trị MSE khá thấp.
- Standard Deviation of MSE (Random Forest) = 5.3803: Mô hình này có sự ổn định khá cao trong việc dự đoán, cho thấy không có hiện tượng overfitting đáng kể trong quá trình huấn luyện.
- Biểu đồ phân phối giá trị thực và giá trị dự đoán (Random Forrest): hai đường phân phối khá giống nhau trong phần lớn phạm vi của trục hoành, điều này cho thấy mô hình có thể dự đoán khá chính xác các giá trị thực tế của biến mục tiêu.



Hình 4. – Biểu đồ phân phối giá trị thực và giá trị dự đoán (Random Forrest)

6. KẾT LUẬN

Bộ dữ liệu các yếu tố ảnh hưởng đến mức tiêu thụ năng lượng khi sạc các dòng xe điện đã được nhóm thực hiện các thao tác theo đúng quy trình đã được đề ra ban đầu. Chúng tôi đã xác định vấn đề; sau khi thu thập được bộ dữ liệu đã mô tả được bộ dữ liệu, thực hiện các thao tác tiền xử lý như điền giá trị khuyết, loại giá trị nhiễu,... sau đó phân tích thăm dò với thống kê mô tả và phân tích ANOVA; trực quan được dữ liệu bằng nhiều kiểu biểu đồ; tìm được mô hình phù hợp với bộ dữ liệu và đánh giá mô hình đó. Nhóm đã đạt được kết quả là rút ra được các nhận xét chi tiết từ các hình thức trực quan, cũng như xây dựng được một mô hình dự đoán có các thông số đánh giá tương đối tốt. Từ việc xem xét kết quả trực quan và mô hình, chúng tôi có thể kết luận các biến ảnh hưởng nhiều đến mức năng lượng tiêu thụ bao gồm dung lượng pin, thời gian sạc, loại sạc, thời gian sạc, kiểu người lái xe, mẫu phương tiện.

Tuy nhiên, nhóm còn mắc một số nhược điểm trong quá trình thực hiện phân tích, như coding convention trong các file notebook còn tùy tiện; đôi lúc có một số thao tác bị lặp lại nhiều lần;...

TÀI LIỆU THAM KHẢO

[1] Electric Vehicle Charging Patterns. Link: https://www.kaggle.com/datasets/valakhorasani/electric-vehicle-charging-patterns (Ngày truy cập: 11/12/2024)

[2] API Reference – scikit-learn 1.6.0 documentation. Link: https://scikit-learn.org/stable/api/index.html (Ngày truy cập: 11/12/2024).

[3] Implementation of Polynomial Regression – GeeksforGeeks. Link: https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/ (Ngày truy cập: 11/12/2024).

[4] API Reference — Matplotlib 3.9.3 documentation. Link: https://matplotlib.org/stable/api/index (Ngày truy cập: 11/12/2024).

[5] API reference — seaborn 0.13.2 documentation. Link: https://seaborn.pydata.org/api.html (Ngày truy cập: 11/12/2024).

[6] Python API reference for plotly – 5.24.1 documentation. Link: https://plotly.com/python-api-reference/ (Ngày truy cập: 11/12/2024).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Cao Thắng	Lên ý tưởng, tìm bộ dữ liệu; phân tích thăm dò (thống kê mô tả), trực quan hóa dữ liệu, đánh giá mô hình; viết báo cáo; chỉnh sửa sau báo cáo.
2	Võ Phi Thân	Xây dựng mô hình dự đoán, đánh giá mô hình, trực quan hóa dữ liệu.
3	Võ Minh Quyền	Tiền xử lý dữ liệu, phân tích thăm dò (hình dạng dữ liệu, xác định tương quan); thiết kế slide thuyết trình.
4	Võ Anh Quân	Phân tích thăm dò (phân tích ANOVA), trực quan hóa dữ liệu; viết báo cáo.