Objective:

Electric Vehicle Sales by State in India: Machine Learning Project

This project aims to analyze and predict the sales of Electric Vehicles (EV) by state in

India using machine learning. The dataset contains the following columns:

● Year: The year of the sales.

● Month\_Name: The month in which sales occurred.

● Date: The specific date of the sales.

● State: The state in India where the sales occurred.

● Vehicle\_Class: The class of the vehicle (e.g., sedan, SUV, etc.).

● Vehicle\_Category: The category of the vehicle (e.g., commercial, passenger).

● Vehicle\_Type: The type of the vehicle (e.g., 2-wheeler, 4-wheeler).

● EV\_Sales\_Quantity: The quantity of EV sales.

Steps Involved:

1. Data Collection: Load and inspect the dataset.

2. Data Preprocessing: Handle missing values, convert date formats, and

perform feature engineering.

3. Exploratory Data Analysis (EDA): Visualize trends and relationships between

variables.

4. Feature Engineering: Create new features from the date column and encode

categorical variables.

5. Modeling: Build a regression model to predict EV sales.

6. Evaluation: Evaluate the model performance and interpret the results.

7. Visualization: Visualize the results and trends using graphs and charts.

Python Code: Step-by-Step

Step 1: Data Collection

Start by loading the dataset. For this example, let's assume the dataset is in CSV

format.

# Import necessary libraries

import pandas as pd

import numpy as np

# Load the dataset

df = pd.read\_csv('ev\_sales\_india.csv')

# Display the first few rows of the dataset

print(df.head())

Step 2: Data Preprocessing

Handle missing values and convert the date column to a proper datetime format.

# Convert 'Date' column to datetime format

df['Date'] = pd.to\_datetime(df['Date'])

# Check for missing values

print(df.isnull().sum())

# Fill missing values (if any) using median for numerical

columns or mode for categorical columns

df['EV\_Sales\_Quantity'].fillna(df['EV\_Sales\_Quantity'].median()

, inplace=True)

df.fillna(df.mode().iloc[0], inplace=True)

Step 3: Exploratory Data Analysis (EDA)

Visualize trends in EV sales over time, across states, and vehicle categories.

import matplotlib.pyplot as plt

import seaborn as sns

# Plot EV sales over the years

plt.figure(figsize=(10, 6))

sns.lineplot(data=df, x='Year', y='EV\_Sales\_Quantity',

hue='State')

plt.title('EV Sales by State over the Years')

plt.show()

# Plot sales by vehicle category

plt.figure(figsize=(10, 6))

sns.barplot(x='Vehicle\_Category', y='EV\_Sales\_Quantity',

data=df, ci=None)

plt.title('EV Sales by Vehicle Category')

plt.show()

Step 4: Feature Engineering

Create new features such as month and day from the Date column and encode

categorical variables.

# Extract Month and Day from the Date column

df['Month'] = df['Date'].dt.month

df['Day'] = df['Date'].dt.day

# Encode categorical variables using one-hot encoding

df\_encoded = pd.get\_dummies(df, columns=['State',

'Vehicle\_Class', 'Vehicle\_Category', 'Vehicle\_Type'],

drop\_first=True)

# Drop unnecessary columns like Date, Month\_Name (if already

extracted into numerical values)

df\_encoded.drop(['Date', 'Month\_Name'], axis=1, inplace=True)

Step 5: Modeling

Use a regression model (e.g., Random Forest Regressor) to predict EV sales.

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_squared\_error

# Split the data into features and target variable

X = df\_encoded.drop('EV\_Sales\_Quantity', axis=1)

y = df\_encoded['EV\_Sales\_Quantity']

# Split the dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y,

test\_size=0.2, random\_state=42)

# Instantiate the model

model = RandomForestRegressor(n\_estimators=100,

random\_state=42)

# Train the model

model.fit(X\_train, y\_train)

# Make predictions

y\_pred = model.predict(X\_test)

# Evaluate the model

mse = mean\_squared\_error(y\_test, y\_pred)

rmse = np.sqrt(mse)

print(f'Root Mean Squared Error: {rmse}')

Step 6: Model Evaluation

Check how well the model performs on the test set.

# Plot actual vs predicted sales

plt.figure(figsize=(10, 6))

plt.scatter(y\_test, y\_pred)

plt.title('Actual vs Predicted EV Sales')

plt.xlabel('Actual EV Sales')

plt.ylabel('Predicted EV Sales')

plt.show()

# Check feature importance

importance = model.feature\_importances\_

feature\_importance = pd.Series(importance,

index=X\_train.columns).sort\_values(ascending=False)

# Plot the most important features

plt.figure(figsize=(10, 6))

feature\_importance.plot(kind='bar')

plt.title('Feature Importance')

plt.show()

Step 7: Conclusion

The machine learning model helps in understanding the factors affecting Electric

Vehicle sales across different states and predicting future sales based on historical

data. Feature importance gives insight into which factors (e.g., State, Vehicle

Category) have the highest impact on sales.

Explanation:

● Data Preprocessing: Cleaned the dataset and handled missing values.

● Feature Engineering: Created new columns from the Date column and

encoded categorical variables.

● Modeling: Built a Random Forest Regressor model to predict EV sales and

evaluated its performance using RMSE (Root Mean Squared Error).

● Visualization: Visualized sales trends and feature importance using bar plots

and scatter plots.

**About Dataset**

This dataset is valuable for analysts, data scientists, and researchers aiming to

understand electric vehicle (EV) adoption trends across India. It is versatile and ideal

for geographic market segmentation, trend analysis, and predictive modeling. By

offering insights into regional EV sales patterns, the dataset supports strategic

decision-making in market planning and infrastructure investment.

The data was meticulously scraped from the Clean Mobility Shift website, and then

thoroughly preprocessed to ensure accuracy and relevance. All null values have been

removed, and the dataset has been cleaned to prepare it for immediate use in

exploration, visualization, and analytical projects. It is particularly valuable for market

trend analysis, infrastructure planning, and policy development within the EV sector.

The dataset is provided in CSV format and is ready for analysis.

Included Files:

EV\_Dataset.csv: Contains state-level data on EV sales, including vehicle types and

categories, offering a comprehensive view of EV distribution across Indian states.

Key Features:

State: Names of Indian states with recorded EV sales data.

Vehicle Type: Classifications of vehicles, such as two-wheelers and four-wheelers.

Vehicle Category: Further classification into segments like commercial and passenger

vehicles.

Electric\_Vehicle\_Sales\_Quantity: The number of EVs sold per state, essential for

analyzing adoption trends.