**Project Report: PowerPulse – Household Energy Usage Forecast**

**Domain: Energy and Utilities**

**Project Overview**

The *PowerPulse* project aims to develop a machine learning model to predict household energy consumption based on historical data. This predictive model provides insights into usage patterns for consumers, helping them optimize their energy consumption, while energy providers can use the model for demand forecasting to enhance operational efficiency.

**Skills Acquired**

* Data Preprocessing
* Feature Engineering
* Regression Modeling
* Evaluation Metrics

**1. Introduction**

Energy consumption is one of the primary concerns for households and energy providers alike. In this project, we focus on predicting household energy consumption based on historical power usage data. By applying machine learning techniques, this project seeks to offer actionable insights to both consumers and energy providers. The objectives of this project include:

* Analyzing patterns and trends in energy usage.
* Developing a predictive model that can forecast future consumption.
* Providing recommendations for optimizing energy use.

**2. Data Understanding and Exploration**

**Dataset Information:**

* **Rows**: 1,048,575
* **Columns**: 9
* **Variables**:
  1. **date**: Date in format dd/mm/yyyy.
  2. **time**: Time in format hh:mm:ss.
  3. **global\_active\_power**: Household global minute-averaged active power (in kilowatts).
  4. **global\_reactive\_power**: Household global minute-averaged reactive power (in kilowatts).
  5. **voltage**: Minute-averaged voltage (in volts).
  6. **global\_intensity**: Household global minute-averaged current intensity (in amperes).
  7. **sub\_metering\_1**: Energy sub-metering No. 1 (in watt-hour of active energy), representing the kitchen appliances.
  8. **sub\_metering\_2**: Energy sub-metering No. 2 (in watt-hour of active energy), representing the laundry room appliances.
  9. **sub\_metering\_3**: Energy sub-metering No. 3 (in watt-hour of active energy), representing the electric water heater and air conditioner.

**Initial Observations:**

* The dataset contains both numerical and categorical data.
* Missing values are present in the dataset (4069 null values).
* Variables like date and time are of object type and need conversion into appropriate formats.
* Outliers are identified in several variables like global\_active\_power, global\_reactive\_power, voltage, and global\_intensity.
* Variables exhibit different ranges of values, hence scaling is necessary.

**Libraries Used:**

python

CopyEdit

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

**3. Exploratory Data Analysis (EDA)**

**Insights from EDA:**

* **Correlation Analysis**: It was observed that global\_active\_power is positively correlated with global\_intensity. No strong correlations were observed between other variables.
* **Outliers**: Variables such as global\_active\_power, global\_reactive\_power, voltage, and global\_intensity had significant upper-range outliers, with global\_intensity exhibiting both upper and lower-range outliers.

**Visualization:**

* **Scatter Plots**: Scatter plots were used to visualize relationships between variables, reinforcing the positive correlation between global\_active\_power and global\_intensity.
* **Box Plots**: Box plots identified the presence of outliers, helping in further preprocessing decisions.

**4. Data Preprocessing**

**4.1 Handling Missing Values:**

* **Missing Data Treatment**: Median imputation was performed to handle missing values in variables other than date and time, as it’s less affected by outliers compared to the mean.

**4.2 Outlier Treatment:**

* Outliers in global\_active\_power, global\_reactive\_power, voltage, and global\_intensity were capped using the Interquartile Range (IQR) method. For more extreme variables (e.g., sub\_metering columns), capping was done at 0.01 and 0.99 percentiles.

**4.3 Feature Scaling:**

* As some variables had different ranges (e.g., some with two-digit values and others with three-digit values), Min-Max scaling was applied to normalize the data.

**4.4 Feature Engineering:**

Several new features were created to enhance the model's predictive power:

1. **Daily Averages**: The average of global\_active\_power was calculated for each day, adding a new feature daily\_avg\_power.
2. **Peak Hours**: Extracted hours from the time column, flagging peak hours (between 18:00 and 23:00) in a new binary column peak\_hour.
3. **Rolling Averages**: A 7-day rolling average for global\_active\_power was computed and added as 7\_day\_rolling\_avg\_power.
4. **Weekend Indicator**: A binary column is\_weekend was introduced to differentiate weekends (1) from weekdays (0).
5. **Season**: Extracted the season based on the month from the date column, categorizing them as Winter, Spring, Summer, or Fall.

**Data Types Conversion:**

* The date and time columns were converted to appropriate formats. The date column was parsed as datetime, and time was converted into a 24-hour time format for further analysis.

**5. Model Selection**

Four models were trained and evaluated to predict household energy consumption:

1. **Linear Regression**: A simple regression model that fits a linear equation to the dataset.
2. **Gradient Boosting**: An ensemble learning method that builds the model incrementally.
3. **Neural Network**: A feed-forward neural network with fully connected layers.
4. **Random Forest**: An ensemble learning method using multiple decision trees.

**6. Model Evaluation**

Each model’s performance was evaluated using the following metrics:

* **Mean Absolute Error (MAE)**: Average of absolute errors.
* **Mean Squared Error (MSE)**: Mean of squared errors.
* **Root Mean Squared Error (RMSE)**: Square root of the mean of squared errors.
* **R-Squared (R²)**: Proportion of the variance in the dependent variable explained by the model.

**6.1 Model Performance on Test Set:**

| **Model** | **MAE** | **MSE** | **RMSE** | **R²** |
| --- | --- | --- | --- | --- |
| **Linear Regression** | 0.007506 | 0.0001411 | 0.01188 | 0.99833 |
| **Gradient Boosting** | 0.005619 | 9.02e-05 | 0.009497 | 0.99893 |
| **Neural Network** | 0.006105 | 8.70e-05 | 0.00933 | 0.99897 |
| **Random Forest** | 0.00325 | 5.38e-05 | 0.00733 | 0.99936 |

**6.2 Interpretation:**

* **Linear Regression**: The test set performance shows a low MAE (0.007506) and a high R² (0.99833), indicating a nearly perfect fit.
* **Gradient Boosting**: This model performed slightly better than Linear Regression, with lower errors (MAE: 0.005619, MSE: 9.02e-05).
* **Neural Network**: The performance is comparable to Gradient Boosting with an R² of 0.99897.
* **Random Forest**: This model achieved the best performance with the lowest MAE (0.00325), MSE (5.38e-05), and the highest R² (0.99936), making it the top performer for this task.

**6.3 Final Conclusion:**

* **Random Forest** demonstrated the best performance in terms of minimizing error and explaining variance in the dataset, making it the most suitable model for predicting household energy consumption in this project.
* The results indicate that the model generalizes well to unseen data with no signs of overfitting or underfitting.

**7. Conclusion**

The *PowerPulse* project successfully developed multiple machine learning models to predict household energy consumption. The Random Forest model was identified as the best performer, achieving the highest R² and the lowest error values, indicating its ability to generalize well to unseen data. The predictive models can be applied to help both consumers and energy providers make informed decisions about energy usage and demand forecasting.

**Business Impact**:

* Consumers can utilize the model to gain insights into their energy usage patterns and optimize their consumption.
* Energy providers can better forecast demand, leading to more efficient energy distribution and management strategies.

**Future Enhancements**:

* Incorporating additional features like weather data or household occupancy can further improve the model’s predictive power.
* Testing other machine learning models like XGBoost or LightGBM could provide additional insights and potentially enhance model performance.

**8. References**

* Pandas Documentation: https://pandas.pydata.org/
* Scikit-learn Documentation: <https://scikit-learn.org/>
* Seaborn Documentation: https://seaborn.pydata.org/