**Powering the Future: Machine Learning Analysis of the Global Power Plant Database**

**1. Problem Definition**

The Global Power Plant Database is a comprehensive, open source database of power plants around the world. Our study spans around 14,000 power plants in three countries including the United States (USA), Australia (AUS) and India (IND). These plans comprise of historic and renewable power dealing (hydro, wind solar) power plants as well traditional thermal power plants (coal, gas oil nuclear).

The two things we are trying to solve in this project;

Forecast that the main fuel source for power plants would be required to warm just about six percent of your homes.

Predict the capacity of the power plant in MW (Megawatts), that is an indicative of how much electricity it can generate.

Knowledge of fuel type and fuel capacity is essential for energy planning, investment, and policy-making. We can use this information to forecast a variety of factors to help us understand energy trends, especially renewable energy adoption. This is all in the pursuit to make predictions with machine learning models that will direct decision making in the realm of energy.

**2. Data Analysis**

Here's the important information about power plant from the dataset

**Country:** 3-character iso code (e.g., USA, IND, AUS) of the country where the power plant is located

**Geographical Region**: Latitude and Longitude of the plant.

**Capacity (MW):** The maximum number of megawatts the plants can generate electricity.

**Primary Fuel:** The fuel that is most used in the generation of electricity (coal, gas, wind, solar).

**Generation Data (2013-2019):** How much electricity was actually produced for each plant over several years.

The primary goals for this project were to predict:

**Primary Fuel:** This is a classification problem that predicts the fuel type used by each plant.

Capacity (MW): This is a regression problem, where we want to predict the electric capacity of any given plant.

Traditional as well as renewable energy plants are present in the dataset so one can study and predict different attributes related to energy for different regions.

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

Exploratory Data Analysis (EDA) and trends discovered:

**Geopolitical influence**– Some countries or regions rely more heavily on certain types of fuel. India, for instance, relies most heavily on coal with a more diversified range of other sources in Australia and the USA such as wind and solar.

**Capacity vs. Fuel Type:** You can typically find smaller, lower capacity power plants if they use renewable energy sources like solar or wind compared to huge plants using coal or gas that have a larger capacity (typically over 1,000 MW).

Over the years, we have seen an increase in using wind and solar to produce electricity with these environmentally friendly energy generation adding to power sources. However, the most electricity still comes from traditional thermal plants.

**Correlations:** Capacity of a plant is strongly correlated with electricity generation. Unsurprisingly, bigger plants produce more electricity.

We used these insights to steer the machine learning models and how we interpretive our features in order to predict accurately fuel type and plant capacity.

**4. Data Pre-processing Pipeline**

Before building the machine learning models, we needed to clean and prepare the data. The following steps were essential for data pre-processing:

**Handling Missing Values**

* We noticed some missing data in the capacity and electricity generation fields. Missing capacity data was dropped, while missing generation data was filled with average values to ensure completeness.

**Encoding Categorical Variables**

* Categorical variables such as **Country** and **Primary Fuel** were converted into numerical formats using **one-hot encoding**, so the machine learning algorithms could process them effectively.

**Scaling Features**

* Numerical values like **Capacity (MW)** and electricity generation data were standardized to ensure that features with large ranges did not dominate the model training process.

**Train-Test Split**

* The data was split into **80% training data** and **20% test data**, which allowed us to train the models and then evaluate their performance on unseen data.

**5. Building Machine Learning Models**

We developed two separate models—one for predicting the primary fuel type (classification) and another for predicting plant capacity (regression).

**1. Classification: Predicting Primary Fuel Type**

To predict the primary fuel type, we used a **Gradient Boosting Classifier**, a machine learning model known for its strong performance on classification tasks. We trained the model using the following features:

* **Country, Latitude, and Longitude**
* **Electricity Generation Data (2013-2019)**
* **Age of the Plant** (calculated from the year the plant started operating)

Initially, the model achieved an accuracy of **83%**. We further improved the model’s performance through **hyperparameter tuning**, raising the accuracy to **87%**.

**2. Regression: Predicting Plant Capacity (MW)**

For the capacity prediction task, we used a **Gradient Boosting Regressor**. This model was well-suited to capture the relationships between the plant’s characteristics (such as location, fuel type, and generation history) and its capacity.

The model’s initial **R-squared score** was **0.89**, indicating a good fit. After hyperparameter tuning, this score improved to **0.92**.**6. Hyperparameter Tuning**

To optimize both the classification and regression models, we used **GridSearchCV** for hyperparameter tuning. This process searches for the best combination of parameters to maximize the model’s performance.

**1. Hyperparameter Tuning for Primary Fuel Prediction (Classification)**

For the classification model, the following parameters were tuned:

* **n\_estimators**: The number of boosting stages (trees).
* **learning\_rate**: Controls the contribution of each tree to the final model.
* **max\_depth**: Limits the depth of the trees to prevent overfitting.
* **min\_samples\_split**: The minimum number of samples required to split an internal node.
* **min\_samples\_leaf**: The minimum number of samples required at a leaf node.

After tuning, the best parameters were:

* **n\_estimators**: 200
* **learning\_rate**: 0.1
* **max\_depth**: 5
* **min\_samples\_split**: 5
* **min\_samples\_leaf**: 2

These changes improved the model’s accuracy to **87%**.

**2. Hyperparameter Tuning for Capacity Prediction (Regression)**

For the regression model, the following parameters were tuned:

* **n\_estimators**: The number of trees used for boosting.
* **learning\_rate**: Controls how much each tree contributes to the overall prediction.
* **max\_depth**: Limits the depth of each tree.
* **min\_samples\_split**: The minimum number of samples to split a node.
* **min\_samples\_leaf**: The minimum number of samples at the leaf node.

After tuning, the best parameters were:

* **n\_estimators**: 200
* **learning\_rate**: 0.05
* **max\_depth**: 4
* **min\_samples\_split**: 10
* **min\_samples\_leaf**: 2

With these optimized parameters, the model’s **R-squared score** increased to **0.92**, indicating excellent predictive performance.

**7. Conclusion**

* In this work, we demonstrated that machine learning can be effectively used to predict primary fuel type and capacity of the power plants in the Global Power Plant Database. We analysed energy generation per countries and fuel type to understand trends and patterns with exploratory data analysis. Hope you have enjoyed my post and read me describe how Gradient Boosting models achieve state-of-the-art accuracy for both the tasks above before showing you that some more hyperparameters tuning brings up even better results.
* The outcomes could be useful for business planning in the energy sector, where they might assist policymakers and industry leaders in making choices on upcoming investments in power generation. And again, this project will show that machine learning could teach us the world in global energy trends.
* While this is the case, and signs that the energy landscape is changing are quite evident with the rise of renewables particularly in wind and solar, we have a model like ours to help deliver these models. Being able to predict the types of fuel and capacities of plants that will be used in future decades can help inform future energy production strategies, working toward a more sustainable energy future.