1Dataset chosen

Wind Turbines in Canada

2.1 A business problem that can be solved with this dataset.

How can I predict the power output of wind turbines (total\_project\_capacity\_mw) based on the the dataset features?

2.2 Identify steps and intermediary questions that need to be resolved to solve

1. Data Preprocessing:

Do we need to handle missing values and outliers? If so, how?

Are there any categorical variables that need to be encoded?

Do we need to normalize or scale the numerical features?

Are there any irrelevant features that can be removed?

1. Feature Selection:

Which features have the most impact on the power output?

Are there any redundant or highly correlated features that should be eliminated?

Can we engineer new features that might enhance the predictive power?

1. Splitting the Data:

How should the dataset be split into training and testing sets?

What is an appropriate ratio for the training and testing data?

1. Model Selection:

Which regression algorithms are suitable for predicting power output?

Should we consider linear regression, decision tree regression, or ensemble methods like random forest regression?

Are there any other models specific to wind turbine performance prediction?

1. Model Training and Evaluation:

How will the selected model be trained on the training data?

What evaluation metrics should be used to assess the model's performance?

How can we avoid overfitting the model to the training data?

1. Hyperparameter Tuning:

Do the chosen models have hyperparameters that need to be tuned?

What techniques can be used to find the optimal hyperparameter values?

Can we use techniques like grid search or randomized search for hyperparameter optimization?

1. Model Performance Evaluation:

How well does the trained model perform on the testing data?

Are there any metrics for evaluating the model's accuracy, such as mean absolute error, mean squared error, or R-squared?

1. Deployment and Prediction:

Once the model is trained and evaluated, how can it be deployed for real-time predictions?

How can the model be used to predict the power output of wind turbines given new input data?

2.4 identify variables and statistical approaches to be used

To predict the total\_project\_capacity\_mw of wind turbines based on the given dataset, the following variables and statistical approaches can be considered:

Variables

Total\_Project\_Capacity\_MW (Target Variable): This is the variable we want to predict, representing the power output of wind turbines in megawatts.

Potential Predictor Variables:

Turbine\_Rated\_Capacity\_kW: The rated capacity of the wind turbine in kilowatts can be an important predictor of the total project capacity.

Rotor\_Diameter\_m: The diameter of the wind turbine rotor in meters can impact its power generation potential.

Hub\_Height\_m: The height of the turbine hub above the ground in meters can affect the wind capture and power generation efficiency.

Manufacturer: The wind turbine manufacturer can have variations in the design, technology, and performance of the turbines.

Model: The specific model of the wind turbine can also contribute to its power output.

Commissioning\_Date: The date when the wind turbine was commissioned can provide insights into technological advancements and potential changes in performance over time.

Latitude and Longitude: While latitude and longitude may not have a direct impact, they can indirectly affect wind resources and environmental conditions as follows:

1. Wind Resource Availability: Latitude and longitude can influence the availability and quality of wind resources at a specific location. Different regions have varying wind patterns and speeds, which can significantly impact the power output of wind turbines. Therefore, the wind resource potential at a particular latitude and longitude can indirectly affect the total\_project\_capacity\_mw.
2. Environmental Conditions: The geographical location of wind turbines can also be associated with different environmental conditions, such as temperature, altitude, terrain, and proximity to bodies of water. These factors can affect the wind turbine's efficiency, maintenance requirements, and power generation capacity.
3. Turbine Siting and Design: The selection of latitude and longitude for wind turbine placement is often influenced by factors like land availability, land use regulations, distance from residential areas, and transmission infrastructure. The siting decisions can impact the turbine's performance, including its rated capacity and total project capacity.

Statistical approaches

Random Forest Regression: Random forest regression is an ensemble method that combines multiple decision trees to make more accurate predictions. It can handle complex interactions and reduce overfitting.  
  
  
Data Import  
The code Below highlights the Wind turbines data set importation