# Business Report: Harnessing the Power of Wind - Insights from Canadian Wind Turbine Data

**BCO6008 – Predictive Analytics – Assessment 3**

# Table of Contents

# Introduction

Wind energy is becoming increasingly popular as a clean and sustainable alternative to traditional fossil fuels. Energy companies are keen to find the most efficient turbines that can produce the highest amount of power while minimizing harm to the environment. In this report, we analyze data from wind turbines in Canada to understand their design and performance. Additionally, we develop a model that can predict the turbines that are best suited for maximizing power generation.

## The Dataset

The Canadian Wind Turbines dataset provides valuable information on wind energy projects in Canada. It includes data related to various aspects of wind turbines such as their capacity, location, specifications, and manufacturer details. This dataset is of great interest to researchers, energy companies, and policymakers who are involved in the renewable energy sector and want to analyze the performance and impact of wind turbines in Canada.

## Business Problem

We are investigating the problem of identifying the best turbines for generating the highest power output. This is an important issue for energy companies who want to invest in wind turbines that are efficient and can produce the maximum amount of power while minimizing their impact on the environment. In this report, our goal is to develop a predictive model that can help solve this problem by accurately predicting which turbines are the most effective for maximizing power generation.

## Methodology

The methodology for this report involves several key steps to achieve the desired outcome. These steps are as follows:

- Loading the dataset: We begin by loading the Canadian wind turbine dataset, which contains information about various turbine specifications such as rotor diameter, hub height, manufacturer, model, and turbine power. The dataset will serve as the foundation for our analysis and modeling.

```
# Load the libraries and the dataset
```{r}
library (tidyverse)
library (tidymodels)
library (janitor)
library (dplyr)
library (knitr)
library (ggplot2)
library (caret)
library (parsnip)
library (tune)

wind_turbine <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-10-27/wind-turbine.csv')
```
```

***Figure 1. Glimpse of code when loading the libraries and dataset***

- Data Cleaning: In order to ensure data accuracy and reliability, we perform data cleaning procedures. This involves identifying and handling missing values, as well as removing any irrelevant or redundant columns from the dataset. By cleaning the data, we can ensure that our analysis is based on high-quality and complete information.

```
# Check whether any missing values observed in the dataset
```{r}
skimr::skim(wind_turbine)

#Missing values are observed in the dataset, even almost all observations in the notes column are missing values. We need to clean it :)
```

# Clean the dataset
```{r}
#Remove the notes column from the dataset and missing values in other column
wind_turbine_cleaned <-
  wind_turbine [, !(names(wind_turbine) == "notes")] %>%
  na.omit()

skimr::skim(wind_turbine_cleaned)
```
```

***Figure 2. Glimpse of code when performing the data cleaning***

- Exploratory Data Analysis (EDA): The next step is to conduct exploratory data analysis (EDA) to gain insights and understand the patterns within the dataset. We will explore various aspects such as the environmental impact of wind energy, manufacturer and model analysis, turbine evolution, and turbine design. EDA will help us identify relationships and trends that can inform our predictive model.

- Predictive Modeling: To address the business problem of determining the best turbines for maximizing power output, we will develop a predictive model. This involves data pre-processing, feature selection, model building, as well as model training and testing. By leveraging machine learning techniques, we aim to build a model that can accurately predict turbine power output based on relevant features such as rotor diameter, hub height, manufacturer, and model.

# Exploratory Data Analysis (EDA)

## Environmental Impact

Wind energy offers a green alternative to traditional fossil fuels, providing less pollution and relying on renewable and never-ending resources. By utilizing wind energy, countries can strive towards energy self-sufficiency, minimizing their reliance on imports which can be influenced by geopolitics. The shift towards wind energy also fosters job creation and boosts the economy. While fossil fuel industries also provide jobs, their future prospects are limited due to the finite nature of fossil fuel resources.

Wind energy is beneficial for the environment in the long term, a stark contrast to fossil fuels, which harm the environment and pose threats to future generations. Canadian wind energy projects have significantly contributed to reducing CO2 emissions. By using wind energy instead of other sources, they have prevented the release of almost half a million metric tons of CO2 into the atmosphere.

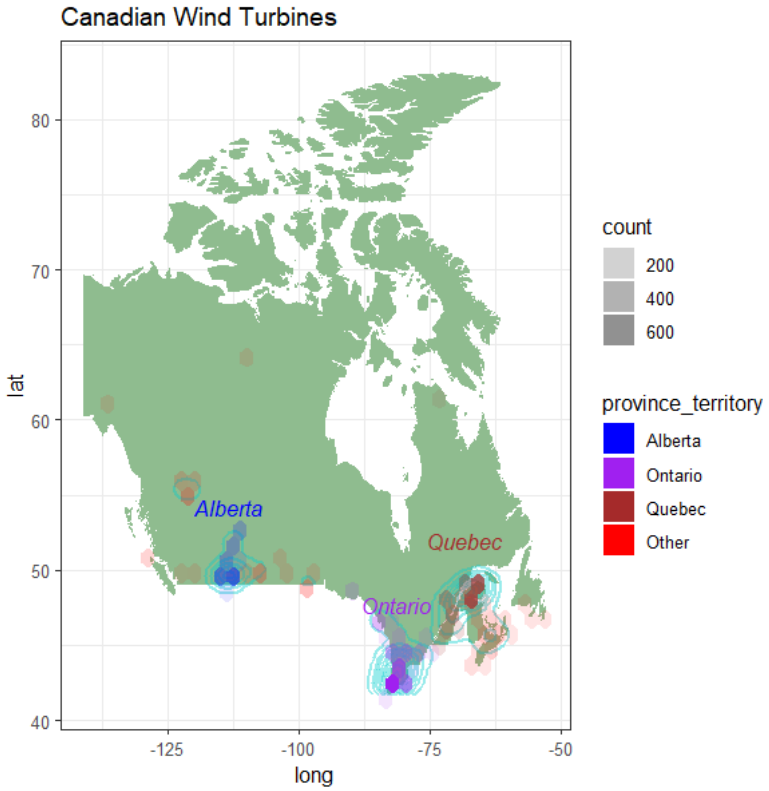



***Figure 3. Wind Turbine Coverage in Canada***

When we look at the Canadian wind energy landscape, we find that the provinces of Alberta, Ontario, and Quebec have played a crucial role. These regions have embraced wind energy and have witnessed the installation of a substantial number of wind turbines. As a result, they have contributed significantly to reducing CO2 emissions and promoting cleaner air.

## Manufacturer and Model Analysis

A detailed analysis of the data reveals that Acciona leads the market in terms of average turbine capacity. Acciona's turbines exhibit higher-rated capacities, which translates to more efficient and productive wind energy generation. Partnering with Acciona provides access to cutting-edge technology, optimized performance, and the ability to generate more clean energy.

The most common model, GE 1.5SLE, has been used in 1011 projects, asserting its dominance in the market. Its widespread adoption signifies its reliability, performance, and acceptance among wind energy developers. Leveraging the GE 1.5SLE model offers benefits such as economies of scale, streamlined maintenance processes, and access to a robust support network. This model alone has achieved a total CO2 emission reduction of 16.2% (68316.75 metric tons per year).

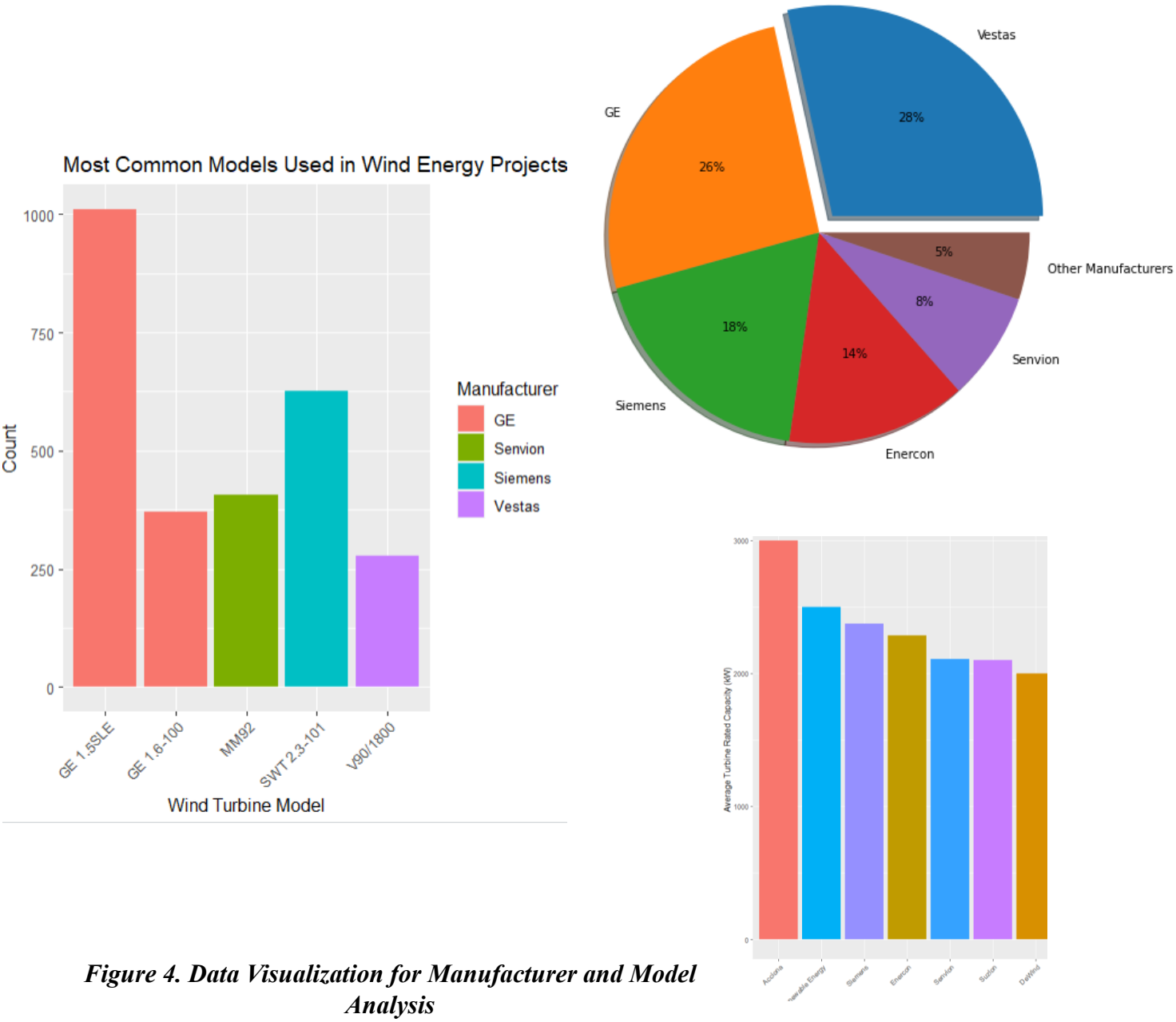



***Figure 4. Data Visualization for Manufacturer and Model Analysis***

## Turbine Evolution

With time, turbines are designed with higher capacities. Larger rotor diameters enable turbines to capture more wind energy, resulting in higher power output. Enhanced turbine design, including aerodynamics and blade advancements, contributes to improved efficiency. Manufacturers develop higher capacity turbines to meet the growing demand and to benefit from economies of scale. Industry standards and competition drive the development of higher capacity turbines. Higher capacity turbines align with environmental goals by reducing reliance on fossil fuels and mitigating climate change.

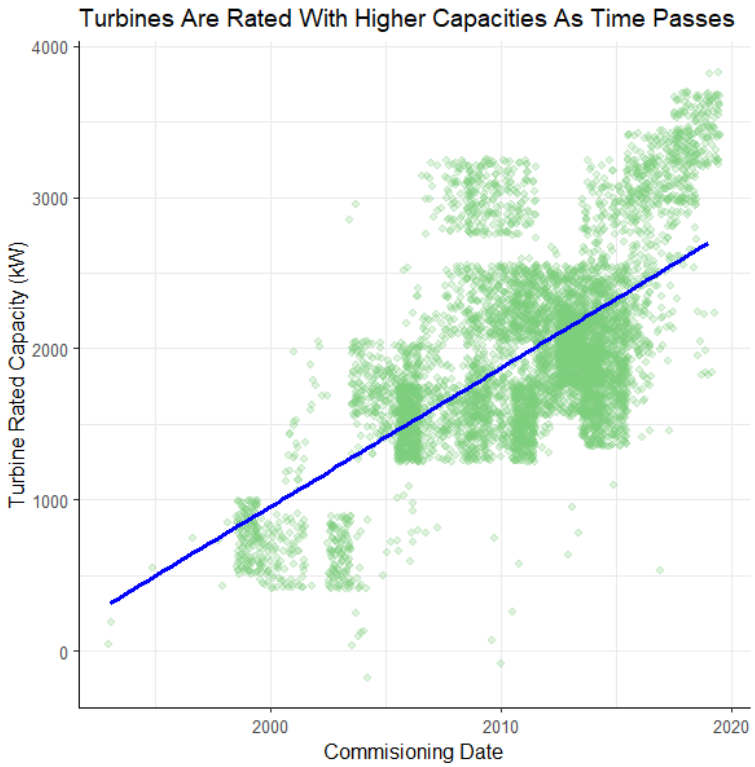
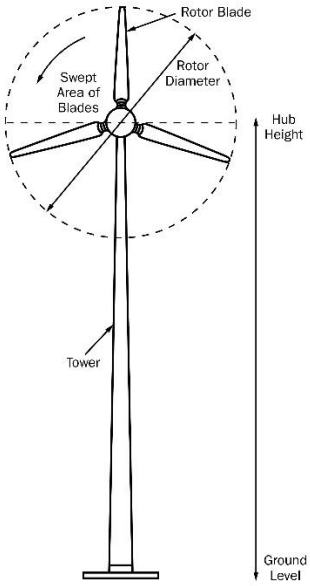



***Figure 5. Turbine Evolution***

## Turbine Design

The turbine features are highly necessary for determining how much electricity can be generated from a single turbine. The features themselves build the turbine outlook, known as the turbine design. They contribute to the maximum capacity a turbine has to provide an electrical output. The term is referred to the turbine capacity. Looking at the dataset, there are two important factors affecting the size of the turbine capacity, i.e. rotor diameter and hub height.




*Source: www.ontario.ca*

### Rotor Diameter

Rotor diameter refers to the length measured from one tip of the rotor blade to another one. Rotor diameter indicates the area of sweep performed by a turbine rotor (Arantegui et al. 2020). The swept area contributes to the amount of wind trespassed. More wind captured will increase the speed of the rotating rotor, resulting in higher electrical energy generated by the wind turbine. A relationship between rotor diameter and the wind turbine capacity was plotted from the provided dataset and shown in the graph below.

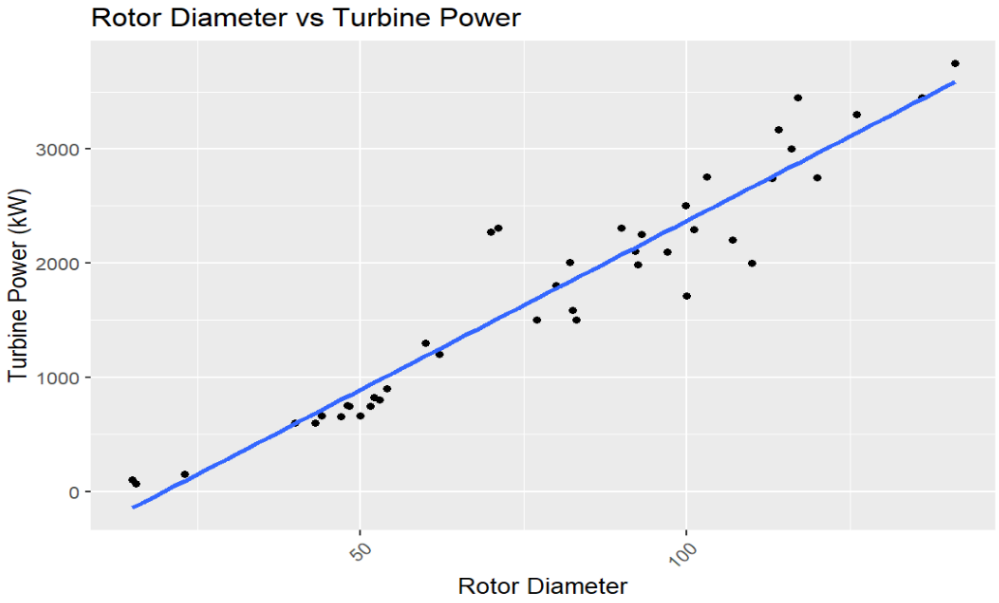



***Figure 6. Relationship between Rotor Diameter and Turbine Power***

From the graph above, it is observed that the larger the rotor diameter, the higher the resulting turbine power (kW) generated from the wind rotor. It is advisable that to design the best turbine producing a higher power, set the wider rotor diameter.

## Hub Height

Hub height refers to the length of a wind turbine pole measured from the group until it reached the wind rotor. Hub height indicates the altitude of the wind turbine, in which a higher altitude will usually have a stronger windblown (Arantegui et al. 2020). A stronger windblown increases the amount of wind passing through the turbine rotor, resulting in higher electrical energy generated from the wind turbine. The relationship between hub height and turbine power was plotted according to the given dataset and provided below.

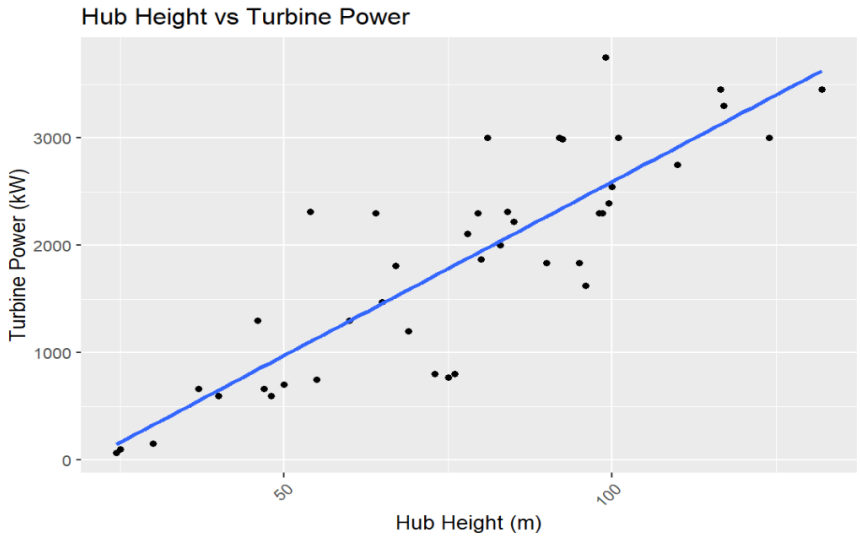
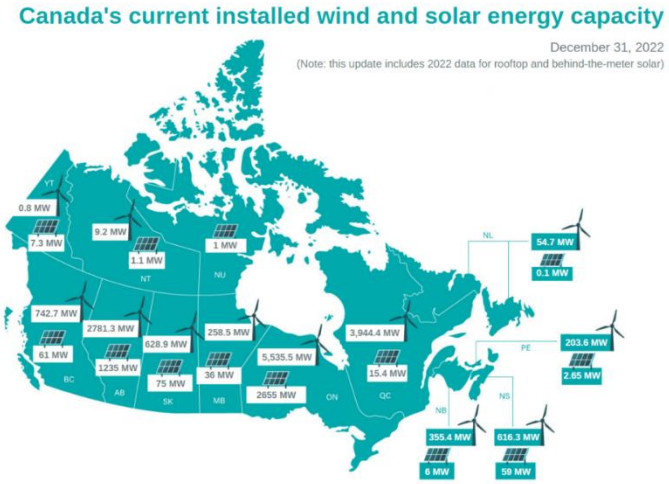



***Figure 7. Relationship between Hub Height and Turbine Power***

From the graph above, it is shown that the higher the hub height, the bigger the resulting turbine power (kW) generated from the wind rotor. It is advisable that to design the best turbine producing a higher power, set the higher hub height of the turbine.

## Geographical Distribution

Canada has successfully put the country in 8th rank for the highest wind energy capacity in the world. It has produced approximately 15,000 MW of electricity from wind energy which covered 7% of the country's demand for electrical energy (CREA, 2022). The wind turbines are located across the country with different intensities in every province.




*Source : Canadian Renewable Energy Association, 2022*

According to the data analysis, there are two questions arising from the geographical distribution, i.e. "What if Canada has more wind turbines?" and "Does more turbines always mean more power?". Two data tables comparison is provided below.

| province | n | province | electric_power |
|---|---|---|---|
| Ontario | 2443 | New Brunswick | 2628.992 |
| Quebec | 1991 | British Columbia | 2456.164 |
| Alberta | 900 | Northwest Territories | 2300.000 |
| Nova Scotia | 310 | Ontario | 2036.720 |
| British Columbia | 292 | Nova Scotia | 2007.161 |
| Saskatchewan | 153 | Newfoundland and Labrador | 1988.077 |
| Manitoba | 133 | Prince Edward Island | 1957.308 |
| New Brunswick | 119 | Quebec | 1943.797 |
| Prince Edward Island | 104 | Manitoba | 1943.233 |
| Newfoundland and Labrador | 26 | Alberta | 1642.278 |
| Northwest Territories | 4 | Saskatchewan | 1577.647 |
| Yukon | 2 | Yukon | 405.000 |

***Figure 8. Table Comparison of Canadian Provinces, Number of Turbines, and Amount of Electricity***

From the above figure, it is observed that Ontario, Quebec, and Alberta are the top 3 provinces with the most frequent wind turbines in Canada. Surprisingly, when looking at the overall amount of electricity generated, Their positions are replaced by New Brunswick, British Columbia, and Northwest Territories as the top electrical energy generator. The analysis indicated that more turbines are not always having more electrical energy generated, meaning that a specific model may be contributed higher producing electricity. Nevertheless, Quebec and especially Alberta have the potency to be exploited. It can be done by adding more turbines using the top model which produces higher electricity.

## Predictive Modelling

The business problem of "How to predict the best turbine producing the highest electric power?" can be answered by setting up a predictive model using a supervised machine learning approach followed by a regression analysis.

The steps to set up the model are summarized below:

- Data Pre-processing
- Set the recipe associated with the feature selections (step_other, step_dummy, step_nzv, and step_normalized)
- Setting up the model by selecting the rand_forest () as the algorithm, "ranger" as the engine, and "regression" as the mode
- Setting up the workflow and running the workflow using fit()
- Augment () the model using the training set
- Visualize the augmented model of the training set
- Evaluate the model by augmenting the model using testing set
- Visualize the augmented model of the testing set
- Set the model evaluation metrics and print out the metric results

The entire code for doing the modeling step was documented in the R script document, which is provided separately as an attachment to this report.

### Data Pre-processing

In this step, some data modification was conducted, shown as below:

- Parsed-number the "commissioning_date" variable, then changed the variable name becomes "commissioning_year"
- Filtered out the turbine power equal or below 2500 kW according to the result of exploratory data analysis
- Set seed for reproducibility
- Split the dataset becoming training and testing set

```r
# Parse Number the commissioning date variable
wind_turbine_cleaned <- wind_turbine_cleaned %>%
  mutate(commissioning_year = parse_number(commissioning_date))

# Filter the dataset to contain only turbine power below or equal 2500
wind_turbine_filtered <- wind_turbine_cleaned %>%
  filter(turbine_rated_capacity_k_w <= 2500)

wind_turbine_filtered$commissioning_date = NULL

# Set seed for reproducibility
set.seed (345)

#Split the dataset
data_split <- initial_split (wind_turbine_filtered)

data_test <- testing (data_split)
data_training <- training (data_split)
```

***Figure 9. Glimpse of code when performing data pre-processing***

## Feature Selection

In this step, some features are selected to increase the accuracy of the model. There are four steps chosen as the additional features, such as step_other, step_dummy, step_nzv, and step_normalize. “Step_other” is selected to remove duplicated observations in the “model” variable. “Step_dummy” is appointed to create a dummy variable to handle analysis involving a categorical variable. “Step_nzv” is selected to remove observations with near zero or no variance at all. “Step_normalize” is chosen to normalize the numerical predictors, which usually use when doing a regression analysis.

```r
# Setting up recipe
data_recipe <- recipe(turbine_rated_capacity_k_w ~ rotor_diameter_m
                      + hub_height_m
                      + commissioning_year
                      + manufacturer
                      + model,
                      data = data_training) %>%
  step_other(model) %>%
  step_dummy(manufacturer) %>%
  step_nzv(all_numeric()) %>%
  step_normalize(all_numeric_predictors())
```

***Figure 10. Glimpse of code when setting up the recipe***

## Model Building

When about to build the model, Random forest is selected as the algorithm. The selection of random forest is based on its advantages, i.e. high predictive accuracy, reduce overfitting, the robustness of the model, and mainly on its flexibility of variable selection (IBM, 2023). The algorithm can generate high-accuracy predictions using multiple trees analysis, taking out the average observations, evaluating it on the testing dataset, and being able to handle both categorical and numerical variables. Thus, the algorithm makes it suitable to analyze the wind turbine dataset which contains both types of variables. “Ranger” was selected as the engine to run the algorithm.

In addition, regression analysis was then proceeded further as the model to predict the electrical output of different wind turbines to answer the business problem question. This analysis was

conducted by the expected outcome of the predictive analysis, i.e. numerical values of turbine power.

```
# Setting up the model
model_1 <- rand_forest() %>% set_engine("ranger") %>%
  set_mode("regression")
```

***Figure 11. Glimpse of code when setting up the predictive model***

## Model Training & Testing

The predictive model was fitted and trained using a training set. To evaluate the model performance, it was evaluated using a testing set.

```
# Run the workflow using fit ()
trained_model <- wf_1 %>% fit(data = data_training)

# Present the results using augment
fit_test_1 <- trained_model %>% augment(new_data = data_training)

#Evaluate the predictive model
fit_test_2 <- trained_model %>% augment(new_data = data_test)
```

***Figure 12. Glimpse of code when training and testing the predictive model***

## Result Interpretation

The result of model training and testing was visualized using ggplot() and shown below.

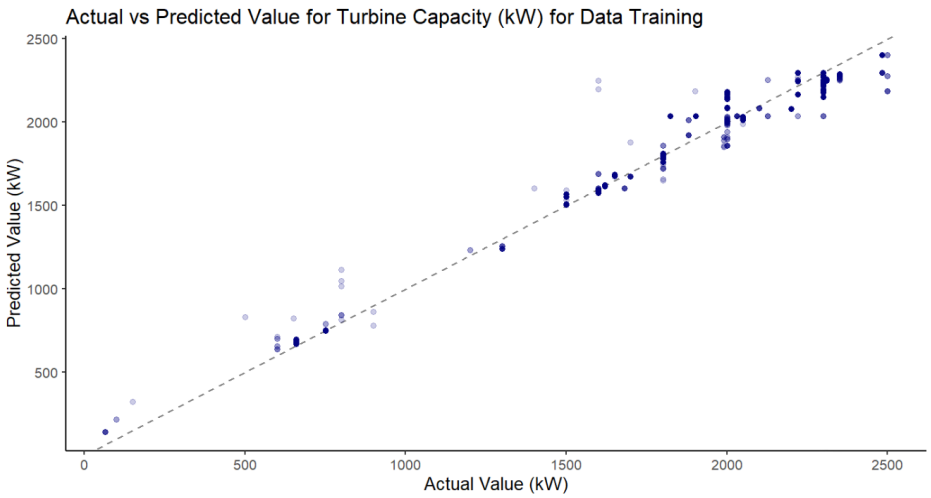



***Figure 13. Result for Model Training***

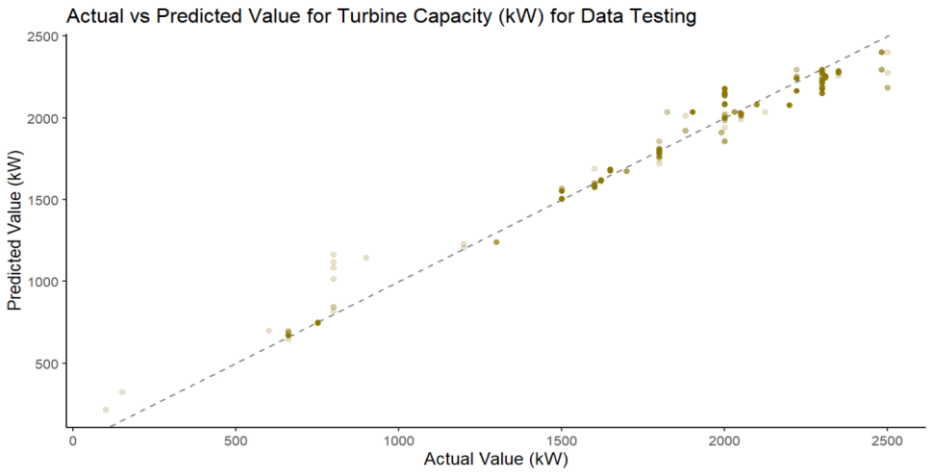



***Figure 14. Result for Model Testing (Evaluation)***

Both figures (***Figure 13*** & ***Figure 14***) show a positive relationship between actual and predicted values. It indicates that the model has successfully generated predictive values from the actual values. The results can be further evaluated using the model evaluation metrics.

| .metric <chr> | .estimator <chr> | .estimate <dbl> |
|---|---|---|
| rmse | standard | 58.858542 |
| mae | standard | 33.478458 |
| rsq | standard | 0.981375 |

***Figure 15. Result of Model Evaluation Metrics for Model Training***

| .metric <chr> | .estimator <chr> | .estimate <dbl> |
|---|---|---|
| rmse | standard | 61.4006594 |
| mae | standard | 35.2875521 |
| rsq | standard | 0.9798097 |

***Figure 16. Result of Model Evaluation Metrics for Model Testing***

Chicco et al. (2021) discussed the definitions and functions of performance metrics in regression analysis. Root-mean squared error (RMSE) is a measurement of deviation between predicted and actual values by taking the square root of the mean difference. Mean absolute error (MAE) is a measurement of the absolute error from the mean difference. R-squared is the measurement of how well a prediction was made from the predictors.

The determination of the model accuracy is based on the values of RMSE and MAE. The lower the RMSE and MAE scores, the more accurate the model is. From both ***Figure 15*** and ***Figure 16***, it is observed that there is only a little difference between RMSE in training and testing the model as well as for the MAE values. It indicates that the model is well-trained and not overfitted, showing good accuracy in predicting the outcome.

Looking at the r-squared values, both model training and testing show a high model fit. By considering the RMSE and MAE values, both r-squared values (above 0.9) indicate a high predictive model performance and accuracy.

## Recommendations

Based on the analysis and findings from the predictive model, the following recommendations can be made:

- Focus on leading manufacturers: Investors should consider partnering with leading manufacturers such as Acciona, as they have demonstrated higher-rated capacities and more efficient turbines. Collaborating with established manufacturers can provide access to cutting-edge technology, optimized performance, and the ability to generate more clean energy.

- Select proven turbine models: The GE 1.5SLE model has shown widespread adoption and a significant impact on CO2 emission reduction. Investors should consider utilizing this model or other proven models to maximize power output and achieve environmental goals. These models often benefit from economies of scale, streamlined maintenance processes, and a robust support network.

- Prioritize larger rotor diameter: Rotor diameter has a positive correlation with turbine power output. It is recommended to prioritize turbines with larger rotor diameters, as they can capture more wind energy and generate higher electrical power. Investors should consider investing in turbines with wider rotor diameters to maximize power generation.

- Optimize hub height: Hub height also has a positive relationship with turbine power output. Higher hub heights result in stronger wind flow, leading to increased energy generation. Investors should consider selecting turbines with higher hub heights to harness stronger wind resources and maximize power production.

- Explore untapped regions: While certain provinces like Ontario, Quebec, and Alberta have seen significant wind turbine installations, other regions such as New Brunswick, British Columbia, and Northwest Territories have demonstrated high electricity generation potential. Investors should explore these untapped regions and consider adding more turbines, especially using the top-performing models identified in the analysis, to further increase power output and contribute to CO2 emission reduction.

- Continuous monitoring and improvement: As wind turbine technology evolves and new models emerge, it is essential for Investors to continuously monitor advancements in the industry. Regularly assessing new turbine designs, technological innovations, and performance metrics can help identify opportunities for further improvement and optimize power generation.

## Conclusion

Overall, the predictive model has successfully answered our primary business problem. It confirmed that the selection of turbines from leading manufacturers such as Acciona and the use of models like GE 1.5SLE can predict higher power output. Moreover, opting for turbines with larger rotor diameters and higher capacities is also indicative of higher power generation. These insights will be valuable for Investors looking to optimize their wind turbine investments and give attractive returns in the long run.

# References


- Arantegui, R, L, Uihlein, A. and Yusta, J, M, 2020, 'Technology effects in repowering wind turbines', *Wind Energy*, viewed 10 June 2023, <https://onlinelibrary.wiley.com/doi/10.1002/we.2450>.
- Chicco, D, Warrens, M, J, Jurman, G, 2021, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, viewed 11 June 2023, < https://peerj.com/articles/cs-623/>.

- CREA, 2022, *Canadian Renewable Energy Association*, CREA, viewed 10 June 2023, < https://renewablesassociation.ca/by-the-numbers/>.

- IBM, 2023, *What is Random Forest?*, IBM, viewed 11 June 2023, <https://www.ibm.com/topics/randomforest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.>.