SEPTEMBER 9, 2024

COMBINED CYCLE POWER PLANT

PREDICTION OF PLANT OUTPUT

AUTHOR: BILL LOUER

LinkedIn Profile

Executive Summary

Combined cycle power plant capacity is heavily dependent on ambient site conditions, as well as the designed operational capability of its equipment. This report focuses on evaluating a dataset that contains ambient site conditions—such as temperature, pressure, humidity and turbine backpressure—and the power output of a combined cycle power plant with the objective of using the relevant variables to predict plant output capability.

The analysis began with a comprehensive evaluation of the dataset to understand its structure and quality. Relationships between the feature variables and MW output were analyzed, identifying key factors that influence plant performance. Following this, various machine learning models were trained and used to predict plant output based on these features. Three different models were trained and validated. All models could predict plant output to a high degree of accuracy. The Random Forest Regression model had the highest R-squared, and lowest error.

Model Type	R Squared	Mean Absolute Error	Mean Squared Error	Accuracy in 95% Confidence Interval
Linear Regression	0.932	3.556	19.674	+/- 1.8%
Gradient Boosting Machine	0.955	2.738	12.953	+/- 1.5%
Random Forest Regression	0.964	2.304	10.396	+1.5%/-1.4%

The dataset was a generic dataset made available by a Kaggle user as mentioned in the acknowledgement section. I believe that further accuracy could be produced with additional feature variables which represent the capacity and operational state of the combined cycle power plant equipment. I could foresee many real-world applications in using this data. These include the following:

- Public data such as this can allow plant owners to benchmark their performance against their competitors.
- Benchmark the performance of power plants based on site conditions so that declines in output or efficiency could be monitored.
- Further refine the model to support more accurate predictions of plant capability
    - Obtain other power plant data such as Combustion Turbine inlet pressure drop, HRSG exhaust differential pressure, HP and IP steam pressures and other plant parameters.
    - Once a more accurate model is generated, small changes in plant output could alarm operators to check on various components. Much of this is already happening on new plant equipment that are monitored by Original Equipment Manufacturers. It is possible that older plant equipment is not being monitored to this level, especially in smaller industries that have limited resources.

Contents

1 Introduction and Objectives

Combined cycle power plant capacity is heavily dependent on ambient site conditions, as well as the designed operational capability of the equipment. Accurately predicting the megawatt (MW) output of combined cycle power plants is essential for improving performance and ensuring efficient energy production.

This report focuses on evaluating a dataset that contains ambient site conditions—such as temperature, pressure, and humidity—and the power output of a combined cycle power plant. The analysis begins with a comprehensive evaluation of the dataset to understand its structure and quality. Relationships between feature variables and MW output were analyzed, identifying key factors that influence plant performance. Following this, various machine learning models were trained and used to predict plant output based on these features. By rigorously assessing the accuracy of these models, we aim to select the most effective one for forecasting MW output. The concluding section of this report communicates the results of the predictive modeling efforts, detailing the accuracy of the chosen model and addressing the potential applications for this type of modeling.

2 Data Collection and Description

The dataset used in this report consists of 9,568 hourly average readings collected from sensors at a Combined Cycle Power Plant. The data includes the following variables:

- Temperature (T):
  - Measurement Unit: Degrees Celsius (°C)
  - Description: Represents the ambient temperature at the time of measurement.
- Ambient Pressure (AP):
  - Measurement Unit: Millibar (mbar)
  - Description: Indicates the atmospheric pressure measured at the plant.
- Relative Humidity (RH):
  - Measurement Unit: Percentage (%)
  - Description: Represents the relative humidity in the air surrounding the plant.
- Exhaust Vacuum (V):
  - Measurement Unit: Centimeters of Mercury (cm Hg)
  - Description: Indicates the level of vacuum in the turbine exhaust.
- Net Hourly Electrical Energy Output (PE):
  - Measurement Unit: Megawatts (MW)
  - Description: The amount of electrical energy produced by the power plant per hour.

This dataset provides the site environmental conditions and energy output, which are crucial for analyzing and predicting the plant's output.

3 Exploratory Data Analysis

The dataset was checked for null values and outliers after basic descriptive statistics were performed. There were 9,568 data points in the dataset representing hourly average observations of the power plant site ambient conditions and output at base load. Mean dry bulb temperatures were about 20C, and mean plant output was 454 MW.

Statistic	Dry Bulb (ºC)	Turbine Back-pressure (cm_Hg)	Ambient Pressure (mbar)	Relative Humidity	Plant Output (MW)
count	9,568	9,568	9,568	9,568	9,568
mean	19.651	54.306	1013.259	73.309	454.365
std	7.452	12.708	5.939	14.6	17.067
min	1.81	25.36	992.89	25.56	420.26
25%	13.51	41.74	1,009.1	63.328	439.75
50%	20.345	52.08	1,012.94	74.975	451.55
75%	25.72	66.54	1,017.26	84.83	468.43
max	37.11	81.56	1,033.3	100.16	495.76

Table 1 Data Descriptive Statistics

The distribution of the data was plotted graphically using box plots as can be seen in the following plots. There were non-significant outliers in the data for ambient pressure and for relative humidity. Given the small number of outliers and the fact that they were only on the cusp of being outliers, they were ignored and used for model training or testing.

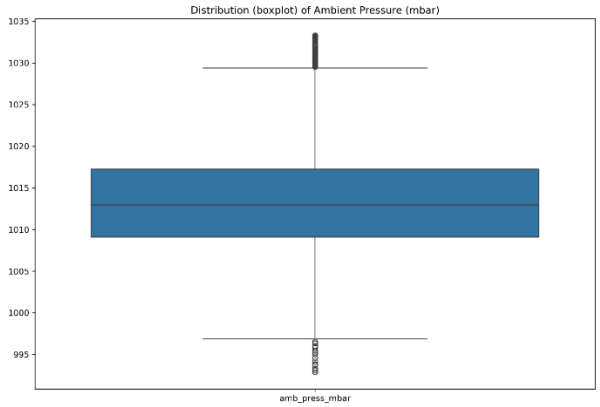

Figure 1 Box Plot (Ambient Pressure)

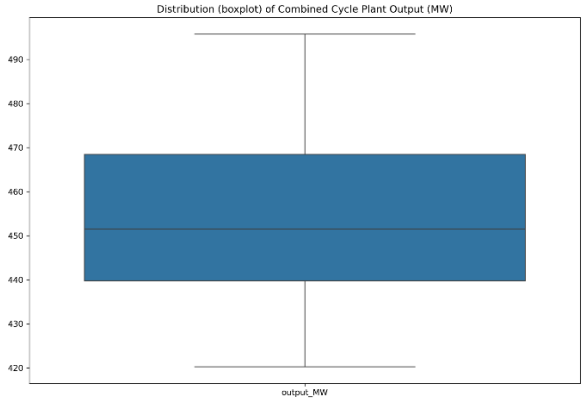

Figure 2 Box Plot (Output MW)

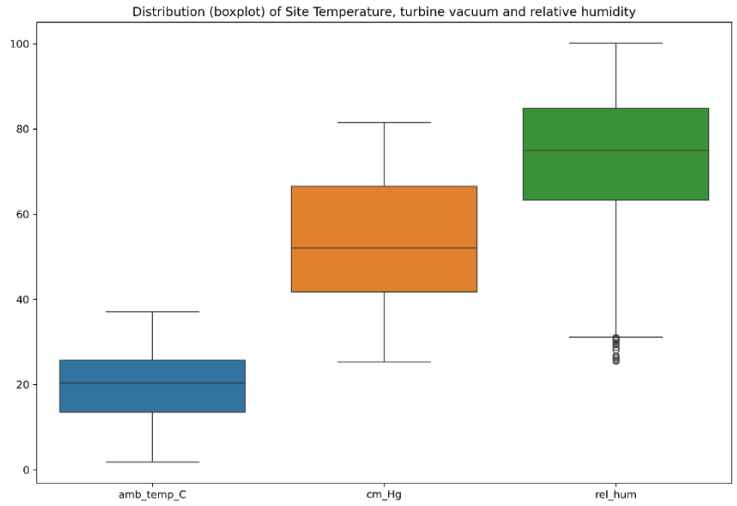

Figure 3 Box Plot (Dry Bulb Temperature, Turbine Backpressure, Relative Humidity)

Variables were evaluated to assess the correlation between feature variables and the plant output. There was a strong correlation between all values, with the highest correlation between temperature, wet-bulb temperature, and turbine backpressure. P-values for each of these correlations were all zero, indicating that the correlation between variables and plant output are all highly statistically significant.

Features	Target Variable	Correlation Coefficient	p_value
amb_temp_C	output_MW	-0.948	0
wet_bulb_temp_C	output_MW	-0.954	0
amb_press_mbar	output_MW	0.518	0
amb_press_PA	output_MW	0.518	0
rel_hum_fraction	output_MW	0.39	0
cm_Hg	output_MW	-0.87	0

Table 2 Correlation Coefficients and P-Values

4 Feature Engineering

There was limited feature engineering used in this dataset. A wet bulb temperature was calculated using each coincident records of dry bulb temperature, relative humidity, and ambient pressure. The feature variables were scaled using Standard Scaler from scikitlearn. The standard scaler feature in scikitlearn works by transforming the independent variables into standardized values, relative to the distance from the mean.

For each feature, the function calculates the mean (average) and standard deviation of the feature values across all samples. It then transforms the variable by subtracting the mean and then dividing by the standard deviation. This results in zero values when the raw variable is equal to the mean and a value of 1, when the variable is exactly one (1) standard deviation from the mean.

5 Modeling

5.1 Multiple Linear Regression

A linear regression model was fitted with the training data and then validated using the test dataset. Eighty (80) percent of the data was used to train the model, and twenty (20) percent was used for model validation for all the modeling scenarios. A linear regression model in scikitlearn was used to fit the data and measure model fit and accuracy.

5.1.1 Model Description

The linear regression model in scikit-learn works by finding the best-fitting line through a set of data points, where the relationship between the input features (independent variables) and the target variable (dependent variable) is assumed to be linear. The model minimizes the sum of squared residuals (the difference between predicted and actual values) to estimate the optimal coefficients (weights) for each input feature. This is done using an ordinary least squares method. Once trained, the model can predict continuous target values for new input data. The model's performance was evaluated using mean squared error (MSE), mean absolute error (MAE) and R^2 to determine how well the model fits the data. In scikit-learn, the Linear Regression class provides a simple interface for training the model, making predictions, and accessing the coefficients of the linear relationship.

5.1.2 Model Results

Model evaluation metrics such as mean squared error, mean absolute error and R-squared were used to evaluate the model from predicted and actual values of the target variables. The results can be seen below.

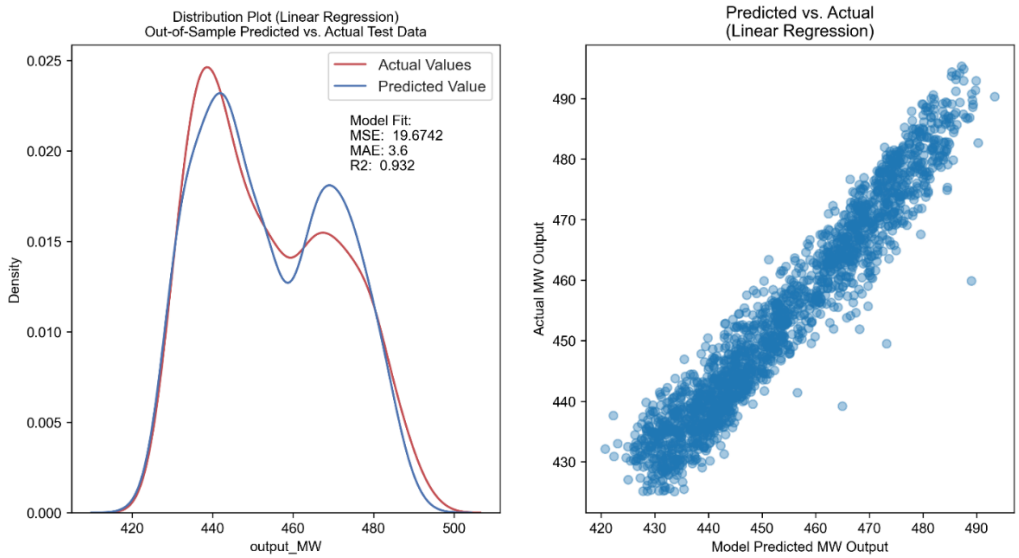

Figure 4 Linear Regression Actual vs. Predicted Distribution Plot

The error of each prediction was calculated as a percent of the actual value and the distribution of errors was plotted on a kernel density estimate plot. The figure below shows the 95% confidence interval of errors, showing a +/- 1.8% error when using the Linear Regression model.

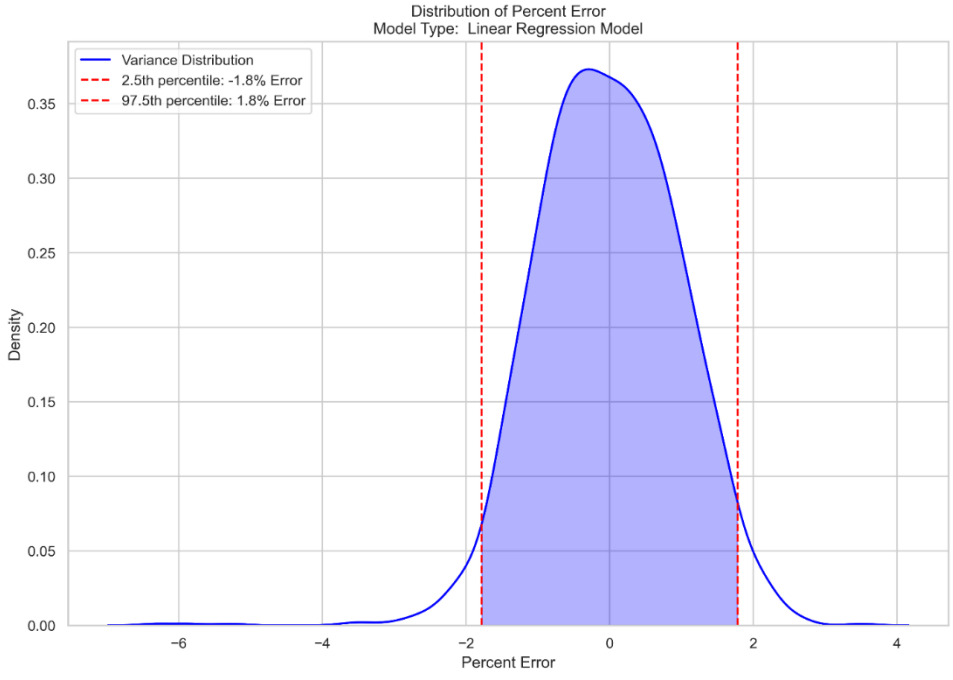

Figure 5 Linear Regression Error Distribution

5.2 Gradient Boosting Machine

A Gradient Boosting Machine (GBM) model was testing under various combinations of parameter values using GridSearchCV. The model was fitted with the training data and then

validated using the test dataset. Th GBM model and GridSearchCV model in scikitlearn were used to fit the data and measure model fit and accuracy.

5.2.1 Model Description

The GBM model is a combination of regression and classification models which works iteratively to continuously improve upon model fit until it reaches a stopping criterion.

An initial model is developed, a prediction made, and residuals (difference between predicted and actual values) calculated. A new model, typically a decision-tree, is trained to predict these residuals. The model learns to correct errors made by the previous model. Predictions are then updated, and the cycle continues. Key components of the model are the learning rate, the number of estimators (number of trees to be built) and the decision tree depth.

5.2.2 Model Results

Model evaluation metrics such as mean squared error, mean absolute error and R-squared were used to evaluate the model from predicted and actual values of the target variables. The results can be seen below.

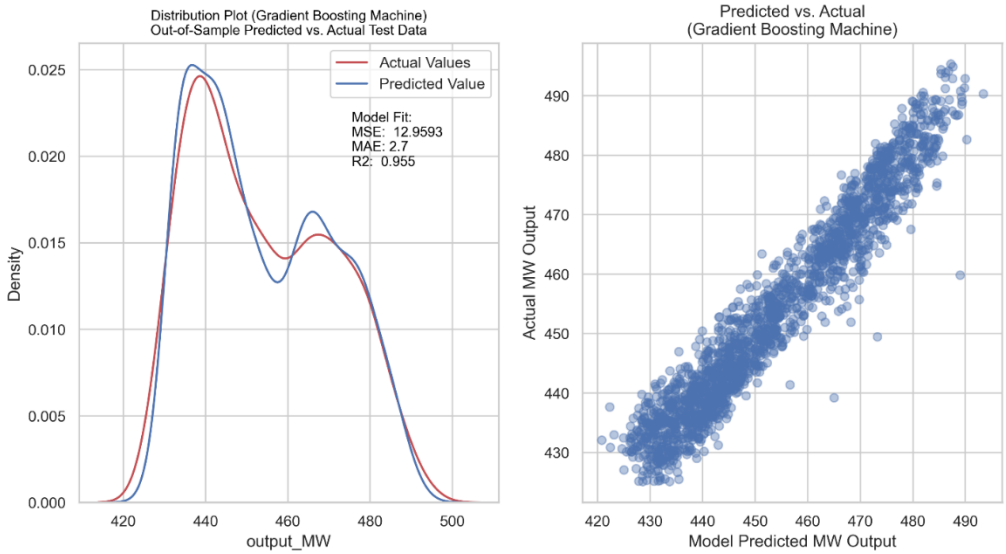

Figure 6 Gradient Boosting Machine Distribution of Actual and Predicted values.

The error of each prediction was calculated as a percent of the actual value and the distribution of errors was plotted on a kernel density estimate plot. The figure below shows the 95% confidence interval of errors, showing a +/- 1.5% error when using the Gradient Boosting Machine model, an improvement from the Linear Regression model.

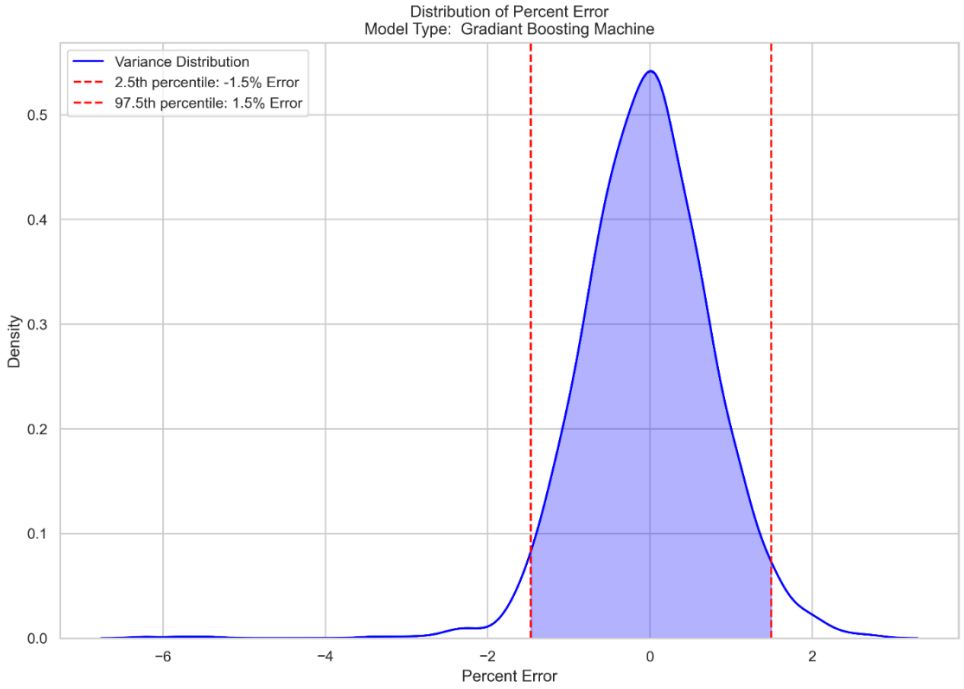

Figure 7 Gradient Boosting Machine Error Distribution

5.3 Random Forest Regression

5.3.1 Model Description

Random forest regression is an ensemble learning method which relies on multiple decision trees. Each decision tree in the forest is trained independently on its own sample of data. Features are considered randomly in each decision tree. This results in a robust set of trees from which predictions are made. In our case, since we are not predicting a categorical variable, the continuous variable is an average of the predictions made from individual trees in the forest.

By using multiple trees, overfitting can be reduced because we are averaging predictions from multiple trees. The process provides insights into feature importance, helping identify which features contribute most to the predictions.

5.3.2 Model Results

As noted previously, model evaluation metrics were used to evaluate the model from predicted and actual values of the target variables. The results can be seen below.

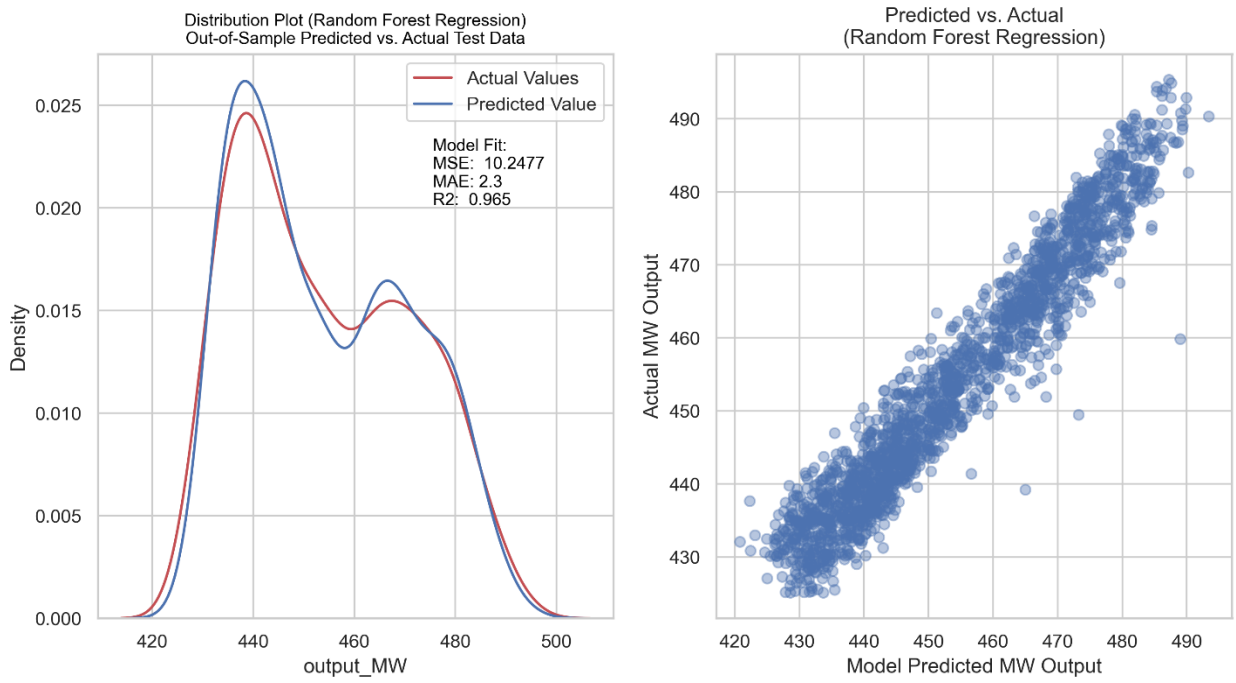

Figure 8 Random Forest Regression Distribution of Actual and Predicted Values

The error of each prediction was calculated as a percent of the actual value and the distribution of errors was plotted on a kernel density estimate plot. The figure below shows the 95% confidence interval of errors, showing a +1.5% and - 1.4% error when using the Randon Forest Regression model, an improvement from the Linear Regression model.

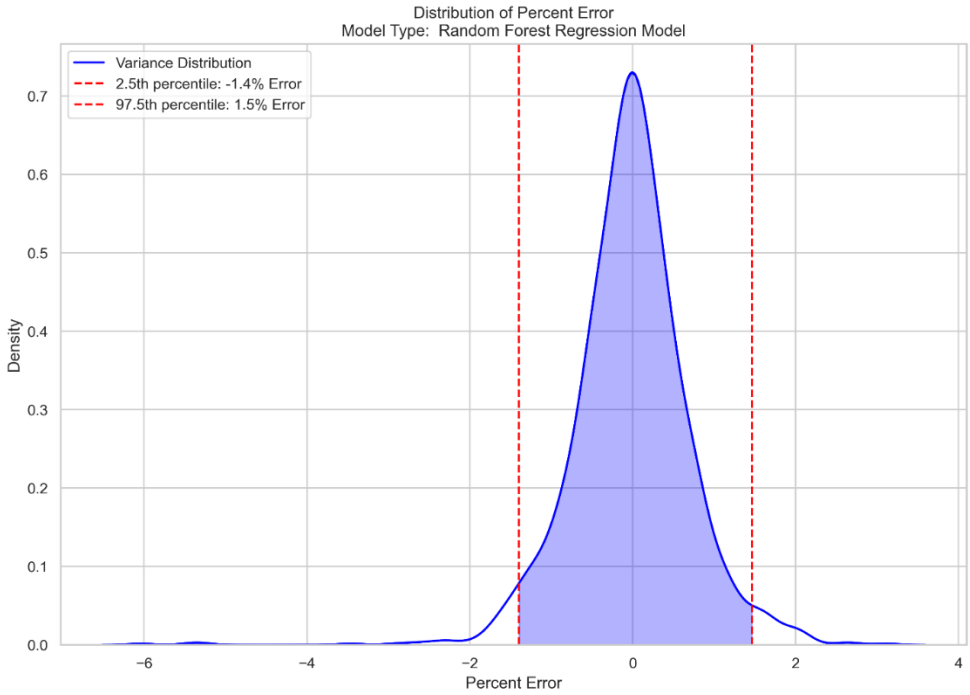

Figure 9 Random Forest Regression Error Distribution

5.4 Model Comparison

Models were evaluated using R-Squared, Mean Absolute Error and Mean Squared Error. In addition, the residuals were calculated for all predicted values and compared to the actual values to assemble an error distribution. A confidence interval of 95% was selected to compare the error bands from each of the models' predictions.

All three (3) models performed well with R-squared values all above 0.90. The Random Forest Regression model had the highest R-squared value and produced an error band of +1.5% and -1.4%.

Model Type	R Squared	Mean Absolute Error	Mean Squared Error	Accuracy in 95% Confidence Interval
Linear Regression	0.932	3.556	19.674	+/- 1.8%
Gradient Boosting Machine	0.955	2.738	12.953	+/- 1.5%
Random Forest Regression	0.964	2.304	10.396	+1.5%/-1.4%

Table 3 Model Fit and Accuracy Comparison

6 Discussion of Results and Next Steps

The evaluation of combined cycle plant output in relation to ambient site conditions and turbine backpressure revealed key insights. It was found that the output of the combined cycle plant is influenced by the ambient dry bulb temperature, ambient wet-bulb temperature, and turbine back

pressure. The dataset was thoroughly reviewed for missing values and outliers using box plots. The number of outliers was minimal, and they were retained in the dataset to ensure robust modeling. Three distinct types of models were assessed, each demonstrating excellent predictive performance. Specifically, the Multiple Linear Regression model achieved an R^2 value of 0.932, with predictions deviating by only 1.8% from actual MW output 95% of the time. The Gradient Boosting Machine model improved this performance with an R^2 value of 0.955, and predictions were within 1.5% of actual MW output 95% of the time. The Random Forest Regression model provided the best fit, achieving an R^2 value of 0.965, with predictions within 1.4% of actual MW output 95% of the time. Based on these results, the Random Forest Regression model is recommended for predicting combined cycle plant performance due to its superior accuracy and reliability.

Although the Random Forest Regression model has achieved an impressive prediction accuracy within ±1.4% of the actual output, there remains potential for further refinement in estimating plant performance. Future work could involve the collection and analysis of additional plant operational data to enhance model accuracy. Specifically, integrating data on evaporative cooling operation status, gas turbine air inlet filter differential pressure, and HRSG blowdown percent or valve position could provide valuable insights. Furthermore, incorporating measurements of high and intermediate pressure steam, as well as temperature, and monitoring steam bleeds or cycle heating circuits that may affect plant output, could further improve the model. An additional review and correlation analysis of these factors could uncover new relationships and refine the predictive capabilities of the model. These steps would contribute to a more comprehensive understanding of the variables influencing plant performance and enhance the reliability of output predictions.

7 Acknowledgements

This work relies on a dataset that was made available on Kaggle from Mr. Aagman Bhatia (https://www.kaggle.com/aagmandeep) with associated references (noted above). Mr. Bhatia's contribution is appreciated.

Dataset source: https://www.kaggle.com/datasets/aagmandeep/combined-cycle-power-plant-dataset-and-prediction