Predicting Electric Output of Combined Cycle Power Plants Using Ensemble Machine Learning Methods

Will Bodeau
Department of Atmospheric and Oceanic Sciences
University of California, Los Angeles

December 21, 2024

Abstract

Predicting the electric output of combined cycle power plants is critical for optimizing energy management and maximizing the efficiency of converting fossil fuels into electricity. This study explores the application of ensemble machine learning methods to forecast net hourly electric output using ambient variables such as temperature and pressure. Using data collected from a Combined Cycle Power Plant operating at full load between 2006 and 2011, an ensemble framework combining ridge regression, random forest, support vector regression, and k-nearest-neighbors was developed and compared to a meta-learner utilizing lasso regression to predict future plant electrical output based on ambient conditions. The model achieved an R-Squared value of 95.5% and a Root Mean Squared Error (RMSE) of 3.62. These findings demonstrate the potential of ensemble learning to improve predictive accuracy in power plant operations, offering valuable insights into key factors that influence performance and help plant operators maintain optimal operating conditions.

1 Introduction

The demand for efficient and sustainable energy production has become increasingly critical in the face of growing global energy needs and environmental concerns. Combined cycle power plants (CCPPs), which utilize both gas and steam turbines to generate electricity, are a cornerstone of modern energy production due to their high efficiency and reduced greenhouse gas emissions compared to traditional fossil fuel power plants. Combined cycle power plants utilize excess heat from gas turbines to power steam turbines, providing higher fuel efficiency. In a gas turbine, natural gas is mixed with air and ignited, spinning a generator. A combined cycle power plant captures exhaust heat from this process through a Heat Recovery Steam Generator (HRSG), which generates steam that is then delivered to a steam turbine. This process allows combined cycle power plants to be up to 50% more efficient than a traditional power plant (General Electric, n.d.). However, their performance is significantly influenced by ambient conditions such as temperature, atmospheric pressure, and humidity, making accurate prediction of power output a vital task. Effective forecasting of electric output is essential for optimizing energy management, and ensuring the economic and environmental sustainability of power plants. It is both economically and environmentally advantageous to utilize the potential energy in fossil fuels more effectively.

Traditional methods of performance analysis, often based on thermodynamic modeling, can be computationally intensive and less adaptable to varying environmental conditions and factors. Machine learning methods are capable of addressing these limitations by offering scalable, data driven solutions which can adapt to complex patterns in large or varied data sets. Power plants can analyze electrical output based on their own in-situ measurements, providing greater insights into optimal operating conditions at a specific power plant.

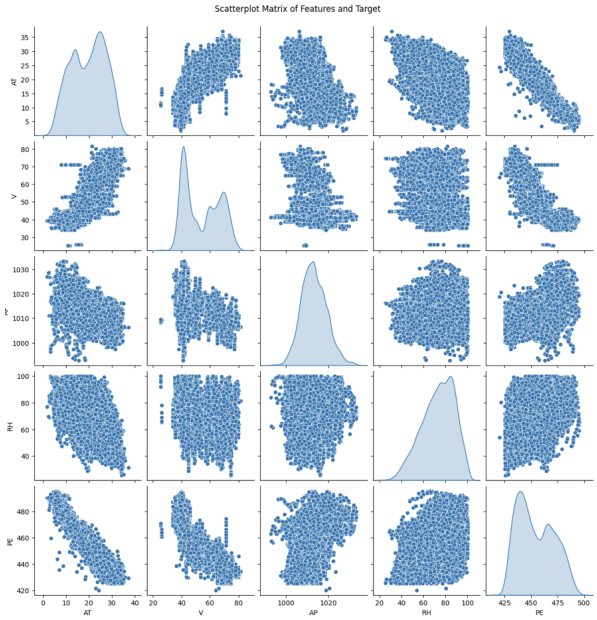
This study aims to address the challenge of accurately predicting the net hourly electric output of CCPPs by utilizing an ensemble machine learning approach. The objectives of this research are threefold: (1) to develop and evaluate an ensemble framework that combines multiple regression techniques for improved predictive accuracy, (2) to analyze the importance of ambient variables in influencing plant performance, and (3) to provide actionable insights that can help operators optimize plant operations under varying conditions. By leveraging ensemble learning methods, this study seeks to advance predictive modeling in energy systems and contribute to the broader adoption of data-driven approaches in the energy sector.

2 Exploratory Data Analysis

The scope of this work includes using data collected from a CCPP operating at full load between 2006 and 2011, which provides a comprehensive dataset for model training and evaluation. Average conditions are recorded from sensors around the power plant that record ambient variables every second, and directly influence the thermodynamic efficiency of the power plant. The features consist of hourly average ambient variables:

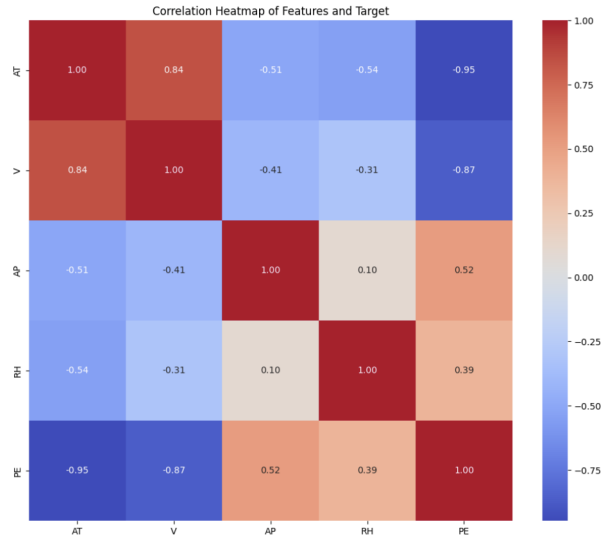
Feature	Range
Ambient Temperature (T)	1.81 °C to 37.1125 °C
Exhaust Vacuum (V)	25.36 to 81.56 cm Hg
Ambient Pressure (AP)	992.89 to 1033.30 milibar
Relative Humidity (RH)	25.36% to 100.16%
Net Hourly Electrical Output (PE)	420.26 to 495.76 MW

The dataset was examined for missing or inconsistent values, but no significant data cleaning was required.



Analysis:
The histogram of PE suggests a bimodal distribution with two prominent peaks around 440 and 470. This implies that the electrical output clusters around two primary values rather than being normally distributed. This could indicate that the power plant operates in two distinct modes or under different conditions, possibly due to variations in demand, input variables, or operational settings. Similarly, variables such as Exhaust Vacuum and Ambient Temperature also exhibit bimodal distributions. Net Hourly Electrical Output and Ambient Temperature show a direct negative linear relationship.

Figure 1: Scatterplot Matrix of Features.



Analysis:
The correlation heat map validates the observed relationship, showing a -94.8% correlation between Ambient Temperature (AT) and Net Hourly Electrical Output. This highlights that AT is inversely related to electrical output such that electrical output decreases at a linear rate as ambient temperature increases. Additionally, the heat map suggests that AT will likely be the most important predictor of electrical output in future predictive models. Understanding this strong negative correlation can help refine feature selection and improve model performance.

Figure 2: Correlation Heat Map.

Feature	Correlation with PE	VIF
PE	1	-
AP	0.518	66.618
RH	0.389	40.704
V	-0.869	74.969
AT	-0.948	39.157

VIF is defined as the inverse of tolerance:

$$VIF = \frac{1}{R_i^2}$$

The Variance Inflation Factor quantifies the inflation in coefficients due to multicollinearity. $VIF = 1 \implies$ no collinearity, and $VIF > 10 \implies$ higher collinearity. All features' variance exceeds the standard threshold of ten, indicating high correlation between features. AT was removed from the model given its high VIF and -94.8% correlation, but led to an increase in residuals due to a loss of predictive power. This is because the ambient conditions are not independent of each other, and the collinear variables collectively explain a significant portion of electrical output's variance. Removing one of the ambient conditions resulted in useful information being discarded due to interactions between features being removed, and less trends captured in the model. Rather than feature removal, regularization is more desirable in order for the model to retain the relative influence of all features, and facilitates relationships relevant under specific operating conditions.

3 Model Development

The data was split into training and testing data using an 80-20 ratio, and an ensemble learning method was utilized for this problem.

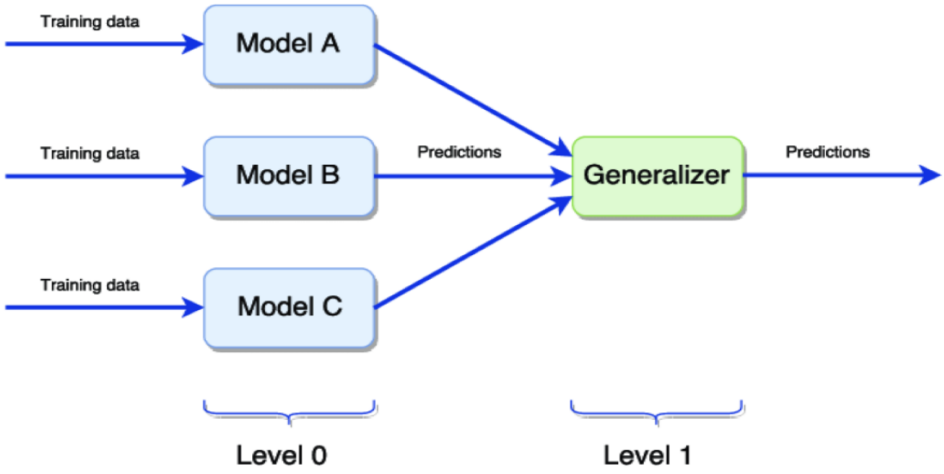


Figure 3: Illustration of stacking. Source: Analytics Vidhya (2020).

A stacked generalization model combines predictions from multiple base learners to improve the overall performance, and uses a meta-learner to best combine the predictions of the base models by minimizing errors. Stacking leverages the strengths of multiple

models, reducing overfitting compared to a single model, making it more generalizable based on different power plants' unique data and trends.

Table 1: Comparison of Training and Test Metrics

Metric	**Training**	**Test**
Mean Squared Error (MSE)	1.759	10.567
Root Mean Squared Error (RMSE)	1.326	3.250
R-squared (R^2)	0.993	0.963

Table 1 summarizes the training and test metrics for the initial model, which included linear regression, ridge regression, decision tree, random forest, support vector regression (SVR), k-nearest neighbors (KNN), and a meta-learner using ridge regularization. The model experienced overfitting, as indicated by the large difference in performance metrics between the training and test data (e.g., an R^2 difference of 3% and an MSE difference of 8.81). To address this, the ensemble was restricted to a diverse set of base learners (ridge regression, random forest, SVR, and KNN), as each base learner can can capture different patterns in the data. Regularization was added for each method in the ensemble, as well as feature scaling for SVR and KNN. These adjustments aim to reduce redundancy and improve generalization.

1. Base Learners

Ridge Regression (f_1)

$$\hat{y}_1 = f_1(X) = \hat{\beta}_\lambda^{ridge} = \arg\min_\beta (y - X\beta)^T (y - X\beta) + \lambda ||\beta||^2$$

The L_2 norm penalty is $||\beta||^2 = \sum_{j=1}^{p} \beta_j^2$. The tuning parameter $\lambda > 0$ controls the trade-off between regression fitting and coefficient shrinkage. The regularization strength λ is set to 1.0. As λ approaches ∞, β approaches zero.

```
('ridge', Ridge(alpha=1.0))
```

Listing 1: Ridge Regression

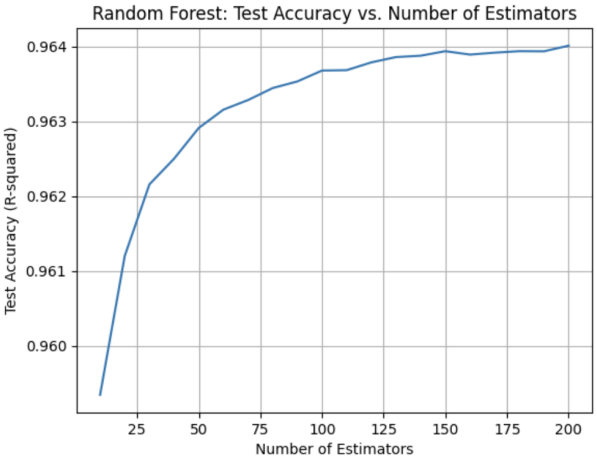
Random Forest (f_2)

$$\hat{y}_2 = f_2(X) = \frac{1}{T} \sum_{t=1}^{T} h_t(X),$$

where h_t are the individual decision trees in the forest and T is the total number of trees. A random forest is desirable due to its easy interpretability and automatic stepwise variable selection. However, it suffers from high variance, which is why it is balanced by other methods in the ensemble.

```
('rf', RandomForestRegressor(n_estimators=100, max_depth=5))
```

Listing 2: Random Forest



Analysis:
Random forest is regularized by limiting the maximum tree depth, and restricting the number of estimators to 150. After 150 estimators, the test accuracy begins to plateau, so 150 estimators balances computational cost and meaningful accuracy increase.

Figure 4: Accuracy Across Estimators.

Additionally, the random forest function in SciKitLearn does not use an impurity function to determine optimal tree splits, but uses a variance reduction technique, minimizing mean squared error, defined as

$$\Delta Var = Var_{parent} - (\frac{n_{left}}{n_{total}} + Var_{left} + \frac{n_{right}}{n_{total}} Var_{right})$$

Support Vector Regressor (SVR, f_3)

$$\hat{y}_3 = f_3(X) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

where α_i, α_i^* are Lagrange multiplies, the radial basis function kernel is defined as $K(x_i, x) = exp(-\gamma ||x_i - x||^2)$. $\gamma = \frac{1}{p \cdot Var(X)}$ and controls the influence of a single training example. Additionally, the data was standardized using $x_j^{scaled} = \frac{x_j - \mu_j}{\sigma_j}$

The SVR algorithm minimizes $\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$ subject to the constraints $y_i - f(x_i) \leq \epsilon + \xi_i^+$ and $f(x_i) - y_i \leq \epsilon \xi_i^-$. C is the regularization parameter for this model.

```
('svr', Pipeline([
    ('scaler', StandardScaler()),
    ('svr', SVR(C=1.0, kernel='rbf', gamma='scale'))]))
```

Listing 3: Support Vector Regression

C controls the balance between model complexity and error toleration. When C is small, the model tolerates larger ϵ_i, and a large C results in a model fitting the training data more closely. $C = 1$ allows the model to minimize training errors, but still prevents extreme overfitting.

k-Nearest Neighbors (KNN, f_4)

$$\hat{y}_4 = f_4(X) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(X)} y_i,$$

where $\mathcal{N}_k(X)$ is the set of indices for the k-nearest neighbors of X.

```
    ('knn', Pipeline([
        ('scaler', StandardScaler()),
        ('knn', KNeighborsRegressor(n_neighbors=50))]))
```

Listing 4: K-Nearest Neighbors

A larger k, represented as the $n_neighbors$ parameter, helps avoid overfitting, as the prediction is based on a larger number of neighbors, resulting in a smoothing effect that reduces the impact of noise. A k of 50 was chosen during the process of hyperparameter tuning:

Table 2: Comparison of Training and Test Metrics

Metric	**Training**	**Test**
Mean Squared Error (MSE)	15.747	15.706
Root Mean Squared Error (RMSE)	3.968	3.963
R-squared (R^2)	0.945	0.945

However, a k of 5 was ultimately chosen because the metrics when considering 50 neighboring points eliminated nearly all aspects of overfitting, it had higher RMSE and less precision as seen in Table 2. The needs of this model are not widely generalizable, as the needs of an individual power plant are unique to that system. Therefore, precision for this data (smaller k) was preferred to more generalizable results (larger k).

2. Meta-Learner (Lasso Regression)

The meta-learner f_m (Lasso regression) takes the predictions from the base learners as inputs. Lasso stands for Least Absolute Shrinkage and Selection Operator, which uses L_1 regularization to shrink less important coefficients to zero.

Let $X \in \mathbb{R}^{n\times 4}$ be the matrix of base learner predictions:

$$X = \begin{bmatrix} \hat{y}_1^{(1)} & \hat{y}_2^{(1)} & \hat{y}_3^{(1)} & \hat{y}_4^{(1)} \\ \hat{y}_1^{(2)} & \hat{y}_2^{(2)} & \hat{y}_3^{(2)} & \hat{y}_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_1^{(n)} & \hat{y}_2^{(n)} & \hat{y}_3^{(n)} & \hat{y}_4^{(n)} \end{bmatrix},$$

where $\hat{y}_j^{(i)}$ is the prediction of the j-th base model for the i-th sample.

Lasso regression is trained on X to minimize:

$$\hat{\beta}^{lasso} = \arg\min_{\beta}(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1 \text{ where } ||\beta||_1 = \sum_{j=1}^{p} |\beta|_j|$$

This provides automatic feature selection, but in its role as the meta-learner, it selects the most important base models based on their contributions to the final prediction. Lasso assigns a weight to each base model's prediction, and shrinks it to zero if it does not contribute significantly to improving the accuracy of the final prediction. This means only the base models with the most useful predictions are retained.

```
    final_estimator = Lasso(alpha=0.1)
```

Listing 5: Lasso Regression

An alpha value of 0.1 was chosen as it applies moderate regularization. The meta-learner balances minimizing prediction error and prevents overfitting, which is encouraged by redundant base learners. Lasso regression will shrink coefficients to near-zero to make them less impactful to the ensemble. This helps generalization for the model, since new ambient temperature data may exhibit different patterns or trends in the data. Thus, these different trends may benefit from certain models in the ensembles over others.

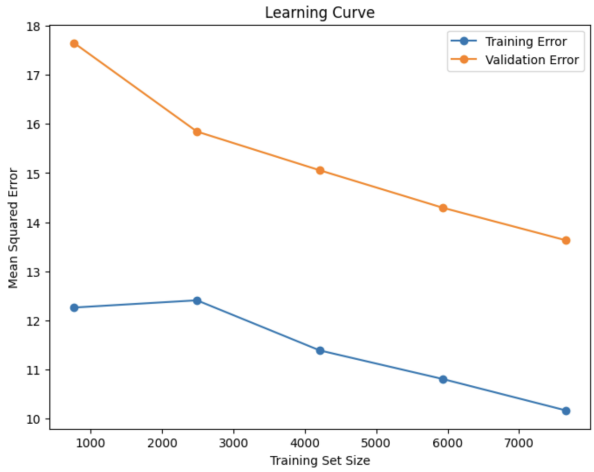
The final stacked model prediction $\hat{y}$ is:

$$\hat{y} = f_m(X) = X\boldsymbol{\beta}_m.$$

```
reg = StackingRegressor(
    estimators=estimators,
    final_estimator=final_estimator)
```

Listing 6: Stacking Regressor

4 Results



Analysis:
Training Error generally decreases as the training set size increases, which is expected because the model can better generalize with more data, and starts to stabilize with a larger training set size. Validation Error decreases as the training set size increases, indicating improved model performance on unseen data as more training data is used. However, a distinct gap between training and validation errors implies some degree of overfitting.

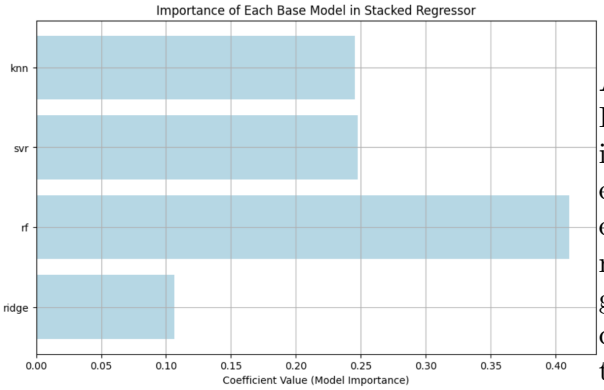
Figure 5: Error Over Training Set Size.

Table 3: Comparison of Training and Test Metrics

Metric	**Training**	**Test**
Mean Squared Error (MSE)	10.4691	13.0994
Root Mean Squared Error (RMSE)	3.2356	3.6193
R-squared (R^2)	0.9641	0.9548

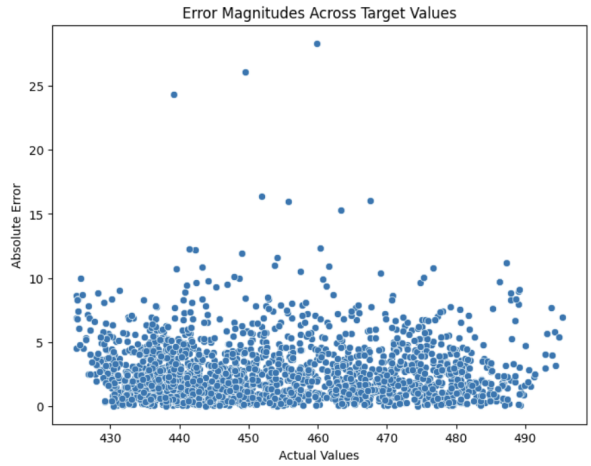
After introducing regularization and removing redundant methods from the ensemble, the model was no longer overfitting to the training data. There was a <1% difference

between the R-Squared of the training and testing data, compared to 3% before regularization. The vertical distance between training and validation error in Figure 3 is quantified in Table 2.



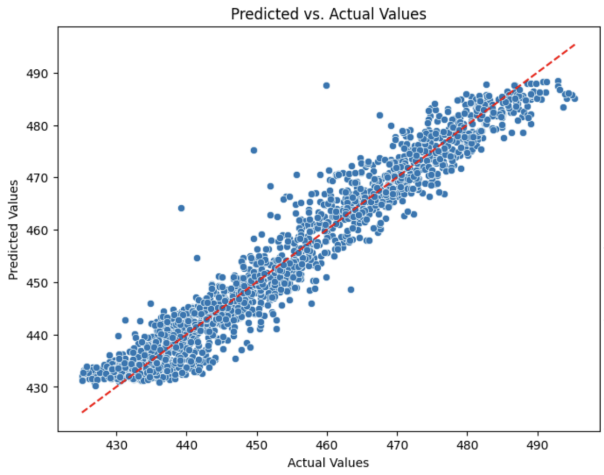
Analysis:
Random Forest is the most important model in the ensemble, with a weight of 40%. It explains a large fraction of the variability of electrical output. KNN and SVR contribute roughly equally to the ensemble, and Ridge Regression contributes only 10%. However, its inclusion helps increase bias and reduce variance, thereby improving generalizability.

Figure 6: Model Importance in Ensemble.



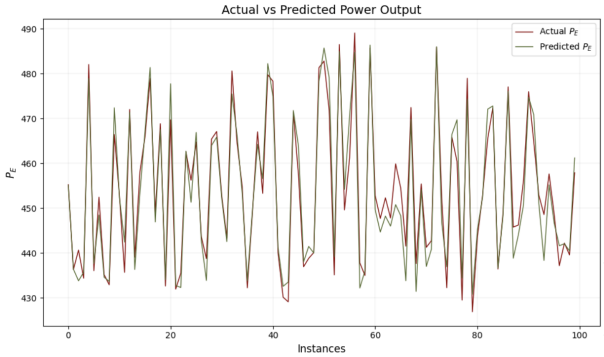
Analysis:
The majority of the absolute errors appear to cluster at lower values, indicating that the model performs reasonably well for most predictions. Outlying errors, specifically when Absolute Error $>$ 20 that may be indicative of poorly predicted data points, or regions where the model struggles.

Figure 7: Error Magnitude of Predictions.



Analysis:
The density of points near the line demonstrates low error for the majority of predictions. There are a few outliers that deviate more significantly from the line, representing predictions with higher errors. The model has minor tails on both ends with higher error rates, meaning that it struggles at both extremes when electrical output is either approaching its maximum or minimum range. This is due to the fact that the residual variance is not uniform across all predictions.

Figure 8: Predictions vs. Actual Values



Analysis:
This model has a strong fit, with predicted values that closely follow actual values. There are no significant deviations between the two lines, aside from minor variations due to noise. This variation due to noise serves as a sign of a good model, indicating a lack of overfitting to the training data.

Figure 9: Predictions vs. Actual Values

Shapley Additive Explanations provide insights into feature importance, and the relationship between feature values, and their impact on the model's predictions. SHAP values represent the impact of each feature on power output. A Shapley value is computed by taking the average difference from all combinations of all features, excluding the feature whose importance is being determined. Fadel, 2022, "Explainable Machine Learning, Game Theory, and Shapley Values: A technical review" SHAP value for a feature f_i in a specific prediction $\hat{y}$ is defined as

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

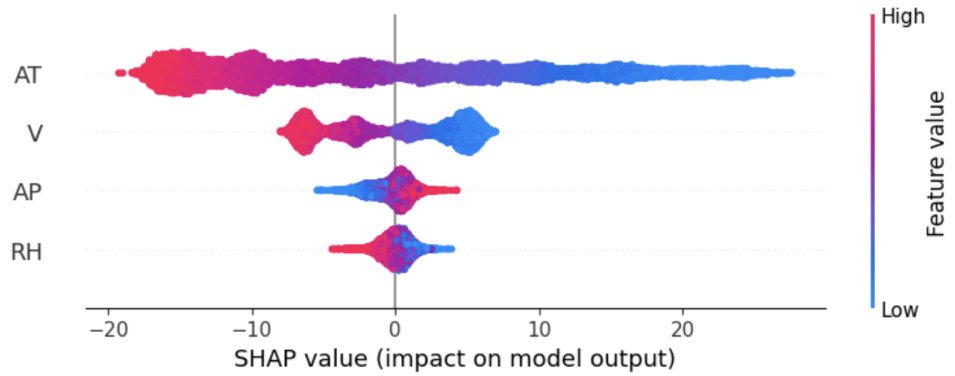
Where:

- $\phi_i(f)$ is the SHAP value for feature f_i.
- S is a subset of features from the set of all features N excluding feature f_i.
- $|S|$ is the size of subset S, and $|N|$ is the total number of features.
- $f(S)$ is the model prediction using the feature set S.
- $f(S \cup \{i\})$ is the model prediction with feature f_i added to subset S.

The final model prediction $\hat{y}$ can be expressed as:

$$\hat{y} = \bar{f} + \sum_{i=1}^{N} \phi_i(f)$$

SHAP values are "a more advanced interpretation system" (Wang, 2022). The purpose of these values is to enhance the transparency of the ensemble model, and provide a direct visualization of the impacts of various ambient conditions on electrical output. This is especially useful in an ensemble model, where interpretability becomes lost with the use of black box models.



Positive SHAP values indicate that a feature increases the model's output, while negative values decrease it. The further the SHAP value is from zero, the greater the influence of that feature on the prediction.

- High Ambient Temperature (AT) values (red) lead to negative SHAP values, reducing power output.
- Low AT values (blue) result in a positive impact, increasing power output.
- Higher Exhaust Vacuum (V) values (red) contribute negatively to power output, while lower values (blue) tend to increase it.
- The effect of Ambient Pressure (AP) is centered around zero, suggesting a relatively weak or balanced impact. However, extreme values show slight variations in SHAP values. High ambient pressure slightly increases power output, and low AP decreases power output.
- Lower Relative Humidity (RH) values (blue) tend to have a slightly positive impact on predictions, while higher RH values (red) cause small negative impacts. Nevertheless, RH has a relatively minor influence compared to AT or V.

AT and V emerge as the dominant predictors, with broader SHAP value distributions and clearer color separations. In contrast, AP and RH show less impact, as indicated by their smaller SHAP value ranges. Overall, power output is primarily influenced by Ambient Temperature and Exhaust Vacuum, suggesting that controlling or monitoring these factors could significantly affect power generation. Ambient Pressure and Relative Humidity have weaker effects, making them less critical for optimization or monitoring efforts.

5 Discussion

Studies used this data to explore strengths and weaknesses of numerous machine learning models. Tüfekci applies 15 regression methods in "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods" and then a series of ANOVA tests to determine the best subset selection and best regression method for this data. This study employs subset selection, while I employed regularization. This is an important difference, since subset selection is a greedy algorithm that does not account for the heavy collinearity seen in these features. It chooses the optimal feature during that specific step in subset selection, which does not account for complex relationships between features. For example, subset selection may select the feature X_1 based on its selection criteria, and then choose X_3 based on its interaction with X_1. However, the model will not have considered the relationship between X_2 and X_3, which could have resulted in a better model than X_1 and X_3. The ensemble method in this study performs roughly equal to the 15 regression methods, achieving an RMSE lower than all 15 models.

The SHAP plots justify why certain features should be prioritized or discarded in future models. For example, Relative Humidity (RH) and Ambient Pressure (AP) show relatively low impact, which could lead to a discussion on the necessity of monitoring them or regulating them in the power plant. Though computationally expensive, SHAP visualizations serve as an essential aspect for decision-makers and non-technical stakeholders, plant managers, engineers, and policy makers, who need to trust model predictions in high-stakes environments like power plant management. SHAP plots enable decision makers to understand the rationale behind operational recommendations of "black box" machine learning models. This transparency is essential in high-stakes environments, where decisions based on model outputs can impact safety, efficiency, and profitability. For instance, if SHAP values indicate that Ambient Temperature (AT) has a high impact on power output, operators can prioritize temperature management strategies. Conversely, if Relative Humidity (RH) shows low impact, stakeholders might consider reallocating resources away from monitoring RH, reducing sensor costs, and focusing on more influential variables. This level of interpretability builds trust in the model, making it easier to integrate machine learning insights into daily decision-making processes.

A limitation of this ensemble method is its sensitivity to noisy or imbalanced data. If sensor measurements in power plants are prone to inaccuracies or missing data, the ensemble model might amplify these issues, leading to biased or unreliable predictions unless robust preprocessing techniques are employed. Ensemble models require significant memory resources, and can be problematic for power plants with limited hardware capacity or when integrating the model into legacy systems.

6 Conclusion

The scope of this model can be expanded to combined cycle power plants equipped with similar ambient condition sensors in order to optimize electrical production based on real-time measurements. By leveraging all features and applying regularization rather than feature removal, the model allows power plants to fine-tune the ensemble method to their specific needs. Additionally, traditional thermodynamic models are computationally expensive for large or complex systems and rely on rigid physical assumptions, making this approach more flexible. Machine learning methods, once trained, can make

rapid predictions when real-time predictions or optimizations are required. Expanding this model to different types of power plants, such as gas-turbine, hydroelectric, or solar, could validate its generalizability and robustness across various energy production methods. Future research could investigate how the model performs under different operational and environmental conditions, such as extreme temperatures or fluctuating energy demands. Additionally, incorporating external factors like fuel costs, market demand, and regulatory policies could enhance predictive accuracy, allowing operators to balance cost-efficiency and sustainability goals.

References

1. Analytics Vidhya. (2020, December 29). *Improve your predictive model score using stacking regressor*. Retrieved from `https://www.analyticsvidhya.com/blog/2020/12/improve-predictive-model-score-stacking-regressor/`

2. Fadel, S. (2022). Explainable Machine Learning, Game Theory, and Shapley Values: A technical review. *Statistics Canada*. Retrieved from `https://www.statcan.gc.ca/en/data-science/network/explainable-learning`

3. General Electric. (n.d.). *Combined cycle power plants*. GE Vernova. Retrieved from `https://www.gevernova.com/gas-power/resources/education/combined-cycle-power-pla`

4. Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *Electrical Power & Energy Systems*, 60, 126–140. `https://doi.org/10.1016/j.ijepes.2014.02.027`

5. Wang, D., Thunéll, S., Lindberg, U., Jiang, L., Trygg, J., & Tysklind, M. (2022). Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management, 301*, 113941. `https://doi.org/10.1016/j.jenvman.2021.113941`