JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# POWER LOAD PREDICTION USING MULTIPLE LINEAR REGRESSION

## *IMPROVING GRID EFFICIENCY AND RESOURCE ALLOCATION*

**[1]Kiran Basavannappagowda, [2]Jaswanth RS**
[1,2] Students
[1,2] MTech in AI
[1,2]Reva University, Bengaluru, KA, India

***Abstract :*** This paper explores how statistical analysis techniques can be leveraged to enhance predictions across diverse domains. Specifically, we focus on comparing different statistical methods to develop a model that accurately predicts the power output of a combined cycle power plant. This prediction is influenced by factors such as atmospheric pressure, humidity, steam pressure, and temperature, which serve as inputs for our model. We analyze various statistical techniques, including multiple linear regression, and employ methods like forward selection to refine our model's accuracy. The objective is to create a more precise predictive model for power output. Our implementation is conducted using the Python programming language due to its extensive array of statistical tools. Utilizing Python enhances the scalability and flexibility of our model, thereby increasing its effectiveness in real-world scenarios..

***IndexTerms* - Predictive analysis, power output, Multiple Linear Regression, Python**

## I. INTRODUCTION

Electricity stands as an indispensable resource for humanity, meeting the vital needs of communities in towns through power plants situated across various locations. However, the escalating occurrence of power fluctuations presents a formidable challenge, arising from factors such as environmental shifts, power surges, and mismanagement. In response to these challenges, combined cycle power plants have emerged as a solution, harnessing both gas and steam turbines to optimize electrical energy generation from a unified fuel source.

In the domain of statistical analysis, Linear Regression serves as a fundamental model, forecasting the value of a dependent variable based on an independent variable. When multiple independent variables influence a single variable, the model evolves into Multiple Linear Regression. This advanced model predicts the output value by considering the most influential independent variables, which may be singular or multiple.

The application of Multiple Linear Regression traditionally entails intricate mathematical techniques, requiring significant effort. However, with the emergence of modern tools such as Python and R Programming, this task has become more accessible. Moreover, these models can accommodate categorical independent variables, where non-numeric data is present. The conversion of such categorical attributes into numerical values is known as Label Encoding.

Mathematically, Multiple Linear Regression is depicted as follows:
$Y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \beta4X4 + \ldots + \beta nXn$
Here, X1, X2, X3…Xn represent independent variables, and Y represents the dependent variable. β0, β1, β3.. βn are algorithm parameters. The objective of the algorithm is to determine the parameter values that yield the best-fitted line for the given dataset.

## II. EXPLORING INSIGHTS AND ANALYSIS IN COMBINED CYCLE POWER PLANTS (CCPPS)

Understanding and Optimizing Combined Cycle Power Plants (CCPPs): The Challenge: Accurately predicting CCPP output often requires complex models that account for a wide range of variables. This complexity has driven researchers to explore alternative methods for enhancing the prediction process.
Statistical Analysis: Statistical analysis offers numerous benefits in various applications, ranging from business and finance to healthcare and scientific research. By leveraging statistical techniques, organizations can derive valuable insights from data, make informed decisions, and drive evidence-based strategies. For instance, in healthcare, statistical analysis enables researchers

to analyze clinical trial data, identify treatment efficacy, and assess patient outcomes. Similarly, in business, statistical analysis aids in market research, forecasting sales trends, and optimizing marketing strategies. Overall, statistical analysis empowers professionals across diverse fields to extract meaningful patterns, quantify uncertainties, and ultimately enhance decision-making processes for better outcomes.

Beyond Prediction: Gas Turbine Analysis: Researchers like Niu et al. focus on applying linearization methods to gas turbines within CCPPs. Such analysis deepens our understanding of turbine behavior and its impact on system output.

Key Gas Turbine Factors: The load on a gas turbine is intrinsically linked to environmental and operational parameters such as temperature, exhaust vacuum, ambient pressure, and relative humidity. These factors directly influence the efficiency of the gas turbine and, consequently, power output.

The Gas Turbine's Role: Within a CCPP, a gas turbine plays the crucial role of compressing air and mixing it with high-temperature fuel. This combustion drives the turbine blades, which then rotate a generator to produce electricity.

Harnessing Waste Heat: The CCPP design cleverly utilizes the heat exhausted from the gas turbine. This heat powers a Heat Recovery Steam Generator (HRSG) to produce steam, which spins a second turbine and generator, resulting in greater power generation efficiency.

## III. ANALYSIS & RESULTS

For our analysis, we opted to utilize the Python programming language due to its extensive support for various packages essential for implementing statistical analysis techniques. The dataset employed in this study was sourced from the UCI Repository. It comprises 9568 instances featuring five distinct attributes: Temperature (T), Exhaust Vacuum (V), Ambient Pressure (AP), Relative Humidity (RH), and the net hourly electrical energy output (EP) of the plant. Among these, EP is considered the dependent variable, while the remaining four attributes serve as independent variables.

To enhance the accuracy of our predictive model and facilitate further analysis, we derived several new features derived from the existing dataset:

Ratio of Pressure to Temperature (AP/AT): This feature was engineered to explore the relationship between atmospheric pressure and ambient temperature.

Heat Index (HI): The calculation of the heat index allows us to evaluate the combined impact of temperature and humidity on the perceived temperature.

HI= c1+c2*T+c3*R+c4*T*R+c5*T^2+c6*R^2+c7*T^2*R+c8*T*R^2+c9*T^2*R^2
In above formula,
HI = heat index in degrees Fahrenheit
R = Relative humidity
T = Temperature in ∘F
c1 = -42.379
c2= -2.04901523
c3 = -10.14333127
c4 = -0.22475541
c5 = -6.83783 x 10−3
c6 = -5.481717 x 10−2
c7 = -1.22874 x 10−3
c8 = 8.5282 x 10−4
c9 = -1.99 x 10−6

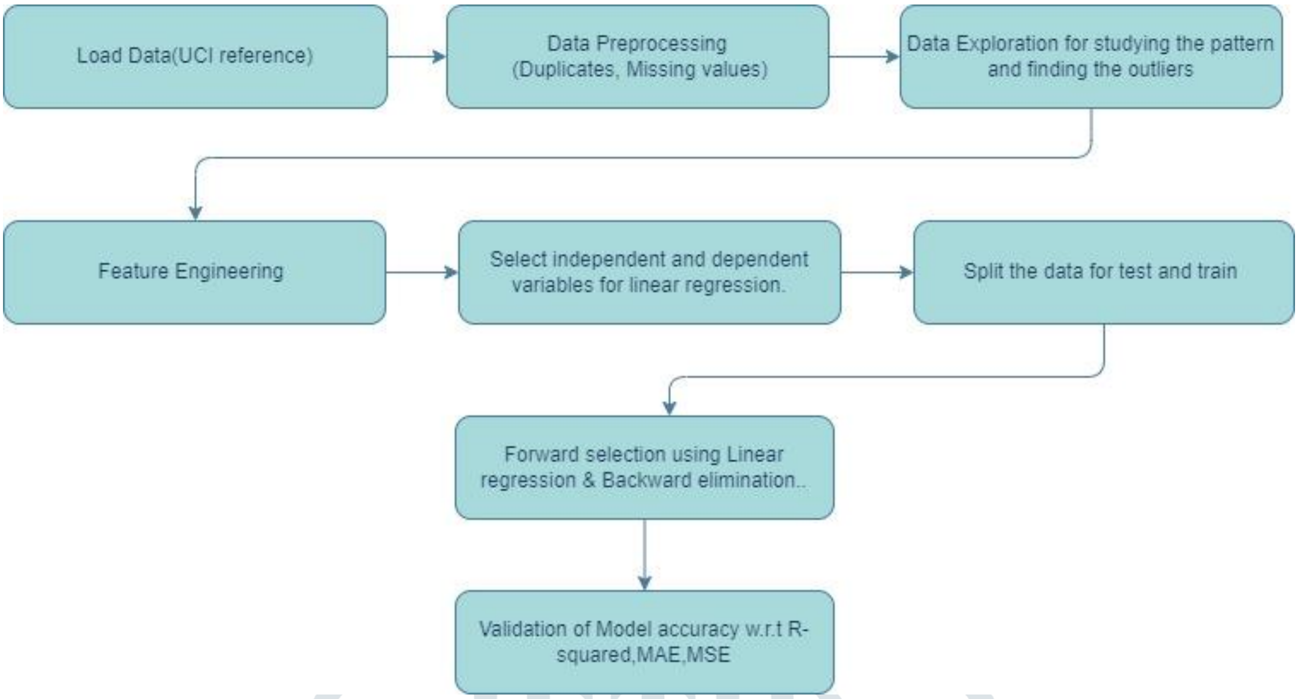
Temperature-Pressure Product (AT * AP): We introduced this interaction term to investigate how the product of temperature and pressure correlates with power efficiency.

Humidity-Temperature Interaction (RH * AT): This feature aims to capture the combined effect of humidity and temperature.

Normalized Temperature Difference Relative to the Logarithm of Vacuum Pressure (AT - HI + Log (AP)): This complex feature is designed to encapsulate the adjusted temperature effect, considering atmospheric pressure.

By incorporating these new features into our analysis, we aim to develop a highly accurate predictive model capable of effectively forecasting the net hourly energy output.

Working model flow:


Load Data(UCI reference)
Data Preprocessing (Duplicates, Missing values)
Data Exploration for studying the pattern and finding the outliers
Feature Engineering
Select independent and dependent variables for linear regression.
Split the data for test and train
Forward selection using Linear regression & Backward elimination..
Validation of Model accuracy w.r.t R-squared,MAE,MSE

The initial phase of any statistical analysis commences with data preprocessing, a pivotal stage aimed at ensuring the dataset's suitability for model training. This process commences by scrutinizing the data for any missing values. In instances where missing values are detected, common practices involve replacing them with the mean or median of the remaining values within the same feature. Fortunately, in the present dataset, no missing values were identified, thereby negating the necessity for such imputation techniques. Furthermore, data preprocessing encompasses the examination of categorical features; should any be present, they are transformed into numerical values through methods such as Label Encoding. However, in this specific dataset, categorical features are absent. Feature Scaling, another vital aspect of statistical analysis, serves to standardize all feature values within a uniform range. Although typically indispensable, the Python linear regression class employed in this study inherently manages feature scaling, thus eliminating the need for additional scaling procedures.

1. **Load Data**: The initial step involves loading the dataset, which contains various parameters that potentially influence power efficiency. Got dataset from UCI titled "Combined Cycle Power Plant" [1]

2. **Data Preprocessing for Duplicates, missing value check** : We perform data cleaning to ensure the quality of our dataset. This includes removing duplicates entries and handling missing values (null checks) to prevent any bias or inaccuracies in the subsequent analysis.
   **Handling Duplicates:**
   Duplicate entries can introduce redundancy and skew the results of subsequent analyses. In Python, the pandas library is commonly used for data manipulation, including the detection and removal of duplicate rows.The drop_duplicates() function in pandas identifies and removes duplicate rows from the DataFrame, providing a cleaner dataset for further analysis.
   **Handling Missing Values:**
   Missing values, or null values, are common challenges in datasets and can lead to biased or inaccurate analyses. The pandas library provides functions to identify and handle missing values.
   By performing these data preprocessing steps in Python, you ensure that your dataset is free from duplicates and handled appropriately regarding missing values. This sets the foundation for more accurate and reliable analyses and machine learning model training.

3. **Data Exploration for Studying the Pattern and Finding the Outliers**: Conducting exploratory data analysis (EDA) in Python is a crucial step to unravel the inherent patterns and distributions within the dataset. EDA serves as a powerful tool for gaining insights into the structure of the data, enabling a comprehensive understanding of its characteristics.

   **Exploratory Data Analysis (EDA):**
   In Python, libraries such as pandas, matplotlib, and seaborn are commonly employed for EDA. Visualizations and statistical summaries are utilized to examine data distributions, central tendencies, and dispersion. This aids in identifying any notable trends, variations, or irregularities that may influence subsequent analyses.

**Identifying and Addressing Outliers:**
Outliers, being data points significantly deviating from the overall pattern, can distort analysis outcomes. In Python, visualization techniques like box plots and statistical methods help pinpoint outliers for further examination.
The seaborn library's boxplot function, for instance, visually represents the distribution of a specific column, aiding in the identification of outliers. Addressing outliers can involve various strategies, such as removing them, transforming the data, or applying robust statistical methods. By engaging in comprehensive EDA in Python, you not only gain a nuanced understanding of your data but also ensure that outliers, if present, are recognized and appropriately managed. This sets the stage for more robust and accurate analyses and enhances the reliability of subsequent modeling and decision-making processes.

4. **Feature Engineering**: Feature engineering is a critical process in data preprocessing that involves crafting new variables to enhance the model's ability to capture intricate relationships within the dataset. This transformative step aims to extract meaningful information from the existing features or generate entirely new ones, ultimately contributing to the model's predictive performance. We engage in feature engineering to create new variables that might better capture the relationships within the data. This step is iterative and may loop back from later stages if additional features are deemed necessary.

**Derived few features:**
Ratio of Pressure to Temperature (PT)
Heat Index (HI)
Temperature-Pressure Product (TP)
Humidity-Temperature Interaction (HTI)
Normalized Temperature Difference Relative to the Logarithm of Vacuum Pressure(THILOGV)

5. **Select dependent and independent variables based on correlation values**:
Variable selection based on correlation values involves assessing the strength and direction of relationships between different variables in a dataset. In Python, pandas and seaborn libraries are used for computing and visualizing correlation values
Compute Correlation Matrix: We used the corr() function in pandas to calculate the correlation matrix for all numerical variables in the dataset.

| | AT | V | AP | RH | PE | PT | HI | TP | HTI | THILOGV |
|---|---|---|---|---|---|---|---|---|---|---|
| **AT** | 1 | 0.843689 | -0.50822 | -0.54395 | -0.94791 | -0.8542 | -0.11865 | 0.999917 | 0.799831 | 0.795558 |
| **V** | 0.843689 | 1 | -0.41572 | -0.31221 | -0.8699 | -0.66023 | -0.03555 | 0.843621 | 0.775156 | 0.635744 |
| **AP** | -0.50822 | -0.41572 | 1 | 0.101631 | 0.518687 | 0.403604 | 0.089108 | -0.49818 | -0.51665 | -0.42081 |
| **RH** | -0.54395 | -0.31221 | 0.101631 | 1 | 0.391175 | 0.404903 | -0.2862 | -0.54629 | 0.033086 | -0.2171 |
| **PE** | -0.94791 | -0.8699 | 0.518687 | 0.391175 | 1 | 0.830933 | 0.207815 | -0.94734 | -0.85428 | -0.8125 |
| **PT** | -0.8542 | -0.66023 | 0.403604 | 0.404903 | 0.830933 | 1 | 0.499205 | -0.85542 | -0.77051 | -0.92426 |
| **HI** | -0.11865 | -0.03555 | 0.089108 | -0.2862 | 0.207815 | 0.499205 | 1 | -0.12069 | -0.36552 | -0.69535 |
| **TP** | 0.999917 | 0.843621 | -0.49818 | -0.54629 | -0.94734 | -0.85542 | -0.12069 | 1 | 0.798414 | 0.796749 |
| **HTI** | 0.799831 | 0.775156 | -0.51665 | 0.033086 | -0.85428 | -0.77051 | -0.36552 | 0.798414 | 1 | 0.802809 |
| **THILOGV** | 0.795558 | 0.635744 | -0.42081 | -0.2171 | -0.8125 | -0.92426 | -0.69535 | 0.796749 | 0.802809 | 1 |

Visualize Correlation Matrix: Visualize the correlation matrix using a heatmap to easily identify strong and weak correlations.
Select Variables Based on Correlation Threshold: Choose a correlation threshold that determines which variables to include or exclude. These features (AT, V, PT, TP, HTI, THILOGV) are selected based on correlation value which are highly influence on building accurate statistical model in identifying PE.

Systematically select variables based on their correlation values and make informed decisions about which features to include in analysis. Adjusting the correlation threshold allows for flexibility in the selection process based on the desired level of correlation strength.

6. **Split the Data for Test and Train:** To develop and evaluate predictive models, it's essential to divide your dataset into training and testing subsets. The model learns patterns from the larger training set and is then tested on the smaller, unseen testing set to assess its ability to generalize.
For our analysis, we've chosen an 80/20 split between training (7654 instances) and testing sets (1914 instances). This provides enough training data while allowing for robust evaluation. During training, the model learns from patterns in the data. The testing set then reveals how well it performs on new, unseen data – a key indicator of real-world success.

7. **Forward Selection Using Linear Regression**: We apply forward selection, a stepwise regression technique, to identify the most significant predictors for our linear regression model. Starting with the most correlated variable, we incrementally add features that improve the model's performance based on statistical criteria.

| Variables | MAE | MSE | R-squared |
|---|---|---|---|
| AT | 4.109 | 28.056 | 0.901 |
| AT, HTI | 3.654 | 22.148 | 0.922 |
| AT, HTI, V | 3.562 | 20.648 | 0.927 |
| AT, HTI, V, THILOGV | 3.514 | 19.989 | 0.929 |
| AT, HTI, V, THILOGV, TP | 3.496 | 19.677 | 0.931 |
| **AT, HTI, V, THILOGV, TP, PT** | **3.490** | **19.634** | **0.931** |

## IV. CONCLUSION

In this study, we applied forward selection , Backward elimination and standard scalar method to systematically identify the most influential predictor variables for predicting the power output (PE) of a system. The forward selection method sequentially added predictor variables to the regression model based on their contribution to reducing the root Mean Absolute Error (MAE) and residual Mean Squared Error (NSE).

Our analysis revealed that the combination of ambient temperature (AT), humidity-temperature index (HTI), vacuum (V), THILOGV(Normalized Temperature Difference Relative to the Logarithm of Vacuum), ambient pressure (AP), and turbine inlet temperature (TP) yielded the most optimal model for predicting power output. This model exhibited the lowest RMSE and RSS values, indicating superior predictive performance compared to other models evaluated in the study.

Furthermore, the coefficient of determination (R-squared) value of 0.931 suggests that approximately 93.1% of the variance in power output can be explained by the selected predictor variables. This high explanatory power underscores the significance of the identified variables in capturing the variability of power output in the system.

Overall, our findings highlight the importance of considering multiple environmental and operational factors, including ambient temperature, humidity, and pressure, when modelling power output prediction in similar systems. The insights gained from this study can aid in optimizing system efficiency, enhancing operational planning, and informing decision-making processes in energy management applications.

Future research could focus on validating the predictive performance of the identified model using additional datasets and exploring the potential integration of advanced modelling techniques to further improve accuracy and robustness.

## REFERENCES

**[1]** Dataset reference from UCI : https://archive.ics.uci.edu/dataset/294/combined+cycle+power+plant

**[2]** A. Dehghani Samani, “Combined cycle power plant with indirect dry cooling tower forecasting using artificial neural network,” Decis. Sci. Lett., vol. 7, no. 2, pp. 131–142, 2018.

**[3]** Elkhawad Elfaki , Ahmed Hassan Ahmed Hassan, “Prediction of Electrical Output Power of Combined Cycle Power Plant Using Regression ANN Model”, DOI: 10.5281/zenodo.1285164.

**[4]** V.Ramireddy, “An overview of combined cycle power plant”,2015,http:// electricalengineeringportal.com/an-overview-of-combined-cycle-power-plant

**[5]** D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, John Wiley & Sons, Hoboken, NJ,aUSA, 2012.

**[6]** L. X. Niu and X. J. Liu, “Multivariable generalized predictive scheme for gas turbine control in combined cycle power plant,” in 2008 IEEE Conference on Cybernetics and Intelligent Systems, 2008, pp. 791–796.

**[7]** H. H. Erdem and S. H. Sevilgen, “Case study: Effect of ambient temperature on the electricity production and fuel consumption of a simple cycle gas turbine in Turkey,” Appl. Therm. Eng., vol. 26, no. 2–3, pp. 320–326, Feb. 2006.

**[8]** H. Kaya, P. Tüfekci, and S. F. Gürgen, “Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine,” in International Conference on Emerging Trends in Computer and Electronics Engineering (ICETCEE 2012), 2012, pp. 13–18.

**[9]** Paper from JETIR title “AN EFFECTIVE MULTIPLE LINEAR REGRESSION MODEL FOR POWER LOAD PREDICTION” by A.Lakshmanarao,G.Vijay Kumar,,T.S.Ravi Kiran. JETIR September 2018, Volume 5, Issue 9