AN EFFECTIVE MULTIPLE LINEAR REGRESSION MODEL FOR POWER LOAD PREDICTION

A.Lakshmanarao[1],G.Vijay Kumar[2],T.S.Ravi Kiran[3]
[1]Associate Professor, [2]Assistant Professor, [3]Assistant Professor
[1,2] Department of CSE, [3]Department of CS
[1,2]Pragati Engineering College, Surampalem, AP, India
[3]P.B.Siddhartha College of Arts & Science,Vijayawada, AP, India

Abstract : ***Predictive analysis is the one of the major machine learning application. In this paper, we compare machine Learning Regression methods to propose an efficient model for predicting net hourly power output of the combined cycle powerplant. The power load of a power plant is effected by the parameters atmospheric pressure, humidity, and exhaust steam pressure, ambient temperature.These four parameters are taken as input parameters for the proposed model.Based on these parameters,we analyze different multiple linear regression machine learning methods and proposed an efficient model which gives better prediction of energy output of the power plant.In the proposed model,we applied forward selection, backward elimination techniques and implement a machine learning model which has lower standard error rate. The implementation was done in Python Language, which provides vast number of packages for machine learning algorithms.***

Index Terms – Predictive analysis, power output, Multiple Linear Regression, Python

I. INTRODUCTION:

The essential resource for human beings is electricity.Electric needs of population were maintained by power plants established in various places in towns.The power fluctuation problems are increasing day by day. The causes for fluctuations are several like environmental changes, overload of power or even misuse of power. A combined cycle power plant is an electric power plant in which a gas turbine and a steam turbine are used to produce more electrical energy from same fuel [3].

Linear Regression is a simple model of machine learning where dependent variable value is predicted based on independent variable value. If there is more than one independent variable affecting a single variable, then it is called as multiple linear regression. In this model, output value (dependent variable value) is predicted considering most significant independent variable.The most significant variable may be one or more. If more than one independent variable influences the dependent variable, then we may consider all those as significant [4]. Implementation of multiple linear regression involves a complex math-based method which will take some effort to be applied.But with latest tools like python,R Programming ,this task can be done easily. The independent variables can be categorical also. Categorical means the features may contain non numeric data. In such cases, we need to convert categorical attributes in to numerical values. This task is known as Label Encoding.

Multiple Linear regression can be represented using the following mathematical notation

$$Y= \beta_0+\beta_1X_1+ \beta_2X_2+ \beta_3X_3+ \beta_4X_4\ldots\ldots \beta_nX_n$$

$X_1,X_2,X_3\ldots X_n$ represents independent variables and Y represents dependent variable. $\beta_{0,}$ $\beta_{1,}$ $\beta_{3..}$ β_n are algorithm parameters. The aim of the algorithms to find the parameter values so that we find best fitted line for the given dataset.

II.RELATED WORK:

The correct prediction of combined cycle power plants using mathematical models needs more parameters [1]. An alternative method for analyzing power plant energy consumption is by using machine learning models which implicitly covers mathemartical models [2]. Kaya et al analysed different types of machine learning models to predict the load electrical power output of a power plant [7]. Niu et al applied linearization methods for analyzing the gas turbine in a combined cycle power plant . The load of a gas turbine is usually based on the parameters temperature, Exhaust Vacuum, Ambient Pressure, Relative Humidity [6]. The role of gas turbine is compression of air and mixing it with a fuel heated to a very high temperature. The mixture transforms through blades making them rotate. Electricity is generated by rotating the gas turbine[5]. The wastage heat is also useful for HRSG for producing stem that spins turbine again. This steam turbine drives a generator to produce more electricity.

III.EXPERIMENTATION & RESULTS:

In our experiments, we opted python language. Because Python language supports several packages for implementing machine learning techniques. The dataset used for this work is taken from UCI Repository. In this dataset, there are 9568 instances of five features namely Temperature(T),Exhaust Vacuum(V),Ambient Pressure(AP),Relative Humidity(RH), net hourly electrical energy output (EP) of the plant.in this the feature EP is taken as dependent variable and remaining four are independent variables.We need to predict the net hourly energy output based on four parameters.

Proposed Model:

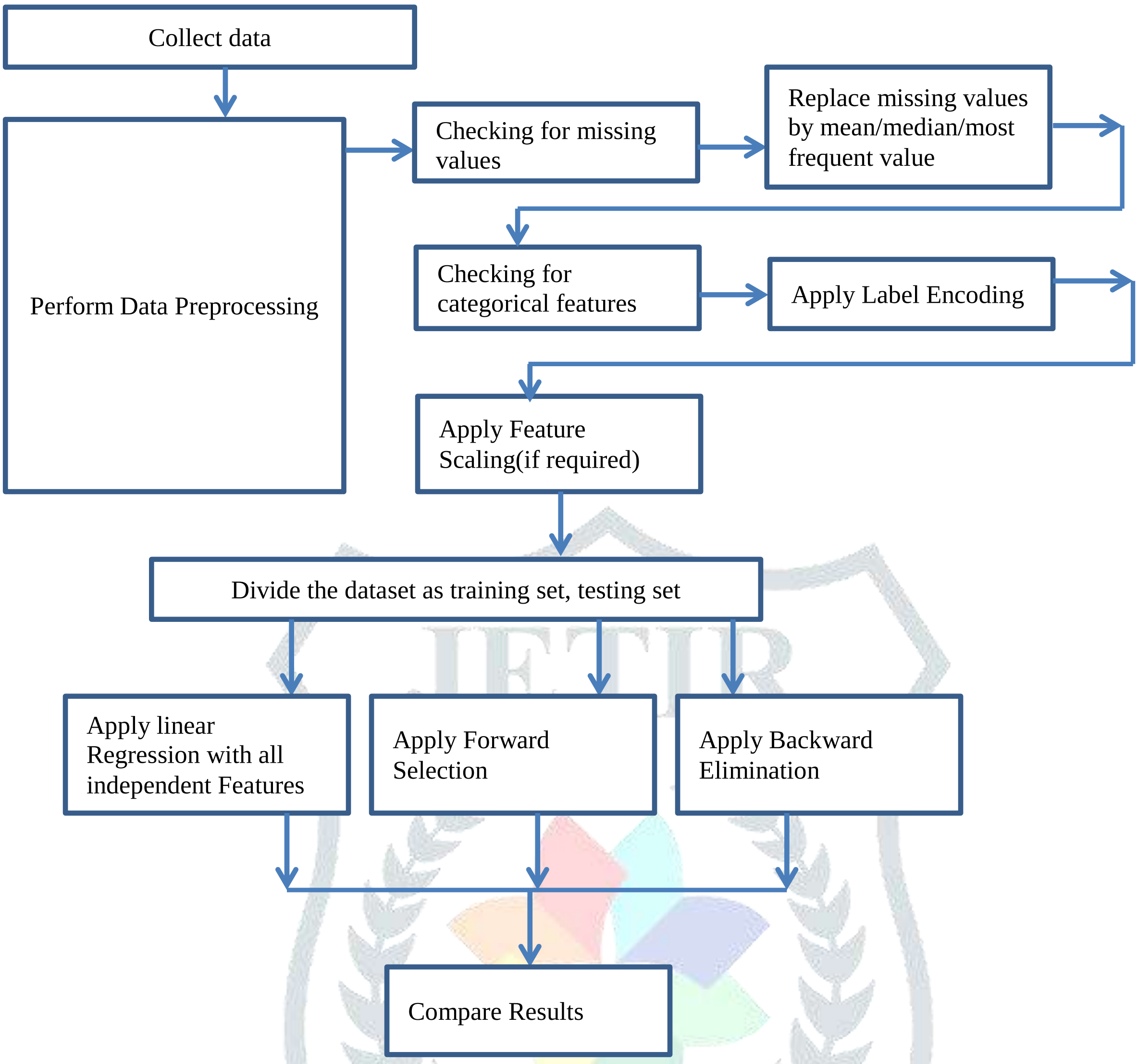
Collect data
Perform Data Preprocessing
Checking for missing values
Replace missing values by mean/median/most frequent value
Checking for categorical features
Apply Label Encoding
Apply Feature Scaling(if required)
Divide the dataset as training set, testing set
Apply linear Regression with all independent Features
Apply Forward Selection
Apply Backward Elimination
Compare Results

Figure 3.1: Proposed model

The first step of any machine learning algorithm is datapreprocessing.In this step, data is verified for missing values. If dataset contains missing values, those values are replaced by mean or median of remaining values of particular feature. In this dataset, there are no missing values. So, there is no need to apply any technique for removing missing values. Data preprocessing also checks categorical features in the dataset. If categorical features are present in the dataset, then we need to change those as numerical numbers by applying Label Encoding. This dataset contains no categorical features. Feature Scaling is one of the important step in machine learning algorithm where values of all features in the given dataset are scaled in the same range. Python linear regression class takes care of feature scaling. So, we are not applying Feature scaling.

In the process of applying machine learning regression, we divide the given dataset into training set and testing set. The model is well learned on training set and it is used for predcitions on test test.The sizes of training set and testing set is user choice. But, usually size of training set is larger than testing set. Preferable sizes for train set and test is 70%,30% or 80%,20% respectively.In our experiment,We consider the ratio of training set and test set as 80%,20%.So training set contains 7654 instances and test set contains 1914 instances.

Multiple Linear Regression can be implemented using different techniques like “All in variables method”, “Forward selection”, “Backward Elimination”.In the “All in variables” method, we consider all the independent variables in the process of predicting dependent variable. In the Forward Selection method, we start with empty list and adding significant features one by one. In Backward Elimination, We start with all independent variables and eliminating one element at a time until there are no more significant independent variables. In this model, we used p-value(or standard error) as a parameter to build the best model.
General significance level of p-value is less than or equal to 0.05.

Data Collection:
In this step, dataset is readed using read_csv.Seperate dependent variable EP (as X), independent variables AT, V, AP, RH (as Y)
Data Preprocessing:
As the dataset taken by us has no missing values and categorical features, there is no need of applying these techniques. Feature scaling is also done by python class automatically.

Dividing dataset:
Dataset is divided into training and test set(80%-20% ratio) using following sample code:

```
from sklearn.model_selection  import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

Model with all independent variables:
Add one's column to the left side of X as per multiple linear regression model (it can be treated as $X_0*\beta_0$ and X_0 is taken as 1).
Implementation:
Apply all Linear Regression model using Linear Regression class from sklearn package.OLS class(Ordinary Least Squares)from statsmodels.formula.api package is useful for finding standard error and other parameters.

```
    import statsmodels.formula.api as sm
    X_opt=X[:,[0,1,2,3,4]]  # original X only
    object=sm.OLS(endog=Y, exog=X_opt).fit()#ordinary least square model step 2 done
     object.summary()
```

Results are tabulated below:

Feature	Coef	Standard Error	R Squared	Adj. R-squared	F-statistic
AT	-1.9775	0.015	0.929	0.929	3.114e+04
V	-0.2339	0.007			
AP	0.0621	0.009			
RH	-0.1581	0.004			

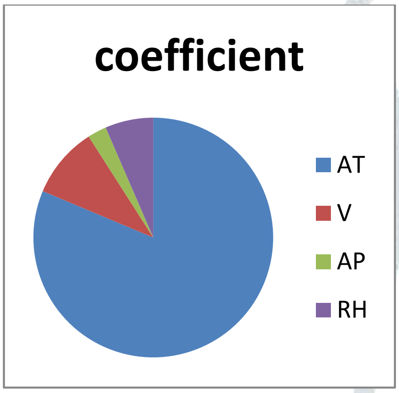

Figure 3.2: Coefficient values

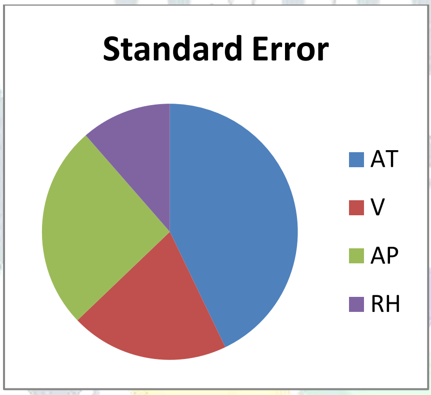

Figure 3.3: Standard error values

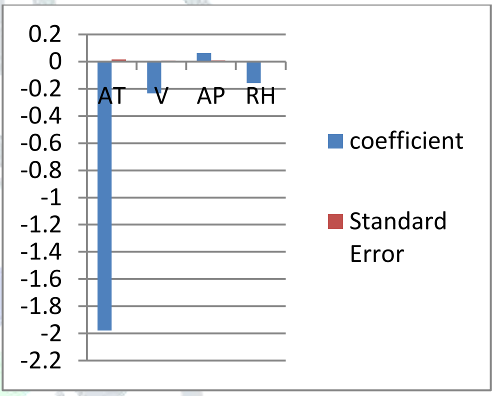

Figure 3.4: Coefficient vs standard error

Energy=454.6093-1.97*AT-0.233*V+0.062*AP-0.158*RH
Mean Energy when all predcitors are zero is 454.6
For every unit increase in temperature, energy decreases by 1.97 units
For every unit increase in Vacuum, energy decreases by 0.23 units
For every unit increase in Pressure, energy increases by 0.62 units
For every unit increase in Humidity, energy decreases by 0.15 units

Weget p value as 0 for all independent variables. As P-value is very less for all Independent Variables. So we can reject the NULL hypothesis. There is a significant linear relationship exists between the dependent variable and the independent variables. F value is much larger than 1 indicates that the variation in group means is not by chance and has statistical significance.

Forward Selection Algorithm:

i. Select a significant level to enter the model (start with empty list)
ii. Fit all simple regression models, select the one lowest p-value (or lowest error rate)
iii.Keep this variable and fit all models with one extra predictor added to the one(s) we already have
iv. Consider the predictor with lowest p-value (or lowest error rate)
v. Repeat steps 3 & 4 until all significant attributes identified.

Implementation:
Apply all Simple Linear Regression model using Linear Regression class from sklearn package.OLS class (Ordinary Least Squares) from statsmodels.formula.api package is useful for finding standard error and other parameters. Here Regression can be applied four times (AT&EP, V&EP, AP&EP, RH&EP).Standard error rates for each individual regression model is tabulated below. We get p-value as 0 for all predictions, so we consider only standard error values. Results are tabulated below:

Feature	Coef	Standard Error	R Squared	Adj. R-squared	Remarks
AT	-2.1713	0.07	0.899	0.899	89.9% of negative variance in energy can be explained by the linear relationship between temperature and energy.
V	-1.1681	0.07	0.757	0.756	75.6% of negative variance in energy can be explained by the linear relationship between vacuum and energy.
AP	1.48	0.025	0.269	0.269	26.9% of variance in energy can be explained by the linear relationship between pressure and energy.
RH	0.4557	0.011	0.152	0.152	15.2% of variance in energy can be explained by the linear relationship between humidity and energy

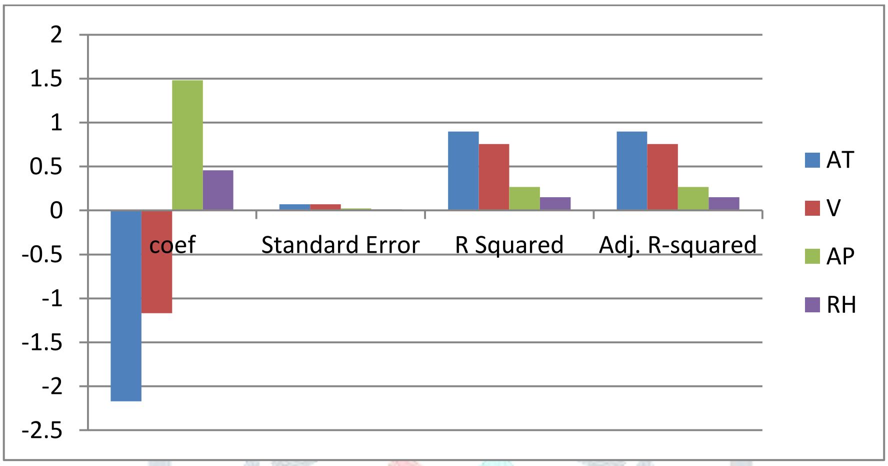

Figure 3.5:parameters values

The above table also suggests following things:

Independent Variable	EP(Dependent Variable)	Remarks
AT	-2.17	Strong negative correlation
V	-1.16	Strong negative correlation
AP	1.48	Strong positive correlation
RH	0.455	Weak positive correlation

Here, Prediction with RH gives less error rate. Keep this variable (step ii) and fit all possible models.
X_opt=X[:,[0,4]] //Keeping variable 4(RH)

Now Apply Linear Regression models with AT, V, RH features. Results are tabulated below:

Feature	Coef	Standard Error	R Squared
AT	-2.3907	0.08	0.921
V	-1.1132	0.07	0.772
AP	1.3292	0.023	0.384

Prediction with AP feature gives less error rate. So include AP and fit all possible models. Results are tabulated below.

Feature	Coef	Standard Error	R Squared
AT	-2.377	0.09	0.921
V	-1.004	0.07	0.804

Prediction with V gives less error rate.So include it& performs regression method.

Feature	Coef	Standard Error	R Squared
AT	-1.0014	0.09	0.929

Backward Elimination Algorithm:

i. Select a significant level to stay in the model (Start with all independent variables).
ii.Fit the full model with all the predictors.
iii.Consider the predictor with highest p-value(or high error rate).If its value>significance level ,goto step iv,else finish
iv.Remove the predictor, Fit the model and goto step iii.

```
X_opt=X[:,[0,1,2,3,4]]   # 0 means 1st column ,1:2nd column(AT) …..4:5th column(RH)
import statsmodels.formula.api as sm
object=sm.OLS(endog=Y,exog=X_opt).fit()#ordinary least square model step 2 done
regressor_OLS.summary()
```

Feature	coef	Standard Error	R Squared	Adj. R-squared
AT	-1.9775	0.015	0.929	0.929
V	-0.2339	0.007		
AP	0.0621	0.009		
RH	-0.1581	0.004		

Now, Remove AT, It has highest error rate.

S0,X_opt=X[:,[0,2,3,4]]

Feature	coef	Standard Error	R Squared	Adj. R-squared
V	-1.0014	0.007	0.804	0.804
AP	0.5645	0.014		
RH	0.1607	0.006		

Now, Remove AP, it has high error rate.

Feature	Coef	Standard Error	R Squared	Adj. R-squared
V	-1.1132	0.007	0.772	0.772
RH	0.1532	0.006		

Now, Remove V.

Feature	coef	Standard Error	R Squared	Adj. R-squared
V	0.4557	0.011	0.152	0.152

IV.CONCLUSION:

In this paper, we analyzed multiple linear regression models for net hourly power output of the combined cycle power plant. Multiple Linear Regression with Forward selection, Backward Elimination are investigated.The best model shows that Temperature(AT) and Vacuum(V) have strong negative correlation with Energy (EP).Pressure(AP) and Humidity(RH) are positively correlated with Energy(EP).But, pressure(AP) has more correlation with Energy(EP) than with humidity(RH) .

REFERENCES:

[1] 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.

[2] 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.

[3] V.Ramireddy, "An overview of combined cycle power plant",2015,http:// electricalengineeringportal.com/an-overview-of-combined-cycle-power-plant

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

[5] 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.

[6] 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.

[7] 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.