AC POWER CONSUMPTION ANALYSIS AND PREDICTION

Submitted By:

Ujjwal Pratik

ACKNOWLEDGMENT

For this study, I sincerely thanks to Zenatix Solutions Private Ltd. There are also some references taken from YouTube videos and google search.

INTRODUCTION

- Business Problem Framing

data contains power for multiple ACs at some hotel in Gurgaon.

Identify patterns/trends in the data?
Which AC was used the most/least?

Relate this power data with the outside temperature of Gurgaon.
Using the power data, predict/forecast the power consumption

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

I have used various Mathematical and Statistical approach for making this project predict better. I have used Statistical summary for checking and analysing various stats like, number of values counts in dataset of each column, mean, median and mode for each column. I have also used various visualization plot for analysing the patterns better. Other than that, stats show outliers and skewness in overall dataset. After analysing the data through Statistical Summary, I have applied some Mathematical tools for removing outliers and skewness present in the dataset, Mathematical tools is also used for checking null values and at last making prediction.

- Data Sources and their formats

The Source of the Data is Zenatix Solutions Pvt Ltd. The Datasets contains 87840 rows and 19 columns.

Dataset format using Pandas

	0	AC 1	AC 2	AC 3	AC 4	AC 5	AC 6	AC 7	AC 8	AC 9	AC 10	AC 11	AC 12	AC 13	AC 14
0	2019-08-01 00:00:00	7.518632	8.788315	0.000000	0.000000	2.617045	4.079041	2.782276	4.624447	5.222060	2.151238	1.585072	0.560373	3.142941	2.749470
1	2019-08-01 00:01:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2019-08-01 00:02:00	7.426114	8.940615	0.000000	0.000000	2.581625	3.781231	2.529366	5.057423	5.349465	2.414715	2.168184	1.818730	3.085110	2.720484
3	2019-08-01 00:03:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	2019-08-01 00:04:00	7.052986	9.161103	0.000000	0.000000	2.592095	3.800127	2.332304	6.322521	3.995392	2.237114	3.345624	2.310409	3.132799	2.676861
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
87835	2019-09-30 23:55:00	0.000000	6.122385	2.192198	2.083315	1.046250	0.000000	3.668421	3.006311	3.614301	1.860847	5.019769	3.154221	3.648026	2.439526
87836	2019-09-30 23:56:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
87837	2019-09-30 23:57:00	1.209176	6.152532	2.211421	0.000000	1.448103	0.000000	3.151248	2.871690	3.417942	2.185493	5.174168	2.772349	3.786657	2.060023

- Data Pre-processing Done

  - Importing Libraries
  - Loading the dataset
  - Checking the shape of the data
  - Checking data types
  - Checking Null values
  - Analysing through Statistical Summary
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis
  - Clustering
  - PCA
  - Splitting the independent and target variable
  - Checking and removing skewness and outliers
  - Building model

- Hardware and Software Requirements and Tools Used

  - Hardware Requirement:
    - Intel core i5
    - 8 GB Ram
  - Software Requirement:
    - Python 3.x with packages
    - Pandas: Data analysis and manipulation tool
    - NumPy: Provide support for mathematical functions, random number etc
    - Selenium: Web scrapping tool
    - Chrome driver: automated Google Chrome
    - Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
    - Seaborn: is a library mostly used for statistical plotting in python.
    - Scikit-Learn: is an open-source Python library that has powerful tools for data
    - analysis and data mining.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

    - R2 score : It is used to evaluate the performance of a linear regression mode.
    - Clustering: The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
    - PCA: Principal Component Analysis or PCA is a widely used technique for dimensionality reduction of the large data set
    - Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
    - Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
    - Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
    - Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
    - Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
    - Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
    - Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
    - Root Mean Squared error: is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
    - Label Encoder: Label Encoding refers to converting the labels into numeric form.
    - Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
    - Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting

- Testing of Identified Approaches (Algorithms)

    - Clustering
    - PCA
    - Train Test Split
    - Linear Regression
    - Lasso Regularization
    - Ridge Regularization
    - Elastic Net Regularization
    - Grid Search CV
    - Cross Validation
    - Hyper Parameter Tuning
    - Random Forest Regressor

- Interpretation of the Results

1- patterns/trends in the data?

So, from above analysis of dataset with the help distribution plot and scatter plot, groupby, clustering and PCA I have a conclusion till here is that dataset is almost normally distributed, not much skewness is present in the dataset and relationship of ACs with Hours and Minutes is also showing much positivity, So the data is now very clean and ready to move towards further process.

2- Which AC was used the most/least?

From calculations, analysis and visualization my conclusion is that AC13 is most used and AC1 is least used

3- So, after Hyper parameter tuning using Grid Search cv, our Random Forest Regressor model is predicting 99.99% r2 score and 99.98% cv score means the performance of the model is extremely good for prediction.

Random Forest Regressor

```
rf=RandomForestRegressor(criterion='mae',max_features='auto')
rf.fit(x_train,y_train)
rf.score(x_train,y_train)
pred_decision=rf.predict(x_test)

rfs=r2_score(y_test,pred_decision)
print('r2 score:', rfs*100)

rfscore=cross_val_score(rf,x,y,cv=5)
rfc=rfscore.mean()
print('cross val score:',rfc*100)
```

```
r2 score: 99.99871560594336
cross val score: 99.98224305489536
```

CONCLUSION

- Key Findings and Conclusions of the Study

During my analysis I got the better accuracy with Lasso Regularization that is 80%, and for better prediction I did hyper parameter tuning and I got minimum difference in accuracy and cross validation is Random Forest Regressor.

Lasso

```
ls=Lasso()

ls.fit(x_train,y_train)
pred_y=ls.predict(x_test)
print('r2 score',r2_score(y_test,pred_y))

print('error')
print('mean absolute error', mean_absolute_error(y_test,pred_y))
print('mean squared error', mean_squared_error(y_test,pred_y))
print('root mean squared error', np.sqrt(mean_squared_error(y_test,pred_y)))
```

```
r2 score 0.8388797168647587
error
mean absolute error 0.7829654272850934
mean squared error 0.9803199333090119
root mean squared error 0.9901110711980813
```

- ## Learning Outcomes of the Study in respect of Data Science

  This study gives me opportunity for lots of learning starting from various types of plotting like distribution plot, boxplot, scatterplot, line chart, grouby and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications

- ## Limitations of this work and Scope for Future Work

  For better performance, I plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.