Predicting Energy Usage In a Steel Factory

1. **Introduction and discovery:**

***Business Domain*** – This project aims to analyze the energy usage of a steel factory based in South Korea. This is collected from a smart small-scale steel industry in South Korea which focuses on producing several types of coils, steel plates, and iron plates. [1] The production of such materials requires heavy machinery and a huge amount of energy.

Typically, when machines require that much energy, they are 3 phase machinery and a phenomenon occurs in which not all energy produced is ‘active’ energy or useful energy, this is because the phases between the current and the voltage end up out of phase due to the strain put on the motors of the machine.

***Framing the problem*** *–* Not having a good power factor becomes expensive and so an

accurate estimation for the energy consumption is required to estimate the methods in which a correction to the power factor can be done. That is why the main purpose of this project is to estimate the energy usage of the steel factory in hopes that an accurate energy prediction can help dimension the size of the solutions that could further improve power factors and overall efficiency in the future of the steel factory.

***Developing Hypothesis*** *–* My initial hypothesis that power factors and reactive powers

will have great impact in the prediction of the energy usage. These determine how much ‘active’ power can be extracted from machinery use therefore these will play a big part and will have high correlations with energy usage.

Also, most probably, emissions produced using the machinery will have a high correlation as well since emissions will only be produced when machines are turned on and running. Of course, as a heavier load is put to the test, the more energy usage and the more emissions will be produced therefore, my hypothesis is that most of the relationships in this data will be linear and positive this will make the model very good for regression type models that can predict based on linear relationships.

1. **Dataset Analysis:**

***Data inventory*** *-* As per UCI Machine Learning Repository - "The information gathered is

from the DAEWOO Steel Co. Ltd in Gwangyang, South Korea. It produces several types of coils, steel plates, and iron plates. The information on electricity consumption is held in a cloud-based system. The information on energy consumption of the industry is stored on the website of the Korea Electric Power Corporation (pccs.kepco.go.kr), and the perspectives on daily, monthly, and annual data are calculated and shown." [2]

The data downloaded at first is in the format shown in figure 1, this format has very poor naming conventions for columns, but we can see that there is a good mix between “Dtypes” since we have float, int and object data types.

Text

Description automatically generated

Figure 1. Data info for dataset Steel\_industry\_data.csv.

***Data processing*** *-* Columns were immediately changed, and the summary statistics can

be seen in figure 2. With this a brief peek of the data can be provided and seen in figure 3.

Table

Description automatically generated

Figure 2. Summary statistics for dataset Steel\_industry\_data.csv.

*Table

Description automatically generated*

Figure 3. Peek at dataset Steel\_industry\_data.csv.

Additional changes were implemented which include:

* Checking for Null and NaN values
* Dropping column ‘nsm’
* Converting tCO2 to kgCO2
* Including hour as a feature and removing all other date aspects
* Revising unique values of object columns

***ELD*** – Plots were made to understand the relationship between the features in question

which include heatmaps (Figure 4) to understand the correlation of the features, scatterplots (Figure 5) to compare energy usage with other features and histograms and boxplots (Figure 6) to understand the distributions of the data through some important features.

Correlations were utilized to determine which features were the most pertinent to the prediction, in this case special attention was put into features relating to usage\_kwh.

Table

Description automatically generated with medium confidence

Figure 4. Correlation heatmap for features used.

In the following figure, notice the energy between difference when power factors drop. There is still some useful energy obtained when we see the lagging power factor compared to the energy in the leading power factor.

This is because most probably, the machinery in the steel production here is mostly machinery that produces a lagging power factor. Measures most have been made to compensate for the power factor, most commmonly used are capacitor banks. These capacitor banck regulate the power factor so that more 'active' power can be obtained, and the power factor penalizations are not met which in the energy industry tend to be quite steep in price.

Graphical user interface, application

Description automatically generated

Figure 5. Scatter plots for energy vs. Power Factors (PFs)

Distribution plot for energy - Very rarely are the machinery used in their maximum outputs, they operate under nominal values most of the time.

Boxplot for energy - Most data is clustered in the 0 to 50 kWh range with some very few values with high kWh.

Chart

Description automatically generated

Figure 6. Histogram and boxplots for Energy Usage.

1. **Model Planning:**

Based on the data and the ELD performed in the code, this model fits well under regression models and so the models that will be used are as follows:

* Linear Regression
* Decision Tree Regressor
* Random Forest Regressor
* AdaBoost Regressor
* XGBoost Regressor
* CatBoost Regressor
* LightGBM Regressor
* SGD Regressor
* Lasso Model

Based on the scores with these models I will choose a model upon further implementation will be made to improve the score. I will utilize Standard Scaler to scale the data that will be used in all this models and a pipeline will be implemented to apply all the models except for Lasso since Lasso will require further testing with its parameters

1. **Model Implementation**

A pipeline is implemented which will include all above-mentioned models, this pipeline intends to minimize the amount of code necessary to test various models. Various parameters were testes and the final parameters chosen to be implemented alongside the pipeline were:

* Scaler: Standard Scaler
* Selector: Random Forest Regressor (200 as number of estimators)
* Random state: 42
* Test size: 0.25
* Maximum iterations (if applicable): 2000
* Learning rates: 0.01
* Maximum depth: 5

The combination of these parameters inside de pipeline facilitates the analysis for the validation of the hypothesis which in turn makes the workflow more efficient. Further analysis was implemented upon the Linear Regression model as there was increased potential for improvement. Additional steps that were taken were:

* Perform Feature Selection
  + Correlation
  + Variance Threshold
  + Select K-Best
* Feature scaling with Standard Scaler
* Feature transformation with second degree polynomial

1. **Result Interpretation and Implications**

Pipeline results are shown in Figure 7 where all models perform very well with scores in the high 90’s with lowest score pertaining CatBoost and the highest scores tied with Decision Tree Regressor, Random Forest Regressor and XBG Regressor by a very minute margin.

**Text

Description automatically generated**

Figure 7. Pipeline results for dataframe with dummies.

Performance improvements in the linear regression model show better results in the model prediction as shown in Figure 8. After this, the best score obtained was with no feature selection and the second-degree polynomial transformation.

Table

Description automatically generated

Figure 8. Linear regression and Lasso model scores.

Model appears to be valid and provides an accurate prediction on the test data. The high score obtained is sufficiently accurate to meet the prediction goal. The output data makes sense since energy predictions are accurate and close to real data. Figure 9 shows scatter plot of predicted energy usage vs. actual energy usage.

Chart, scatter chart

Description automatically generated

Figure 9. Scatter plot for Actual vs. Predicted Energy Usage .

1. **Out of Sample Predictions**

For the out of sample predictions, I used the data frame statistics that belonged to the 50 and 75th percentiles and modified them to assimilate real-world data. Predictions turned out correctly since the predictions accurately predicted the energy usage in terms of rank, that is the one that had the medium load and less emissions and reactive power had less energy usage than the one who had a heavy load and higher emission and reactive power. No negative values or strange predictions were obtained while doing the synthetic dataset for testing.

1. **Concluding Remarks**

The final model used was linear regression with all features after the data cleaning and with a second-degree polynomial transformation. Surprisingly enough, the biggest coefficients don't come from kgCO2 which has the highest correlation with the energy usage instead they come from lagging reactive power [kvarh] which has much to do with an accurate prediction as most of the most positive coefficients belong to this power.

It’s interesting to know that the coefficients coincide with positive values while negative values point towards the power factor and leading reactive power mostly. Also, power factor, as hypothesized, has quite the impact on the prediction as seen with the coefficients.

The method that performed best was the one that used the most features out of all the datasets used and I believe that plays a high part into the prediction since it has more information to make an accurate. A clear positive linear relationship can be seen in the prediction plot and the values predicted are very similar or on par with the test values inputted into the model which can be attested by the model accuracy score

Finally, there were too many features used to make a very accurate prediction, one would really have to consider the value of getting 0.99 as accuracy rather than sacrifice accuracy over speed and computer memory usage. It would really depend the usage of this model, since the model can predict the energy usage accurately it can be used to dimension corrective measures for power factor correction, factory expansion measures, analyze performance of the machinery versus the expected outcome as a preventive maintenance strategy or to simply predict the energy usage accurately for energy efficiency purposes.

1. **References**

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Sathishkumar V E (2019), Department of Information and Communication Engineering,

Sunchon National University, Suncheon. Republic of Korea.