Data Analytics

Steel Industry Energy Consumption In a smart small-scale steel industry in South Korea

Vithushan Vadivel

September, 2022

Table of content

Introduction

Today in this world the electrical energy has become the essential energy source in this modern world. Everywhere and every place consumption of electrical energy is very high. Mainly in Industry sectors, factories and house hold. Even we can say without electricity our life will be very much difficult to find out another source instead of electricity. Having learnt some tools of data analysis, now I want to do some sort of analysis in electrical consumptions in small scale steel industry.

For this analysis I have chosen a data set of electrical energy consumptions in a smart small scale steel industry in South Korea for the year of 2018. With this database I want to be able to make the analysis I want to gather the insights I am looking for regarding correlation between the factors which influence in the energy consumption in this steel industry with the load type whether it is maximum or medium or light load.

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.

This data is mainly based on electrical energy consumptions in a small-scale steel industry in South Korea. I have decided to do the supervised machine learning on this dataset in order to find out the load types whether it will be light load or medium load or maximum load. With the help of this supervised learning, it will be possible to determine what kind of load type during which weekdays or weekend and during which time the load type will be maximum, medium and minimum.

I have retrieved the raw data from UCI repository and I received an overview of the data and then I perform exploratory data analysis and then cleaned it regarding of my observations. Then I really found a challenge in setting up database structures in MySQL and how to create entity-relationship diagram. Finally, I have imported cleaned data to MySQL and got some insights in database too.

Data and data sources

What is the Data?

This is the data regarding the electrical energy consumptions in a smart small scale steel industry in South Korea. The data is about 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.
This data set has data for the whole year 2018.

Data Attribute Information:

- Data Variables Type Measurement
- Industry Energy Consumption Continuous kWh
- Lagging Current reactive power Continuous kVarh
- Leading Current reactive power Continuous kVarh
- tCO2(CO2) Continuous ppm
- Lagging Current power factor Continuous %
- Leading Current Power factor Continuous %
- Number of Seconds from midnight Continuous S
- Week status Categorical (Weekend (0) or a Weekday(1))
- Day of week Categorical Sunday, Monday….Saturday
- Load Type Categorical Light Load, Medium Load, Maximum Load

Data Source

https://archive.ics.uci.edu/ml/datasets/Steel+Industry+Energy+Consumption+Dataset

By : Sathishkumar V E,
Department of Information and Communication Engineering,
Sunchon National University, Suncheon.
Republic of Korea.
Email: srisathishkumarve@gmail.com

Data collection

Data collection is the essential part of this project. I have collected the Data from the websites of UCI which is the Repository of datasets for machine learning and intelligent systems

I have selected this data to do the project in the supervised machine learning as it has 11 features to determine the Load Type of electrical energy consumptions.

Data cleaning and exploratory data analysis

Before starting the data cleaning I have imported all the necessary libraries like pandas, numpy, seaborn, matplotlib and sklearn for later visualization and model testing.
Then I have imported and visualized the data set through head function in python. Data set sample is looked like as follows.

```
#Visualizing the data set using head functions
energy.head()
```

	date	Usage_kWh	Lagging_Current_Reactive.Power_kVarh	Leading_Current_Reactive_Power_kVarh	CO2(tCO2)	Lagging_Current_Power_Factor	Leac
0	01/01/2018 00:15	3.17	2.95	0.0	0.0	73.21	
1	01/01/2018 00:30	4.00	4.46	0.0	0.0	66.77	
2	01/01/2018 00:45	3.24	3.28	0.0	0.0	70.28	
3	01/01/2018 01:00	3.31	3.56	0.0	0.0	68.09	
4	01/01/2018 01:15	3.82	4.50	0.0	0.0	64.72	

The shape of the dataset is 35040 rows and 11 columns.
I have checked the null values in the whole dataset columns and fortunately didn't find any missing values. And then checked the datatypes for every columns.

```
#Checking for the Missing Values / Find the Null values
energy.isna().sum()
```

```
date                                  0
Usage_kWh                             0
Lagging_Current_Reactive.Power_kVarh  0
Leading_Current_Reactive_Power_kVarh  0
CO2(tCO2)                             0
Lagging_Current_Power_Factor          0
Leading_Current_Power_Factor          0
NSM                                   0
WeekStatus                            0
Day_of_week                           0
Load_Type                             0
dtype: int64
```

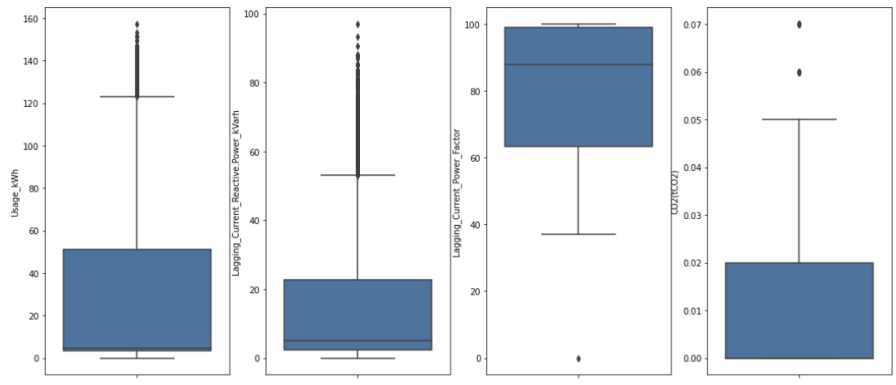
```
#Checking for the data types
energy.dtypes
```

```
date                                  object
Usage_kWh                            float64
Lagging_Current_Reactive.Power_kVarh float64
Leading_Current_Reactive_Power_kVarh float64
CO2(tCO2)                            float64
Lagging_Current_Power_Factor         float64
Leading_Current_Power_Factor         float64
NSM                                    int64
WeekStatus                            object
Day_of_week                           object
Load_Type                             object
dtype: object
```

In order to find out the outliers I have used box plots and found some outliers in lagging_current_power_factor and CO2(tCO2) columns and I decide to drop those outliers as these are very small part of the dataset by using the drop function in python.

```
#Creating Boxplots to analyse the outliers in the dataset or not
f, axes = plt.subplots(1, 4, figsize=(18, 8), sharex=True)
sns.boxplot(data=energy, y="Usage_kWh", ax=axes[0])
sns.boxplot(data=energy, y='Lagging_Current_Reactive.Power_kVarh', ax=axes[1])
sns.boxplot(data=energy, y='Lagging_Current_Power_Factor', ax=axes[2])
sns.boxplot(data=energy, y='CO2(tCO2)', ax=axes[3])
```

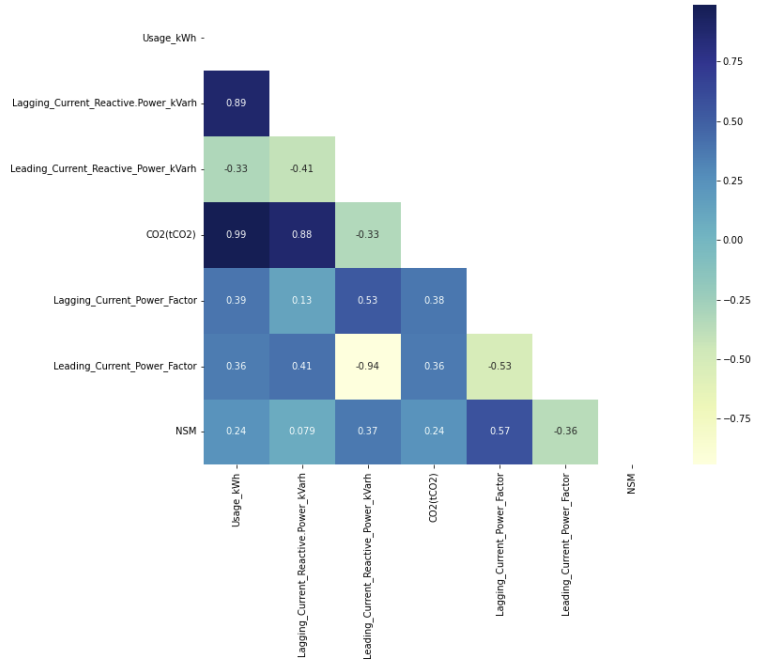
```
<AxesSubplot:ylabel='CO2(tCO2)'>
```



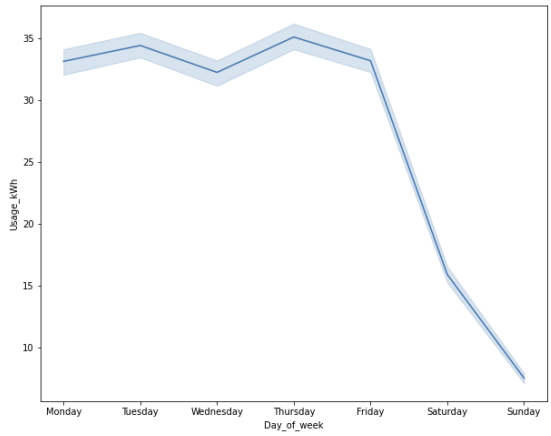
Then I decided to see the correlations in between each factors, so that I have used correlation function and heat map to see the correlations in between them.

```
#Finding Correlations in between the columns in Dataset
corr = energy.corr()

# Generate a mask for the upper triangle
mask = np.triu(np.ones_like(corr, dtype=bool))

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(11, 9))

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(energy.corr(), mask=mask, cmap="YlGnBu", annot=True)
```

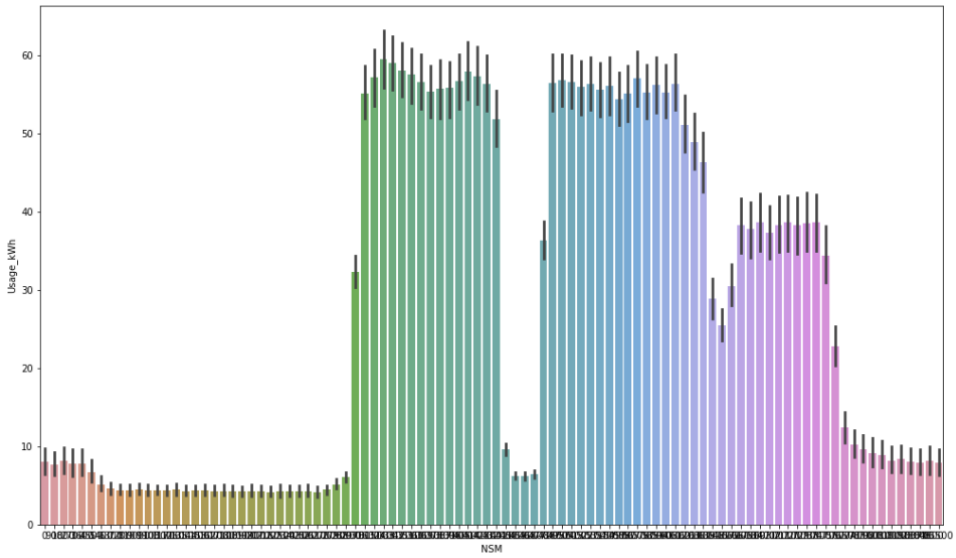
```
<AxesSubplot:>
```



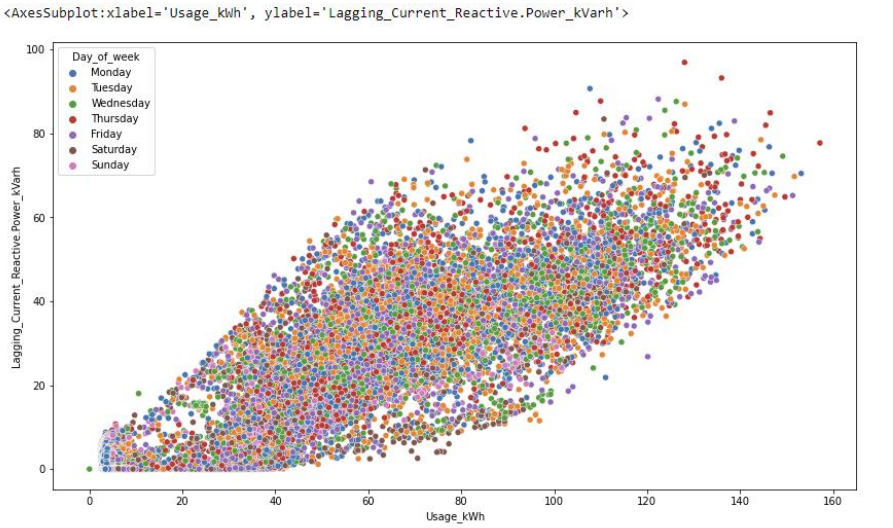
Correlation in the heatmap will give more insights that the CO2 is highly correlated to Usage_kWh and Lagging current reactive power. Lagging current reactive power and Usage _kWh is also highly correlated in between them.



This plot shows how the consumption of electricity is changes during in weekdays from Monday to Sunday. Though the machine is running continuously, as because there is less productions in weekend, the energy consumptions is very low in weekends.



This plot is showing that the usage of electricity with the time during the month of January 2018. It gives some insight that there is more consumptions in electricity in the middle of the month where the maximum load can be experienced. I think the steel productions in the factory will be much higher in the middle 10 days of the month causing high consumptions of electrical energy.



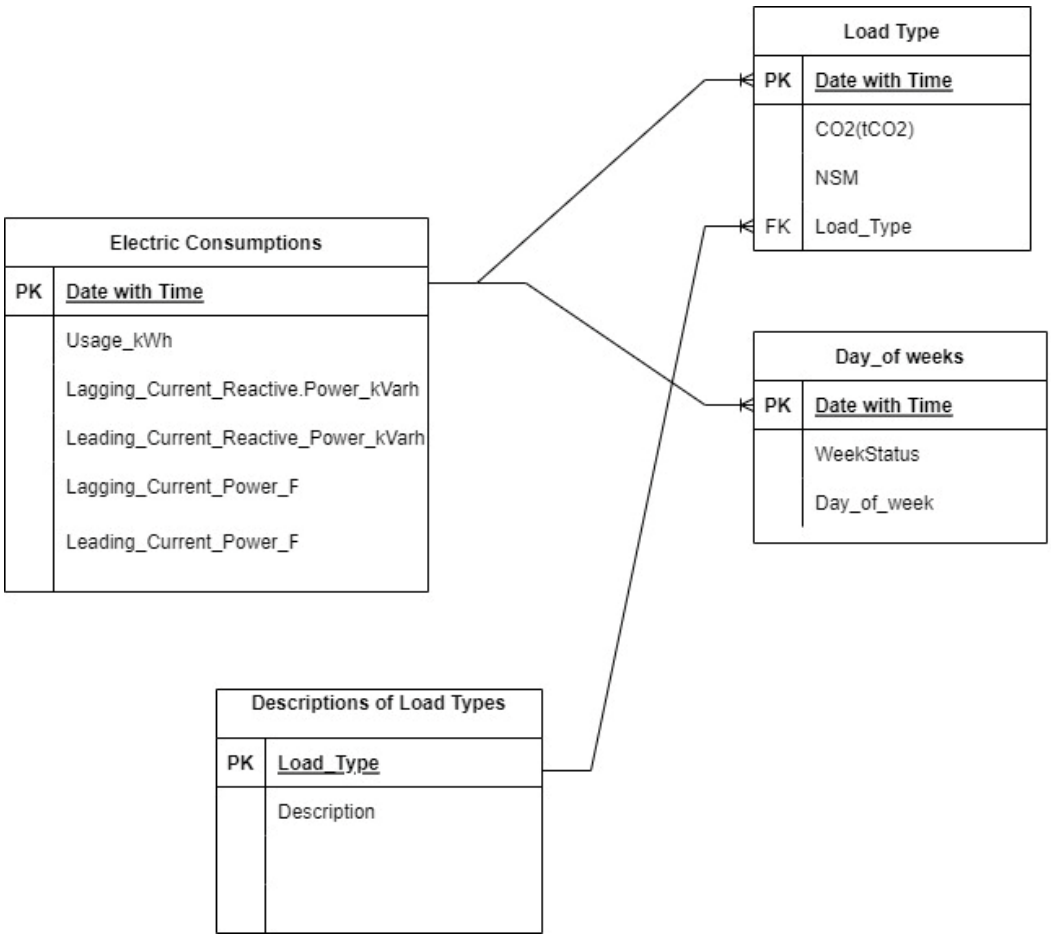
There is a highly correlations in between electrical energy usage and Lagging current reactive power. It's a positive correlations and value is approximately 0.89.

Deciding on what type of database to use

SQL	NoSQL
can work with smaller amounts of data	can work with big amount of data
supports transactions on cell level	does not support atomic transactions
slower querying, but faster updating	slower update, faster querying
Structured data	unstructured or semi structured data
OOP unfriendly (object oriented programming)	OOP friendly

Following my plan, the next step is creating my database. I decided to use SQL because this type of database allows me to use my cleaned and converted dataframes (structured data) and perform queries on a small level. Furthermore, I can link the different tables with each other using primary and foreign keys which allows me to quickly make queries and join different information from various tables. In my situation I also do not care about how friendly the system is towards object-oriented programming and writing longer queries is acceptable. The most important factor is to be able to work with the output of my data cleaning process: structured data. The relational database I will use is MySQL. In order to prepare the foundation for the creation of my database I created the following entity relationship model to clarify the different entity tables I will use and specify their relationships.

Entities Relationship Diagram. ERD

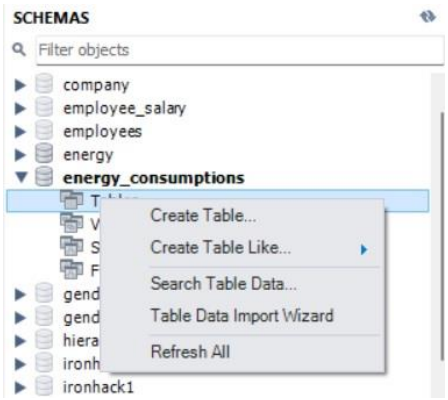


The following entity represents the results of my data cleaning process. Each entity has the unique primary key. For three table of entities the primary key is the same as date with time as its unique. All the values of data is measured in a unique time. The first entity gives the electrical energy consumptions and factors influencing the electrical energy consumptions. Second entity is giving data about load type in a unique time and CO2 emissions in that unique time. Third entity diagram gives some more information about day of week like Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday for that unique id date with time. Even it gives some information whether that date is a weekday or weekend. Fourth entity is giving descriptions of load types. For example if the load is light, then the energy consumption is low, if the load is medium then energy consumptions is medium level and if it maximum, then consumption of energy is very high. This relationship between all entities will later allow me to perform various queries on correlations between the data of the different entities.

Creation of the database and data importation

I began creating my relational Database in MySQL Workbench and created new database with the name of energy_consumptions.

I had a huge data and cannot able to insert it by manually. So that I have used Table Data Import Wizard to import .csv files of those entities.



Then I had joined these four entities as a one Table in MySQL by using the command line below

```
select * from `electrical consumptions`
join load_types on `electrical consumptions`.date = load_types.date_2
join `day_of weeks` on `day_of weeks`.date_3 =load_types.date_2
join `descriptions of load types`on `descriptions of load types`.Load_Type1=load_types.Load_Type
group by `electrical consumptions`.date;
```

Then I have created new table and dropped down the duplicates columns like date_2,date_3, Load_Type1

```
Create table joined_electrical_consmptions as (
select * from `electrical consumptions`
join load_types on `electrical consumptions`.date = load_types.date_2
join `day_of weeks` on `day_of weeks`.date_3 =load_types.date_2
join `descriptions of load types`on `descriptions of load types`.Load_Type1=load_types.Load_Type
group by `electrical consumptions`.date);

alter table joined_electrical_consmptions drop column date_2;
alter table joined_electrical_consmptions drop column date_3;
alter table joined_electrical_consmptions drop column Load_Type1;
```

I have finished in setting my database for the energy consumptions and now I am going to write queries to find out some useful insights

```
select Day_of_week,  avg(Usage_kWh), avg(`Lagging_Current_Reactive.Power_kVarh`), avg(Lagging_Current_Power_Factor)
from joined_electrical_consmptions
```

The Output of the above query is given a results below

Day_of_week	avg(Usage_kWh)	avg(`Lagging_Current_Reactive.Power_kVarh`)	avg(Lagging_Current_Power_Factor)
Monday	33.14393474842783	16.106470125786156	79.61819378930795
Tuesday	34.4276141826926	16.610396634615388	80.08050881410237
Wednesday	32.25423477564114	15.465582932692328	80.5575560897438
Thursday	35.11208333333361	17.35670673076924	79.56191706730785
Friday	33.195014022436176	16.103950320512865	79.84841947115399
Saturday	15.919020432691978	6.3098858173077	82.22658253205113
Sunday	7.545633012820356	3.2356330128205184	82.17167467948714

This results shows that average use of electricity in week days is approximately same and less amount of energy is used in weekends.

Conclusion

I have collected this data from UCI repository and I founds lots of challenges in doing this analysis. I didn't find big difficulties in data cleaning but as because this data is very huge, every time when I run the model testing for supervised machine learning, it took lots of time to display the results. It's a big challenge to join these four entities in one query and when I joined and can't able to create a new table as there was a duplicate column. So, I have renamed the primary key columns and joined then after dropped those duplicated columns which have been renamed. I did a forecast by using prophet is a new challenge for me because I have used this library for the first time and I have succeeded in it. Based on the data and the queries I did, can make a conclusion that there will be maximum consumptions of electricity in weekdays rather than in week end. If we analyse it in a day then, maximum consumption is in day time of the day as because this Industry is doing more productions in day time.

Links

Github repository:

https://github.com/Vithun93/Final-Project---Electrical-Energy-Consumptions