Segmenting the Stake Holders of Power plants

Data: fuel\_receipts\_costs\_eia923.csv is the data for this case study

Executive summary:

The analysis of Power Generation in the USA. It is one of the major industrial sectors in the USA. It supplies energy to different types of utilities. The power plants run on different types of fuels, which will cost the fuel different factors involved. Recent industrial revolutions from past decades and brought many changes in power generation methods such as renewable energy generation by solar, nuclear, and solar. And also fuel availability is decreasing day by day. This made a revenue loss for some of the thermal power plants. Then with the source of the PUDL database, we can find factors that are more affecting the fuel cost. I analyzed significant variables using K-means clustering to find out which cluster power plants are operating at a lower cost of fuel and are also eco-friendly. With my understanding of clusters, I suggested some of the changes that should make in buying quality fuel, implementing the advanced working principle of boilers to use maximum efficiency of fuel. So that Power plants will achieve maximum profits.

Power Plants Business Situation

Power plants in the United States are 11925 utility-scale major industries generating electricity. In this power sector, power plants run with different fuels such as coal, Natural gas, and Petroleum-coke. From the History of power generation, a major part of the fuel was coal used to run power plants to generate steam by burning coal, and by running the turbine. Later due to ecological factors operating, the maintenance cost of power plants is increasing. One of the stakeholders wants to reduce a power plant operating cost they need an understanding of the business to that which type of measures are needed to adopt and invest money on a long-term basis and short-term. And what are the financial and economic measures required to adopt advanced techniques to run power plants with less operating cost, to make profits, and to compete with renewable energy power generating Companies?

Then we came up with the data required to be available in PUDL. It is an open-source Public utility data liberation project that helps to understand information about Electric Power Generation, Transformation, and Distribution companies in the United States including financial statements, fuel received units, and contents of fuel that decides the quality of fuel which affect the cost of fuel and distribution information help to understand company performance, growth potential, and competitive pressure. For this case, I choose one specific table that contains the monthly fuel contract information, purchases, and costs reported in EIA-923 Schedule 2, Part A, for our analysis in the PUDL database.

The objective of our project is to illustrate the graphical and mathematical summaries that help to better understand where each company is compared to other companies that are performing well over the years. Now this time is competing with renewable energy-generating companies like wind, solar, and water This is a business problem because the stakeholder goes forward to success in business, and profits simultaneously go forward with it in short term as well as long term.

Approach:

For this, I had taken the data from PUDL’s EIA-923 Schedule-2 Part A data has 31 variables of categorical data and numerical data that represents factors that affect the cost of fuel to run the power plant with fuel such as coal, natural gas, and petroleum coke.

It has several missing, Nan values in this large data set. To start the analysis of data we should clean and identify the required variables that help for analysis.

In that, I have chosen to focus on 10 columns that are significant variables and dropped Nan values then taken 2% of the data. Then split the data into test and train data into ratio 75:25 and normalized the train data in jupyter notebook to convert values between 0 to 1.

Technique

After observing the data, I have chosen clustering analysis which k means the clustering method was applied to the data. The elbow method gives SSE values between points. In the graphical representation, it showed 2,3,4 possible chances of forming optimum clusters in that I have chosen 3 by observing the value of SSE. K=3 is the optimum value forming the cluster

Result

After applying the kmeans=3, I was able to create an accurate and easy analysis of the clusters of how power plants in terms of offensive performance best with low fuel cost are in cluster 0, the rest of them in the other two clusters. I also choose a bar chart as those are the easiest way to readable and approachable plots for stakeholders. I have tested many other clusters to determine the best number is 3 fit for data.

Conclusion

After forming clusters0, cluster1, and cluset2. Power plants in cluster 0 are performing well in with low operating costs in terms of fuel cost per MMBtu. It has fewer contract suppliers compared to other clusters which means if we buy a bulk amount of fuel might get fuel with less price.

Cluster0 has fewer mines when compared to other clusters it’s also one of the factors that help in reducing fuel costs.

Coming to the emission of pollutants cluster 0 emits less content of ash, sulphur, and chlorine into the atmosphere states the quality of fuel they are using. It helps to run power plants with low maintenance costs.

By observing the three clusters I would like to suggest my stakeholder start to adopt the measures and also implement the advanced working principles to use the maximum calorific value of the fuel. Which helps to run a power plant profitably.

Appendix

<https://catalyst.coop/pudl/>

<https://scikit-learn.org/stable/modules/clustering.html#clustering>