PUDL – Project Report

(MIS-64060-002)

Pranay Kumar Kodeboyina

811251177

pkodeboy@kent.edu

INTRODUCTION

The document provides a clear introduction to the aim and objective of the analysis, which is to use machine learning algorithms to analyze the PUDL dataset and provide insights on power generation in the US. The focus is on finding a type of fuel that is both cost-efficient and environmentally friendly. The use of K-means clustering algorithm to segment the data is also mentioned, which is a good approach to understand patterns in the data.

The U.S. government is planning to preserve fossil fuels. This project analyses factors associated with power generation in the U.S.A. to suggest a fuel type that can be excluded from the power generation process.

The government spends the least on oil, so it is assumed to be used the least and can be saved for the coming decades. Using a machine learning algorithm, data is grouped into 3 groups based on fuel type. Further analysis is done based on average fuel cost, number of units of fuel received at power plants, and their chemical composition.

The results show that gas is the fuel type on which the government is spending the most. Oil is the one on which the U.S. government is spending the least. Coal is not widely used despite having lesser per unit cost for MMBtu. This could be because of the presence of impurities like ash, mercury, and sulphur in it.

The data contains information about monthly fuel contracts, purchases, and costs. It has more than 20 variables, including the mine_ids from which fuels are to be supplied to different power plants by various suppliers, and the type of transportation used to supply these fuels.

Problem Statement:

Burning fossil fuels to generate power leads to the release of harmful gases and greenhouse gases that cause climate change. Fossil fuels such as coal, oil, and gas are the main contributors to greenhouse gas emissions, which cause health problems like respiratory issues, heart attacks, stroke, and early mortality. The use of fossil fuels is also linked to Alzheimer's disease and autism spectrum disorder. The main goal of this project is to analyse the data and provide solutions to reduce the negative impact of fossil fuels and help the US power generation industry to identify the best fossil fuel for power generation that is both cost-effective and environmentally friendly.

The US Power generation Unit has employed a data analyst to analyse historical data on monthly fuel contracts, purchases, and costs to determine which type of fossil fuel they spend the least amount of money on. They aim to reduce expenses on this type of fossil fuel and eliminate it from power generation. The objective is to conserve fossil fuels for future use that are not currently in high demand.

Technique:

To begin the data analysis process, the dataset was first screened for variables that contained either missing or redundant data. These variables were then excluded from the dataset to ensure that the data used for analysis was as accurate and relevant as possible. Additionally, to improve the interpretability of the data, a random sample of 2% of the total dataset was taken using the set.seed() function. Lastly, the fuel type code PUDL variable was converted into a numerical variable by creating three new dummy variables to represent the three types of fuel present in the dataset. This process ensured that the fuel type variable could be used in further analysis and modelling.

Initially, only a small percentage (2%) of the available data was randomly selected for analysis. This was done to improve the interpretation of the data, since handling a large dataset could be impractical. The data was then split into two sets - a training set, which comprised 75% (9128) of the data, and a validation set, which comprised 25% (3043) of the data. The K-Means clustering algorithm was applied to the training set to segment the data into groups, with the optimal number of clusters determined using the WSS and Silhouette methods.
Due to overlapping of the K-means cluster, DBSCAN method is used to implement the clustering on the train data and the value of K is obtained, the optimal value of k was found to be 3, as the silhouette score was highest at this value. The DBSCAN algorithm seeks to minimize the total within-cluster variation by defining k clusters. Accordingly, the data was divided into 3 clusters based on their similarities. This method is particularly useful for clustering large datasets in a quick and efficient manner.

Analysis and Discussion:

Cluster 1 - Gas

The analysis provides insights into the characteristics of Cluster 1, which represents the Gas fuel type. The analysis suggests that Gas has the lowest average fuel cost per MMBtu compared to other fuel types, which could make it a cost-effective option for energy generation. Gas has the lowest average fuel MMBtu per unit, indicating that it generates less heat content per unit compared to other fuel types. Gas does not contain any ash, sulfur, and mercury content, which could make it a clean and environmentally friendly fuel option. Natural gas is the energy source code for Gas, and pipelines are the most commonly used transportation type to supply this fuel type. This information could be useful in assessing the infrastructure requirements and associated costs for using Gas as a fuel type. The graph indicates that Gas is mainly purchased on spot rather than through contracts, which could be an important factor to consider for procurement strategies.

Cluster 2 - Oil

Oil is the most expensive type of fuel and is purchased only on spot. This suggests that the buyers are not willing to enter into long-term contracts for purchasing oil, possibly due to the high cost and uncertainty of future prices. Oil has a relatively low average units received in comparison to gas and coal could be due to its high cost. Customers may prefer to use other fuel types that are cheaper and more cost-effective. Sulfur content in oil is important as it indicates that oil may have negative environmental impacts such as air pollution

Cluster 3 - Coal

Coal has the lowest average cost per unit of fuel, it also has the highest average fuel MMBtu per unit, indicating that it produces more heat energy per unit than gas or oil. This could explain why it is widely used in the US despite its environmental impact. Coal is purchased both on spot and on contract, but there are more spot purchases than contract purchases. Coal include BIT and SUB, which stand for Bituminous Coal and Sub bituminous Coal, respectively. These are two types of coal that differ in their energy content and chemical properties, and they are both widely used in the US for electricity generation and other industrial purposes. Presence of ash, sulfur, and mercury in coal can have significant environmental and health impacts, including air pollution, acid rain, and water pollution. It is important to note that the use of coal is declining in the US and many other countries due to these impacts, as well as the increasing availability and affordability of renewable energy sources.

Conclusion:

The US Power generation unit has made a decision to exclude a type of fossil fuel from power generation in order to preserve it for future generations. The chosen fuel to be excluded is Coal, and this decision was made based on the fact that the amount of money spent on Coal is significantly higher compared to Gas and Oil. However, I would like to suggest a different approach. Instead of focusing on excluding the fuel that the company spends the least on, they should consider redirecting their funds towards a more efficient and environmentally friendly fuel.

It is important to note that Coal is not the best type of fuel for the environment, as it emits high levels of pollutants and greenhouse gases. In contrast, Oil is a cleaner and more efficient type of fuel that can be used for multiple purposes. In addition, Oil exists in its purest form and does not require extensive processing like Coal, which makes it a more cost-effective option in the long run.

Therefore, it would be wise for the government to redirect their spending towards Oil, rather than continuing to invest heavily in Coal. By doing so, they can not only reduce their expenses, but also promote the use of a more sustainable and efficient source of energy.

Reference:

https://www.iea.org/fuels-and-technologies/coal
https://stats.stackexchange.com/questions/23472/how-to-decide-on-the-correct-number-of-clusters
https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb