FML PROJECT REPORT

ZACHARIAH ALEX
zalex@kent.edu

EXECUTIVE SUMMARY

As a business analyst working for Galaxy Green organization, the primary aim is to understand how power is generated in US using fossil fuels. Raw data is downloaded from PUDL open-source website and it contained about 608565 rows and 30 columns. The data was cleansed by removing variables with more than 50% null values and by further analysis using Hierarchical clustering algorithm the main fossil fuels used for power generations was found out to be gas, coal and oil in cluster 1, cluster2 and cluster 3 respectively. Each of the characteristics of the fuels in each cluster is obtained. From the analysis it is clear that Coal is the fuel which causes highest impact on environment as it has constituents like ash, mercury and sulphur. Since natural gas is the purest fossil fuel, in spite of their high price, fuel received from suppliers is comparatively very high than that of other fuels. This gives us the general idea that the most preferred fuel in US for power generation is Gas. These fuels are mostly transported through pipelines and purchased in contract or on spot purchases. The United States natural gas output has substantially grown because to technological advancements, and has encouraged numerous industrial and utilities to migrate from coal to natural gas due to the global warming concern. If cost is a concern, there are other alternatives like oil/petroleum which causes significantly less impact to environment on emission of greenhouse gases compared to coal.

INTRODUCTION

The data is obtained from PUDL which is an open-source data processing pipeline that makes US energy data easier to access and use programmatically. The data consist of 608,565 rows and 30 columns with several missing values. We are assuming here that columns with more than 50% of missing values will significantly affect the performance of our clustering algorithm. Therefore, we will be removing those columns to enhance our algorithm performance on clustering data. We are considering only one categorical variable column that is fuel type code pudl. Since fuel type code pudl column is a categorical variable, it is converted to factor levels to perform clustering. The filtered data is sampled 2% with set seed value 8578 and split into test and train in 75:25 ratio.

PROBLEM STATEMENT

Galaxy Green, a non-profitable organization focused on advancing a culture of saving earth by reducing greenhouse gas emissions. They do this by reframing the way the world uses fossil fuels for power generation. As a member of this organization, it is my responsibility to find out ways to minimize the cause for this so as to contribute towards developing a sustainable future for future generations. We will be analysing the data on how power generation is done in US using different types of fossil fuels to understand the general idea about what type of fuel is being used the most and try to suggest a possible solution to tackle the effect of greenhouse gas emissions.

Questions:

- What are main types of fossil fuels used for electricity generation in US?
- Which among the fuels is the possible cause for increasing greenhouse gas emission?
- What are the possible solutions for tackling this greenhouse gas emissions?

ANALYSIS AND DISCUSSION

Hierarchical clustering algorithm is used to perform clustering in the data. This is because of mainly 2 reasons:

- ✓ Hierarchical clustering algorithm does not require pre-specified number of clusters.
- ✓ Can specify what distance metric to use.

Here we are choosing k =3 because we are classifying the fuel types into mainly 3 categories. Wards distance is used as the distance metric because Ward's minimum variance criterion minimizes the total within-cluster variance. The clusters formed are named as follows:

- ❖ CLUSTER 1: GAS
- ❖ CLUSTER 2: COAL
- ❖ CLUSTER 3: OIL

FINDINGS

- ▶ The three major categories of energy for electricity generation are fossil fuels like coal, natural gas, and petroleum/oil.
- ▶ Coal has constituents like ash, mercury and sulphur whereas gas is the purest form of fuel. We can say that coal is the fuel which causes the maximum greenhouse gas emission compared to the other two fuels.
- ▶ We should be promoting usage of cluster 1 fuels that is the Gas. This form of fuels has zero percentage of chemical constituents and emission of greenhouse gases are minimal. Although high cost is a factor, to have a sustainable future, we should bring down global warming effect by minimizing the usage of other fossil fuels.
- ▶ Coal is the cheapest fuel compared to gas and oil and has high heat content per unit.
- ▶ Gas is the most used fuel for electricity generation in US.
- ▶ Gas is mainly transported through pipelines whereas coal can be transported via Rail-roads and Trucks.
- ▶ For Gas, fuel heat content is less therefore this is purchased in large quantity compared to other fuels.
- ▶ Generally, fuels are bought in contract and spot purchases in US. Coal is purchased on contract basis and gas is spot purchased whereas oil is bought in both contract and spot purchase.

CONCLUSION

There are mainly 3 types of fossil fuels namely gas, coal and oil. Greenhouse gases are released into the atmosphere in massive quantities when fossil fuels are burned. Global warming is caused by greenhouse gases, which trap heat in our atmosphere. The most environmentally friendly fossil fuel to utilize is natural gas whereas coal has one of the worst effects. More than any other fossil fuel, coal has the greatest overall environmental impact. The effects of coal mining on the environment cannot be exaggerated, particularly with regard to open-pit mining. Coal mining has a significant negative influence on the environment, including the eradication of forests, the degradation of water quality, and the irreversible alteration of the terrain. There are other alternatives like oil/petroleum which is a better substituent for coal. There should be a shift from usage of fossil fuels like coal to gas or oil for power generation to have a sustainable future for future generations.