Final Exam

Kristian Lavinder

Kent State University

BA-64060-002: FUNDAMENTALS OF MACHINE LEARNING

Dr. Li Liu, Ph.D.

December 17th, 2023

Executive summary: The data associated with fuel receipts can be mined for many connections. With data specific to dates, there can be significant trends identified, but eliminating the dates and combining all the data can reveal connections that support current initiatives. With the large amount of data associated with contaminates (i.e. sulfur, ash, mercury) the data can be useful in a cost-benefit analysis. Natural gas is cleaner and its prices are competitive with coal, however coal makes considerably more heat. Petroleum has some advantages over coal but in still not as clean as natural gas. The cost of petroleum and lack of significant advantage make its use not justifiable.

Introduction: There was a significant amount of data to use. There were 663,572 observations with 30 different variables. Some of the fields were categorical and incomplete, meaning there were blanks in some of the variables for some of the observations. The field specific to fuel type was converted to numerical data, then the data was partitioned down to 3136 observations. Many of the fields were redundant or had blank entries and were reduced from the original data set. The final fields that remained were Fuel Type, Quantity, Thermal Content, Sulfur Content, Ash Content, Mercury Content, and finally Cost per Thermal Unit. The data was normalized using the ward method, verified by comparing the various methods of "average", " single", "complete" and "ward". Seven clusters divided the partitioned data adequately. The clusters had unique differences despite one very large and one very small cluster.

Problem statement: What is the association between fuel types, cost, thermal production, and contamination in the given data set?

Analysis and Discussion: The seven clusters produced some clear differentiations. Coal carries the most contaminants; however it is the least expensive.

Coal can be differentiated by its mix of contaminants. The clustering shows that mercury has minimal effect on thermal content, the same cost and thermal production can occur without mercury as a side effect.

Petroleum produces higher thermal content than natural gas and is cleaner than coal, but is significantly more expensive than Natural Gas. This may explain the lower volume of use.

Natural gas has little to no sulfur, ash, and mercury but there are three clear price categories: $3.62, $7.49, and $31,052. The large outlier is a miniscule portion of the data, but the discovery is curious. The thermal production isn't significantly greater, but perhaps there is something particularly unique (e.g. ultra clean?) or the location of final use is remote and difficult to access.

Natural gas predominates the data, which favors environmental initiatives. A further analysis could account for the dates of the data to determine if there are trends associated with clusters.

- **Cluster 1:**
  - Predominate Fuel Type: Coal
  - Quantity (1.8 % of data)
  - Thermal Content (Very High)
  - Sulfur Content: High
  - Ash Content: High
  - Mercury Content: High
  - Fuel Cost Per thermal units $2.31
- **Cluster 2:**
  - Predominate Fuel Type: Natural Gas
  - Quantity (4.59 % of data)
  - Thermal Content (Low)
  - Sulfur Content: None/Scant
  - Ash Content: None/Scant
  - Mercury Content: None/Scant
  - Fuel Cost Per thermal units $7.49
- **Cluster 3:**
  - Predominate Fuel Type: Petroleum
  - Quantity (0.2 % of data)
  - Thermal Content (Moderate)
  - Sulfur Content: Low
  - Ash Content: None/Scant
  - Mercury Content: None/Scant
  - Fuel Cost Per thermal units $16.78
- **Cluster 4:**
  - Predominate Fuel Type: Natural Gas
  - Quantity (90.33 % of data)
  - Thermal Content (Low)
  - Sulfur Content: None/Scant
  - Ash Content: None/Scant
  - Mercury Content: None/Scant
  - Fuel Cost Per thermal units $3.62
- **Cluster 5:**
  - Predominate Fuel Type: Coal
  - Quantity (1.77 % of data)
  - Thermal Content (Very High)
  - Sulfur Content: Moderate
  - Ash Content: Higher
  - Mercury Content: Low
  - Fuel Cost Per thermal units $2.45
- **Cluster 6:**
  - Predominate Fuel Type: Coal (mixed)
  - Quantity (1.27 % of data)
  - Thermal Content (Very High)
  - Sulfur Content: High
  - Ash Content: Highest
  - Mercury Content: None/Scant
  - Fuel Cost Per thermal units $2.37
- **Cluster 7:**
  - Predominate Fuel Type: Natural Gas
  - Quantity (0.013 % of data)
  - Thermal Content (Low, higher than other Natural Gas)
  - Sulfur Content: None/Scant
  - Ash Content: None/Scant
  - Mercury Content: None/Scant
  - Fuel Cost Per thermal units $31,05

Conclusions: The data clustering shows that cleaner fuel that is lower in sulfur, ash, and mercury is unfortunately more expensive and produces lower thermal units per weight. Coal can be differentiated into cleaner and dirtier coal based on the content of sulfur and ash. Coal with mercury shows no benefit in cost and use should be eliminated as alternative options are available in considerable volumes.

PRESENTATION VIDEO:

https://video.kent.edu/media/klavinde_final_exam_Video/1_svfqaulr