Assignment Instructions: MIS 64060 Final Exam

Purpose

The objectives of this assignment are threefold:

1. To use real-world data
2. To use the appropriate machine learning techniques for the business problem, and
3. To present the solution to top-level management.

Learning Outcomes

This assignment covers all learning outcomes in this class.

What is PUDL?

The PUDL Project is an open-source data processing pipeline that makes US energy data easier to access and use programmatically.

Hundreds of gigabytes of valuable data are published by US government agencies, but it's often difficult to work with. PUDL takes the original spreadsheets, CSV files, and databases and turns them into a unified resource. This allows users to spend more time on novel analysis and less time on data preparation.

For this project, we will use one specific table, the monthly fuel contract information, purchases, and costs reported in EIA-923 Schedule 2, Part A, for our analysis. This table contains 608,565 rows and 20 variables, though note that several variables have significant missing values. You can interact with the table using an open-source data explorer called Datasette, though for your final analysis, you will use R or Python.

Task

Your objective is to use any of the methods we have learned in this class as a primary means of understanding and analyzing the data. For example,

- What information is revealed by clustering? Use any and all variables to describe and identify the clusters.
- How should the value for the number of clusters be chosen?
- Select the best segmentation and describe the clusters. How would your segmentation help understand power generation in the US?

The above are only sample questions. This is an open-ended task, as such, you may answer questions that you formulate based on your insight and analysis.

Data

While the data itself is clean, the dataset contains several variables that have significant missing values. Follows these steps:

1. Remove all variables that have significant missing values.

2. Ensure that the variables have the right attributes. For example, numerical or categorical.
3. To ensure that both the data, and the analysis are unique to each student, randomly sample about 2% of your data using a random 4-digit number as the seed to sample the data. Use 75% of the sampled data as the training set, and the rest as the test set (if needed). This should yield a training set of about 9000 and a test of about 3000.

What to Turn In?

Your final submission will consist of both a report, and a short presentation in class. The report should contain the following sections:

- Executive summary - Write a summary of your findings, and recommendations about power generation in the US. This summary should not be longer than a paragraph or two.
- Introduction - A brief description of the data and any data cleaning you might have done. For example, removing variables, recoding, etc.
- Problem statement - What questions will you be answering
- Analysis and Discussion - Present your findings. Make sure to specifically address the questions you raised in the previous section
- Conclusions - State any assumptions, and any final thoughts you have

Final Grade

There is no one answer to this project. Your final grade depends on the following:

- The insight you show in developing your questions
- The analysis. Did you apply the techniques correctly? Was the analysis in-depth?
- Quality of your report - Both written, and correctness
- Quality of your final presentation

Extra Credit

- Use multiple-linear regression to determine the best set of variables to predict fuel_cost_per_mmbtu.
    - Check its prediction on the test set
- Append your data with the cluster information.
    - Predict the cluster for the test sample
    - Rerun the regression model with the chosen variables + cluster information. Check the prediction on the test set
- Did adding cluster information improve your prediction?