Course Project 2: Customer Classification

By: David Baker, Paulina Nelson, Morgan Shepherd,
Addison Thalmer, Isaac Torres M, Zephyr Zink

A. Data Processing

We selected and extracted the following attributes from our time series data set, as we believe these would be parameters of concern for the utility when deciding how to meet each customer's demand. We verified any overlaps in attributes by comparing the correlations among variables, excluding those with high correlations to other attributes:

Attribute	Justification
Peak Demand (Max) [W]	Represents the most power this customer would be demanding in a given day. Indicative of peak demand.
Minimum Demand [W]	Represents the least amount of power this customer would demand in a given day. Indicative of lowest demand.
Average Demand [W]	Over the span of a day, representative of the typical amount of power this customer would be demanding at any given time of day.
Difference (Max-Min) [W]	The largest increase / decrease in demand that this customer would require in a day.
Demand Standard Deviation [W]	How much this customer's demand would vary from their average, or if their power usage is highly varying or consistent.
Max Ramp Rate [dW/dt]	The most a generator would have to speed up to meet this customer's increase in demand.
Min Ramp Rate [dW/dt]	The most a generator would have to slow down to meet this customer's decrease in demand.

- Multidimensional dataset in xslx or csv format: see attached excel file

B. Customer Clustering

For our analysis, we decided to use a K-Means clustering algorithm. Our goal is to determine the similarity between customers, therefore grouping them by the nearest representative among the aforementioned attributes makes the most sense – with advantages being this technique's ability to handle large data sets and the interpretability of the resulting clusters. K-means clustering works best with spherical groupings of data as opposed to nonlinearity, which works for our approach since the attributes of power demand mostly appear uniformly dispersed. Results of the clustering analysis are presented in *Figure* 1. To accommodate bias, we tested against a range of clusters and assessed the associated silhouette score for each. We determined that the optimal number of clusters for our analysis was 3, with a silhouette score of 0.54. The

silhouette score represents how close a given data instance is to its own cluster vs the other clusters, and ranges from –1 to 1.

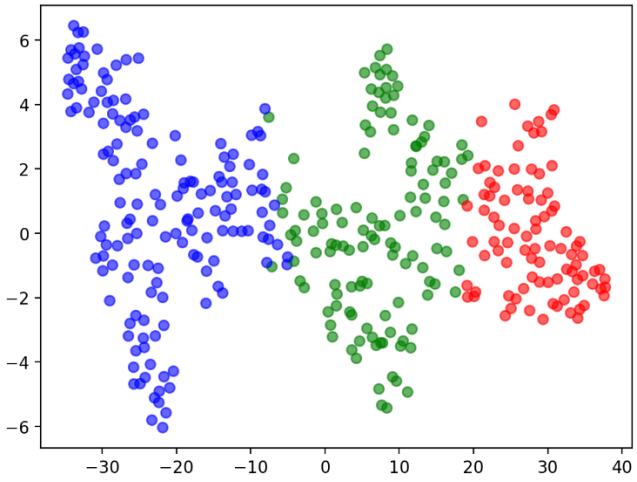

Figure 1: Results of k-means clustering analysis with k=3 visualized using t-distributed stochastic neighbor embedding.

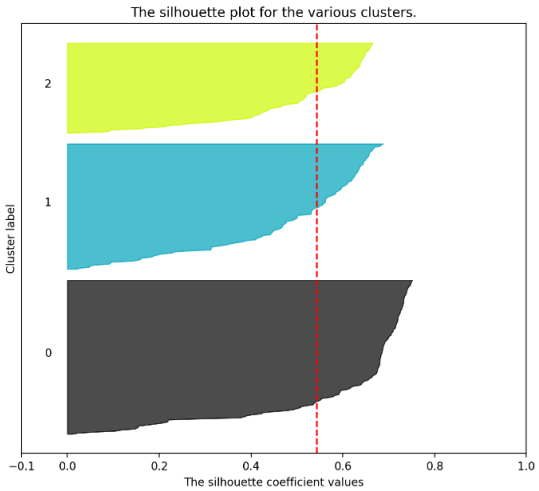

Figure 2: Silhouette plot of the clusters determined in figure 1.

C. Customer Classification

To classify customers, we analyzed three models: a logistic regression classifier, k-neighbors classifier, and linear discriminant analysis (LDA). K-neighbors classifier and LDA had the highest accuracy. For our analysis, we chose k-neighbors because it requires no knowledge of underlying patterns or distribution of the data which is not the case for LDA. In addition to needing to know underlying information, LDA requires a lot of assumptions to be met which are not necessarily applicable to real-world data hence our choice to use k-neighbors classifier. To validate k neighbors, we developed a confusion matrix as shown in Figure 4. Class one had the highest classification truth with 100% true classification and 0% false classification. However, this class only had

one data point in the true and expected datasets. Class two had 89% true classification, and three had 95% true classification and 5% false classification. Although the classifier misclassified a data point belonging to class 3 as belonging to class 2, this classifier seems to be highly effective with high accuracy and low error.

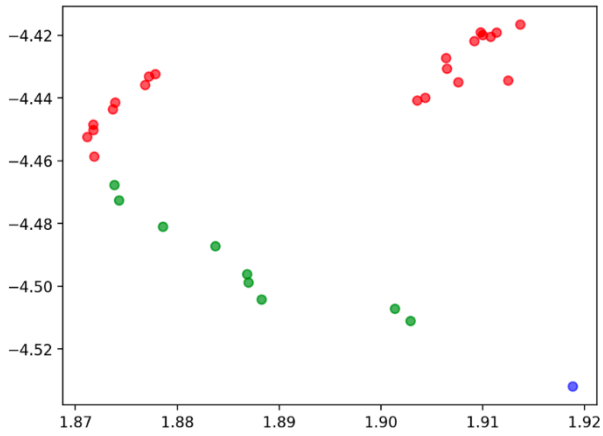

Figure 3: Classification results from the model proposed in Part C visualized using t-distributed stochastic neighbor embedding.

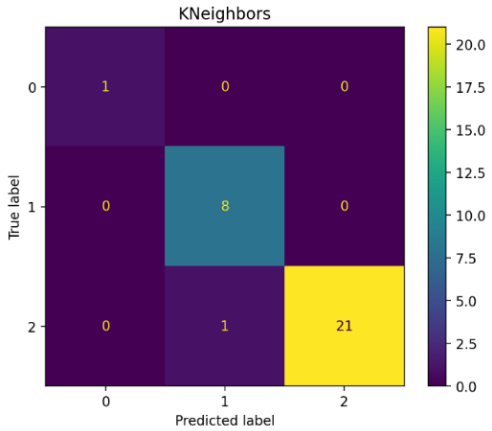

Figure 4: Confusion matrix for the k-nearest neighbor classifier developed in Part C.

Appendix

```
Logistic Regression Classifier
        prediction [2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 2 1 1 2 2 2 2 1 1 1 1 1 0]
        actual     [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 1 2 1 1 0]
        score on train 0.9700598802395209
        score on test 0.9354838709677419
        0.9354838709677419

K Neighbors Classifer
        prediction [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 1 1 1 1 0]
        actual     [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 1 2 1 1 0]
        score on train 0.9910179640718563
        score on test 0.967741935483871

Linear Discriminant Analysis
        prediction [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 1 1 1 1 0]
        actual     [2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 1 2 1 1 0]
        score on train 0.9640718562874252
        score on test 0.967741935483871
```

Figure 5: Training and testing scores for the logistic regression classifier, k-nearest neighbors classifier, and LDA classifier.

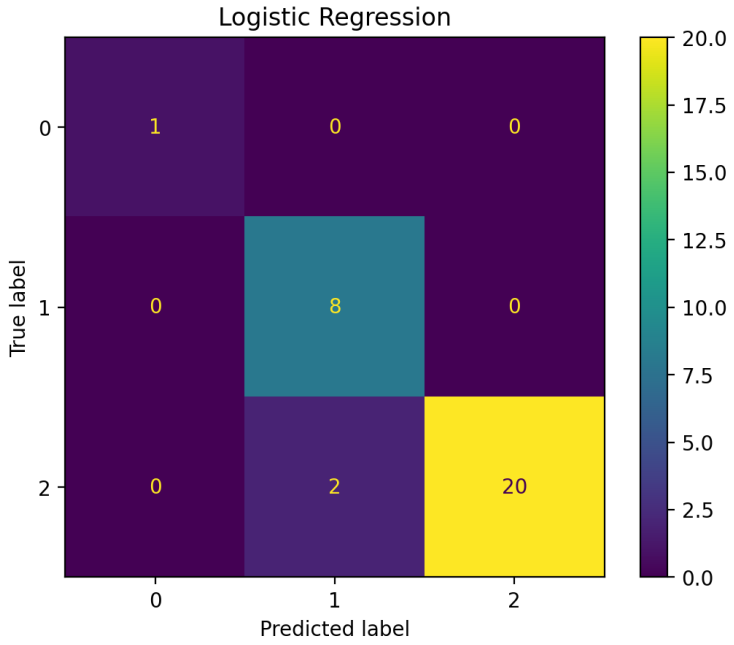

Figure 6: Confusion matrix for logistic regression classifier.

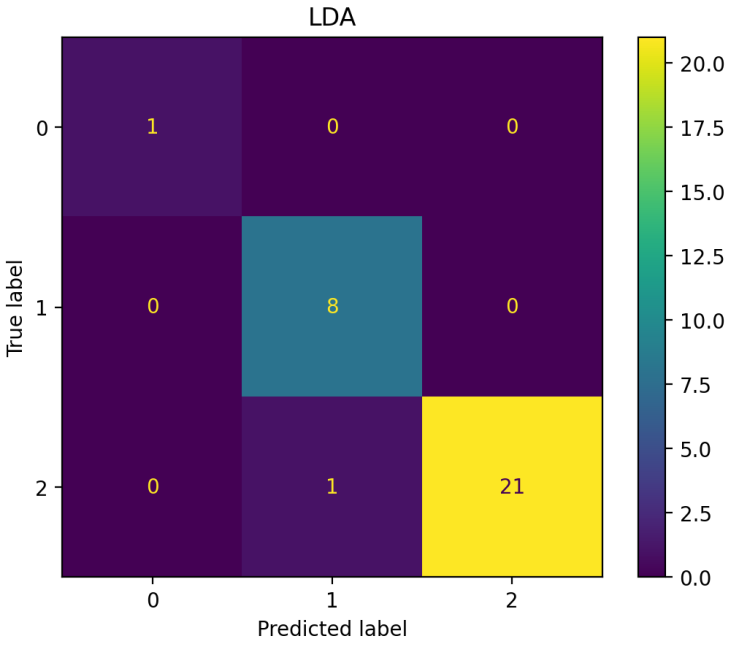

Figure 7: Confusion matrix for LDA classifier.