Sean Arthur Wright

Michigan State University

Wells Fargo Data Analysis

Deliverable 1

My program utilizes a K-Means unsupervised machine learning algorithm. The data is initially broken down into individual pandas dataframes based off of each activity. Each of these subsequent dataframes has a resulting column for how much carbon was consumed by the individual for the given activity. This was calculated by multiplying the consumption value by the carbon footprint factor. This proved to be most difficult because many individuals provided a means of consumption that did not have a valid carbon factor to link to. If the user only provided one valid means of consumption it was considered the best case scenario. If the user provided an invalid means of consumption the factor value was dropped. If the user provided no valid means of consumption or multiple valid means of consumption then an average was taken for the provided carbon factors. The Quality of Life column was added without modifying any values because we do not want to reduce the happiness of any individual. The next column calculated was the Carbon Consumed per Quality of Life unit Index (CCpQ). This value is simply the quotient of the previous two columns and is used to determine exactly how much carbon each individual consumed in order to attain one “happiness” unit. This value is used to determine exactly how efficient each individual is with their consumption methods. The algorithm then takes each of these 27 activity dataframes and constructs a single dataframe with each activity’s CCpQ. A new column is added which calculates the total CCpQ for each individual. Lastly, an efficiency column is added which simply calculates whether the individual is in the top (marked as ‘1’) or bottom (marked as ‘0’) 50% of consumers. Using the sklearn python library the algorithm passes this final dataframe without the efficiency column because this will be used to see if our machine learning program is calculating the correct result. The K-Means algorithm will attempt to split our data into N groups (in our case 2), and cluster the data points into proper groups based on how efficient the individual was. When the model was ran the machine learning algorithm was able to predict whether an individual was a high or low carbon consumer with around ~99% accuracy.

When paired with my application the intent is to be able to inform consumers whether or not they are high carbon consumers on a daily basis. The application would be provided through every app store and each consumer will fill in information regarding the means of consumption they use (electric vs gas water heater etc). Daily push notifications will be sent to the use with three questions in order to gain more information about their consumption habits on a daily basis. The individuals consumption pattern is then passed on to the machine learning algorithm and the user is informed about what percentile they fall under and if they are high carbon consumers or not. I believe if people can see exactly how much they consume compared to their peers it has the potential to majorly influence the ways in which we consume.