# Reducing Household Carbon Emissions using Machine Learning

Stepheny Perez, Omid Jafari, Tanya Olivas Rico

Fall 2018

## 1 Introduction

Our team found an analytic challenge appropriate for this project. This is a real-world example with a tangible reward other than the experience we will acquire using techniques learned during this course. The dataset provided contains information about households and their power consumption and associated carbon footprint. In addition to consumption, quality of life importance (or QoLI) is provided for each household, which indicates how important that particular activity is to that household. Using this information, we want to suggest consumption modifications to households to lower their carbon footprint without reducing their quality of life.

By analyzing said data, the team will be able to describe how the data product succeeds mathematically in minimizing an individual's carbon footprint with minimal negative impact on their utility. The end product is implemented operational code that can be run with any data set provided by the client in order to make accurate predictions for said read-world problem.

## 2 Motivation

Environmental sustainability is an important goal for our generation. In this project, we hope to show how small changes in a household can make a large difference in overall carbon footprint. This is a difficult task because most people don't want to give up their lifestyle, which in this case we represent by "Quality of Life". We are trying to find a balance between high quality of life and low carbon footprint, so both the earth and the individual are happy. We found this idea on a campus challenge. *Campus Analytics Challenge: Live Green and Live Happy*

# 3 Methodology

## 3.1 Dataset

We are given a data set describing 1000 individuals' daily activities; there are 27 activities which include units of consumption and are given a Quality of Life factor ranging 1-100 each. There is also a list of equipment each individual might use for each activity. We are also given the Carbon footprint of each equipment for each activity.

Below is a simplified representation of the dataset provided to us for this challenge.

| | | | |
|---|---|---|---|
| household1 | a1 consumption | a1 QoLI | a1 CF |
| household1 | a1 consumption | a1 QoLI | a2 CF |
| household1 | a3 consumption | a3 QoLI | a2 CF |
| ... | ... | ... | ... |
| household2 | a1 consumption | a1 QoLI | a1 CF |
| household2 | a1 consumption | a1 QoLI | a2 CF |
| household2 | a3 consumption | a3 QoLI | a2 CF |
| ... | ... | ... | ... |

## 3.2 Preprocessing

The first step is preprocessing the data. The data was provided in an Excel file with two sheets, where the second sheet is simply a lookup table containing the carbon footprint of certain activities. We combine this data with the full dataset so the carbon footprint of an activity is readily available for analysis.

The second step is to replace all empty cells with '0' (after concluding this was the intention of an empty cell), and to normalize the data.

The third step is to prepare for clustering. As it currently stands, each household has a row for each activity in the dataset. We combine and re-arrange these values so each household has only one row in the dataset, along with consumption for each activity and QoLI for each activity. This also makes it easier to read and work with the data later, because the previous layout was difficult because of the multiple rows for each sample.

## 3.3 Clustering

The second step is to cluster the rearranged data. This should give us clusters of households with similar usage patterns. We will be using the k-means clustering method, though this may change in the future if we learn new clustering methods that are more appropriate. We will determine our k value by

using the elbow method Raschka, 2017, because we don't know exactly how many clusters will be needed to represent similar households. We don't have a specific metric by which we are clustering, we are just hoping that they are similar enough that we can use the same consumption modification suggestions on the entire group. You can see above in the figure, the elbow method suggested that we use 2-3 clusters, so we chose 3.

### 3.4 Predicting missing attributes

We thought it would also be useful to be able to predict any missing attributes in a particular sample. For example, let's say a particular household didn't want to disclose their vehicle consumption use. We can use this model to predict what that value probably is, by training a separate regressor for each attribute, where that attribute is $y$.

### 3.5 Consumption Suggestions

Next, we will suggest consumption modifications to each cluster of households. Because each cluster represents similar households in terms of consumption and QoLI, we can assume that any consumption modifications will similarly affect each household in that cluster and therefore be appropriate for each household in that cluster. This is not a machine learning problem, but an optimization problem, where we want to minimize one variable (carbon footprint) while maximizing another (total quality of life for that household). Because we didn't learn about optimizations in this class, we chose to go a more trivial route. First, we find the household in each clustering with the highest quality of life, which is the sum of the product of the consumption of all actions and the QolI of that same action. Then, we find the household with the lowest carbon footprint in each cluster, which is done by summing the total carbon footprint for each household. Then, we average their consumption of each activity, and that's the new suggestion. For example, if Household 1 has consumption of the household heating ¿= 70% as 10.0, and Household 2 has the same action consumption as 2.0, and Household 1 has the highest happiness and Household 2 has the lowest carbon footprint, we will suggest to all households in the cluster to consume 6.0 household heating instead, assuming that is lower than their current consumption.

### 3.6 Results

As you can see above, we have done clustering to group each household together with other households that have similar behavioral and preference patterns. Now, in the next steps, we can suggest consumption modifications to each group dependent upon their consumption and preferences. We can assume

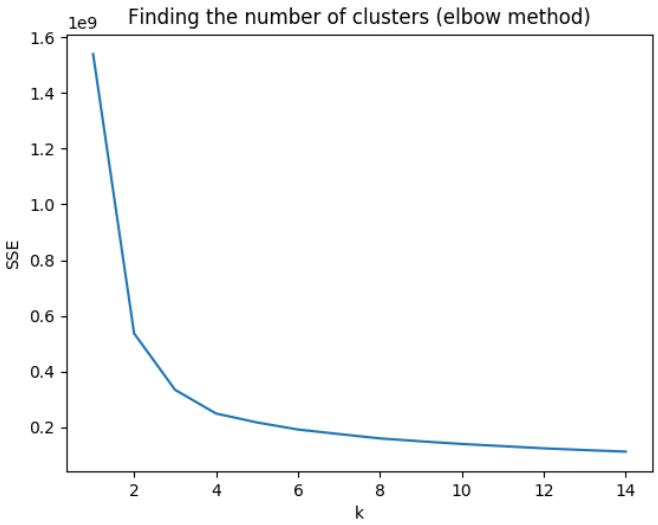


Figure 1: Elbow method

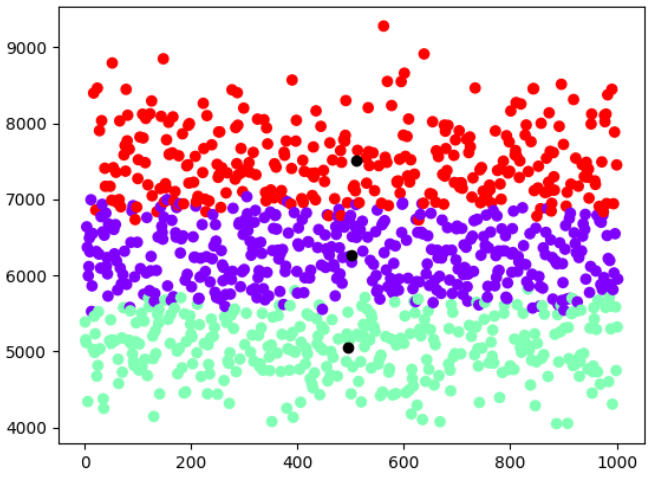


Figure 2: Clustering results k = 3

that the households in each cluster are similar to one another and can be given similar instructions.

You can also see above that we ran the elbow method and got k = 3. This clearly worked well, because the clustering seems somewhat logical, though it's not as clear of a cluster as the examples from class.

Another result we have to share is the regression results. As mentioned before, we ran a regression for every unique parameter (except ID) meaning there were 28 total regressions run.

* Regression model for "Household heating =¿ 70F" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.918
Regression model for "Household heating ¡ 70F" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.064
Regression model for "Use of heat pump" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.931
Regression model for "Use of air conditioner" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.010
Regression model for "shower - short" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.991
Regression model for "shower - long (¿ 3 min)" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.114
Regression model for "bath" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.690
Regression model for "wash-up" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.056
Regression model for "use of dishwasher" *
– Best parameters values –
'alpha': 1.0

– Mean squared error: 1.225
Regression model for "use of clothes washer" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.095
Regression model for "use of clothes dryer" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.999
Regression model for "use of cooking range" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.840
Regression model for "use of oven" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.115
Regression model for "use of self-clean feature of electric oven" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.123

* Regression model for "Small kitchen appliance in the home" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.180

* Regression model for "TV/computer use" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.027

* Regression model for "air travel - large plane" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.131

* Regression model for "air travel - small plane (¡50 seats)" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.856

* Regression model for "car trips- self only" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.946

* Regression model for "car trips - driver and self" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.982

* Regression model for "car trips - 2+ people with multiple end points" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.077

* Regression model for "trips using public ground transportation" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.018

* Regression model for "bags of garbage disposed" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.992

* Regression model for "bags of recycling deposited (negative CF)" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.151

* Regression model for "bags of compost deposited (negative CF)" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.945

* Regression model for "hazardous or electric items disposed" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 1.118

* Regression model for "large items disposed" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.996

* Regression model for "QoLI" *
– Best parameters values –
'alpha': 1.0
– Mean squared error: 0.133

Total number of regressors: 28

# References

*Campus Analytics Challenge: Live Green and Live Happy*. `https://www.mindsumo.com/contests/campus-analytics-challenge-2018`. Accessed: 2018.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Raschka, Sebastian (2017). *Python Machine Learning. Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition*. Packt.