Roy Siebers | 2565072 | 5 december 2017

Machine learning mini project

MINOR APPLIED DATA SCIENCe

The assignment

Describe the data

Name of the dataset: Smart meters in London

Link: <https://www.kaggle.com/jeanmidev/smart-meters-in-london>

Domain: Weather, Energy, Home, Smart homes.

Explore the data

**What is in the dataset?**

From the Kaggle page the following should be inside the dataset:

There is 19 files in this dataset :

**informations\_households.csv** : this file that contains all the information on the households in the panel (their acorn group, their tariff) and in which block.csv.gz file their data are stored

**halfhourly\_dataset.zip**: Zip file that contains the block files with the half-hourly smart meter measurement

**daily\_dataset.zip**: Zip file that contains the block files with the daily information like the number of measures, minimum, maximum, mean, median, sum and std.

**acorn\_details.csv** : Details on the acorn groups and their profile of the people in the group, it's come from this [xlsx spreadsheet](https://acorn.caci.co.uk/what-is-acorn).The first three columns are the attributes studied, the ACORN-X is the index of the attribute. At a national scale, the index is 100 if for one column the value is 150 it means that there are 1.5 times more people with this attribute in the ACORN group than at the national scale. You can find an explanation on the [CACI website](https://acorn.caci.co.uk/what-is-acorn)

**weather\_daily\_darksky.csv** : that contains the daily data from [darksky api](https://darksky.net/dev). You can find more details about the parameters in [the documentation of the api](https://darksky.net/dev/docs#response-format)

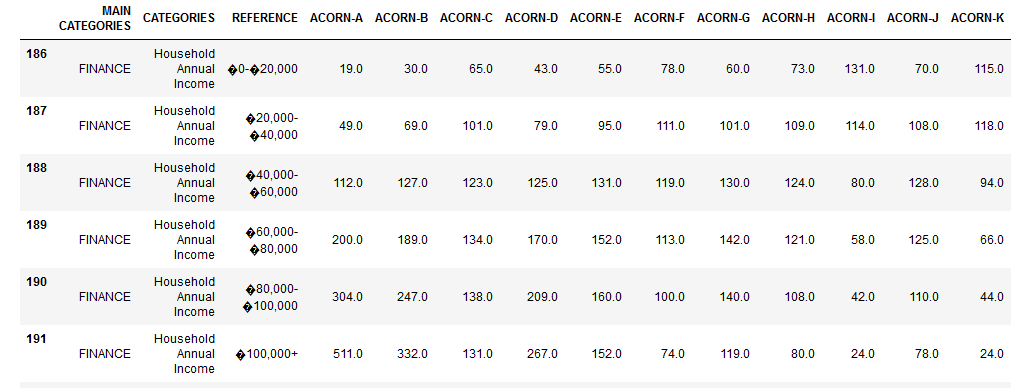
**weather\_hourly\_darksky.csv** : that contains the hourly data from [darksky api](https://darksky.net/dev). You can find more details about the parameters in [the documentation of the api](https://darksky.net/dev/docs#response-format)

For more exploration of the data see **IndividueleOpdrachtExploration.ipynb**

**Next step**

So what I wanted to do was create an machine learning tool to predict what different acorns would use in power, when only given what their average income compared to the rest of london.

Example:



In the image above, we can see that ACORN-A has 511% of the national average 100.000+ household annual income.

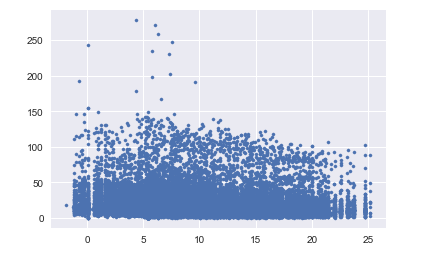
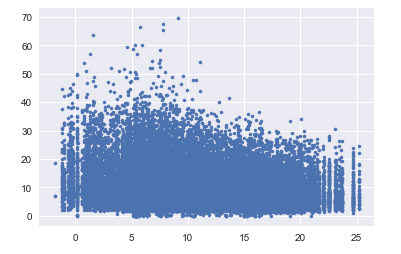
**Deploying the machine-learning**

Looking at the data I would like to show the difference in usage of electricity in different Acorns.

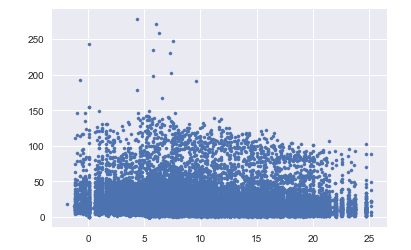
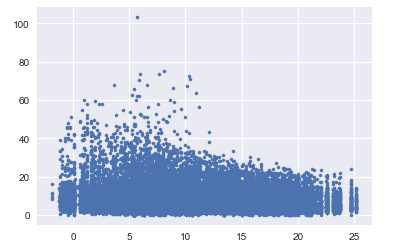
To do that I have split the data of different acorns that I believe would be interesting to use.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Acorn** | **Training set** | **Test set** |
| A | Block\_0 | Block\_2 |
| C | Block\_4 | Block\_5 |
| D | Block\_7 | Block\_8 |
| F | Block\_50 | Block\_52 |

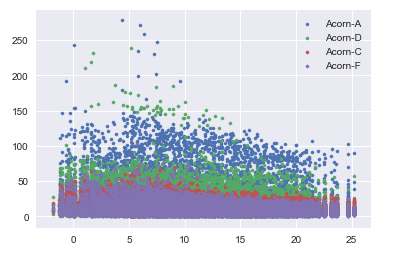
Below shows the data of each of the testsets.



Figuur 1: Acorn-A Figuur 2: Acorn-C

Figuur 3: Acorn-D Figuur 4: Acorn-F

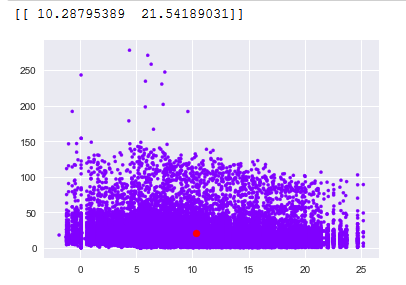


The scatter plot above shows the energy usage of 4 different acorns plotted into one graph,

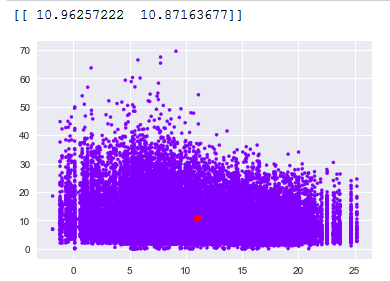
When we take the income data we come to the hypotheses that higher income means more energy consumed.

What I wanted to do show in what acorn someone lives according to their energy consumption.

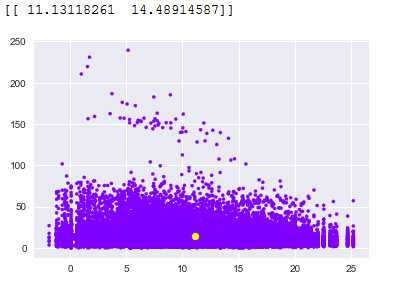
So what I did was I used k-means to determine what the middle was for each cluster. As seen below.



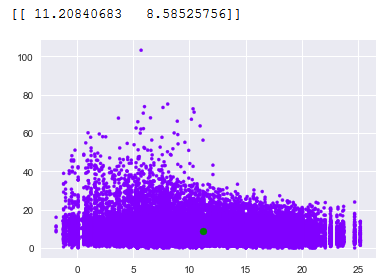
Figuur : Acorn-A



Figuur :Acorn-C



Figuur :Acorn-D



Figuur 7:Acorn-F

With this done, I want to see if the average of my test data from each acorn is a good indication to what acorn a household belongs to.

To do this I’ve used the data from the test set.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Acorn** | **Train cluster center** | **Test cluster center** |
| A | [[ 10.28795389 21.54189031]] | [[ 11.13697452 18.71338757]] |
| C | [[ 10.96257222 10.87163677]] | [[ 10.70207193 13.39257876]] |
| D | [[ 11.13118261 14.48914587]] | [[ 11.083244 14.20186307]] |
| F | [[ 11.20840683 8.58525756]] | [[ 11.2550947 9.51569166]] |

As we can see in above data, the in 3 out of 4 Acorns the cluster center of the train data was closest to cluster of the test data. But this means very little. The only thing we can say is that we have seen that the higher income an acorn is that they probably consume more energy.

Conclusion

In the end we could see that the energy usage was close to each other in each Acorn, this allowed me to use the center of the cluster as a measuring tool. This however is not the right way to use K-means. Because all it did was show the average energy consumption per day. Instead I should have used different ways to show the data. Because of a lack of knowledge about what K-means is supposed to do I thought it would be a good one to use. Nevertheless I used it the wrong way, if I was going to use it correctly I should have at least labeled the data with that acorn they belong to.

Reflection

Let’s start with saying that I made some really bad decisions during the course of this project. I started late, I didn’t ask for feedback on my project and I am really unhappy with the end result.

When I started this course I was very excited to learn about big data and machine learning but when the course started to progress I started to become demotivated to put the work in, with the noisy classrooms to having to relocate a lot during lessons. My motivation dwindled and I started to not do my homework anymore. These were things the department could do nothing about and I do not blame them for it, I just wanted to tell my reason for my motivation dropping.

With me falling behind on the schedule I was afraid of starting on my mini project. I felt like if I push it away I don’t have to feel terrible right now for not knowing the things I was supposed to know. While in hindsight if I just went and talked with Mr. Lambooij I am sure that we could have worked things out.

In this semester it has once again shown me that I have very visible shortcomings. I have a lot of trouble with doing a bit of work every week instead of doing it all in a couple of days. I am scared to ask for feedback on things that I am supposed to know, I always feel like it will be alright anyway. Maybe that is because I’ve been able to get away with it so many times that it just snuck in as natural for me.

But this time I hit a brick wall. This project was impossible to successfully complete. I simply did not have the knowledge to do this in ‘a couple of days’ like I normally do.

Even when I had to show my midway-portfolio I knew in the back of my head that I didn’t do enough, but still told myself everything will be ok. Because it always had been. But this time went it went different.

The last problem I had was my chaotic agenda. I had no idea the project was supposed to be done the 11th I always had in mind that we had at least till the 18th. When I found this out I put every moment of my time into this project. The first few days I worked from 10:00 till 20:00 for 3 days straight, and after that from 10:00 till 23:00.

My main problem is that I always think everything will be OK, until its too late.

I feel like I can do better but have not been able to prove that in this project.