How the data was split

Feature selection ‘athome’ and ‘unit’

Why was datetime not used why was it serialised

Date acted like an iteration, of what is happening every minute. It acted as a series so for simpler access it was converted into a series

The first unit represents the first minute of *falana* day and the 36270th unite representing the last minute of *falana* day, denoting the end of the data set.

General concerns:

The hardest part about the dissertation was understanding the dissertation itself. The sub-sections such as literature review, methodology, analysis, results and conclusions had straight forward names and explanations that seemed to be even more understandable. We were told what was expected of us in each section and their implementations but it was the actual implementation that was challenging. Related work was to be read and analysed in the literature review but exactly what areas where crucial to discuss, what information had to be in the summary and what was unnecessary. Although the initial understanding was quite clear it wasn’t until I got to typing did I realise I had so many questions and concerns. They were handled and rectified in due time. One very important thing that helped with this would be the early submission of the Project Proposal which had a mini Literature Review, abstract and client motivation which helped getting us in the mindset of understanding what and how the dissertation worked. Although not much from the initial project proposal was made use of in the final project hand in, it was still of immense importance when it came to grasp the idea of the professionalism and formal nature of the report.

Dataset:

Projects that involve data mining often struggle with finding the perfect data set. I was lucky in terms that the I wanted a minimalist data set. An extremely simple one with very little and necessary attributes only. The struggle came in choosing between the two datasets that I narrowed it down to. One dataset was a public one from Kaggle. It included the required data but had other additional data and that too of 30 different households in a neighbourhood. This data was comprehensive yet a bit too intricate for the intentions of my project. Another aspect of this data set was that power reading from the household were taken every 30 minutes which made it undesirable to be worked with given the urgency and demand of my project application.

The other dataset that I chose came from my supervisor’s personal household. It changed the course of my ethics application and some formalities and consent forms had to be fulfilled but the dataset was perfect when it came to the requisites of the project. This data contained 3 attributes in total, datetime, power and a numeric variable to denote the number of people at home. Minimum being no people and a maximum of two people. A significant problem that was faced during the importing and using of data within python was the “datetime” attribute. Since there were power readings for every minute of every day, that meant there were 24 x 60 = 1440 readings for a single day and the data set was collected over a period of 2 months (approximately 60 days). The splitting and usage of the “datetime” attribute was not impossible but laborious. And given a situation where data had to be well broken down, the “datetime” field was left as it is as a reference and another serialised attribute by the name “unit” was added to the dataset. This field was used as a unit of time in minutes to denote the first minute of the 2nd of January 2018 at 15:26 and the 36270th value representing the last minute of the data set at the 23rd of February 2018 at 11:12.

The importance of the occupancy prediction weighed out over the patterns of electricity data usage as per the nature of days and the types of days (for example Mondays, weekdays, weekends etc). It is crucial when you think of it, it is crucial to help refine the system, a false negative works just as well to indicate there is room for concern.

The split for the training and testing data was an additional area that drew some time. It was advised in the end of term presentation feedback that the split wasn’t right down the middle or that the training included all the dates of January and some of February and the testing data of just February rather make the split more random.

The percentage of the split was also a concern and there is debate when it comes to a fair value. Some say an 80:20 is a fair split, whereas in some cases and scenarios even a split of 70:30 is justified. But then again, these split ratios are also highly dependent on the size of the data set.

I personally landed at using a 67:33 training to testing ratio. Given 36210 instances, it left me with a seemingly healthy split of 24300 learnable instances to 11970 predicting/testing instances. The split was also ensured to be random. The python built-in data split had a function of splitting the data randomly, not right down the middle and as shown in the figure below, the first 20 instances of both the testing and the training data was random.

|  |  |
| --- | --- |
| **Training dataset (67%)** | **Testing dataset (33%)** |
|  |  |

\*The numbers on the right-hand side is the additional field that was added to the data set to serialise the headings of the power and ‘athome’ variable readings. Although the first 20 values only aren’t sufficient to display the randomness in the collection and split of the data it is still apparent that the data was not collected serially.

This part of the code is responsible for the split. The split ratio and the randomness of the split. The sklearn library has a function “model\_selection**.**train\_test\_split” which takes in attributes such as the output variables, the size of the testing data and the random state.

X\_train**,** X\_test**,** Y\_train**,** Y\_test **=** sklearn**.**model\_selection**.**train\_test\_split**(**X**,** y**,** test\_size **=** 0.33**,** random\_state **=** 4**)**

Implementation/Algorithms:

The implementation of the machine learning algorithms seemed to be frightening. And it was at first. There was a learning curve for the understanding of how they work. Small things such the training set, the testing set, supervised and unsupervised learning, different types and techniques to do the same thing. The availability of so many different techniques is what encouraged me to try out different ones and statistically analyse which one worked best. The coding and actual implementation of the code for the k- NN regression and linear regression didn’t take as long as understanding their concepts did.

The applications of this project, although motivated towards the senior and disabled to inspire and help along their independent living, can also be implemented in different scenarios. Like in the case of an instances where we predict a false positive, where the prediction tells us there is someone home and the known data is telling us they aren’t, this could in some ways signify a breaking in and the presence of a burglar etc.

This issues paper explores voluntary euthanasia. It is not intended to be exhaustive, however it aims to add to considerations of this very complex and sensitive topic through analysis of the domestic regulatory environment relating to both passive and active forms of voluntary euthanasia, and of relevant international laws by way of comparison with domestic regulation. It concludes with a human rights-based analysis of voluntary euthanasia and...