INFO411: Data Mining and Knowledge Discovery

Project 9

Instructions:

This task is a real-world data mining problem. You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) your proposed data mining approach and methodology; (3) the strengths and weaknesses of your proposed approach; (4) the performance measures that can evaluate your data mining results; (5) the results and a brief discussion. Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Energy Rating Data for Air Conditioners

Background:
The Energy Rating Label (ERL) is a mandatory comparative energy label that provides consumers with product energy performance information at point-of-sale on a range of appliances. Attached to each appliance, allows comparison between similar appliance models through a star rating system (the greater the number of stars, the higher the efficiency) and the annual energy consumption. Further details are available from

`http://www.energyrating.gov.au`

Data has been collected from suppliers when they register appliances that are going to be sold in Australia and New Zealand, available from

`http://data.gov.au/dataset/energy-rating-for-household-appliances`.

A new star rating system has been used since 2010, but the data also includes values for the previous SRI system used prior to 2010. Estimated yearly power consumption is also provided for each appliance. Some fields such as Model Number/name will not be relevant for this data mining task, other than for labelling purposes. Some fields may require pre-processing.

A factor that distinguishes air conditioners is that, unlike most other appliances, have two ratings: one for their efficiency when cooling and the other with their efficiency as heating. (Though, not all models can cool and heat.)

Requirements:

1. Download the latest airconditioning dataset from the above website, and restrict to models that are available (using the Available Status field) and to models that are sold in Australia. You can identify models that are sold in Australia by whether "Australia" appears in the Sold_in field. This can be done using the `grep` function in R as follows:

   grep returns true false vector

   ```
   subdata <- alldata[grep("Australia",alldata$Sold_in),]
   ```

2. You are required to propose a model for predicting the SRI values (quantitative) of air conditioners according to the new star rating system used since 2010, using other relevant variables apart from the star ratings. Do this for both their cooling and their heating ratings.

3. Also, propose a model to predict the yearly energy consumption of the air conditioner in heating and a model to predict it in cooling.

4. Also, describe, with summaries, visualisations, and by other means, the relationship, if any, between an air conditioner's rating when cooling and its rating when heating. Is there evidence to believe that the relationship holds after accounting for the size, design, maker, and other properties of the AC?

5. Discuss any data preprocessing and selection of attributes which have been applied.

6. You need to provide the performance measures of your prediction results.

7. Present all of your models, with particular attention to informing consumers about important factors which affect energy efficiency.

8. Discuss the variables which are most important for the purpose of predicting energy rating.