

# COMP3425 and COMP8410 Data Mining 2019

## Assignment 2

### Description of Data

The data supplied for the assignment is in one 4.1 MB file, called **Stock\_prices\_2018.csv**. It is a tabular, comma-separated values, newline-terminated format, with a first line of attribute names. It comprises daily stock prices and volumes for 128 not-very-randomly-selected Australian stocks, over the period 1 January 2018 to 31 December 2018, with a line for each stock on each day.

The data was derived by screen-scraping the Wall Street Journal stock prices at [quotes.wsj.com](http://quotes.wsj.com). For example for Australian company BHP, this page provides most of the data:

<http://quotes.wsj.com/AU/XASX/BHP/historical-prices>.

The *Date*, *Open*, *Close*, *High*, *Low* and *Volume* attributes for each stock come directly from there, with necessary transformation to present as

- a *yyyymmdd* format for the *Date*;
- decimal numbers for the prices in \$AU; and
- a non-negative integer for the *Volume*.

The “BHP” in the url path fragment above is a stock code that identifies a stock for those attributes, and is called *Code* in the data.

In addition, the data has been supplemented with *Sector* and *Subsector* categories for each stock. This information was also taken from the Wall Street Journal web site, from <http://quotes.wsj.com/company-list>.

Several attributes have been derived from the *Date* for each line: *Weekday*, *DayofMonth*, *Month*, *Year*, *WeekofYear*, and *DayofYear*. Of these, *Weekday* and *Month* are textual and the others are respectively 2,4,2 and 3 digit positive numbers with leading zeros.

The remaining attributes have been derived from the prices and *Volume* as follows.

- *Close-Open*:  $Close - Open$
- *Change*: “up” if *Close-Open* is zero or positive, “down” otherwise
- *High-Low*:  $High - Low$
- *HMLOL*:  $(High - Low) / Low$
- *PriorClose*: The *Close* value of the stock on the *Date* immediately prior to this one. On the earliest day given for each stock, the *PriorClose* is set instead to the value of *Close* from the same day.

**Hint:** This dataset is quite large for the R implementations of several data mining algorithms. Be prepared to reduce the training size when necessary (you can use the Partition feature in Rattle to set the training data set size).