COMP3425 Data Mining 2019

Assignment 2

Maximum marks	**100**
Weight	**20% of the total marks for the course**
Length	**Maximum of 10 pages, excluding cover sheet, bibliography and appendices.**
Layout	**A4 margin, at least 11 point type size, use of typeface, margins and headings consistent with a professional style.**
Submission deadline	**9:00am, Monday, 20 May**
Submission mode	**Electronic, via Wattle**
Estimated time	**15 hours**
Penalty for lateness	**100% after the deadline has passed**
First posted:	**1st April, 9am**
Last modified:	**1st April, 9am**
Questions to:	**Wattle Discussion Forum**

This assignment specification may be updated to reflect clarifications and modifications after it is first issued.

It is strongly suggested that you start working on the assignment right away. You can submit as many times as you like. Only the most recent submission at the due date will be assessed.

In this assignment, you are required to submit a single **report** in the form of a PDF file. You may also attach supporting information (appendices) as one or more identified sections at the end of the same PDF file. Appendices will not be marked but may be treated as supporting information to your report. Please use a **cover sheet** at the front that identifies you as author of the work using your u-number and name, and identifies this as your submission for COMP3425 Assignment 2. The cover sheet and appendices do not contribute to the page limit. You are expected to write in a style appropriate to a professional report. You may refer to http://www.anu.edu.au/students/learning-development/writing-assessment/report-writing for some useful stylistic advice. You are expected to use the question and sub-question numbering in this assignment to identify the relevant answers in your report.

No particular layout is specified, but you should follow a professional style and use no smaller than 11 point typeface and stay within the maximum specified page count. Page margins, heading sizes, paragraph breaks and so forth are not specified but a professional style must be maintained. Text beyond the page limit will be treated as non-existent.

This is a single-person assignment and should be completed **on your own**. Make certain you carefully reference all the material that you use, although the nature of this assignment suggests few references will be needed. It is unacceptable to cut and paste another author's work and pass it off as your own. Anyone found doing this, from whatever source, will get a mark of zero for the assignment and, in addition, CECS procedures for plagiarism will apply.

No particular referencing style is required. However, you are expected to reference conventionally, conveniently, and consistently. References are not included in the page limit. Due to the context in which this assignment is placed, you may refer to the course notes or course software where appropriate (e.g. "*For this experiment Rattle was used*"), without formal reference to original sources, unless you **copy text** which always requires a formal reference to the source.

An assessment rubric is provided. The rubric will be used to mark your assignment. You are advised to use it to supplement your understanding of what is expected for the assignment and to direct your effort towards the most rewarding parts of the work.

Your assignment submission will be treated confidentially. It will be available to ANU staff involved in the course for the purposes of marking.

Task

You are to complete the following exercises. For simplicity, the exercises are expressed using the assumption that you are using Rattle, however you are free to use R directly or any other data mining platform you choose that can deliver the required functions. You are expected, in your own words, to interpret *selected* tool output in the context of the learning task. Write just what is needed to explain the results you see. Similarly, you should describe the methods used in terms of the language of data mining, not in the terms of commands you typed or buttons you selected.

1. Platform

Briefly describe the platform for your experiments in terms of memory, CPU, operating system, and software that you use for the exercises. If your platform is not consistent throughout, you must describe it for each exercise. This is to ensure your results are reproducible.

2. Data

Look at the pairwise correlation amongst all the numeric variables using Pearson product-moment correlation.

(a) Explain **why** you would expect *DayofYear* and *WeekofYear* to be highly correlated.

(b) Qualitatively **describe the correlations** amongst the variables *High*, *Low*, *Open*, *Close* and *Volume.* **Explain** what you see in terms of the source of the data.

3. Association mining: What factors affect volume of sales?

(a) **Compute** and give the 5-number summary for *Volume*. Qualitatively **describe** what it tells you about *Volume*.

(b) For this exercise, bin *Volume* into 5 categories using Quantiles. **Why** is quantile binning appropriate for association mining with *Volume*?

When you have completed this question 3, remove the extra variable you created so they do not interfere with other exercises.

(c) Generate association rules, adjusting min support and min confidence parameters as you need. **What** parameters do you use? Bearing in mind we are looking for insight into what factors affect *Volume*, **find 3 interesting rules**, and explain both **objectively** and **subjectively** why they are interesting.

(d) **Comment** on whether, in general, association mining could be a useful technique on this data.

4. Study a very simple classification task

Aim to build a model to classify *Change*. Use *Change* as the target class and set every other variable as Input (independent). Using sensible defaults for model parameters is fine for this exercise where we aim to compare methods rather than optimise them.

(a) This should be a very easy task for a learner. **Why?** *Hint:* Think how *Change* is defined.

(b) Train each of a Linear, Decision tree, SVM and Neural Net classifier, so you have 4 classifiers. Hint: Because the dataset is large, begin with a small training set, 20%, and where run-time speeds are acceptable, move up to a 70% training set. **Evaluate** each of these 4 classifiers, using a confusion matrix and interpreting the results for the context of the learning task.

(c) **Inspect** the models themselves where that is possible to assist in your evaluation and to explain the performance results. Which learner(s) performed best and **why**?

5. Predict a Numeric Variable

One investment strategy could rely on the previous day's price to predict the opening price for a stock, enabling you to place a buy or sell offer overnight ready for the next day. To predict the opening price for a day, you cannot use any of the other prices or *Volume* for that same day as that information is not available until the close of the day, when it is too late. So, using the variables you have in the dataset, but ignoring all of *High, Low, Close, Volume, Close-Open, Change, High-Low,* and *HMLOL*, aim to predict the opening price *Open* using the previous day's closing price *PriorClose.* and the date and stock-related variables. Use a regression tree or a neural net.

(a) Explain which you chose of a regression tree or neural net and **justify** your choice.

(b) Train your chosen model and tune by setting controllable parameters to achieve a reasonable performance. **Explain** what parameters you varied and how, and the values you chose finally.

(c) **Assess** the performance of your best result using the subjective and objective evaluation appropriate for the method you chose, and **justify** why you settled with that result.

6. More Complex Classification

An alternative investment strategy might be to predict where there will be a big proportional change in the price over a day, once the day has opened, and so a good opportunity for a short term gain (or loss) that day.

(a) Transform HMLOL to a categoric class variable by binning into 2 classes, using a k-means clustering of the HMLOL. The skewed distribution of HMLOL is helpful here as the higher values we need for investing are relatively rare. When you have completed this question 6, remove the extra variable you created so it does not interfere with other exercises. Now, be sure to ignore most of the current day price and volume variables, that is ignore all of *High, Low, Close, Volume, Close-Open, Change, High-Low,* and *HMLOL*. **Explain** why *HMLOL* should be ignored. This time, use the *Open* price, *PriorClose* and all the date and company description variables for learning.

Initially, use a small training set, 20%, and where run-time speeds are acceptable, experiment with a larger training set. **Explain** how you will partition the available dataset to train and validate classification models below.

(b) Train a Decision Tree Classifier. You will need to adjust default parameters to obtain optimal performance. **State** what parameters you varied and (briefly) their effect on your results. **Evaluate** your optimal classifier using the **error matrix**, **ROC**, and any quality information specific to the classifier method.

(c) Train an SVM Classifier. Then proceed as for (b) Decision Tree above.

(d) Train a Neural Net classifier. Then proceed as for (b) Decision Tree above.

7. Clustering

(a) Restore the dataset to its original distributed form, removing any new variables you have constructed above. For clustering, use only the five raw variables, *Date*, *Open*, *High*, *Low* and *Volume* and remove all of the others.

Experiment with clustering using the k-means algorithm. Rescale the variables to fall in the range 0-1 prior to clustering. Use the full dataset for clustering (do not partition) by building cluster models for each of *k*= 2, 5 and the recommended default (i.e. squareroot (n/2) for dataset of size n) clusters. Choose your preferred *k* and it's cluster model for k-means to answer the following.

(a) **Justify** your choice of *k* as your preferred (*Hint:* have look at parts b-d below for each cluster model).

(b) **Calculate** the sum of the within-cluster-sum–of-squares for your chosen model. The *within-cluster-sum–of-squares* is the sum of the squares of the Euclidean distance of each object from its cluster mean. **Discuss** why this is interesting.

(c) Look at the cluster centres for each variable. Using this information, **discuss** qualitatively how each cluster differs from the others.

(d) Use a scatterplot to plot (a sample of) the objects projected on to each combination of 2 variables with objects mapped to each cluster by colour (*Hint:* The Data button on Rattle's Cluster

tab can do this). **Describe** what you can see as the major influences on clustering. **Include** the image in your answer.

8. Qualitative Summary of Findings (approx 1/2 page)

Would *you* use these results to advise your investment decisions?

Comparatively **evaluate** the techniques you have used and their suitability or not for mining this data. This should be a *qualitative* opinion that draws on what you have found already doing the exercises above. For example, what can you say about training and classification speeds, the size or other aspects of the training data, or the predictive power of the models built? Finally, what else would you **propose** to investigate to assist your investment decisions?

Assessment Rubric

This rubric will be used to mark your assignment. You are advised to use it to supplement your understanding of what is expected for the assignment and to direct your effort towards the most rewarding parts of the work. Your assignment will be marked out of 100, and marks will be scaled back to contribute to the defined weighting for assessment of the course.

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
1. Platform & 2. Data	10	9-10 1.Platform description complete (memory, CPU, operating system, software). 2. All required correlations (*Year/DayofYear* & all pairs of *High*, *Low*, *Open*, *Close* and *Volume*) clearly explained in terms of the data domain, in the correct directions and for correct reasons demonstrating an understanding of the data.		7-8 1. Platform description complete. 2a partial or unclear 2b partial or unclear 2b partial explanation	5-6 1. Platform description complete. 2a attempt but with correlation direction wrong 2b Partial description of 4 variables or unclear 2b Partial explanation	0-4 1. Platform description incomplete. 2a. correlation reason missing or unrelated to *DayofYear* and *WeekofYear* 2b. Description unrelated to *High*, *Low*, *Open*, *Close* and *Volume* 2b. Explanation unrelated to data source
3. Association mining	10	9-10 Answers demonstrate deep understanding of association mining, by the careful selection of interesting and differentiated rules and clear rationale for interestingness. Comment shows original and insightful analysis of association mining on the problem.		7-8 a 5-number summary complete b 5-number explanation for Volume clear b Binning understood c Support and confidence clear c 3 interesting rules given c objective interestingness is given for all 3 c subjective interestingness attempted d Comment makes sense	5-6 a 5number summary ok b 5-number explanation for Volume poor b Binning misunderstood c Support or confidence not clear c < 3 interesting rules given c objective interestingness is incomplete c subjective interestingness is incomplete d Comment is cursory or off-track	0-4 Required information not provided and/or incorrect or misleading, demonstrating lack of engagement with the problem

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
4. Simple classification	10	9-10 Deep understanding of the 4 models demonstrated through analysis of performance on the change task.		7-8 a correctly explains why definition of change makes it seem easy b 4 confusion matrixes given b confusion matrixes explained in terms of the data and the method and the model learnt. c some evidence of understanding what the models are doing c reasoning for comparative performance demonstrating understanding of the methods behind them	5-6 a partially explains why definition of change makes it seem easy b 4 confusion matrixes given b confusion matrixes explained at face value only c weak understanding of learnt models c comparative performance only cursorily presented c reason for comparative performance is shallow	0-4 a inadequate explanation b confusion matrix missing b confusion matrix misunderstood c Interpretation of confusion matrix missing c no apparent understanding of what the models are doing c missing or unexplained comparative analysis
5. Prediction	20	17-20 Approach to problem demonstrates effort to produce good results and a deep understanding of the relative benefits of the 2 models in the context of the problem domain. Results are interpreted in the context of the problem domain.	14-16 a justification for choice shows understanding of the comparative benefits of each and extensive experiments with performance. b parameter variations shows a combination of experimentation and understanding of the parameters with justification for stopping at selected parameters. c several subjective and objective evaluation measures used as appropriate to method chosen c justification for stopping demonstrates awareness of appropriateness of best	12-13 a justification for choice shows understanding of the comparative benefits of each and experiments with performance. b parameter variations shows a combination of experimentation and understanding of the parameters c multiple subjective and objective evaluation measures used as appropriate to method chosen c justification for stopping demonstrates awareness of appropriateness of best result	10-11 a justification for choice shows some understanding of the comparative benefits of each or experiments with performance. b parameter variation demonstrates some experimentation c cursory evaluation given c justification for stopping perfunctory	0-9 a weak justification for choice b parameter variation insufficient c evaluation fails to demonstrate effort or understanding of evaluation c justification for stopping effectively absent

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
			result and scope of potential for further improvement			
6. Complex Classification	30	26-30 Exemplary use of classification models with comprehensive and fit-for-purpose performance analysis on the problem	22-25 a explanation correct b,c,d parameter variation clear and extensive demonstrating understanding of effect in all 3 methods b.c.d error matrix and ROC correctly interpreted in all 3 methods b,c,d extensive use of specific evaluation methods used and significance clearly explained in all 3 methods	18-21 a explanation correct a satisfactory approach to dataset partitioning b parameter variation clear and sufficient for good results b error matrix correctly interpreted b ROC correctly interpreted b some specific evaluation methods used c parameter variation clear and sufficient for good results c error matrix correctly interpreted c ROC correctly interpreted c some specific evaluation methods used d parameter variation clear and sufficient for good results d error matrix correctly interpreted d ROC correctly interpreted d some specific evaluation methods used	15-17 a explanation correct a satisfactory approach to dataset partitioning b parameter variation perfunctory b error matrix given b ROC given b few specific evaluation methods used c parameter variation perfunctory c error matrix given c ROC given c few specific evaluation methods used d parameter variation perfunctory d error matrix given d ROC given d few specific evaluation methods used	0-14 a explanation incorrect a unsound use of training/testing/validation data b no parameter variation b no error matrix b no or faulty ROC b specific evaluation methods missing c no parameter variation c no error matrix c no or faulty ROC c specific evaluation methods missing d no parameter variation d no error matrix d no or faulty ROC d specific evaluation methods missing

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
7. Clustering	10	9-10 The application of k-means algorithm to the dataset and its evaluation demonstrates exemplary understanding of the algorithm, its evaluation, and its limitations. Suitable evaluation methods or clustering experiments beyond those required here may be used.		7-8 a justification convincing b measure calculated correctly. Discussion recognises value and limitations c discussion on centres reflects numeric results and emphasises the interesting parts that relate to the significance in domain terms d correct image included and description shows understanding linked to data domain	5-6 a justification offered but not clear or unconvincing b measure calculated correctly c discussion on centres reflects numeric results d correct image included	0-4 Clustering experimentation and discussion inadequate
8. Qualitative Summary	10	9-10 Many aspects of evaluation are discussed and a clear conclusion is drawn, with direct reference to potential goals of the domain of the data. Proposal for further investigation demonstrates creativity and thoughtful engagement with the problem, clearly building on the work reported.	8 A clear conclusion is drawn from the work reported and a defended proposal for further investigation is proposed, with clear links to both the work reported and the domain of application.	7 A rounded, balanced summary of the work is presented with a justified proposal given.	6 A summary of the work is presented and a proposal made.	0-4 Answer does not demonstrate adequate engagement with the problem nor a qualitative understanding of the work reported.