31250 Introduction to Data Analytics

32130 Fundamentals of Data Analytics

Assignment 3: Data Mining in Action

Due date	**Friday, 11.59 pm, 12 October 2018**
Marks	Out of 100, weighted to 50% of your final mark.
Submission format	A report in PDF (preferable) or MS Word Doc and an oral defence.
Filename	ida_a3_xx.pdf or ida_a3_xx.doc where xx is your student id.
Report format	Around 10 pages for students in 31250 and 12 pages for student in 32130 with the information described below. Use 11 or 12 point Times or Arial fonts.
Submit to	UTS Online assignment submission button. **Please, make sure to call the filenames as described above.**

Scenario

This assignment is a practical data analytics project that follows on from the data exploration you did in assignment 2.

You will be acting as a data scientist at a consultant company and you need to make a prediction on a dataset. The dataset can be found on the UTSOnline site as well as on the Kaggle website (using the link that we will send to you by email).

You will be given one of the classifiers studied in class to work with. You need to build a classifier using that technique to predict the class attribute. At the very minimum, you need to produce a classifier using that method for the data. However, if you explore the problem very thoroughly (as you should do in industry), preprocessing the data, looking at different methods, choosing their best parameter settings and identifying the best classifier in a principled and explainable way, then you should be able to get a better mark. If you choose to use KNIME and you show 'expert' use (i.e. exploring multiple classifiers, with different settings, choosing the best in a principled way and being able to explain why you built the model the way you did), this will attract a better mark. If you choose to use R or Python to build, optimise and test different models, this will also attract better marks.

You need to write a short report describing how you solved the problem and the results you found. See below for requirements.

You also need to attend a short oral defense of your classifier of around 5 minutes where you show the classifier (e.g. using the KNIME workflow or Python/R code) and answer some questions about it. Details about oral defences will be given by email and in class.

Kaggle

For this assignment you will use the Kaggle website (kaggle.com) to download your assignment data and also submit to your assignment solution. The report itself will be submitted on UTSOnline as for the other assignments. We will provide you a link by email, which contains an invitation for the project. You need to use the link to access the project, because the project is private only for 31250/32130. Sharing the competition with anyone irrelevant to the subject is strictly prohibited. When you submit to Kaggle you will need to make a login to Kaggle using your **UTS email address**. You won't be able to submit to the competition if you use a non-UTS email address.

You will find 3 datasets: a training set for training your model (it contains the target values), a test set for testing the model (it does not have the target values – you need to predict them) and a submission sample which shows you what the submission file to Kaggle should look like.

The assessment is real time. This means that as soon as you submit the file, Kaggle will assess the accuracy of your system based on AUC and provide you the result. You can submit multiple times, but Kaggle has a limit for the number of times you can do this per day.

Do not use the accuracy or AUC reported from Kaggle as a measure of your test error in the final competition and optimise to it. This is because Kaggle has two measures: a public measure, which it reports to you, and a private measure, which it keeps hidden. Instead, develop several models and estimate the test error yourself before submitting to Kaggle. Remember that your estimate of test error is just that: an estimate. The actual private measure will probably be a little bit different.

Data sets

You can find the training dataset and the test dataset to evaluate the accuracy of your classifier on the Kaggle website using the link that we will provide to you by email.

Classification Task

Build a classifier that classifies the "**QUALIFIED**" attribute. You can do different data pre-processing and transformations (e.g. grouping values of attributes, converting them to binary, etc.), providing explanation why you have chosen to do that. You may need to split the training set into a training, validation and test sets to accurately set the parameters and evaluate the quality of the classifier.

You can use KNIME to build classifiers. Feel free to use any other tool such as other classifiers in R, Weka, Python, Orange, scikit-learn or other pieces of software. If you do this, though, please explain more about your classifier - and be sure that you are producing valid results! You don't need to limit yourself to the classifiers we used in class, but if you do you need to describe about them in your report and make sure you are producing valid results.

A hint: usually it's not a case of having a 'better' classifier that will produce good results. Rather, it's a case of identifying or generating good features that can be used to solve the problem.

Assignment report

In your report include the following information (3-5 pages):

- The **data mining problem, inputs, output;**
- The **data preprocessing and transformations** you did (if any).
- **How you went about the problem**
- **Classification techniques used and summary of the results and parameter settings**
- The **actual classifier that you selected** - the type, its performance, how it solved the problem (if it makes sense for that kind of classifier), and reasons for selecting it.

- **Reflection** section - ***for those students enrolled in 32130 only!*** Each student enrolled in 32130 additionally must write up to two pages more reflecting on their learning in assignment 3. What did you learn about data mining and yourself as a result of doing the assignment? How would you approach the problem differently if you were to do it again? The more incisive and thoughtful your reflection is, the better your mark.

On average each student will require between 24 and 36 hours to complete this assignment.

Prize

There will be a **class prize** for the submission that is at the first place in the assignment scoreboard on Kaggle.

Assessment

This assignment is assessed as individual work.

The **report** contributes up to 30 marks out of the 50. The marking criteria can be found on UTSOnline.

The **oral defense** contributes up to 20 marks of the 50. At the oral defense students need to submit their report to the examiner and answer questions about their solution showing the workflow (in KNIME) or the working code in Python, R or other tools. Students receive either 0, 10 or 20 marks in the following way.

Students showing the baseline classifier they were assigned in KNIME and are able to satisfactorily answer questions about it will receive 10 out of 20.

Students showing an in-depth investigation using many classifiers (in KNIME), with valid data preprocessing, parameter estimation and model evaluation, will receive 20 marks out of 20.

Students showing an in-depth investigation using Python or R (multiple classifiers, some data preprocessing, parameter estimation and model evaluation) will receive 20 marks out of 20.

Students who fail the oral defense will be permitted to undertake it once again. If they pass, they will receive a maximum of 10 out of 20.

Relationship to Objectives

This assignment addresses subject objectives 3, 4, 5 and 6.

Return of Assignments

Marks for the oral defense will be received a short time following the defense itself. The mark for the written report will be given within 3 weeks of submission. Feedback on the report will be given only for students requesting it. Emails will be sent when marking is complete.

Academic Standards

All text in your assignment should be paraphrased into your own words and referenced using the Harvard referencing style. Please refer to the Subject Outline for details about penalties for Academic Misconduct.

Late Penalties

Refer to the Subject Outline for details of the Late Penalty that may be applied to submitted work unless prior arrangements have been made with the subject coordinator.

Special Consideration

You may apply for special consideration (SC) due to unforeseen circumstances, either before or after the due date, at http://www.sau.uts.edu.au/assessment/consideration/online.html. The three basic reasons for SC are health, family, or work problems; "I haven't finished yet" is not a valid reason. You must provide documentary evidence to support your claim, such as a doctor's certificate, a statutory declaration, or a letter from your employer.

Note

The assignments may be checked through the Turnitin ® Plagiarism Prevention system, for identifying unoriginal material, copied (without reference to the source) from an electronic source on the Internet, electronic libraries, other assignments.