31250 Introduction to Data Analytics

32130 Fundamentals of Data Analytics

Assignment 3: Data Mining in Action

Marks	Out of 100, weighted to 40% of your final mark.
Submission format	A report in Adobe PDF (preferable) or MS Word Doc and an Excel spreadsheet.
Filename	ida_a3_group_xx.pdf or ida_a3_group_xx.doc where xx is your group number. ida_a3_group_xx.xls for the spreadsheet.
Report format	Around 30---40 pages with the information described below. Use 11 or 12 point Times or Arial fonts.

Scenario

This assignment is a group data analytics project with individual components within the group work. It follows on from the individual work you did in assignment 2.

You will form groups of 4 or 5 students. Your group will be acting as a data analytics consultant company. The dataset can be found on the site. Group members are expected to work together on common tasks, such as preprocessing (most of it you have already done in Assignment 2), and the development of the report. The individual part of the assignment is connected with the application of the data mining methods.

As in real world projects you need to explore different data mining techniques. Each member of the group should work with at least one mining method, data preprocessing and the corresponding tool(s). The results will be included in the individual parts of the report. At the end of the day, you as a group (company) decide which model is the best for the task.

You should follow the CRISP---DM methodology in the way you structure your study and the sections of the report. CRISP---DM is used by many analytic companies to ensure the quality of documenting the results and the accurate repetition of the steps of the investigation, if necessary.

Group

You need to form a group on your own. Once you have formed your group, email

- group name
- group members

Don't leave this to the last minute because we won't give you the link until you form your group! There is a strong correlation between low marks in this assignment and the length of time before forming a group.

Data sets

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

31250 / 32130 Assignment 3

Classification Task

Build a classifier that classifies the "salary" 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 follow the structure provided by CRISP---DM methodology and include the following sections:

- the **data mining task, inputs, output;**
- the **data preprocessing and transformations** you did (if any).

  Explain which attribute(s) you have processed, what kind of preprocessing/transformation technique(s) you have used, why you have used those techniques. If you selected a subset of attributes, explain the rationale behind the selection. In this section you can reuse as much as you need from the individual work that you did in Assignment 2. There is no need to copy the whole text from the report on Assignment 2 into the report for Assignment 3. You just need to refer to the particular section in the report for Assignment 2, from where you are taking the results.

- **classification techniques used**

  Each group member should apply at least one technique and report on it as well as the results obtained, their interpretation of those results and a summary of the capabilities and limitations of each model. For example, in what situations will your model will be useful, how it can be used, and for which tasks it may not be suitable.

- the **actual classifier (model) that the group has selected** --- the type, its performance, and errors on the training set and on the cross---validation set, and reasons for selection in comparison to the other techniques that you have tried.

- **summary** section --- At the end of the paper include a summary section that can be presented to the senior management staff. It should include

the model that you offer, what it will be good for, its limitations and recommendation what to do next!

- **reflection** section --- ***for those students enrolled in 32130 only!*** Each student enrolled in 32130 additionally must write a 1 page individual reflection on their learning in assignment 3 and recommendations for how they would approach the task differently (better) if they were to do it again.

You should clearly state in the report which group member has worked on which classifier.

Your report will likely be between 30---40 pages in length using an 11 or 12 point font, including title page and graphs, although this could be more pages if you explore lots of methods. On average each student will require between 24 and 36 hours to complete this assignment.

Assessment

This assignment is assessed as a work with individual and group work components.

The assessment criteria (for the individual mark of 50%) for students in 31250:

- quality of the analysis that each individual member did ------ 25%
- depth of understanding by each individual member and communication skills. ------ 25%

The assessment criteria (for the individual mark of 50%) for students in 32130:

- quality of the analysis that each individual member did ------ 15%
- depth of understanding by each individual member and communication skills. ------ 15%
- thoughtfulness of the individual reflection and recommendations ------ 20%

and for the (group component) (50%):

- quality of selected classifier ------ 10%
- clarity of the interpretation of the results ------ 20%
- quality of the summary section ------ 20%

Relationship to Objectives

This assignment addresses subject objectives 1---6.

Return of Assignments

We plan to return marked assignments including comments within 3 weeks of submission. 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.

Statement on Groupwork

For the group work part of this assignment students will be assessed as a team, which means each group member will normally receive the same mark for the group work section.

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.

31250 / 32130 Assignment 3