Ivor Buquetlowd
Managing Director
World of Bargains
Bargain House
Milton Keynes
MK5 6LP

Dear ITNPBD6 Student,

Let me introduce myself, I am Ivor Buquetlowd, the owner of a chain of over 100 shops in the UK and I would like you to help me run them more efficiently. The shops are all similar, but they make amounts of revenue varying from £1 million to nearly £5 million per year. I need a computer program that lets me enter the details of a shop and gives me a prediction of how much money that shop should make. This will help me choose new locations for shops and assess the performance of existing branches.

I have data describing the following aspects of each store:

- Town
- Store ID
- Manager name
- Staff numbers
- Floor Space and Window Space
- Car park (yes or no)
- Demographic score
- Location (Shopping Centre, High Street, Retail Park)
- 40 min, 30 min, 20 min and 10 min drive time population size
- Store age
- Clearance space in store
- Competition number (how many competing stores are near ours)
- Competition score (from how good the competing stores are)
- Revenue from last year

I have asked my IT department to make the data for my stores available to you. You can download it from

http://www.cs.stir.ac.uk/courses/ITNPBD6/assignment

Please remember that this data is commercially sensitive and you should not show it to anybody or discuss your findings with anybody except Kevin Swingler.

I would appreciate it if you could use the Weka software to build a predictor capable of estimating the revenue that a store should generate based on its values for the variables listed in the data. Remember, it costs a lot of money to collect each value, so the fewer variables you use, the better. The population data is particularly expensive to collect, so I would be grateful if you could tell me whether or not I should continue to use it. However, the system will only be of use to me if the average size of the error made on predictions is less than a million pounds.

I would particularly like to know whether regression logistic regression, a multi-layer perceptron or a decision tree could help, as my IT director has some experience of these techniques.

I will need one regression model, which predicts income and one classifier. For the classifier, I have classified each store by its performance as either poor, reasonable, good, or excellent. These should be the target classes.

Please send me a report detailing the following:

- An introduction to the task and the techniques you used. Please explain why data mining is suitable for this task;
- Which variables you chose to investigate, and why you chose them;
- A detailed description of the process you followed and the results that you found. This section should contain a lot of detail. Make sure you include the following:
    - What pre-processing you performed on the data and what problems, if any, you found and fixed;
    - A technical description of the techniques I've asked you to investigate – how they learn, how they represent the solution and how they make predictions. Please explain them carefully so that I can understand the concept and the maths behind it;
    - The process you used to arrive at the solution – how did you test it and how did you establish that it is the best possible solution from the data you had? Include and explain accuracy measures for each technique and each subset of variables that you tried;
- A full analysis of the results gained by the best model you could find – how accurate is it? Where does it make errors?
- A recommendation for which variables I should not pay to collect in future, or in other words, which variables are sufficient for this task? Describe how you came to this conclusion;
- A recommendation of which technique I should use and why.

The report is all I need you to send me. If I like the results, I can use Weka to reproduce them myself. Please make sure your report contains sufficient detail to allow me to do so.

Good luck and thanks for your hard work.

Best regards,

Ivor Buquetlowd
M.D World of Bargains

P.S. Please see http://www.cs.stir.ac.uk/courses/ITNPBD6/syllabus.php for notes on late submission and plagiarism.