

# Machine Learning in the Real World

Worth: 10% of total grade [30 marks]

Due Date: Monday September 24<sup>th</sup> 2007

Team Size: 3

## Synopsis:

You have now graduated (Yea!). And are the sole employees of ``Data Are Us'', a snazzy new startup company guaranteed to make you Millionaires, if you could just attract your first paying customers! You have just been giving your big break! A unit of a big internet company called ``Webs Are Us'' has given you a dataset on internet traffic. They want to know what you can find in their internet database.

The internet dataset will be provided. I will bring it to class on a CD. (Beware it is very large).

The goal is to find something that will convince this company to sign you up to do all its analyses and put you on your path to fame and fortune.

You can use any algorithms you can find on the net.

Don't wait until

the last moment. The hardest part of this assignment will be preparing the dataset. You will \*NOT\* be able to run on the whole dataset, you will have to throw parts out. Can you run on part of the dataset to bootstrap you to find the most important parts of the dataset? What question are you going to try to answer?

This assignment has been left intentionally vague. All grades will be based on the writeup! Remember you are trying to convince this company to give you their money and not your rival company ``Data-dredging Are Us''. So convince me! This should look like a business pitch including all your data findings.

Make sure you explain what algorithms you used, our company is old and conservative and we don't like black magic.

Also if you don't tell me you did something, I will assume you didn't! So make sure you tell me all the wonderful things you did in term of methodology and exploration experiments.

Make sure all group members' names are included in the writeup and each member must send me email saying what percentage of the work they felt each member did. No grades will be given if I do not receive email from each member!

Please remember this assignment is meant to be fun, fun, fun! But it is only worth 10% of your grade, don't kill yourself over it. Are we having fun yet?

Marking is based on the following material:

`\begin{enumerate}`

`\item A marketing report describing your data analysis and why this company should give you its lucrative contract.\end{enumerate}`