Actuaries Institute Data Analytics Competition 2015

What is the Competition?

The Actuaries Institute Data Analytics Working Group has set up a competition as an introduction to studying data analytics. The competition is fairly simple and is based entirely on publicly available information from the Australian Bureau of Statistics.

Participation is open to all Members. Those who are attending the Data Analytics Seminar being held on 19 October 2015 will be invited to try out some of their new skills by taking part in the competition. Closing date for the competition is Monday 30 November 2015.

How do I join?

The competition is hosted by Kaggle, which is a company which hosts Data Analytics competitions. The competition follows the same structure as all Kaggle competitions. This particular competition is not open to the public, but is by invitation only. The invitation is distributed as a web link, and anyone with the web link can join the competition. It is not intended that we get entrants from outside the actuarial community, but we are not preventing anyone from outside the community participating.

In order to participate in the competition, you need to:

- Click on the link and log in with your existing Kaggle account or set up a new Kaggle account. Kaggle verify your account by SMS, so have your mobile phone ready.

- Click on the link again and read and agree to the competition rules. You can participate in the competition as an individual or as a team. There is no limit on team size.

- Download the data (which is in the form of csv files) and use the training data, together with whatever programs and algorithms you like to derive you own test data predictions. One of the data files is a sample results file which shows you the format of the file you should upload.

- Upload you predictions and these will be automatically marked to give you a position on the leader board. You can then improve your model to get a higher position, subject to the constraint of only two uploads per day.

What is being predicted?

The test data consists of a randomly chosen 60% of the census areas (SA2) in Australia from the 2011 census, with data relating to the population age and demographic structure as well as some education income and housing information. It also includes the number of deaths in the census area in 2011.

The test data contains 40% of the areas but without the deaths column. You need to build a model that predicts the deaths for the 40% of census areas in the test data.

Although the task is to predict deaths, the prediction column could have been anything. The task is not specific to life insurance and the challenge is to consider what other factors in the training data, or derivable from the data are likely to be relevant to the variable being predicted. In data science this is called “feature creation”.

How will it be marked?

The test set has been further split into two subsets. The first of these subsets is scored against your submissions by calculating the root mean squared error in the absolute number of deaths. This determines your position on the leader board. When the competition deadline is reached the other subset is used to determine your final ranking on the leader board. So if you overfit your model to the exposed portion of the test set this won’t help the final rankings.

How will the prize winners be determined?

The teams at the top of the leader board will be asked to submit a paper on how they did it. This is not an arduous thing, but could be the basis for a magazine article. Provided the winning team contains at least one member of the Actuaries Institute, and they didn’t cheat in deriving their model, then they will win the prize. The first prize is a $200 book voucher, second prize a $100 book voucher and third prize a $50 book voucher.

What constitutes cheating?

The data underlying the competition is all in the public domain. So although the census areas are not identified in the data it is possible to work out which census areas are in the test set and create an algorithm that essentially get a perfect result. The paper describing the process followed to arrive at the best entry should allow the judging panel to eliminate any entries that have cheated.

Using other data sources, though, is acceptable, provided these are identified in the paper and would generally be available to someone working in the field. For example the Australian Life Table 2010-2012 could be incorporated in a winning model, as could the geographic locations of cities or hospital emergency departments.

Where do I start?

In essence this is a simple competition. You can get a reasonable answer using only Excel. There is lots of help on Kaggle (and elsewhere) including scripts written by other Kagglers.

Ready to start?

https://kaggle.com/join/aiDataAnalysis2015