**[](http://www.google.com/url?sa=i&rct=j&q=&esrc=s&frm=1&source=images&cd=&cad=rja&docid=faOoPIJKk-fFMM&tbnid=9V61LGszONYSlM:&ved=0CAUQjRw&url=http://openclipart.org/detail/61759/kangaroo-sign-by-djcowan&ei=zY1lUvaaLKWIyAHgpYHIDA&bvm=bv.54934254,d.aWc&psig=AFQjCNGRj6SURfbCux9miXfA_jCBwZrpSw&ust=1382473517407550)UCONN vs. UMASS Analytics Case Competition**

**Kangaroo Auto Insurance Company Modeling Problem**

**Oct 2016**

1. **Business problem**

You work for Kangaroo Auto Insurance Company, an Australian company.  Your business partner, who is not familiar with statistics at all, and would like you to create a rating plan based on the historical auto claim data.  Your business partner is concerned about segmentation as well as competitiveness, as there are several other competitors in the market.

For this case competition, your group’s task is to provide a method for predicting the claim cost for each policy, and to convince your business partner that your predictions will work well.

1. **Data Description**

Modeling data will be sent out via email. The Kangaroo data set is based on one-year vehicle insurance policies from 2004 to 2005. There are 67856 policies, of which 4624 (6.8%) had at least one claim. The data will be split to three parts, training data, validation data and hold out data. In hold out data, claimcst0, clm and numclaims will be set to NA. You can build your model on training data and test the model on validation data. In the end, you can use your best model to score the hold out data. We will evaluate your model based on your hold out data prediction.

Below is the variable information in the data.

ID: policy identity

Veh\_value: market value of the vehicle

Exposure: The basic unit underlying an insurance premium

Clm: Occurrence of claim (0=no, 1=yes)

Numclaims: The number of claims

Claimcst0: Claim amount (our response variable)

Veh\_body: Type of vehicles

Veh\_age: Age of vehicles (1=youngest, 4=oldest)

Gender: Gender of driver

Area: Driving area of residence

Agecat: Driver’s age category from young (1) to old (6)

1. **Modeling**

Each group may have at most 5 people and will:

1. Work together within group on data analysis, **but not between groups**
2. Build a model to predict the claim cost (claimcst0) and submit the predicted cost for validation data. You can build the model with any software
3. Submit the prediction of hold out data as a csv file. Format of submission will be provided with data
4. Prepare presentation slides to summary the prediction and analysis results. You do not need to explain the problem, just summarize what you did and what you found
5. Work with your group members and ask for clarification from the email at the end of this introduction
6. **Model Evaluation**

The model will be evaluated by gini index. We will calculate your score once you submit your result.

1. **Timeline** 
   1. Kick-off: The week of Oct 24
   2. Competition will start right after kick-off meeting
   3. First Q&A session: The week of Oct 31
   4. First submission: Mon. Nov 7
   5. Second Q&A session: The week of Nov 7(Optional)
   6. Final submission: Mon. Nov 24
   7. Group presentation & UCONN winners announced: TBD
   8. Winners’ Job Shadow day at Travelers: TBD
2. **Contacts:**
3. Daren Mauro: [DMAURO@travelers.com](mailto:DMAURO@travelers.com)
4. Yi Cao: YCAO@travelers.com
5. Tiran Chen: TCHEN3@travelers.com
6. Matthew Lamoureux: MLAMOUR2@travelers.com
7. **Benchmark model**

The benchmark model is a gbm model. Sample code will be provided. You can get the benchmark score by running the sample code.

1. **Presentation instructions**

Submit your code with documentation along with a report answering the following questions:

* 1. What methods did you consider (you don’t have to actually try all of these methods; just ones that you think would work for this problem)?
  2. What method did you choose in the end, and why?
  3. How did you do you variable selection?
  4. What variables help explain pure premium (explain to a non-statistician; please include this in your presentation for your business partner)?
  5. What other variables not in the data set do you think might be useful?

The teams scoring better than the benchmark will move to the second stage, the live presentation. Each qualifying team will give a 7min presentation on the above questions and 3min Q&A session. Top teams will be eligible for consideration as the overall UCONN winner.

The winning UCONN team will join the winning UMASS teams for a job shadow day at the Travelers Hartford campus, and make final presentation to a panel who will determine the ultimate winner.