## Homework 2

In this assignment you will build at least three models to predict whether an automobile has been in an accident using the CLAIM\_FLAG column. 1 means the automobile was in a crash. A zero means the vehicle was not in a crash. All predictor variables used must be significant for each of your top three models.

Use the **carInsurance.csv** file for this assignment. The data file can be found in the **Datasets** folder.

For this project, points will be awarded for a professional approach taken to develop and report about the predictive models.

**DELIVERABLES:**

* Your report in a word document (4 to 8 pages maximum).
* Your source file which builds your models.

Your report will have four sections and each are described below:

## Report: Exploratory Data Analysis (10 %)

Describe the variable highlights in a way so your not-so-technical manager can understand it. Explain why particular variables are of interest. What kind of distributions and correlations do you see? Be creative with how you summarize the data. Use good judgement here since time and space are limited. However, do not leave this section empty or your manager will think you that you did not do your job.

## Report: Data Preparation (10 %*)*

* Explain what you imputed and how you imputed. Explain why you chose your selected method for imputing each variable. (Remember to create flag variables for imputed values)
* What kind of transformations or scaling did you use and why?
* What kind of binning did you try?

## Report: Building the Models 20%

* Explain how you selected the model variables.
* Ensure that all variables are significant and provide a summary of the model variable significance (using the chi test) and coefficients for each of the top three models including your selected model.
* Discuss the models and provide as much interpretation as you can within reason. Do they make sense? For example, if a person had their license revoked before it would be expected they would be a candidate for an accident with their current vehicle.

## Report: Model Selection 30%

* For all models, show a summary of accuracy, precision, recall, f1 scores along with averages and standard deviations of these scores for all folds.
* Show typical confusion matrices for your top three models.
* Explain which model you selected and why.
* Draw a cumulative gains chart and ROC curve for your selected model. Interpret these results.
* Interpret the model coefficients of your selected model as best you can – explain why some variable coefficients have a higher weighting.
* Caution your boss about uncertainties around your selected model.

## Code File: 30%

* Include your code in one separate file. It must run without error.
* Cross fold validation is required. Please see the **updated notes and examples** in the day11 folder.
* Build at least one model that uses automated dummy variable, binning and recursive feature elimination. Keep this code in your application.
* Use k-fold validation for all models.