Regression Prediction Competition

Due date: Tuesday, December 10 at 6:30pm

(During our normally scheduled final exam)

**Teaching Goals**

* Understand relationships inherent in an archival data set
* Create an automated interest rate assessment tool
* Identify variables most commonly used to assess an interest rate
* Connect the results of text analysis to the

**Deliverables**

* R data frame or CSV file with predicted values for interest rates on the prediction data
* R script showing your final predictive regression equation and transformations / data manipulation required to achieve the results
* A summary of the best fitting *descriptive* regression model possible
* A summary of the process you used to reach your final model. This should also include:
  + A description of the 3 variables that are the most significant predictors of the DV
  + Justification for any data manipulation, cleaning or transformation activities you used.
* 2-3 minute informal explanation in-class discussing your approach and headaches you encountered (no materials required.)

**Summary of the Data**

Data used for this project is provided by the professor. It includes information on 31,543 colleges, universities and other institutions of post-secondary education. The observational unit of the data is recorded at the institution level. The training data includes 25,203 observations that have a variable recorded describing the average annual cost of attendance for each observation. The prediction data contains the remaining 6,340 observations that are missing the values for the average cost of attendance. All other variables are shared between the two datasets and you can assume the two samples are homogeneous in all other ways.

**Objective**

The goal of this project is build a regression model for predicting the average annual cost of attendance for each observation in the prediction dataset. I would also like to see a summary of your approach. Verbose detail is not required, but I would like to know more than “we cleaned the data, then built a regression model.”

**Questions you should consider answering in the summary:**

* How did you clean the data?
* What analytical methods did you try?
* How much of the variability in average annual cost were you able to predict?
* What frustrations did you encounter?

**Joel’s recommended strategies for improving your predictions**

* Cleaning your data from outliers and extremely rare nominal categories may drastically improve fit.   
  You may also consider using the prediction set as your guide.
* Be extremely careful of how you manage missing values. A solution with many missing predictions will rarely be the best overall fit.
* Consider creating dummy variable intervals from continuous variables and applying transformations to continuous variables.
* The best fitting strategies will require a careful consideration of the regression assumptions, particularly the linearity and multicollinearity assumptions.

**Grading Rubric**

* **Accuracy of the prediction (50%)**
  + All submissions will be compared using RMSE. Missing values will be replaced with a zero as a penalty to estimate RMSE.
  + The model with the lowest RMSE ‘wins’ and will be given a perfect score (A) for this portion of the assignment as well as some prize donated by our corporate partners.
  + All other submissions will receive a scaled score. If all groups are very close, everyone will receive an A for this portion. If all groups are above my pre-determined standard of excellence, everyone receives an A for this portion.
* **Quality of the Analysis: (30%)**
  + Process specified is correct and free of egregious mistakes
  + Most significant variables are properly identified.
  + Fit and methodological approach of the descriptive regression is robust
* **Equation / routine is submitted and matches the output (20%)**
  + Whatever final prediction equation should match the predicted values in the test set submissions.
* **Team Member evaluation (optional): Scaled portion of the final group score**