DNSC 6279
Assignment 2

In this assignment you will prepare a data set and build a logistic regression model on the Kaggle Allstate Claim Prediction Challenge data. Your finished model could be used for predicting what types of Allstate car insurance policies will have future insurance claims. **Please complete with your homework group.**

Download and use the original contest data: https://www.kaggle.com/c/ClaimPredictionChallenge/data. You don't need the test data. **YOU MAY NOT UPLOAD THE DATA TO THE SAS CLOUD ENVIRONMENT.**

To receive full credit on this assignment you must:

1. **(1 pt.)** Convert the numeric claim_amount target to a new binary variable that will be used as a target for logistic regression. Any claim_amount > 0 should be considered a claim event.
2. **(1 pt.)** Over- or under-sample the data to create a more balanced target distribution for training your logistic regression model.
3. **(1 pt.)** Partition your data appropriately. Create a 30% test partition along with other partitions you may need.
4. **(1 pt.)** Impute any missing values.
5. **(2 pts.)** Create numeric features for the blind_make, blind_model, and blind_submodel high cardinality categorical variables and use them in your model.
6. **(1 pt.)** Train a logistic regression model with variable selection based on minimizing validation error or maximizing another suitable validation accuracy measure.
7. **(1 pt.)** Report your test AUC and give its exact interpretation.
8. **(2 pts.)** Explain the 3 most important variables in your model in terms of the odds that a policy will have a claim associated with it.

You may takes steps beyond those listed above. The group with the highest test AUC in the class, **that also follows correct practices**, will receive 3 extra credit points. Similarly, the group with the second highest score will receive 2 extra credit points. The group with the third highest score will receive 1 extra point. If multiple groups have the same test AUC, extra credit will be awarded at the discretion of the instructor.

Create a document that briefly describes and justifies all 8 steps, or more, in your modeling process. This document must contain screenshots from software that indicate the number of observations in your test set and your test AUC. Place this document in a folder with any code or EM diagrams. Zip this folder and turn it in to blackboard.