Assignment 2

Please complete in groups of 4 or less.

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. You have two options for acquiring the data:

- In the SAS on Demand for Academics environment, you may copy the SAS data set from <your home directory>/my_courses/jphall0/train_subset.sas7bdat into your home directory and use it in Enterprise Miner in the cloud environment.

- For extra credit you can 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 ORIGINAL CONTEST DATA TO THE SAS CLOUD ENVIRONMENT. *****

To receive full credit on this assignment you must:

- 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. (This task has been completed for you in the train_subset data set.)
- Over- or under-sample the data to create a more balanced target distribution for training your logistic regression model. (This task has been completed for you in the train_subset data set.)
- Partition your data appropriately. Create a 30% test partition along with other partitions you may need.
- Impute any missing values.
- Train a logistic regression model with variable selection based on minimizing validation error.
- Report your test AUC.
- Explain the most important variables in your model in terms of the odds that a policy will have a claim associated with it.

To receive extra credit you may:

- Create numeric features for the blind_make, blind_model, and blind_submodel high cardinality categorical variables and use them in your model.
- Turn in the model with the highest test AUC in the class – that also follows correct practices.

Create a document that briefly describes and justifies the steps 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. **PLEASE INDICATE ALL GROUP MEMBERS NAMES IN THIS DOCUMENT**. Place this document in a folder with any code or EM diagrams. Zip this folder and turn it in to blackboard.