2017 USC Marshall Statistics PhD Screening Exam
Part II: Take Home (Due at 6PM, June 6, 2017)

The `spam` data set is built in R and can be obtained by first installing the package `kernlab` and then using the R commands

```
library(kernlab)
data(spam)
```

This data set consists of 4601 observations and 58 variables. The first 48 variables contain the frequency of the variable name (e.g., business) in the e-mail. If the variable name starts with num (e.g., num650) then it indicates the frequency of the corresponding number (e.g., 650). The variables 49-54 indicate the frequency of the characters ;, (, [, !, $, and #. The variables 55-57 contain the average, longest and total run-length of capital letters. Variable 58 indicates the type of the mail and is either "nonspam" or "spam", i.e. unsolicited commercial e-mail. Our purpose is to build a classification model which can accurately predicts spam emails yet interpretable. To this end, separate the data into training and test sets by using the first 2800 observations to build the model and the remaining observations to evaluate the model. The following questions only provide rough guidelines. Most of these questions are open-ended ones. Feel free to use additional model building ideas if you feel they can improve the classification power of your model.

1. Build a logistic regression model using the training data with all 58 variables. Calculate the classification error using the test data. How do you interpret the model? Which variables are most important ones in predicting spam emails? List the top 5 important variables suggested by the model and interpret.

2. Now add the variable selection component into the logistic regression model above and calculate the test error again. How do you compare this model with the model in question 1)? List the top 5 important variables. How do you interpret this new model?

3. Construct pairwise interactions using the original 58 variables. Build a new logistic regression model including both the original 58 variables and all two-way interactions. Did you encounter any problems in building such

a model? How can you fix the problem? Compare your new model with the previous two and interpret.

4. Again add the variable selection component into the logistic regression model in question 3). Compare this model with the previous models and interpret.

5. Can you come up with any additional ideas on how to improve the classification power of your model? This is an open-ended question and you can try anything you can think of to improve your existing model. Or you can even try a completely new model if you believe it would do a better job. But in any case, provide interpretations.