STA 6704 – Data Preparation Final Project
(Fall 2014)

The data set PHY_TRAIN used in this project came from 2004 KDD CUP competition. It is a perfect data set for building predictive models using regression or neural network techniques. Since data preparation requirement for building both types of models is high, it is a good data set to end data preparation class.

To save your time, a file with useful information is included for your reference. In addition, all variables are listed in the appendix as well and I offer a suggestion procedure for you to complete a predictive model using regression technique. The suggestion procedure is as follows:

- Step 1: Understand the data by studying the reference file "Variables" and using distribution macro discussed in Lecture 03 and visualization tools discussed in Lecture 02.
- Step 2: Since there is not any categorical variables, categorical variables' smoothing or clustering are not necessary
- Step 3: Create missing value indicators and MVP variable using SAS Code Node
- Step 4: MVP variables need to be clustered or smoothed using techniques discussed in Lecture 04
- Step 5: Missing value imputation is necessary since logistic regression are implemented using complete cases only. Idea missing value imputation methods for this study including tree imputation, regression imputation or cluster imputation.
- Step 6: Transformation are necessary and possible transformation methods are discussed in Lecture 07
- Step 7: Variable selection might not as critical if you use regression to fit the model
- Step 8: Search over the model space.
    - ■ Types of Model structure (regression models)
    - ■ Search for idea model need to consider
        - ◆ Second order terms
        - ◆ Interaction terms
        - ◆ Model selection techniques: Forward, Backward, Stepwise, LASSO, or Ridge
        - ◆ Model Assessment: Mean Square Errors, Misclassification Rate, and others
- Step 9: If you decide to fit more than one models, you need to use Model Comparisons Node to select the best model
- Step 10: Prepare your presentation to communicate your findings to others.

The goal of this project is to give you a chance to review all topics discussed in the class and to follow possible sequence to search the best predictive model for a given data set.

STA 6704 – Data Preparation Final Project
(Fall 2014)

You can do anything if you think it is relevant and interesting in developing effective predictive models for this data set! Just think of this as an open project. You can ignore all of the above and take your own way. Only requirement is using the data set PHY_Train and to find the best model to predict the target variable "Target".

Note: You can build a "bad" logistic regression model with just 10 variables to achieve 70% accuracy and build a "bad" neural network model with 10 variables to achieve 72% accuracy easily. Therefore, you will not get high grade if your model is not better than these bad models mentioned.

Please turn in your final report (12 pages (appendix not included)) and make a final presentation (15 minutes) in class on the final date (4/23/2014) and feel free to ask me any questions. Your final report should include at least the following section:

(1) Executive Summary (1 Page)
(2) Data Preparation:
    a. Exploration
    b. Missing Values Indicators and MVP
    c. Missing Value Imputation
    d. Transformations
    e. Outliers
(3) Model Building:
(4) Model Performance
    a. Model performance evaluation (ROC curve, c statistics, and Validation sum squares)
    b. Variables importance
    c. Variable impact to the target variable
(5) Conclusions
(6) Appendix: All figures, SAS Code, and tables should be in Appendix

In your presentation, you need to try to explain your findings to someone who does not have any knowledge on data mining and statistics but pay the bill.

Grading standard:

(1) Presentation: 50 Points
(2) Report: 100 Points
(3) Model performance: 50 Points (Based on the relative rank with your classmate)

Appendix Variables in the data set

Alphabetic List of Variables and Attributes			
#	**Variable**	**Type**	**Len**
1	exampleid	Num	8
3	feat1	Num	8
4	feat2	Num	8
5	feat3	Num	8
6	feat4	Num	8
7	feat5	Num	8
8	feat6	Num	8
9	feat7	Num	8
10	feat8	Num	8
11	feat9	Num	8
12	feat10	Num	8
13	feat11	Num	8
14	feat12	Num	8
15	feat13	Num	8
16	feat14	Num	8
17	feat15	Num	8
18	feat16	Num	8

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
19	feat17	Num	8
20	feat18	Num	8
21	feat19	Num	8
22	feat20	Num	8
23	feat21	Num	8
24	feat22	Num	8
25	feat23	Num	8
26	feat24	Num	8
27	feat25	Num	8
28	feat26	Num	8
29	feat27	Num	8
30	feat28	Num	8
31	feat29	Num	8
32	feat30	Num	8
33	feat31	Num	8
34	feat32	Num	8
35	feat33	Num	8
36	feat34	Num	8
37	feat35	Num	8

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
38	feat36	Num	8
39	feat37	Num	8
40	feat38	Num	8
41	feat39	Num	8
42	feat40	Num	8
43	feat41	Num	8
44	feat42	Num	8
45	feat43	Num	8
46	feat44	Num	8
47	feat45	Num	8
48	feat46	Num	8
49	feat47	Num	8
50	feat48	Num	8
51	feat49	Num	8
52	feat50	Num	8
53	feat51	Num	8
54	feat52	Num	8
55	feat53	Num	8
56	feat54	Num	8

Alphabetic List of Variables and Attributes			
#	**Variable**	**Type**	**Len**
57	feat55	Num	8
58	feat56	Num	8
59	feat57	Num	8
60	feat58	Num	8
61	feat59	Num	8
62	feat60	Num	8
63	feat61	Num	8
64	feat62	Num	8
65	feat63	Num	8
66	feat64	Num	8
67	feat65	Num	8
68	feat66	Num	8
69	feat67	Num	8
70	feat68	Num	8
71	feat69	Num	8
72	feat70	Num	8
73	feat71	Num	8
74	feat72	Num	8
75	feat73	Num	8

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
76	feat74	Num	8
77	feat75	Num	8
78	feat76	Num	8
79	feat77	Num	8
80	feat78	Num	8
2	target	Num	8