Adult Data Set Statistical Analysis

**Introduction:**

The following analysis is based on the Adult dataset provided by X.

**Data Description:**

X.

Additional details

NOTE 1: ALL ANALYSIS MUST BE DONE IN SAS OR R and all code must be placed in the appendix of your report. I’m okay with data cleaning steps and EDA being provided using other tools such as Python.

NOTE 2: Do not forget about organization among your group. Divide and conquer is always great, but there is “one report to rule them all” so make sure that it flows as you are stitching things together. I don’t see this as a big problem as project 1 was pretty good across the board.

**Required Information and SAMPLE FORMAT**

Required deliverables in the complete report. The format of your paper (headers, sections, etc) is flexible although should contain the following information.

PAGE LIMIT: I do not necesarrily require a page limit, but you should definitely be shooting for know more than 7 pages written. It of course can blow up quite larger than that due to graphics and tables, but good projects are clear, concise, to the point. You do not need to show output for every model you considered. (You may put supporting plots/charts/tables etc. in the appendix if you want, just make sure you label and reference them appropriately.)

Introduction **Required**

Data Description **Required**

Exploratory Analysis **Required**

Addressing Objective 1:

Restatement of Problem and the overall approach to solve it **Required**

Model Selection **Required**

Type of Selection

**Any or all**: LASSO, RIDGE, ELASTIC NET,

Stepwise, Forward, Backward

Manual / Intuition

Checking Assumptions **Required**

**Optional** Lack of fit test

Influential point analysis (Cook’s D and Leverage)

**Optional** Residual Plots

Parameter Interpretation

Interpretation **Required**

Confidence Intervals **Required**

Final conclusions from the analyses of Objective 1 **Required**

Addressing Objective 2

Make sure it is clear how many models were created to compete against the one in Objective 1. Make note of any tuning parameters that were used and how you came up with them (knn and random forest logistics, CV for penalty of lasso, etc) **Required**

Main Analysis Content **Required**

Overall report of the error metrics on a test set or CV run as well as ROC curve information. Also if the two best models have error rates of .05 and .045, can we really say that one model is outperforming the other?

Conclusion/Discussion **Required**

The conclusion should reprise the questions and conclusions of objective 2 with recommendations of the final model, what could be done to help analysis and model building in the future, and any insight as to why one method outshined all the rest if that is indeed the case. If they all are similar why did you go with your final model?

Appendix **Required**

Well commented SAS/R Code **Required**

Graphics and summary tables (Can be placed in the appendix or in the written report itself.)