MSDS 6372 Project 2 Description

There are two main objectives for Project 2. Since each group will be using their own data set, there will be a little flexibility in what needs to be delivered. Below is a summary of what is absolutely necessary as part of your report.

Objective 1: Display the ability to perform EDA and build a logisitc regression model.

- Perform your logistic regression analysis and provide interpretation of the regression coefficients including hypothesis testing, and confidence intervals. For simplicity sake, you do not need to include interactions with this model. Comment on the practical vs statistical significance of the deemed important factors..

Logistical Considerations.

- Just like last time, this does not have to be extremely fancy in terms of the model building approach, let EDA, feature selection, and overall intuition guide you.

Objective 2: With a simple logistic regression model as a baseline, perform additional competing models to improve on prediction performance metrics. Which metrics are up to you and your given data set.

- Record the predictive performance metrics from your simple, highly interpretable model from Objective 1.
-
- You must include one additional model which is also a more complicated logistic regression model than in Objective 1. By complicated, I do not mean that you include more predictors (that will be somewhat sorted out in Objective 1), but rather model complexity through interaction terms, new variables created by the group, or transformations.

- Create another competing model using just the continuous predictors and use LDA or QDA.

- Use a nonparameteric model approach as a competing model. Random forest for predictors that are both categorical and continuous or a k-nearest neighbors approach if just working with continuous predictors.

- Provide a summary table of the performance across the competing methods. Summarize the overall findings. A really great report will also give insight as to why the "best" model won out. This is where a thorough EDA will always help.

Logistical Considerations.

- Don't forget PCA can be helpful in various ways throughout your analysis as well as other unsupervised tools such as heatmaps and cluster analysis from Unit 13.

- I think a good course of action is to tackle Objective 1 in SAS. The selection tools are really straight forward to run and the output is a little bit easier to grab. For objective 2, its better to go with R for this reason…..to ensure performance metrics are comparable make sure that the models are run on the exact same training and test sets (or through a CV approach).

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.

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.

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**
    - 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) **Required**

Main Analysis Content **Required**

Overall report of the error metrics on a test set or CV run. Also if the two best models have error rates of .05 and .045, can we really say that one model is outperforming the other? What other tools that we learned in the second half of this class that could help us get at that?

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.)