MSDS 6372 Project 2 Description

For this project I’m going to let you guys decide what data set you want to use. You may choose your own data set or select one of the three data sets below.

1. <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing> Predicting if a customer will subscribe to a term deposit.
2. <https://archive.ics.uci.edu/ml/datasets/Adult> Predicting if someone makes over 50k
3. R package aplore3, use the glow\_bonemed data set. Assessing risk factors and predicting if a woman with osteoperosis will have a bone fracture within the first year after joining the study. ?glow\_bonemed for data description of variables.

For this project, again you will only need to have a train/test split. Depending on the data set you choose, the groups may have to discuss a strategy to effectively split the data set to deal with issues like unbalanced response levels (too many yes not enough no) and/or small sample sizes could make splitting the data difficult. I’m open for discussion when if the groups wants to bounce ideas off of me.

Similar to Project 1, 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 logistic 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.
* If you feel like interactions are absolutely necessary to capture what is going on, then contact me so we can discuss an overall strategy of how to provide interpretations.
* When it makes sense, utilize PCA and clustering during EDA.

***Objective 2: With a simple logistic regression model as a baseline, perform additional competing models to improve on prediction performance metrics. Which metrics to compare models and evaluate performance 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 logistic regression 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, transformations or additions through polynomials.
* Create another competing model using just the continuous predictors and use LDA or QDA.
* (Optional/Bonus) Use a nonparameteric model approach as a competing model. Random forest or decision tree 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. Its not necessarily expected, but if your EDA is light, think about using these tools to get practice even if its not necessarily practical for your analysis.
* For feature selection for objective one, make sure you use lasso, but create your final model using a glm call so that you can obtain all the necessary statistical information and tests. For objective two, I expect groups to provide ROC curves, discuss selection of an appropriate prediction cutoff, and reporting confusion matrix results like overall accuracy, sensitivity, and specificity (all from the test set).

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.

**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. Tie things back to what your eda shows. **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? Comment on these things if they come up. **Required Metrics from Test set or CV: Accuracy, Sensitivity, Specificity,ROC/AUC**

I also highly recommend a plot comparing ROC curves across the methods tried.

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.)

Classifying run versus pass- 1st, 2nd, 3rd downs

* Yards needed for first down
* Yards until end zone

Time in the game- 16 units

Current Score

Strength of the defense

Weather- continuous variable- Fahrenheit

* Stadium / Home versus Away

Field- levels or continuous yards

* Maybe run both and determine which produces a better ASE

2017 to 2021 has 243131 rows

* Decided on 2019 to 2021 season (be about ½ the # of rows before further filtering and selection

How do you make the call to go with levels or continuous? (yards, for example)

Data selection/collection

* Filtering, transforming

EDA

* Narrow down predictors (PCA)

Complete EDA by end of this week (3/20)

* Following step: covering logistic regression

Due date is April 10th