MSDS 6372 Project 2 Description

For this project I’m going to let you guys decide what data set you want to use. If you don’t want to worry about finding a data set, or are concerned that you might pick a data set that is too ambitious, then you can use one of the data sets below as I know they are doable.

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.
4. There are many binary classification data sets available on Kaggle as well as the UCI data base. Feel free to talk to Dr. Turner if you feel that one of those would be good and want to run it past him.

For this project, you will need to have a train/validation split. 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. For this project, each group will be submitting a presentation recording (via zoom, youtube, etc). Groups should still submit their final codes along with an mp4 or weblink of their presentation.

The recorded presentations should be no more than 30 mins long and address the following objectives:

***Objective 1: Display the ability to perform EDA and build a logistic regression model for interpretation purposes.***

* Perform your multiple 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 project 1, this does not have to be extremely fancy in terms of the model building approach, let EDA, feature selection, and/or overall intuition guide you. Keep in mind previous project feedback and note the difference between a model that is interpretable versus a model that is complex. (A model with a lot of predictors can still be interpretable)
* Interactions models shouldn’t be used here as you should display your ability to interpret the regression coefficients. Effects plots may be used in addition too but not at the exclusions of coefficient interpretation.

***Objective 2: With a simple logistic regression model as a baseline, perform additional competing models to improve on prediction performance metrics.***

* In this section you will build 3 additional classification models to compare to your model in objective 1. The goal of this objective is to build a model where prediction performance is prioritized. One model should be an attempt at a complex logistic regression model including interaction terms or polynomial terms. One model should be an LDA or QDA. And the final model should be a nonparametric model such as knn, random forest, classification tree, etc.
* There should be a quick discussion on the reasoning for your complexity you tried/used in the complex logistic model. Perhaps another slide on a deeper dive of EDA or discussion on logistical points based on your own knowledge of the variables.
* For each model, 6 metrics should be used and reported on the validation set. The Sensitivity, Specificity, Prevalence, PPV, NPV, and AUROC. Reporting the logloss metric would also be helpful to include if used during training. It should be well communicated on what metrics you feel are more important for comparing your models and describing their performance. Summarize the overall findings and your recommendations for what model you should go with for making future predictions. Make sure to communicate what threshold you are using that derives many of the metrics. Excellent projects will effectively communicate why some error metrics may be more important to optimize given the practical use cases of the data set used.
* Feature selection should be implemented in the logistic model here unless it was done in objective 1 already. You can always use it in both.

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 and 14. 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.
* If using feature selection for objective one, if you are using lasso/glmnet, create your final model using a glm call so that you can obtain all the necessary statistical information and tests. For objective two, do not forget ROC curves can provide comparison of the models in addition to the error metric table.

Additional details

NOTE 1: ALL ANALYSIS MUST BE DONE IN 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 as long as it is well communicated.

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**

Presentation slides should be submitted along with an mp4 recording of the presentation. All group members should participate in the presentation. The final codes used to perform the analysis should also be submitted. The report should take 30 mins or less. The general format of the presentation should follow the guidelines below:

Introduction and Objective Summary

Data Description / Processing Summary

Exploratory Data Analysis

Objective 1: Must include a high level explanation of the model fitting approach.

Which variables were included/exclude along with how/why?

Feature Selection Summary if applicable

The final model should be clearly defined.

Summary table of coefficients via Odds Ratio

Example of at least two regression coefficient interpretations (preferably one categorical and one numeric)

Objective 2: Summary of approach to include complexity in the logistic model

Discussion of what metrics really make the most sense to evaluate your models

Model comparisons

ROC curves

Insights as to why one tool worked better than the other. Or perhaps why they are all equally good…or bad?

Conclusion: Quick summary of the findings in objective one

Quick summary of the findings and recommendations in objective 2

Additional discussion points: Scope of inference? What would you do if given more time? Why one method performs better than the other for objective 2? Why is one or two metrics more important/informative compared to the others?