MSDS 6372 Project 1 Description

There are two main objectives for Project 1. Below is a summary of what is absolutely necessary as part of your report.

***Objective 1: Display the ability to build regression models using the skills and discussions from Unit 1 and 2 with the purpose of identifying key relationships, interpreting those relationships, and making good predictions.***

* Build a model with the main goal to identify key relationships and is highly interpretable. Perform your regression analysis and report the predictive ability of your model using a test set or some other means through CV. Be sure to provide metrics if you compare multiple models.
* Provide interpretation of the regression coefficients in the model including hypothesis testing, interpretation of regression coefficients, and confidence intervals. It’s also good to mention the Practical vs Statistical significance of the predictors.
* Fit a second model with the goal to produce the best predictions possible. Interpretation is no longer important so you can get as complicated as you like. Use feature selection techniques to avoid under/over fitting. Compare this model with your first, highly interpretable model, and comment on if this second model brings additional value or the first model is preferred.

Practical Consideration:

EDA, EDA, EDA! It helps you on so many fronts so use it to your advantage. When writing a concise report, you do not have to literally step out every single step of your model building process. I know you guys are going to being iterating on things many many times. That does not all have to be there or you can summarize in a paragraph.

What is key in the report is that you develop a “story” of your analysis. Keep in mind that when you are finished with your analysis. You know how it is going to end (what the final models look like). You can use this to your advantage when selecting what parts of the EDA and additional information. For example, if you know that predictor X7 is in your final model and it is one of the stronger relationships, that is probably a good one to show and discuss in the EDA part. You would show the reader, “Hey look at these interesting trends”, “Hey look at these that are not”, etc. When you report your final model and you are bringing back up the predictors discussed in EDA, it helps build the confidence of the reader in what you are doing is making sense.

You are required to produce at least two models to compare. In addition to these two, if your team has other ideas and would like to compare additional models based on manual intuition etc, go for it!

***Objective 2 (New): As mentioned before the advantages of regression is the interpretation of the model produced. It also works very well when the number of predictors is large and the sample size is small. Disadvantages to regression is the fact that not all data sets are going to yield “linear” relationships. If the relationship is quite complex, it will be difficult to fit a good model without a lot of clever engineering of new predictor variables.***

***Nonparameteric regression models are models which do not assume the residuals follow any distribution and do not assume any sort of “linear” structure on what the trend between the response and the predictors could be. For Objective 2, I want you to use the ISLR text book and read up on one nonparametric technique. I want you to select from k-nearest neighbors’ regression or regression trees. Splines could also be a topic but personally I think k-nn or trees would be much easier to digest for a short turn around project such as this.***

<http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf>

***After reading up on this topic, I want you to use a set of predictors used from your previous linear regression models and fit your nonparametric model to the data, keeping in mind what the particular model calls for in terms of what type of predictors can be used and parameters that control for model complexity. The ISLR book provides code at the end of each chapter and I will also try provide a small script for you to have so you’re not starting from scratch. In your write up after the regression models, I want you to include a small section that includes the following:***

1. A brief description of your nonparametric model’s strategy to make a prediction. Include Pros and Cons.
2. Provide any additional details that you feel might be necessary to report.
3. Report the test ASE using this nonparametric model so we can see how well it does compared to regression.

***Additional details***

NOTE: ALL ANALYSIS MUST BE DONE IN SAS OR R and all code must be placed in the appendix of your report. Python is okay for quick formatting of data and data visualization, but analysis should be in R or SAS.

***Required Information and SAMPLE FORMAT***

PAGE LIMIT: I do not necessarily require a page limit, but you should definitely be shooting for no more than 7-8 pages written for the main report (not including graphics and codes). It of course will blow up quite larger than that due to graphics, tables, and code but good projects are clear, concise, and 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.). Effective communication is critical here.

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

Introduction **Required**

Data Description **Required**

Exploratory Data Analysis **Required**

Addressing Objective 1:

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

Model Selection **Required**

Type of Selection

**Options**: LASSO, RIDGE, ELASTIC NET,

Stepwise, Forward, Backward,

Manual / Intuition,

A mix of all of the above.

Checking Assumptions **Required**

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Compare Competing Models **Required**

Via: Training and test set split or CV

**Possible Metrics**: ASE (**Required**), AIC, BIC, adj R2, are all welcomed addittions

Parameter Interpretation (Simple model only)

Interpretation

Confidence Intervals

Additional Details on a more complicated regression model

Objective 2 Deliverable (see above)

Final conclusions from the analyses of Objective 1’s interpretable model and include comments on what model would be recommended if prediction was the only goal (comparing all models considered).

Appendix

Well commented SAS/R Code

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

In addition to the overall conclusions, feel free to include additional insights or concerns gleaned from the analysis. What needs to be done next or how could we do it better next time or if we had more time?