MSDS 6372 Project 1 Description

For this project, we are going to be using a data set provided in Kaggle titled “Life Expectancy (WHO): Statistical Analysis on factors influencing life expectancy”. The information was collected over numerous years and, because logistics in recording changes over time, there will be some missing data issues that the group must discuss and develop a strategy to proceed.

Data, along with information on the variables included in the file can be found at <https://www.kaggle.com/kumarajarshi/life-expectancy-who>.

There are two main objectives for Project 1. Before providing the details to the objectives, the groups should investigate the data and make any additional logistic decisions like using all of the data or a subset, dealing with missing data ,cleaning up variable names, changing their data types, etc. Once that is complete, I want each group to subset the data set into 2 data sets. Suggested: 85% train, 15% test.

***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 and interpreting those relationships.***

* Build a model with the main goal to identify key relationships and that is highly interpretable. Provide detailed information on summary statistics, EDA, and your model building process.
* Provide interpretation of the regression coefficients of your final 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. Answer any additional questions using your model that you deem are relevant.
* The training data set can be used for EDA and model fitting while the test set can be used to help compare models to make a final call.

Practical Consideration for Objective 1:

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. You can summarize that iteration stuff 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 to show. 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. I will discuss presentation strategies “Dos and Donts” during our dead week (Unit 6).

***Objective 2: While your model from Objective 1 may be interpretable there may be some additional complexity that you could incorporate to your model so that it can predict better at the expense of interpretations. The purpose of this objective is to go through a process to compare multiple models with the goal of developing a model that can predict the best and do well on future data.***

* Use the training and test set to build at least one additional multiple linear regression model that attempts to find a model with additional complexity than the interpretable model of Objective 1. The point here is to make sure we understand how to add complexity to a linear regression model. Hint: It’s not just including a model with predictors that you’ve eliminated from Objective 1.
* I want you to use the ISLR text book below (and the google machine) and read up on one nonparametric technique to build a regression model. I want you to select from k-nearest neighbors’ regression or regression trees. There is a chapter on trees in the ISLR book. For information on knn regression, see Section 3.5. It is important to know that knn can be applied to regression as well as classification problems. Make sure your implementation in R is using the knn regression versions rather than the classification. See me on this if you need help or reassurance. You will use the training and test sets here to help determine the most appropriate tree or the most appropriate “k” in knn.
* <http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf>
* At this point you should have at least 3 models, 2 linear regression models and 1 nonparameteric model. For each of the three models, provide the primary measure of fit for comparisons: test ASE or via CV. You may also include additional metrics for completeness like R squared/Adjusted Rsquared, AIC, and BIC where applicable (remember these are only derived from the training set not the test set).

Practical Consideration for Objective 2:

Part of Objective 2 is to get comfortable with figuring new things out. As you learn about knn or regression trees, consider providing a brief description of how that model is used to make a prediction so your audience knows the basic idea. It’s clearly different from multiple linear regression. It doesn’t have to be technical. They are both pretty intuitive approaches and can be explained in a single paragraph.

When all we care about is predictions, don’t be afraid to try things. Transformations, interactions, creating new variables from old variables. Let your EDA help you in coming up with outside the box ideas. Note: This tip could be applied to Objective 1 as well, as long as it yielded an interpretable model.

Many students will eventually migrate to more automated packages in R like Caret for Objective 2. These packages may have the ability to run CV on all of the models simultaneously and provide a CV press statistic for each one of the models. If you choose to learn the caret package, you can combine the train and test sets together to fit your models, just report the CV ASE/PRESS instead of the test ASE.

Another graph that is helpful for just getting a sense for how your predictions are behaving is to plot the predicted test set values versus the true values in a scatterplot. The closer the data is to the 45 degree line the better and should. Note: You won’t be able to do this effectively with the caret approach listed above.

***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 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.

1. Introduction **Required**
2. Data Description **Required**
3. Exploratory Data Analysis **Required**
4. Addressing Objective 1: **Required**

* Restatement of Problem and the overall approach to solve it
* Model Selection

Type of Selection

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

Stepwise, Forward, Backward,

Manual / Intuition,

A mix of all of the above.

* Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

* Parameter Interpretation

Interpretation

Confidence Intervals **Not Required, but use if beneficial to the discussion.**

1. Addressing Objective 2: **Required**

* Restatement of problem and the overall objective
* Description of the approach for building a complex regression model. Feature selection must be used here if only manually model fitting approaches were used in Objective 1.
* Brief description of how the nonparametric tool works intuitively. Can this model overfit? How?
* Comparison of model results

Table of test ASE and any other relevant model fitting metrics.

Discussion and insight as to what the results suggest. Why does one fit better than the other? Or perhaps why does it appear that all the models appear to be performing about the same?

1. Final summary **Required**

* Quick recap of Objective 1 and Objective 2 findings
* Provide any additional details and comments on the implications of the models. Scope of inference? What other data would this model be good/poor to apply to? Problems/concerns with the data or data collection? What would you do if you have more time? What else would you collect? etc.

1. Appendix **Required**

* Well commented SAS/R Code. You may send me a github link if you wish.
* Graphics and summary tables (Can be placed in the appendix or in the written report itself.)
* Make sure you include figure labels and reference them in the report if you are using an appendix to communicate figures.