MSDS 6372 Project 1 Description

For this project I’m going to let you guys have some flexibility on what data set to use. You may use your own or I’ve provided a list of a few data sets below if you do not want to worry about finding a data set. If you do find your own data set, make sure it is suitable for regression (continuous response) and there are a reasonable number of predictors to work with (I would say at least 10-15). Sample size should be around the 200-300 range at a minimum.

1. <https://www.kaggle.com/kumarajarshi/life-expectancy-who>. Modeling a countries life expectancy based on various economic and health factors.
2. Some may find it interesting to go back and give the Ames housing market data set another try using the new approaches and discussion covered in class.
3. The Boston housing data set from the ISLR package. There are multiple ways to approach this in terms of defining the response variable. Original response is the median value of owner-occupied homes. However, from an environmental perspective it may be interesting to predict and model nitrous oxide levels.
4. Predicting medical costs for insurance claims. Insurance.csv in the FILES folder

Note that many publicly available data set have some data entry errors. While some cleaning is always required, students should not spend all of their time cleaning. The focus should be on the objectives listed below. When it comes to data cleaning, just make sure you communicate what you did, and if there are still concerns with the data just point them out in the discussions at the end of the presentation. It is totally fine (for this project) to delete variables that are not reliable or delete rows that are suspect or missing and you do not have the time (or its simply impossible) to replace them with more effective values ways

***Project Objectives***

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 dealing with missing data if there is any, cleaning up variable names, changing their data types, etc. I also want you to split the data into a train/validation split first, and then use the training data to go through the objectives.

***Objective 1: Display the ability to build regression models using the skills and discussions from Unit 1, 2, and 3 with the purpose of identifying key relationships and interpreting those relationships in an organized and clear fashion (Unit 4) .***

* Build a model with the main goal to identify key relationships and 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.

Practical Consideration for Objective 1:

* EDA, EDA, EDA! It helps you on so many fronts so use it to your advantage. When presenting a concise report, you do not have to literally step out every single step of your model building process or data cleaning. Make sure there is a focus on the discussion of trends between the response and predictors. 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 single slide.
* Objective 1 should include interpretation of at least a subset of the predictors in your final model. You should demonstrate your ability to perform tests on coefficients, assumption checking, and interpreting those coefficients. You should do what makes sense based on the data you are using and the questions that naturally would come up. For example, you may decide that a group variable is important and each group needs its own slope. If that makes sense then the student is expected to model that appropriately and perform test that make the most sense for interpretation.
* Feature selection does not necessarily have to be a part of this objective, but it may make sense to use it to help make some final calls. Again do what makes sense.

***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.***

* In this objective you may take either a train/validation approach or a CV approach for model comparisons. The main point here is to make sure we understand how to add complexity to a linear regression model. Hint: Its not just including a model with predictors that you’ve eliminated from Objective 1.
* You are familiar with the implementation of KNN regression, regression trees, and random forest from discussions in class. I want you to run at least one nonparametric model to compare to your complex model.
* At this point you should have at least 3 models, 2 linear regression models and 1 nonparametric. For each of the three models, provide measures of fit for comparisons: such as validation MSE. You may also include additional metrics for completeness like R squared/Adjusted Rsquared, AIC, and BIC where applicable. This final to-do is the only point where the validation set is being used. This is because we are truly validating our decisions that have made using the training. It is important to do this step last when you have completed all other tasks. Do not continue to update models to get your validation MSE to be smaller. Just report it and offer a recap of the comparison of the results suggests and provide a discussion on your findings and what you would recommend using moving forward to make predictions. Additional insight as to why one model is better than the other, or why they are all the same is encouraged or discussions on how practical the models are is encouraged.

Practical Consideration for Objective 2:

Feature selection and CV is a must here to assess the bias/variance trade off of some potentially super complicated models you try. 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.

I recommend that you use caret to implement things in this objective as it is standardized. It will help ensure model comparisons are on a “apples to apples” level of comparison. Consider setting the seed each time you run a new model and perform CV.

***Project Deliverables***

Groups should provide a presentation of their analysis to address objective 1 and 2. The presentation should be recorded (with all group members contributing) and either a zoom link or hard file (mp4) should be provided as part of submission. The presentation should be no more than 30 minutes.

The group should also submit an R or Rmarkdown file containing the analyses presented in the presentation and ONLY the things presented in the report.

Only one person needs to submit the project. In summary there should be 3 items.

1. Presenation file
2. Video recording of presentation (youtube, mp4, etc)
3. R file of work

All students will be asked to fill out Peer Reviews. These should NOT be submitted with the project. There will be a separate process for Peer Review submission. Dr. Turner will provide more detail later.

**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

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

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

Perhaps additional EDA to motivate (situational)

Model comparisons

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? Recommendations moving forward? Insight the model gave? Etc.

Appendix: Do not need to present but if you think there are additional tables or graphics worthy of reference, then place them in the back and mention them while presenting.