## Final Project Instructions

Respond to the following in the six project sections:

1. **Introduction (5 points)**
   1. It is recommended that you complete the introduction after you have finished all the other parts.
   2. State whether you solved your project as a regression or classification problem.
   3. The general proposal comments including instructions for how to convert a regression problem to a classification problem if that is what you are interested in working on.
   4. Describe the major findings you have identified based on your analysis.
   5. Which inputs/features seem to influence the response/outcome the most?
   6. What supported your conclusion? Was it only through predictive models?
   7. Could EDA help identify similar trends/relationships?
   8. Was clustering consistent with any conclusions from the predictive models?
   9. What skills did you learn from going through this project?
   10. This is not related to application or project inputs/outputs directly. What general skills can you take away from the project to apply to applications more specific to your area of interest?
2. **EDA (10 points)**
   1. Basic Information
      1. You must provide important information about your data. Although you gave that information in the proposal, it is always important to provide those details when working with data.
         1. Show the number of rows and columns
         2. The variable names
         3. The data types
         4. Number of missing values per variable
         5. Number of unique values per variable
   2. Visualization
      1. In the main report, you must include useful visualizations that represent:
         1. Counts of categorical variables
         2. Distributions of continuous variables
         3. Relationships between continuous variables
         4. Summaries of the continuous variables grouped by categorical variables
         5. Consider visualizing relationships between continuous inputs using scatter plots, pair plots, joint density plots, and correlation plots.
         6. **If you are working on a regression problem:** Visualize scatter plots between the continuous response and the continuous inputs. Summarize the response with boxplots for the unique values of the categorical inputs. Consider using residual plots to assess the fit of your regression model and identify any patterns in residuals.
         7. **If you are working on a classification problem:** Visualize histograms and relationships between continuous inputs broken up by the outcome unique values. Count the number of observations for each combination of outcome unique value and the categorical input unique values. Use ROC curves or precision-recall curves to assess the performance of classification models, alongside confusion matrices.
   3. Reporting
      1. In the main report you must include what you feel are the most important figures to describe the behavior in your data.
      2. Avoid overcrowding your main report with unnecessary figures. Each figure should serve a clear purpose in supporting your analysis. Essentially, describe the important conclusions that should be drawn from that figure. The description does not have to be long. You should provide enough to detail to make it obvious you understand what the figure is showing. Thus, your main report does not need a full EDA—You already performed a full EDA with the proposal.
      3. Failure to draw inferences from the figures included in your main report may result in a penalty. Ensure that your discussion clearly reflects an understanding of what the figure indicates and its significance to your analysis.
      4. The supporting document for EDA can include many more figures. In this way, your support document can be your practice notebook where you try out different groupings and visualizations. You do not need to include discussion text in your support document. If you feel your proposal included all the figures you tried out, you may submit your proposal again as the supporting document. The supporting document allows you the opportunity to show you explored other options after receiving the proposal feedback. For example, you should visualize distributions of transformed variables as recommended in the proposal feedback.
   4. Considerations
      1. Based on the proposal feedback, explore the different methods available to handle MISSING data and ensure that you visualize the impact of each method. Ensure that your EDA and reporting are thorough, and remember that inferences drawn from your visualizations are essential to demonstrate your analytical skills.
3. **Clustering (15 points)**

**Note**: You executed CLUSTERING as part of the PROPOSAL. However, you have learned more about clustering, summarizing, and visualization since the proposal was submitted. You must retry clustering to incorporate what you have learned since your initial attempt.

* 1. Considerations
     1. If you have a mix of continuous and categorical inputs in your data set, cluster the data based on the continuous inputs alone. After identifying the optimal number of clusters, compare the cluster assignments to unique values of several of the categorical inputs.
     2. Summarize the continuous inputs associated with each of the cluster assignments.
     3. You should consider using boxplots, violins, scatter plots, conditional density plots, pairs plots, and/or conditional joint density plots.
     4. **If you are working on a classification problem,** compare your cluster assignments to the outcome unique values.
     5. If your EDA revealed that some continuous variables are not “Gaussian-like,” apply a necessary transformation (such as the log transformation) to those variables before performing the clustering.
     6. If your EDA revealed that the continuous variables have vastly different scales, you must standardize the variables before clustering. If your EDA revealed that the continuous inputs are highly correlated, consider clustering using the original variables.
     7. Failure to draw inferences from the figures included in your main report may result in a penalty. Ensure that your discussion clearly reflects an understanding of what the figure indicates and its significance to your analysis.
  2. Reporting
     1. Include how you identified the optimal number of clusters, and the required interpretation/visualization of the cluster results in the main report.
     2. The supporting document can include as much as you want. Again, treat the supporting the document as practice. This way you can have a notebook dedicated to clustering and keep that separate from the EDA and predictive modeling practice. Ensure that your EDA and reporting are thorough and remember that inferences drawn from your visualizations are essential to demonstrate your analytical skills.

1. **Models: Fitting and Interpretation (30 points)**

Note: BEFORE FITTING THE MODELS. If your EDA revealed that some continuous variables are not “Gaussian-like”, apply the natural log or square root transformation to those variables before modeling. If your EDA revealed that the continuous variables have vastly different scales, you MUST standardize the variables before modeling.

* 1. You must fit multiple models to predict the response of interest.
  2. If you are working on a regression problem, you must use linear regression models. If you are working on a classification problem, you must use logistic regression models. You must use statsmodels to fit these models.
  3. You will consider multiple formulations (FEATURES included in the model) and study the coefficients and their confidence intervals to interpret feature importance and relationships to the response.
  4. Formulations you must consider:
     1. No INPUTS! Fit an INTERCEPT-ONLY or CONSTANT AVERAGE model.
     2. Categorical inputs with additive features.
     3. Continuous inputs with linear additive features.
     4. All inputs (continuous and categorical) with linear additive features.
     5. Continuous inputs with linear main effect and pair-wise interactions.
     6. Interact the categorical inputs with the continuous inputs. This model must include the linear main effects as well (the formula interface helps with the interactions).
  5. You MUST also try 2 additional formulations. The additional formulations are up to you, but base you choices on your EDA. So what could you try? Here are several examples:
     1. If you noticed a potential non-linear relationship between the output and an INPUT you should include polynomial features in the model. You may want to try interacting categorical inputs with linear (main effect) and polynomial features derived from the continuous inputs. This checks if the non-linear relationship changes across categories.
     2. If you noticed the output has cyclical behavior or a repeating pattern with respect to an input then you could apply the SIN() function to inputs. You should standardize the inputs BEFORE applying SIN(). Interacting the SIN() function with categorical inputs allows you to check if the repeating pattern changes across categories.
     3. You will fit a total of 8 models. The Spotify Songs application has a single main output. As stated in the project feedback, you can treat this as a regression task after applying an appropriate transformation. Alternatively, you can convert the problem to a classification task. You may decide which approach you want to work on.
  6. For each model that you fit you must answer the following questions associated with the regression coefficients:
     1. How many coefficients were estimated?
     2. How many coefficients (and thus features) are STATISTICALLY SIGNIFICANT using commonly accepted thresholds?
     3. WHICH coefficients (and thus features) are STATISTICALLY SIGNIFICANT and what are the coefficients POSITIVE or NEGATIVE for those features?
     4. Which two STATISTICALLY SIGNIFICANT coefficients (and thus features) have the highest MAGNITUDE coefficient values?25
  7. For each model that you fit you must show the performance on the training set. What you show depends on whether you are working on a REGRESSION or CLASSIFICATION task.
     1. REGRESSION
        1. For each model show the predicted vs observed figure for the training set, and the R-squared and RMSE on the training set.
        2. Which model has the best performance on the training set? Is the best model according to R-squared the SAME as the best model according to RMSE? Is the best model better than the INTERCEPT-ONLY model? How many coefficients are associated with the BEST model?
     2. CLASSIFICATION
        1. For each model show the CONFUSION MATRIX on the training set assuming a 0.5 threshold; the Accuracy, Sensitivity, Specificity, and FPR on the training set assuming a 0.5 threshold; the ROC curve on the training set; and the Area Under the ROC curve (ROC AUC) value.
        2. Which model has the best performance on the training set? Is the best model different when considering Accuracy vs ROC AUC? Is the best model better than the INTERCEPT-ONLY model? How many coefficients are associated with the BEST model?

1. **Models: Predictions (15 points)**
   1. You must make predictions with two models.
      1. You must predict with the model with ALL inputs and linear additive features.
      2. You must predict with the best model on the training set.
   2. You must create data sets to support visualizing the predictions. You must NOT make predictions of the training set.
   3. REGRESSION tasks must create 2 visualization grids while CLASSIFICATION tasks must make 1 visualization grid.
      1. Regression Prediction Visualizations: Grid 1
         1. You must identify the continuous input that you feel is the MOST important based on the statistically significant coefficients in your models. The MOST important input must have 101 unique values between the minimum and maximum training set values.
         2. ALL other inputs must be set to CONSTANT values. Continuous inputs should use a CENTRAL value like the MEAN or MEDIAN. Categorical inputs should use the MOST frequent category.
         3. Make predictions with BOTH models on the visualization grid. You MUST visualize the AVERAGE OUTPUT as a line, the CONFIDENCE INTERVAL as a grey ribbon, and the PREDICTION INTERVAL as an orange ribbon with respect to the most important input.
         4. Comment on whether the TREND and UNCERTAINTY appear different between the two models.
      2. Regression Prediction Visualizations: Grid 2
         1. You must identify the THREE most important inputs based on the statistically significant coefficients in your models. One of the three must be the SAME input you selected for grid 1. This input will have 101 unique values in the grid 2, just as it did in grid 1. The two other inputs must have fewer unique values in the grid. Categorical inputs should use all unique values (categories). Continuous inputs should use 5 unique values between the training set minimum and maximum values.
         2. ALL other inputs must be set to CONSTANT values. Continuous inputs should use a CENTRAL value like the MEAN or MEDIAN. Categorical inputs should use the MOST frequent category.
         3. Make predictions with BOTH models on the visualization grid. You MUST visualize the AVERAGE OUTPUT as a line with respect to the input with the 101 unique values in the visualization grid. The line must be colored by one of the two other inputs with non-constant values. The third input must be associated with the facets. It is your choice as to which input is associated with the line color (hue) versus the facets. If you color by a continuous variable, you should use a diverging color palette. You may use the default color palette for coloring by a categorical variable.
         4. Comment on the trends between the two models.
      3. Classification Prediction Visualizations
         1. You must identify the THREE most important inputs based on the statistically significant coefficients in your models. However, one of the three inputs must be a continuous input. The most important continuous input of the three must have 101 unique values in the grid between the training set minimum and maximum values. The two other inputs must have fewer unique values in the grid. Categorical inputs should use all unique values (categories). Continuous inputs should use 5 unique values between the training set minimum and maximum values.
         2. ALL other inputs must be set to CONSTANT values. Continuous inputs should use a CENTRAL value like the MEAN or MEDIAN. Categorical inputs should use the MOST frequent category.
         3. Make predictions of the EVENT PROBABILITY with BOTH models on the visualization grid. You MUST visualize the EVENT PROBABILITY as a line with respect to the input with the 101 unique values in the visualization grid. The line must be colored by one of the two other inputs with non-constant values. The third input must be associated with the facets. It is your choice as to which input is associated with the line color (hue) versus the facets. If you color by a continuous variable, you should use a diverging color palette. You may use the default color palette for coloring by a categorical variable.
         4. Comment on the trends between the two models.
   4. Reporting
      1. The creation of your visualization grid and figure showing the predictive trends should be given in the main report.
2. **Models: Performance and Validation (25 points)**

Note: Part D focused on fitting models on the training set and studying the coefficients. Performance was assessed ONLY on the training set. Part F is focused on identifying the model that will perform the best on NEW data. You will use CROSS-VALIDATION to calculate a more reliable estimate of the model performance.

* 1. You will use 3 of the formulations you considered in Part D.
     1. You must select the formulation that was the best model on the training set.
     2. You must choose 2 additional formulations. One should be simple (few features), and one should be of medium to high complexity.
  2. You must use CROSS-VALIDATION to evaluate the performance for each of the 3 models. Using “regular” K-fold OR stratified K-fold depends on whether you are working on a REGRESSION or CLASSIFICATION problem. You may choose 5-fold, 10-fold, or repeated cross-validation. You must identify an appropriate performance metric to focus on based on the output data type.
  3. Visualize the CROSS-VALIDATION results by showing the AVERAGE CROSS-VALIDATION performance metric with the 95% confidence interval for each model.
  4. Which model is the BEST according to CROSS-VALIDATION?
  5. Is this model DIFFERENT from the model identified as the BEST according to the training set?
  6. How many regression coefficients are associated with the best model?
  7. Important considerations
     1. PRE-PROCESSING actions like standardization must be applied within each FOLD and CANNOT be performed BEFORE applying CROSS-VALIDATION (discussed in Weeks 13-14).
     2. PRE-PROCESSING actions like applying LOG-TRANSFORMATIONS can be performed BEFORE applying CROSS-VALIDATION.
     3. There are multiple ways to approach executing the CROSS-VALIDATION. Select the approach you feel is best for you.
     4. You may use scikit-learn to manage the data splitting required for CROSS-VALIDATION.
     5. You may fit the models with statsmodels OR scikit-learn based functions.
     6. You may use the Pipeline module to manage all aspects of preprocessing, model fitting, evaluating, and running cross-validation.
     7. Include discussion text in your main report associated with applying CROSS-VALIDATION to identify the best model.
     8. Your supporting notebook can be used for practice. This lets you practice the CROSS-VALIDATION away from the rest of the work for the project. This approach is recommended to make your work more modular.

## Optional Project Inclusions

The following are optional inclusions to your final project.

1. Ridge
   1. Apply the RIDGE penalty to the MOST complex model you fit even if CROSS-VALIDATION says it is NOT the best model.
   2. Train and tune a RIDGE penalized model via CROSS-VALIDATION.
   3. Examine the final tuned coefficient estimates.
2. Lasso
   1. Apply the Lasso penalty to the MOST complex model you fit even if CROSS-VALIDATION says it is NOT the best model.
   2. Train and tune a LASSO penalized model via CROSS-VALIDATION.
   3. Examine the final tuned coefficients estimates – are any zero?
3. Elastic Net
   1. Apply the ELASTIC NET penalty to the MOST complex model you fit even if CROSS-VALIDATION says it is NOT the best model.
   2. Train and tune an ELASTIC NET penalized model by optimizing BOTH tuning parameters via CROSS-VALIDATION.
   3. Examine the final tuned coefficients estimates. Is the tuned model closer to RIDGE or LASSO? Are the final tuned coefficients zero?
4. Model of your choice.
   1. We have focused on (generalized) linear models (linear regression and logistic regression). You may select a non-linear model of your choice, such as Support Vector Machines, Ensembles, or neural networks. The syntax for fitting them is VERY similar to the scikit-learn linear model and logistic regression syntax (Weeks 13-14). Use cross-validation to tune and assess performance of that model.