## Course Project Instructions

One of the main objectives of this course is to help you gain hands-on experience in communicating insightful and impactful findings to stakeholders. In this project you will use the tools and techniques you learned throughout this course to train a few classification models on a data set that you feel passionate about, select the regression that best suits your needs, and communicate insights you found from your modeling exercise.

After going through some guided steps, you will have insights that either explain or predict your outcome variable. As a main deliverable, you will submit a report that helps you focus on highlighting your analytical skills and thought process.

Write your report from the perspective of a Chief Data Officer or the Head of Analytics and will assess whether your final linear regression went through all the necessary steps to achieve the main objective of your analysis.

Yes, you are expected to leverage a wide variety of tools, but this report should focus on presenting findings, insights, and next steps. You may include some visuals from your code output, but this report is intended as a summary of your findings, not a code review. Optionally, you can submit your code as a python notebook or as a print out in the appendix of your document.

The grading will center around 5 main points:

* Does the report include a section describing the data?
* Does the report include a paragraph detailing the main objective(s) of this analysis?
* Does the report include a section with variations of classifier models and specifies which one is the model that best suits the main objective(s) of this analysis?
* Does the report include a clear and well presented section with key findings related to the main objective(s) of the analysis?
* Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different predictive modeling techniques?

**Setup instructions:**

Before you begin, you will need to choose a data set that you feel passionate about. You can brainstorm with your peers about great public data sets using the discussion board in this module.

**How to submit:**

The format of your work must adhere to the following guidelines. **The report should be submitted as a pdf.** Optionally, you can include a python notebook with code.

Make sure to include mainly insights and findings on your report. There is no need to include code, unless you want to.

### Project

### *Optional: find your own data set*

As a suggested first step, spend some time finding a data set that you are really passionate about. This can be a data set similar to the data you have available at work or data you have always wanted to analyze. For some people this will be sports data sets, while some other folks prefer to focus on data from a datathon or data for good.

### *Optional: participate in a discussion board*

As an optional step, go into a discussion board and brainstorm with your peers great data sets to analyze. If you prefer to skip this step, feel free to use the Ames housing data set or the Churn phone data set that we used throughout the course.

### Required

Once you have selected a data set, you will produce the deliverables listed below and submit them to one of your peers for review. Treat this exercise as an opportunity to produce analysis that are ready to highlight your analytical skills for a senior audience, for example, the Chief Data Officer, or the Head of Analytics at your company.

Sections required in your report:

* Main objective of the analysis that specifies whether **your model will be focused on prediction or interpretation** and the benefits that your analysis provides to the business or stakeholders of this data.
* Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.
* Brief summary of data exploration and actions taken for data cleaning and feature engineering.
* Summary of training at least three different classifier models, preferably of different nature in explainability and predictability. For example, you can start with a simple logistic regression as a baseline, adding other models or ensemble models. Preferably, all your models use the same training and test splits, or the same cross-validation method.
* A paragraph explaining which of your classifier models you recommend as a final model that best fits your needs in terms of accuracy and explainability.
* Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your classifier model.
* Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model after adding specific data features that may help you achieve a better explanation or a better prediction.

### FAQs

Here are frequently asked questions about the assignment and review process. Please read these before starting your assignment.

Do I have to come up with my own data set?

* You are highly encouraged to find a data set you feel really passionate about. This will help you showcase analytical work that truly matches your skills. But if you prefer, you can use some of the data sets from this course.

Is it OK to choose the same data set as someone else?

* Yes, more than one person can analyze the same data set. Most likely your insights will be different from your peers and you will still be able to showcase your own talent as a unique solution.

Do I have to train more than 3 different classifiers?

* It is highly recommended that you try at least three different classifiers to highlight which tool or technique improved your prediction or interpretation.

Is this an individual assignment?

* You can ask for help or assistance on technical issues and general direction of your analysis, but the interpretation of the analytical output and the writing of the report should be your own.

Interpretation:

* In some cases, the primary objective is to train a model to find insights from the data.
* In yp = f(omega,x), the interpretation approach uses omega to give us insight into a system.
* Understand the underlying the features that are provide the most value in predicting our outcome variable.
* Commong workflow:
  + Gather x, y; train model by finding the omega that gives the best prediction yp = f(omega, x)
  + Focus on omega (parameters) rather than the best prediction or yp, to generate insights.
* Exmple interpretation exercises:
  + X = customer demographics, y = sales data; examine omega to understand loyalty by segments
    - Whats driving sales as opposed what will the sales be
  + X = car safety features, y = traffic accidents; examine omega to understand what makes cars safer
    - Find right safety features rather than predicting the number of accidents a car will have.
  + X = marketing budge, y = movie revenue; examine omega to understand marketing effectivness
    - How much should we spend, rather than predicting a movies box office revenue.

Prediction:

* In some cases, the primary objective is to make the brest prediction
* In yp = f(omega, x), the prediction approach compares yp with y
* The focus is on performance metrics, which measure the quality of the model’s predictions
  + Performanc emeterics usually involve some measure of closeness between yp with y.
  + Without focusing on interpretability, we risk having a black-box model.
* Example prediction exercises:
* X = customer purchase history, y = customer churn; focus on predicting customer churn
  + Correctly predicting the likelyhood of a particular customer will churn
* X = financial information, y = flagged default/non-default; focus on predicting load default
* X = purchase history, y = next purchase; focus on predicting the next purchase
  + Predict future purchase

Tradeoff between high level of interpretability or high level of predictability