Execution Checklist

This checklist will help you complete your project with confidence. These are mainly guidelines, so feel free to adjust them as needed for your project.

Section #1: Introduction

A proper introduction gives visitors the context needed to understand the project. This is also where you should introduce the dataset and explain the challenge you're solving.

❏ What is the main challenge or topic your project addresses?

❏ Which dataset are you using and how did you acquire it? (*Also summarize the types of features and variables available in the dataset.*)

❏ What are the most important findings from your project? (*Preview your results and draw visitors in.*)

❏ How does your project address the challenge? (*Which data science and machine learning techniques do you use?*)

❏ Who are you and why is this project important or valuable to you? (*What is your motivation for doing this project?*)

Section #2: Library Imports

This section is very straightforward. Simply import the relevant libraries, modules, and algorithms you'll need for your project. Then, read in your dataset.

❏ **Tip:** Review the previous Cornerstone Projects if you need a refresher.

❏ **Tip:** You don't need to get this perfect on the first try. If you forget an import, it's no big deal. Just come back here, add it, and then re-run the code cell.

Section #3: Exploratory Analysis

Exploratory analysis is all about scouting ahead. Make sure you can answer the following questions for yourself. Review the Cornerstone Projects if you need a refresher on the code and visualizations.

Basic Information

- ❏ How many observations and features does your dataset have?
- ❏ Do you understand each feature intuitively?
- ❏ Do the values for each feature make sense? Are they on the right scale?
- ❏ Do you anticipate issues with missing data?
- ❏ Were your features read in as the correct datatype?

Distributions

- ❏ Do each of the distributions make intuitive sense to you?
- ❏ Do you anticipate any issues with outliers or sparse data?
- ❏ Are there any surprising distributions you should take a closer look at?
- ❏ Do the summary statistics confirm what you've seen from the charts?

Feature Relationships (Segmentations & Correlations)

- ❏ Have you segmented key categorical features and/or the target variable?
- ❏ What have you learned about the relationships between your features?
- ❏ Are there any surprising correlations (or non-correlations)?
- ❏ Have you visualized your correlation matrix for easier reference?
- ❏ Do you anticipate any helpful new features to engineer?

Section #4: Data Cleaning

Remember, proper data cleaning can make or break your project. Better data > fancier algorithms. Review the Cornerstone Projects if you need a refresher on the code.

Unwanted Observations

- ❏ Have you dropped duplicate observations?
- ❏ Have you dropped irrelevant observations?

Structural Errors

- ❏ Are there any features that should be encoded as binary indicator variables?
- ❏ Have you fixed typos and inconsistent capitalization in your categorical features?
- ❏ Are there any classes in your categorical features that refer to the same thing? (*e.g. "N/A" and "Not Applicable" appearing as two different classes*)

Outliers

- ❏ Have you visually checked for any potential outliers to remove in your features?
- ❏ Do you have a good reason to remove each outlier? (*e.g. suspicious measurements, different population, different application*)

Missing Data

- ❏ Have you labeled missing values in categorical features?
- ❏ Have you flagged and filled missing values in numeric features?
- ❏ **Tip:** There are certain situations where dropping observations with missing values is appropriate, such as if you only care about predicting observations that have a given feature value. For example, you might *only* wish to predict housing prices for single-family homes, in which case you would simply drop any observations that weren't for single-family homes (including those with missing values for property type).

Section #5: Feature Engineering

Feature engineering is one of the best ways data scientists can improve model performance and add value into the applied machine learning process.

Domain Knowledge

- ❏ Do you have prior expertise in your chosen domain? If not, have you done sufficient reading / research / preparation to understand it better?

- ❏ Do you know anyone else in your network who also has domain expertise?

- ❏ Based on your knowledge of the domain, are there any features you could engineer that would *potentially* improve the performance of your model?

Heuristics

- ❏ Are there any interaction features you could create?

- ❏ Are there any indicator features you could create?

- ❏ Have you grouped sparse classes in your categorical features?

- ❏ Do you need to do any form of data wrangling, such as aggregating data (i.e. rolling it up)?

- ❏ Are there any ordinal categorical features you could encode as numeric?

- ❏ Are there any potentially useful outside datasets you could merge in?

Preparing the ABT

- ❏ Have you created dummy variables for your categorical features?

- ❏ Have you dropped unused and/or redundant features? (*e.g. ID columns, features that wouldn't be available, text descriptions and metadata, etc.*)

Section #6: Algorithm Selection

For your Capstone Project, you've probably already imported the relevant algorithms at the start of the project, under the Library Imports section. Therefore, you should use this section to explain your choices and showcase your understanding.

- ❏ Why did you choose those algorithms?
- ❏ What are their practical benefits?
- ❏ What are the key hyperparameters to tune for your chosen algorithms?

Section #7: Model Training

Once you've done the steps leading up to this one, model training should be straightforward and formulaic. Review the Cornerstone Projects if you need a refresher on the code.

Data Spending

- ❏ Have you split your dataset into separate training and test sets?
- ❏ Have you set a random seed for replicable results?
- ❏ Do you understand the purpose and use-case of cross-validation?

Pre-Processing & Pipelines

- ❏ Have you set up your modeling pipelines with the proper preprocessing steps?
- ❏ Have you set random states for each algorithm to ensure replicable results?

Hyperparameter Tuning

- ❏ Have you declared hyperparameter grids with reasonable hyperparameter values to try for each of your algorithms?
- ❏ Have you set up GridSearchCV objects for each of your algorithms to perform cross-validation and tune hyperparameters?
- ❏ Have you fit models using each of your algorithms?

Winner Selection

- ❏ Which of your models had the best cross-validated score?
- ❏ Which of your models performs the best on the test set?
- ❏ Were you able to satisfy your win-condition for this project?
- ❏ Do you need to use any additional performance metrics to evaluate your model?

Section #8: Insights & Analysis

Many people miss this crucial Insights & Analysis section. This comes at the end of your project, and it's really there to tie everything together. This is where you'd summarize your results, discuss your most important findings, and even explain how you would expand upon your project if you had more time and resources.

- ❏ What were your key findings and results?
- ❏ What was your winning model (if applicable)?
- ❏ What did you *personally* learn by completing this project?
- ❏ How would you expand upon or improve this project if you had more time and/or resources?
- ❏ Are there any additional datasets that you would wish to acquire?
- ❏ Were there any useful references that helped you complete your project? If so, you should add citations at the bottom.

And voilà! Take your time, don't rush, and make sure you're all caught up. You may need to do a few more iterations or tweaks, but after completing these steps, you should now have a pretty kick-ass project that is UNIQUE to you… how awesome is that!

In the next module, we'll look at polishing it up a bit more, hosting it online, and adding it your own personal portfolio!