Introduction to Data Science: Final Project

For the final project, we will be building upon the midterm and using our more advanced toolset of statistics analysis and machine learning. As such, you should begin by reusing your previous dataset or acquiring a new one in order to answer a practical question or application. From there, you will practice a standard data science workflow, exploring the initial data, cleaning null values and transforming features as appropriate. (For example, you might have to create dummy variables from a categorical variable if you wish to use that information within a regression model.) Finally you will then apply a machine learning technique whether it be regression, classification or clustering in order to yield further insights and a useful model for analysis or predictive purposes.

	Below Standards	Meets Standards	Exceeds Standards
Exploratory Data Analysis	Does little to no exploratory analysis of the data before jumping into machine learning techniques.	Performs initial exploration of data, exploring correlation and multicollinearity between variables and distribution of individual features.	Creates heatmaps or investigates further relationships and trends between subsets of the dataset.
Preprocessing	Does not perform standard preprocessing techniques such as normalization or filling null values.	Performs standard preprocessing techniques including filling (or dropping null values) and normalizing data features to a standardized scale.	Performs additional preprocessing techniques such as creating dummy variables.
Feature Selection/ Engineering; Appropriate Setup	Does not choose features that are appropriate for the problem domain.	Chooses appropriate features to feed into a machine learning pipeline. Engineers at least one complex feature in an attempt to improve model performance. (This effect should be measured before/after including the new feature.)	Goes through an iterative process, testing hypotheses and the impact of various features on the model, ultimately selecting the most impactful features.
Machine Learning	Does not employ machine learning techniques.	Effectively employs at least one machine learning technique discussed. If supervised, uses train test split appropriately, and compares results across the two.	Employs multiple machine learning techniques or tests the effect of several tuning parameters on a single algorithm.
Presentation	I have made some initial graphs or statistics, but have failed to present those as a cohesive story.	I have a defined problem and have at least a preliminary analysis of the question. I have summarized this as a blog post or presentation.	I have a defined problem and have analyzed the problem from multiple angles, synthesizing this into a convincing point of view.

Project Deadlines

Wednesday August 29th -- First Draft / Outline Due
Wednesday September 5th -- Projects Due/Presentations