Introduction to Data Science: Midterm Project

For the midterm project, 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 regression techniques in order to build and tune a predictive model.

	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. Presents a minimum of 3 insightful visuals.	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.
Regression Techniques	Does not employ machine learning techniques.	Effectively employs regression techniques including a train test split or cross validation.	Employs multiple regression 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 Timeline / Outline

Class 5 - 7: Final Skills and techniques to be used for midterm will be introduced and discussed

- Simple Regression Models
- Train / Test Split
- Lasso / Ridge Regression
- Cross Validation / Tuning

Class 8: Lab practice

- This in class exercise will be very similar to the midterm project and will review the full data science process from loading data to model tuning

Class 9: Devoted time to work on project

- Class time will be devoted to work on your midterm project
- Homework: Finish Project

Class 10 : Project Due + Class Discussions