Machine Learning Project

Objectives:

Synthesis of machine learning material is the goal of this project. While the dataset is drawn from a Kaggle competition, the purpose should be oriented to demonstrating of understanding machine learning theory.

For this project, your primary task is to employ machine learning techniques to accurately make predictions given a dataset. The framework will be through the lens of the House Prices: Advanced Regression Techniques from Kaggle. While the primary goal of Kaggle competitions is generally focused on predictive accuracy, you will be expected to lead your audience through descriptive insights as well. For the purposes of your project you will aim to not only create a model that predicts well, but also allow yourself to describe data insights drawn from exploration.

1. Submission in respect to the deadline.
2. Background knowledge of dataset(s).
3. Communication of motivation: why do we care?
4. Research questions of interest: what do you want to find out?
5. Answers to research questions: what have you uncovered?
6. Presentation skills.
7. Time management (not going over the allotted time).
8. Ability to answer audience questions effectively and efficiently.
9. Balance of complexity and simplicity.
10. Explanation of future work: what would you do if given more time, data, etc.?
11. Demonstration of EDA skills (Numeric/graphic methodology)
12. Demonstration of machine learning skills (Supervised/unsupervised methodology)
13. Ability to assess model weaknesses and identify improvements.
14. Ability to manage a team workflow.

Meeting Jonathan Presley:

* EDA and Preprocessing:
  + Figure out what kind of variables (categorical, integer, ordinal) so I can transform them as needed.
  + Missingness to see if I should impute.
  + Correlation to find if potential multicollinearity or strong correlation with target.
  + Outliers, to see if I treat them differently.
  + If there is a skew to try to normalize, to have a more normal distribution of residuals, maybe one category is heavily unbalanced.
* Feature engineering (creating new variables with what I have).
* Modeling, get scores and compare.
  + From simpler to more complex models.
  + Score my models on root mean square errors because Kaggle will rate it that way.