Project Proposal

Team:

1. **Michael Valentino**
2. **Jie Yeh (Henry)**
3. **Jianxiong Zhao (Jason)**

Project Plan and description

Link:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Quick stats of the data-set:

- 1460 training, and 1460 test for submission on Kaggle
- 80 features, half of which are categorical
- Comes with descriptive text file of features
- Contains NaN values.
- Target labels are numerical.

What we will learn:
The training set data is not so large, but is a suitable size for learning and experimenting with different Regression techniques. The inclusion of many categorical features will force us to explore ways of encoding them to numerics. The NaN values will also present opportunities to try different techniques for dealing with missing information. Especially considering the NaN values appear in both numerical and categorical features. Using the description file we can learn how to try and better understand data we are working with. Additionally, we can utilize data visualization to more rigorously analyse and pull out patterns/correlations between features and the target labels.

What we will/may use:

- Linear and Polynomial model regression
- Gradient Boosting
    - Using sklearn's default implementation
    - Implementing our own simple gradient boost on other methods.
- Ensemble based methods
    - Random Forests
    - Building ensembles Linear and Polynomial regressors
- Techniques for preventing overfitting
    - LASSO
    - Ridge
    - Early stopping