Working with Advanced Regression Techniques

I. Title

Participating in a Kaggle Competition titled *"House Prices: Advanced Regression Techniques"*.

II. Abstract

Based on the leaderboards for this competition, our goal will be to obtain the lowest possible root mean squared logarithmic error (current leaderboards ~.10). Our goal is to also practice creative engineering and improve our algorithm development skills.

Looking at the data, we expect to need to preprocess the data because there exist numerous 'NA' values, numerous categorical values and values which are irrelevant to the algorithms price prediction(i.e 'Utilities' which are 100% 'AllPub'. Therefore this field probably doesn't offer much insight on the price). We will be attempting to employ various machine learning algorithms, including linear regression, random forests and boosting to achieve this task. Scikit-learn provides all we will need for linear regression and random forest. XGboost will help increase accuracy on linear regression.

III. Hypothesis

Regularizing the missing data in the tree we will lead to a better result than ignoring missing data points.

IV. Specific Aims

We plan to test several methods of learning, attempting to combine them fluidly, to see if regularizing the data leads to improved learning results. Possible methods to test include decision trees, decision forests, and neural networks.

V. Potential pitfalls and alternative strategies

We see a pitfall in that, if the data regularization is inaccurate, it will likely corrupt the results more than it helps them. The alternative path is to simply preemptively sweep the data and ignore these NA data points.

VI. Team Members

Jesse Hazard, Pawel Linek, Joseph Burns.

VII. Dataset

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