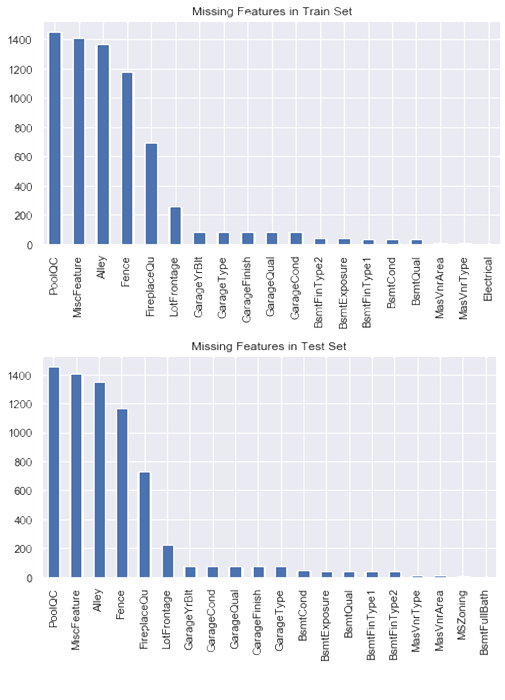
Missingness

Only a few columns had any missing values, but of those, some were missing 90% or more of the entries. Looking at the top three missingness columns, we can see that the high rates of missingness are due to the feature in question not being relevant or present. Most columns do not have a swimming pool, so swimming pool quality is expected to be blank. The same goes for fences and alleys. These missing entries merely need a placeholder value to represent “none,” so we imputed zeroes. On further analysis, this prescription was found to be appropriate for all of the columns with missingness but for “LotFrontage,” which is common to all properties, but also missing nearly 1 time in 7. This required a careful choice of imputation technique so as to preserve the statistical integrity of the variable. Because lot frontage was found to be characteristic of the neighborhoods, and many neighborhoods were found to have either a single very common value or a narrow distribution of common values, we resolved this by imputing the mean lot frontage value for the neighborhood of the property.



Stacking

In order to get the most out of each of our models, we need to combine them. The simplest way to do this would be to average the predictions, giving them each equal weight. This approach has the advantage of simplicity, but it fails to use the varying accuracy of the models to the greatest advantage. In order to get the best results from our combined models, we want to be able to tune the relative weights so as to get the greatest possible accuracy. If we set up the problem as a linear combination with varying weights, it becomes a simple linear regression problem, with the features being the models, and the dependent variable being the actual sale prices. This approach enables us to combine several models with given rmses and extract a prediction with a lower rmse than any one of the input models could achieve. This works best when we have several models with different strengths and weaknesses, rather than having all of the models be of one type. In our case, we have both tree-based, and linear models to work with.

