Michael Moran

Apply Machine Learning for Capstone Project

20 June 2018

Using Machine Learning to Fit a Better Model

To begin applying machine learning to the Housing Price data set, certain preliminary determinations must be made with respect to the data. Since there are many available input variables to use with Sale Price as an output variable, this is a supervised machine learning problem. A regression model is the appropriate way to frame this machine learning question since the goal of this project is to predict the price of a house and price is a continuous variable.

The machine learning techniques that were applied to the data are linear regression, Random Forest, and Support Vector Machines. The results from linear regression were evaluated based on their R-squared values and p-values. While these metrics did indicate that the selected independent variables were significant, the model performed poorly as it showed little sensitivity to including or removing different variables. To explore if a different model would better fit the data, Random Forest and Support Vector Machines were then run over the same set of variables as were used for linear regression.

The results of Random Forest and Support Vector Machines were validated by comparing the Root Mean Squared Error and the Mean Absolute Error of each model run over the test set and the train set. These validations methods were chosen because the model is predicting a continuous variable, and in the case of the Mean Absolute Error, because previous data analysis demonstrated the presence of outliers in the dataset.

The Root Mean Squared Error and the Mean Absolute Error were calculated for the initial runs of Random Forest and Support Vector Machines. Both models were then tuned using the variables identified by Random Forest as having greater importance, and the Root Mean Square Error and the Mean Absolute Error were calculated for the new models. A second revision was applied to the model using a higher threshold for feature importance, and the Root Mean Squared Error and the Mean Absolute Error were found for the updated models.

After two revisions, the Random Forest model was found to yield the best performance between the training set and the test set (as measured by the percent error of each set’s Mean Absolute Error of 25.66%), and to predict the sale price of house to within plus or minus 15.41% of the sale price for the average home. Further revision of the model was found to decrease the model’s performance, decreasing the Root Mean Square Error and MAE but increasing the percent error between the training and test values of each metric. Additional feature engineering on the Square Footage variable was performed, but it did not produce a more accurate model.