Sean O’Hara

Dr. Garcia

DSCI 401\_01

11/13/2017

Assignment 2

1. Data Preparation

In order to clean the data up in both datasets I had to perform a couple of data transformations before building a model. Upon looking at the data I noticed that there were four columns that had their observations measured in years. These columns are naturally integers in python and are thus treated as quantitative variables. Since one year is not quantitatively worth more than another, I chose to turn these columns into strings to make them categorical. Before I turned them into strings I first had to change each of the column names and replace the “.” with a “\_” so that it would be compatible with the integers to strings process. I then used the get\_dummies method to assign dummy variables to each of the categorical variables. After converting the categorical variables to quantitative variables, I replaced the missing observations in the data frame with values using the mean method which would impute new values for the missing one.

1. Exploratory Analysis

When I made an initial scan of the data a couple of things stood out to me. Variables such as Overall.Qual, Overall.Cond, X1st.Flr.SF, Total.Bsmt.SF, Garage.Area, as well as Gr.Liv.Area all had general positive trends when plotted with SalePrice as the response variable. Among these Gr.Liv.Area (General Living Area) and Overall.Cond (Overall Condition) seemed to have the strongest correlation with house sales price based on the charts plotted. I think that the overall condition and the general living room of the house will be strong predictors of the sales price. I don’t think that the months it takes to sell the house will be a good predictor since various types of houses take various lengths of time to sell depending on many factors one of which might include the condition of the housing market in a year.

1. Model Building

I used the Lasso approach to train my model. This method allows for the selection of variables for predictors by the shrinking of coefficients to exactly zero. Thus, the model limits the predictors we use to a subset of the predictors we start with. Initially I ended up with an R^2 of .90. Then I realized that I had removed some variables from the DataFrame that could have been left alone since the Lasso method performs the sub selection of predictors for me. After I added these variables back in I found that the best alpha value was 0.25 since it offered the highest R^2 value. As a result, my model performed with a R^2 of 0.999, a MSE of 782,422, and a MAE of 474. For comparison, my baseline linear regression model resulted in a R^2 of 0.946. The Lasso method provided a far better model then the baseline model.

1. Predicting and Validating

My best model performed on the test data with a R^2 of 0.986, and a MAE of 4,080.95 as well as a EVS of 0.986. In comparison, the R^2 for the baseline linear regression model performed with a 0.985 R^2. Compared to the baseline model, my model was only a small improvement. I think that the model is reasonably good. Naturally it didn’t perform as well as on the training dataset. Looking at the actual vs predicted values for the house sales price, in some cases the prediction was spot on, deviating by hundreds of dollars, in others it deviated in the tens of thousands. The MAE is 4,080.95, meaning that the predicted values deviate from the actual values by an average of $4,080.95 dollars. Before I would start to trust this model to make predictions that I could rely on consistently I would like to see the MAE shrink to at least 1,000. In conclusion, my model could use some work in order to more accurately predict the actual values of houses sales prices.