7.4 Applying Machine Learning

Jason Wilcox

August 6, 2018

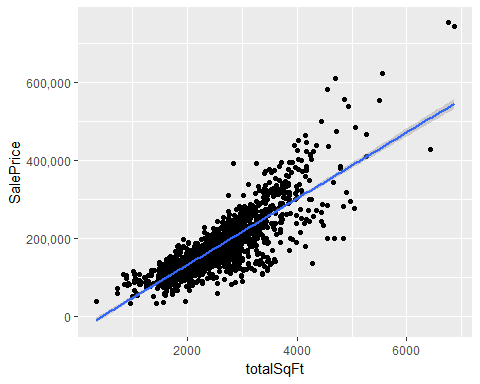
## R Machine Learning to Predict Housing SalePrice

The data frame I’m working with consisted of a few key parts. Each row was a different house, identified by Id and each of these houses had information from 1st floor square footage, kitchen quality, to neighborhood. These were the variables that held all the information and provided tools. The last variable was SalePrice, which was what I was trying to predict using a linear regression model. This is also called a supervised learning because I’m trying to predict something.

Some of the key components were totSqFt, OverallQual, GrLivArea, ExterQual, and kitchenQual. I engineered totSqFt by adding GrLivArea with TotalBsmtSF to get a better, stronger, and more significant variable. As seen below, I graphed totalSqFt vs SalePrice to see the correlation (which I checked with cor()) and got 83%. This shows it’s a significant varaible in relation to SalePrice.

combined$totalSqFt <- combined$GrLivArea + combined$TotalBsmtSF  
cor(combined$SalePrice, combined$totalSqFt, use = "pairwise.complete.obs")

## [1] 0.829042



Using a correlation plot, I was able to find the most important variables, on several of the top ones were those based on the quality of something. OverallQual being of the highest correlation. I made sure include all the variables like this, and used a number system converted from the original ratings of “None” up to “Ex” (for excellent) to 0-5. This also converted several factor variables into numeric, making it a better fit for my linear regression model.

I designed several models, ranging from one using everything, to using almost nothing. I added and removed variables to check for the model’s R^2 value and adjusted R^2 value to adjust up or down. I wanted to have enough variables to give the model enough data to be accurate, but not too much so it was overfit and wouldn’t perform well. I also used the accuracy() command (from forecast) which takes the model and training values and finds various stats such as RMSE, MAE, MAPE etc. I used this to check several different models to try and get a low RMSE as well as MAPE. It appeared as the first model which used everything, was the best model. It had the lowest RMSE and MAPE.

round(accuracy(prediction, trueValues), 3)

## ME RMSE MAE MPE MAPE  
## Test set 0.003 20975.7 14563.54 -0.546 8.749

Using this model, I’ll submit it to the Kaggle competition to check the RMSE of my linear model vs the test data. If there was more time, there would be several things I would try to do to improve the predictive model. I’d feature engineer a few more traits such as age and total bathrooms since their smaller constituants were considered significant. I’d expect the engineered variables may also have an increased significance when combined like the totalSqFt was. I’d also want to use a skewing method to reduce the significant right side skew present, seen in the Q-Q plot below as well as address the lack of homoscedasticity.

