Initial Model

Thomas Kwok

April 18, 2019

training$SalePrice <- log(training$SalePrice)  
test$SalePrice <- log(test$SalePrice)

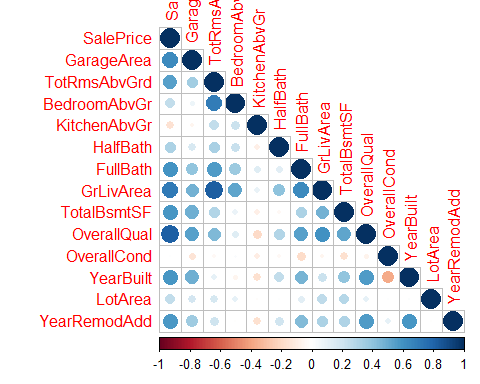
For my Final Project, looking at the Sales Price of housings in Ames Iowa compared to the predictors given from the training data set, I saw some variables that immediately stood out to me. My initial models involved focusing mainly on the regression analysis of numerical data more than on classification data. I looked at Sales Price and compared it to twelve variables which are listed below.

select\_var <- c('SalePrice', 'GarageArea', 'TotRmsAbvGrd', 'BedroomAbvGr','KitchenAbvGr', 'HalfBath', 'FullBath', 'GrLivArea', 'TotalBsmtSF', 'OverallQual', 'OverallCond', 'YearBuilt', 'LotArea', 'YearRemodAdd')  
select\_training <- training[,select\_var]  
df\_train <-na.roughfix(select\_training)

The variables I chose mainly dealt with the size of the houses that were sold, as I looked at the number of total rooms, the number of bedrooms, the number of kitchens, the number of bathrooms and half bathrooms for example. The reason for these variables are that most of the times when people look into houses, as described in shows on the HGTV network they usually talk about number of bedrooms, kitchens, and bathrooms. That made me know that these were immediately good factors. Next I looked into area of the house as that was usually the next big criteria, as the more spacious the house was, the better it should sell. Thus, I looked into Garage Area, Living area, Basement square footage, and lot area. These are good indicators of the size of the house, with a bigger house usually means more expensive house. Lastly, I looked into the condition and quality of the house and year it was built, and year it was remodeled. I thought these were important factors as houses that were newer or remodeled recently would likely sell more as the owners likely remodeled it for a purpose.

After that I ran a correlation plot to see how each variable correlated with Sales Price. The data is listed below.

correlations <- cor(select\_training, use="everything")  
corrplot(correlations, method="circle", type="lower", sig.level = 0.05, insig = "blank")



From the data, it seemed that a lot of the variables I chose did have a correlation with Sales Price. The Living Area, Overall Quality, and Total Rooms seemed to have the most correlated data with Sales Price. Also the Living Area and Total Rooms were correlated with one another, which made sense because as if there are more rooms there should be a bigger area. One thing I found interesting was the little impact that Overall Condition had on Sales Price.

Next I ran three models, a multiple linear regression, a random forest, and a tree to see how the data would compare. For the results, I normalized the Sales Price by taking the log of it, as price was sold in the hundreds of thousands, so the error would look huge if I did not normalize, which would have given a false Root Mean Square error. From my results, I found that the RMSE of the linear model was .1349 which meant that 13.49% of the deviation was created from the linear model.

lm1 <- lm(SalePrice ~ ., data = select\_training)  
prediction <- predict(lm1, test, type="response")  
RMSE(test$SalePrice, prediction)

## [1] 0.1349054

The tree model gave me a RMSE of 20.95% which was higher than my linear model.

set.seed(1234)  
tree <- rpart(select\_training$SalePrice ~ ., data = select\_training)  
prediction2 <- predict(tree, newdata=test)  
RMSE(test$SalePrice, prediction2)

## [1] 0.2095764

And lastly I ran a random forest which gave me the best RMSE which was 12.89% deviation.

rf1 <- randomForest(SalePrice ~., data=select\_training, ntree=500)  
prediction3 <- predict(rf1, newdata=test)  
RMSE(test$SalePrice, prediction3)

## [1] 0.12892

From these initial models, I then wanted to look at how the model sale price compared to the sale price of the test data specifically. So I ran them side by side to compare in the end for each three model.

lm1 <- lm(SalePrice ~ ., data = select\_training)  
pred\_lm1 <- predict(lm1, newdata=test)  
pred\_lm1\_df <- data.frame(pred\_lm1)  
test\_lm1\_df <- cbind(test$SalePrice, pred\_lm1\_df)  
head(test\_lm1\_df)

## test$SalePrice pred\_lm1  
## 1 181500 183035.8  
## 2 143000 164139.3  
## 3 307000 259364.0  
## 4 129900 173082.2  
## 5 144000 116321.0  
## 6 149000 160603.3

set.seed(1234)  
tree <- rpart(select\_training$SalePrice ~ ., data = select\_training)  
pred\_tree <- predict(tree, newdata=test)  
pred\_tree\_df <- data.frame(pred\_tree)  
test\_tree\_df <- cbind(test$SalePrice, pred\_tree\_df)  
head(test\_tree\_df)

## test$SalePrice pred\_tree  
## 1 181500 147015.0  
## 2 143000 112070.5  
## 3 307000 251553.4  
## 4 129900 197237.9  
## 5 144000 112070.5  
## 6 149000 112070.5

rf1 <- randomForest(SalePrice ~., data=select\_training, ntree=500)  
pred\_rf <- predict(rf1, newdata=test)  
pred\_rf\_df <- data.frame(pred\_rf)  
test\_rf\_df <- cbind(test$SalePrice, pred\_rf\_df)  
head(test\_rf\_df)

## test$SalePrice pred\_rf  
## 1 181500 168843.2  
## 2 143000 158963.6  
## 3 307000 252720.1  
## 4 129900 160444.7  
## 5 144000 128253.6  
## 6 149000 144816.8

Looking at six houses for each of them, it seems that all of them could definitely be better in predicting house prices. Mainly because they each overpredicted houses that cost very little (129,900) and severely underpredicted houses that cost a lot (307,000).

My next step would be looking into using some classification data and seeing how I would use them. For example, I feel thatNeighborhood would be a good indicator of Sale Price also, and so would Building Type and Housing Style (though they would probably correlate too much with area of house). Kitchen quality is another variable I could look into and I could also add in some amendities like pool to see if it plays a role or not in the data.

For these initial models, I personally chose my variables instead of using code to see which models are best for the model, so that is something to keep in mind for future modelling.