Capstone Project Milestone Report

Michael A Moran

15 April 2018

Using Data Science to Evaluate Housing Prices

The original focus of this project was to look at how to find economic ‘diamonds in the rough,’ neighborhoods within increasingly expensive cities that are more stable in terms of the price of a one-bedroom apartment than the rate of rent increase in the city overall. The goal was to identify neighborhoods where young professionals could find an apartment with some measure of confidence that they wouldn’t be priced out of the area soon. The Zillow Rent Index was acquired through Kaggle.com, but the data set required a level of tidying beyond my skill level. The data set was created in such a way that the month and year, which needed to be variables of their own, were entered as column names for rents of various neighborhoods of different cities. After consulting with my mentor, Jeff Lipkowitz, we decided that I should try to run a regression problem using the House Prices data set, also available via Kaggle, whose data wrangling needs were more within the scope of the techniques covered in this course.

To prepare the data for regression analysis, three major tasks had to be accomplished: the NA values needed to be converted to zeroes, the categorical variables needed to be converted to dummy variables, and the original categorical column needed to be removed from the data set. Replacing the NA values with zeroes was accomplished simply using the replace function:

newframe <- train %>% replace(is.na(train), 0)

To create the dummy variables, I used the dummy function of the dummies package to convert each categorical variable into dummy form, creating a data frame that I would then attach to the original data frame with the cbind function. Finally, I used list(NULL) to remove the original data column. I selected a group of thirty categorical variables that I thought could affect the final price of a house, then cut that group in half, and proceeded to convert each variable one by one. An example of that code is included below:

daf <- dummy(train$MSSubClass)

daf2 <- cbind(newframe, daf)

daf2[,c("MSSubClass")] <- list(NULL)

This was a time-consuming process however, and inefficient with respect to coding. Jeff suggested a different application of the dummies package in which each variable could be selected by column name in one piece of code and converted to dummy variables without the need to convert each variable individually or code to remove the original variable column:

dafn <- dummy.data.frame(as.data.frame(newframe), names = c("MSSubClass","LotConfig","Condition1","BldgType","HouseStyle","RoofStyle","ExterCond","Foundation","BsmtCond","Heating","KitchenQual","GarageType","PoolQC","Fence","SaleCondition") , sep = ".")

By using the dummy.data.frame function, the excess coding is avoided and the same output is achieved with less time and effort. After writing the cleaned date to an Excel file using the xlsx package for review and further analysis, I uploaded the cleaned data and documentation to GitHub, uploaded the repository link to SpringBoard for Jeff’s review and began to move onto the statistical analysis and data visualization. I started my statistical analysis by running some base R stats over the Sale Price variable, looking at the mean, quartiles, variance, and standard deviation, the output of which is included below:

. > mean(dafn$SalePrice)

[1] 180921.2

> summary(dafn$SalePrice)

Min. 1st Qu. Median Mean 3rd Qu. Max.

34900 129975 163000 180921 214000 755000

> var(dafn$SalePrice)

[1] 6311111264

> sd(dafn$SalePrice)

[1] 79442.5

I generated a base R histogram of Sale Price and boxplots for Sale Price as a function of FullBath, the number of full bathrooms in the house; Neighborhood, the neighborhood within Ames, IA where the house is located; Sale Condition, the financial conditions surrounding the sale; and Condition 1, the proximity of the house to a main road. I generated scatter plots for Sale Price as a function of the total number of bedrooms above ground and as a function of Condition 1, and I generated quartile plots for the sale prices by number of bathrooms and number of bedrooms above ground.

The output plots for these visualizations were plain and unattractive, and involved a lot of coding to generate. Additionally, many of the ways in which I wanted to look at the data proved difficult to represent with visualization in R. On Jeff’s recommendation, I tried using Tableau, which enabled me to upload the cleaned Excel data file and drag and drop variables to generate data visualizations with ease. I selected twelve variables (MSSubClass, the type of dwelling involved; FullBath, the number of full bathrooms; BedroomAbvGr, the number of bedrooms; GarargeType, the location of the garage; BldgType, another dwelling classification; SaleCond, the condition of the sale; OverallCond, the overall condition rating; LotConfig, the lot configuration; BsmtCond, the overall condition of the basement; RoofStyle, the type of roof attached; Condition1, the proximity to a main road or artery; and GrLivArea, the total square footage above ground) from the entire data set that I thought would factor into the final sale price and generated plots of their relationship to the sale price or median sale price depending on the variable involved.

I chose the numbers of bathrooms, the number of bedrooms, the total square footage, and the overall condition of the house for use in regression based on their linear relationship with the sale price. I also chose the type of garage and basement condition variables, as well as MSSubClass (which describes the building with respect to both the type of structure and when it was constructed) as the frequency for the sale price of each category grouped around a distinct mode. Last, I selected the Neighborhood variable given the large differences in median sale price as a function of the neighborhood. To include this in my regression analysis, I will need to create dummy variables for Neighborhood. After consulting with Jeff regarding my output and conclusions, we decided that I should begin regression by increasing the variables, run regression again on the variables that the corrplot function shows as greater than 50% correlated to the sale price. Then, I will run Random Forest and varImpPlot to see which features are most important and use that output to enhance the regression model.