Bonus Unit 4 – Alan Kessler, PREDICT 410, Section 57

1. Can you do a dimension reduction using PCA and make the model more intuitive?

In and of itself, PCA did not necessarily make the model more interpretable for the client. The dimensionality reduction that it accomplishes assists in training the model and reaching convergence, but the individual components still have to be interpreted themselves. This process is time consuming and largely speculative when dealing with the large number of predictors present in this data. Using my 129 final predictors prior to forwards selection, the first two principal components only explain around 12% of the variance in the data. We can speculate about the characteristics and variables that have heavier weights in the component but this does not result in clean breaks. Additionally, depending on the client's goals in implementing this model, using PCA could result in a more complicated implementation due to the scaling and application of factor loadings.

This is not to say that PCA cannot be used to increase interpretability. A different way of incorporating this technique would be to combine it with clustering of variables. The technique used in the SAS VARCLUS procedure is to cluster variables and then extract the first principal component from each cluster as a new feature to use in modeling. The exact algorithm is not available in Python, so I used scikit learn's feature agglomeration class to group variables prior to performing PCA on each cluster. Dummy variables were created first, so using this technique, levels may end up in separate clusters. The resulting principal components did not result in any improvement to interpretability at least in a way that a client with little statistical background would benefit from.

2. Will a cluster analysis result in a realignment of the neighborhoods?

K-means clustering is the approach I took for grouping neighborhoods in the second assignment. When doing this clustering, I made sure to have a sufficient number of observations in each cluster. This resulted in selecting three clusters total. One way to look at the clusters more closely is to observe the relationship between Sale Price and total square footage. Well defined clusters should show that the linear relationship between those two variables should vary by cluster.

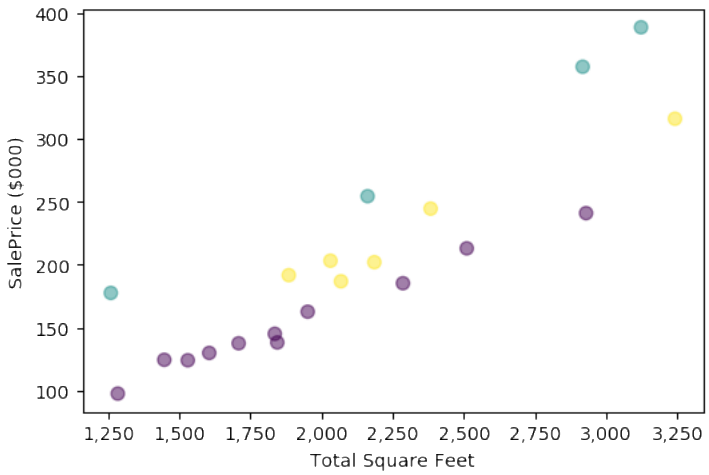

Figure 1 – Sale Price vs. Total Square Feet by Neighborhood Cluster

Figure 1 shows that each cluster shows a linear relationship between square footage and Sale Price but that this relationship varies by cluster. For example, the cluster shown in blue would have a much higher intercept than the cluster shown in purple. This characteristic makes the cluster assignments a significant variable in the model.

3. Will a PCA or FA set of variables provide a model improvement?

In using the first principal component from each cluster, I did not observe an improvement in model fit. This indicates that I should consider using additional components in each cluster. I did observe a slight improvement over the model from the second assignment when applying PCA to all of the features without any clustering. This makes sense because for the second assignment, I stopped the forwards selection when the improvement in fit increased more slowly. That means that there was something to be gained from this additional data.