BA 820 Team Project: Mid-Project Report

Team 8: Huiying Ba, Danting Huang, Ziqin Ma, Jiao Sun, Senbo Zhang

Data Inspection and Cleaning

The dataset we are using contains house prices from Ames, Iowa, along with 79 exploratory variables, and the goal of this project is to find the optimal supervised machine learning model to predict house prices in the "future" test set. A brief glimpse of columns shows there are 43 character variables and 36 numeric variables. We first extracted two scale variables, `OverallQual` and `OverallCond`, leaving the rest numeric vectors to clustering and other techniques. Inside character columns, some variables record the levels of characteristics in scale words (for example, "excellent" vs."fair", "regular" vs. "irregular"). Based on the variable description file in the dataset, we factored these scale words into integers from 0. For those columns with non-scale words but still indicating levels, we manually assigned integers to values following the extent of conditions given. Then we combined all scale variables to a sub-frame for factor analysis (removing one row with missing values). For data cleaning, we substituted all numeric NAs with means of corresponding columns, though we are still figuring out how to deal with missing categorical values. At this time, we leave type variables aside and focus on scales and numbers. This report includes all the methods we tried at this stage. Please refer to "Plot.pdf" to view the related plots in each method.

Clustering: Hierarchical

We started with hierarchical trees. Our first guess in 3 clusters and the cluster plot looks good, while we observed that there are 2 sub-branches each under the top stem. We naturally moved to 4 clusters, in which case the tree gives a better segmentation. We further tried 5 and 6 and decided to stick on **4.**

Clustering: KMeans

We used the total WSS and silhouette methods to identify the optimal number of clusters: the elbow is 4 while the highest silhouette is at 2. We tried from k = 2 and noticed some outliers. After removing these rows, we re-plot the WSS and silhouette scores and observed 2 and 4 as possible k. Both attempts give average silhouette scores of 0.14 with some negative values. In the plots, we can still observe some overlaps near the boundary of clusters, while the sizes are not even. Since we only did clustering on non-scale numeric variables, the result will probably improve if we add scale numeric variables. For now, we choose 4.

Dimension Reduction: PCA

For PCA, we are more accustomed to combining scree plot and eigenvalues > 1 (when the eigenvalue is greater than 1, it means that the component is more stable) approaches to determine the number of Principal components we will use. From the scree plot, we got the first 12 PCA explaining nearly 70% of the variance. The eigenvalue graph shows that dim 1 to 12 have eigenvalues greater than 1. Then we retained the first 12 PCA as our new artificial variables.

Dimension Reduction: EFA

In determining the number of factors, we applied several techniques and the results vary a little. Obvious elbows in the screen plot are 3, 6, 14. Parallel analysis suggests 6, and eigenvalue result (> 0.7) gives 3. We decided to evaluate the 3 factors indexes. After the first `fa`, we removed variables not assigned or assigned to multiple factors. In the second attempt, TFI, RMSR, and RMSEA indexes are excellent. Then we created the FA diagram. We concluded that utility quality, garage, and amenities are the **three** underlying factors related to house prices.

In this exploratory analysis, we tried some unsupervised techniques to find clustering patterns in our training set for different types of variables. The next step is to fit these variables into different analytical techniques to build prediction models for the sales price.