House Price Analysis and Prediction

The Problem

The housing market is going through a series of turbulence in recent years, both in the US and the rest of the world. Thus, understanding the factors that contribute to the changes in house prices is more important than ever before. For house sellers and buyers, it is crucial to know the pricing methodology behind in order to evaluate the sale price. For real estate agencies and developers, the analysis we provided will be able to guide them through the process of housing construction and pricing.

In this housing analysis project, we would like to use unsupervised and supervised machine learning models to predict future house prices based on various housing features. Ideally, we would also be able to provide inferences on influential factors on property prices.

Dataset Inspection and Cleaning

The training dataset we are using contains house prices from Ames, Iowa, along with 79 exploratory variables, with 43 character columns and 36 numeric columns. 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 ordinal levels. For those columns with non-scale words but still indicating levels, we manually assigned integers to values following the levels of conditions given. Then we combined all scale variables to a sub-frame for factor analysis (removing one row with missing values). In terms of missing values, we substituted all numeric NAs with means of corresponding columns, type NAs with a string “N/A”, and scale NAs with a separate level called “NA”.

Our entire working process started from unsupervised techniques to uncover factors beneath scales and reduce column dimensions. Then we would import principal components and category clusters to supervised models to find the best one with minimum validation mean squared error. Combing all techniques we would handle a shell script to predict house prices with features in the model.

Modeling

- KMeans

By running total WSS and silhouette attempts, we tried from 2 clusters and noticed some outliers. After removing these rows, we re-plotted the WSS and silhouette scores, giving 4 clusters with the highest average silhouette scores of 0.15. In the plots, we can still observe some overlaps near the boundary of clusters, while the sizes are not even. Since we only clustered on non-scale numeric variables, we would not parse the result for prediction. The purpose of KMeans for our model is primarily identifying outliers, hence reducing the variance from outliers when running PCA.

- Principal Component Analysis
  In determining the number of PCs, we valued more on the cumulative variance explained rather than eigenvalues > 1. Choosing the first 22 dimensions gives us approximately 93% of the cumulative variance explained, and so we profiled 22 principal components back to our dataset for prediction models.

- Partitioning using Gower Distance
  We would like to reduce the number of character columns to run supervised machine learning models at the end, though in this dataset converting them to dummy variables is not a practical scenario. Most of our character variables have 5 to 6 unique values; if we would convert them all to dummies, the computation cost would intensively increase, even canceling the benefits of PCs. Regarding of our goal to predict prices, we decided to sacrifice interpretability for character variables by grouping them into clusters. In later steps, we would coerce the cluster assignment as dummy variables for supervised techniques.

  Instead of Euclidian and Jacobian distance, we selected Gower distance as the input of clustering. Although the output of this step is to reduce the dimensions of character variables ONLY, we still want to use all values of each observation when assigning clusters. Gower distance can measure the dissimilarities among all types of vectors, and we partitioned the dissimilarity matrix into 2 clusters (with the highest silhouette width). At this step, we created a new dataset with 22 principal components and clusters. Adding `SalePrice` back, we introduced this new `train` dataset to train prediction models.

- Supervised Models
  We divided the original train dataset into two parts: input and validation then tried three models, Lasso Regression, Random Forests, and xgboost. After fitting models in the input dataset and validating in the validation set, we finally decided to use the xgboost, which returned the smallest validation MSE (854,295,790).

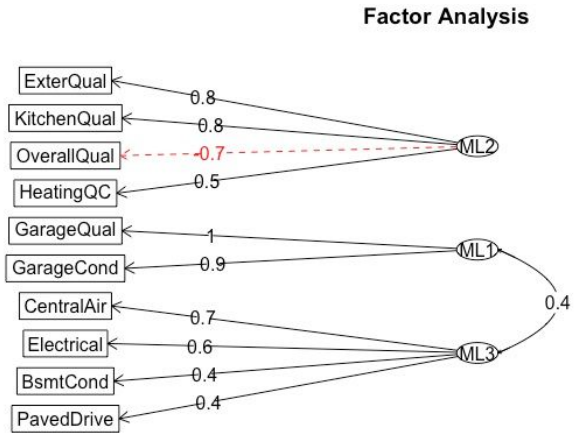
Though we did not choose LASSO regression as our predicting model, we were still interested in which columns in the input dataset are significant to the house price. We checked the coefficients of LASSO regression and noticed that a unit increase `PC1` can boost the price by $25,294. We will expand on that in Price Inference section.

- Outcome
  Finally, we developed a model shell script. By importing a new data frame with property features, we would be able to provide a price estimate. We tried the original test set and successfully got estimates for each listing.

Price Inference

From LASSO regression, we noticed that PC1 has a significant impact on the prediction. It turns out that housing year, size and number of rooms/facilities are the three predictors that affect the housing price most. Housing year includes the above-ground house age as well as garage age. The size is also broken down into living area size and garage size. In addition, since Iowa is cold and snowy during the long winter, people in Iowa also focus a lot on the garage conditions (year built, square feet, number of cars fit) and number of fireplaces. The material of the house is also a crucial factor that affects the price, as masonry material is more stable and has a blocking effect for fire and cold.

Except for the high contribution variables from principal components, we also wanted to find some underlying factors across variables, specifically, columns for scales. Balancing among different techniques, we decided to evaluate 3-factor indexes. After removing overlying and insignificant variables, we ended up with the fa-diagram below:



According to major loadings of each factor, we named MLs as “Garage”, “Qualities” and “Amenities” respectively. Also, We were surprised to observe that the overall quality of the property is negatively correlated to Qualities, as well as the correlation between “Garage” and “Amenities”. With this diagram, we would like to summarize garage, property qualities and amenities as latent constructions driving the motivation of how people evaluate property features. This result can be partially explained by the bad winter conditions in Iowa. When residents cannot fo outside because of heavy snow, they have to live on property necessities (heating, kitchen, electrical power, etc.), so do their cars.

Discussion

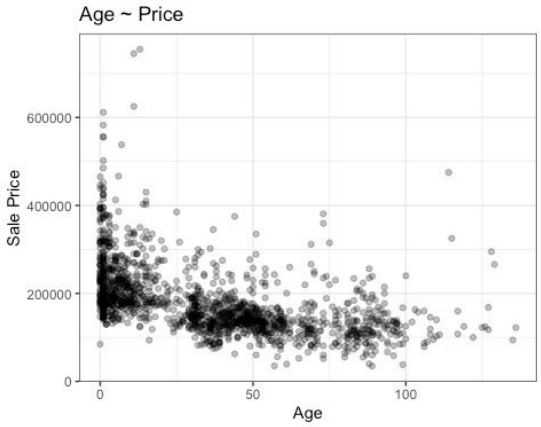
We found some limitations through our working process, especially when dealing with character variables. As mentioned in partitioning, we converted all 34 character columns into one column with 2 cluster assignments, gaining in the speed of regression trees but losing the interpretability of features. If we do want to know which variable is a significant contributor in the dissimilarity matrix, we could remove one character column and re-partition the rest, repeating 34 times to see the difference in optimal k, average in-group silhouette score, and cluster results. It is a trade-off between interpretability and predictability of supervised models; since our primary goal is to predict price, we did not code that model, thus losing some insights into character predictors.

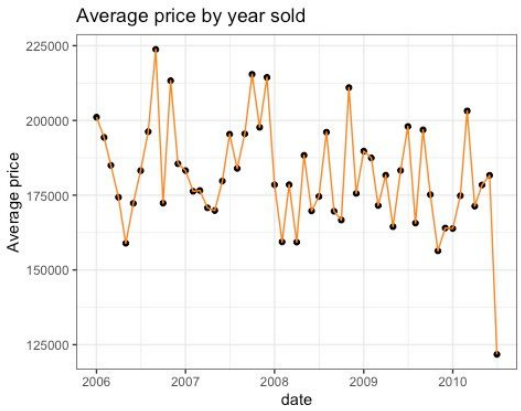
In real-life modeling, joining geographic data can be an approach to gain more interpretability in type columns. A lot of type variables describe the environment features around the house, for example, alley access to the property and paved drive condition. These features are almost the same within clocks/communities; if we have the district information, we could allocate listings to their district by their shared features, consequently removing some variables. Potential problems could still appear in supervised models, but at least we could learn some level of similarities.

In data cleaning, we parsed `YrSold` columns to the age of property when sold by taking the difference between year-sold and year-built. We also mutated month-sold column to quarters in a year as a categorical variable. On the other hand, we can merge year and month into a data vector for supervised models. Below are the price trends by age and date sold.

The left plot shows a decreasing price trend as the age of a house increases, while the right provides a clear seasonality of price variation. In the end, we chose to stick on age to evaluate years for two reasons. First, we agreed that the age of a property is more important than the year sold, given we already know the effect of age on price.

Moreover, we are not confident to run supervised models in time-series, To assure the success of shell script, we chose not to parse year/month as the date this time.





Conclusion

We trained a model on house price dataset with property features. Several machine learning techniques were involved for dimension reduction and output prediction, with some constraints in interpreting significant variables. We also proposed other approaches to aggregate character variables.

The final shell script can apply widely to other regions and multiple facets in the market. Buyers and sellers can be able to look into the features under a number, while real estate agencies can learn the price construction and significant factor.

Reference:

https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#external-measures-for-clustering-validation
https://medium.com/@rumman1988/clustering-categorical-and-numerical-datatype-using-gower-distance-ab89b3aa90d9
https://towardsdatascience.com/hierarchical-clustering-on-categorical-data-in-r-a27e578f2995
https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/
https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3