Group Project

Leslie An, Camille Little, Sara Bolf, Zev Lee

Introduction

https://www.kaggle.com/ruiqurm/lianjia/downloads/lianjia.zip/2 **Data set.)**

The data set comes from Kaggle. The data set is the final price of houses bought in Beijing from 2011 to 2017. There are 318852 entries, and roughly 25 variables. Most variables are quantitative, but there are a few binary variables such as elevator, and categorical data such as district. The data is stored in a single CSV file. strictly speaking, d = 26 and n = 318852. However, since one column is the original site, one column is the trading id, and the other one is unreadable, d = 23 in actuality. Below is a list of all the variables in the raw data.

- url - this is the link and will not be used
- id - this is the trading id and will not be used
- Lng - longitude, quantitative
- Lat - latitude, quantitative
- Cid - community id, categorical
- tradetime - time variable
- DOM - active days on market, quantitative
- followers - quantitative
- totalPrice - quantitative
- Price - quantitative
- square - quantitative
- livingroom - quantitative
- drawingroom - quantitative
- kitchen - quantitative

Learning outcomes:
Author(s): Leslie An, Camille Little, Sara Bolf, Zev Lee

- bathroom - quantitative
- floor - this is unreadable
- buildingType - categorical
- constructionTime - year, quantitative
- renovationCondition - categorical
- buildingStructure - categorical
- ladderRatio - quantitative
- elevator - binary
- fiveYearsProperty - binary
- subway - binary for easy subway access or not
- district - categorical for categories
- communityAverage - continuous variable

In order to process the data, we

- download the data from the url above.
- It is a zip file and we need to unzip it.
- It is a singular csv file so it can be read directly in R.
- To split our data, we think our data is time sensitive, so the first 20 % will be used for DATA1 and the rest will be DATA2. There are 10 stray points that are too early for any good, so we will remove them. They are supposedly in 2004, which is about 7 years before any other transaction.

Goals

I think the data is interesting because it studies the housing prices of one of the biggest metropolitan areas in the world. It is also interesting because the data is taken from a website, which functions similar to a social media website. For every deal, there can be followers that observe the current price. This can also yield some interesting insights about crowd wisdom.

The five preliminary questions we try to answer are

- What are the effects of the social media nature of the website on price?
- How do people's opinion of the house change over time?

- How do traditional housing metrics, such as number of bedrooms, stack up against more nuanced metrics, such as the five year property rule or community?
- Do people value different aspects of a house in different times?
- Throughout all variations, what are the top 5 most important factors of a high price trade?

Question 1 is more straight forward. We can use price as response and see how social media followers affect the price. This is more of a regression problem, but classification is important because we might need to include other elements. We do not see any issues at the moment, but which additional predictors to include can be challenging.

Question 2 has the followers as a response variable, and days on the market as well as house metrics such as bedrooms as predictors. This is a regression, and we try to see how a house that is on the market long compare to houses that is new to the market in terms of followers. The potential issue is that we need to understand the follow mechanic better at this website and see if comparisons can be made.

Question 3 is a regression problem. We try to see the importance of all the variables with price being the response. The most important thing in this question is how important each of the metrics is. While we know a bigger house is likely to sell more, how important is it in terms of decision making? We might use ridge to start, and try elastic net to see how each parameters shrink as we increase penalty.

Question 4 is a regression problem, and we want to see that given that the data is span over 6 years, we try to see if people's preference changed. price is the response again, and all other variables can be used for predicting. The accuracy that we can project DATA2 using DATA1 would be major here. Potential challenges is big with this problem, as there are a lot of possible contributors to inaccuracies, and we might need to compare our results with other divisions of the original data to see more clearly.

Question 5 is a regression problem. We try to see what are the top 5 dictating variables in predicting a house's price. Again, ridge can be used to start. However, lasso could be better as we want to eliminate the lesser variables. If this is not possible, perhaps we can try and see what are the top 5 predictors each year is.