Modelling Competition 2021-2022

Real estate transactions

When an individual or a family wishes to sell their house, they often choose to get help from a real estate agent [1]. The agent helps the family decide on a price for the house called the list price. The family signs a contract (called a listing agreement) with the real estate agent. The contract says that if the agent finds a buyer for the house who will pay the list price, then the family will sell the house at that price. The actual price when the house is sold is called the sale price. Usually the sale price is less than the list price, but not always.

An important part of the real estate agent's job is to help fix the list price. The agent looks carefully at the characteristics of the house, including its size, location, age, the amount of property that comes with the house (called the lot size), the number of bedrooms and bathrooms, and whether the house has various desirable features. In fixing the list price the agent considers the prices of similar houses that have been sold recently.

The real estate agent provides all of this information for a large book that is shared by real estate agents. Each page describes one house, including the characteristics that are important in establishing the list price, along with the list price. Then, real estate agents who are helping families who want to buy a house can use this information to help guide the prospective buyers to the house that might be right for them. The data for this project was gathered from several of these books.

House pricing models

The characteristics of a house strongly affect the sale price. The direction of the effect of most of these characteristics is obvious: larger houses sell for more than smaller houses, houses with more bedrooms and bathrooms sell for more than houses with fewer bedrooms than bathrooms, the presence of a garage tends to raise the sale price of a house, and so on. On the other hand, these features by no means provide a perfect prediction of the sale price of a house. In part, this is because some features are not recorded systematically by real estate agents. For example, whether there is a busy and noisy street in front of the house will matter, whether the house has been kept in good repair is important, and so on. Location usually quite important. For example, the location of the house determines what public school any children living there will attend, and some schools are regarded as better than other schools. Even if there were a complete list of all the features there still would not be a perfect description of the price, because the price is also affected by who buys the house. Houses and buyers are all different. If a family looks at a house that has just been advertised for sale and that house is "just right" for them, it may well sell for the list price or even a little higher. On the other hand, if the house has been for sale for a long time and a family sees that the house will meet their needs only with some changes then it may well sell for quite a bit less than the list price.

[1] This is true in the majority of cases but there is a substantial minority of cases in which individuals or families sell their house without using a real estate agent.

Hedonic pricing model

A hedonic pricing model is a regression equation that relates the actual sale price of the house to its characteristics. Hedonic pricing models have a number of important uses.

Real estate agents find them useful in deciding on the list price of a house. In most communities, owners of houses pay property taxes. The property tax is determined by the market value of the house. Each year the owner of the house gets a tax bill indicating what the local government has decided the house is worth, and from that how much tax must be paid. An important task for local government is to decide how much each house is worth each year. For houses that have been sold in the same year, the market value is the selling price, or quite close to the selling price. But for houses that have not been sold recently, the task is more difficult. (If local government does a poor job in estimating house prices for property taxes, this can become a political problem. House owners may be upset if they think the local government has decided their house is worth more than it actually is.) Sending someone around to look at each house each year and estimate its value is expensive. It is much cheaper to look at the record of the characteristics of each house, and then use a hedonic pricing model to estimate the value of the house. But the model must be pretty good: if the estimate is too low there is a loss of tax revenue, and if it is too high the house owner will complain.

Project Data

The aim is to predict the house prices in King County (Washington State , USA). The variables in the datasets are:

- `id`: Identification number of the property (this variable is irrelevant for the analysis)
- `date`: Date house was sold
- `price`: Price (the prediction target)
- `bedrooms`: Number of Bedrooms/House
- `bathrooms`: Number of Bathrooms/Bedrooms
- `sqft_living`: square footage of the home
- `sqft_lot`: square footage of the lot
- `floors`: Total floors (levels) in house
- `waterfront`: House which has a view to a waterfront
- `view`: Has been viewed
- `condition`: How good the condition is (larger values mean better condition)
- `grade`: overall grade given to the housing unit, based on King County grading system (larger grades are better)

- sqft_above: square footage of house apart from basement
- sqft_basement: square footage of the basement
- yr_built: Built Year
- yr_renovated: Year when house was renovated
- zipcode: zip code
- lat: Latitude coordinate
- long: Longitude coordinate
- sqft_living15: Living room area in 2015 (implies? some renovations) This might or might not have affected the lotsize area
- sqft_lot15: lotSize area in 2015 (implies? some renovations)

1 Assignment

Imagine that you work for a statistical consulting firm that has been asked to create and estimate a hedonic pricing model for King County.

The objective of the model is to predict the selling prices of houses with the same mix characteristics shown in the data set. There are many models that could be set up.

You will need to think about all the technical problems that can arise in regression and decide which variables should be treated as categorical. You also will probably wish to carry out some hypothesis tests as part of your work, check for interactions etc...

1.1 Report

The written report of your project assignment should have, at least:

1. Model-building steps (exploratory analysis, model fitting, variable selection, etc.).
2. Best model/models selected (There are more than one good model, so give reasons to justify your election).
3. Predictive power of your model within sample and out of sample (using some form of cross-validation)
4. Other questions to be assessed:
   - What is the variable that contributes the most to explain the variability in price?, what are the characteristics of the cheapest/most expensive houses?
   - Any other question that you consider interesting

1.2 Evaluation

The grade on the assignment will depend on three things:

1. **Clarity** The report must clearly indicate the steps you follow in your work, including which models are considered. Being able to communicate effectively is extremely important for a researcher. Excellent work is of no value if no other than the investigator can understand it.

2. **Content** The analysis should use the tools developed in the course in an appropriate and correct manner. The report should anticipate questions that a critical reader might ask.

3. **Presentation** The report should not be too long, and should be easy to read, use tables and figures that are relevant and omit unnecessary graphs and/or code in the text.

YOU MUST UPLOAD THE REPORT AND R CODE BY EMAIL BEFORE 19th OF JANUARY AT 23:00