Can we predict house prices using known features of each house and a supervised learning approach?

Florence Galliers

18/10/2020

Background:

House prices are an important part of the economy... Motivate the story... What do we already know? What lessons will the story teach us?

When purchasing a house people are more likely to look only at the number of bedrooms and number of bathrooms, rather than the square foot of the house. This may be because they do not understand how square foot is calculated, or find it difficult to visualise how this translates into the house size.

Objectives:

- Understand which independent variables in the data set can be used to predict house price (the dependent variable)
- Minimize the differences between predicted and actual house prices by using model selection to choose the most accurate model.

Data:

The data I am going to use for this analysis contains house prices and information regarding the features of each house. It has 4600 entries. The original data set downloaded from Kaggle had 17 independent variables, however I felt that 7 of these were not relevant to this analysis and so they have been removed. The dependent variable in this analysis will be the house price (in US dollars).

Although it is from 2014 and USA data, I feel it is a really interesting data set that contains a large amount of information and number of variables. I was not able to find a similar data set from a more recent year or from the UK, and so I am staying with this data set as I think it will produce some interesting results ... or something like this.

Data Preparation:

- There was originally a variable called sqft_basement which gave the size of the basement if present, however a lot of houses did not have basement so I felt it would be more useful to turn this variable from numeric into binary. It now shows 0 = no basement, 1 = basement.
- I also changed the variable year_renovated into a binary variable, as again, not all the houses had been renovated. 0 = not renovated, 1 = renovated.
- dummy variables for condition variable??
- prices are large numbers, change to show in thousands rather than pounds?

```
##     price bed bath sqft_living sqft_total floors condition basement yr_built
## 1  599999   9 4.50        3830       6988    2.5         3        1     1938
## 2  340000   8 2.75        2790       6695    1.0         3        1     1977
## 3 1970000   8 3.50        4440       6480    2.0         5        1     1959
## 4 2280000   7 8.00       13540     307752    3.0         3        1     1999
```

```
## 5  840000   7 4.50        4290      37607    1.5         5        0     1982
## 6  999000   7 4.00        3150      34830    1.0         3        0     1957
##   renovated      city
## 1         1   Seattle
## 2         1 Shoreline
## 3         0   Seattle
## 4         0   Redmond
## 5         0  Issaquah
## 6         1  Bellevue
```

Methods:

As house price is a continuous variable I have taken a supervised learning approach and will be using regression to look at the relationship between house price and features of each house.

Review approaches tried or considered

Summary of final approach and justification of why this approach was chosen:

Results:

Summary of major results, graphs, diagnostic outputs

Strictly relevant to the objectives

Must include a link to the Github repository containing a fully reproducible and documented analysis Reported in scientific style.

Conclusions:

~1 paragraph

Literature Cited:

3-5 peer reviewed references

-
-
-
-