Predicting Boston House Prices

The volatility in price fluctuation has caused an issue for the real-estate and property owners in Boston. Anyone wanting to buy a house here is perplexed in choosing between multitudes of criteria and its effect on the final price of the house. In order to help these new-home buyers of Boston, our company has collected data from 506 different locations and categorized them under 14 different headers based on crime rate, land utilization, presence of non-retail business, age of the buildings, air quality, number of rooms, proximity to water bodies and highways, demography, property tax rates and median value of the homes in that location. Based on the collection of this sample data, we shall be able to predict and better suggest our clients the type of homes that they can purchase as per their requirements.

Initially we map the correlation characteristics to understand how they interact and influence one another independently. Based on the summary (mean and median values) and the co-relations between the entities, we model an initial linear regression model to predict the median cost of house in any locality. Number of rooms in a house and the population status shows higher linearity with our median house cost. The findings show that the age of the building and presence of non-retail industries do not have any relation with the cost. Thus, we do multiple iterations and generate a model wherein we do not consider the earlier characteristics (and black population and the zonal land), but instead the interaction between number of rooms, population status and accessibility to highway. Based on this number, we are able to explain around 84% of the variation in our response i.e. the median cost in a locality.

Following equation is generated to predict the house cost:

log(Median House Value) = 1.954 + 0.377 (number of rooms) + 0.103 (accessibility to highway) + 0.06 (lower status of population) – 0.8(nitrogen oxide levels) – 0.03(distance to employment centers) - 0.03 (pupil teacher ratio) – 0.01[(number of rooms) (lower status of population) + (number of rooms) (accessibility to highway)] – 0.002 (lower status of population) (accessibility to highway)

Note: We use the natural log for regression model in R as it reduces the skewness in model and makes it resemble normal distribution.

One interesting thing to note is that there is a significant negative impact of air quality on the price of the house. For all other similar criteria, if a tenant wants to move to a location with minimum nitrogen oxide levels, then on an average they will have to pay an excess of 7000 USD. Similarly there are multiple tradeoffs as well. Suppose there is a family with 2 kids, it will be their priority to live in safer neighborhood with good air quality and in a better school district, i.e. where pupil to teach ratio is low. In such a condition, if they are looking for a 3 room apartment near the highway, then it costs them between 8000 and 22000 USD with a mean price of 14000, whereas if they choose to drive a bit extra and stay further away, then they will get it in between 2500 and 15,000 USD.

Boston being one of the financial hubs in the US, let us consider a baseline scenario which would be preferred by most of the working professionals. We assume they require a 2 bedroom apartment, closer to the employment centers, moderate crime rates with mean air quality. Proximity to school districts and population status wouldn't be major criteria in their decision making. The estimate price would be around 10,500 USD and with a range in between 10,200 and 12000 USD. Overall it has been observed that air quality and lstat are the two main attributes in our model, which are the driving factors for the top 5% houses in various location.

In conclusion we have developed a model which would help home buyers make a better decision towards not only selecting a particular location to buy a new house, but will also help them negotiate the correct price with the seller based on their preferences