Julian Murillo

CPSC 392-02

Chapman University

Dataset Selection and Questions to Consider

Boston Housing Data Source:
https://www.kaggle.com/kyasar/boston-housing#boston_housing.csv

General Dataset Information:

This Dataset contains 14 features (13 continuous and 1 binary) and 506 observations/records. The reason I chose this dataset is because I have always been interested in housing and housing prices. The dataset does not contain any missing values and is ideal for regression analysis as well as classification type methods.

Attribute Information:

1.) Crim: per capita crime rate by town
2.) Zn: proportion of residential land zoned for lots over 25,000 sq.ft.
3.) indus: proportion of non-retail business acres per town
4.) Chas : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5.) Nox: nitric oxides concentration (parts per 10 million)
6.) Rm: average number of rooms per housing
7.) Age: proportion of owner-occupied units built prior to 1940
8.) Dis: weighted distances to five Boston employment centres
9.) Rad: index of accessibility to radial highways
10.) Tax: full-value property-tax rate per $10,000
11.) Ptratio: pupil-teacher ratio by town
12.) Black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13.) Lstat: % lower status of the population
14.) medv: Median value of owner-occupied homes in $1000's

Questions from a Business Perspective:

There are many groups/individuals in the North-East that would be eager to derive information from raw data concerning housing. This is because geographic and demographic data is powerful in deciding where to operate as a new or expanding business. Finding which neighborhoods, boroughs, or districts have the most disposable income will drastically increase the likelihood of revenue growth. For instance, if I were an investment firm looking to open a new office for financial consulting, I would like to know as precisely as possible which

neighborhoods or urban centers within Boston have the things such as low crime-rates (less risk), higher property values (more appeal from high earners), closer proximities to dense work areas (foot-traffic), and a higher-class citizens (more exposure to individuals with greater disposable income). These four attributes alone contribute immensely to client traffic/growth, better talent applying to your firm, increases in current employee salaries and bonuses, and further overall growth of the firm. A few questions we might have regarding the data itself and what it might mean intuitively are:

1. How might we go about determining whether or not a Boston home is expensive or not with the data we have available? Could we group (cluster) to gather specific information on a specific neighborhood?
2. Are median values of homes and the proportion of lower-income individuals inversely related? What would we expect?
3. What are the most important variables when it comes to determining the median property value? How would we determine this? Why should we open an office there?
4. In the city of Boston, is there a higher crime rate when the median values of homes are lower? Does this mean higher risk in those areas for our firm?
5. Does property tax in Boston decrease if the home is in a bad area? Why? Because of the area, or because of the number of median property value?
6. Does the proportion of non-retail business acres per town matter as to whether or not a home is priced above average? How might this determine the competition in the area?

7. Does the number of rooms per house increase or decrease with an increase in the parent-teacher ratio in Boston? Do we move into an area with more children so that we can manage more trusts/savings accounts?
8. How might the age of the homes determine property value? Will we have to set up our offices in a similar building style to appeal to potential clients?

Honorable Mentions:

California housing Data:
https://www.kaggle.com/camnugent/california-housing-prices/version/1

Mashable News Popularity Dataset:
http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

1994 Census on Adults:
http://archive.ics.uci.edu/ml/datasets/adult

Audit/Risky Business Dataset:
http://archive.ics.uci.edu/ml/datasets/Audit+Data