Predictions on California House Values and Incomes

Alex Beeston, Thomas Brower

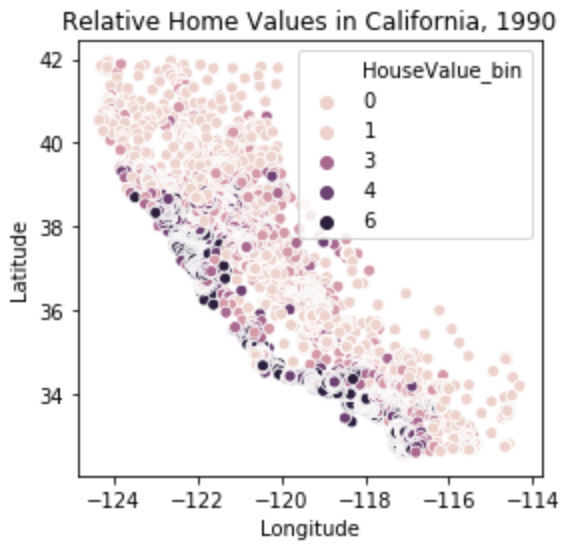
Introduction

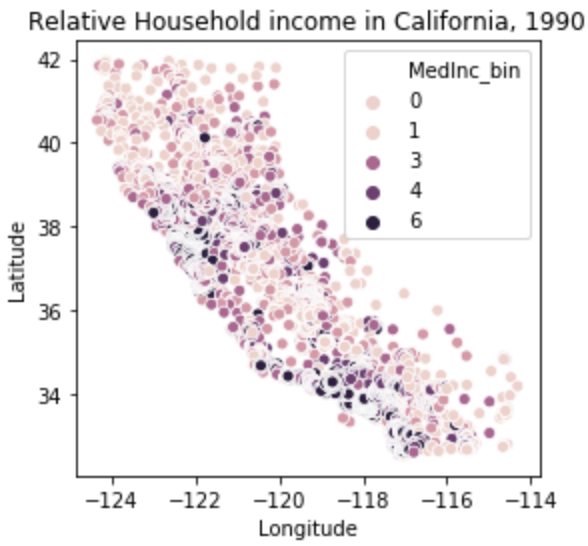
Dataset

The dataset has nine attributes that describe residential block groups in California. A block group is the "smallest geographical unit for which the U.S. Census Bureau publishes sample data," and it "typically has a population of 600 to 3,000 people" [1]. The nine attributes that describe a block of houses are:

- Median household income
- Median house age
- Average number of rooms
- Average number of bedrooms
- Block population
- Average house occupancy
- Median house value
- Latitude
- Longitude

The median house value and median household income were binned into five categories according to their percentile rankings such that data points in the 20th percentile were placed in category one, data points between the 20th and 40th percentile were placed in category two, etc. The maps below show the relative home values and household incomes binned into these five categories. The data was obtained from StatLib Repository. Note that, unfortunately, units were not documented for median household income and median house value. However, because the goal is to classify these metrics into categories based on percentile scores, the units are not necessary.





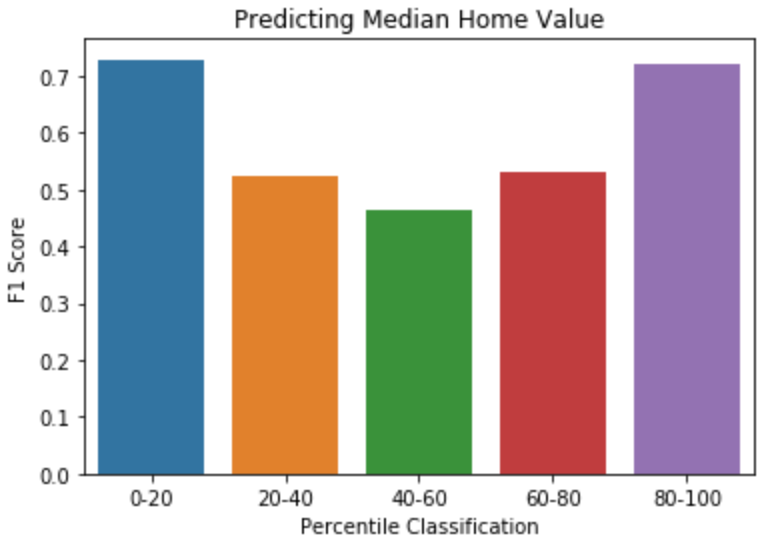
Analysis Techniques

Decision tree classification and neural networks were used to predict the response variables.

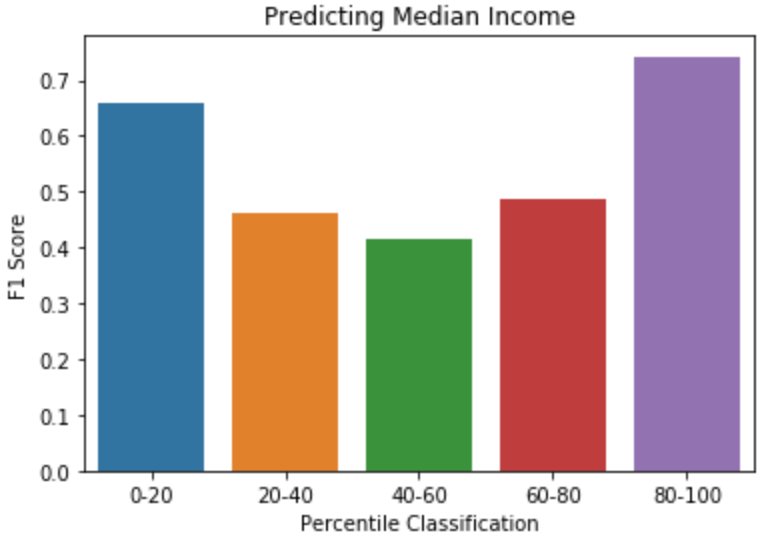
Results - Decision Trees

The decision tree method was used to classify the median home value and median household income. For both response variables, the tree method performed best on classifying values in the first and fifth categories. The tree classified median house values in the first and fifth categories with F1 scores of 0.728 and 0.722, respectively, and median household incomes in the first and fifth categories with F1 scores of 0.659 and 0.743, respectively. The average F1 score for classifying median house values and median household incomes was 0.593 and 0.553, respectively. Hence, the tree performed slightly better when classifying median house values.

When classifying the median house value, the tree first separates on the median income with a threshold of 4.071. The tree then separates on median income again, with thresholds of 2.518 and 6.323 for blocks with median incomes less than and greater than 4.071, respectively. The gini numbers on the leaf vertices range from 0 to about 0.75, with many of the leaves containing gini scores of around 0.2



When classifying the median household income, the tree first separates on the average number of rooms with a threshold of 5.83 rooms, much to the authors' surprise. The tree then separates on median house value, with thresholds of 1.251 and 2.176 for blocks with less than and greater than 5.83 rooms, respectively. The gini numbers on the leaf vertices range from 0 to about 0.75, with many of the leaves containing gini scores of 0.



It is concluded that using decision trees is a moderately reliable method for classifying the relative median incomes and home values of census blocks in California in 1990.

References

[1] *User's guide to Scikit-learn sample data set. Available at* https://scikit-learn.org/stable/datasets/index.html#california-housing-dataset