Predictions on California House Values and Incomes

Alex Beeston, Thomas Brower

Introduction

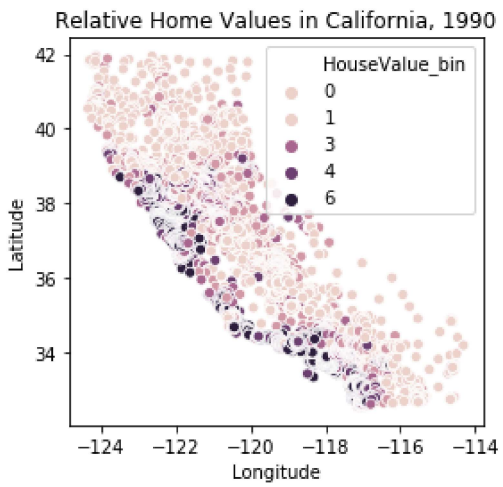
Our analysis on the California Housing data set will help inform businesses and government how the housing market has changed within the last 30 years. This information is incredibly valuable as it helps the government determine the needs of their constituents and informs businesses where their efforts should be focused to gain the most profit. Presentation slides available here: https://docs.google.com/presentation/d/1tVwqIGQ1YUDQIr3Dqtc4-hkZ0xGM8fVe6UAzAPLP9DE/edit#slide=id.g722ada4d44_0_12

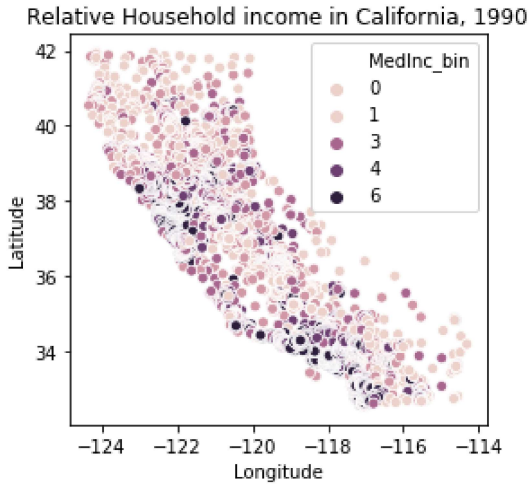
Dataset

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

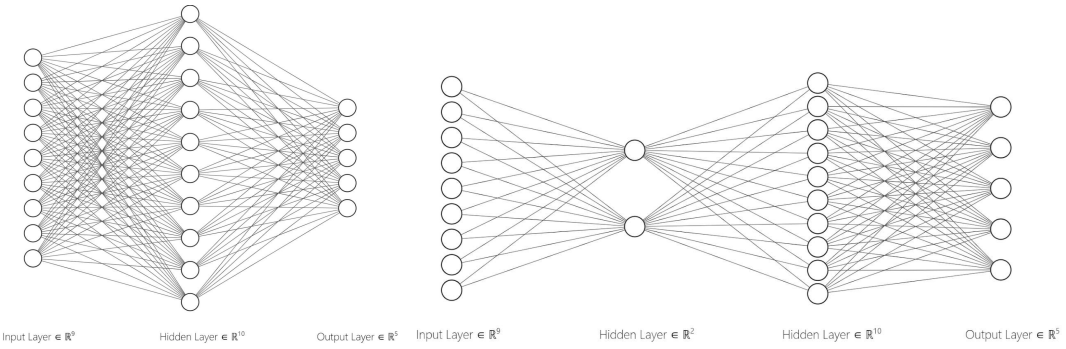
The median house value and median household income were binned into five categories according to their percentile rankings such that data points in the 20th percentile were placed in category one, data points between the 20th and 40th percentile were placed in category two, etc. The maps below show the 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

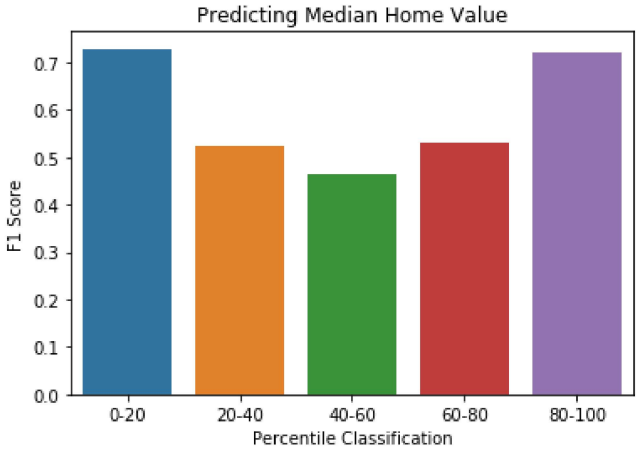
Decision tree classification and neural networks were used to predict the response variables. We used a grid search on our neural networks to find the best number of nodes for a layer. This method was computationally expensive so a different method of fine tuning would need to be explored in the future if further tuning of the parameters was desired. Finally the max iteration was extended from the standard 200 iterations to 500 iterations in hopes of each network being able to converge. Due to the data and network construction, however, most of the networks did still not converge after 500 iterations. Since a variety of constructed networks were considered in our grid search, here are two examples that were used.



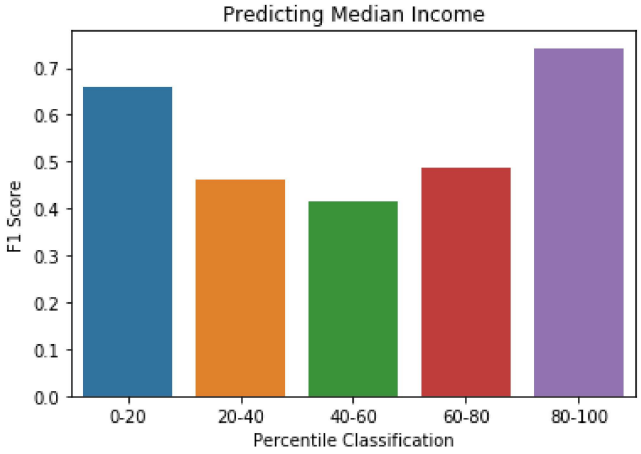
Results - Decision Trees

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.

Results - Neural Networks

The following is a table where the f scores are compared to different structures of a neural network with respect to each of the five percentiles.

Amount of layers	Number of Nodes	F-score
1	[239, 250, 198, 172, 156]	[**0.776,** 0.60, 0.50, 0.55, 0.76]
1	[150, 150, 95, 140, 70]	[**0.781**, 0.55, 0.50, 0.55, 0.76]
2	[(2, 253), (2, 227), (2, 243), (2, 258), (2, 222)]	[0.716, 0.466, 0.458, 0.52, **0.73**]
2	[(105, 140), (105, 55), (65, 75), (40, 140), (135, 10)]	[0.73, 0.51, 0.48, 0.53, **0.74**]

The results received here were similar to the results we received using a tree classification, but it is clear that a neural network can give better results given time and attention to tuning the parameters. Had more time been allowed further tuning would have been explored with different activation functions and weight parameters. Due to the computationally expensive nature of neural networks, we were unable to run the same models when considering the median income like the tree classification. This shows that tree classification can be more valuable than neural networks when time is an issue and when accuracy is not as important as getting results.

Conclusion

Although we did not get amazing results with a lot of our percentiles, our information could still be beneficial for both government and business entities. Their first and fifth percentile did very well in both of our models which represent the least expensive and most expensive houses in california. Having the information about the highest and lowest houses in the market is vital information as everything else has to, quite literally, fall in between what information you've gained. Data from a more recent year would help confirm or deny some of these models and be worthwhile in exploring.

References

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