Housing Price Prediction using Regression Techniques

Reyden Harris, Ian Abeyta, Mishal Raza-Taimuri

Problem Statement

It has been estimated as of July 1, 2022 that the state of California has the highest population of people in the United States with nearly 39 million residents saturated in a land mass of 155 thousand square miles and the median value of owner occupied housing in California is estimated at $659,300 (Census.gov, 2024). Compared to other locations in the United States, California is one of the most expensive places to live.

The current state of housing in California is 2.5 times higher than the national average, and housing prices throughout California are rising faster than the average income of its residents (Berkeley 2023). There are a number of regulatory and economic factors that influence high rates of housing prices, however there exists a lack of clear understanding, not only, on how these factors correlate to housing prices but also, how these will impact future trends. California has a lot to offer to its residents including, breath-taking sceneries and diversity of cultures hence, it is imperative to predict housing prices for logistic analysis as well as awareness for future home buyers in California.

Objectives

The goal of this project is to develop a machine learning model that can accurately predict housing prices based on various features. A linear regression algorithm was utilized to make predictions on the price of houses by using the California housing dataset from scikit-learn. A high accuracy model is developed which presents the projected outcomes for families interested in purchasing a home in the state of California.

Data Description & Preprocessing

The dataset consists of 8 different features including: median Income of homeowner, age of the home, average number of rooms, population of the city in which the home resides, average number of household members, latitude and longitude (*Figure 1*).

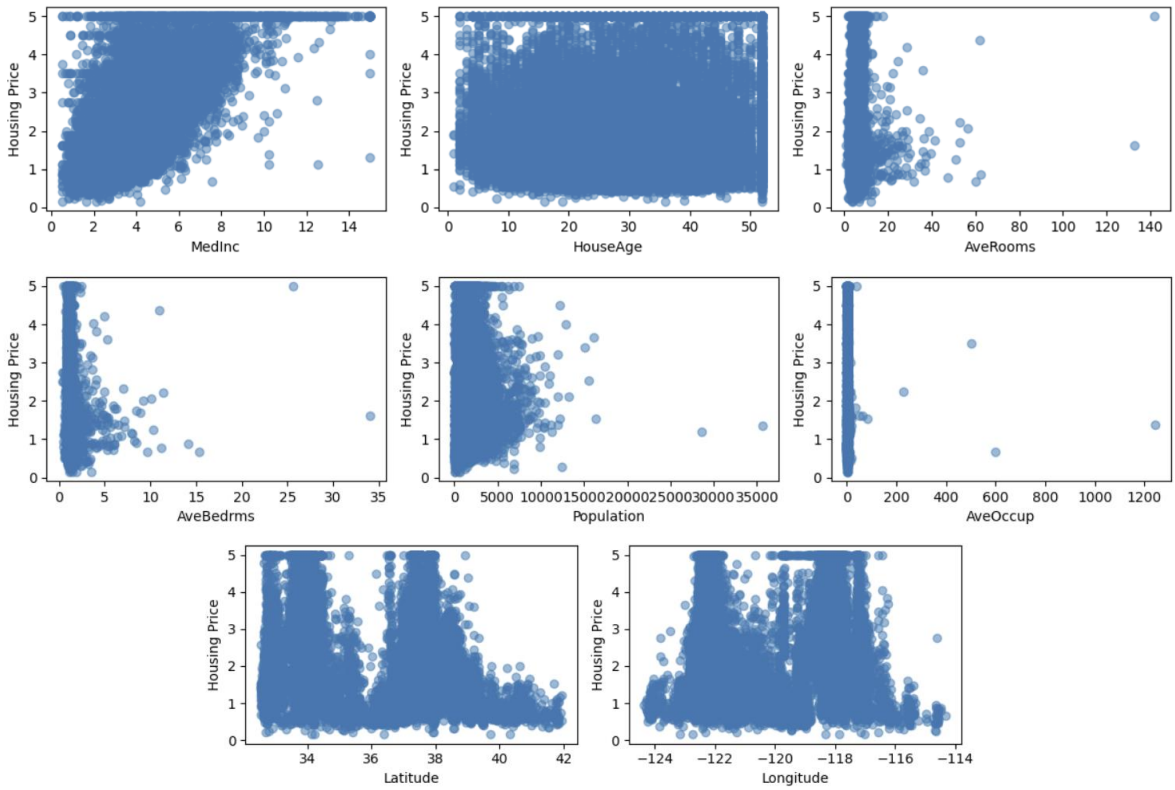

Figure 1: Dataset features

The target variable for this dataset is the median house value with a minimum value of 0.149 and a maximum value of 5.00001. There are a total of 20,640 samples within this dataset, and to preprocess this dataset we are going to check for and handle missing or NULL values in the dataset. Any missing NULLs were replaced with the median value for that feature.

Additionally, another preprocessing technique we will showcase is outlier handling. Using an alpha value of 0.05, we will exclude the 1st and 99th percentile values, this will promote a dataset that does not have an unbalanced influence with data skewed to the left or the right.

Feature Engineering

For feature engineering, the VectorAssembler from the pyspark.ml.feature library was employed. All the features were combined into a single vector to demonstrate the linear regression. From this point, we assembled the vectors, selected the features and ‘target’ (our median home price), and pulled those into a dataframe for testing, training and validation.

Model Building & Evaluation

To build this model, we split the data into 3 divisions, training data, testing data and validation data. The linear regression model was trained on the training data-split by setting the parameters as the features vector created and the target column from the dataset, which is the dependent variable.

The model was evaluated by the Root Mean Squared Error [RMSE] on the test data. A score closer to 0 is an indication that the predicted value and the actual value are extremely close. Then, k-fold cross validation was performed on this dataset, and the RegressionEvaluator module was imported to estimate the root mean squared error metrics.

Results & Analysis

Evaluation by the Root Mean Squared Error [RMSE] on the test data obtained a score of 0.718, which is a great score. The RegressionEvaluator module used to estimate the root mean squared error metrics provided an average value of 0.756.

The results of this regression are excellent and it can be confirmed with great confidence that the predicted home value will be very similar to the actual home value (*Figure 2*). A scatter plot visualization was generated to represent the distribution of our data across a perfect prediction line. This shows a positive correlation between the claimed predictions and reality. The results were also validated with a separate code which utilized the scikit-learn feature library and results obtained were identical.

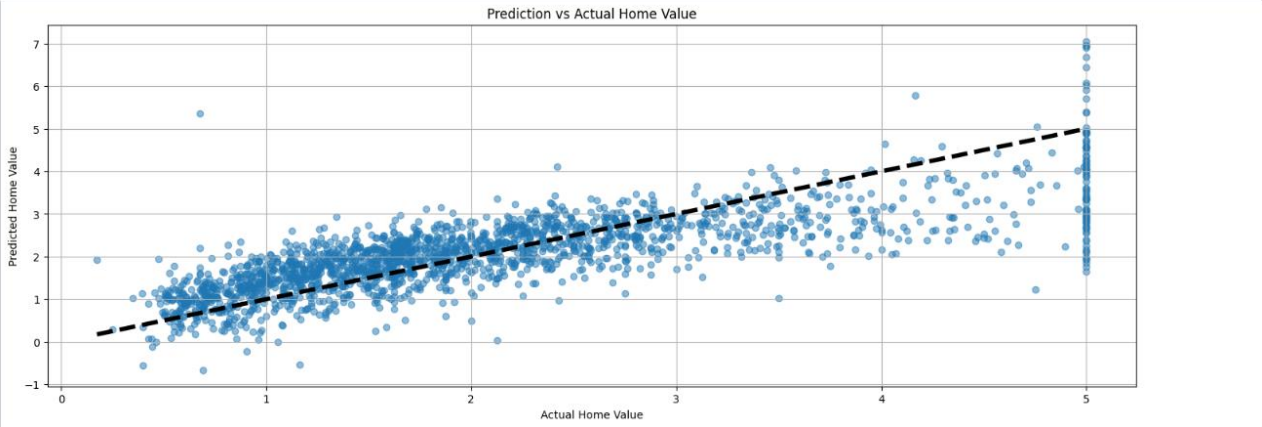

Figure 2: Regression model outcome

Conclusion

Using California as representation of increasing housing prices and population, a machine learning model was developed that accurately predicted housing prices based on various features. A linear regression algorithm was utilized to make predictions on the price of houses, using the California housing dataset from scikit-learn.

After evaluating the provided dataset and the linear regression model, a very strong positive correlation is depicted. If all variables remain constant, we can expect that housing prices throughout California will continue to rise. This information combined with the latest data from the United States Census Bureau shows that, families at or below the national average household income range may not have the ability to purchase a home in California, in the immediate future, as projected by the designed model.

References

4th Quarter 2023 National Market Summary. (2023). *U.S. Department of Housing and Urban Development*. Office of Policy Development and Research.

Berkeley's Economic Review (2023). California Housing Crisis: a Necessary Crisis to Solve Amidst High Interest Rates.

Jauregui, A., Fan, Q., Curry, J. (2023). House Price Capitalization of Stormwater Retention Basins: Evidence from Fresno-Clovis Metropolitan Area in California. Journal of Real Estate Finance and Economics. 67(4). https://doi-org.libproxy.library.unt.edu/10.1007/s11146-021-09860-2

Means, T., Stringham, E. (2012) Unintended or Intended Consequences? The Effect of Below–market Housing Mandates on Housing Markets in California. *Journal of Public Finance and Public Choice*. 30(1-3) DOI: https://doi-org.libproxy.library.unt.edu/10.1332/251569212X15664519360461

Michieka, Nyakundi M; Gearhart, Richard S, III; Ampatzidis, Yiannis. (2021). Oil Prices, the Housing Market, and Spillover Effects: Evidence from California's Central Valley. *Journal of Housing Research*; Washington 30(1), 77-97. DOI:10.1080/10527001.2021.1901546

Pinto, R. (2023). Housing market trends: analyzing housing dynamics, evaluating mortgage risk, and understanding the impact of filtering on affordability. *Business Economics*. 58(3) 149-157. DOI: https://doi-org.libproxy.library.unt.edu/10.1057/s11369-023-00328-0

ReydenHarrisPy. (n.d.). *Reydenharrispy/HousingPricePredictionML: Using a linear regression model to predict housing prices in California*. GitHub. https://github.com/ReydenHarrisPy/HousingPricePredictionML

United States Census Bureau (2024). QuickFacts: California; Florida; New York; Texas. https://www.census.gov/quickfacts/fact/table/CA,FL,NY,TX/PST045222