Real Estate Investment & California Price Prediction

Thip Rattanavilay

DSC680 - Fall 2021

https://github.com/thiprattanavilay/DSC680

Which Domain?

I have selected to dive into Real Estate investment in the state that I currently live in which is California. The median price of a single-family home in California dipped to $811,170 in July, a minor shift from the month before in what could be the start of a cooling off period in a booming market. I do have a few investments but yet to investment in California because of the high cost of living. This project will help me decide if I should invest in real estate here in California and my prediction will inform me where and when to invest into this real estate market.

Here I have listed 10 references that I will be gathering my information from to help with this project.

1. https://www.kaggle.com/camnugent/california-housing-prices
2. https://www.kaggle.com/fedesoriano/california-housing-prices-data-extra-features
3. https://www.car.org/
4. https://www.redfin.com/city/11203/CA/Los-Angeles?utm_source=google&utm_medium=ppc&utm_term=kwd-10537751&utm_content=424422438903&utm_campaign=1014388&gclid=Cj0KCQjwt-6LBhDlARIsAIPRQcJpr3EYbARLHkKSJcMiExh6AKW0JfAS9gpDUKH9GQK8vU1aJT3rSRQaAs1mEALw_wcB
5. https://www.zillow.com/homes/92805_rb/
6. https://www.kaggle.com/alexgude/california-traffic-collision-data-from-switrs
7. https://developers.google.com/machine-learning/crash-course/california-housing-data-description
8. https://www.car.org/marketdata/data
9. https://www.car.org/en/marketdata/data/countysalesactivity

10. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html
11. https://www.noradarealestate.com/blog/california-housing-market/

Which Data?

Data collected by the California Association of Realtors similarly show that "sales remain solid but the momentum continues to increase but soon to slow." Seventy percent of homes still sold above the asking price and listings continue "flying off the shelf," according to the association. I will also pull datasets from Kaggle and Zillow to add to my finding. This should provide more than enough data to start with.

Variable and Description

- **longitude** (signed numeric - float) Longitude value for the block in California, USA
- **latitude** (numeric - float) Latitude value for the block in California, USA
- **housing_median_age** (numeric - int) Median age of the house in the block
- **total_rooms** (numeric - int) Count of the total number of rooms (excluding bedrooms) in all houses in the block
- **total_bedrooms** (numeric - float) Count of the total number of bedrooms in all houses in the block
- **population** (numeric - int) Count of the total number of population in the block
- **households** (numeric - int) Count of the total number of households in the block
- **median_income** (numeric - float) Median of the total household income of all the houses in the block
- **ocean_proximity** (numeric - categorical) Type of the landscape of the block
- **Unique Values** ‘'NEAR BAY', 'OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND' ›
- **Median_house_value** (numeric - int) ∫†Median of the household prices of all the houses in the block

Dataset Size: 20640 rows x 10 columns

Dataset Links

https://www.kaggle.com/camnugent/california-housing-prices

https://www.kaggle.com/fedesoriano/california-housing-prices-data-extra-features

https://www.car.org/marketdata/data

Research Questions? Benefits? Why analyzes these data?

California Census Data which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The dataset also serves as an input for project scoping and tries to specify the functional and nonfunctional requirements for it.

Some of the key factors to consider when analyzing a real estate market include:

- Property types with the greatest demand
- Most active agents and investors
- Who the local home wholesalers are
- Percentage of renter-occupied households
- Housing inventory stock
- Where the biggest employers are located

1. Build a model of housing prices to predict median house values in California using the. provided dataset.

2. Train the model to learn from the data to predict the median housing price in any district, given all the other metrics.

3. Predict housing prices based on median_income and plot the regression chart for it.

4. What area is safe to live with low cost?

5. Is the cost of living more expensive in northern California?

6. Is the cos of living cheaper in the central valley?

7. Southern Californian less expensive than Northern California?

8. What city or county has the worst crimes?

9. What is more in demand, Single family, Multi family, apartment or condos?

10. What is the market so high in California?

What Method?

I will need to do exploratory data analysis and do the following manipulations on data.

* Creating new features
* Removing outliers
* Transforming skewed features
* Checking for multicollinearity

Training machine learning algorithms: Here, I have trained various machine learning algorithms like:

Linear Regression

Linear regression models assume that the relationship between a dependent continuous variable Y and one or more explanatory (independent) variables X is linear (that is, a straight line). It's used to predict values within a continuous range (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). Linear regression models can be divided into two main types

Simple Linear Regression

Simple linear regression uses a traditional slope-intercept form, where a and b are the coefficients that we try to "learn" and produce the most accurate predictions. X represents our input data and Y is our prediction.

Multivariable Regression

A more complex, multi-variable linear equation might look like this, where w represents the coefficients or weights, our model will try to learn.

Mean Squared Error (MSE) Cost Function

The MSE measures how much the average model predictions vary from the correct values. The number is higher when the model is performing "bad" on our training data.

One Half Mean Squared Error (OHMSE)

We will apply a small modification to the MSE — multiply by 1/2 so when we take the derivative, the 2s cancel out

Potential Issues?

One obvious issue that comes to mind is how to deal with wildly different results from using different approaches that might have many limitations.

I do fear that I won't complete the coding portion effectively with the given timeline. The data found on Kaggle are both numerous and high quality. There's no missing or fails data that needs to be removed, and some of the noise where a wait time will take longer and in the course of a couple minutes is smoothed by the prediction model.

Concluding Remarks

The project also aims at building a model of housing prices in California using the California census data. The data has metrics such as the population, median income, median housing price, and so on for each block group in California. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics. Districts or block groups are the smallest geographical units for which the US Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). There are 20,640 districts in the project dataset. I will take this data to help with my future investments or for my audiance.