Housing Prices in California: Trying to predict prices of housing in CA

JUSTIN CHU, University of California- San Diego, USA

Fig. 1. Panorama of Los Angeles

Housing prices in California have always been higher than in other states in the nation but within the state, there is a variance in housing prices depending on the region as well. Today, we look into if we can use a regression model to predict the prices of houses for sale during the first six months of 2021 in the state of California.

Additional Key Words and Phrases: datasets, regression, machine learning

ACM Reference Format:

Justin Chu. 2022. Housing Prices in California: Trying to predict prices of housing in CA. 1, 1 (March 2022), 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

When I was looking for data sets that intrigued me, I stumbled upon a data set about housing prices in California for the first half of 2021. Being born and raised in LA County, I had noticed that the prices of housing had increased by quite a bit over the years. For me, I thought that prices of housing was unpredictable but I thought to myself, why not try it? So I used that data set that I found and decided to try to make a regression model to try to predict the prices of houses based on the information from the data set. The question that I strove to answer was how can I best predict the price of houses for sale using the information given in the data set?

2 DATA SET OVERVIEW

The data set that I used was found on Kaggle at https://www.kaggle.com/yellowj4acket/real-estate-california. It has 39 columns and 31,238 unique rows, with each row representing a house on the real estate market. The columns includes

Author's address: Justin Chu, j8chu@ucsd.edu, University of California- San Diego, 9500 Gilman Dr., La Jolla, California, USA, 92093.

Manuscript submitted to ACM

information about each listing such as price, type of listing, the location of the listing (city, county, longitude, latitude, etc.), and the type of home.

3 EXPLORATORY DATA ANALYSIS

With the data set imported, I used pandas to clean up the data set and explore what information in the data set could and could not be used.

3.1 Cleaning the data

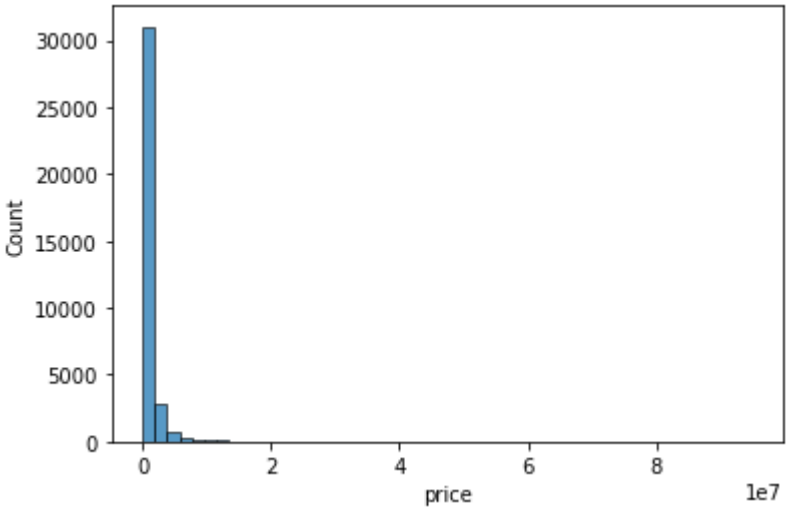
When I first looked at the columns of the data set, I instantly noticed there were a bunch of columns that I could not use or did not need. I removed those columns, which were the index column (as there is no need for a separate column for the indexes), id, stateId, countyId, cityId, country, longitude, latitude, currency, and time.

Later, after exploring the data a bit, I noticed that there were a lot of things within the columns that were not useful for answering my question. For example, I removed all of the empty lots from the data set, as the question was looking at prices of houses, not those of the land. I also narrowed down my data set so that I was only looking at the houses that were for sale, not the houses that had been removed from the market, for instance.

Inside of the data, there were also values that seemed to be incorrect. For example, there was a bunch of houses that were listed as having 0 or 1 square feet of area but were listed at high prices, and that would surely mess with the final prediction.

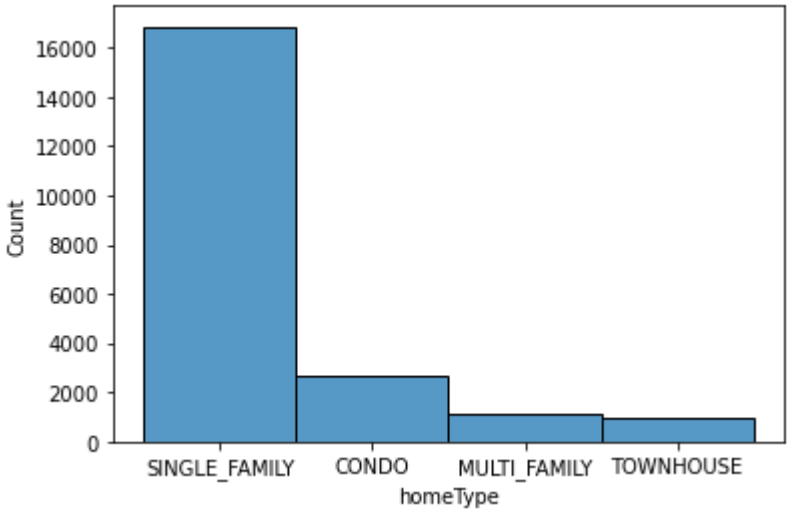
3.2 Exploring the data

After the initial cleaning of the data set, I decided to take a quick look at the prices to see what the spread looked like and this was the result.



As we can see in the graph above, there are a lot on the left end of the graph but the graph also seems to have a really large outlier so I found the max value, which is 9.5 e +07 or $ 95,000,000. Next, I looked at the county column and figured that would be a factor in the price of houses as well, so I did a one hot encoding of the column so that I could

use it in the regression model. I did the same for home type as well so that I could use that in the regression model too. This graph below shows the amount of each kind of house in the data set after it is cleaned.

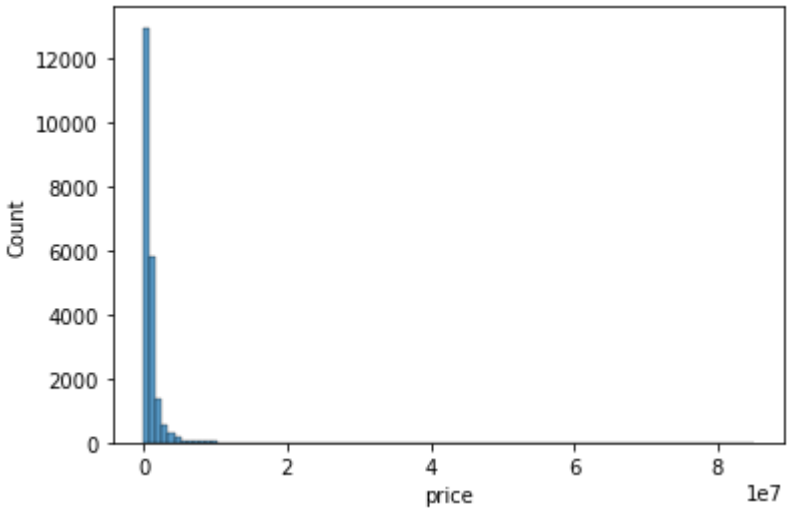


When looking at the city column, I decided that I should split it between the major cities (Los Angeles, San Francisco, San Jose, San Diego, Sacramento) and not major cities since prices in the cities are usually really high.

With the year built column, there were a lot of strange numbers as well, where there were strange years like 9999 and 0. When looking deeper into the ones with the year built as 0, we can see that most of the listings that have the label are lots which usually means that there is not a house built on there, hence why the year built is not real. On the other hand, the rows with the year built labeled as 9999 has a mix of single family and multi family homes, meaning there were a bunch of mistakes in the data set.

For the living area column, I saw a bunch of listings that had square footage of 0 or 1 square feet, which did not make sense as I had already removed all of the lots and it is quite hard to build or sell a 0 square foot or 1 square foot house for thousands to millions of dollars. Thus I had them removed as those are also most likely a case of an error in the data set.

I also took value counts of a lot of the columns that I felt was useful to see if I could use them to make the regression model be better. Finally, I looked at the price column one last time to see how the distribution changed after I did all of the cleaning.



4 PREDICTIVE TASK AND MODEL

I first split my model into training and test sets, with 0.75 as the training size, using the features large cities, the year built, the One Hot encoding of home type, if it is a new construction, living area, the number of bedrooms and bathrooms, and if it had a pool.

I chose these feature because in my mind, they made sense. Usually, the living area (square footage of the house), the year the house was built the number of bedrooms and bathrooms, and pool are huge indicators of the amount of money a house will sell for. However, since I did go house hunting with my parents, I know that things like cities and the type of house also influence the value of the house in terms of price quite a lot.

I used a few baselines to determine if my model, which I decided would be a Decision Tree Regressor, was the best course of action to predict the price. Personally, I chose the Decision Tree Regressor because I believe that the decision tree would be able to adjust more for each feature that is being used for the regression. For the baseline models, I used Linear Regression, which gave a mean absolute error of 589643.1663394197, Support Vector Machine, which gave a mean absolute error of 655780.1080130442, and Gaussian Naive Bayes, which gave a mean absolute error of 764618.0499907425. The model I used, the Decision Tree Regressor, gave a mean absolute error of 586737.7238474357. Thus, the model that I chose was the most accurate at predicting the price of housing. I used mean absolute error as the metric to determine accuracy because I was looking at the documentation for sklearn and it stated that it was the best for regression.

5 LITERATURE

This task has been accomplished by people before as there are always articles about housing prices increasing or decreasing as a prediction. But that is pretty general and is not exactly targeting the question that I was tackling. When I was looking at what other people had tried to accomplish, I saw that most of them were looking at general trends for housing prices as a whole but no one had tried to do what I was trying to. However, that makes sense because each house is different and there is no real way to use historical data to try to predict how much a house will cost accurately. I believe that this is due to a multitude of factors including the presentation of the house, the real estate agent pitching

the house, monetary inflation, clientele, and changes in neighborhood, amongst other things. For example, I have a friend who lives in my hometown that was trying to sell their house. However, due to the neighborhood changing because of the new light rail line that was installed in the area and the decrease of foreign buyers due to the pandemic, the house was not able to get the same price that it would have gotten if sold just a year before. I think what I could have done instead of doing a regression like this is to do a classification style problem, where I would tackle it the same way but have price bins instead of trying to get the exact price. I think that with a better data set with less mistakes and better surveying, such as making the assessment of levels be more uniform (i.e. 1, 2, 3, 4+ vs one-story, two, 3, 3-4 +), the model that I was trying to achieve would be much more successful.

6 RESULTS

As we can see above, the models actually don't seem to do quite well but the decision tree regressor does the best out of all of the models. I think overall, the data set does not have enough proper information to be able to properly predict the prices of housing in California. There are often a lot of other circumstances that increase or decrease the value of property, such as school district, accessibility to transportation (i.e. public transport, freeways/highways), and quality of the house. I know this from first hand experience as the neighborhood I used to live in had housing prices skyrocket because the school district had a notoriously good reputation. Even within the city, there were certain parts of it with housing prices that were a lot lower because it was very far from the main part of the town (i.e. the school, restaurants, etc.) and even though the size of the houses were comparable, the prices were quite a bit lower. I think that the feature that I had added were important for many people when it comes to buying houses but the data set did not have enough information for me to create a better model. I think there was also a huge problem with the data set because there are a lot of feature that were important that had mistakes in them, meaning even though I cut out the 0 square feet houses, there could be other incorrect numbers which would lead to incorrect predictions on the model. I think what we can gather from this is that the prediction models can have bad numbers when it comes to evaluation time but it doesn't mean that it is necessarily a problem with the model. As we can see, there are a lot of external circumstances and issues with the data set that could cause the model to not come out as well as we hoped. However, given those circumstances, the model did perform beyond expectation and with better data, I firmly believe that the model would be much more successful.