Californian Housing Price Estimations Using Random Forest Regression

Kent Kenyon
Texas State University
kck46@txstate.edu

Beckett Allen
Texas State University
ra1205@txstate.edu

Abstract

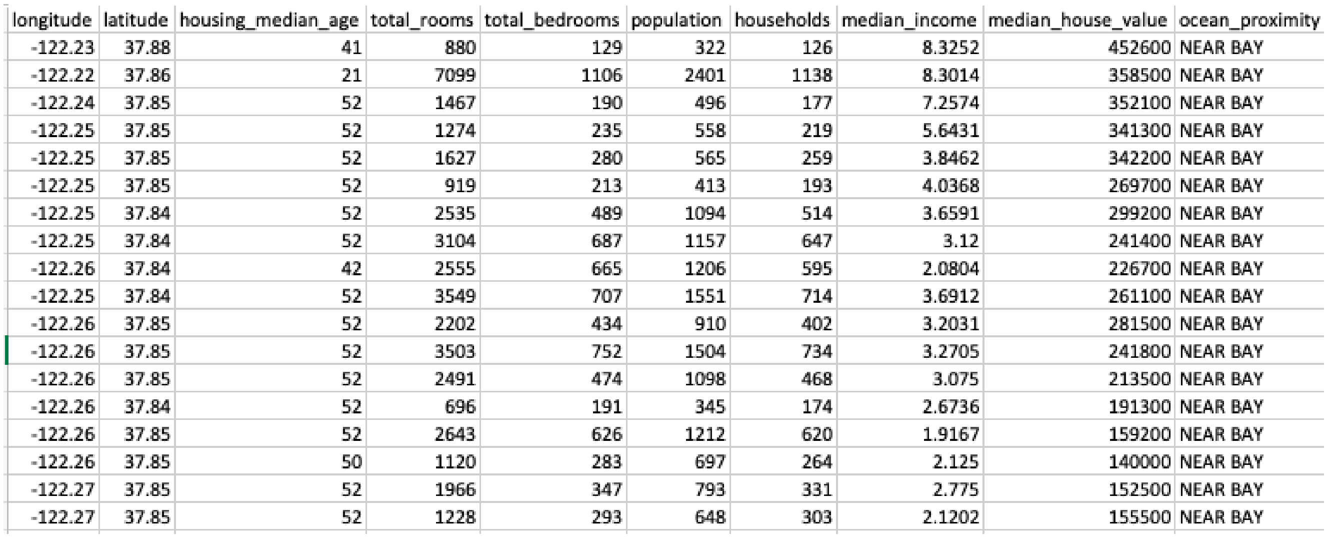
Most people buy a home at some point in their life and it can be hard to find a home that is valued at a reasonable price. In this paper, we discuss the applications of machine learning on property valuation on a data set of 20,640 Californian homes. We used a Random Forest Regression to analyze our data. Our regression takes in many physical aspects of a home, yet our research was more interested in how geographical location could affect a properties value. With our initial interest in geographical price analysis through Zip-Codes and not being able to find reasonable data, we included geographical parameters in our regression including longitude, latitude, and the proximity of the home to the ocean. With this, we were able to successfully analyze the most important factors in buying a home, the usefulness of machine learning in home valuation, and derive some theoretical properties from our random forest regression analysis.

1. Introduction

Buying a home is easily one of the most important and stressful investments a person will make in their lifetime. With buying a home being part of the "American Dream", most people will have to buy, or even sell, a home at some point in their lifetime. It is important that buyers and sellers are well informed on their properties valuation to ensure that they are not spending most of their finances on an overpriced property.

The United States home-ownership rate reached "67.9 percent in the second quarter of 2020" [1] with millennials, also known as Generation Y, increasingly entering the housing market. Millenials have had to face many "major obstacles to home-ownership such as college costs that are growing faster than wages and ever-mounting student loan balances." [3] With these challenges, the average age to buy a home as increased to the age of 32, which sits right in the middle of the 72.1 million millenials who are between the ages of 23 to 39. [3] With this anticipated increase home buyers in the marketplace, property valuation is extremely important to today's homeowner.

With this influx in possible home buyers and the importance of freedom seen in the United States society that comes from buying a home, we looked at the most important factors that motivate a person to buy a home. It all comes down to the famous quote by Harold Samuel, "location, location, location." When buying a home, "81 percent of Americans agree that they value the location of their home over the size, with all generations in agreement." [1] Americans also reported that the top factors "that influenced neighborhood choice were the quality of the neighborhood (58percent, convenience to job (44percent) and the overall af-



longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY
-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY
-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY
-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR BAY
-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR BAY
-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR BAY
-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR BAY
-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR BAY
-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR BAY
-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR BAY
-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR BAY

Figure 1. Small sample of our data set for Californian housing.

fordability of the homes (41percent).” [1]

Keying in on location and unable to find the appropriate data for a Zip-Code analysis, we turned toward California, the most populous state in the United States. We found a California housing data set that included the geographical components we were interested in. These geographical components are important as ”the most unaffordable area to live (when factoring in median income) is the Los Angeles-Long Beach-Anaheim, California area, [where] only 12.75 percent of properties are affordable to the median household.” [1] We assessed the geographical components of the longitude, the latitude, and the proximity of the house to the ocean for our regression analysis of Californian housing prices.

2. Data

The data set that we used comes from Kaggle, which is a subsidiary of Google LLC and is one of the largest data science communities. The data set includes the data of 20,640 homes located throughout a California district based off of a government census. Each entry of the data set includes: longitude of the home, latitude of the home, the housing median age, the total number of rooms, the total number of bedrooms, the population size of the area, the number of households in the area, the median income of residents, the median house value, and the proximity of the home to the ocean.

This data set was chosen mainly due to the geographical parameters included in the data set, as well as the being the best fit for the type of data we were looking to model. The data allows us to get some analysis in terms of location and includes many of the main factors that are used in determining a houses valuation.

3. Problem Description

This problem is simply a regression problem where we are taking in many given parameters through our input to estimate a houses value as our output. The main parameters that we want to address are the longitude, latitude, and ocean proximity with how these geographical parameters can impact a houses price. We will look at the importance of each of these parameters to determine if geographical location plays a significant factor in a houses valuation. To implement our machine learning algorithm we used scikit learn, as well as pandas to handle and process the data we used. We used a random forest regression on our data to predict house values and compare these values to the real cost of each home.

Our workflow is as follows:

1. Read our data in from the pre-processed housing.csv file from our GitHub repository that gives us our important parameters.

2. Visualize the data by plotting the housing price density, comparing parameters to the housing price, and plotting the parameters.

3. Split the data into training and testing sets to train and test the regression.

4. Get the importance of the parameters and tune the parameters and hyperparameters of the random forest regression to fit our data.

5. Print the results of the random forest regression.

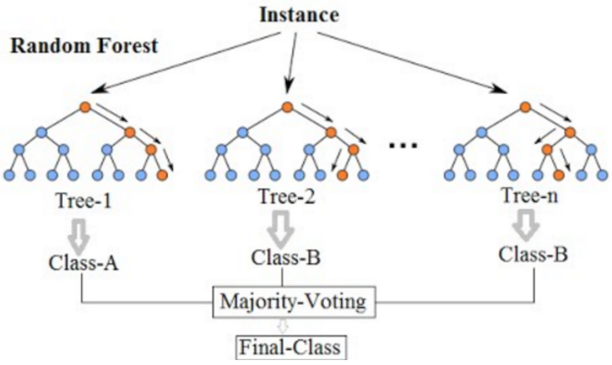


Figure 2. Random Forest Trees

4. Methodology

For the purpose of this project, we did not look into what the house price was listed as, only what the house was valued at through the median house value. By valuing a house off of the parameters, we are able to use a regression model to best fit this data. To better understand machine learning we researched and used a random forest regression model for our regression. We did not account for any extraneous factors in this regression analysis, such as large-scale economic impact, but we should still be able to successfully model a random forest regression.

We used a random forest regression due to its ease of use because we were most interested in making sure we could visualize the project and understand the many aspects of machine learning. We chose the random forest regression as it is a readily available ensemble learning regression that avoids many mistakes, such as overfitting, through combining the predictions of multiple machine learning algorithms. The random forest algorithm does this through the creation of trees, that can be adjusted in the hyperparameters, and creates an average of all these predictions through aggregation in a supervised learning approach.

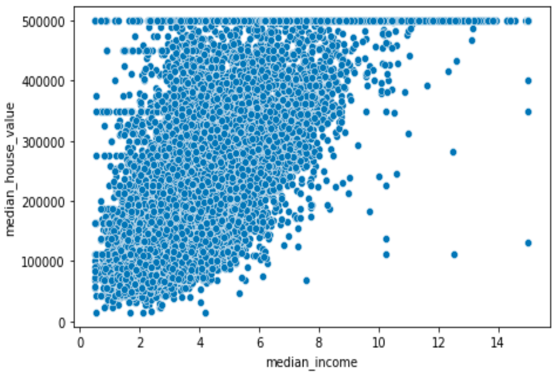


Figure 3. Example Visualisation of Data for Median Income of an Area Compared to Median House Value

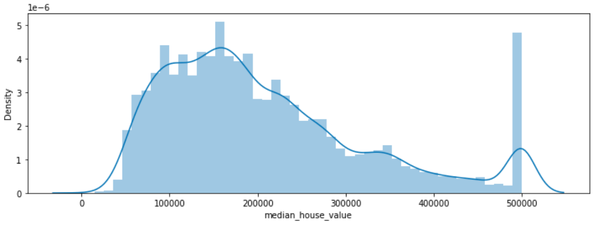


Figure 4. The Density of the Median House Value

Before we could begin using the random forest regression, we first sought to visualize the data we had imported from our housing.csv file. We visualized a smaller subset of the data in some instances to better view the data. This is a good first step in implementing a robust regression model

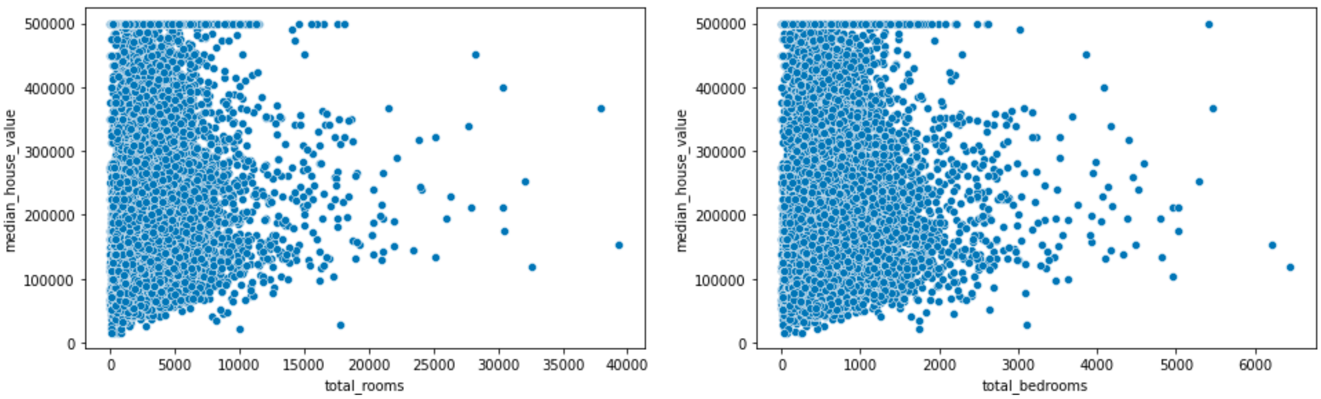


Figure 5. This side-by-side of Total Rooms(left) and Total Bedrooms(right) helps to demonstrate the overfitting that might occur in a normal regression if not weighted accordingly as these graphs are very similar.

so we understand the data we are working with before we decide to train it.

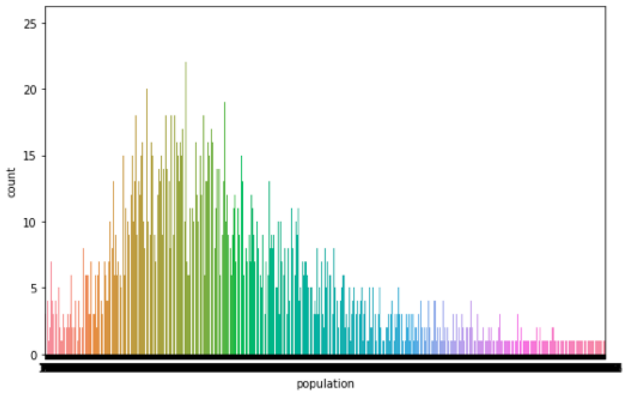


Figure 6. Population count of each area

After looking at the visualizations (Figures 3-6) of our data and drawing further conclusions from these graphs, we needed to ensure the data was clean (e.g. eliminating data where features were missing). Seeing the data was cleaned, we then split the data into two parts. We had 80 percent of the data set to train our regression made and then we allowed the other 20 percent to test the accuracy of the regression model that was made from the training data. This will show that the random forest algorithm is a great tool to use in estimating a houses value, thus giving home buyers the ability to determine if a house is over or undervalued. The non-linear nature of a random forest regression algorithm gives it a higher prowess over linear algorithms, but we must keep in mind that random forest can't extrapolate outside the given data due to the nature of decision trees.

The regression was trained with around 16,512 entries. We trained the regression with all of the included features as we believed that the random tree regression would account for any overfitting that may occur in other regression algorithms, such as a linear regression, due to it being an ensemble machine learning regression. In fitting our trained random forest regression model to the training set, we found that the random forest regression model fit the training set with a precision of 97.65 percent, thus fitting the data much better than a linear regression model.

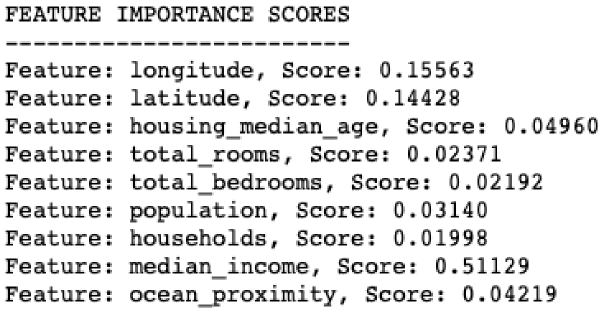

```
FEATURE IMPORTANCE SCORES
-------------------------
Feature: longitude, Score: 0.15563
Feature: latitude, Score: 0.14428
Feature: housing_median_age, Score: 0.04960
Feature: total_rooms, Score: 0.02371
Feature: total_bedrooms, Score: 0.02192
Feature: population, Score: 0.03140
Feature: households, Score: 0.01998
Feature: median_income, Score: 0.51129
Feature: ocean_proximity, Score: 0.04219
```

Figure 7. Importance Scores of our Features

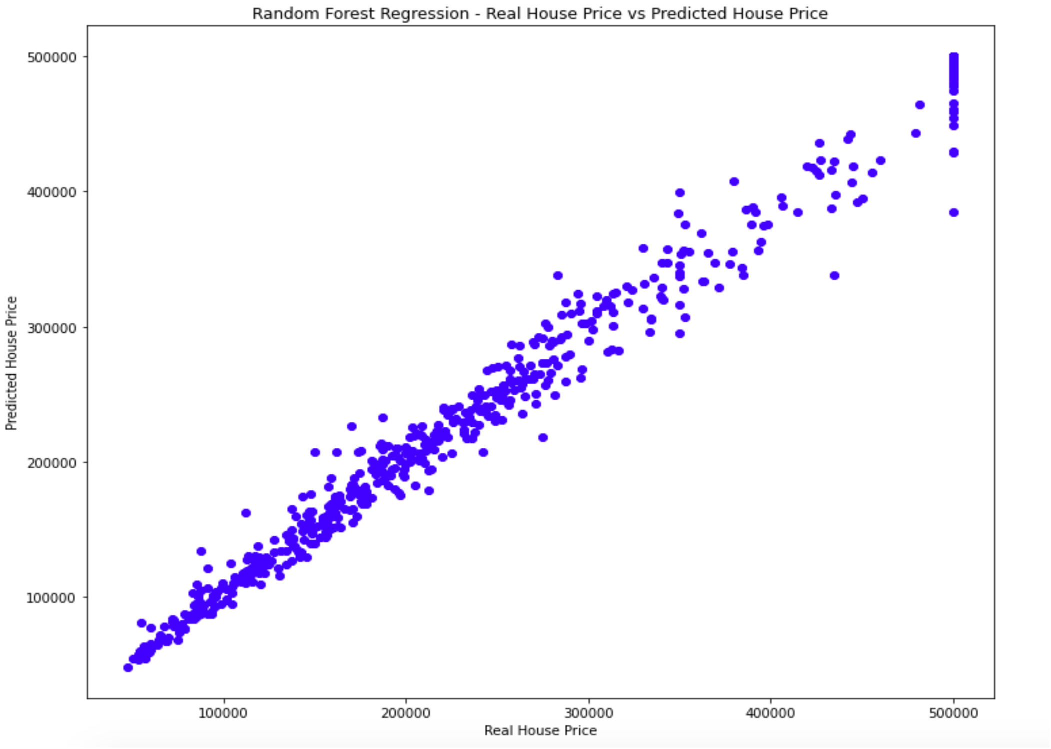


Figure 8. Predicted Value of a Home Vs. Real Value of a Home

We continually tuned and adjusted the features included and the hyperparameters of the random forest regression to get this precision on the trained data, but our results came more from the tested data. We used importance scores from each feature to help with our understanding of how each feature impacted the random forest regression as seen in Figure 7. We determined that if the random forest regression model could get an accuracy for the housing valuation above 75 percent on the testing data, then we could consider the random forest regression to be an accurate way to evaluate housing prices and set a standard for housing prices. This would make understanding if a house is overvalued or undervalued much easier within the housing marketplace by assigning values to these houses through given inputs.

5. Results

For our testing set, we included 4,128 entries to determine how well the trained regression model would fit to other housing valuations within our data set. We got the trained random forest regression we made in our methods to accurately depict the test data with 82.07 precision, which would tell us that the random forest regression model was a good fit for analyzing this data and is capable of being used to appropriately assign a homes value. We used the regression to assign values to the houses and compared these predicted values to the real values of the houses in the test data set as seen by a small subset in Figure 8. We found that the random forest regression made would slightly increase with only including the parameters of Median Income, Longitude, and Latitude, but these changes were not substantial. We also were able to understand more and notice how regularization occurs through penalizing the selection of newer features, over features having similar gain that were already selected. The process of hyperparameter tuning was also very helpful in understanding fully how the random forest regression worked as we tuned the depth of the trees made through max_depth, determined how many trees to make before the averaging of predictions through n_estimators, and many other hyperparameters that were important in forming the regression. We found that making the number of trees too small would lead to overfitting the data.

6. Conclusions and Future Work

The primary intent of this project was to explore the use of a known machine learning algorithm to predict a homes value through geographical factors. Although we were not able to properly use Zip-Codes the way we initially intended at the start of our research process, we successfully applied a machine learning algorithm to housing data with a prediction of the housing valuation. On a large scale, we could repeat this process of using the random forest regression algorithm on housing in a wide variety of Zip-Codes if we were given the data. The ability to reliably estimate and assign a houses value has many possible implications. For example, it would allow buyers/sellers to accurately appraise the value of their home without the need of for third parties to be involved. By using machine learning we can further dive into problems like this to cut out the time, money, and energy put into tasks faced by most Americans.

In this project, we solely used the random forest regression algorithm to predict a houses value, yet we understand that the random forest regression algorithm is a very powerful algorithm that can be used on a variety of machine learning tasks and random forest can also be used for classification problems. By using machine learning algorithms to solve common problems of the average American, such as buying a home, we can effectively save time and money to put towards other aspects of our society.

It is possible to expand this project in future works to include more features that impact a homes price like the state taxes, number of neighbors, style of home, etc. We can expand this process to include homes from a multitude of Zip-Codes and apply this on a global scale to accurately predict home values around the world. This work could also be extended by adding a wider range and variety of houses as random forest regression is not the best with extrapolation. This ease of knowing a homes value would significantly minimize the risk in investment of housing and reduce stress in the home buying process. A robust housing valuation estimator would minimize the number of buyers overpaying for a home, minimize the number of homeowners underselling a home, and further maximize the efficiency of the home buying process.

Along with any technology that impacts the way we interact, there are challenges that arise from assuring that the home valuation model stays fair. It is important that we recognize that this regression model is supposed to supplement our decision on whether a house has a good valuation as each home buyer will need to adjust the home they buy to their own qualitative analysis of the situation based on personal preferences and needs. Furthermore, it may be beneficial to a home owner to adjust and train this model again to fit their own personal needs that match their criterion for what they desire in a home.

7. Bonus: Theoretical Properties

7.1. Property 1

We can see from Figure 5 that the random forest regression is able to ignore input variables that aren't independent of one another as we can tell that there is a strong correlation between the number of total rooms in a house and the number of total bedrooms in a house. This is much different from other algorithms like a linear regression that requires the independent variables to have no multicollinearity. This derives the mean decrease impurity where the data set can have two or more correlated features and any of them can be used as the predictor with no preference over the others. Once one of the correlated features is used, it reduces the importance of the others since the impurity they can remove has already been accounted for by the first feature.

7.2. Property 2

Random forest is well known for being able to handle thousands of input variables without deletion. Due to this, theoretically, we could add significantly more features to the random forest algorithm and it would not change the algorithm too drastically as long as we have a lot of trees being created and have set good hyperparameters. If we only had one tree in the random forest regression this would most likely overfit the data as this would be the same as only having a single decision tree. Random forest regressions are able to reduce variance due to selecting a random subset of features and training on different samples of the data. The random forest regression would still be able to apply an importance score to each of the features effectively as the trees do have that randomization through bagging and have ensemble output with low bias since the trees are not correlated. Through this feature importance we are also able to look at the most important factors that attribute to a houses valuation, which would be important to homeowners trying to market their home to potential buyers. We cannot fully determine which factors are most important though if we leave factors that correlate, so to determine factors with the most importance we would need to make sure each was independent. This can be seen in our results section as we were able to get nearly the same results by adding all the features in comparison to just having three of the features.

7.3. Property 3

We assume that the model does not accurately describe data outside of the given range seen in the regression model due to the inability of random forest regression to accurately extend the model into other valuations outside of the given housing value range. With random forest incapable of extrapolating values outside of the given range of housing data, we can derive the idea that with more values adding to the range of the model we should expect to see an increase in the range of values that the model is able to accurately give an appraisal value to. This was accurately seen in the experimentation for the project as the more we increased the size of the training data, thus giving the training data a better range, the better our trained regression was able to assign good valuations to the houses. We can see this slightly in Figure 8 as the density on the top-right of the graph is lower and more spread out, yet this can be seen more by giving input variables outside the range covered.

8. GitHub

Below is the link to our GitHub repository. Here we have posted the proposal, the project code, the housing.csv data, and any other parts of the project we deemed necessary.

```
https://git.txstate.edu/ra1205/
CS4347_Project
```

References

[1] V. Araj. 87 real estate and home buying statistics for 2020, 2020.

[2] L. Breiman. Random forests, 2001.

[3] SoFi. What is the average age to buy a house?, 2019.