Real Estate Investment & California Price Prediction

Thip Rattanavilay, Master of Science, Data Science

Bellevue University, DSC-680

Abstract

I explore how predictive modeling can be applied in housing sale price prediction by analyzing the housing dataset and use machine learning models. Actually, I tried different models, namely, Multilayer linear regression, regression tree, lasso regression, random forest, and ridge just to name a few. Additionally, as the data have lots of variables with many missing values, I spend much time dealing with the data. I have performed explorer data analysis, feature engineering before model fitting. And then using rmse and R-squared (R2) to measure the model performance. After I tried the different models, I got some results.

Keywords: home value, log error, linear regression, decision tree, random forest, model regression, ridge, r-square (R2)

Real Estate Investment & California Price Prediction

Introduction

Housing costs have been on the rise in California, which has impacted affordability. Only less than 25% of home buyers can afford to buy a median-priced home in the Golden State. According to C.A.R.'s Traditional Housing Affordability Index (HAI), the percentage of home buyers who could afford to buy a median-priced existing single-family home in California in the second quarter of 2021 fell to 23 percent from 27 percent in the first quarter of 2021 and 33 percent in the second quarter of 2020. I am going to dive into the data and provide guidance on weather home buyer or investor should invest in California.

Business Problem Statement / Hypothesis

The California real estate market has had a record-breaking year in 2021, and we still have 1 months to go. Home prices in the Golden State have risen to all-time highs over the past year or so. This is largely due to an ongoing supply shortage within the real estate market. There are plenty of buyers out there, but not enough properties to satisfy demand.

The project aims at building a model of housing prices to predict median house values in California using the provided data from California Association of Realtors, Kaggle and Zillow dataset. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics. The motivation is to improve the accuracy in home value prediction with advanced methods. The main contribution is that it finds that traditional linear models are not predictive for complex home value data sets, while tree based non-linear models are most accurate with the lowest mean square errors. Data collected 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. This solution should provide insight for investors, new home buyer or REIT (stock market) on whether to invest or not invest in a certain area and district.

Dataset

After parsing through the initial dataset, I can see there are 20640 entries and 14 data columns. All data represents a block within a neighborhood. Initial descriptive statistics of the data. While not particularly useful now, it is important to see variances in the data. For initial pre-processing, let's group by data. Cont_col is all data, Loc_col is geographical and dist_col is distance. The target variable is Median_House_Value (Figure 1).

Zillow variables

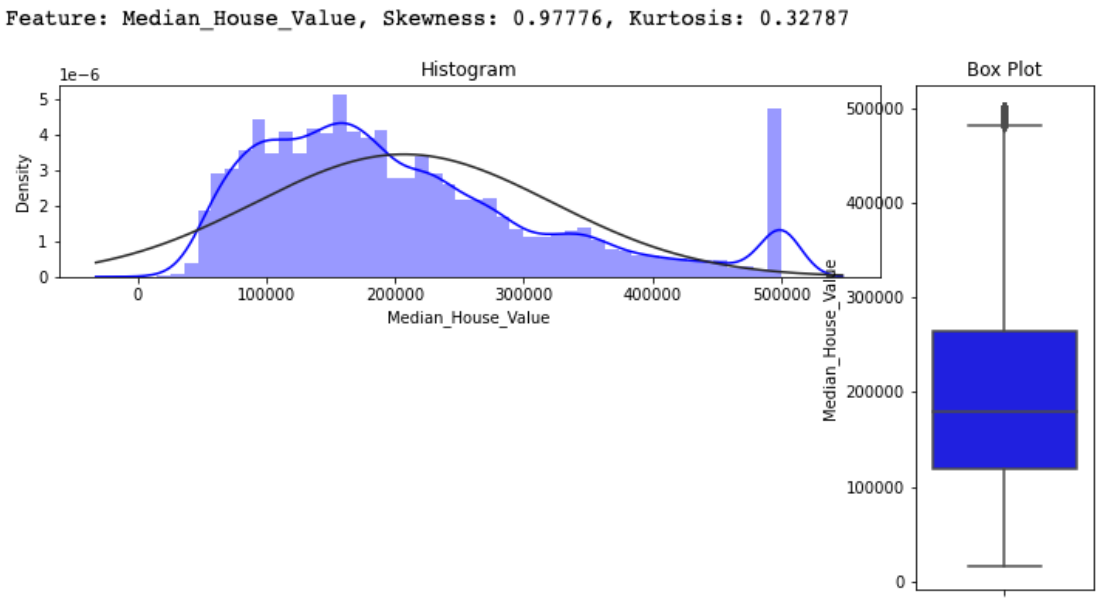
1. UnemploymentRate
2. MedianMortageRate
3. MedianMortageRate
4. MedianSoldPrice_AllHomes.California

California housing variables

1. Median_House_Value
2. Median_Income
3. Median_Age
4. Tot_Rooms

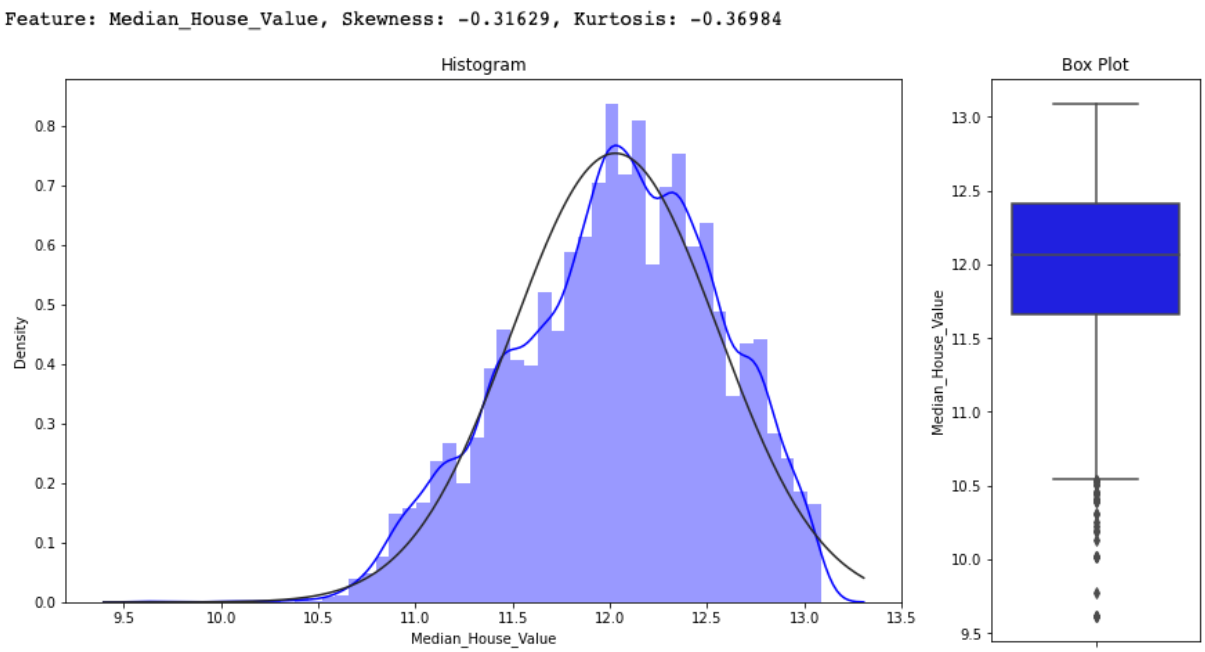
5. Tot_Bedrooms
6. Population
7. Households
8. Latitude
9. Longitude
10. Distance_to_coast
11. Distance_to_SanDiego
12. Distance_to_SanJose
13. Distance_to_SanFrancisco

Figure 1.



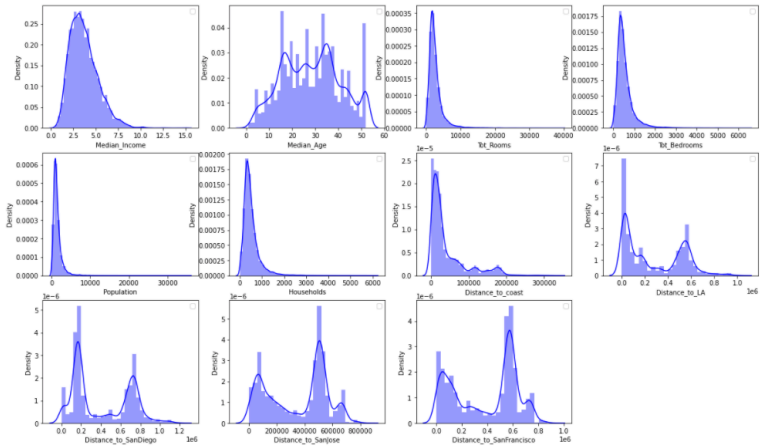
You can see the outliers on the higher end of the median house spectrum. We will now remove these according IQR. 5.19% Outlier removed in Median_House_Value according to IQR. This visualization is to check that all outliers were removed properly (Figure 2).

Figure 2.



As you can see, all data columns have no extreme outliers. Exploratory Data Analysis can begin now (Figure 3).

Figure 3.



Method

Chosen approach:

Business understanding: Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.

Data understanding: Examine the data and document its surface properties like data format, number of records, or field identities. Dig deeper into the data. Query it, visualize it, and identify relationships among the data.

Data preparation: Determine which data sets will be used and document reasons for inclusion/exclusion. A common practice during this task is to correct, impute, or remove erroneous values.

Modeling: Determine which algorithms to try (e.g. regression, neural net). Pending your modeling approach, you might need to split the data into training, test, and validation sets. Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design

Evaluation: Do the models meet the business success criteria? Which one(s) should we approve for the business? Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

Deployment: Report final results. Develop and document a plan for deploying the model. Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.

For each category of model, I'd tried 5 different parameters with a test train and validation set which gave me 5 different outcomes for each algorithm. From these I'd selected the ones with lowest RMSE. This gave me different candidate models. So, I had 5 different candidate models, one from each algorithm. From these I selected the one with the lowest RMSE as my final model. I then used this final model to check for the RMSE on the test set. I listed out my approach to achieve my findings.

One of the challenges in dealing with this data is, it had 35% of the attributes missing more than 90% of values. We handled imputation using 2 strategies given below

Approach 1:

- Imputation 2 datasets (California housing and Zillow)

Approach 2

- Imputation before operation (on split records)

  The purpose of running the imputation model using second approach is to take advantage of the records to find better model

Feature Engineering

As part of feature engineering, I created new variables using variable interaction and also did feature selection using different approaches.

Zillow variables

5. UnemploymentRate
6. MedianMortageRate
7. MedianMortageRate
8. MedianSoldPrice_AllHomes.California

California housing variables

14. Median_House_Value
15. Median_Income
16. Median_Age
17. Tot_Rooms
18. Tot_Bedrooms
19. Population
20. Households
21. Latitude
22. Longitude
23. Distance_to_coast
24. Distance_to_SanDiego
25. Distance_to_SanJose
26. Distance_to_SanFrancisco

- **Feature Selection & Dimensionality Reduction**

  The below mentioned techniques are used to make feature selection and dimensionality reduction.

- **Model Building & Fine Tuning**

All the ML algorithms used in this project are tuned with using scikit-learn algorithms and other algorithms.

  - Multilayer Linear regression
  - Regression tree
  - Lasso regression
  - Random forest
  - Ridge regression
  - The whole data is divided into train 90% and test 10%, models are built on train data using cross-validation techniques

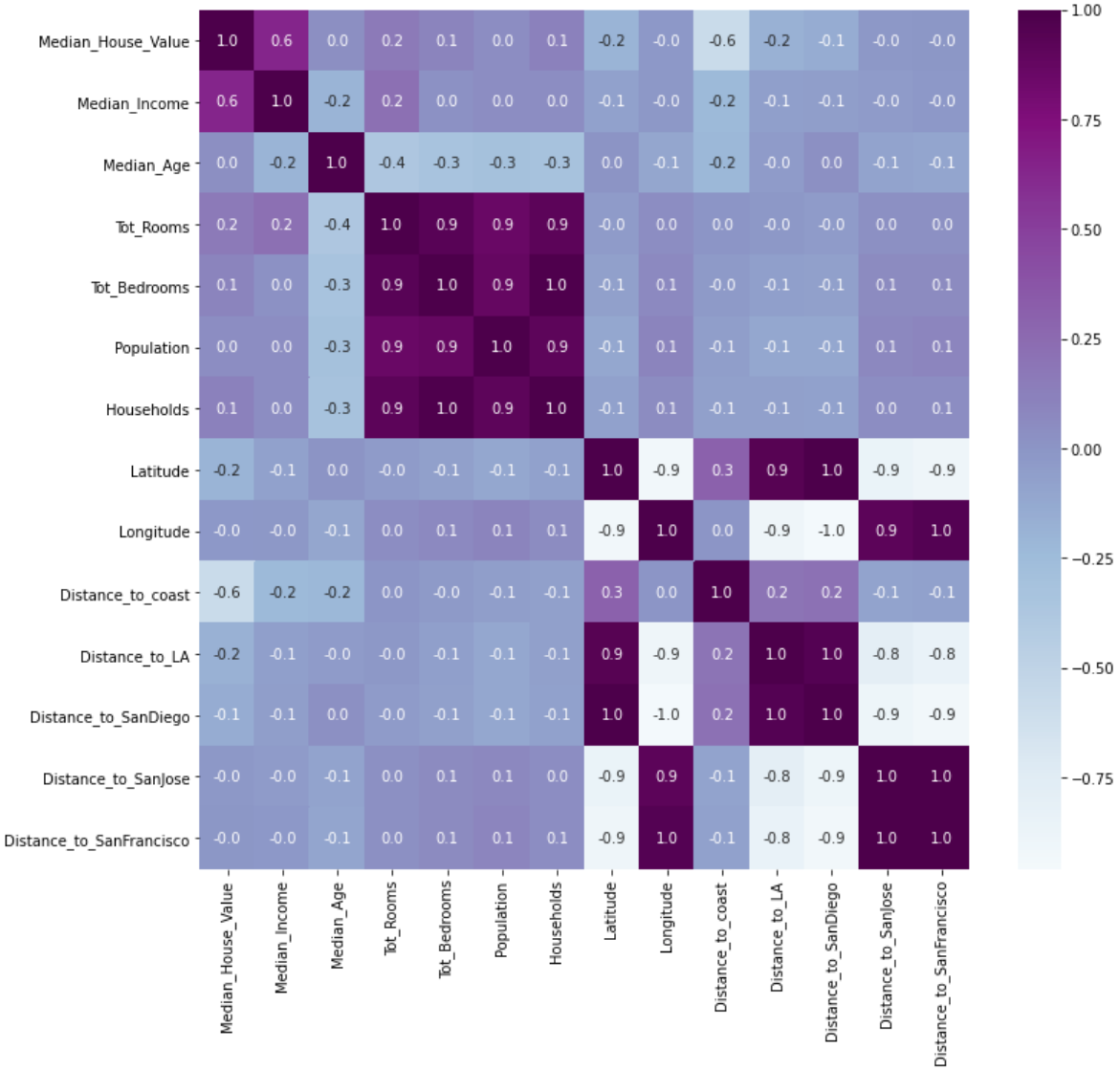
California Real Estate Investment and Prediction / Data Visualization

The project aims at building a model of housing prices in California using the California data from Kaggle and Zillow. 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 investments, home buyer and RIET.

Data Selection

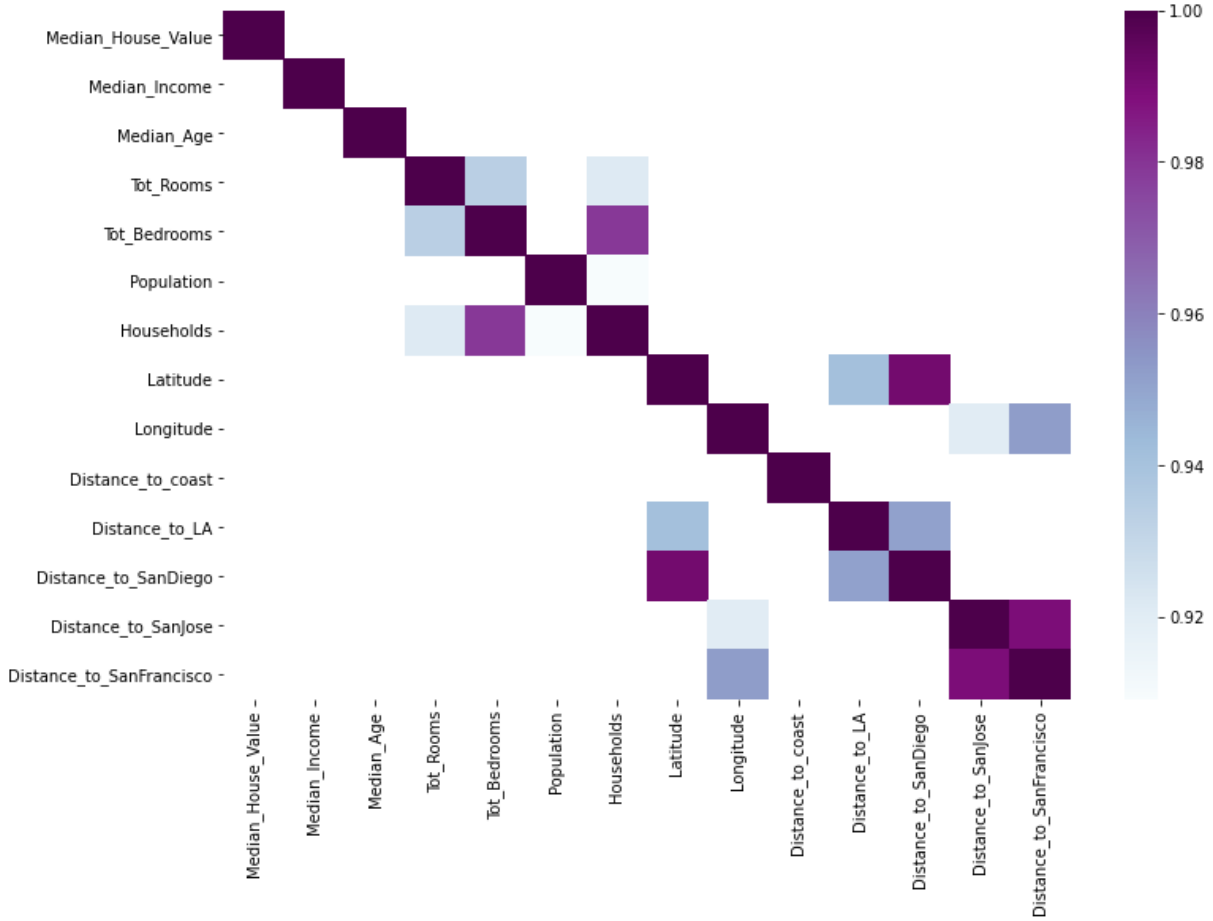
Highly correlated variables (>0.9) were observed among Tot_Rooms, Tot_Bedrooms, Households, and Population. Tot_Room, Tot_Bedrooms, Households will be excluded since Household variable alone can represent them (Figure 4)

Figure 4



Another set of high correlation between geographical variables: Latitude, Longitude, Distance_to_SanDiego, Distance_to_SanJose, and Distance_to_SanFrancisco. Thus, Latitude, Longitude, Distance_to_SanJose, and Distance_to_SanDiego will be dropped. Keeping only Distance_to_LA will cover Distance_to_SanDiego and Distance_to_SanFrancisco will cover Distance_to_SanJose since they are in close proximity. Figure 5

Figure 5.



	Median_House_Value	Median_Income	Median_Age	Households	Distance_to_coast	Distance_to_LA	Distance_to_SanFrancisco
0	13.022764	8.3252	41	126	9263.040773	556529.158342	21250.213767
1	12.789684	8.3014	21	1138	10225.733072	554279.850069	20880.600400
2	12.771671	7.2574	52	177	8259.085109	554610.717069	18811.487450
3	12.740517	5.6431	52	219	7768.086571	555194.266086	18031.047568
4	12.743151	3.8462	52	259	7768.086571	555194.266086	18031.047568
...	...	...	...	...	...	...	...
20635	11.265745	1.5603	25	330	162031.481121	654530.186299	222619.890417
20636	11.252859	2.5568	18	114	160445.433537	659747.068444	218314.424634
20637	11.432799	1.7000	17	433	153754.341182	654042.214020	212097.936232
20638	11.346871	1.8672	18	349	152005.022239	657698.007703	207923.199166
20639	11.400876	2.3886	16	530	146866.196892	648723.337126	205473.376575

19569 rows × 7 columns

Training / Validation and Testing sets

There were total records of 19569 and seven attributes after removing the outliers and highly correlated variables. The data set was then partitioned into training, validation, and test data sets (40%, 35%, and 25% respectively). There were 7828 records in training, 6849 records in validation, and 4620 records in test data sets with 7 variables. Figure 6

Figure 6.

```
Training   :  (7828, 7)
Validation :  (6849, 7)
Test :  (4892, 7)
```

MODEL BUILDING STRATEGIES

Regression Analysis is used in many studies including the predictions of house prices. Since the purpose of this study was to predict the numeric and continuous target variable of house prices, Multiple Linear Regression, Regression Tree, Random Forest, and Ridge and Lasso Regression models were selected.

Model will be trained on the train data set and evaluated with validation data set for better performance. The best model will be selected by comparing the metrics of each model on the test data set.

The selected models were built to predict the target outcome of Median House Value by using the independent predictors: 'Median_Income', 'Median_Age', 'Households', 'Distance_to_coast', 'Distance_to_LA', 'Distance_to_SanFrancisco'. The target outcome was the Median House Value. Figure 7

Figure 7

```
Variables: Median_Income, Median_Age, Households, Distance_to_coast, Distance_to_LA, Distance_to_SanFrancisco
Start: score=0.28, constant
Step: score=0.16, add Median_Income
Step: score=0.11, add Distance_to_coast
Step: score=0.11, add Households
Step: score=0.11, add Median_Age
Step: score=0.11, add Distance_to_LA
Step: score=0.11, add Distance_to_SanFrancisco
Step: score=0.11, unchanged None
['Median_Income', 'Distance_to_coast', 'Households', 'Median_Age', 'Distance_to_LA', 'Distance_to_SanFrancisco']
```

The best predictor was Median_Income with a score of 01.6, and the rest were Distance_to_coast, Households, Median_Age, Distance_to_LA, Distance_to_San_Francisco with the same score of 0.11. All the predictors were used for training the model since all revealed similar scores.

RESULTS AND FINAL MODEL SELECTION

Regression statistics of different models based on the test data. According to regression statistics and residual errors plot, Random Forest deem the best model. Figure 8

Figure 8.

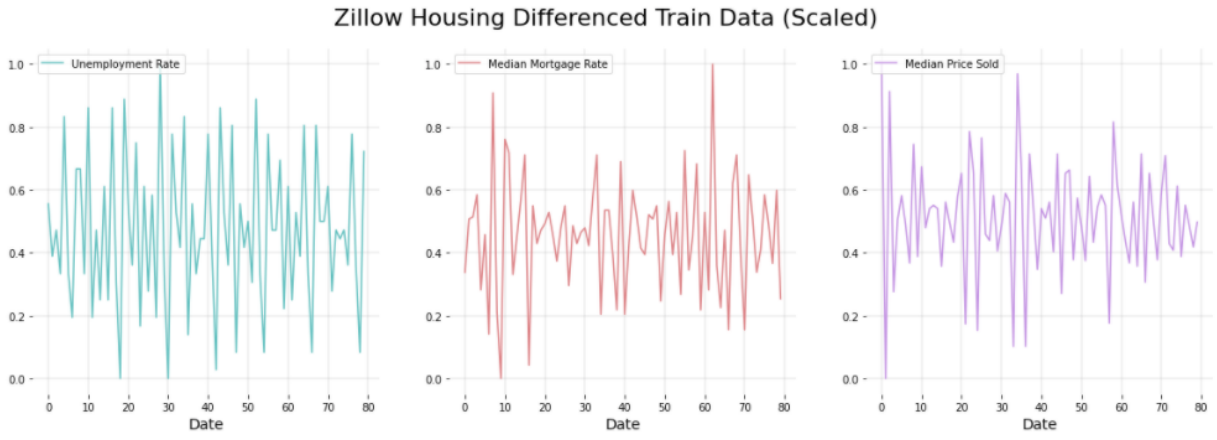
```
MULTIPLE LINEAR REGRESSION:

Regression statistics

                        Mean Error (ME) : 0.0045
      Root Mean Squared Error (RMSE) : 0.3183
           Mean Absolute Error (MAE) : 0.2474
         Mean Percentage Error (MPE) : -0.0365
Mean Absolute Percentage Error (MAPE) : 2.0652
LASSO:

Regression statistics

                        Mean Error (ME) : 0.0045
      Root Mean Squared Error (RMSE) : 0.3183
           Mean Absolute Error (MAE) : 0.2474
         Mean Percentage Error (MPE) : -0.0365
Mean Absolute Percentage Error (MAPE) : 2.0652
RIDGE:

Regression statistics

                        Mean Error (ME) : 0.0045
      Root Mean Squared Error (RMSE) : 0.3183
           Mean Absolute Error (MAE) : 0.2474
         Mean Percentage Error (MPE) : -0.0369
Mean Absolute Percentage Error (MAPE) : 2.0653
REGRESSION TREE MODEl:

Regression statistics

                        Mean Error (ME) : -0.0049
      Root Mean Squared Error (RMSE) : 0.4417
           Mean Absolute Error (MAE) : 0.3343
         Mean Percentage Error (MPE) : -0.1088
Mean Absolute Percentage Error (MAPE) : 2.7865
RANDOM FOREST:

Regression statistics

                        Mean Error (ME) : 0.0007
      Root Mean Squared Error (RMSE) : 0.2300
           Mean Absolute Error (MAE) : 0.1647
         Mean Percentage Error (MPE) : -0.0341
Mean Absolute Percentage Error (MAPE) : 1.3736
```

```
Multiple Linear Regression: (0.6282688626237674, 0.31829977423969413)
Random Forest: (0.8058548489465904, 0.23003034691641602)
Regression Tree: (0.28431003078211403, 0.4416561939597983)
Lasso Regression: (0.6282666086798013, 0.31830073922311347)
Ridge Regression: (0.6282273903822614, 0.3183175293261262)
```

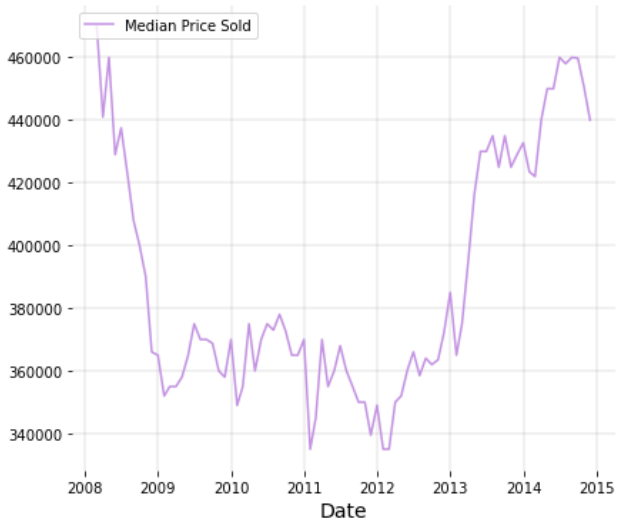
Summary Metrics Table of each model on test data set

	Model	R2	ME	RMSE	MAE	MPE	MAPE
0	MLR	0.6283	0.0007	0.2300	0.1647	-0.0341	1.3736
1	Lasso	0.6283	0.0045	0.3183	0.2474	-0.0365	2.0652
2	Ridge	0.6282	0.0045	0.3183	0.2474	-0.0369	2.0653
3	Regression Tree	0.2886	-0.0059	0.4403	0.3330	-0.1172	2.7772
4	Random Forest	0.8059	0.0007	0.2300	0.1647	-0.0341	1.3736

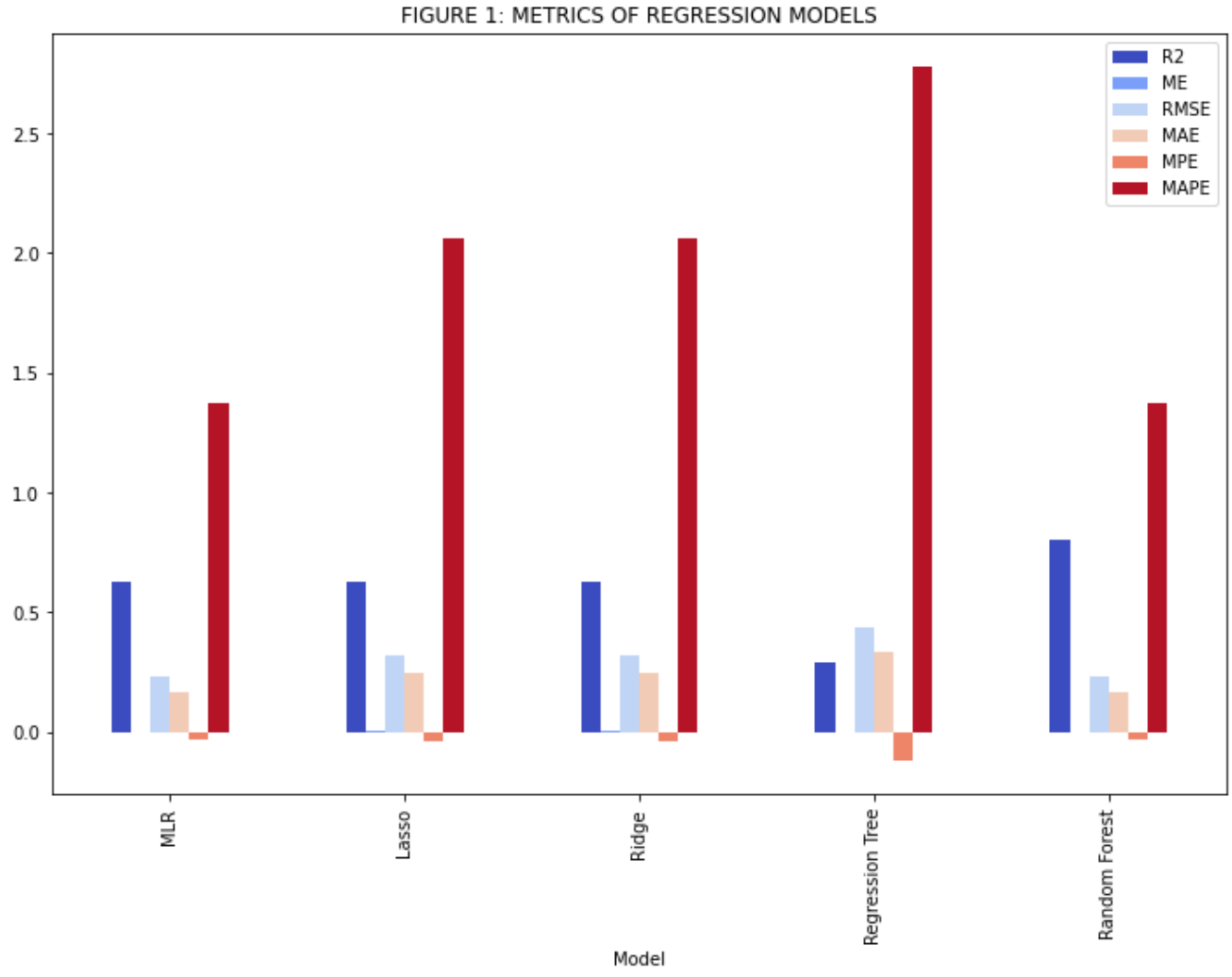
Train data on Zillow



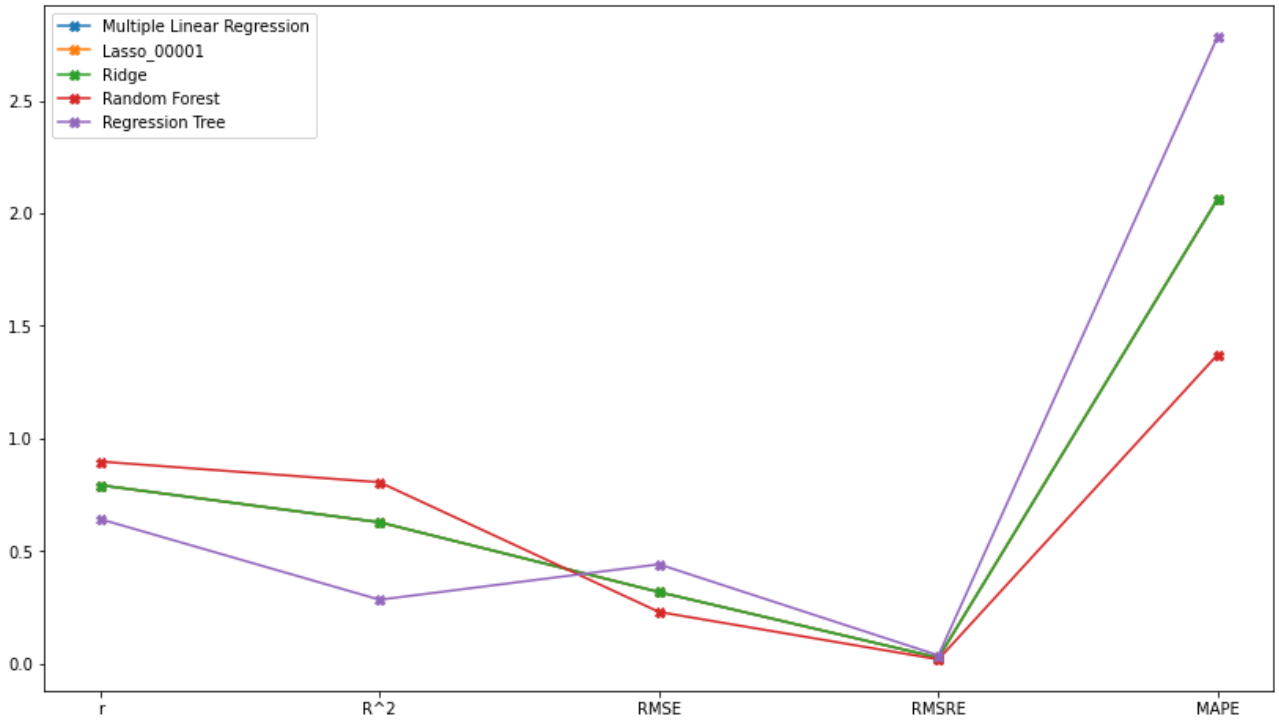

Unemployment Rate
Median Mortgage Rate
9
8
7
6
5
4
3
2008
2009
2010
2011
2012
2013
2014
2015
Date


Median Price Sold
460000
440000
420000
400000
380000
360000
340000
2008
2009
2010
2011
2012
2013
2014
2015
Date

Comparison of Metrics for Each Model



Plotting graphs model and similarity between MLR, LASSO,RF, RT and RIDGE



From the Metric Table, metrics of Multiple Linear Regression, Ridge, and Lasso were almost identical with R2 value of 0.6282 (62.82%). Thus, in the above graph, all three overlap together and doesn't show any difference between these three models.

The metrics of Regression Tree for the train data set showed zero percent error revealing the overfitting behavior and the R2 value of 28.86%, the lowest score out of all the models.

Among them, Random Forest scored the highest R2 value of 80.59% best fitting the data set. The metrics for the Random Forest also ranked the lowest percentage errors of ME, RMSE, MAE, MPE, and MAPE as shown in the above figures. Thus, Random Forest deemed the best model for the prediction of California house prices.

Recommended Next Steps

The Random Forest scored the highest R2 value of 80.59% best fitting the data set. The metrics for the Random Forest also ranked the lowest percentage errors of ME, RMSE, MAE, MPE, and MAPE. Therefore, we would use this algorithm for all machine learning processes. The application of this research can be applied to applications across the real estate industry. The models had high precision outputs and could be utilized to create deliverables that can identify real estate evaluation to answer questions such as: Which house should I buy or build to maximize my return? Where or when should I do so? What is its optimum rent or sale price?

Appendix.

Zillow variables:

UnemploymentRate

MedianMortageRate

MedianMortageRate

MedianSoldPrice_AllHomes.California

California housing variables:

Median_House_Value

Median_Income

Median_Age

Tot_Rooms

Tot_Bedrooms

Population

Households

Latitude

Longitude

Distance_to_coast

Distance_to_SanDiego

Distance_to_SanJose

Distance_to_SanFrancisco

House_Area	House_Details	Region	Tax
area_basement	num_unit	region_county	tax_total
area_patio	num_story	region_city	tax_building
area_shed	num_room	region_zip	tax_land
area_pool	num_bathroom		tax_property

1. Total rooms = bath_count + bedroom_count
2. Average room size = total_finished_living_area_sqft/roomcnt
3. Ratio of structure tax to land tax = structure_tax/land_tax
4. ExtraSpace = lot_area_sqft - total_finished_living_area_sqft

10 Questions

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

The dataset that I gather was rich in data and it took time to understand it.

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

Great insight on the median house hold it clearly stated that a normal earned income cannot afforded a home is California for a single-home

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

Prices are too high right now in California with a median income. The price is currently at $800K at this moment.

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

The safest area in California is currently in the Central Valley and Coast area.

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

The cost of living in Northern California is at 90% higher than other States or Counties, you would have to earn about $200k to live there as a single person.

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

The cost of living in the Central Valley is about 50% higher than other States, you would have to earn $70K-$80K.

7. Southern Californian less expensive than Northern California?

I would have to agree with this statement. About 30% cheaper.

8. What city or county has the worst crimes?

The data and news outlet say Bay area and some parts in Southern California.

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

After reviewing the data from what was gathered. It seems to be all of the above.

10. Why is the market so high in California?

I think that it has to do with the media outlet and investors like Zillow and Redfin that are buying everything and reselling it with a higher price tag.

Conclusion

Overall, our model underestimated the values in each of the California counties by an average of about 7%. This was in contrast to the overall upward trend of the pricing index which would have forecast an upward trend of anywhere from 5-10% year over year. The R squared in this model of training set is very good, but in the test set the R squared is relatively low, which may show the random forest model is a little bit overfitting. The overall Mean squared error on our testing datasets was just

under 5%. The overall loss function averaged 10% for the validation data set, but around 33% for the training data set, indicating our training model could still be improved.

The reason for the model shortcomings likely stems from the difficulty the model faced in projecting a large number of varying time series over each income, zip code, rather than a large number of observations from a single time series (as would be the case for a financial time series. Perhaps with more data points in future, or additional features, I may be able to resolve these issues. My recommendation for investors and homebuyer is to wait while to see if the market will dip because the housing prices are just too high right now and it looks like it will die down at some point in time.

References

Fedesoriano. (2021, July 3). *California housing prices data (5 new features!)*. Kaggle. Retrieved November 9, 2021, from https://www.kaggle.com/fedesoriano/california-housing-prices-data-extra-features.

Housing Data. Zillow Research. (2021, March 25). Retrieved November 9, 2021, from https://www.zillow.com/research/data/.

" House Prices: Advanced Regression Techniques ". Kaggle, 2020, https://www.kaggle.com/c/house-prices-advanced-regression-techniques.

"Financial crisis 07-08". Wikipedia, 2020, https://en.wikipedia.org/wiki/Financial_crisis_of_2007–08.

"Half-Bath". Relator, 2020, https://www.realtor.com/advice/buy/what-is-a-half-bath/.

"RMSE". Wikipedia, 2020, https://en.wikipedia.org/wiki/Root-mean-square_deviation/.

"R-squared". Wikipedia, 2020, https://en.wikipedia.org/wiki/Coefficient_of_determination/.

"Linear Regression". Wikipedia, 2020, https://en.wikipedia.org/wiki/Linear_regression/.

"Lasso Regression". Wikipedia, 2020, https://en.wikipedia.org/wiki/Lasso_(statistics)

"Understanding Random Forest". Medium, 2020, https://towardsdatascience.com/understandingrandom-forest-58381e0602d2/.

"Random Forest". Wikipedia, 2020, https://en.wikipedia.org/wiki/Random_forest.