Project 3 Draft Milestone 3: 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, 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)

Project 3 Draft Milestone 3: Real Estate Investment & California Price Prediction

Business Problem

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.

Method

For each category of model, I'd tried 5 different parameters with a test train and validation set which gave me 5 different outcome 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.

Chosen approach:

- Exploratory data analysis
- Impute missing values
- Feature Engineering
- Model Building & Fine Tuning

Impute Missing values

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

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

- **Feature Selection & Dimensionality Reduction**

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

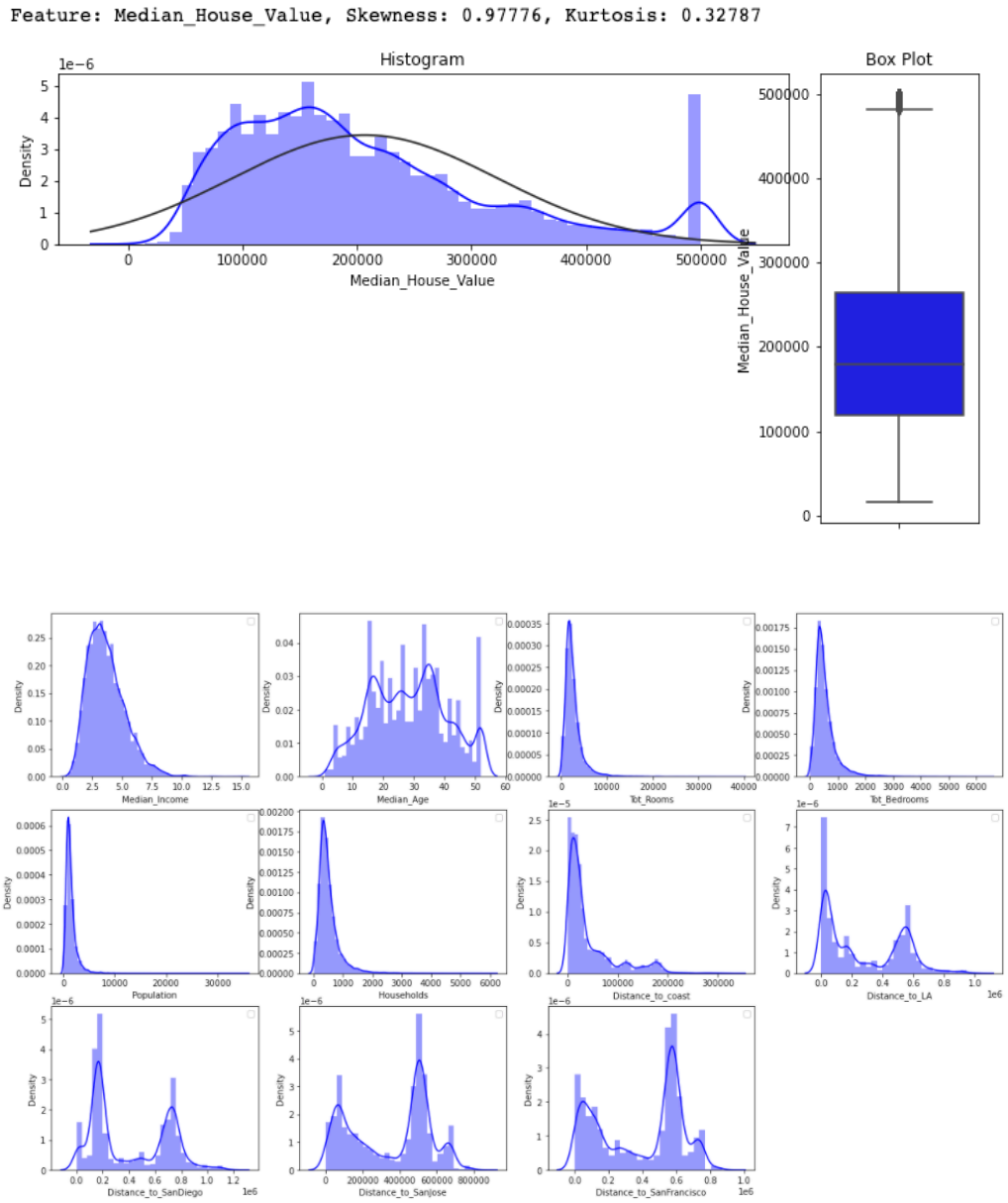
- **Model Building & Fine Tuning**

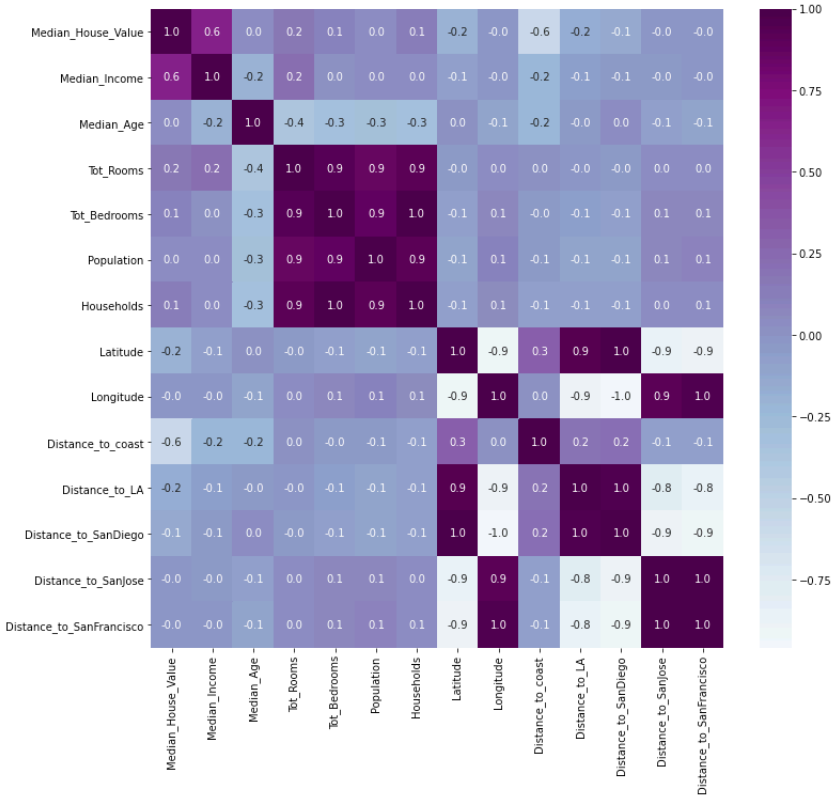
  All the ML algorithms used in this project are tuned with using scikit-learn algorithms

  - 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

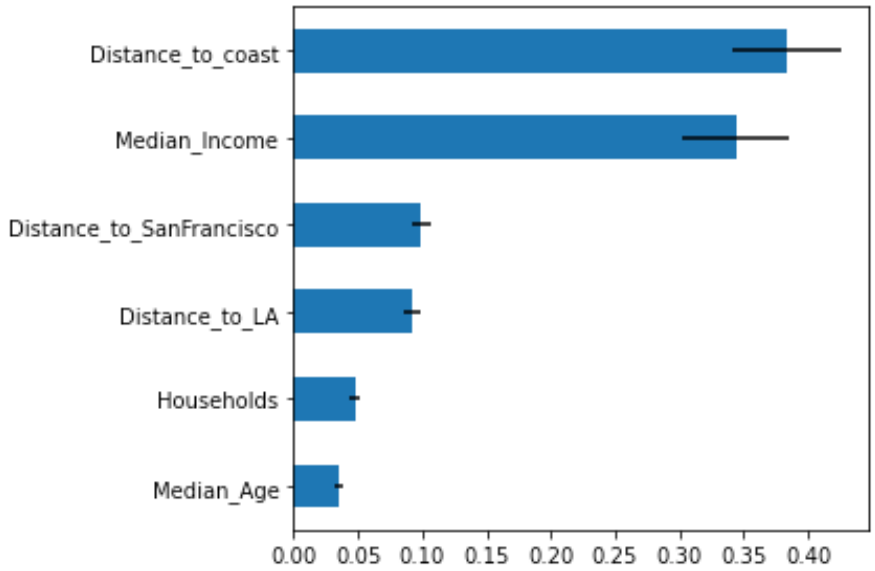
Illustrations

California_House_Prices_Prediction

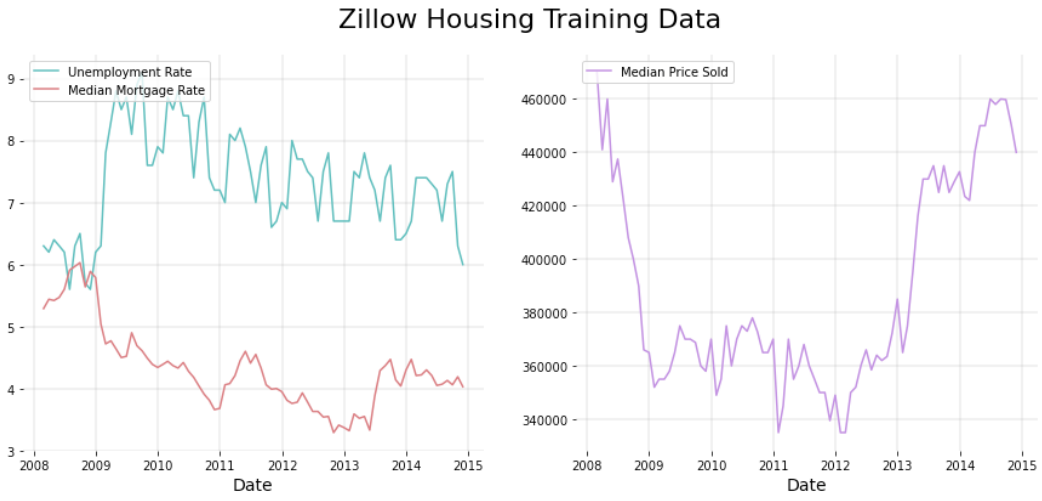


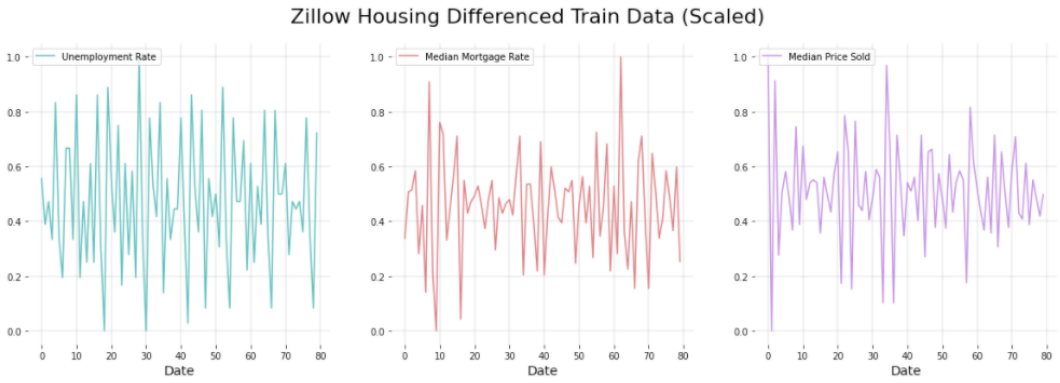

Median_House_Value 1.0 0.6 0.0 0.2 0.1 0.0 0.1 -0.2 -0.0 -0.6 -0.2 -0.1 -0.0 -0.0
Median_Income 0.6 1.0 -0.2 0.2 0.0 0.0 0.0 -0.1 -0.0 -0.2 -0.1 -0.1 -0.0 -0.0
Median_Age 0.0 -0.2 1.0 -0.4 -0.3 -0.3 -0.3 0.0 -0.1 -0.2 -0.0 0.0 -0.1 -0.1
Tot_Rooms 0.2 0.2 -0.4 1.0 0.9 0.9 0.9 -0.0 0.0 0.0 -0.0 -0.0 0.0 0.0
Tot_Bedrooms 0.1 0.0 -0.3 0.9 1.0 0.9 1.0 -0.1 0.1 -0.0 -0.1 -0.1 0.1 0.1
Population 0.0 0.0 -0.3 0.9 0.9 1.0 0.9 -0.1 0.1 -0.1 -0.1 -0.1 0.1 0.1
Households 0.1 0.0 -0.3 0.9 1.0 0.9 1.0 -0.1 0.1 -0.1 -0.1 -0.1 0.0 0.1
Latitude -0.2 -0.1 0.0 -0.0 -0.1 -0.1 -0.1 1.0 -0.9 0.3 0.9 1.0 -0.9 -0.9
Longitude -0.0 -0.0 -0.1 0.0 0.1 0.1 0.1 -0.9 1.0 0.0 -0.9 -1.0 0.9 1.0
Distance_to_coast -0.6 -0.2 -0.2 0.0 -0.0 -0.1 -0.1 0.3 0.0 1.0 0.2 0.2 -0.1 -0.1
Distance_to_LA -0.2 -0.1 -0.0 -0.0 -0.1 -0.1 -0.1 0.9 -0.9 0.2 1.0 1.0 -0.8 -0.8
Distance_to_SanDiego -0.1 -0.1 0.0 -0.0 -0.1 -0.1 -0.1 1.0 -1.0 0.2 1.0 1.0 -0.9 -0.9
Distance_to_SanJose -0.0 -0.0 -0.1 0.0 0.1 0.1 0.0 -0.9 0.9 -0.1 -0.8 -0.9 1.0 1.0
Distance_to_SanFrancisco -0.0 -0.0 -0.1 0.0 0.1 0.1 0.1 -0.9 1.0 -0.1 -0.8 -0.9 1.0 1.0
Median_House_Value
Median_Income
Median_Age
Tot_Rooms
Tot_Bedrooms
Population
Households
Latitude
Longitude
Distance_to_coast
Distance_to_LA
Distance_to_SanDiego
Distance_to_SanJose
Distance_to_SanFrancisco
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

	feature	importance	std
1	Median_Age	0.034889	0.003350
2	Households	0.047419	0.004691
4	Distance_to_LA	0.091673	0.006557
5	Distance_to_SanFrancisco	0.099123	0.006975
0	Median_Income	0.343972	0.041653
3	Distance_to_coast	0.382924	0.042084



Zillow_California_prices





Appendix.

PREDICTORS:

- Median_Income
- Median_Age
- Total_Rooms
- Total_Bedrooms
- Population
- Households
- Latitude
- Longitude

- Distance to LA
- Distance to SanDiego
- Distance to SanJose

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.
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? (Last Name, Year)

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.