Predicting California House Prices

Daniel Pickett
December 2022

Section 1: Problem Statement

The price of housing has increasingly become a hot-button issue in the United States. Over the last two to three decades, the median pricing of units, regardless of whether to own or to rent, has more and more outpaced median wages, such that the amount of time worked or salary earned in order to afford it continues to increase with each year. This has been further exacerbated by the recent COVID pandemic: when many jobs in higher-priced cities transitioned to remote work, many of the employees relocated to more affordable cities, increasing their cost of living in turn. California, understandably, was a large source of these outgoing employees. Many of these jobs were located in its Bay Area or Greater Los Angeles Area, two of the most expensive locations to live in the country. Their access to pleasant weather, abundant nature, and diverse culture, which attracted wealth and talent to the state, much of which established Silicon Valley, previously helped cement California's cost of living as among the highest of all states.

As a result of these effects on pricing both long- and short-term, homeownership in California is among the lowest in the country. But what if house price could be predicted ahead of time?

Provided a 1990 dataset on California house characteristics (such as age and number of bedrooms) and location (such as geographic coordinates and distance from different major cities) versus their median price within their given block, can we create a model to estimate the value of a California house not listed within the data?

Section 2: Data Wrangling

The raw dataset contained 20,640 records with 14 columns of data, including the dependent variable, median house price. Each record represented a block of houses within California, whose location was defined by independent variables for latitude and longitude.

Superficial checks revealed little wrong with the data. All columns contained non-categorical numerical data. No values were missing or of the incorrect data type. Additionally, there were no duplicate records – although there were several that had identical combinations of latitude and longitude, their other values were different enough that most likely these "duplicates" were simply blocks near each other. However, many records did share the specific maximum values of median house price, median income, and median age, suggesting that these "maximums" were actually placeholders for missing data. These incomes and ages were subsequently imputed with the respective means. However, as the dependent variable, the house prices could not be substituted without significantly affecting model results, so their records were removed entirely.

All variables other than latitude and longitude appeared to have some degree of right skewness. This is consistent with the concept that there are generally fewer expensive houses, and any traits associated with them are less common as well. After the treatment for the missing values, there did not appear to be any other type of placeholder values.

Five variables describing the distance from the block to either the coast or a major California city were provided in kilometers. These were converted to miles for easier comprehensibility. Additionally, median house price was provided in USD, but median income was provided in tens of thousands of USD. For easier comparability, median income was converted to single USD.

There were two other noteworthy observations that did not merit immediate action:

- **At least one block contains fewer people than households.** This could be because some of the houses in the block are unoccupied, and the occupied houses have low occupancy.

- **10 blocks have a median house value of $25,000 or less**. Each of these blocks was close to a coast or major California city, which historically has highly correlated to value, so such a low median value seemed suspicious. However, the data is over twenty years old, there are any number of other factors that could contribute to the lower values, and there is no other reason suggesting why these values could be incorrect.

The wrangled dataset contained 19,675 records with all 14 columns of data.

Section 3: Exploratory Data Analysis

A heat map was created to visualize correlations between any of the variables (Figure 1).

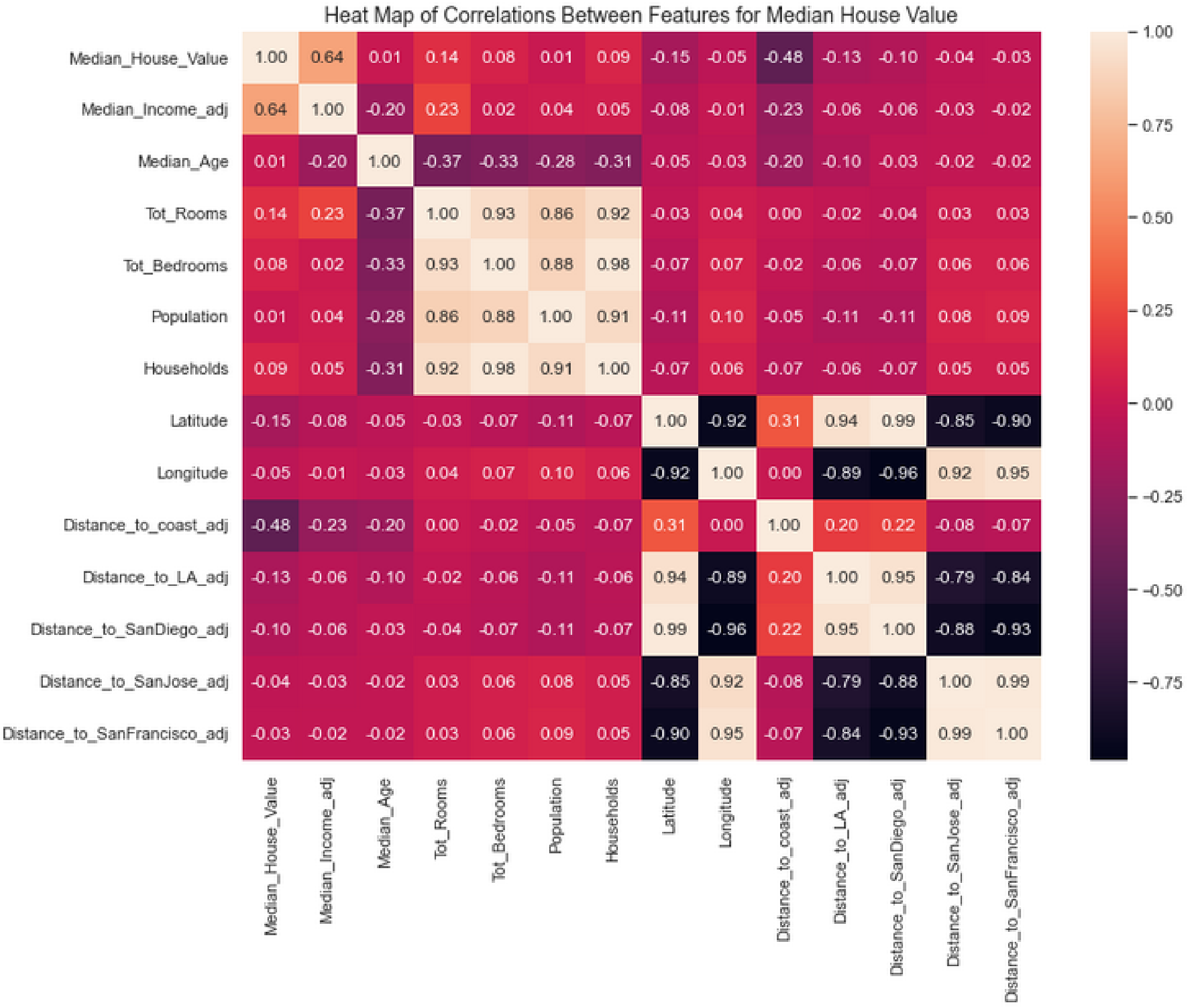


Figure 1: Correlation heat map between all variables provided in the dataset.

According to the heat map, only two variables have any significant correlation with median house value, median income (positively correlated) and distance to the coast (oddly, negatively correlated). Total rooms, bedrooms, population, and number of households within a block, while seemingly uncorrelated with house value, are all very strongly positively correlated with each other. This made sense conceptually: more houses directly results in more rooms and more people, yet none of these are necessarily indicative of a measure of wealth within the block. However, perhaps the densities of people per room or household or of rooms per household could be – perhaps either fewer people within a living space indicated that each person had a larger space to themselves, or more rooms within a household indicated that the houses were larger, both of which support the notion that the houses would be more expensive. A second heat map was created, this time containing five new density variables (Figure 2):

- Population per room within a block
- Population per bedroom within a block
- Population per household within a block
- Rooms per household within a block
- Bedrooms per household within a block

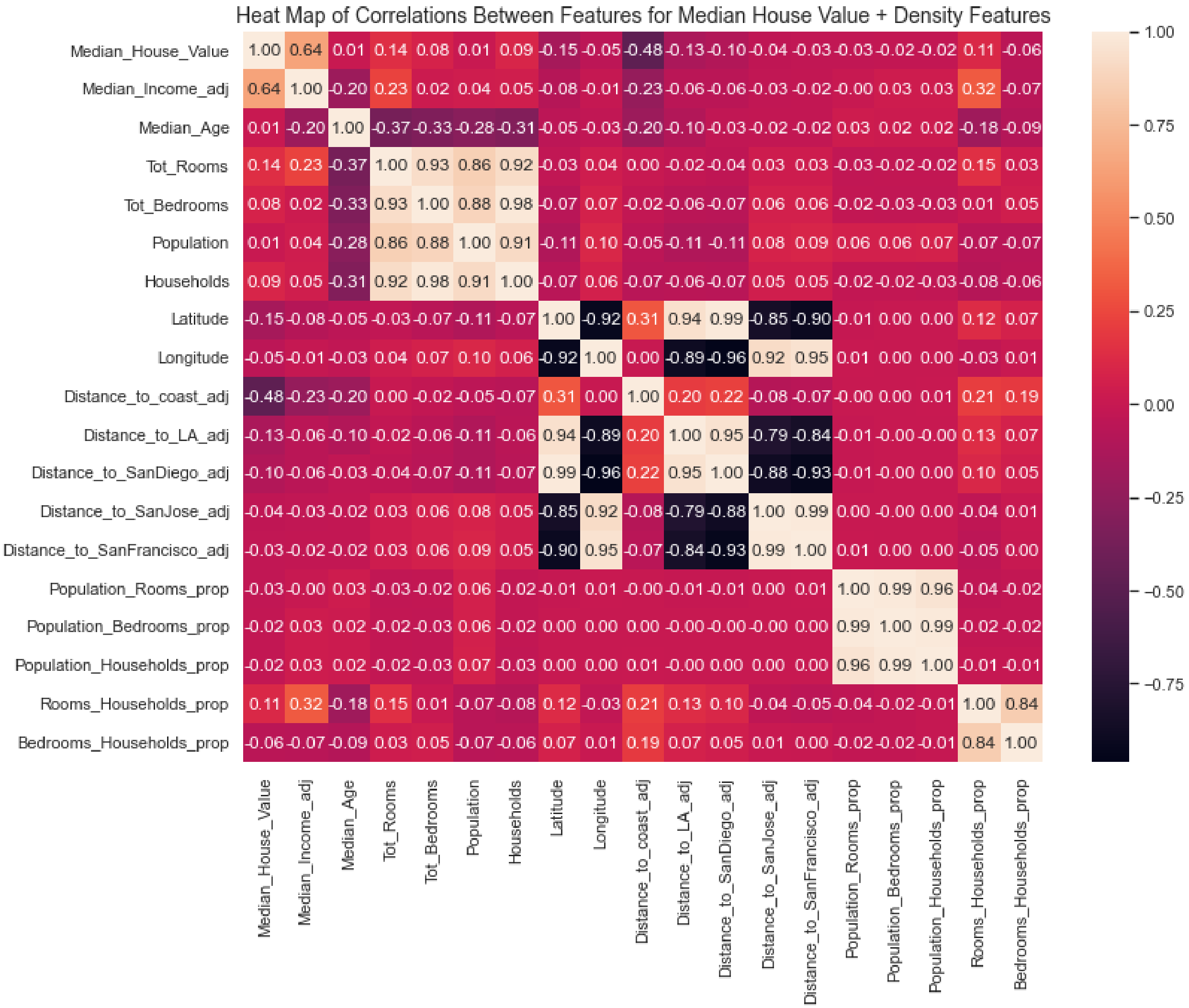


Figure 2: Correlation heat map between all variables provided in the dataset, plus five new density variables (ending in "_prop").

As shown above, none of these density variables were significantly correlated with median house value, and they were subsequently dropped from the analysis.

Next, each field other than latitude and longitude was checked for outliers. This was visualized by boxplot and checked using its outlier threshold of 1.5 times the interquartile range beyond either the 1st or 3rd quartile.

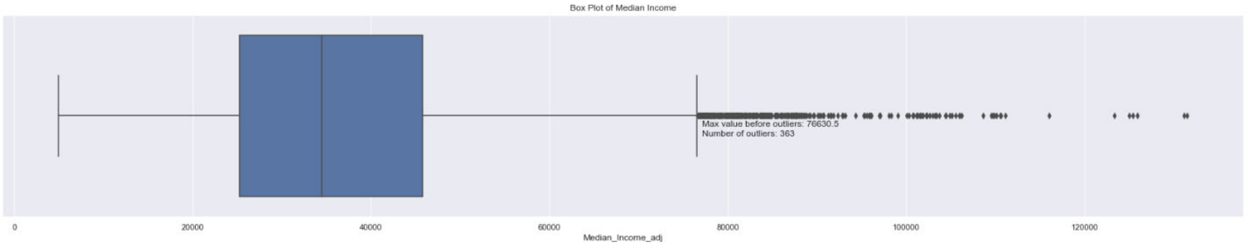


Figure 3: One of the boxplots created to check for outliers, in this case for median income within a block.

Based on this rule, each variable appears to have some number of outliers, and in each case, all of them are high value instead of low value. However, as mentioned in Section 2, each variable is distributed at least somewhat skewed to the right, so these greater outliers are expected. Additionally, they appear to only technically be outliers: although they do surpass the outlier threshold, most of them do not appear to be significantly greater than the highest-value records that aren't outliers. Therefore, none of these outliers were adjusted or removed on this basis.

There was one noteworthy observation that did not merit immediate action:

- **20 blocks have a very high median house value (at least 450,000) but also have a very low median income (less than 25,000).** Each of these blocks had varying closeness to a major city, but nearly all of them were close to the coast (fewer than 15 miles), which theoretically would positively correlate with house value. However, most of these blocks are also quite old – 18 of the 20 blocks have a median age of at least 19 years. The relationship between age and house value could be viewed in several different ways: the houses were purchased at a time when one with a much lower income could afford them; the age supports higher value due to the house's historicity; or the age supports *lower* value due to the house falling into disrepair. Regardless, there is not enough information or reason to be suspicious to look further into these records.

Section 4: Pre-Processing and Training Data Development

Few preparatory steps remained for the data at this point, as the data was largely clean to begin with and most issues were already solved to properly undertake Section 3. As none of the variables are categorical, they required no transformation into dummy or indicator variables. Scaling was also considered for each of the variables other than latitude and longitude. Because none of them were normally distributed, it would not have made sense to scale them to a normal distribution, so they were instead scaled to a range from 0 to 1. Finally, the data was split into training and testing sets at a ratio of 7 to 3. The resulting X and y training sets each contained 13,772 records, and the X and y testing sets each contained 5,903 records.

Section 5: Modeling

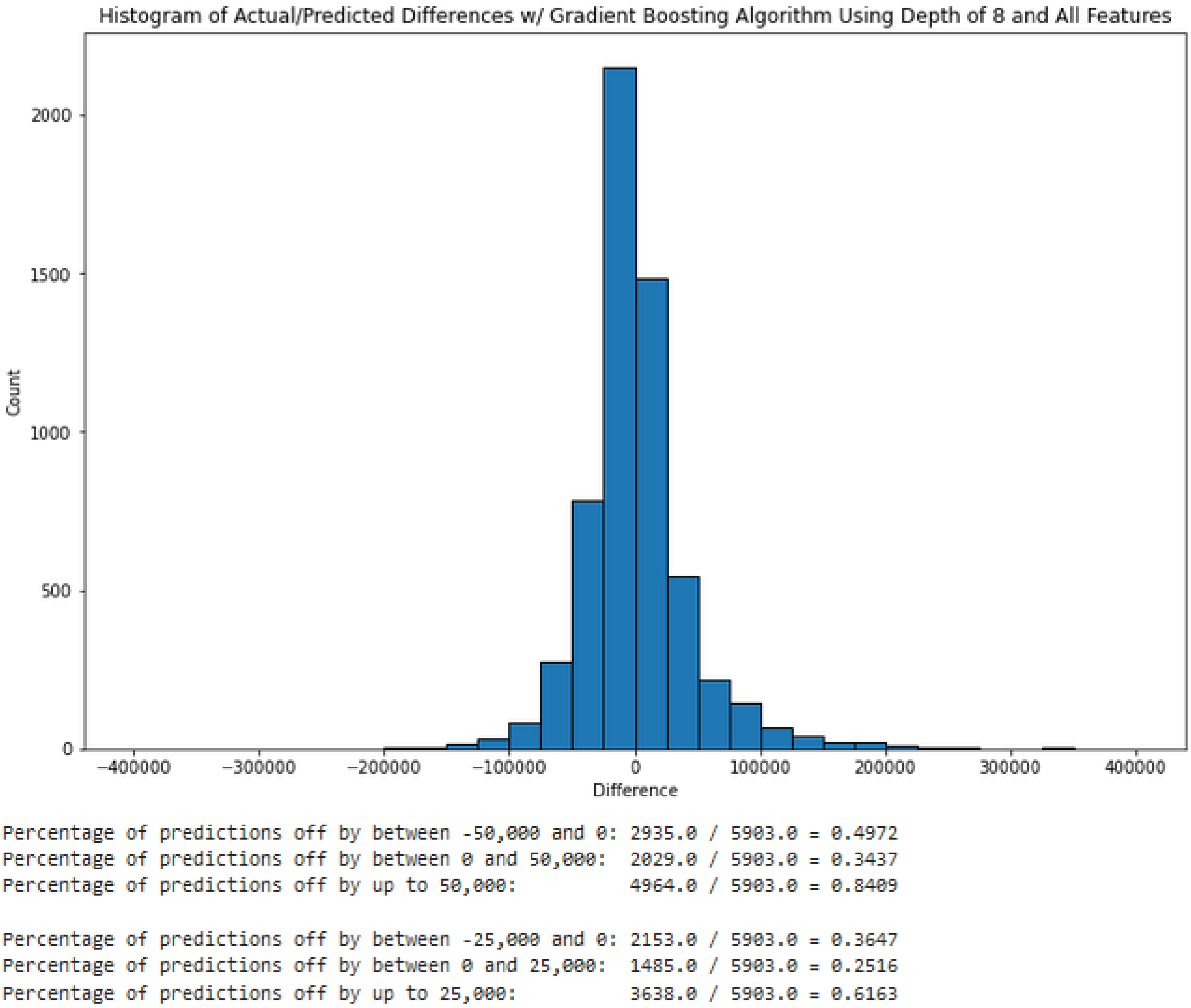
Now that it was cleaned and pre-processed, the data was ready for model fitting. The dependent variable, median house price, is a quantitative variable presumably correlated to some combination of the other provided features, and whose values are continuous rather than discrete or categorical. This therefore required some type of regression model.

A total of nine models were fitted onto the X and y training sets, then used to predict median house prices from the X testing set, which were finally compared against the y testing set:

- Model 1: linear regression model, using all 13 dependent variables
- Model 2: linear regression model, using the 2 dependent variables previously determined to have significant correlation with median house price (median income and distance to the coast)
- Model 3: linear regression model, using the same 2 highly correlated dependent variables, while also taking the log of distance to the coast (to potentially address certain higher-value underestimations)
- Model 4: linear regression model, using the best 10 features as determined by the SelectKBest algorithm

- Model 5: K-nearest neighbors model, using the 2 highly correlated variables, with 20 neighbors as determined by the GridSearchCV algorithm
- Model 6: K-nearest neighbors model, using the best k features as determined in Model 4 by the SelectKBest algorithm, with 9 neighbors as determined by the GridSearchCV algorithm
- Model 7: decision tree model with a tree depth of 10 as determined via cross-validation scoring of different k-depth models
- Model 8: random forest model with a tree depth of 15 as determined via cross-validation scoring of different k-depth models
- Model 9: gradient boosting model with a tree depth of 8 as determined via cross-validation scoring of different k-depth models

For each model, the root-mean-squared error (RMSE) was calculated between the actual testing values and the predicted values. This would be interpreted as the average value that the predicted value differed from the actual. Additionally, a histogram of the differences between the predicted and actual values was created to better visualize their distributions; the ideal histogram would lack skewness and have a tall center and small thin tails, indicating that more predictions were not far from their actual values, and that there was no discernable pattern in how they differed. From each histogram, it was noted what percentage of predictions differed by a maximum of 50,000 and by a maximum of 25,000. Along with RMSE, these were the chosen “ballpark” metrics of the model’s accuracy: the former would be the minimum threshold that the predictions were usable, and the latter would be the threshold at which the prediction was “very good”.

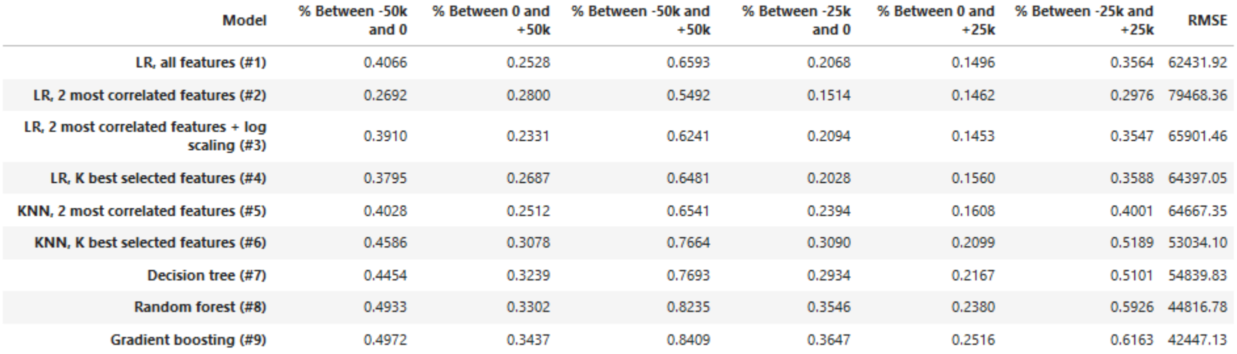


```
Percentage of predictions off by between -50,000 and 0: 2935.0 / 5903.0 = 0.4972
Percentage of predictions off by between 0 and 50,000:  2029.0 / 5903.0 = 0.3437
Percentage of predictions off by up to 50,000:          4964.0 / 5903.0 = 0.8409

Percentage of predictions off by between -25,000 and 0: 2153.0 / 5903.0 = 0.3647
Percentage of predictions off by between 0 and 25,000:  1485.0 / 5903.0 = 0.2516
Percentage of predictions off by up to 25,000:          3638.0 / 5903.0 = 0.6163
```

Figure 4: One of the histograms of the differences between predicted and actual values, in this case for Model 9.

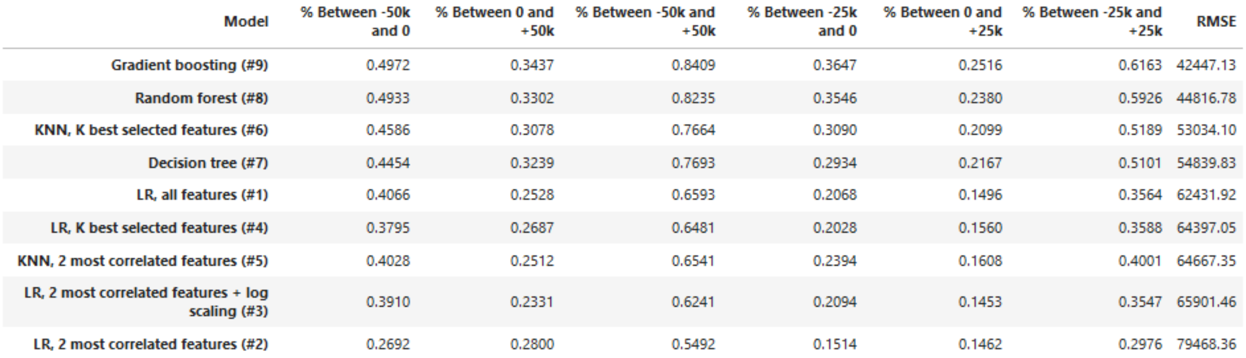
These results were compiled into the following summary table (Figure 5):



Model	% Between -50k and 0	% Between 0 and +50k	% Between -50k and +50k	% Between -25k and 0	% Between 0 and +25k	% Between -25k and +25k	RMSE
LR, all features (#1)	0.4066	0.2528	0.6593	0.2068	0.1496	0.3564	62431.92
LR, 2 most correlated features (#2)	0.2692	0.2800	0.5492	0.1514	0.1462	0.2976	79468.36
LR, 2 most correlated features + log scaling (#3)	0.3910	0.2331	0.6241	0.2094	0.1453	0.3547	65901.46
LR, K best selected features (#4)	0.3795	0.2687	0.6481	0.2028	0.1560	0.3588	64397.05
KNN, 2 most correlated features (#5)	0.4028	0.2512	0.6541	0.2394	0.1608	0.4001	64667.35
KNN, K best selected features (#6)	0.4586	0.3078	0.7664	0.3090	0.2099	0.5189	53034.10
Decision tree (#7)	0.4454	0.3239	0.7693	0.2934	0.2167	0.5101	54839.83
Random forest (#8)	0.4933	0.3302	0.8235	0.3546	0.2380	0.5926	44816.78
Gradient boosting (#9)	0.4972	0.3437	0.8409	0.3647	0.2516	0.6163	42447.13

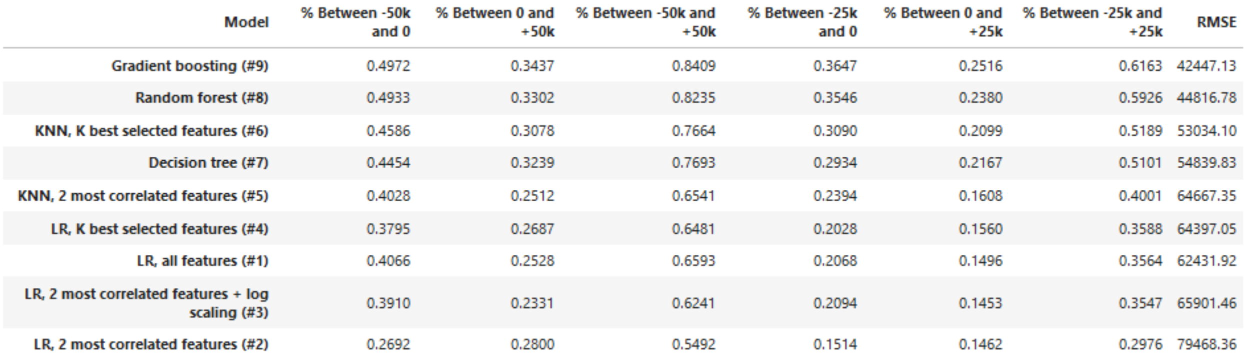
Figure 5: Summary table of the accuracy metrics of the nine models.

The table was then sorted by rank of RMSE (Figure 6) and of percentage of predictions differing by a maximum of 25,000 (Figure 7), the most important of the ballpark metrics:



Model	% Between -50k and 0	% Between 0 and +50k	% Between -50k and +50k	% Between -25k and 0	% Between 0 and +25k	% Between -25k and +25k	RMSE
Gradient boosting (#9)	0.4972	0.3437	0.8409	0.3647	0.2516	0.6163	42447.13
Random forest (#8)	0.4933	0.3302	0.8235	0.3546	0.2380	0.5926	44816.78
KNN, K best selected features (#6)	0.4586	0.3078	0.7664	0.3090	0.2099	0.5189	53034.10
Decision tree (#7)	0.4454	0.3239	0.7693	0.2934	0.2167	0.5101	54839.83
LR, all features (#1)	0.4066	0.2528	0.6593	0.2068	0.1496	0.3564	62431.92
LR, K best selected features (#4)	0.3795	0.2687	0.6481	0.2028	0.1560	0.3588	64397.05
KNN, 2 most correlated features (#5)	0.4028	0.2512	0.6541	0.2394	0.1608	0.4001	64667.35
LR, 2 most correlated features + log scaling (#3)	0.3910	0.2331	0.6241	0.2094	0.1453	0.3547	65901.46
LR, 2 most correlated features (#2)	0.2692	0.2800	0.5492	0.1514	0.1462	0.2976	79468.36

Figure 6: Summary table sorted by RMSE.



Model	% Between -50k and 0	% Between 0 and +50k	% Between -50k and +50k	% Between -25k and 0	% Between 0 and +25k	% Between -25k and +25k	RMSE
Gradient boosting (#9)	0.4972	0.3437	0.8409	0.3647	0.2516	0.6163	42447.13
Random forest (#8)	0.4933	0.3302	0.8235	0.3546	0.2380	0.5926	44816.78
KNN, K best selected features (#6)	0.4586	0.3078	0.7664	0.3090	0.2099	0.5189	53034.10
Decision tree (#7)	0.4454	0.3239	0.7693	0.2934	0.2167	0.5101	54839.83
KNN, 2 most correlated features (#5)	0.4028	0.2512	0.6541	0.2394	0.1608	0.4001	64667.35
LR, K best selected features (#4)	0.3795	0.2687	0.6481	0.2028	0.1560	0.3588	64397.05
LR, all features (#1)	0.4066	0.2528	0.6593	0.2068	0.1496	0.3564	62431.92
LR, 2 most correlated features + log scaling (#3)	0.3910	0.2331	0.6241	0.2094	0.1453	0.3547	65901.46
LR, 2 most correlated features (#2)	0.2692	0.2800	0.5492	0.1514	0.1462	0.2976	79468.36

Figure 7: Summary table sorted by percentage of predictions differing by a maximum of 25,000.

A clear hierarchy can be seen from both of the sorted tables. The gradient boosting and random forest models clearly perform best by all metrics, with RMSEs less than 50,000 (the average prediction is at least “usable”), approximately 60% of all predictions within 25,000 of the actual value (that is, “very good”), and over 80% of all predictions within 50,000. All four of the linear regression models and one of the K-nearest neighbor models vary in rank between the tables, but regardless perform the worst of all models, with RMSEs between 60,000 and 80,000, and at most 40% of all predictions within 25,000 of the actual value. The decision tree and remaining K-nearest neighbor model perform significantly better than these five models but are still overshadowed by the first two.

Although both the gradient boosting and random forest models perform very well, the gradient boosting model performs slightly better across all metrics, and is therefore the recommended model for this dataset.

Section 6: Conclusion

The tested gradient boosting model is a good fit for predicting median house prices within this dataset. Although this data is from 1990, the model can easily be applied to a more modern dataset, with uses in areas such as:

- Short-term arbitrage: An investor who sees a large difference between the listed price of a for-sale house and the predicted true price could make a quick profit flipping the house.
- Long-term net profit: A developer can determine whether the costs of building certain styles of houses in an area will be worth their future potential values.

Notes for future research:

- A few more dependent variables could be conducive to both achieving better predictions and understanding conceptually why those predictions were more accurate. For example, a variable for median square footage, under the implication that larger houses are more expensive, could have potentially led to a more correlated density variable in Section 3. California is also known for many natural features other than the coast, such as large forests and skiable mountains, whose vicinity to a block could also affect its house price.
- A more modern dataset would allow the model to prove if it really is robust against the effect of over three decades – much could change in what drives the housing market since then.
- It would be interesting to see if this (or another tested) model would be useful among similar datasets for other states, containing replacement variables for distances to their own major cities and natural features, for example.