Analysis and Forecasting of California Housing

Yucong Chen

Ulster University, Belfast, BT1 2LT, United Kingdom

Abstract. House prices have significant impact on people's daily life, and it is essential for people to have fixed abode, to live, work and social prosperity and stability. Hence predicting House price is a meaningful and big challenge. To achieve this goal, we use California Census dataset in this project to how distinctive features (attributes) can make the house price higher or lower. The main idea of this project is to build a Regression Model that can learn from this data and make predictions of the price of a house in any block, given some useful features provided in the datasets. In the regression task, we applied cross-validation and K-Fold method on Ridege Model, Random Forest, Gradient Boosting models to select the optimal hyperparameters. Then we apply the best selected model on test set, the results show decent performance for Random Forest and Gradient Boosting. The Random Forest performs the best with MSE (Mean Squared Error) 0.290, while it takes training time 14.7 seconds. Although the Gradient Boosting takes the result of MSE is 0.295, it took a shorter training time (2.91s).

Keywords: California Housing; K-Fold Method; Random Forest.

1. Introduction

California Housing Data [Li15] was first used in the paper Pace, R. Kelley, and Ronald Barry." Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297. [PB97]. It contains useful information on the house price to help us understand how does the location impacts. How about the size of the house? And the age? It contains a lot of information that can help us to find the answer. [SBD+22] It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables, and sits at an optimal size between being too toyish and too cumbersome.

In this project, our aim is to predict the median price of household in a block given suitable features (or information) provided.

2. Data Preprocessing and Analysis

Data analysis is important to help us better understand the data structure and the relation between features and target variables. Data preprocessing can turn intractable features like text or categorical features into information that is easier for computers to understand and deal with, which further eases the regression task.

2.1 Data Summary

Here we first summarize the California Housing dataset using visualization and some basic statistics. As showed in Figure 1California Housing dataset contains 20640 rows and each one of them stores information about a specific block. It contains 10 columns with 9 features and one target variable-*median house value*.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Figure 1. Head rows of the California Housing dataset

It is easy to notice that all variables are numerical variables. However, we cannot directly put the whole data into model for training, as some of the variables are Geospatial data. To deal with this problem, we need more work to clean the data.

2.2 Cleaning Data

Table 1 shows the number of NAN values of all variables, we can notice that *total* $_{bedroom}$ has 207 NAN values.

A solution for this problem is to fill the empty values with the median.

Hence, we first found that the variable that is most correlated to "*total bedrooms*" is "*households*."

Therefore, we can use them to divide the records grouped in "blocks" of 20 units. Then we fill the NAN values with the median of the group it belongs to.

The correlation between each variable. The Table 2 shows the correlation values.

The relationship between median income and median house value. The median income is the most correlated variable to median house value with 0.69 correlation coefficient.

A second problem is that we can see some horizontal lines in some particular points, for example 450000, 350000, 350000 and 275000.

2.3 Geospatial Data and Ocean Proximity

The households by sea are usually more expensive than those inland. So, it is reasonable to take the Geospatial data into consideration when predicting the price of household. First, we plot the categorical feature" ocean *proximity*" in Figure 3. So, we plot median house value in a Figure 4 to see how our target behaves. The dark blue spots (marking the more expensive houses) are near the ocean, which highlights the significance of ocean proximity in our analysis. Moving from the ocean to the interior, the price of houses drops substantially. And the northern state approaching does not have dark blue spots, even close to the coast.

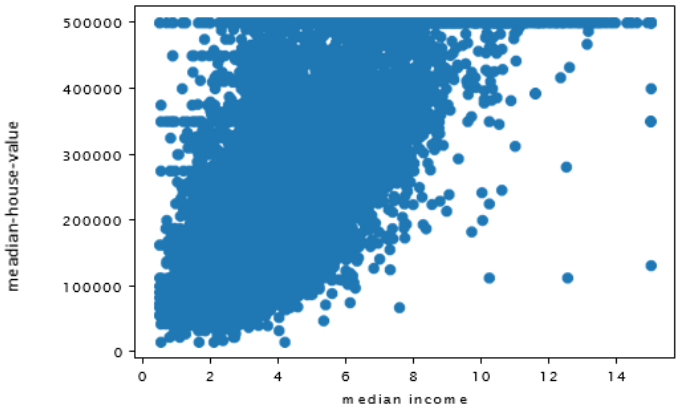

Figure 2. Median house value VS Median income, x

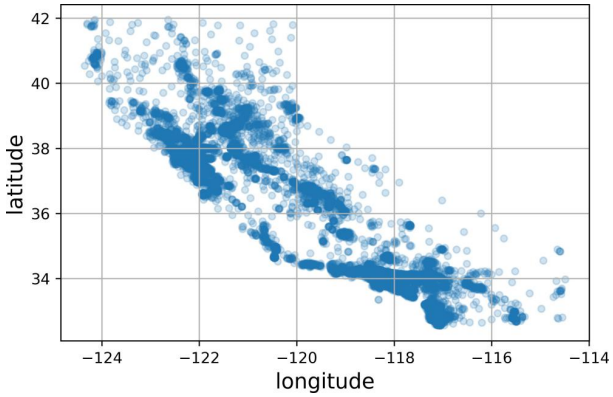

Figure 3. Longitude and Latitude in the map

Table 1. Number of NAN values of all Variables

Variable	Description
longitude	0
latitude	0
housing median age	0
total rooms	0
total bedrooms	207
population	0
households	0
median income	0
ocean proximity	0
median house value	0

Table 2. Correlation of all variables with median house price.

Variable	Description
median house value	1.000000
latitude	0
median income	0.688075
rooms per household	0.151948
total rooms	0.134153
housing median age	0.105623
households	0.065843
households gp	0.065705
total bedrooms	0.050761
population per household	-0.023737
population	-0.024650
longitude	-0.045967
bedrooms per household	-0.046517
latitude	-0.144160

In this model, the categorical value "ocean proximity" is really important, because it is difficult for some models to learn from latitude/longitude. From the observation above, we can see that houses with the categories "*next to bay*," "*near ocean" and "less* 1*h to ocean" are* more expensive than those inland. But they also have a wide range of prices, we can do better than that by introducing new categories. As discussed above, dark blue spots all along the coast except the northern part of the state, where we had cheaper houses even on the coast.

That inspires us to introduce a new category to separate the north coast from the rest: Above the latitude of 38.20, we will create new categories for the "*near ocean*" and "*less* 1*h to ocean*". Hence, we use the additional categorical variable to represent the household in the north or south.

We can see from Figure 5 that the additional variable "*near ocean north*" and" less 1*hour to ocean north" is* incredibly supportive in separating the house value near ocean and less 1hour to ocean it north and south into separate range.

2.4 Normality of the House Price

From Figure 6 that the distribution is severely skewed and not normal. Also, the values are clipped somewhere near 500 000. We can check it statistically.

In statistics, a Q–Q plot [GW68] (quantile-quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. We plot the QQ-plot in Figure 8, and we can see that the tail distribution is far from that of a normal distribution.

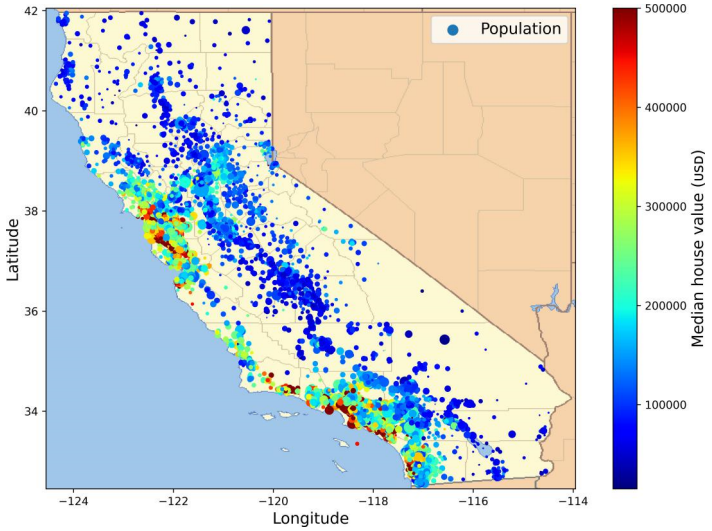

Figure 4. Median house value in the map

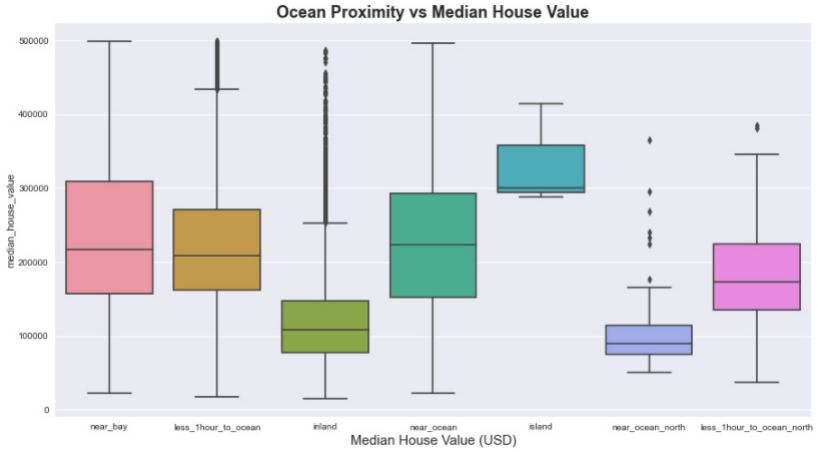

Figure 5. Housing Values per Category after changes

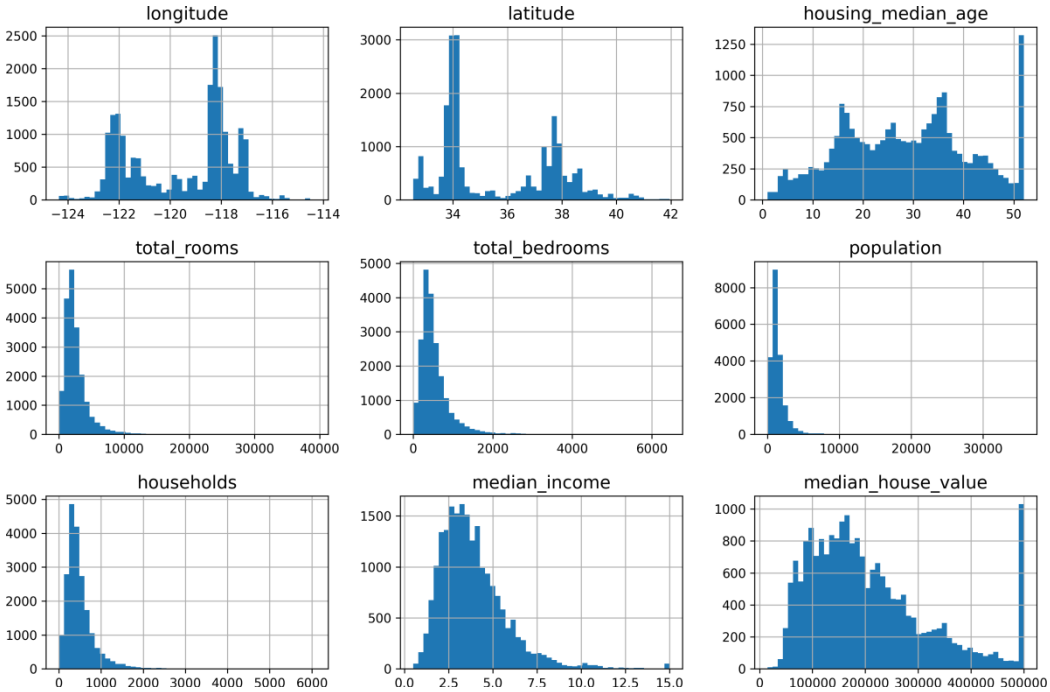

Figure 6. Distribution of House Price

As for statistical test for normality test, D'Agostino's K2 test, named for Ralph D'Agostino, is a goodness-of-fit measure of departure from normality, that is the test aims to gauge the compatibility of given data with the null hypothesis that the data is a realization of independent, identically distributed Gaussian random variables. The test is based on transformations of the sample kurtosis and skewness and has power only against the alternatives that the distribution is skewed and/or kurtic. By applying D'Agostino's K2 test, we get a p-value that is close to 0, hence we reject the null hypothesis test, and conclude that the House value is not normal.

The non-normality of the house value seems frustrating. Luckily, there are some ways to transfer the non-normal variable into normal ones such as logarithm transformation. Applying the logarithm on the house value, we then plot the histogram of the house price in Figure 7. We can see that the distribution is more like a normal distribution with 0 skewness. Although it still has a heavy tail on the extremely low value, and the statistical test still rejects the normality of the log house value, taking the log is still helpful for our regression task. Similarly, some features like 'households,' 'median income,' 'population,' 'total bedrooms,' 'total rooms' in the dataset is also skewed, the log trick can help, and turn the feature distribution into centered one with 0 skewness.

In summary, we have done the following:

Data cleaning was performed. Fill some NAN value with the median of proper selected group;

The median income is highly correlated with median house price with correlation coefficient 0.69;

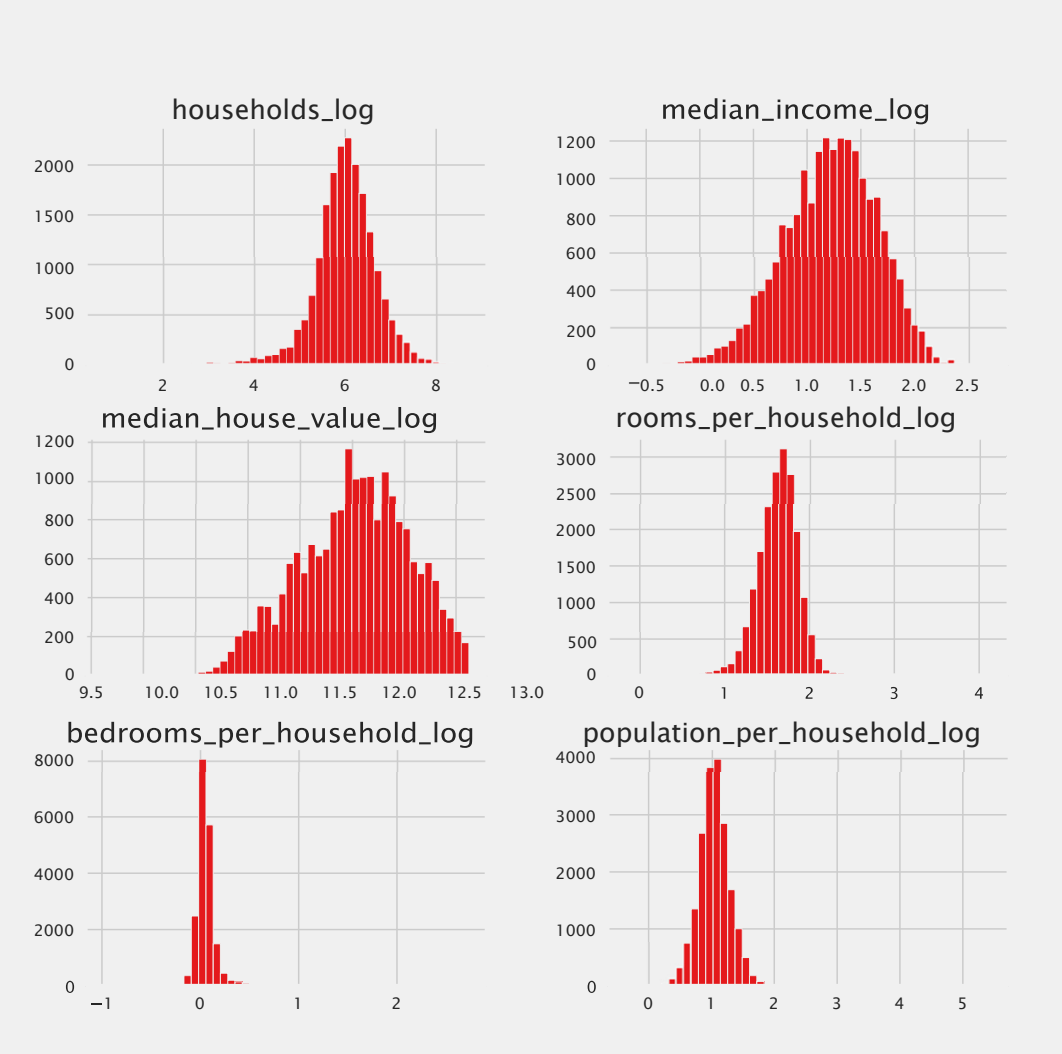

Figure 7. Distribution of Log House Price

Analyzed the impact of the location on the house price and found that households by the sea and in south tend to have higher prices and created additional variable for the north near the sea and north less-1-hour to sea;

Analyzed the normality of the house price and found some log-norm distributed among them;

The corresponding log features are created. Analyzed the distribution of the target feature and concluded that it may be useful to predict log of it;

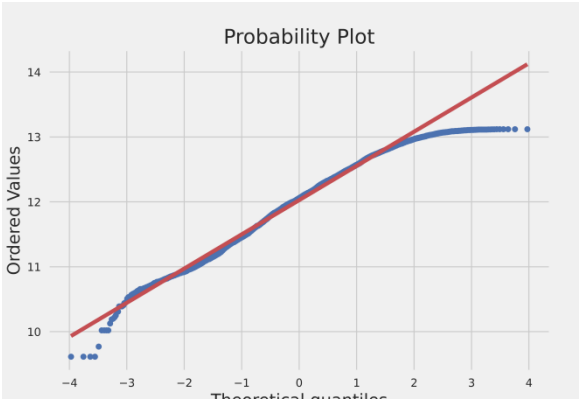

Figure 8. QQ-plot of the House Value

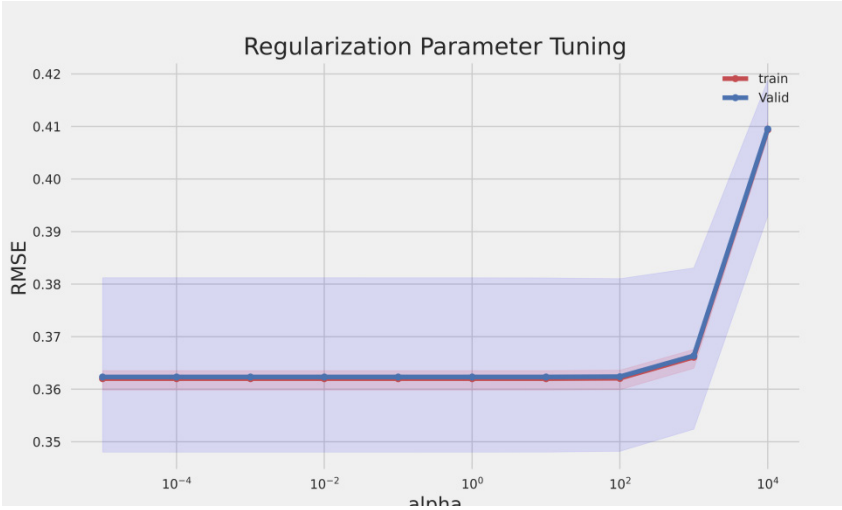

Figure 9. Scores of Ridge Model against different alpha values.

3. Numerical Experiment

In this section, we do the regression task on the dataset to predict the median house value.

3.1 Model Selection

To select a proper model, we compare the model performance by using different models, that is Ridge Linear Model, Random Forest, and Gradient boosting.

Firstly, we introduce these three models:

Ridge Linear regression [HS77] is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It adds a penalty term that penalizes the l2 norm of the coefficients, and hence can be used to prevent overfitting problem.

Random forests [Ho95] or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

Gradient boosting [HTF09] is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.

3.2 Numerical Results

In the numerical experiment, we use train/test split with 20% of the whole dataset as test set. When training these three models, we use cross-validation method to select the best hyper-parameters of each model, then we use the best hyper-parameters for the model to get the score and make prediction on test set.

For Ridge Linear regression model, we use K-fold [JWHT13] method to validate the performance of the linear regression with respect to different choice of alpha parameters. As shown in Figure 9, we can see that curves for train and CV [All74] are remarkably close to each other, it is a sign of underfitting. The difference between the curves does not change along with change in alpha this mean that we should try more complex models comparing to linear regression or add more new features (f.e. polynomial ones)

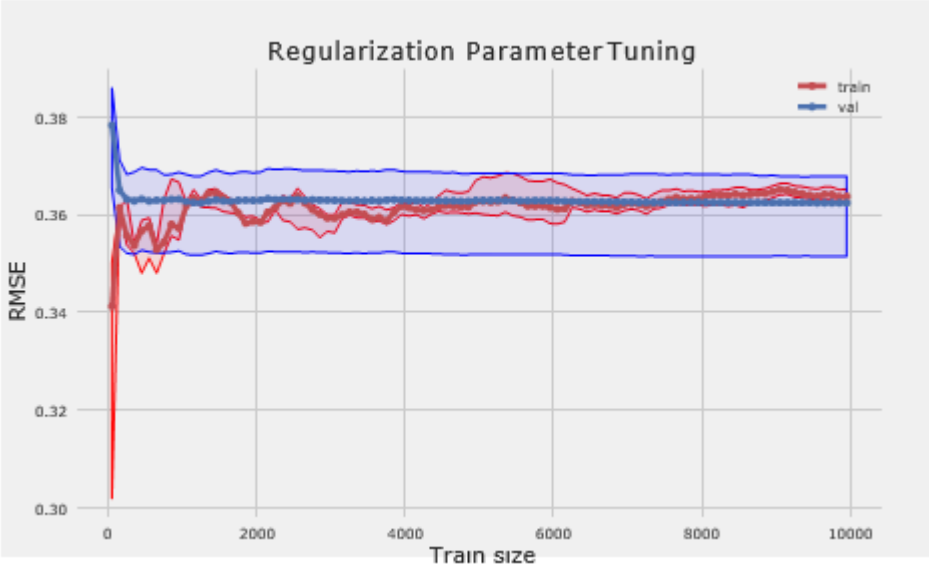

Figure 10. Learning Curve of the Ridge Model with alpha=1

Table 3. Best results for each model

Model	MSE	Time(s)
Ridge	0.37310	**0.17**
Random Forest	**0.29093**	14.7
Gradient boosting	0.29484	2.91

But our prediction does not change when alpha goes below 1. As shown Figure 10 in Learning curves indicate high bias of the model - this means we will not improve our model by adding more data, but we can try to use more complex models or add more features to improve the results.

Similarly, we apply cross-validation and K-fold method to Random Forest model and Gradient boosting model to select the best hyper parameters for those models.

Finally, we train these three model with optimal hyperparameters on the train set and make the prediction on the test, the metrics is computed the mean squared error (MSE) [BD15] between prediction and actual value as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i)^2$$

where y^{pred} is the predicted log value of the median house price and y_i is the actual log value of median house price. We summarize the performance of each model on test set in Table 3. We can see that Random Forest performs the best with MSE 0.290, while it also takes the longest training time, that is 14.7 seconds.

4. Conclusion

To predict the California House Price with the given dataset. Before training process, we have done lot of work on the dataset: We have done data cleaning where we fill some NAN value with the median of proper selected group; We have found that the median income is highly correlated with median house price with correlation coefficient 0.69; We have analyzed the impact of the location on the house price, and household by the sea and in south tends to have higher price and created additional variable for the north near the sea and north less-1-hour to sea. We have analyzed the normality of the house price and found some log-norm distributed among them. We have created corresponding log features. We have analyzed the distribution of the target feature and concluded that it may be useful to predict log of it.

For the model selection, we applied cross-validation and K-Fold method on Ridege Model, Random Forest, Gradient Boosting models to select the optimal hyperparameters. Then we apply the best selected model on test set, the results show decent performance for Random Forest and Gradient Boosting.

References

[1] [All74] David M Allen. The relationship between variable selection and data augmentation and a method for prediction. techno metrics, 16(1):125–127, 1974.

[2] [BD15] Peter J Bickel and Kjell A Doksum. Mathematical statistics: basic ideas and selected topics, volumes I-II package. Chapman and Hall/CRC, 2015.

[3] [GW68] Ramanathan Gnanadesikan and Martin B Wilk. Probability plotting methods for the analysis of data. Biometrika, 55(1):1–17, 1968.

[4] [Ho95] Tin Kam Ho. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE, 1995.

[5] [HS77] Donald E Hilt and Donald W Seegrist. Ridge, a computer program for calculating ridge regression estimates. Department of Agriculture, Forest Service, Northeastern Forest Ex- periment, 1977.

[6] [HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Boosting and additive trees. In The elements of statistical learning, pages 337–387. Springer, 2009.

[7] [JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning, volume 112. Springer, 2013.

[8] [Li15] Yuming Li. The asymmetric house price dynamics: Evidence from the california market. Regional Science and Urban Economics, 52:1–12, 2015.

[9] [PB97] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. Statistics & Probability Letters, 33(3):291–297, 1997.

[10] [SBD+22] Saptarsi Sanyal, Saroj Kumar Biswas, Dolly Das, Manomita Chakraborty, and Biswajit Purkayastha. Boston house price prediction using regression models. In 2022 2nd Interna- tional Conference on Intelligent Technologies (CONIT), pages 1–6. IEEE, 2022.