Application of machine learning on the California Housing dataset to investigate/predict the Median Housing Value

Department of Mathematics and Statistical Science

Python for Machine Learning (CS 577)

SPRING 2022

Moscow, ID, US
kark6289@vandals.uidaho.edu

Abstract— Machine learning models allow users to create and train models using training sets of data and make predictions using test sets. It aids the system in developing artificial intelligence by spotting patterns and boosting the accuracy of the foregoing predictions. I'll be using the California Housing dataset in this paper to make predictions on the Median Housing Value using three machine learning models: Linear Regression, Decision Tree, and Random Forest, as well as preliminary data analysis results. We discovered that of the three models, Random Forest predicted the findings with the highest accuracy, which could be owing to the model's architecture, which uses numerous trees to avoid overfitting the data.

Keywords—Machine learning, California Housing, Prediction, Accuracy, Regression

I. Introduction

According to Forbes, the median housing prices for homes in California have increased by 10 % from February 2021 to February 2022 which shows the current market inflation. [1] So, predicting housing prices would be important asset for homeowners and buyers so they can determine an appropriate time to buy or sell homes. Although the data I will be using for this project is from 1990 U.S. census [5], it gives us some insight into the ways in which machine learning models can be used to make housing value prediction and its importance.

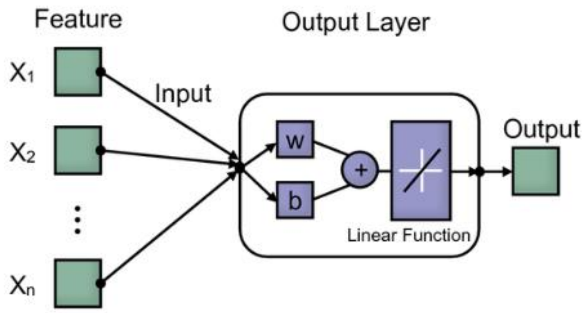
The data consists of housing information in 20640 districts of California which are the instances (rows) along with 9 attributes (columns): longitude, latitude, housing_median_age, total_rooms, total_bedrooms, population, households, median_income, and median_house_value. The data source is in **sklearn.datasets.fetch_california_housing** function according to https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset. The target variable for this dataset is the median_house_value in hundred thousand dollars ($100,000) and I will be using machine learning models to predict this dependent variable. The rest of the variables will be the independent variables.

II. METHOD

This is a supervised learning model because the input and output labels are provided and since the response variable is continuous, I am using regression models.

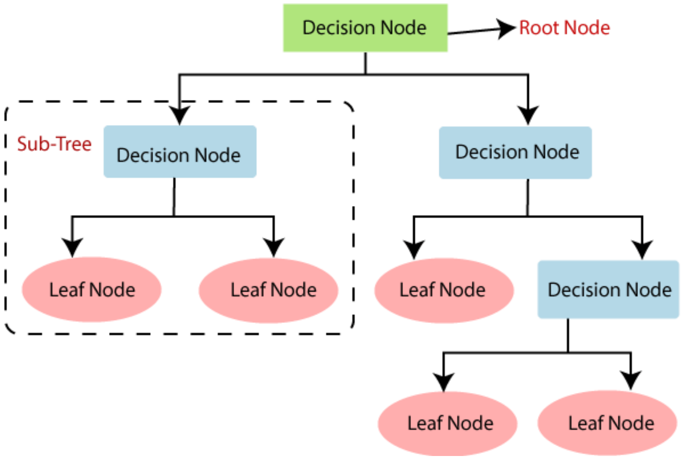
Machine Learning algorithms:

A. *Linear Regression*



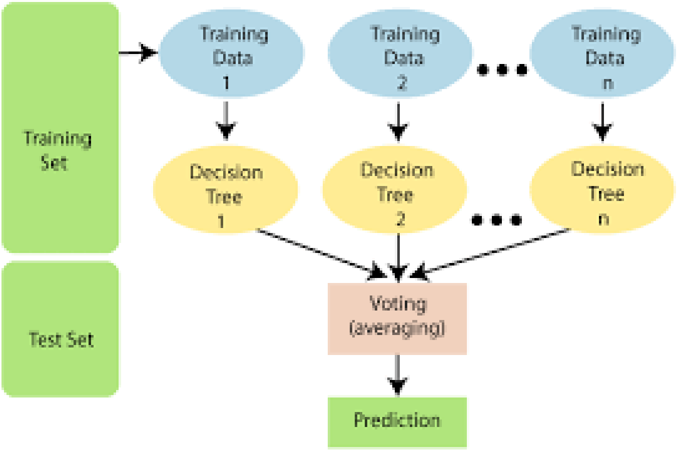
Source:https://www.researchgate.net/figure/Architecture-of-machine-learning-algorithm-a-linear-regression-b-neural-network_fig1_350319555

B. *Decision Trees*



Source:https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

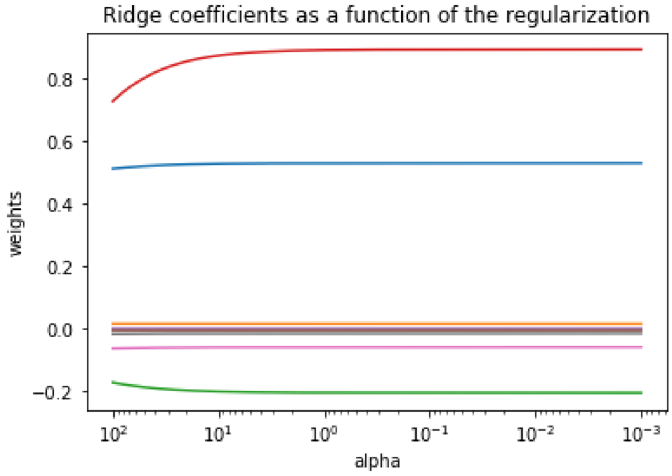
C. *Random Forest*



Source:https://www.javatpoint.com/machine-learning-random-forest-algorithm

I will be using these algorithms to design and train the model and to see how accurately the median housing value in California can be predicted and if there are optimizers that can produce the lowest mean square error (MSE). I will use libraries such as Numpy and Scikitlearn throughout the analysis and matplotlib and plotly for visualization. Finally, I will be comparing the MSE and R-square value of the different models that I will incorporate to see which one works the best to predict the median housing value.

D. *Parameter Setting:*



Models	
Linear Regression (Ordinary Least Square with Ridge Regression)	Alpha = 1*

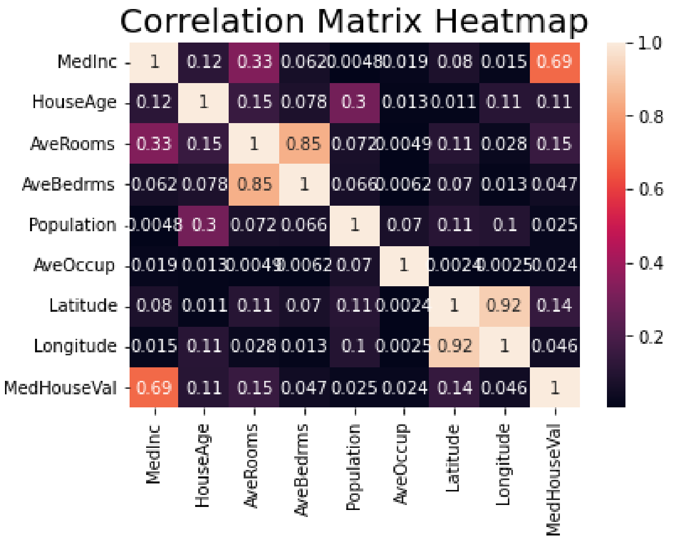
Decision Tree	criterion = 'mse',
	max_depth = 8

Random Forest	criterion = 'mse',
	n_estimators = 13,
	max_features = 3,
	max_depth = 8

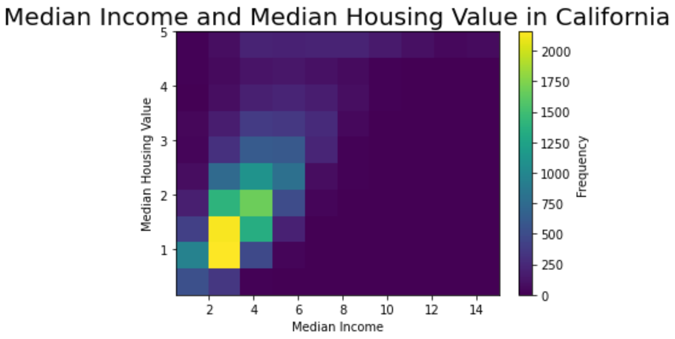
*The result does not change even if we increase the value of alpha which means that there are no outliers to work with.

III. Experimental Results

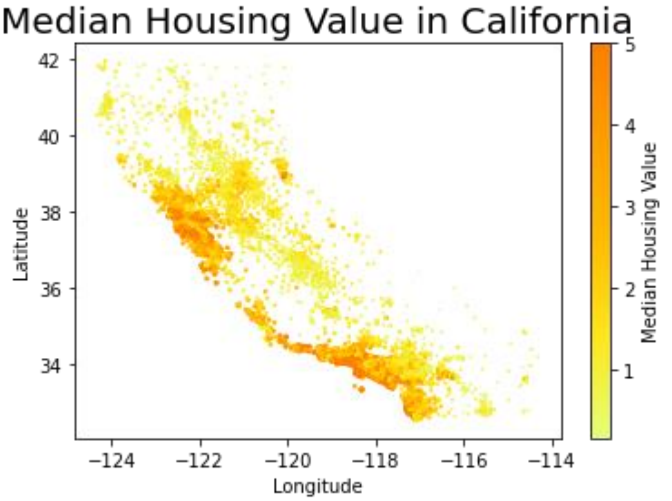
A. *Fig 1: The correlation matrix shows how strong or weak the relationship is between each variable to every other variable provided in the data. Since our response variable is Median Housing Value, we can see that it is highly correlated with Median Income and slightly correlated with Average rooms and Average Bedrooms.*



B. *Fig 2: To further investigate the relationship between Median Income and Median Housing Value, we look at the 2d histogram and see that as the median income increases, more people tend to have houses with higher value even there are very few people that actually have high income and high housing value. Most number of people lie around the lower income range and lower housing value.*



C. Fig 3: The scatterplot shows the distribution of Median Housing Value in different districts of California. We can see that the coastal areas (the areas closer to the ocean) generally have a higher Median Housing Value compared to the rest of the areas.



D. Testing and Training Performance

Models	MSE T rain	MSE Test	R^2 Train	R^2 Test
Linear Regression (Ordinary Least Square with Ridge Regression)	0.517	0.543	0.611	0.593
Decision Tree	0.325	0.443	0.756	0.667
Random Forest	0.264	0.333	0.802	0.751

IV. Conclusion and Discussion

As a result, we can see that Random Forest outperforms the other two models in terms of accuracy and has the lowest Mean Squared Error (MSE) since it avoids overfitting the data by using numerous trees and calculating the average of those trees. It also focuses on feature selection, focusing on the most essential features that contribute the most to the data and the accuracy of prediction. According to published reports, the greatest R2 test score for Random Forest is 0.8, which is similar to my results. [2] However, utilizing m-fold cross validation to randomly select the data into training and testing sets to provide greater model training and testing accuracy could improve my results even further.

Acknowledgment

I would like to thank Dr. Min Xian for providing guidance and comments throughout the class and for the project.

References

[1] DePietro, Andrew. “California Housing Market Report 2022.” Forbes, Forbes Magazine, 14 Apr. 2022.

[2] Eliezer, Daniel. “End-to-End Maching Learning Project: Predicting House Prices in California.” Medium, Analytics Vidhya, 13 Sept. 2021, https://medium.com/analytics-vidhya/end-to-end-maching-learning-project-predicting-house-prices-in-california-2e95171d49cc.

[3] Hunter, John, et al. “Matplotlib.pyplot.hist2d¶.” Matplotlib.pyplot.hist2d - Matplotlib 3.5.0 Documentation.

[4] InfoPython. “How to Plot Longitude and Latitude Data Using Python.” InfoPython, 8 Jan. 2022.

[5] Pedregosa et al. Scikit-learn: Machine Learning in Python JMLR 12, pp. 2825-2830, 2011.

[6] Scikit. “Sklearn.model_selection.GRIDSEARCHCV.” Scikit, 2011

[7] Szabo, Bibor. “How to Create a Seaborn Correlation Heatmap in Python?” Medium, Medium, 26 May 2020.