Predicting Median House Value

Caleb Kelly

Computer Science Department, University of The District of Columbia

CSCI 421: Machine Learning

Timothy Oladunni

April 30, 2021

**Abstract**

The study conducted in this paper investigates the accuracy of predictive models when predicting the median value of a house. A better understanding of which models predict the median value the most accurately will improve future research projects reducing the possibility of missing important models. Five models will be used in total, these models will be executed once with forward selection processing and once without to compare which process created a more accurate model. The accuracy of each model will be compared with the model’s root mean squared error. The data I used comes from Boston Massachusetts. The results show that sometimes if possible, using all the attributes provided in the data can lead to more accurate results.

*Keywords:* Predictive models, accuracy, median value, forward selection, Boston

1. INTRODUCTION

The real estate market is a market that has been studied since it was created. Just like the stock market there are people all over the world trying to predict the value of a house. Predicting this number can be difficult, many reputable companies are able to predict small percentage increases over long amounts of time due to market trends but predicting a specific houses value can be more challenging. Like any market the real estate market can be very unpredictable. Being able to predict the market could lead to a lot of profit made from real estate so it is a common thing to try and predict what a house could be sold for.

Buy low and sell high is the name of the game when it comes to real estate but understanding when a house is considered low or high pricewise can be difficult to predict. Over the course of the fall semester in my machine learning class we learned how to use many different mathematical prediction models and how to gauge their accuracy in what the model is predicting. I am going to run the Boston housing dataset in two different processes and compare the results to see what predictive models from which process are most accurate at predicting the median house value.

1. METHODOLOGY

*2.1 Data Exploration*

The dataset being used comes out of Boston Massachusetts collected by the U.S Census Service. The dataset is small with only 506 cases, but this was selected for that reason as going through data that has samples from multiple different locations could lead to a higher root mean squared error and could lead to unnecessary attributes from the data. The first step to creating this process is labeling the data and selecting your response variable, for our research the response variable selected is the MEDV (median house value). Once the response variable is set, we put the data through two different processes.

*2.2 Process Explanation*

The first process is using Forward Selection which is a method that selects the attributes that increase the model’s performance. Forward Selection goes through rounds of running the predictive model adding an attribute each round, if the attribute decreases the performance of the model the attribute is not selected. If the attribute increases the performance of the model, it is kept and continued throughout the process with those attributes. Depending on the dataset and what is being predicted Forward Selection is a method of increasing the prediction models accuracy. Once Forward Selection has been executed fully and the attributes are selected, we then run the selected attributes through the five models I chose for this research.

*2.3 Linear Regression*

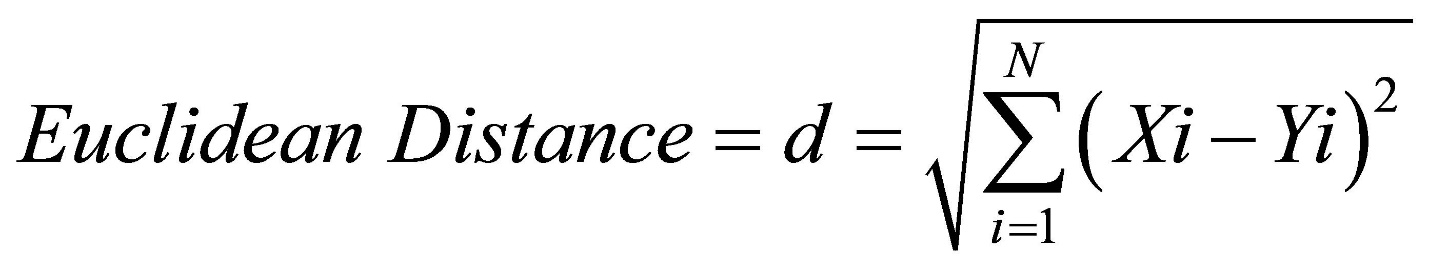
Linear Regression is the first model the data gets tested on. It is a mathematical formula that shows the correlation or relationship between two variables plotting it on an X and Y graph. Linear Regression is mathematically expressed as such;

ŷ = b0 + b1x

Where ŷ represents the predicted value of the dependent variable. X is the independent variable, b0 is a constant, and b1 is the regression coefficient.

*2.4 KNN*

The second mathematical model the data goes through is k-NN (K-nearest neighbor). KNN is a regression and classification algorithm that uses points nearby to generate a prediction. KNN takes a point and finds the K-nearest points allowing it to predict a label for that point. The value of K is up to the user for our example we chose five nearest points. KNN can be mathematically explained using the Euclidean distance concept. This concept is a mathematical formula to measure how far our first data point is from the training set and then calculate the Euclidean distance for the prediction data. Euclidean distance is mathematically expressed as such;



Where X and Y are the test points and data points, and “i” is the number of attributes from the data set.

*2.5 LASSO*

The third model the data is used for is LASSO (least absolute shrinkage and selection operator). Lasso is very similar to linear regression, but lasso uses a technique called shrinkage. Shrinkage is when you shrink or regularize the coefficients to prevent overfitting, this allows LASSO to work better on different datasets. We used LASSO to compare with the other regression models and get a better insight of our data.

*2.6 Ridge Regression*

The fourth model the data is used for is Ridge Regression. Ridge regression is a regularization technique used to analyze multicollinearity data. Multicollinearity is a collinear association between to variables. It is regarding situations where more than two variables in a regression model are very linearly related. Ridge Regression is used in these situations because the mean square estimators are usually smaller than the least square estimators derived in other regression models. Similar to LASSO it is a regression model that is beneficial to use in certain situations depending on the data.

*2.7 Decision Tree*

The fifth and final model the data goes through is a decision tree. A decision tree is an algorithm where nodes represent the attributes from the data creating if-else conditions. The top node is the root node depending on the conditions of the data you move through the tree branches following the if-else conditions until you reach a leaf node. 

1. CONCLUSION

In order to judge the accuracy of each model we have to compare the root mean squared error. This number gauges how closely our models were able to predict MEDV.

|  |  |
| --- | --- |
| MODEL | Root Mean Squared Error (lower = better) |
| Linear Regression | 5.431 |
| LASSO | 5.431 |
| Ridge Regression | 5.431 |
| Decision Tree | 2.778 |
| KNN | 5.076 |
| Linear Regression W/ Forward Selection | 5.651 |
| LASSO W/ Forward Selection | 5.651 |
| Ridge Regression W/ Forward Selection | 5.651 |
| Decision Tree W/ Forward Selection | 3.922 |
| KNN W/ Forward Selection | 5.137 |

Based on the Root Mean Squared Error we can see the best performing model is our Decision Tree model without using forward selection. All models from the process that does not use forward selection were more accurate than the ones that did use forward selection. It is surprising that forward selection caused our models to perform less accurately, this could be because each attribute holds a significant correlation to the response variable and leaving any out leads to a less accurate prediction. Regardless of the cause the performance speaks for itself the simpler regression models were outperformed by the Decision Tree model and KNN.