XGBoost for Boston Real Estate: A Baseline Model

January 23, 2023

1 Introduction

In this study, we present a baseline model for predicting housing prices in Boston using the XGBoost algorithm. The Boston Housing Prices dataset contains a variety of information on properties in the Boston area, including features such as crime rate, average number of rooms per dwelling, and property tax rate. We will use this dataset to train a XGBoost model and make predictions on the median value of owner-occupied homes. This will serve as a baseline as we work to improve model results.

2 Analytic Approach

- Target Variable

  MEDV: The median value of owner-occupied homes in $1000's

- Input variables:

  CRIM: per capita crime rate by town. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.. INDUS: proportion of non-retail business acres per town. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). NOX: nitric oxides concentration (parts per 10 million). RM: average number of rooms per dwelling. AGE: proportion of owner-occupied units built prior to 1940. DIS: weighted distances to five Boston employment centres. RAD: index of accessibility to radial highways. TAX: full-value property-tax rate per $10,000. PTRATIO: pupil-teacher ratio by town. B: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town. LSTAT: % lower status of the population

3 Model Description

XGBoost (eXtreme Gradient Boosting) is a powerful and popular machine learning algorithm that is used for both regression and classification problems. It is an implementation of the gradient boosting algorithm and is known for its speed and accuracy. The basic idea behind XGBoost is to train a series of simple decision trees and then combine them to create a more powerful model. Each decision tree is trained to predict the residual error of the previous tree, i.e. the difference between the true value and the predicted value of the previous tree. By training multiple decision trees to predict the residual error, the algorithm is able to capture complex relationships in the data and improve overall accuracy of the model.

The model has several important parameters that can be adjusted to improve performance.

- Model Parameters
  - learning rate: the step size at which the algorithm learns from the data.
  - max depth: maximum depth of each decision tree in the ensemble.
  - min child weight: minimum number of instances needed to be in each node.

- subsample: the fraction of the data used in each iteration.
- colsample bytree: the fraction of features used in each iteration. .
- objective: the loss function used to train the model.
- eval metric: the metric used to evaluate the model's performance.
- reg alpha : L1 regularization, penalizes the features which increase cost function.

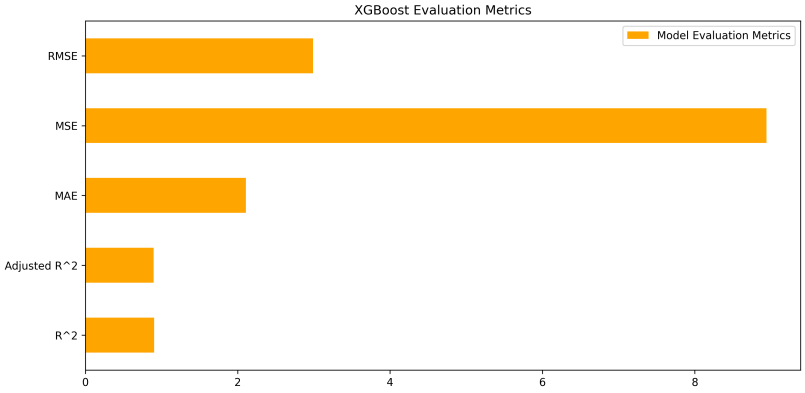
- Model Flow

4 Results (Model Performance)

The baseline model performance on the test set resulted in a MSE of 8.9, MAE of 2.1, RMSE of 2.99, R^2 of 0.9 and adjusted R^2 of 0.89, these values are plotted below for visual reference.

The relatively low Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error indicate that the model predictions are close to the true values. The high R^2 and adjusted R^2 values mean the model has a good fit to the data and explains a large proportion of the variance in the housing prices. It's worth noting that a high R^2 value does not necessarily guarantee a good model, it can be affected by the number of features, overfitting, and even randomness.

Based on these metrics, the XGBoost baseline model seems to have a good performance on the Boston Housing dataset.

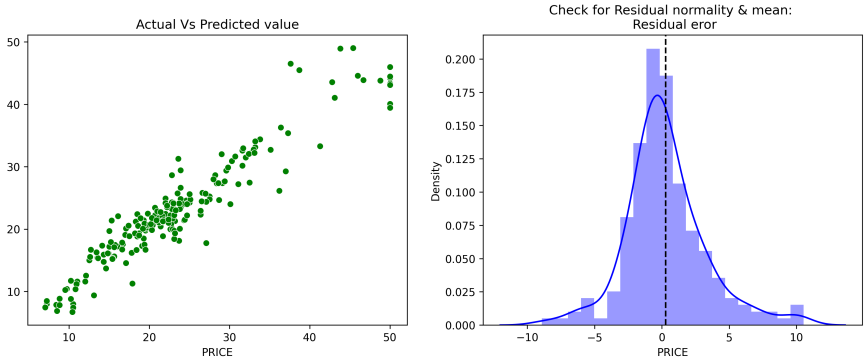


5 Model Understanding

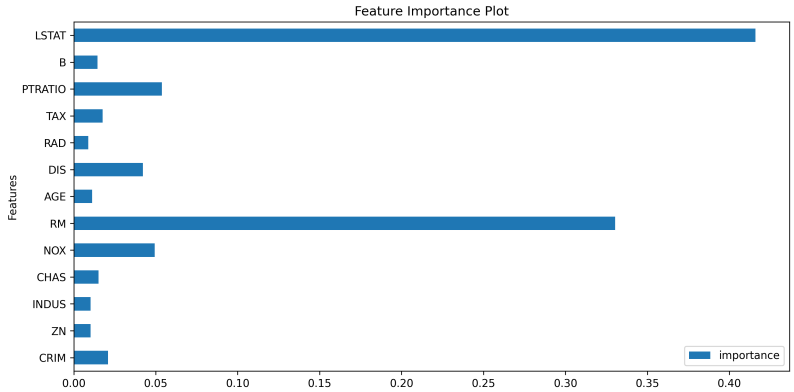
Checking for residual normality and a mean value of zero refers to evaluating the distribution of the residuals (i.e., the difference between the predicted values and the true values) of the model and verifying they center around zero.

A normal distribution of the residuals is an indication that the model is correctly capturing the underlying patterns in the data while a mean value of zero indicates that the model is not systematically underestimating or overestimating the true values.

The residual normality plot below confirms this check on the baseline model meaning it is appropriately predicting housing prices. Furthermore the scatter plot confirms that the model accurately predicts housing prices as the points appear nearly in a straight line.



The models feature importance tells us which variables are most correlated with the housing prices. The feature importance values are normalized and can be used as an indicator of how much each feature contributed to the model's predictions. According to the plot below we see LSTAT (% lower status of the population) and RM (average number of rooms per dwelling) have the highest feature importance values of 0.41 and 0.33 respectively. We can infer therefore that LSTAT and RM play a large role for this baseline model.



6 Conclusion and Discussions for Next Steps

Based on the evaluation metrics and visual representation of the model performance, it can be concluded that the XGBoost model is a feasible solution for the Boston Housing dataset. The model has a good fit to the data and explains a large proportion of the variance in the housing prices. However, it's important to note that these values are only one aspect of the model performance and should be considered in conjunction with other factors such as model interpretability, overfitting, and generalization performance.

XGBoost, like any other machine learning algorithm, is susceptible to overfitting. This occurs when a model is too complex and captures noise in the data, leading to poor generalization performance on unseen data. However, from the evaluation metrics provided it doesn't seem that this model is overfitting. To ensure that the model does not overfit, it's important to regularly evaluate the model's performance on unseen data and use techniques such as early stopping, regularization, and cross-validation.

Some other features which can be generated from the current data include the ratio of the house's age to the average age of the houses in the neighborhood or the proportion of the house's price to the average price of houses in the neighborhood.

Other relevant data sources that could be used to help the modeling include demographic data on the population of the neighborhood, employment and income data for the area, transportation data such as the proximity to major roads, public transportation, and airports, and lastly, additional data on the house's condition and amenities, such as the number of bathrooms, square footage, and presence of a garage or pool.

This baseline model and its performance evaluation metrics serves as a standard to measure our own implementation.