Udacity Machine Learning Engineer Capstone

Predicting Sale Prices of Boston Homes

By Greg Gompers

Domain Background

Predicting the price of goods for sale is in the Regression, Machine learning field. This is a field of research uses historical data to predict future outcomes. In the case of regression these outcomes are usually in the form of a number, or price, as in the case of my project of predicting sale price of homes in Boston.

Problem Statement

The problem is that it is very hard to predict the sale price of a home. This can be solved with a regression algorithm, when given the right data about previous homes and their sale prices. I can measure this by the accuracy of my regression machine learning model.

Datasets and Inputs

The Data set I will be using as input for my regression machine learning model is the Boston Housing Dataset. This is originally a UCI machine learning repository dataset made in 1978, which I found on the Kaggle website. The Data set has 506 entries, each with the final sale price, and 13 features which measure relevant information about each house.I will use the relationships between the 13 features of the houses to create a regression learning model, to predict the final sale price.

Link to dataset download and information:

https://www.kaggle.com/c/boston-housing

Solution Statement

My solution is to use a machine learning regression model to predict the final sale price of the homes, based on the relationships between the recorded relevant features of the homes

Benchmark Model

At this Kaggle link: https://www.kaggle.com/c/boston-housing/leaderboard

multiple benchmark models can be found, from many teams across the world, to show their performance on the same Boston Housing Dataset

Evaluation Metrics

I will be using Root mean Squared error, or RMSE as an evaluation metric to check my models performance on a testing set of data which has never been seen by the model. This will explain the average accuracy of the model in correctly predicting the sale price of a home.

Project Design

I will be loading in the dataset, I will analyze the summary statistics of the columns in the dataset, then I will visualize the data distributions of each column, then I will create a list of about 15 regression machine learning models, and I will evaluate them with K-fold cross validation, optimizing for RMSE value. I will then visualize all models performance, then choose the top 3 models to perform hyperparameter tuning on, and then I will select the final model as the one which performs best on an unseen test set of 20% of the data