Udacity Machine Learning Engineer Capstone

Blog Post Submission: Predicting Sale Prices of Boston Homes

By Greg Gompers

Project Overview

In this project, I will be able to predict the sale price of homes in Boston.

I will be able to do this by using a data set of 14 measurements, recorded on 506 homes, by the UCI machine learning repository in 1978.

I will examine the relationships between these measurements, first with visualizations of these relationships, and then,, with multiple regression machine learning algorithms. I will use RMSE as my evaluation metric with 10-fold Cross Validation. I will graph the results, and then I will hyperparameter tune the highest performing models, and finally, use an unseen test set of 20% of the data, to choose the highest performing model overall.

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.

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.

Data Exploration

To explore the data in the set, I first look at the top 20 rows using the pandas .head(20) function, then I use the .describe() function to give me a list of eight summary statistics for each feature column.

After this, I Visually display the data using histograms, density plots, and a correlation matrix to be able to visually understand relationships in the data

(All data explorations and visualizations can be found in Notebook File)

Exploratory visualization

While exploring the data visually, I was able to see that multiple of the feature columns had normal distributions, and some had exponential distributions.

This is what Inspired me to run the regression algorithms before and after standardizing the data, to see you how this could improve their ability to find patterns in the data. This ended up being a very good idea, as standardizing the data set created a very high performing KNN model

Algorithms and Techniques

For this Project, I have chosen a list of about 15 different regression machine learning algorithms.

Because it is difficult to know which algorithm will work best before trial and error, I created a loop in which all 15 regression machine learnIng models were tested sequentially using 10 fold cross validation.

I also saved the last 20% of the data set, to be used as a on final test on unseen data.

Benchmark Models

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

Data preprocessing

The Boston housing data set from the UCI machine learning repository, created in 1978, did not require any pre-processing steps to be usable.

This is because there are no missing values, and all of the data is already in numerical format.

Implementation

The implementation of the list of models, and the K-Fold cross validation used in the project can be seen in the included python Notebook

Refinement

After going through the list of about 15 regression models, I followed through to hyperparameter tuning on the scaled KNN model, the gradient boosted model, and the extra trees model.

Are used to grid search to find the best hyper parameters of these models, To run with on the unseen set of test data.

Model Evaluation and Validation

I used 10 fold cross validation to initially test the models, and then tested again on a unseen 20% of the data set.

Justification

The list of benchmark models can be found on the cargo website, which also show the competition results using root mean squared error.

As my model my final gradient boosted model registered in at an MSE of 11.97, this is a statistically significant result as opposed to making a random guess over the continuous value range of home prices. A random guess would have a much higher RMSE value because The range of continuous values Is so large.