Anush Pogosyan

Analysis of model predicting Boston house prices:

1) Statistical Analysis and Data Exploration

There are 506 houses in a data set.
There are 13 features in a data set.
Minimum price of a house: $5K.
Maximum price of a house: $50K.
Mean price of a house: $22K.
Median price of a house: $21K.
Standart deviation $9.19K.

2) Evaluating Model Performance

Mean squared error was chosen as a metric to analyze errors. Although mean absolute error can also be used, squaring the error imposes a higher error gradient for higher deviations from the mean than for smaller deviations, thus, forcing the model to chose parameters such that it accounts for high deviations from the data more (such as outliers). Similarly to the mean absolute error, median absolute error was not chosen as a performance metric (even though the model is more robust to outliers, it still does not enforce precision as strictly as squared error). Explained variance and coefficient of determination give similar information - on what fraction of data can be explained by the model. However, they would be applicable if data was concentrated more close to one line - making the residuals smaller. In this case, when all residuals are relatively small, it would be more preferable to use coefficient of determination than MSE, because the latter will not make small errors significant enough. In addition, R squared is not an error metric, but a measure of goodness of fit, there fore it would not be applicable. Thus, in the case where data is more spread rather than aligned to one line, mean squared error is the best metric to estimate accuracy of the model, since it shows how on average the predicted values deviate from true values, by penalizing bigger deviations more.

Splitting data into training and testing sets is the key point in machine learning. This procedure helps avoiding overfitting model, which might happen if the model is validated on the same dataset, as well as will provide false optimistic results on how well the model will perform on new data points. To help evaluate the performance of the model, training data is usually further split into training and validation sets (this process is called cross validation), this operation is performed multiple times using random split, which helps reduce overfitting.

Grid search provides a mechanism to run a model using provided set of ranges of parameters and chose the best fit among them. The default value for CV parameter in grid search is 3, which means that the data will be split in 3 parts, 2 of them will be used as part of training set and one left for validation purposes, this algorithm will be run 3 times to increase the probability of each point to be included in testing as well as training sets. CV parameter can also be increased, in this case the variance of the resulting model will decrease. The best assessment of the model can be achieved by leave-one-out cross validation, however, it is an expensive procedure, since the model will be evaluated k-1 times. More commonly used parameters for k are 5 or 10. In this case 5 was chosen as a tuning parameter to perform more sampling from training data and thus, more accurately reflect model's performance. Grid search uses provided scoring function to estimate the fit of the model. Some of the scoring function can be minimization of mean squared error or maximization of coefficient of determination.

3) Analyzing Model Performance

The increase of training size leads to decrease of testing error and increase in training error up to a point where both training and testing errors converge - when there almost as many data points from the population that the size of a training set is reaching the total dataset. Few data points lead to small training error, since it is easier for the model to find the coefficients of a function that will describe all the points (ex., having 2 data points gives a perfect linear fit and will result in 0 training error). However, when the model is tested on testing data, the testing error is high, since the initial few points may not be representative of the whole dataset. With the increase of data points, training error is increasing, because for a fixed model complexity, the model cannot be tuned to describe each data point in the training set. However, the testing error is decreasing, since increasing number of data points approach the real data set and thus, the model can be tuned to describe more data and testing it on other datapoint from the testing set will not give high error. Eventually, training and testing error converge to a point where testing error cannot decrease anymore, because of the noise in the data and bias, i.e., the model is not ideal for each data point as well as data points are not ideally fit to any models.

When the model is fully trained, in case of fewer features (max depth = 1), it suffers from high bias or underfitting, which means that the model is not complex enough to describe all data points even in the training set, hence the training error converges to some value significantly greater than 0. On the contrary, in case of the max number of features used (max depth = 10), the model suffers from high variance, or overfitting. Since the model is complex enough to describe most of the data points in the training set, it is very sensitive to every single observation, and even though the training error converges approximately to 0, with new training data the model will look differently, because it will adjust to certain data points, hence the testing error is increasing starting from certain model complexity, indicating that the model became too overfit to be able to generalize correctly.

Increasing model complexity leads to decrease of testing error and hight variance - overfitting of the model. This can be interpreted in a way that each data point has a significant impact to the model and if some data points are removed, the model will change dramatically, which is undesirable for general model. Test error will be decreasing with the increase of model complexity till some point, and then will start increasing again. The best fit will be number of features (model complexity) that results in the minimal testing error. Based on analysis of the model complexity graph, testing error is minimal when max depth is equal to 6 or 7. That means that there are just enough features to describe the training data relatively precisely (training data is relatively low) and minimizes testing error.

4) Model Prediction

To test the model grid search was used with max depth parameters from 1 to 10. CV parameter was set to 5, which splits data into 5 sets, one of each is used for validation, this algorithm is run 5 times and by the end of the last iteration the model will have the best estimator based on scoring function (minimization of mean squared error). However, since the process of splitting data is randomized, doing this once might not show the best result, so the GridSearch is set to run 10 times, each time the best parameter and predicted value are stored in corresponding arrays, then the mean of the arrays is taken to evaluated average best parameter and average predicted value.

The predicted value of the house with given parameters is $20468, which stays in line to the statistics of the data: the mean is $22000 and standard deviation $9190, thus the predicted price is very close to the mean and lies within 1 standard deviation from the mean.