Supervised Learning - Analysis and prediction of house prices

Poddaturi, Dinesh R

1/30/2022

In this project I use multiple supervised learning methods to carefully analyze and predict house prices. Originally this project was a part of Kaggle data science competition. Although I didn't win the competition, I learned a lot during and after the competition. I believe in the mantra "learning by doing" (coined by an American philosopher *John Dewey*). So I have been working on this project by making multiple changes, trying different algotithms, eventually improving the prediction power.

About the data: The data are free to download from Kaggel website. I downloaded the data a while ago and has been working with the same data. There are two different files of data; (1) Training data containing the price of the house and house characteristics (e.g., the year it was constructed, number of bed rooms, number of bathrooms, latitude and so on) and (2) Testing data containing only the characteristics of the house. Our goal simply is to predict the house prices in the testing data. In order to do that, first we analyse the training data using multiple supervised learning methods and use the models to predict the house price in the testing data.

One could immediately say this is not a classification problem. Note that the house prices are not boolean i.e., 0 or 1, instead they are discrete variable (one could argue it is a continuous variable, but I stay away from that argument in this analysis). So this could be considered as a regression problem. From my rigorus training in Economics, I can immediately identify a problem with the data. People choose to live in a particular area. It could depend on the school district (people with kids), location (near to a metro or a mall), clean air, less noise pollution and so on. We cannot observe all these in the data (although we have latitude and zip code, we cannot observe on what an individual makes choices). Hence, there could be endogeneity in the data. In this work, I do not focus on that issue. The primary objective of this work is to use supervised learning methods to predict the house prices. So, I stay away from endogeneity and self selection issues (coming mostly from economics).