Abstract

This paper seeks to predict house prices using Multiple Linear Regression using four variables from the Ames Housing Dataset De Cock (2011). The Ames Housing Dataset contains 2930 housing records with 80 explanatory variables from the Ames, Iowa Assessor's Office. The question we want to ask is: Can we reasonably predict house prices only using a few fields that most people could easily access. The fields we chose both are external factors, such as distance from school and subdivision, and internal factors such as Ground Floor area, Basement area, overall condition, etc. We will explore if linear regression is a suitable model for home price prediction and what attributes contribute the most to the home sale price.

Introduction

Home prices are an important factor in a society's overall economic health. Historically, purchasing a home gives an individual tremendous economic advantage over a rentor, since the home buyer can build equity into their property. The housing market is a societal and economic indicator by showing how that particular society values homes and home ownership. Indeed, according to Olga and Antonios (2019), the housing market can provide key insights into the economic health of that society beyond normal banking indicators. Hence, the ability to accurately predict the price of a house will allow the prospective buyer to better understand what attributes of a house are most valuable, and also what attributes of a home contributes most to its sale. The housing dataset we will be using is the Ames Dataset (De Cock (2011)), a dataset of over 2900 houses and attributes of the home such as condition, location, ground floor area, sales price, etc. The dataset can be found from the Ames, Iowa assessor's office and is readily accessible for anyone to use.

Predictive model

While more advanced machine learning techniques such as random forest regresser and extreme gradient boosting often have better results for predicting home prices, these techniques require a more advanced understanding of machine learning principles that a layperson would not necessarily have. We will investigate internal home factors such as ground floor living area and basement area and external factors such as distance to school and subdivision.

The model we will use is a simple multivariate linear regression model. A linear regression model can be described as $y^{(i)} = W_1x_1^{(i)} + W_2x_2^{(i)} + \cdots + W_nx_n^{(i)} + b^{(i)} + \epsilon^{(i)}$, where $y^{(i)}$ is one of the desired predicted variables, $x_k^{(i)}$ is the corresponding explanitory variable, $b^{(i)}$ is the intercept/bias, and $\epsilon^{(i)}$ is the residual between $W_1x_1 + W_2x_2 + \cdots + W_nx_n + b$ and y. Linear regression is relatively easy to understand and requires zero hyperparameter tuning. It's a type of model that someone could easily build in excel or another spreadsheet type of software. Housing data such as the Ames dataset is also publically available through a city or county assessors office, thus almost anyone can run this type of model with only a little bit of research.

References

De Cock, Dean. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education* 19 (3).

Olga, Bormpotsialou, and Rovolis Antonios. 2019. "Housing Construction as a Leading Economic Indicator." *Studies in Business and Economics* 14 (3): 33–49.