House Price Analysis
Fall 2023

Background

Using the USA house price data, determine the factors that influence house price.

Dataset

Use the dataset USA-housing.xlsx spreadsheet. This data is a Kaggle dataset on U.S. house prices.

Housing dataset

SalePrice	final sale price
LotFrontage	width of lot on street
LotArea	square feet of lot
OverallQual	quality of house on scale of 1 to 10
OverallCond	condition of house on scale of 1 to 10
YearBuilt	calendar year of construction
BsmtFin	square feet of finished basement
TotalBsmtSF	total square feet of basement (finished and unfinished)
1stFlrSF	square feet of first floor
2ndFlrSF	square feet of second floor
LivArea	square feet of living area
BsmtFullBath	number of full bathrooms in basement
BsmtHalfBath	number of half bathrooms in basement
FullBath	number of full bathrooms above ground
HalfBath	number of half bathrooms above ground
Bedroom	number of bedrooms
Kitchen	number of kitchens
TotRmsAbvGrd	number of rooms above ground
Fireplaces	number of fireplaces
GarageCars	number of garage spaces for cars
GarageArea	square feet of garage
WoodDeckSF	square feet of wood deck
PoolArea	square feet of pool area

Assignment

What's due:

PowerPoint presentation due before class on Monday, November 13, 2023. The expected length of the presentation is 15-20 minutes, approximately 10-20 slides. Please send me the slides at least one hour before class. You can describe the slides from your seat.

Outline

Using the house price dataset, perform an analysis of the following aspects of the data.

1. Visualization
    a. Develop an overall view of relationship of the continuous dependent variable (SalePrice) with all continuous X-variables
    b. Highlight at least two graphs where there are strong relationships between the X-variable and SalePrice
2. Variance Inflation Factor (VIF)
    a. Perform a VIF analysis
    b. Identify variables with VIF less than 10
3. Neural networks (one hidden layer) – continuous output (note: normalize the data before running the NN)
    a. Develop neural networks (one hidden layer with 1 to 3 nodes) with continuous output price using 70% training data, randomly selected, and variables with VIF less than 10
    b. Test the neural networks with the remaining 30% testing data
    c. Compare the accuracy of all neural networks
4. Neural networks (two hidden layers) – continuous output (note: normalize the data before running the NN)
    a. Develop neural networks (two hidden layers with 1 to 3 nodes each) with continuous output price using 70% training data, randomly selected, and variables with VIF less than 10
    b. Test the neural networks with the remaining 30% testing data
    c. Compare the accuracy of all neural networks
5. Identify a list of lessons learned
    a. When do visualizations help? When do they not help?
    b. Which neural network was the most accurate?