Alexis Mekueko
Data622, Final ProJet
05/24/2022

Predicting a Sale Price of a House

Using Multiple Linear Regression and Neutral Network to Predict the sale price of a house in Ames city

Let's say someone is interested in buying a house and don't know how to prepare for it. Imagine this person who is looking for a dream home and don't even know where to start. Buying a house start with knowing the sale price. If this person who is wondering about buying house and has no clue, then there is a problem. Buying a house is a lifetime experience for many first-time home buyers. It is an exciting moment, but it can be a bad one if things turn out unexpectedly. Thus, the question is: how can a home buyer know about the sale price of a house? Meaning, how to predict a sale price of a house? Well! Machine learning has the answer to this question. Machine learning algorithm is a computer algorithm that teaches itself a pattern based on data. In other words, machine learning builds a pattern that helps identify the future event occurring.

When we said machine learning is a computer algorithm to build a pattern for a sale price of a house that is not fully correct. It is not correct because there are many algorithms, and the pattern relies heavily on the data. If there is no data, then there is no price prediction. Therefore, data plays an important role in predicting the price of a something with the machine learning algorithm. The dataset used to predict the sale price of a house is from the Ames housing division in Iowa. The dataset is available at Kaggle.com. The data has about 1460 observations and 81 features with sale price (the target variable) included. These features describe just about everything that is part of a house. We are talking about the lot size, the year the house was built, the garage condition, fireplace conditions, etc. Although there are enough variables in this dataset, we still need to clean the data before fitting it to the model for analysis. The cleaning involved removing all unnecessary variables, the variables that have equivalent meaning with other variables or variables that have way more missing values by visual inspection. Finally, we used correlation technique to find which variable has strong influence on the target variable.

The purpose of cleaning the data is to ensure that we have the right data for the method used for the analysis. Let's recall that data analysis is the process of collecting, modeling, and analyzing data to extract meaningful information that ultimately leads to decision-making. In other words, the leadership want to make decisions based on solid facts. Building these solid facts is where the methodology comes to play. There are several methods and techniques use in analyzing the data. The choice of the method and technique will depend on the aspect of the data (quantitative or qualitative). This Ames Housing data has quantitative aspect because it has finite number (discrete data) of values resulted from measurements collected on each house individually. For example, the number of bedrooms in the house, the number of car garage attached to the house and the size of the lot. Although some values are qualitative such as the condition of the garage (excellent, good, fair and none), it still refers to quantitative data. These data are taken as quantitative because this is not a survey where the responses are not uniformed. It is measurable data with standard in place. This type of measurements requires expertise that would apply the standards and unformalized for all houses. Cleaning the data is a step that precedes the analysis. We used multiple linear regression and Neutral Network method to analyze the data. The technique used is basically fitting the data to the model. The model here refers to the computerized formulas for each machine learning algorithm.

The purpose of performing the analysis is to predict the sale price of a house. Initially, we had a question about how one can predict the sale price of a house. Well! This is where the magic happens. The accuracy of predicting this price depends on the formulas. The choice of multiple linear regression is because while preparing the data for the analysis, we plotted few predictors against the target variable and saw some linear relationship. The second choice of neutral network is because we wanted to see how an unsupervised machine learning will perform against a supervised one. At the end, the verdict was tight because each algorithm performed well. Since the neutral network seems to be popular in neuron studies (complicated topic), we thought it would have easily outperformed the multiple linear regression. The assumption was not valid because this regression analysis showed better results with a low mean squared error value compared with the one from neutral network. In addition, based on the output values such as p-value, the chosen explanatory variables explained well the sale price of a house. Furthermore, we like the regression performance better because we can apply it into a business. Knowing a set of features along with the machine learning algorithm that can directly influence the sale price of a house, we can build an application that can tell customers the sale price of a house. We think this can be a starting point for many home buyers. There are other factors such as speculation (demand and supply) and inflation that influence the sale price of o house but we believe the home buyers want to know this price first.