House Price Prediction

Satish Agrawal
DSC 680
https://satishagrawal.github.io/

Business Domain Overview (Which Domain?)

Real state is one of the biggest industries and have millions of transactions every year. Real state includes a commercial complex, a multi-family building, a residential house, a plot etc. [James Gheen]. This project aims to look at only the price of a house. House price fluctuates a lot and there are variety of factors that impact the final price of a house [National Association of Realtors]. The size of the house/lot, amenities inside the house, upgrades, age and quality of materials used are few of the factors impact the price of a house. Other than what's included in the house, there are multiple external factors that impact the price of the house heavily, like market condition, job market, season and the time of the year to list a few.

There are various listing agencies helping list houses in the market and let user choose a property from a website. These companies have estimated price (for example zestimate from Zillow [3 Zillow]) and seller also chooses a listing price which may or may not be the same as estimated price. There is need of a mechanism to come up with a great estimated price given all (or most) of the important factors. This solution will provide a consistent way for seller, buyer and listing companies to see a fair price for the house and make their own decision.

In this project I will build a machine learning model that takes important factors for consideration and trains a model which can accurately predict the price of a house.

Data Understanding (Which Data?)

For this project, I have selected a dataset from Kaggle [Kaggle data set]. This data set has 81 different attributes about houses sold recently, which includes the sale price. As I am working to train a model that can predict the sale price, that is the target variable for this project. The sales price is a continuous number in US dollars.

All other 80 variables are explanatory variables that help determine the sales price. Below is the list of main variables in the data set:

1. Basement information: There are around nine attributes defining the condition and the quality of the basement in the house. The condition of the basement is an important factor which impacts the final price and the valuation of the house. Few attributes are: is basement finished? What is the square footage of finished area? Is it a walkout basement? Etc.
2. Garage information: There are at least seven attributes about the condition of garage(s) in the house. Garage also play a vital role in the house price. For example, house with three car garages is going to be more expensive than two or one car garage.

3. Lot information: There are a few attributes that have the information around the lot for example, lot size, lot shape, lot frontage etc. Lot details is one of the many important attributes that determine the house price.
4. Year built: This is when the house construction was finished. Older house may have older appliances and may have other issues. Year build is a key to help determine the house price.
5. There are quite a few important attributes that impact the final price of the house for example, number of bathrooms (full bath versus half bathrooms), size of the living room, heating mechanism, cooling mechanism, electrical, number of fireplaces, number of fans in the house etc. This data set has many of these attributes included and will be useful in determining the final price.

The data set at hand has a lot of variables that I assume are going to be strongly correlated with the final house price. Initial analysis has revealed that only a few columns has missing data and quality of data is acceptable.

Problem Statement and Hypothesis (Research Questions? Benefits? Why analyze these data?)

The sale price of a house is of interest for various parties including:

- Buyer
- Seller
- Marketplace like Zillow, Redfin [3], [4]
- County/City
- Title companies
- Insurance companies [Mark Chalon Smith]

A good model to predict the price of a house is going to be helpful to all the above and many other entities

Approach (What Method?)

The target variable for this project is the sale price for a house. The price for a house is a whole number represented in dollars. It is a continuous variable which may range from a few thousands to millions of dollars. I plan to train regression models for this.

I will spend some time looking at data and analyzing variables which are strongly correlated to the target variable. I may choose to drop a few unrelated variables and also derive new variables from already present attributes, for example, from year built and year sold, I can easily calculate the age of the house at the time of sale. I will look for data quality like any garbage values in some of the fields or even missing values. I will also check for any outliers which can be data recording issues. I will also scale and encode the data as needed.

Once I have a cleanedup dataset, I will train a few regression models on the train set and implement model selection and evaluation techniques while tuning the hyper parameters. Few of the algorithm that I

will certainly try are GradientBoostingRegressor, XGBRegressor, RandomForestRegressor, Ridge and Lasso.

Finally, when I have identified the best performing model with best hyper parameters, I will train a final model with all of the data at hand and get it ready for production deployment.

Potential Issues?

Some of the attributes have excessive missing values. I am concerned if those are highly correlated to sales price, it may not be acceptable to simply drop those columns. On the other hand, it is also not going to be easy to impute those missing values. I normally do not like attributes with more than 50% of the missing values. I will explore efficient ways to impute those attributes.

The price of a house varies a lot through out the year. The attributes in the dataset are the element describing the condition of the house. It has no information of the market situation which is a very important factor. Currently, it is a sellers-market and houses are being sold way above asking price. This model as planned at this point may not be as effective in the marker like this. However, it is still going to set the benchmark for all the parties interested.

Concluding Remarks

House is one of the common needs for human being. As population grew, so did the need for housing and amount of transaction every year [Norada]. Based on the business need, financial status, job change and personal preferences, people move from one place to another and in turn buy and sell houses. It is very important to have an estimated price for a house for all the parties involved to look at as starting point to negotiate up or down based on market condition and individual needs. Its promising to see what machine learning algorithm can do with data. The sales records we have are representation of what has happened already in housing market and each record represents one transaction. This project will use that historical knowledge and train an effective model to predict a sale price for a house. The final sale price may be a result of negotiations between all parties involved but a guideline sale price is going to be really helpful and game changer if we can achieve great accuracy.

References:

[1] National Association of Realtors - https://www.nar.realtor/research-and-statistics/housing-statistics/existing-home-sales

[2] Kaggle data set - https://www.kaggle.com/bakar31/eda-house-price-prediction?select=train.csv

[3] Zillow https://www.zillow.com/homedetails/22803-Hansen-Ave-Elkhorn-NE-68022/91935264_zpid/

[4] Redfin - https://www.redfin.com/NE/Omaha/16823-Pasadena-Ct-68130/home/63901488

[5] Mark Chalon Smith - https://www.insurance.com/home-and-renters-insurance/home-insurance-basics/5-factors-that-affect-rates.html

[6] Norada - https://www.noradarealestate.com/blog/housing-market-predictions/

[7] James Gheen - https://www.biggerinvesting.com/6-types-of-real-property-infographic-real-estate-investing/