Image Processing Approach for House Price Estimation

Allen Ansari

Author Note

I would like to express my very great appreciation to Blake Arensdorf for his valuable and constructive suggestions during the planning and development of this research work.

Abstract

Most of approaches to predict house pricing were based on features like total house sq, number of rooms, bathroom, lot sq etc. Real estate companies provide such estimates using this formula which sometimes close to the sale prices and in some cases they are inaccurate. In this research, I am investigating interior and exterior appearance impact to estimate house price by extracting house images features using CNN and combined with traditional estimating model. I developed three different models to compare affecting of image to house price prediction and was able to improve median absolute percentage error by 1%.

Keywords: computer vision, house price, real state, convolutional neural networks

## Introduction

## When you are searching for a house the first things that catch your eyes are the house's images posted along with other information. Predicting house prices has always been one of the interesting projects for most data scientists. There are many features that are found to be correlated with house price and lots of approaches and models to address this problem. Real estate websites like Zillow are using house general features like total house sq, total lot sq, number of bedrooms/bathrooms, kitchen sq, school zone, street address, state, zip code and others which are entered by owners/agents, to predict the price. The Zestimate is Zillow’s estimated market value for houses. The model trained on, about 110 million homes information in the United States by using factors like physical attributes, tax assessments and prior transactions. The Zestimate has a median error rate of 7.9%[[1]](#footnote-1). But interior/exterior houses appearance is not considering in this model. However, the house vision features are the keys to this market values. In this research I used both categories of features to estimate house price by comparing 3 different models.

## Dataset Description

## The collected data included 50 features such as Listing Price, Address, City, State, Zip Code, County, Subdivision, Legal Description, Property Type, Bedrooms, Baths, Garage(s), Stories, Style, Year Built, Building Sqft., Lot Size, Maintenance Fee, Living, Family Room, etc., for 6348 houses listed in [www.HAR.com](http://www.har.com) within Houston, TX zip codes. This dataset contains more than 204k images of those houses posted by owners/agents. Data wrangling was most time consuming part of this project since there were lots of missing value and some information entered in different formats for example rooms dimensions had different type of formats like:

## ['1st', '1st']

## ['18x15, 1st', '5.49 x 4.57(m)']

['20\'5"x14\'9\', 1st', '20\'5",14\'9\', 1st']

## and I calculate area for each room by using those dimensions.

Table 1 shows some statistics about dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature Name | min | max | mean | std |
| Listing Price | $60000 | $14500000 | $516073 | 636683 |
| Build Sqft | 754 | 21032 | 2838.53 | 1332.93 |
| Tax Rate | 0.61 | 3.9 | 2.63 | 0.3 |
| Age | 1 | 145 | 34 | 24 |
| Lot Size | 1104 | 1306800 | 11350 | 28261 |
| No. of Bed | 1 | 10 | 3.7 | 0.81 |

Table 1 – Dataset Statistics Summary

A picture containing sky, outdoor, building, tree

Description automatically generatedA bathroom with a large bathtub

Description automatically generated with low confidenceA kitchen with wooden cabinets

Description automatically generated with medium confidenceA picture containing indoor, ceiling, room, sofa

Description automatically generated

Figure 1. Sample house from (www.HAR.com, 2020), where itis represented by 4 images for the exterior side, the bathroom, kitchen and the living room

## Proposed baseline

The main aim of this research is to test the impact of including visual features of houses to be used for the house’s prices estimation. Since the images downloaded from HAR website were not labeled I had to use image classification model to classify them based on room type. I used a trained model[[2]](#footnote-2) to categorize pictures based on the categories shown in Table 2.

|  |  |
| --- | --- |
| Room type | quantity |
| Exterior | 35000 |
| Bathroom | 45497 |
| Bedroom | 29738 |
| Kitchen | 1 |
| Dining room | 1104 |
| Interior | 9000 |
| Living room | 1 |

Table 2 – Number of images per room category

This had about 90% accuracy and for sure there are some images classified in wrong category.

## Price estimation model

My first approach was a neural network model using house textual features to estimate the house price. I used MinMaxScaler() for numeric features and BinaryEncoder() for categorical variables to normalized the data. To train the network, I use the listing price of recently posted houses for sale which is the price asked by owner for that house. So, we need to consider that this price is not sold price and it may add more bias in the model. Considering this the median absolute percentage error I got for this model was 9.95% and median absolute error was $ 34194.34 which means 50% of the price predicted are closer than the error percentage and other 50% are farther off. In other word half of estimated price has close to $34194.4 off from listed price.

For the second model I randomly selected one exterior, one bathroom, one kitchen and one living room image form image classifier result for each house and combined them as one image as shown in figure 2. After applying several filters, the best result, I got was 33.71% for median absolute percentage and $109041.11 for median absolute error. Please note that we may need to consider the error from image classifier which categorized about 10% of images in wrong category.

The next model is to combine these 2 models together as shown in figure 3. For this model the result improved to 8.10% for the median absolute percentage and $ 27292 for median absolute error. Result for each model shown in table 3.



Figure 2. Sample house image input structure to CNN model

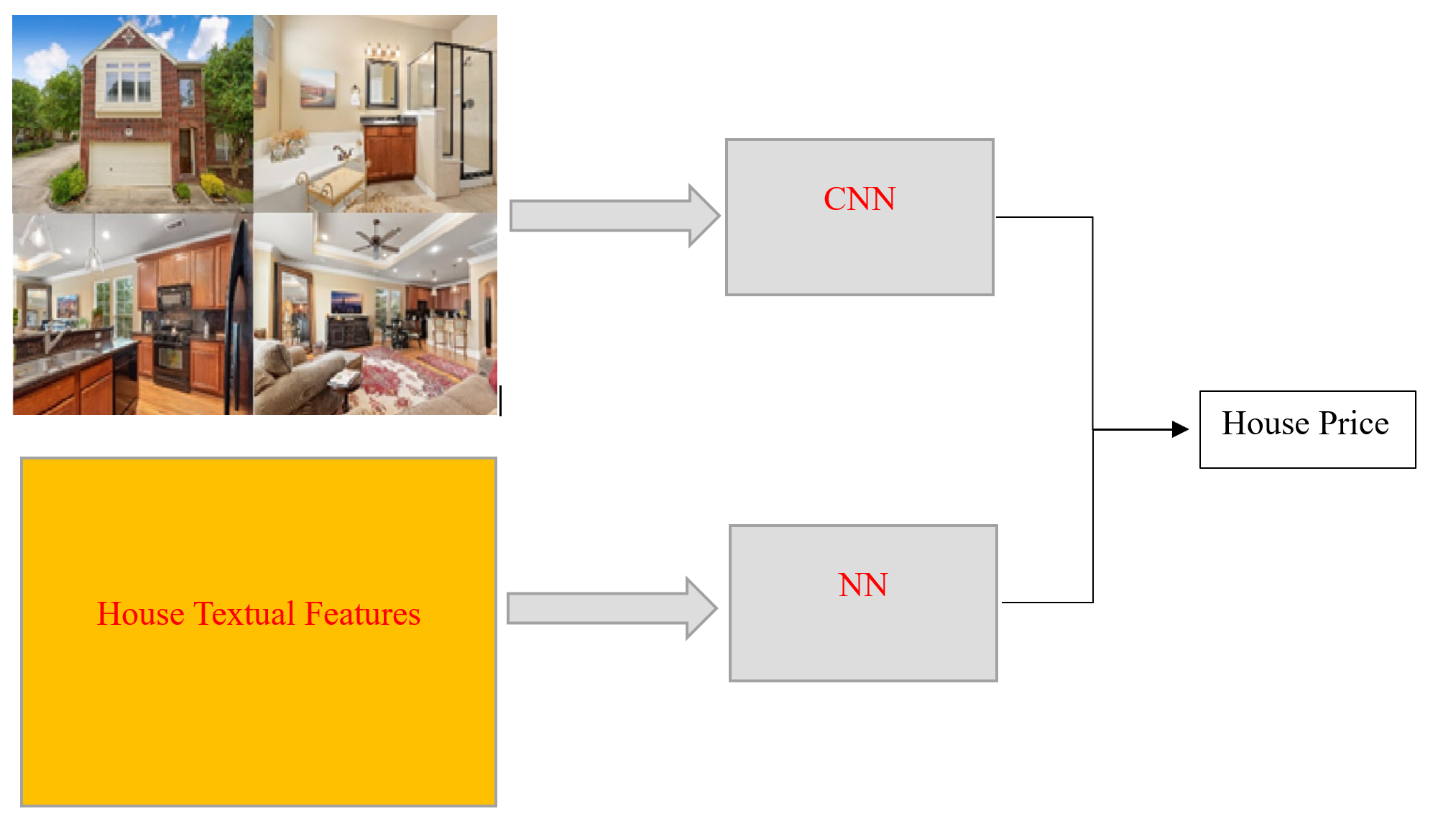


Figure 3. Ensemble Model Structure

|  |  |  |
| --- | --- | --- |
| Model Name | Median % error | Median absolute error $ |
| MLP | 9.95 | 34194.34 |
| CNN | 33.71 | 109041.11 |
| MLP + CNN | 8.1 | 27292 |

Table 3 – Median error rate of different methods

## Conclusion

I have presented a novel algorithm to consider the impact of appearance on the value of residential properties. After collecting large datasets of real estate photos and metadata for houses in Houston area and using ensemble model median percentage error improved by almost 2% (almost $7000) which is a considerable improvement to estimate house price. We need to note that this improvement happened even with 10% error in image classifier model and I expected to get better improvement if all images categorized correctly and run the model on more larger dataset.

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

1. https://www.zillow.com/zestimate/#acc [↑](#footnote-ref-1)
2. https://towardsdatascience.com/image-classifier-house-room-type-classification-using-monk-library-d633795a42ef [↑](#footnote-ref-2)