Average House Price Prediction in the U.S.

Qijun Ma
Crean lutheran High School,CA 92618, USA.

Abstract: This would be a research focusing on the growth of house price as a trend in long terms and most importantly look at the model of which the house price would grow in, whether its linear model, exponential model, or any other possible models. The prediction would be done using the methods of regression through machine learning and the algorithm of gradient descent.

Keywords: Machine Learning; House Price Prediction; Linear Regression; Polynomial Regression

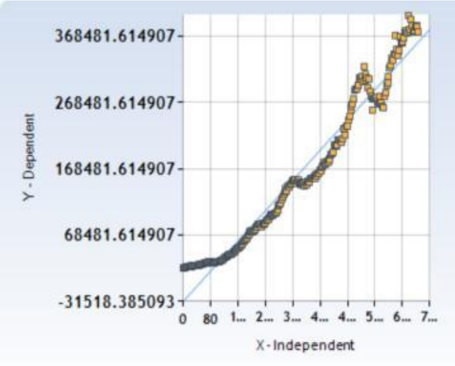
Introduction

Machine learning is a tool that can be used to analyze big data and make predictions. And its predictions … can be about all kinds of things – customer churn likelihood, possible fraudulent activity, and more[1]. One of these things people use machine learning to predict is house prices. Businesses such as zillow make predictions on individual house prices and they have machine learning models that analyze hundreds of data points on each property[2]. Zillow forecasts 7.8% home value growth over the next 12 months (July 2022-June 2023) using their multi-factor prediction method[3]. There are a lot of other house price predictions similar to this that are being made by economists and investors[4]. This study would not be focusing on these areas because there is already a lot of progress made by big companies. Instead, this research is going to look at the big historical trend of the average home price growth/decay in the U.S. over the last sixty years and try to find a model that can best understand this trend. This research would test out a lot of different types of relationships between time and the growth of home prices to see which model best represents the actual world. This research is aimed to serve as a reference to people who are uncertain when to buy houses in the long run, let's say five to ten years by extending the datasets deep into history. It definitely would not serve as an accurate model to predict the price of houses in a one or two months future. This research is also aimed to evaluate the healthiness of the house price growth and see if its longer impact into the economy and society.

Methods

To begin with, my data comes from the archival economic data website which provides average sales price of houses in the U.S. every quarter, from 1963 to June of 20225 . There are a total of around 200 data points. Throughout different models, I would manipulate the data to fit different types of models, and I would use gradient descent to calculate the best fit line throughout different models. Because I want to look at the sequence, my independent variable would be time and my dependent variable would be price. To measure how good the model is, I look at both the cost and the average deviation of the last 5 data points from the line that I did not train my model with.

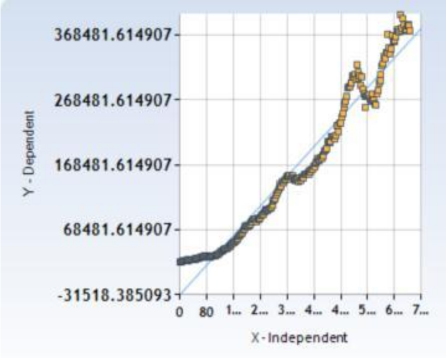
The first method that I used was linear regression, this most likely wouldn't be the best model, but I still tested it out. Using gradient descent, I found that 1 always converges to approximately 500 no matter what the starting point is. But 0 changes dramatically depending on the starting point. So I tested out different 0 values while putting 1 to 500. The smallest cost from my testing is 1169530448.3685331, in which the line has a y-intercept of- 19955 and a slope of 544. The average prediction error of thlast five points is - 124154. I graphed all the data points out to take a closer look. And its clearly seen that the line is not a best fit as the housing data does not seem linear at all.



The second method I choose is taking logarithms on both x - values and y -values because if the graph is a standard polynomial regression, it would be:

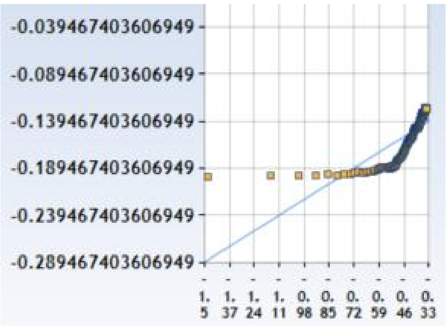
$$y= a^{x} ,\log(y)=\log(x)=n \log(ax)=>\log(y)=n \log(ax)$$

Although we cannot know the direct relationship between y and x from this equation, we can know the main exponential power of the polynomial regression. After trying different values in gradient descent, the smallest cost of the new logarithm function I found is around 0.06 in which 0 is around 2.90 and 1 is around 0.90. This looks pretty good at first and if this is the correct model, the exponential power in the original equation would be 0.90. However, the graph of this shows there needs to be further investigation. This set of data points clearly does not show a linear model, it rather looks like an exponentially growing data set. To further understand what the exponent is in this model. I would need to take logarithms on both the axis-again.



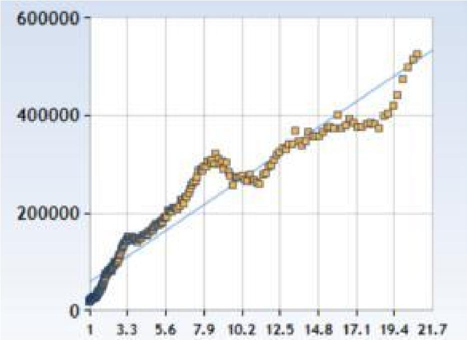
The equation that symbolize this graph most likely looks something like this:

log(y)= (n log(ax). And it should be linearized if I did another logarithm on it. However, the result shows otherwise. After another round of linearization, it still looks like an exponentially growing curve. This means that the original curve definitely isn’ t a simple polynomial. I did one last trail of taking logarithms on both sides to ensure that this trying direction would not work. As shown in the figure to the left, there would not be a further linearization using logarithm because all the values on this graph are already negative, and you cannot take the logarithm of a negative number, so another direction on this problem is needed. From what I see in the graphs, I believe the correct model to represent this graph is probably something like y versus x to the power of x because if the polynomial is any constant, the graph would be linearized in the first try. Only in the case of xx, the equation being logged would not be constant like this:



$y= a^x$,log(y)=log(x)=x log(ax)=log(y)=x log(ax)

log(y) verse x is definitely an exponential equation. And if we continue to do logarithm on log(y) vs. x,it would just become log(log(y)) vs log(x) and log(log(log(y))) vs log(log(x)) which would be exponential no matter how many times we logged. So what I did is I did a yny vs x graph. Where n is a constant.I didn’t find any methods to directly find the value of n, so I tried multiple n values and the reasonable range of n values is from 0.001 to 0.0001. The value that worked out the best for the prediction is0.00065, and the graph looks like this figure on the left. When the value is 0.00065. The prediction error is around 50,000. And this great reduction in the prediction error means this is a better model.



Conclusion

When I begin this research I have expected that the model of average house price would either be linear or polynomial. However, in this research, I found out that the actual model is something similar to xnx or x!. This means house price is growing at a rate that is actually not too fast right now because currently the nx is still a small number due to the small n value. However, it can be foreseen with this model that in the far future, house prices can grow faster than exponentially.

References

[1] What does Prediction mean in Machine Learning? Available from:https://www. datarobot. com/wiki/prediction/#:~:text=Machine%20learning%20model%20predictions%20allow,possible%20fraudulent%20activity%2C%20an d%20more.

[2] Andrew Martin, Bin, Cat N, K Nielsen, Maggie, Wendy Kan,(2017) Zillow Prize: Zillow’s Home Value Prediction (Zestimate), Kaggle.

[3] Zillow Home Value and Sales Forecast: July 2022, available from: https://www. zillow.com/research/home-value-sales-forecast-july-2022-31240/.

[4] ALFRED Graph, available from: https://alfred.stlouisfed.org/graph/?g= SNMG.

About the author: Qijun Ma (2005.07), male, Han Nationality, Shanghai native, student, high school degree, Crean lutheran High School, research direction: Economics,Machine Learning.