Boston House Price Predictions

Course: Math Concepts for Deep Learning

Group Members: Javier Jimenez,
Luis Alberto Sosa Quintero,
Ting Fu Kevin Tsai,
Olanrewaju Yusuf

Date: February 7, 2022

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

1.0 Introduction

The real estate market is competitive and the market value can depend on various factors. The ability to accurately predict is valuable to anyone entering the real estate market.

1.1 Problem Statement

This project investigates the methodological approach of using neural networks to predict house prices in Boston, Massachusetts in the 1970s. During that period, the median home price rose from $23,000 to $55,700, an average annual gain of 9.9%.

1.2 Data

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston, Massachusetts. The dataset is small in size with only 506 cases and 14 variables. Originally, the data was collected in order to solve problems associated with the willingness to pay for clean air.

The independent variables include two structural attribute variables (S), eight neighbourhood variables (N), two accessibility variables (A), and one air pollution variable (P) as in Table 1.

Table 1. Variable description

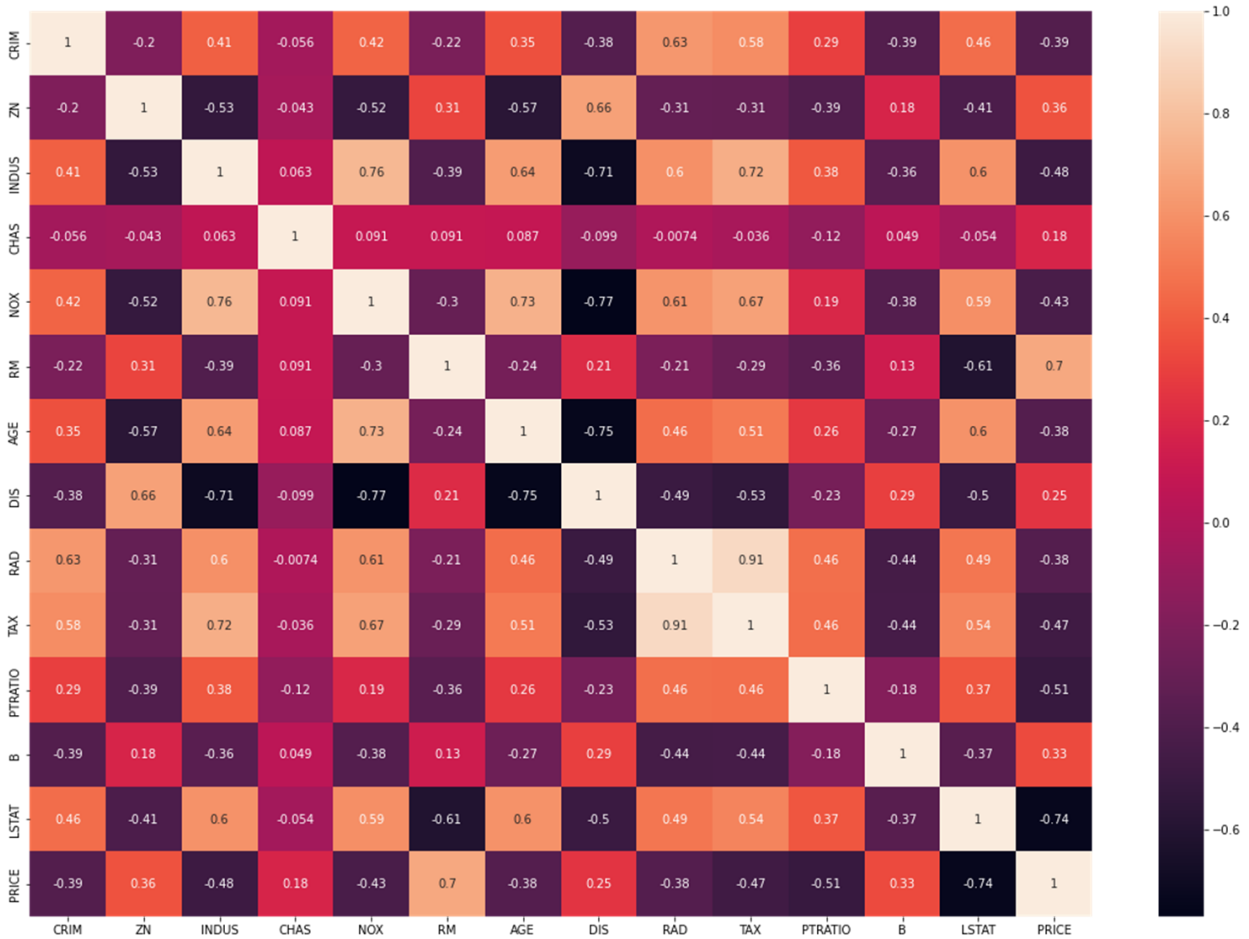
Variables	Example	Type
CRIM - per capita crime rate by town	0.00632	N
ZN - proportion of residential land zoned for lots over 25,000 sq.ft.	18.0	N
INDUS - proportion of non-retail business acres per town.	2.31	N
CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)	0	N
NOX - nitric oxides concentration (parts per 10 million)	0.538	P
RM - average number of rooms per dwelling	6.575	S
AGE - proportion of owner-occupied units built prior to 1940	65.2	S
DIS - weighted distances to five Boston employment centres	4.0900	A
RAD - index of accessibility to radial highways	1	A
TAX - full-value property-tax rate per $10,000	296.0	N

PTRATIO - pupil-teacher ratio by town	15.3	N
B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town	396.90	N
LSTAT - % lower status of the population	4.98	N
MEDV - Median value of owner-occupied homes is $1000's	24.0*	target

*24.0 corresponding to a median price of $24,000 – ($24,000 in 1970 is equivalent to $172,454 in 2022 adjusting for inflation)

Using a correlation matrix, RM (number of rooms) and LSTAT (percentage of lower status population) can be identified as highly correlated variables to the price of a particular house. Number of rooms (RM) has a positive correlation, whereas percentage of lower status population (LSTAT) has a negative correlation.

Figure 1. Correlation matrix of housing variables



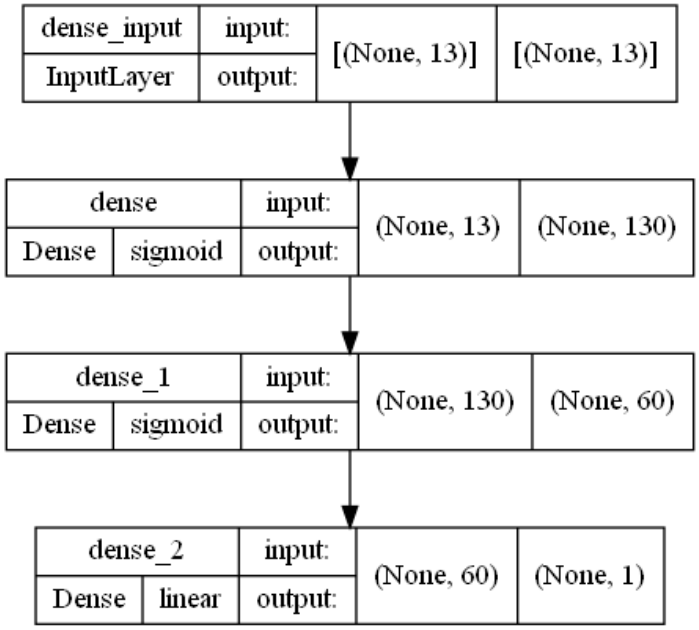
2.0 Sequential Models

As baseline models, data were scaled using a standard scalar before being split into train and test subsets and used to train a neural network. All models use linear activation for the output layer and mean square error as the loss function.

2.1 Sigmoid Activation

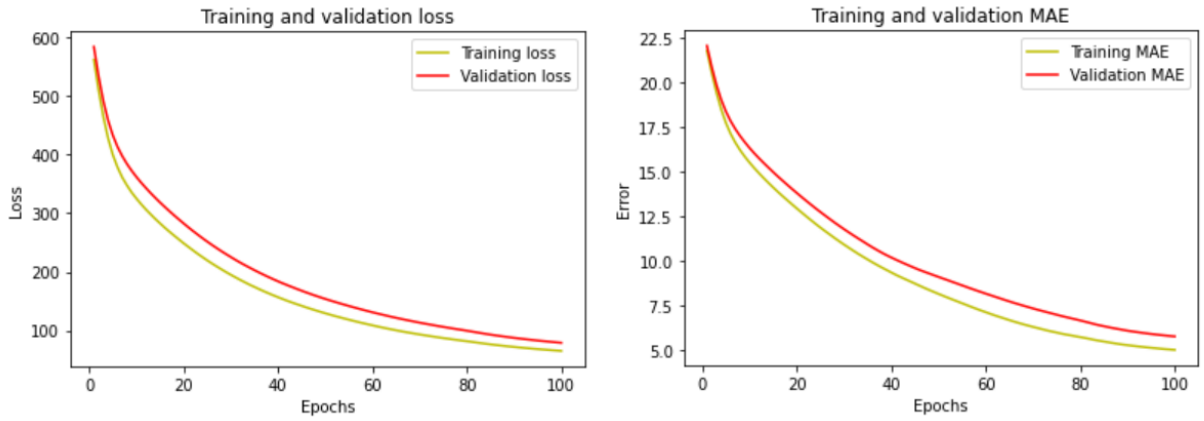
The model consists of 3 dense layers with 130, 60, and 1 neuron respectively, using sigmoid activation for the first and second layers.

Figure 2. Sigmoid activation model



The loss was 48.2741 and the mean absolute error was 4.2625.

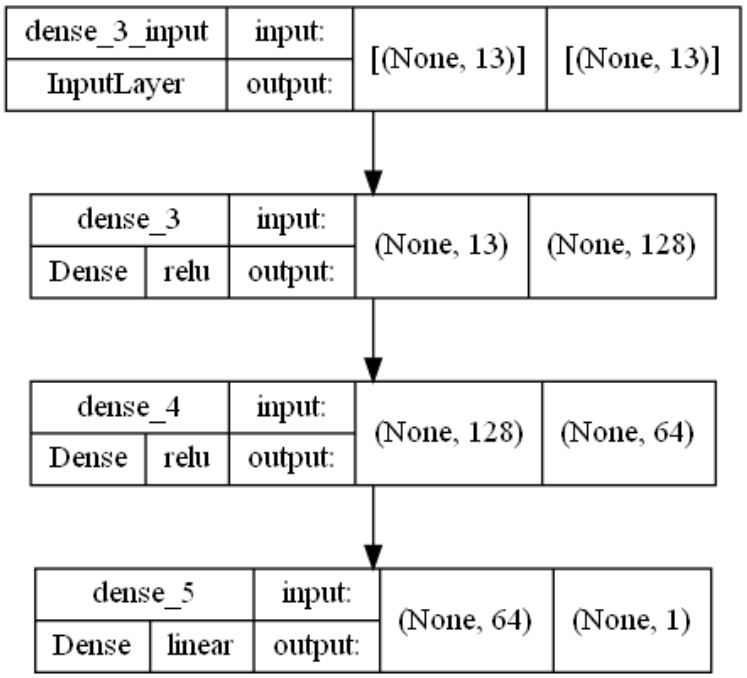
Figure 3. Loss and mean absolute error of sigmoid model during training



2.2 Relu Activation

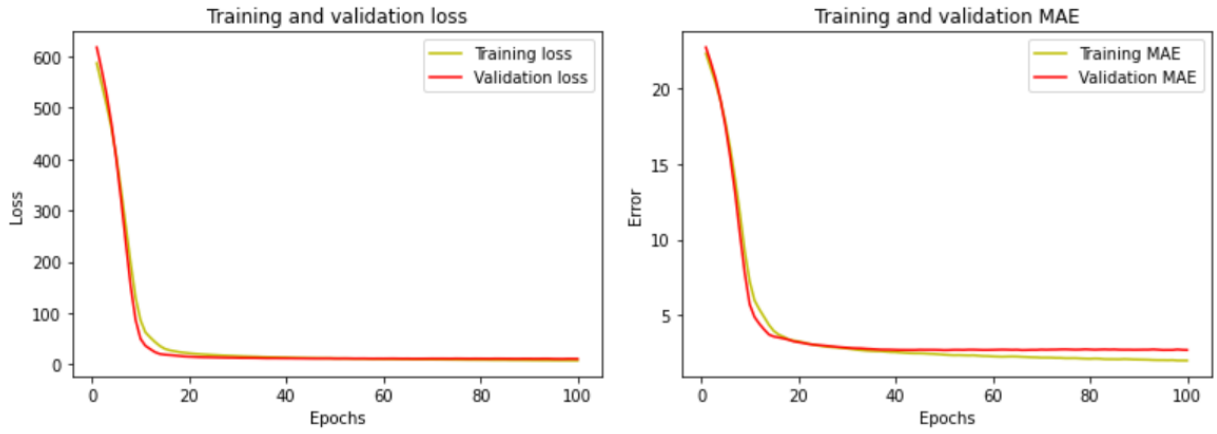
As a second baseline model, relu activation was used instead of the previously used sigmoid. The model consists of 3 dense layers with 128, 64, and 1 neuron respectively.

Figure 4. Relu activation model



The loss was 14.7645 and the mean absolute error was 2.8664.

Figure 5. Loss and mean absolute error of relu model during training

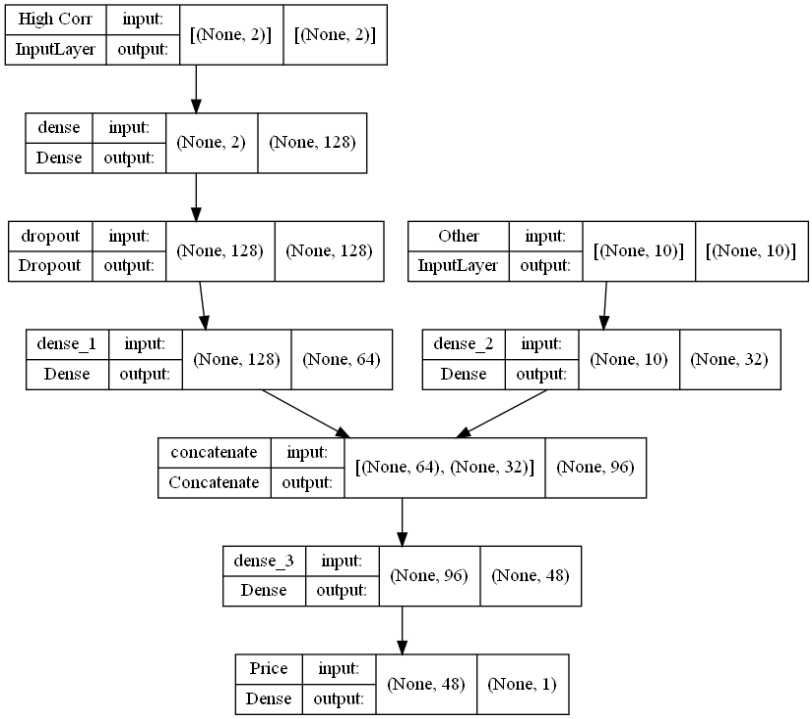


The relu model has a better performance compared to the sigmoid model. It was also much faster to train as seen in both the loss and mean absolute error graphs where the curve starts very steep and begins to flatten out after 20 epochs. In the graphs of the sigmoid model, the curves only start to level around 80 to 100 epochs.

3.0 Multi-input Model

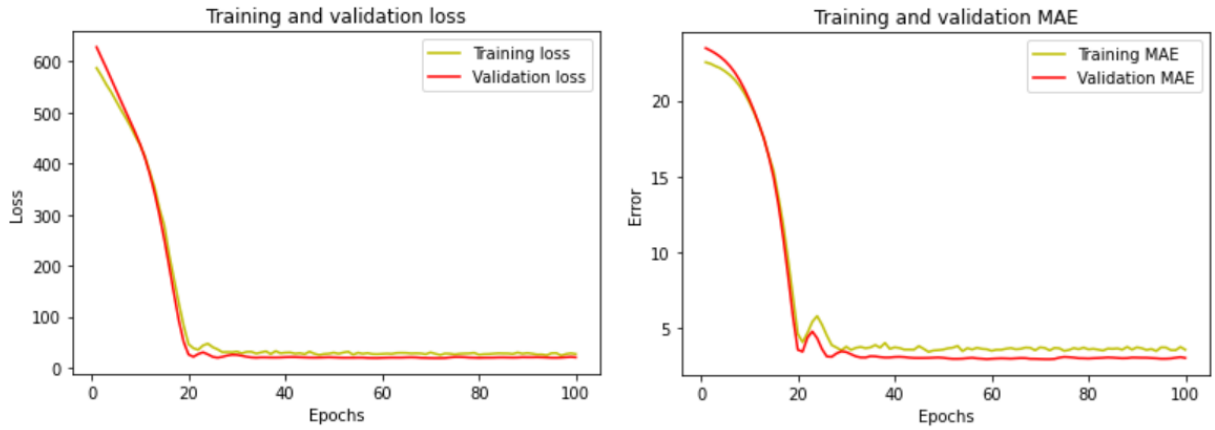
Using the findings from the correlation matrix, 2 variables with high correlation to the price are isolated from other variables. A multi-input model was built to take high correlation variables as one set of input and the remaining variables as another set. They are passed through different hidden dense layers and concatenated before producing a single price output.

Figure 6. Multi-input model



This model had similar results to the relu baseline model, with 16.8569 loss and 3.0071 mean absolute error.

Figure 7. Loss and mean absolute error of multi-input model during training



4.0 Classical Machine Learning

Several regression models are built and compared to neural network models and the results are shown in Table 2.

Table 2. Model Comparison

Model	Mean Absolute Error	Parameters
Sigmoid NN	4.2625	
Relu NN	2.8664	
Multi-input NN	3.0071	
Linear Regression	3.0559	LinearRegression()
Decision Tree	3.3971	DecisionTreeRegressor(criterion="squared_error")
Random Forest	2.4082	RandomForestRegressor(n_estimators=30, random_state=30)

Without the need of hyperparameter tuning, all 3 regression models have similar performance as the neural network models. Random forest regressor even outperforms all other models with the lowest mean absolute error of 2.4802.

Lastly, we have the predictions based on the Sigmoid model as shown in Table 2.

The differences between real and predicted prices are minimum. The sigmoid activation function works well with this sample of houses. The dollar had an average inflation rate of 3.87% per year between 1970 and today, producing a cumulative price increase of 618.56%.

Observing the data we see that the price difference between the real and the predicted price is always positive, which indicates that the model tends to predict a little below the real price.

Table 3. Model Predictions

Real($)	Predicted ($) in 1970	Predicted ($) in 2022
21.2	18.51	152,335
20.6	20.62	148,023
21.5	20.94	154,490
21.7	21.00	155,927
13.14	12.63	96,287

5.0 Conclusion

Though trying to achieve better performance by using insights from the correlation matrix and a more complex model, the multi-input model failed to do so, by being similar to, if not worse than the relu model. Using relu or linear activation is more suited for price prediction, as sigmoid is more commonly used for probability predictions.

Neural networks do not allow us to see the ranking feature, thus making it more complex to interpret than classical machine learning models. In some cases, it may be better to use models such as random forest over neural networks, as seen by the better result achieved in this report.

In order to improve the predictions of the models, a larger dataset can be used, as 506 samples is a small dataset for deep learning. Though mean squared error is the default loss function for regression problems, the effect of other loss functions such as mean squared logarithmic error, which penalize the model less for larger mistakes, can also be investigated.

6.0 References

Harrison, David & Rubinfeld, Daniel. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management. https://www.researchgate.net/profile/Daniel-Rubinfeld/publication/4974606_Hedonic_housing_prices_and_the_demand_for_clean_air/links/5c38ce85458515a4c71e3a64/Hedonic-housing-prices-and-the-demand-for-clean-air.pdf

Omrshahid. “Deep Learning Regression.” *Kaggle*, Kaggle, 1 Jan. 2022, https://www.kaggle.com/omrshahid/deep-learning-regression

“The Boston Housing Dataset.” Boston Dataset, https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html