House Price Estimation with CNN

Ngudup Tsering
1104377
MSc Computer Science
Lakehead University, Thunder Bay

***Abstract*—Most automatic house price estimation is based on the multilayer Neural Network or with traditional machine learning techniques such and Multi Linear Regressor, Support Vector Regressor, Random Forest etc. This paper proposes a 1Dimensional Convolutional Neural Network using California Housing dataset for house price prediction.**

***Index Terms*—CNN, MLR, SVR, NN, convolution, pooling, L1 Loss, R^{2}score**

I. Introduction

Real estate market plays an important part in defining the growth of the market. Housing renovation and construction stimulates the market by increasing the rate of house sales, employment and spendings [4]. It also impacts demand from many other relevant industries such as building materials, household consumer goods, etc. In addition, the housing sector serves as a critical indicator of the real estate sector of the economy and of the asset prices that help to forecast inflation and supply. Existing price prediction relies on excruciating visits to agents and long searches and comparisons that is time-consuming and tiring, and often futile as it lacks a standard. Therefore, a reliable, scientific, impartial and automatic price predictor plays an important role in helping authorities to design policies to control inflation. Furthermore, it also helps individuals to make informed investments. Predicting house price is a complex mechanism as it depends on many factors and different aspects of the economy. According to [1], the house price gets affected by some factors like its neighbourhood, age of the house, and the number of bedrooms. The more bedrooms and bathrooms the house has, and higher the price. Therefore, I have relied on these factors to estimate the price, along with median-income that was empirically found directly correlated with the house-value.

II. Literature Review

With the explosion of technology on account of the exponential increase in power of computation and memory, many tedious and difficult tasks are being attempted to be resolved with the help of Neural Network. As the standard of living witnessed tremendous growth in the last decade, the potential for the investment in buying assets has also increased with renewed interest in real estate sectors. There is some work that has been done to automate the real estate price evaluation process which can be broadly classified into Data Disaggregation based models and Data Aggregation based models. The Data Disaggregation based models predict the house's price with respect to each attribute alone like the Hedonic Price Theory. The approach taken in this paper belongs to Data aggregation model, where the model depends on all the attributes to estimate the price. Flitcher et al in (Fletcher et al., 2000) explored the best way to estimate the property price comparing the results of aggregation and disaggregation of data. They found that the results of the aggregation are more accurate. Furthermore, they also discovered that the property's geographical location plays an important role in determining the price.

Nevertheless, the Neural Network Model does report some drawbacks, as the estimated price is not the actual one. One of the reason can be attributed to the difficulty in obtaining the actual data from the market. Moreover, the property price is affected by many other economic factors that are hard to include during the estimation process. The author of [4] compared the performance of Multiple Linear Regression (MLR) model and Neural Network model to estimate house prices and concluded that using NN increases R2 score by about 26.47 [3] explored similar research by comparing the Support Vector Regressor (SVR), and Neural Network (NN) on a dataset collected by themselves that combines both visual and textual features for house price estimation. The conclusion was similar to the [4] with a 6% increase in the R2 score which can be explained with augmented visual data used along with textual data.

All approaches that exist either use MLR, SVR or MLP as NN to solve this regression problem. This proposal is an attempt to look at the regression from Convolutional Neural Network (CNN) [5] approach and offer a fresh perspective on the novelty of 1D CNN and it's capacity to tackle it.

III. Methodology

A. Dataset

The California Housing dataset consists of data collected from the California 1990 Census. The dataset has 20640 examples of blocks groups, each containing on average 1425,5 individuals living in a geographically compact area. There are 10 different variables per instance. The model uses 8 features amongst them for prediction. Figure 1 show the variation of respective features for the first 500 instances from the dataset.

A block group typically has a population of 600 to 3,000 people. Most districts have around 100–500 households with a maximum around 4800. Majority of the houses in each district are more than 50 years old. Latitude and longitude define the geographical area. Most districts have bedrooms between

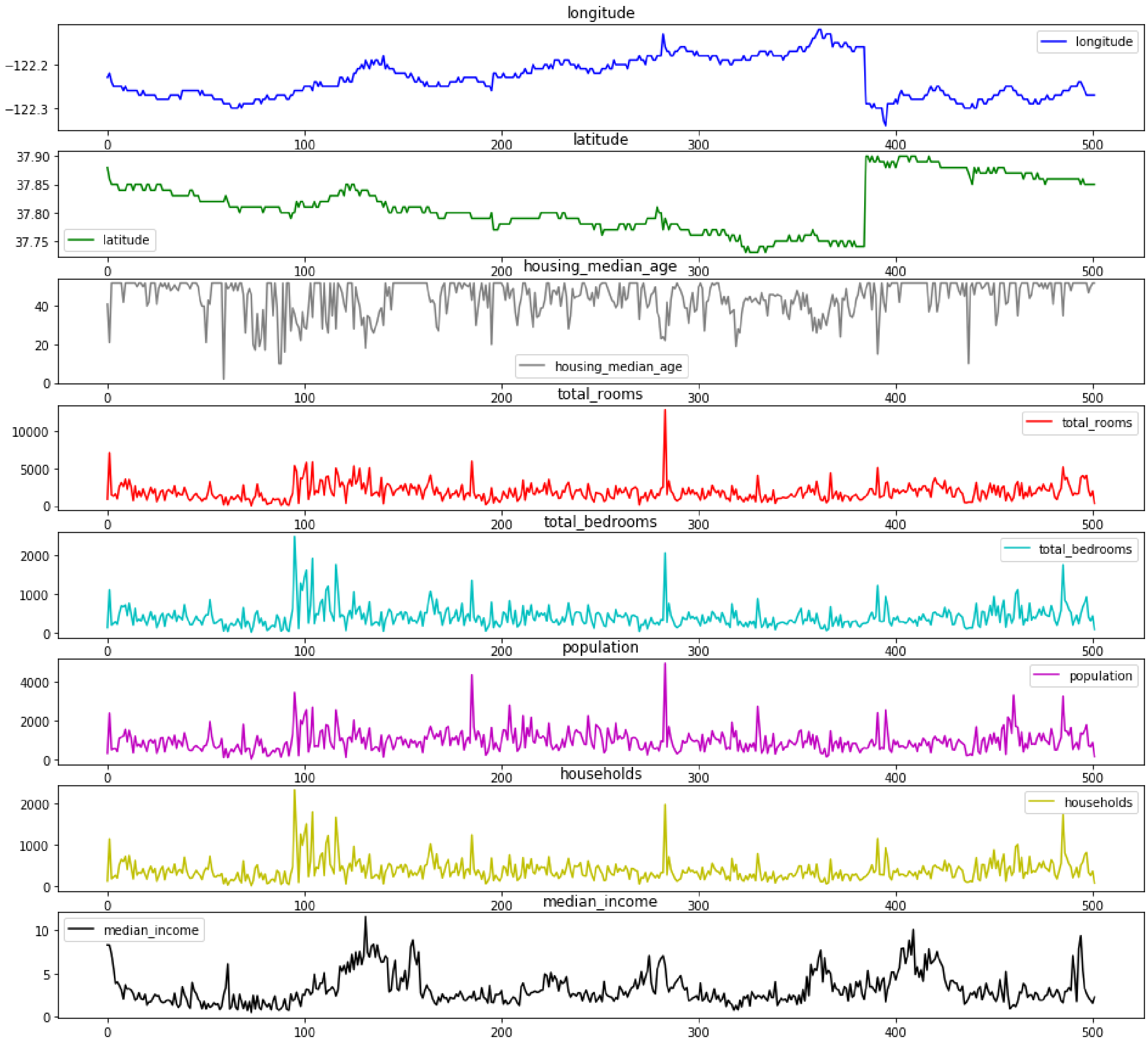

Fig. 1. Variation of respective features.

300–600. total rooms in the district are around 3000. Median-income was directly correlated with the house-value and had a Gaussian distribution unlike other features in the dataset. The all the missing data is removed from the dataset in the preprocessing.

B. *Proposed Network Architecture*

A CNN recognises simple patterns within the data, which can then be used in higher layers to shape more complex patterns. When the position of the feature within the segment is not of great significance, a 1D CNN is powerful in extracting interesting features of the dataset from its shorter segments. It can be applied in the analysis of time sequences of sensor data and Natural Language Processing.

- **Input Layer:** The input data has been preprocessed and doesn't contain any missing values that can affect the prediction model. Every instance of the dataset has eight features, hence the input is 8 to provide a holistic impression per instance. The first layer in the network must reshape it to a single dimension.
- **Batch Normalisation:** Batch normalization is a technique for training very deep neural networks that standardizes the inputs to a layer for each mini-batch. This has the effect of stabilizing the learning process and dramatically reducing the number of training epochs

Models	Training L1 Loss	Traning R score	Testing L1 Loss	Testing R score
Kamis's Model	1.293 E9	0.9039	1.713 E9	0.87392
Emans's Model	5.9223 E6	0.9653	9.708 E6	0.9348
Proposed Model	5.1961 E4	0.6369	5.195 E4	0.6095
Linear Regression	-	0.651	-	0.629
Random Forest	-	0.813	-	0.732

TABLE I
COMPARISON BETWEEN DIFFERENT MODELS.

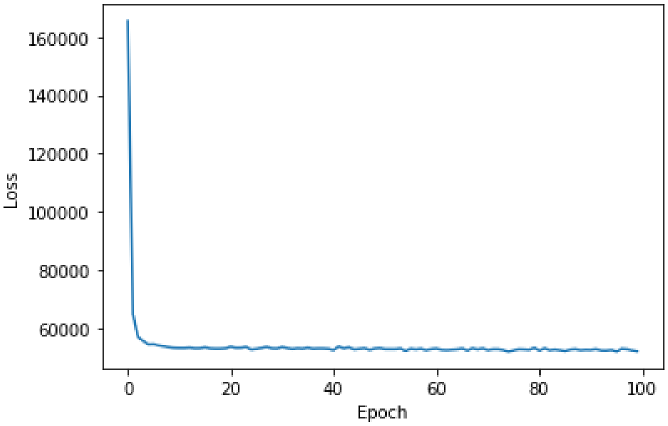

Fig. 2. LAD v/s Epoch

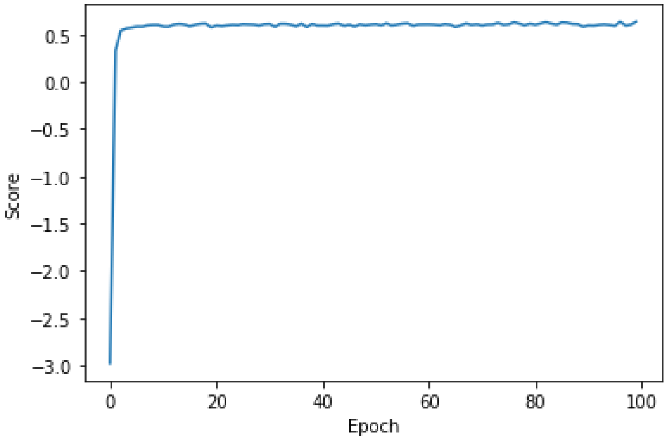

Fig. 3. R^2 vs Epoch

required to train deep networks.

- **1D CNN layer:** The first layer defines a kernel of height 1. One filter would allow the neural network to learn a single feature in the first layer which is not adequate, hence 64 filters are defined to detect 64 features.
- **Pooling layer:** A pooling layer is usually used after a CNN layer in order to subsample the input to provide a simpler output that helps prevent overfitting of the data. The proposed model uses max-pooling layer after two consecutive CNN layers followed by an average-pooling layer after two consecutive CNN layers.
- **Fully Connected Layers:** The FC layers use the flattened output from the CNN layers in the vector of height 64. The proposed model has three FC layers and the last FC layer is responsible for providing a single output since we have a single value to predict, the house price (i.e. regression).

A total of 25729 hyperparameters will be learning during training epochs for the proposed model.

GitHub link to the said model is https://github.com/NTsering/housing-price.git

IV. PERFORMANCE EVALUATION:

A. *Least Absolute Deviation*

L1 Loss function stands for Least Absolute Deviations, also known as LAD. L1 Loss Function is used to minimize the error which is the sum of the all the absolute differences between the true value and the predicted value. It is measured by:

$$LAD = \frac{1}{n}\sum_{i=1}^{n} \mid \widehat{y} - y \mid \tag{1}$$

where $\widehat{y}$ is the estimated value from the regression and y is the actual value. The smaller the LAD, better the estimation model.

B. *The coefficient of determination R^2*

The coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is calculated a set of various errors:

$$SSE = \sum_{i=1}^{n} (\widehat{y}_i - y_i)^2 \tag{2}$$

$$SST = \sum_{i=1}^{n} (\overline{y} - y_i)^2 \tag{3}$$

SSE is the Sum of Squares of Error and SST is the Sum of Squares Total. The R-squared value is calculated by:

$$R^2 = 1 - \frac{SST}{SSE} \tag{4}$$

The value of R ^{2} ranges between $-\infty$ and 1, the higher the value, the more accurate the estimation model.

V. Results

We compared the LAD and R^2 value in both training and testing and our model underperformed with respect to other popular models. The purpose of the paper was exploring the CNN model for regression and the performance indicates that CNN can be used with appropriate design choices. With proper hyperparameter turning, it is plausible to propose that the performance can be improved.

Figures 2 and 3 show that the performance of the model increases and the LAD decreases with successive training cycles.

The inference time for the model is 0.373s

The results tabulated in table 1 show that our model at this stage performs similar to a Linear Regressor (LR). The LR was implemented and tested for California Housing dataset. At the same time, Random Forest, though a slow, performed better than most. The [3]-and [4] model used a different dataset, and therefore, one-to one comparison is not meaningful.

VI. Conclusion

This paper provides a fresh approach to regression using 1D CNN for house price prediction. Further experiments and fine-tuning will improve the model incrementally. Additionally, it evinces the capacity of CNN in regression and tabular data analysis which is virtually confined to image and video-related applications.

References

[1] Limsombunc, V., Gan, C., and Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. American Journal of Applied Sciences, 1(3):193–201.

[2] Fletcher, M., Gallimore, P., and Mangan, J. (2000). The modelling of housing submarkets. Journal of Property Investment and Finance, 18(4):473–487.

[3] Eman H. Ahmed, Mohamed N. Moustafa. House price estimation from visual and textual features. arXiv:1609.08399v1 [cs.CV] 27 Sep 2016.

[4] Khamis and Kamarudin (2014). Comparative Study On Estimate House Price Using Statistical And Neural Network Model. International Journal Of Scientific Technology Research Volume 3, Issue 12, December, 2014.

[5] Yann LeCun, Leon Bottou, Yoshua Bengio and Patrick Hafner. Gradient Based Learning Applied to Document. Proc. of the IEEE, November 1998.