Draft Technical Paper Working Title

Elizabeth Gabel
Indiana University
eligabel@iu.edu

Alexander Mervar
Indiana University
amervar@iu.edu

Aidan Rosberg
Indiana University
arosberg@iu.edu

Abstract

This paper presents a neural network method for predicting the average home cost of a given neighborhood. It builds from previous work in the area, and refines neural parameters and feature space to increase accuracy, and compares performance to a linear regression model and support vector model. It serves to demonstrate the usefulness of neural networks in the problem space.

1 Introduction

The Californian housing market is impacted not only by the features of a given house, but by the unique features of the surrounding landscape. This can mean that buyers do not only look at the qualities of a given home of interest, but also the qualities of the neighborhood the homes reside in that may be unique to the coastal state. While previous papers have focused on individual home costs and features, this paper seeks to demonstrate the efficacy of a neural network in predicting the average cost of homes in a neighborhood given the neighborhood's proximity to the ocean, median income, population, and households, along with the average age and home features of the homes themselves.

Previous work examines housing market prediction with both pre-neural and neural methods. In pre-neural methods, regression models such as linear regression, support vector models (SVMs), kohonen neural networks (KNNs), and random forest models have all been used to various degrees of success. Pow et al. (**?**) demonstrated that KNNs and random forest models outperform baseline linear regression and SVMs, and speculate this is likely due to ability to consider a higher vector space and draw connections beyond a linear plane. Later studies, such as Ćetković et al (**?**), examined the efficacy of neural network methods for market prediction, and found the results reasonable enough to continue refinement of parameters and further development of neural network methods in this field.

We seek to continue investigation into how best to harness the processing abilities of neural networks to solve the problem. This paper demonstrates that our neural network outperforms linear regression models, and shows how to refine the parameters of the neural network for best results.

2 Dataset

The dataset that we have selected is a selection of house data used in the second chapter of Aurélien Géron's book 'Hands-On Machine learning with Scikit-Learn and TensorFlow' (**?**). The data contains information from the 1990 California census. Although this not necessarily mean that our model can predict current housing prices, due to the volatility of the current economy due to the COVID-19 pandemic and the nature of this project, it was decided that applying this methodology to current datasets would be outside the scope of this paper.

2.1 Preprocessing Methods

While the dataset we selected had already been preprocessed to an extent, we still took steps to ensure the efficacy of our models. The original dataset contained categorical values as a way to express the distance of the given neighborhood from the coast. We applied one-hot encoding to extract these values to individual features. One-hot encoding was chosen as a way to reduce bias introduced by other methods (Original encoding, Hashing). Another method employed to reduce bias was standardization. Because the ranges between the numerical values in the original data were high we believed our neural model performance could be improved by scaling the values. We also applied mean imputation to estimate missing data.

3 Methods

3.1 Linear Regression Model

To test the validity of our neural network is was essential to create a linear regression to act as the baseline for comparison. By using Python packages pandas and statsmodels, we were able to fit the variables of the dataset to the "median_house_value" variable using a standard linear regression method. The following coefficient table was generated.

variable	coef	std err	t	P>|t|	[0.025	0.975]
longitude	-0.4594	0.018	-26.092	0.000	-0.494	-0.425
latitude	-0.4664	0.019	-25.200	0.000	-0.503	-0.430
housing_median_age	0.1154	0.005	24.205	0.000	0.106	0.125
total_rooms	-0.0902	0.015	-6.186	0.000	-0.119	-0.062
total_bedrooms	0.2626	0.022	12.080	0.000	0.220	0.305
population	-0.3853	0.010	-36.896	0.000	-0.406	-0.365
households	0.2549	0.022	11.490	0.000	0.211	0.298
median_income	0.6384	0.005	116.635	0.000	0.628	0.649
ocean_proximity_1H OCEAN	0.1072	0.007	14.790	0.000	0.093	0.121
ocean_proximity_INLAND	-0.2370	0.011	-21.579	0.000	-0.259	-0.216
ocean_proximity_ISLAND	1.4594	0.267	5.473	0.000	0.937	1.982
ocean_proximity_NEAR BAY	0.0752	0.015	5.048	0.000	0.046	0.104
ocean_proximity_NEAR OCEAN	0.1483	0.013	11.488	0.000	0.123	0.174

3.2 SVMs????

3.2.1 Neural Network

4 Results

4.1 Linear Regression Results

4.2 SVM????? Results

4.3 Neural Network Results

5 Discussion

6 Conclusion

Acknowledgements