Using Neural Networks To Predict the California Housing Market

Alexander Mervar, Aidan Rosberg

Abstract

This research 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 linear regression and support vector models. It serves to demonstrate the usefulness of neural networks in the problem space

Question: How can we use a neural network to predict the housing market within California?

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. (2014) demonstrated that KNNs and random forest models outperform baseline linear regression and SVMs, and speculate this is likely due to the ability to consider a higher vector space and draw connections beyond a linear plane. Later studies, such as Ćetković et al (2018), 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.

Methodology

Pre-Processing

The original dataset contained categorical values 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 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.

Linear Regression

Linear regression is a statistical method that is often used to model the relationship between a dependent variable and one or more independent variables. In the context of the housing market, linear regression could be used to predict the value of a house based on factors such as its location, size, age, and other characteristics.

To test the validity of our neural network it 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.

Neural Network

We created our artificial neural network with the Python package TensorFlow. Our model consists of three 64 node hidden layers a dropout layer to prevent over fitting, and we use Adam as our gradient decent optimizer. We experimented with multiple types of neural networks including convolutional and recurrent neural networks, but our results had shown a conventional feed-forward artificial neural network provided the best results.

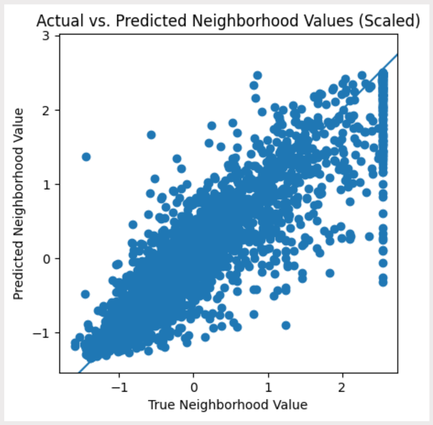
Results

Linear Regression Results

Coefficient Table

variable	coef	std err	t	P>|t|	[0.025	0.975]
longitude	-0.4537	0.042	-10.749	0.000	-0.536	-0.371
latitude	-0.4610	0.045	-10.247	0.000	-0.549	-0.373
housing_median_age	0.1135	0.012	9.657	0.000	0.090	0.137
total_rooms	-0.1826	0.038	-4.758	0.000	-0.258	-0.107
total_bedrooms	0.3079	0.055	5.572	0.000	0.200	0.416
population	-0.4674	0.028	-16.558	0.000	-0.523	-0.412
households	0.3849	0.057	6.770	0.000	0.273	0.496
median_income	0.6684	0.014	48.123	0.000	0.641	0.696
ocean_proximity_<1H OCEAN	0.1115	0.018	6.110	0.000	0.076	0.147
ocean_proximity_INLAND	-0.2170	0.027	-8.092	0.000	-0.270	-0.164
ocean_proximity_ISLAND	0.4481	0.574	0.781	0.435	-0.677	1.573
ocean_proximity_NEAR BAY	0.0405	0.038	1.079	0.281	-0.033	0.114
ocean_proximity_NEAR OCEAN	0.1453	0.034	4.287	0.000	0.079	0.212

Neural Network Results



We created our artificial neural network with the Python package TensorFlow. Our model consists of three 64 node hidden layers a dropout layer to prevent over fitting, and we use Adam as our gradient decent optimizer.
We experimented with multiple types of neural networks including convolutional and recurrent neural networks, but our results had shown a conventional feed-forward artificial neural network provided the best results.

Conclusion

When running the two models against one an- other, the linear regression scored an R2 score of 0.6738328671015846 and the neural network scored an R2 score of 0.816898481006914.

An R2 score, also known as the coefficient of determination, is a measure of the goodness of fit of a model, where a value of 0 indicates that the model does not explain any of the variance in the data, and a value of 1 indicates that the model perfectly explains the variance in the data. A linear regression model with an R2 score of 0.67 means that the model explains about 67% of the variance in the data. An R2 score of 0.82 for a neural network indicates that the model explains about 82% of the variance in the data.

Overall, while both a linear regression model and a neural network can be used for predictive modeling, a neural network generally has a higher capacity to model complex relationships in the data and make more accurate predictions, as demon- strated by its higher R2 score.

The use of a neural network over a linear regression offers several key advantages, which has been illustrated by the application here. It can model complex, non-linear relationships between variables, handle large amounts of data, and improve over time. These advantages make neural networks a valuable tool for making accurate predictions in the California housing market.

Acknowledgements

Jasmina Cetkovic, Slobodan Laki ´c, Marijana Lazarevska, Miloš Žarkovi ´c, Saša Vujoševi ´c, Jelena Cvijovi ´c, and Mladen Gogi ´c. 2018. Assessment of the real estate market value in the european market by artificial neural networks application. Complexity, 2018:1–10.

Aurélien Géron. 2017. Hands-on machine learning with scikit-learn and tensorflow. Tools, and Techniques to build intelligent systems.

Nissan Pow, Emil Janulewicz, and L. Liu. 2014. Applied machine learning project 4 prediction of real estate property prices in montréal.