1. Introduction

A house value is simply more than location and square footage. Like the features that make up a person, we would want to know all aspects that give a house its value. This project is going to be focused on solving the problem of predicting house prices for house buyers and house sellers.

In the project, we will evaluate the performance and predictive power of models that has been trained and tested on data collected from houses in suburbs of Boston, Massachusetts. Models trained on this data that is seen as a good fit could then be used to make certain predictions about a house – in particular, its monetary value. The models would prove to be invaluable for someone like a real estate agent who could make use of such information on a daily basis, or someone like you and me who would like to find their dream home with a reasonable price tag.

If you're studying data science, you probably heard of the Boston housing dataset, which is a very ubiquitous dataset. This dataset contains information about 506 census tracts of Boston from the 1970 census. And each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts, including average number of rooms per dwelling, pupil-teacher ratio, per capita crime rate and so on.

Like 'hello world', the Boston Housing Dataset has become part of a common vocabulary. And it will remain so, not only because thoroughly labeled datasets for machine learning are still not that easy to find, but because using the same dataset for decades to test different algorithms has allowed scientists to control for that variable and highlight the differences in algorithm performance. However, the Boston housing dataset is small, especially in today's age of big data. If the amount of data is too small, the model is likely to appear overfitting, resulting in poor model performance. In order to reduce the impact of this problem on the training model, we intend to optimize the data preprocessing and model selection by, for example, reducing the number of features, cross-validation, and reducing the complexity of the model.

We are going to break everything into logical steps that allow us to ensure the cleanest, most realistic data for our model to make accurate predictions from:

1. Load Data and Packages
2. Preprocessing of Dataset
   1.1.Analyzing
   1.2.Impute Missing Data and Clean Data
   1.3.Feature Transformation/Engineering
3. Modeling and Predictions
   1.1.Linear Regression
   1.2.Neural Network
   1.3.Random Forest
   1.4.Gradient Boost

Models

Neural Network

3. Algorithm

Neural Networks (Like the network shown in Figure 1) are a set of algorithms, modeled loosely on the human brain, a neural net consists of thousands or even millions of simple processing nodes that are densely interconnected. Most of today's neural nets are organized into layers of nodes, and they're "feed-forward," meaning that data moves through them in only one direction. An individual node might be connected to several nodes in the layer beneath it, from which it receives data, and several nodes in the layer above it, to which it sends data.

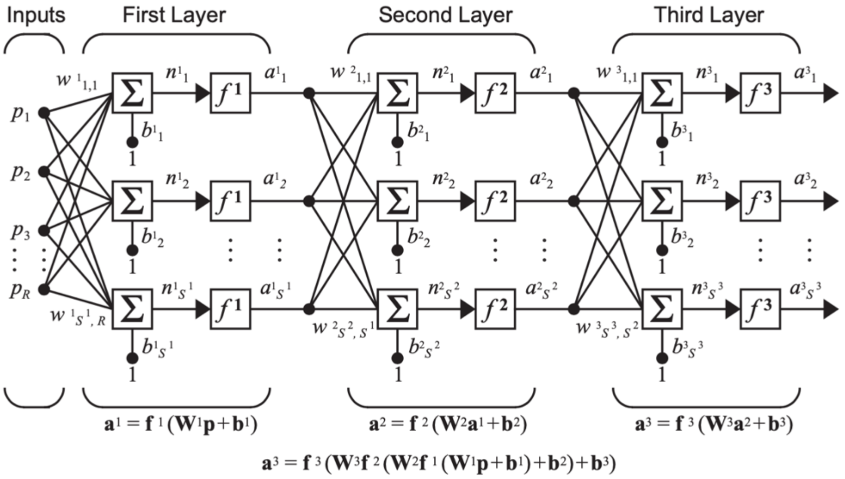

Figure 1. A Three-Layer Network

To each of its incoming connections, a node will assign a number known as a "weight." When the network is active, the node receives a different data item — a different number — over each of its connections and multiplies it by the associated weight. It then adds the resulting products together, yielding a single number, which is called net input, as the Equation 1showing.

$$n = \mathbf{W}\mathbf{p} + b \qquad \text{Equation 1}$$

Then, the neuron output can be calculated as Equation 2, depending on the particular transfer function that is chosen.

$$a = f(\mathbf{W}\mathbf{p} + b) \qquad \text{Equation 2}$$

When a neural net is being trained, all of its weights and biases are initially set to random values. Training data is fed to the bottom layer — the input layer — and it passes through the hidden layers, getting multiplied and added together in complex ways, until it finally arrives, radically transformed, at the output layer. During training, the weights and biases are continually adjusted until training data with the same labels consistently yield similar outputs.

4. Experimental Setup

For the case of the project, after preprocessing data, we selected ‘MEDV’ as target, and rest of variables as features. Then, ‘train_test_split()’ function is used to split data to train and test data. The parameter ‘random_state’ of this step in all models is set as 1 so that the results of all models are comparable.

Here, ‘LMPRegressor()’ is chosen to be used as Neural Network model in PyCharm. It contains many parameters like ‘activation’, ‘solver’, ‘max_iter’, ‘batch_size’, ‘early_stopping’ and so on.

To get the best parameters, a new python file is written to adjust the different parameters mentioned above. Here we use both ‘GridSearchCV’ and ‘RandomizedSearchCV’ methods to find the best parameters.

‘GridSearchCV’ is grid search and cross-validation. Grid search, search for parameters, that is, adjust the parameters in order according to the step size within the specified parameter range, use the adjusted parameters to train the learner, and find the parameters with the highest accuracy on the verification set from all the parameters, which is actually a cycle and comparison process. It can guarantee to find the most accurate parameters within the specified parameter range, but this is also the defect of grid search. It requires traversing all possible parameter combinations, which is very time-consuming in the face of large data sets and multiple parameters.

‘RandomizedSearchCV’ is used in the same way as ‘GridSearchCV’, but it replaces ‘GridSearchCV’'s grid search for parameters by randomly sampling in the parameter space. For parameters with continuous variables, ‘RandomizedSearchCV’ will sample it as a distribution. This is what grid search cannot do. Thus, we use ‘RandomizedSearchCV’ finally. Its search ability depends on the set ‘n_iter’ parameter.

Then, we train the model using the best parameters by the customized function ‘train_model()’ and get MSE and R^2 score, which are used to judge the performance of model.

Finally, the return value ‘loss_curve_’ and prediction are used to plot the results.

5. Results

In adjusting parameter part, we focus on ‘activation’, activation function for the hidden layer; ‘solver’, the solver for weight optimization; ‘batch_size’, size of minibatches for stochastic optimizers, which is useful when the model is not converge; ‘max_iter’, the maximum number of iterations, which is also an important parameter for convergence; ‘early_stopping’, whether to use early stopping to terminate training when validation score is not improving.

After using cross validation, we get the best parameters shown in Table 1.

Table 1
Best Parameters

Parameters	activation	solver	batch_size	max_iter	early_stopping
Value	relu	sgd	70	11000	True

Note: The number of sampling without replacement ‘n_iter’ in ‘RandomizedSearchCV’ is 350. Best score in ‘RandomizedSearchCV’ is 0.622.

Parameter ‘relu’ is the rectified linear unit function as Equation 3 expresses

$$f(x) = \max(0, x) \qquad \text{Equation 3}$$

The advantage of this activation is that the convergence rate of SGD obtained by using ReLU will be much faster than sigmoid/tanh. It can get the activation value without calculating a lot of complicated operations, compared to sigmoid/tanh.

Parameter ‘sgd’ refers to stochastic gradient descent. Since it is not a loss function on all training data, but a loss function on a certain training data in each iteration, so that the update speed of each round of parameters is greatly accelerated.

In results after training model part, we get MSE and R^2 score to see whether the model is good.

MSE is a risk function, which can be calculated by Equation 4, corresponding to the expected value of the squared error loss. MSE is almost always strictly positive. It is a measure of the quality of an estimator. The closer to zero the value is, the better the model is.

$$MSE = \frac{1}{n}\sum_{i}(y_i - y_{\text{pred}})^2 \qquad \text{Equation 4}$$

Here, MSE of neural network model is 10.352, which is close to zero, indicating that the model performs well.

R squared score is a statistical measure of how close the data pairs in a set are to their fitted regression line. This measure ranges from 0 to 1, indicating the extent to which the dependent variable in a data set is predictable. A R squared of 0 means that the dependent variable cannot be predicted by the independent variable, while 1 means that it can be predicted without error. Thus, the larger R squared score is, the more predictable the dependent variables are. The result of neural network model is 0.896, which also shows that the model performs well.

What’s more, two images are also improved to judge the performance.

Firstly, we plot the image of error and epochs (Figure 2). We can see that as the number of iterations increases, the mean square error of the model gradually decreases. When the number of iterations reaches about 150, the mean square error of the model has been reduced to 0. This indicates that the model has converged at this time, which is the same as the ‘n_iter’ of model.

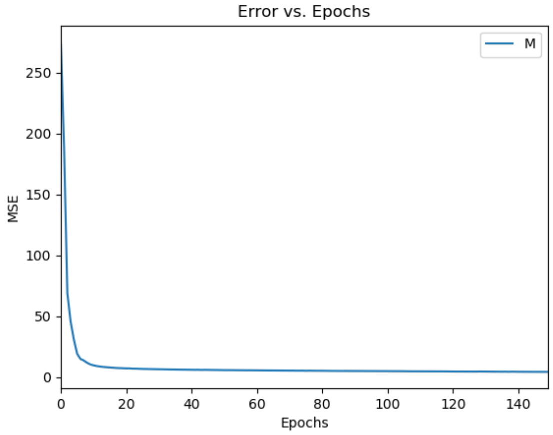

Figure 2. Error vs. Epochs

Secondly, we plot the fitted curve and true value curve (Figure 3). The blue line in Figure 3 represents real value of house price, while red line represents the fitted value predicted by neural network model. We can see that the red line and the blue line are roughly coincident, which directs that the performance of model is good and it can predict most house price correctly.

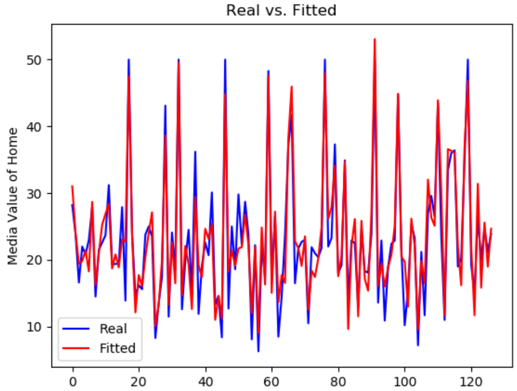

Figure 3. Real Value Curve vs. Fitted Curve

7. Reference

1. “A Beginner's Guide to Neural Networks and Deep Learning.” Pathmind, pathmind.com/wiki/neural-network#:~:text=Neural networks are a set,labeling or clustering raw input.
2. “API Reference¶.” Scikit, scikit-learn.org/stable/modules/classes.html.
3. Cantaro, M. (2020, January 19). Things You Didn't Know About the Boston Housing Dataset. Retrieved from https://towardsdatascience.com/things-you-didnt-know-about-the-boston-housing-dataset-2e87a6f960e8
4. “Code Faster with Line-of-Code Completions, Cloudless Processing.” Kite, kite.com/python/answers/how-to-calculate-r-squared-with-numpy-in-python.
5. Hagan, Martin T., et al. Neural Network Design. s. n., 2016.
6. Larry Hardesty | MIT News Office. “Explained: Neural Networks.” MIT News, 14 Apr. 2017, news.mit.edu/2017/explained-neural-networks-deep-learning-0414.
7. Naya, Gabriel. “Metrics and Python.” Medium, Towards Data Science, 26 Nov. 2019, towardsdatascience.com/metrics-and-python-850b60710e0c.
8. Samratp. (2018, December 07). Boston Housing Prices - Evaluation & Validation. Retrieved from https://www.kaggle.com/samratp/boston-housing-prices-evaluation-validation