Submitted to the Annals of Applied Statistics

FROM BLACK-BOX TO TRANSPARENCY: EXPLAINING REGRESSION MODELS FOR BOSTON HOUSING PRICES USING SHAP

BY DISEN LIAO[a],

University of Waterloo, [a]*d7liao@uwaterloo.ca*

Regression models are essential tools for predictive tasks, but there is often a trade-off between model complexity and interpretability. Complex models, such as Neural Networks (NN) and ensemble methods like Random Forest (RF), typically provide stronger predictive performance but are often difficult to interpret due to their intricate structures. In this study, we analyze the Boston Housing dataset, which comprises 506 entries and 13 features. We fit various regression models to the data and compare their performance in predicting housing prices. Besides, we use SHAP (SHapley Additive exPlanations) to measure feature importance and explain the models. Our aim is to elucidate the factors influencing housing prices and demonstrate how SHAP can enhance the interpretability of complex models.

1. Introduction. Housing prices are a significant concern for many people. A model that can accurately predict housing prices is highly desirable, but the utility of such a model extends beyond mere prediction. In addition to forecasting prices, regression models can be used to infer the importance of various features. This allows us to gain a deeper understanding of the factors influencing housing prices and the mechanisms by which these prices are determined.

In this study, we analyze the Boston Housing dataset, containing information about housing in the Boston area by Harrison (1978) . We fit various regression models to the data and evaluate the performance of each model using two metrics: Mean Squared Error (MSE) and R-squared (R^2). We will demonstrate that complex models, such as multi-layer perceptron (MLP), exhibit superior predictive power compared to traditional machine learning models.

Furthermore, we utilize these models to identify important features contributing to housing price predictions, which is essential for understanding the model. Many related works have addressed feature importance across different models. Altmann et al. (2010) used the permutation of features in a fitted model to observe the resulting degradation of the model's performance to evaluate feature importance. Gini importance, as described by Breiman (2001), explains Random Forest models by using a splitting function called "Gini impurity" to determine which attribute to split on during the tree learning phase. A feature's Gini importance is defined as the mean decrease in impurity from parent to children over all nodes where the specific variable is used for splitting.

For linear models, if the coefficients are on the same scale, they can also be interpreted as feature importance. Based on this, Muthukrishnan (2016) demonstrated how to use LASSO regression to identify important features, which remain non-zero under the l_1 penalty on the coefficients.

However, traditional model explanation approaches are challenging for complex models like Neural Networks due to their large scale. Shapley values, proposed by Shapley et al. (1953), originated from game theory, where they define the marginal contribution that a player has to a game. These values have become a popular tool in the field of explainable Machine Learning. SHAP (SHapley Additive exPlanations) by Lundberg et al. (2017) is an efficient approximation technique that determines the expected marginal contribution of a feature to any feature set not containing the feature. This method can be universally used to interpret any model. We will show an example of using SHAP to explain the Neural Network and compare feature importance given by SHAP with other methods.

2. Dataset. The Boston Housing Dataset originates from data collected by the U.S. Census Service concerning housing in the Boston, Massachusetts area. This dataset was developed to illustrate the challenges associated with using housing market data to measure consumer willingness to pay for clean air. It comprises 506 observations and 13 attributes. The response variable is the median value of owner-occupied homes (MEDV). For detailed descriptions of the explanatory variables, we refer readers to the dataset documentation.

Before fitting the models, we performed a log-transformation to normalize the distribution of all features and scaled them to the range 0-1 to ensure comparability. In our analysis, we calculated the correlation between MEDV and the other explanatory variables, as shown in Figure 1. The results indicate that the percentage of lower-status population (LSTAT) and the number of rooms (RM) are highly correlated with housing prices. Consequently, we anticipate that LSTAT and RM will be significant predictors in our modeling efforts.

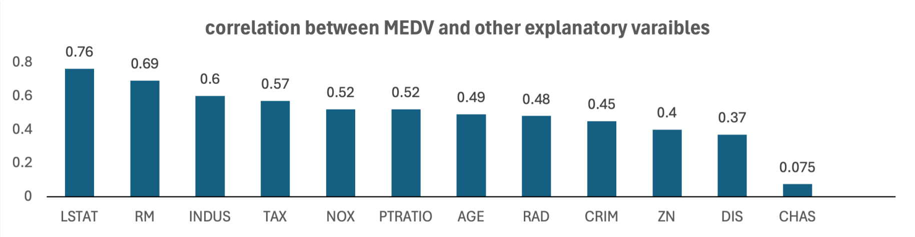


Fig 1: Correlation between MEDV and other explanatory variables, LSTAT and RM are highly correlated with MEDV.

3. Methods. We outline what models we have fitted, and how we evaluated model performance as well as the feature importances.

3.1. *Model Fitting.* In this section, we describe the methodologies used for fitting various regression models to the Boston Housing dataset. We cover Linear Regression, Ridge Regression, Lasso Regression, Random Forest, and Neural Network. For each model, a grid search method was implemented to find the best hyper-parameters when fitting the data.

3.1.1. *Linear Regression.* Linear regression models the relationship between a dependent variable $\mathbf{y}$ and one or more independent variables $\mathbf{X}$. The model is given by:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where β represents the coefficients to be estimated and ϵ is the error term. The ordinary least squares (OLS) method is used to estimate β by minimizing the sum of squared residuals:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

3.1.2. *Ridge Regression.* Ridge regression, also known as Tikhonov regularization, addresses multicollinearity by adding an $l2$ penalty to the loss function. The model is given by:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

The coefficients are estimated by minimizing the penalized sum of squared residuals:

$$\hat{\beta} = \arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \right)$$

where λ is the regularization parameter controlling the strength of the penalty.

3.1.3. *Lasso Regression.* Lasso regression introduces an $l1$ penalty, which encourages sparsity in the coefficient estimates. The model is:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

The coefficients are estimated by minimizing the sum of squared residuals with an $l1$ penalty:

$$\hat{\beta} = \arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right)$$

where $\|\beta\|_1 = \sum_j |\beta_j|$ and λ is the regularization parameter.

3.1.4. *Random Forest.* Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees. The model reduces variance and improves predictive accuracy. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, the Random Forest algorithm can be summarized as:

1. Draw B bootstrap samples from the original dataset.
2. For each bootstrap sample, grow a regression tree by recursively splitting the data on the best feature selected from a random subset of features.
3. Aggregate the predictions of all trees by averaging them to form the final regression prediction.

3.1.5. *Multi-layer Perceptron (MLP).* The Multi-layer Perceptron (MLP) is a fundamental type of neural network inspired by the architecture of the human brain, consisting of interconnected layers of neurons. For regression tasks, the network aims to minimize prediction error by adjusting the weights of the connections through a process called backpropagation. A typical neural network with one hidden layer can be expressed as:

$$\hat{\mathbf{y}} = f(W_2 \sigma(W_1 \mathbf{X} + b_1) + b_2)$$

where $\mathbf{X}$ is the input data, W_1 and W_2 are weight matrices, b_1 and b_2 are bias vectors, σ is the activation function, and f is the output activation function (identity for regression). In our implementation, we used two hidden layers, matrices $W_1 \in \mathbb{R}^{12\times 64}$ and $W_2 \in \mathbb{R}^{64\times 128}$ map the input vector into a 128-dimensional embedding space, $W_3 \in \mathbb{R}^{128\times 1}$ projects this embedding back to a single real number representing the predicted housing price. The weights are updated using backpropagation to minimize the loss function:

$$L(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{2n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ is the predicted value and y_i is the actual value for the i-th instance, and n is the number of instances.

3.2. *Model Evaluation.* To evaluate the performance of the regression models, we use two metrics: Mean Squared Error (MSE) and R-squared (R^2). MSE measures the average of the squares of the errors, which are the differences between the observed and predicted values. The MSE is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i represents the observed values, $\hat{y}_i$ represents the predicted values, and n is the number of observations. Lower MSE values indicate better model performance. R^2 is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

TABLE 1

Comparison of different models based on MSE and R²

Model	MSE	MSE std	R^2	R^2 std
Linear	24.23	13.71	0.54	0.17
RF	21.39	11.25	0.57	0.19
Ridge	24.13	13.78	0.54	0.17
Lasso	31.07	17.41	0.42	0.25
MLP*	12.61	1.67	0.84	0.01

where $\bar{y}$ is the mean of the observed values y_i. R^2 values range from 0 to 1, with higher values indicating better model performance. An R^2 value of 1 indicates that the model explains all the variability of the response data around its mean.

We estimate the generalization MSE and R^2 using 5-fold cross-validation. Cross-validation is a technique that divides training data into folds to assess the generalizability of a model. We run each model 10 times with different random seeds (0-9) to ensure the robustness of our results.

3.3. *SHAP Feature Importance.* SHAP (SHapley Additive exPlanations) provides a unified approach to measure feature importance based on Shapley values from game theory. Shapley values represent the contribution of each feature to the model's prediction by considering all possible combinations of features.

For a given model f and an input feature i, Shapley values are calculated as:

$$\phi_i(x) = \sum_{S \subseteq \{1,\ldots,n\} \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \left(f_{S\cup\{i\}}(x) - f_S(x) \right),$$

where S is a subset of all features excluding i, x is one observation of the data, and $f_S(x)$ represents the model prediction using features in S. $\phi_i(x)$ can only explain one data observation, to measure the importance of feature i, we average $\phi_i(x)$ on all $x \in \mathbf{X}$.

SHAP uses background data to estimate the expected marginal contribution of each feature efficiently. This method can be applied universally to interpret any model, providing insights into feature importance that are consistent across different model types. We will use SHAP to explain our fitted models and compare the feature importance results with those obtained from other methods.

4. Results. We present the results for model fitting and feature importance analysis.

4.1. *Model Fitting Results.* We present the evaluation results of fitted models in Table 1. The MLP outperforms other regression methods on this dataset, achieving an MSE of only 12.61, whereas all other models have an MSE exceeding 20. Additionally, the MLP regressor achieves an R^2 value of 0.84, while the R^2 values of the other models do not exceed 0.6. Moreover, the MLP demonstrates much smaller standard errors for both metrics compared to the other models, indicating the stability of the MLP model. The superior performance of the MLP can be attributed to its complexity. Our MLP model contains approximately 10,000 parameters, which is significantly more than the size of the dataset. This phenomenon is known as "over-parameterization". Allen-Zhu et al. (2019) provided a convergence theory for neural networks based on "over-parameterization" and explained why deep models often outperform traditional models. However, the complexity of MLP makes it very hard to explain, and in the next section, we will open the "black-box" using SHAP.

4.2. *Feature Importance Results.* To make the different importance scoring systems comparable, we normalized them so that the importance scores of all features sum to 1. SHAP provides estimated Shapley values for each feature, representing its importance. Figure 2 displays the SHAP feature importance for our MLP model. The MLP model assigns the highest importance to the index of accessibility to radial highways (RAD), followed by the number of rooms (RM) and the crime rate (CRIM). Meanwhile, the MLP model assigns relatively equal importance to several features, highlighting its complexity. In contrast, the Lasso model demonstrates sparsity by setting the coefficients of some features to zero.

We also applied SHAP to other models and compared the feature importance given by SHAP with other approaches. In Figure 3a, we present the top three important features of the Random Forest (RF) model, as determined by SHAP and Gini importance separately. In Figure 3b, we show the top three important features of Ridge regression, as determined by SHAP and the model coefficients. From these figures, it is evident that the feature importance rankings under different valuation systems are quite similar, differing only by minor numerical values. The ranking is nearly identical. However, the feature importance varies more significantly across different models. The RF model primarily considers the number of rooms (RM) and the percentage of lower status population (LSTAT), while giving other features very low importance. In contrast, the Ridge regression model, besides LSTAT, also considers the distance to Boston employment centers (DIS) and the concentration of nitric oxide (NOX).

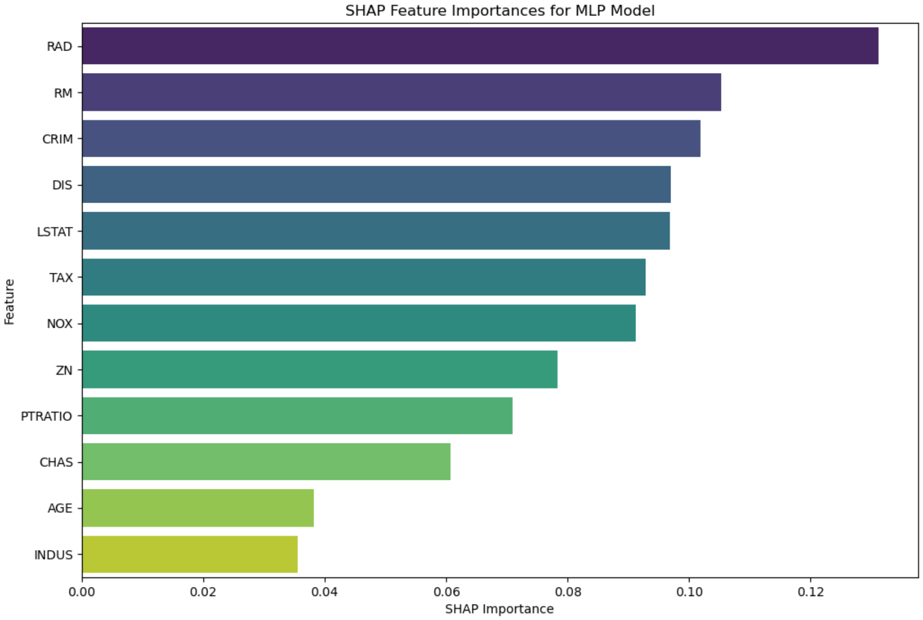
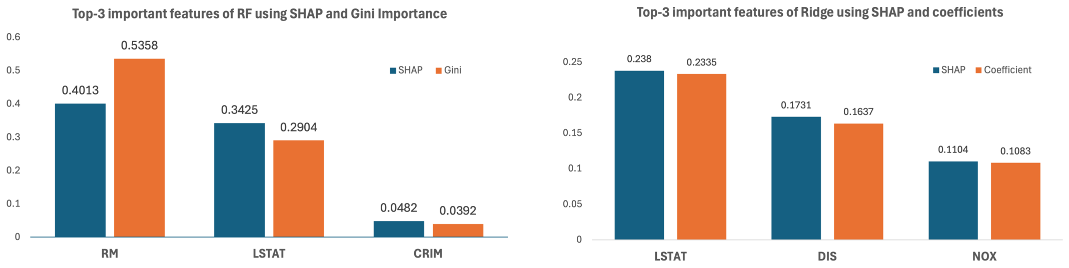


Fig 2: Feature importance of all features used by the MLP model as explained by SHAP. The results indicate that RAD has the highest feature importance. All features are assigned varying degrees of importance, highlighting the complexity and nuanced decision-making process of the MLP model.



(a) RF explained by SHAP and Gini importance

(b) Ridge explained by SHAP and coefficient

Fig 3: Comparison of feature importance for Random Forest (RF) and Ridge regression, explained by SHAP and their respective feature importance metrics. The SHAP values and traditional importance metrics show similar importance rankings within each model. However, the most important features identified differ significantly between the two models.

5. Conclusion. In this paper, we applied multiple regression models to the Boston housing dataset to analyze and predict housing prices. Among the models tested, the Multi-Layer Perceptron (MLP) significantly outperformed others, demonstrating the effectiveness of "over-parameterization" in capturing complex patterns.

To understand individual feature contributions, we employed SHAP (SHapley Additive exPlanations), starting with the MLP model. Interestingly, when SHAP was applied to other models, such as linear regression, ridge regression, and ensemble methods, it revealed a consistent pattern in feature importance rankings, indicating its robustness across different modeling approaches.

Different models evaluate feature importance in distinct ways. Linear regression models assign importance based on linear relationships, while tree-based models like random forests focus on features that reduce prediction error the most. Despite these differences, SHAP provides a unified framework that allows for a consistent comparison of feature importance across various models.

Across all models, certain features consistently emerged as key determinants of housing prices, including the number of rooms (RM), lower status of the population (LSTAT), crime rate (CRIM), and distance to employment centers (DIS). The prominence of these features suggests that both the physical characteristics of the houses (such as size, as indicated by the number of rooms) and locational factors (such as socio-economic status of the neighborhood, safety, and proximity to employment opportunities) are crucial determinants of housing prices.

References.

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song (2019). “A convergence theory for deep learning via over-parameterization”. In: *International conference on machine learning*. PMLR, pp. 242–252.

Altmann, André et al. (2010). “Permutation importance: a corrected feature importance measure”. In: *Bioinformatics* 26.10, pp. 1340–1347.

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45, pp. 5–32.

Harrison, D (1978). “Hedonic prices and the demand for clean air”. In: *Journal of environmental economics and management* 5, pp. 81–102.

Lundberg, Scott M and Su-In Lee (2017). “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30.

Muthukrishnan, Ramakrishnan and R Rohini (2016). “LASSO: A feature selection technique in predictive modeling for machine learning”. In: *2016 IEEE international conference on advances in computer applications (ICACA)*. Ieee, pp. 18–20.

Shapley, Lloyd S et al. (1953). “A value for n-person games”. In.