Nimita Gaggar
Berkeley, USA
+1 (510)-502-5253 , nimita_gaggar@berkeley.edu December 11, 2023

PREDICTING HOUSE PRICES USING MACHINE LEARNING

Executive Summary:

- **Objective**: This paper aims to address the challenges in accurately predicting house prices by employing advanced Machine Learning models on the Boston house dataset.

- **Methodology:** Three regression models—Linear Regression, Lasso Regression, and Regression Tree—are selected for their suitability in predicting house prices. Evaluation metrics such as R-Squared, MSE, AIC, BIC, MSPE, and Cross-Validation are utilized to measure model accuracy.

- **Dataset:** The study utilizes the Boston house dataset from the UCI Machine Learning Repository, featuring 14 attributes of homes, including crime rate, land zoning, and nitric oxides concentration.

- **Key Findings:**

❖ Linear Regression using all predictors reveals potential non-significant variables (Indus and Age).

❖ Lasso Regression outperforms Linear Regression across multiple metrics.

❖ Regression Trees provide a non-linear approach but need careful pruning for practicality.

❖ Conclusion: Lasso Regression with optimal lambda performs slightly better than Linear Regression and pruned Regression Tree. Model choice should consider interpretability and predictive accuracy.

Problem: The problem addressed in this research is the accurate prediction of house prices, considering the challenges of over- or under-evaluation in the real estate market. The question the study aims to answer is which regression model among Linear Regression, Lasso Regression, and Regression Tree is most effective in predicting future house prices based on the Boston house dataset.

Introduction: Precisely estimating property values remains crucial for homeowners, buyers, and real estate stakeholders. Challenges like over- or under-evaluation persist due to inadequate detection measures, making property cost assessment a formidable task. This paper addresses these issues through a comprehensive analysis using Machine Learning, a subset of Artificial Intelligence (AI). As technology advances, AI becomes integral to various fields, including real estate, where it aids in predicting property values. The study primarily focuses on Machine Learning within the realm of AI to create predictive models based on the Boston house dataset. Past studies on predicting house prices have employed diverse methodologies, techniques, and datasets. Research has explored the correlation between house prices and local amenities, street-based local areas, and macroeconomic factors. Model selection is a critical aspect of ML, and this paper focuses on three regression models—Linear Regression, Lasso Regression, and Regression Tree—applied to the Boston house dataset. The study emphasizes the importance of accurate model selection for achieving reliable predictions.

A. Model Selection: The study focuses on three regression models: Linear Regression, Lasso Regression, and Regression Tree. The choice of these models is motivated by their suitability for regression tasks and their potential to handle the complexities of predicting house prices.

I] **Linear Regression using all predictors:**

- Linear Regression is a fundamental and widely used model for predicting numerical outcomes. It assumes a linear relationship between the predictor variables and the target variable.
- All predictors in the dataset are utilized to train the model, offering a baseline for comparison.

II] **LASSO Variable Selection:**

- Lasso Regression is a variant of Linear Regression that includes a regularization term. It is effective in feature selection, encouraging sparsity in the model by penalizing irrelevant predictors.
- Lasso Regression is applied to the dataset to identify and emphasize the most influential predictors, potentially improving prediction accuracy.

III] **Regression Trees:**

- Regression Trees provide a non-linear approach to modeling relationships between predictors and the target variable. They can capture complex interactions and patterns in the data.
- Regression Trees are employed to explore non-linear relationships within the dataset, providing an alternative perspective on predicting house prices. Model accuracy is assessed by dividing the dataset into training and test sets, with 80% used for training and 20% for testing. Evaluation metrics help gauge the effectiveness of the selected models.

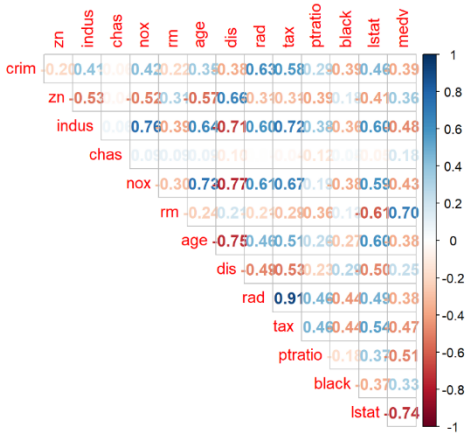
Evaluation metrics such as R-Squared, Adjusted r-Square, Mean Squared Error (MSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mean Squared Prediction Error (MSPE), and Cross-Validation are utilized to measure model accuracy. The study also explores attribute correlations in the Boston dataset through a heatmap, identifying influential factors. Outliers are addressed to enhance model accuracy, revealing Lasso Regression's superior performance across all metrics compared to Linear Regression.

Data Description: The dataset utilized originates from the UCI Machine Learning Repository, featuring housing values in Boston suburbs collected in 1978. With 506 entries detailing 14 attributes of homes, including crime rate, land zoning, and nitric oxides concentration, the dataset forms the basis for training Machine Learning models. The target variable, MEDV, represents the median value of owner-occupied homes, while other attributes contribute to model training. The statistics of these features offer insights into the dataset's composition, setting the foundation for robust predictive modeling. The rest of the variables are used for training the models. Statistics of the features are described in the table below:

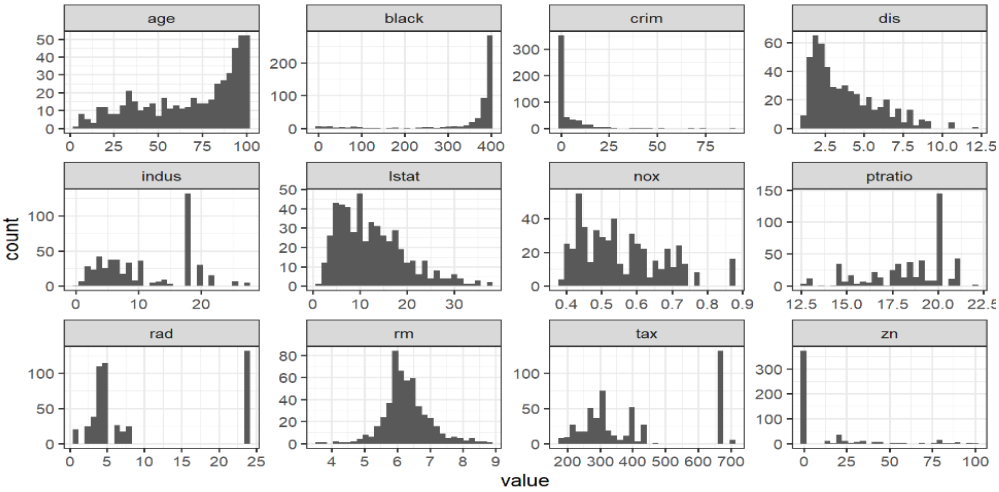
- No null values or duplicate rows are present in the dataset, ensuring data integrity.
- The 25th and 75th percentiles of the attribute "ZN" are both 0, indicating high skewness. This skewness is attributed to "ZN" being a conditional variable.
- The 25th, 50th, and 75th percentiles for the attribute "CHAS" are all 0, signifying high skewness. This skewness is a result of "CHAS" being a categorical variable with values 0 or 1.
- The maximum value of "MEDV" is 50.00, suggesting censorship at this value (corresponding to a median price of $50,000).
- The attributes “CRIM,” “ZN,” “RM,” and “B” are identified to have outliers.

S. No.	Attribute	Mean	Minimum	Maximum
1.	CRIM	3.61	0.006	88.97
2.	ZN	11.36	0.00	100.00
3.	INDUS	11.13	0.46	27.74
4.	CHAS	0.069	0.00	1.00
5.	NOX	0.55	0.385	0.871
6.	RM	6.284	3.56	8.78
7.	AGE	68.574	2.90	100.00
8.	DIS	3.795	1.129	12.126
9.	RAD	9.549	1.00	24.00
10.	TAX	408.23	187.00	711.00
12.	PTRATIO	18.455	12.60	22.00
13.	B	356.67	0.32	396.90
14.	LSTAT	12.65	1.73	37.97
15	MEDV	22.53	5.00	50.00

A heatmap is employed to assess the correlation between different attributes in the dataset (Diag1).



Observations (refer diagram below): * Proportion of owner-occupied units built prior to 1940 (age) and proportion of blacks by town (black) is heavily skewed to left, while per capita crime rate in town (crim) and weighted mean of distances to five Boston employment centres (dis) is heavily skewed to right. * rm is normally distributed with mean of approximately 6. * Most of the properties are situated close to the five Boston employment centres (dis skewed to right) * There is a high proportion of owner-occupied units built prior to 1940 (age skewed to left) and blacks in town (black skewed to right) * From scatter plots, it is seen that lstat and rm show strong correlation with medv. * 93% of the properties are away from Charles River. The properties bordering the river seems to have higher median prices.



Interpretation:

I] Linear Regression using all predictors: Indus and age have very high p-value and seem to be non-significant. The estimated coefficients are as follows:

```
Residuals:
    Min      1Q  Median      3Q     Max
-16.5396  -2.7188  -0.4474   1.9110  25.3554

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.080382   5.770019   6.253 1.06e-09 ***
crim         -0.106513   0.034771  -3.063 0.002341 **
zn            0.037695   0.015618   2.414 0.016255 *
indus        -0.029509   0.069130  -0.427 0.669717
chas          3.486921   0.994154   3.507 0.000505 ***
nox         -16.101065   4.230945  -3.806 0.000164 ***
rm            3.641921   0.489093   7.446 6.22e-13 ***
age           0.008683   0.015360   0.565 0.572185
dis          -1.389889   0.224246  -6.198 1.46e-09 ***
rad           0.280221   0.072862   3.846 0.000140 ***
tax          -0.009817   0.004088  -2.401 0.016809 *
ptratio      -0.940402   0.146160  -6.434 3.65e-10 ***
black         0.008994   0.002978   3.020 0.002692 **
lstat        -0.584341   0.058634  -9.966  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Building model with all variables

- Residual standard error: 4.784 on 390 degrees of freedom
- Multiple R-squared: 0.736,
- Adjusted R-squared: 0.7272
- F-statistic: 83.64 on 13 and 390 DF, p-value: < 2.2e-16

Building model without variables indus and age

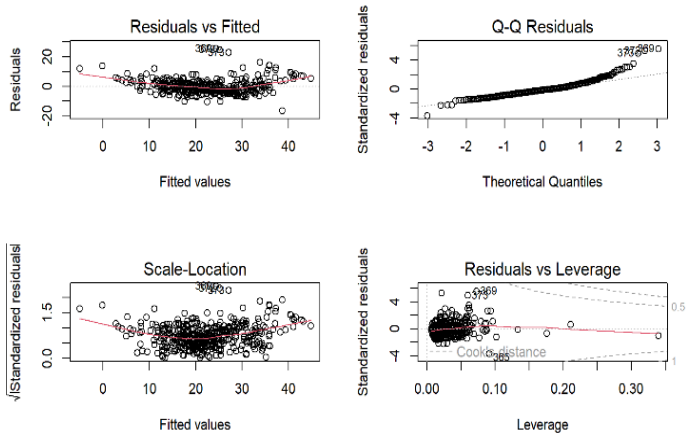
- Residual standard error: 4.774 on 392 degrees of freedom
- Multiple R-squared: 0.7357,Adjusted R-squared: 0.7283
- F-statistic: 99.18 on 11 and 392 DF, p-value: < 2.2e-16

Variable selection: Best subset (13 variable) and stepwise (forward, backward, both) techniques of variable selection were used to come up with the best linear regression model for the dependent variable medv.

> AIC(model.step) : [1] 2423.432 > BIC(model.step): [1] 2475.45

MODEL ASSESSMENT

- Residuals vs Fitted plot shows that the relationship between medv and predictors is not completely linear
- The variance is not completely constant and hence the assumption of constant variance is not totally satisfied
- From the q-q plot we see that it is not completely normal and a little skewed to the right.

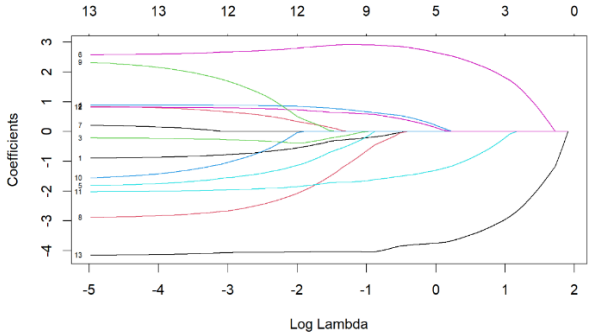


REGRESSION ANALYSIS

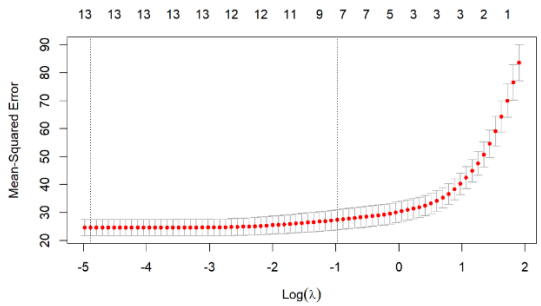
	MODEL 2	MODEL 1
Model equation	-age -indus	all
Model mse	22.79436	22.88217
R-squared	0.7356717	0.7360073
Adj r-squared	0.7282544	0.7272076
Aic	2423.432	2426.919
Mspe(out-sample)	21.69374	22.01144
Cross validation	23.58146	23.52317

- Based on AIC criteria and adjusted R square values, model 2 is slightly better than model 1. In-sample MSE is nearly the same for both models.
- We need to check out-of-sample MSPE for both models. Based on out-of-sample prediction error, model 2 is slightly better than model 1. MSPE of model 1 is 22.01144 while that of model 2 is 21.69374. Based on cross validation also, model 2 performs better.

II] LASSO Variable Selection: A. Now we use LASSO variable selection technique. Here lambda is the p enalty factor which helps in variable selection and so higher the lambda, lesser will be the significant vari ables included in the model. B. Using cross-validation we now find the appropriate lambda value using er ror versus lambda plot. We take the value with the least error as well as the error value which is one stand ard deviation away from the lowest error value. we then build models on the basis of both of these. For th e higher error value , the number of variables selected decreases.



A

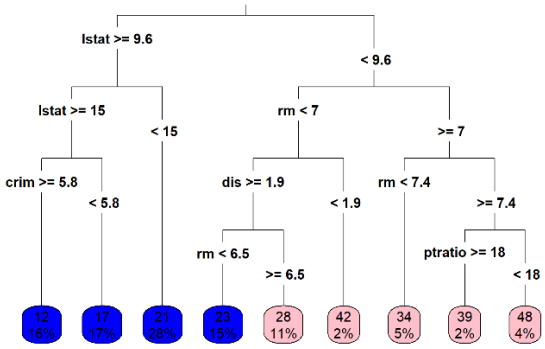


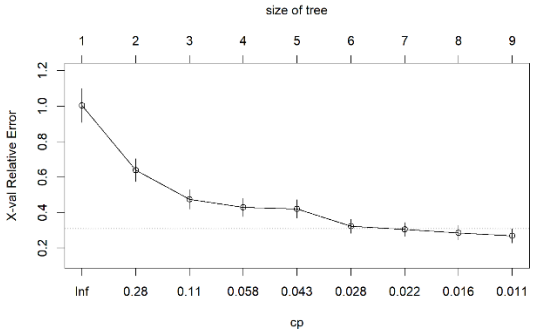
B

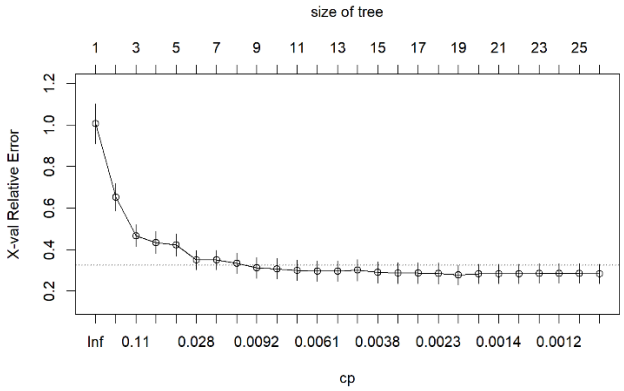
III] Regression Trees:

Following regression trees were fitted to the training data. * Using default value, cp = 0.01 and no additional constraints. This resulted in a tree with 8 terminal nodes. * Making cp = 0.001 and allowing the tree to grow large This results in a tree with 27 terminal nodes. A plot of cp values vs error rates (fig 7) shows that a cp value of 0.0072 would reduce the complexity of the model. * Finally, with cp = 0.0072. This results in a tree with 10 terminal nodes.

Variables actually used in tree construction:



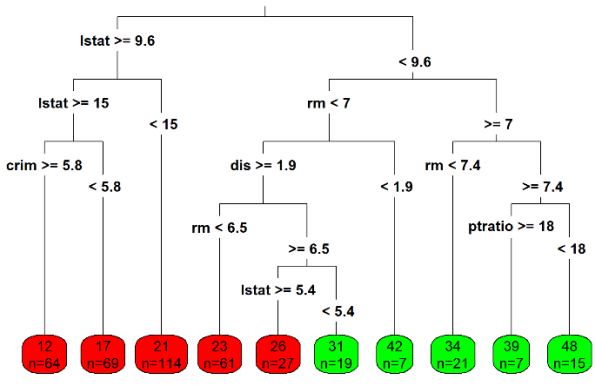




However, from plot cp, we observe that a tree with more than 7 to 9 splits is not very helpful.

Further pruning the tree to limit to 9 splits; corresponding cp value from plot is 0.0072.

From plot, we can observe that cross validation error does not always go down when tree becomes more complex. We can see that a tree with more than 7 to 9 splits is not very helpful.
The large tree results in lowest in-sample and out-of-sample prediction error compared to other two trees. But there is a large difference between in-sample and out-of-sample performance for this tree. In contrast, trees with default cp value and pruned tree have lower difference between in-sample and out-of-sample prediction errors. Further, they are easier to read and interpret and do not sacrifice much in terms of out-of-sample prediction error. Since pruned has lower prediction errors compared to tree with default cp value, it is the chosen tree.



RESULT:

	GLM (STEPWISE VARIABLE SELECTION)	LASSO REGRESSION (model.lasso.min)	REGRESSION TREE
MODEL EQUATION	-age -indus	- age -indus	crim + dis + lstat + ptratio + rm
MODEL MSE	22.79436	22.8858	14.8156
R-SQUARED	0.7356717	0.7359651	
ADJ R-SQUARED	0.7282544	0.727861	
AIC	2423.432		
MSPE(IN-SAMPLE)	22.03	22.02	12.69
MSPE(OUT-SAMPLE)	21.69374	22.032814	15.79491

CONCLUSION:

It is very important to predict house price accurately. A lot of time people pay overprice from the actual market price for a real estate property, similarly a lot of time sellers get very low price compared to the actual market price of the property. Not only people, but various estate agencies also face the same problem where they are not sure whether to invest toward a certain property or not. They are confused as they are not able to predict what the price of the house can be in future. The main purpose of this paper is to help people who are facing these issues to predict the house price in future years. In this paper an intelligent system is made using the Regressor models which are Linear Regression, Lasso Regression and Regression Tree on the Boston House Dataset to predict the house price. The study concludes that, based on multiple evaluation criteria, the Lasso Regression model with optimal lambda performs slightly better than Linear Regression and Regression Tree. Model choice depends on the balance between interpretability and predictive accuracy. The Regression Tree, particularly pruned for simplicity, provides a reasonable compromise.

This research provides insights into the strengths and limitations of different regression models for predicting house prices, helping stakeholders make informed decisions in real estate valuation. Further research could explore ensemble methods or advanced algorithms for improved predictive performance.

References:

- Holly, S., Pesaran, M. H., & Yamagata, T. (2010). A spatio-temporal model of house prices in the USA. Journal of Econometrics, 158(1), 160-173.
- Bahia, I. S. H. (2013). A data mining model by using ANN for predicting real estate market: Comparative study. International journal of intelligence science, 3(04), 162.
- Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. International Journal of Computer Applications, 115(9).
- Anandharajan, T. R. V., Hariharan, G. A., Vignajeth, K. K., & Jijendiran, R. (2016, January). Weather monitoring using artificial intelligence. In 2016 2nd international conference on computational intelligence and networks (CINE) (pp. 106-111). IEEE.
- Nadikattu, R. R. (2016). The emerging role of artificial intelligence in modern society. International Journal of Creative Research Thoughts.
- Bosamia, M. (2013). Positive and negative impacts of information and communication technology in our everyday life. Dostupno na: https://www. Research gate. net/publication/325570282_Positive_and_Negative_Imp acts_of_Information_and_Communication_Technology_in_our_Everyday_Life [30. kolovoza 2021.].